AI Journal of Business

Volume 1 Number 1 November 2024

A Study on Methods for Summarizing Information from YouTube Channels
Hyoung-sik Park

AI-Driven Stock Prediction and Investment Strategy Analysis for the Magnificent 7 Tech Giants
Jumin Lee

Study on the Effect of Data Augmentation on Image Learning
Tae-hwan Choi

The Prioritization of Evaluation Indicators for Artificial Intelligence (AI) Business Capabilities in the Financial Sector: Focusing on the Global AI Index
Suho Ahn, Tae Yeon Oh, Yongkyu Kim

Reducing Training Time and Enhancing Efficiency in Large Language Models: Ternary Output Quantization (TOQ)
Do-hoon Lee

A Study on the LLM-Based Software Architecture for Multi Robot Control
Sangho Choi

AI Business Academy
www.aiba.or.kr

9 773058 699007 02
ISSN 3058-6992

Contents

Chairman's Greetings ---------- 2

Papers

1. A Study on Methods for Summarizing Information from YouTube Channels ---------- 5
2. I-Driven Stock Prediction and Investment Strategy Analysis for the Magnificent 7 Tech Giants ---------- 14
3. Study on the Effect of Data Augmentation on Image Learning ---------- 26
4. The Prioritization of Evaluation Indicators for Artificial Intelligence Business Capabilities in the Financial Sector: Focusing on the Global AI Index ---------- 38
5. Reducing Training Time and Enhancing Efficiency in Large Language Models: Ternary Output Quantization(TOQ) ---------- 52
6. A Study on the LLM-Based Software Architecture for Multi Robot Control ---------- 60

Informantions

Articles of Association for “AI Business Academy” ---------- 71

“AI Business Academy” Research Ethics Regulations ---------- 79

Call for Papers of “AI Journal of Business” ---------- 85

Chairman's Greetings

Foreword

Welcome to the inaugural issue of *AI Journal of Business*, a platform dedicated to exploring the dynamic intersection of artificial intelligence (AI) and business. As the Founding President of AI Business Academy, the institution behind this journal, I am honored to present our first edition and share the vision that guides this publication.

We are living in an era where artificial intelligence is reshaping industries, redefining business models, and transforming decision-making processes. No longer a distant concept, AI has rapidly evolved into a vital force driving innovation across the global economy. From data-driven insights to automation, from personalized customer experiences to intelligent supply chains, the applications of AI are vast and continue to grow at an unprecedented pace.

However, amid this transformation, it has become clear that there is a need for a dedicated forum where business leaders, academics, and researchers can come together to share knowledge, ideas, and innovations that will drive the future of business. The *AI Journal of Business* was born from this need—a commitment to providing a premier platform for the exchange of groundbreaking research, case studies, and critical discussions around the role of AI in business.

Why the AI Journal of Business?

The *AI Journal of Business* was founded to fill a critical gap in academic and industry discourse. While much has been written about the technical aspects of AI, there remains a need for deeper exploration of its practical applications in business. Our mission is to bridge this divide by presenting research that is both rigorous and relevant, addressing not just theoretical advancements but also the real-world challenges and opportunities that AI presents for businesses.

As businesses worldwide face increasingly complex challenges, from global competition to rapidly changing markets, AI is emerging as a key solution.

However, the successful integration of AI into business strategies requires not just technical expertise but also a thorough understanding of its broader implications, including ethical, economic, and organizational considerations. This journal is designed to address these multi-dimensional aspects of AI in business, providing insights that are not only academic but also actionable.

Our Mission

The mission of the *AI Journal of Business* is to advance knowledge, foster innovation, and promote critical thinking in the realm of AI as it relates to business. We aim to do this through:

1. **Collaborative Research**: The journal will provide a platform for interdisciplinary collaboration. By bringing together experts in AI, business, economics, and management, we hope to create a rich dialogue that explores AI's impact across sectors and industries.
2. **Thought Leadership**: Beyond academic research, the journal will feature perspectives from industry leaders who are at the forefront of AI innovation. These contributors will share their experiences, successes, and lessons learned, offering invaluable insights into how AI is being used to transform business practices globally.
3. **Practical Insights**: Our goal is to ensure that the cutting-edge research published in this journal is accessible and useful for business leaders and policymakers. We will highlight case studies and real-world applications of AI that can serve as blueprints for others looking to implement AI-driven strategies in their organizations.

Embracing the Future Responsibly

While the opportunities presented by AI are immense, we must also acknowledge the responsibility that comes with its adoption. AI's potential to revolutionize business is clear, but it also raises important ethical, societal, and governance questions. As we move forward, it is crucial that AI development and implementation are approached with a sense of responsibility and foresight.

The *AI Journal of Business* will actively engage with these critical issues. Whether it is addressing concerns about data privacy, algorithmic bias, job displacement, or the broader social impacts of AI, this journal will serve as a

platform for thoughtful, informed debate. Our goal is to not only advance the use of AI in business but also to ensure that it is done in a way that benefits society as a whole.

A Collaborative Effort

The success of the *AI Journal of Business* would not have been possible without the efforts of many individuals. I would like to extend my heartfelt thanks to our editorial team, advisory board, contributors, and partners who have worked tirelessly to make this journal a reality. It is through their hard work and dedication that we are able to launch this platform today.

We are excited to embark on this journey and invite you, our readers, to join us in this important conversation. Whether you are a researcher, a business leader, a policymaker, or a student, the *AI Journal of Business* is a place for you to explore, contribute, and engage with the most pressing issues at the intersection of AI and business.

Conclusion

As we launch this journal, we do so with a clear purpose: to shape the future of AI in business. The *AI Journal of Business* represents not just a collection of research papers, but a vision for a world where AI drives meaningful, responsible, and lasting change in the business landscape. We look forward to the journey ahead and to the many exciting discoveries that await us.

Thank you for joining us.

Dongsung Cho
Founding Chairman, AI Business Academy
Chair Professor, aSSIST University
Professor Emeritus, Seoul National University

A Study on Methods for Summarizing Information from YouTube Channels

Hyoung-Sik Park

ABSTRACT

This study explored the possibility of summarizing information from YouTube channels. To achieve this, the research compared and analyzed speech-to-text (STT) data from top-viewed videos and comment data from recent videos. The keyword network analysis revealed that comment data could represent the main content of the videos, thereby validating the feasibility of information summarization using only comments.
To enhance the reliability of the study, text similarity analysis and statistical verification were conducted. The results of this research are expected to contribute to the utilization of influencer marketing by businesses and to help individual subscribers quickly grasp the main content of channels. Furthermore, it is anticipated that this study will make a practical contribution to the development of efficient processing and analysis methodologies for digital media data.

☞ keyword : YouTube Channels, Information Summarization, STT (Speech-to-Text), Keyword Network, Similarity Analysis, Text Analysis, Comment Analysis, Influencer Marketing

1. Introduction

Digital media, particularly social media platforms like YouTube, have established themselves as a crucial means of information delivery and marketing tools in modern society. The diverse content on YouTube channels significantly influences the opinion formation, consumption patterns, and cultural trends of millions of users. Especially with the growing interest in healthcare and health-related content, such information plays a vital role in users' health-related decision-making (Mohamed & Shoufan, 2024). However, the vast amount of content generated on YouTube channels is reducing the efficiency of information delivery. While the need for effective video summarization methods has emerged, existing research has mainly focused on summarizing individual videos or converting audio information to text, with limited research on methods to summarize information across entire channels.

The purpose of this study is to present and evaluate the validity of a method for comprehensively summarizing information from large amounts of video and comment data generated on YouTube channels. Specifically, for YouTube channels providing health and dietary supplement-related content in Korean, this study explores the possibility of summarization using only comments by comparing and analyzing speech-to-text (STT) data from top-viewed videos with recent video comments.

This research moves beyond fragmentary video summarization to present a new methodology for understanding the characteristics of entire channels, which can help businesses analyze YouTube content more quickly and develop influencer marketing strategies. Additionally, it provides ways for individual users to effectively grasp the core content of subscribed channels and offers foundational data that can be used in developing influencer marketing and personalized content recommendation systems. This study is expected to make a practical contribution to the development of efficient processing and analysis methodologies for digital media data.

2. Theoretical Background

2.1 Previous Research

Existing YouTube summarization research has primarily followed three approaches: summarizing videos themselves, summarizing captions uploaded by content creators, and extracting audio information to convert it to text before applying summarization techniques. The video summarization method typically involves condensing videos longer than an hour into approximately 10-minute versions. However, this approach has limitations in terms of information consistency and efficiency from a channel perspective, where multiple videos with similar content need to be quickly and effectively summarized, as it provides summaries through different views of video and text.

Previous research on text summarization of YouTube videos has utilized two methods commonly used in literature information and text sentence summarization: Extractive Summarization and Abstractive Summarization. Extractive summarization identifies and extracts important sentences or phrases from the original text, while abstractive summarization understands the meaning of the original text and generates new sentences to create summaries. While earlier studies primarily used extractive summarization methods, recent research has adopted abstractive summarization approaches using artificial neural network-based models, particularly Transformer models, to generate more natural and human-readable summaries. Recently, information summarization using Large Language Models (LLMs) based on Transformer architecture has seen significant growth.

In video summarization research, Kaiyang Zhou et al. (2018) proposed a reinforcement learning-based unsupervised video summarization method which considers the diversity and representativeness of video frames. Their study used a network combining CNN and LSTM to select video frames and evaluated summary quality through various reward functions. Experimental results showed superior performance compared to other unsupervised learning methods.

Research on summarizing video information through caption or audio text conversion includes the work of Lavish Mangal et al. (2022), who developed a YouTube Transcript Summarizer that extracts video captions and generates summaries using abstractive summarization. Their research uses the Hugging Face Transformer model for caption summarization and provides results through a REST API. Additionally, Sulochana Devi et al. (2023) proposed a model applying abstractive summarization to extracted YouTube video captions. Their study presents methods for converting video audio to text and summarizing the converted text to effectively provide users with necessary information. They also explain the two main approaches to text summarization and demonstrate situations where abstractive summarization may be more useful than extractive summarization.

KIM, JI YEON (2023) proposed a new abstractive summarization model emphasizing the importance of summarization without information loss. This enables searching and identifying appropriate information from vast amounts of text data in the big data era. In similar research on literature information summarization, Lee Jong-won, Kim Tae-hyeon, Shin Dong-gu, and Cho Woo-seung (2022) proposed a Korean information summarization system for summarizing national R&D project information. This system developed a preprocessor analyzing Korean linguistic characteristics and a natural language processing-based project information summarization model.

However, Transformer models commonly used in abstractive summarization have fundamental limitations in input token count, or input sentence size, making them impractical for summarizing large amounts of information without additional summarization methodologies. To summarize videos longer than an hour, methods must be devised to divide the video into 10-minute segments, extract and convert audio information to text, create summaries within the token size limits of

Transformer models, and merge these results while preserving the original content.

This study focuses on developing ways for users to quickly grasp necessary information, considering both the advantages and limitations of existing research methods, to address the problem of extensive information volume and time-consuming comprehension in YouTube channels. The aim is to help users understand main content without watching videos while improving information consistency and transmission efficiency.

2.2 Theoretical Background

In this paper, we aim to verify hypotheses about information summarization methods based on key theoretical foundations including information summarization, keyword network analysis, and document similarity analysis. Therefore, we will provide an overview of these major theoretical frameworks..

2.2.1 Information Summarization Methodology

The theoretical background of information summarization can be broadly categorized into two methods: Extractive Summarization and Abstractive Summarization.

Extractive summarization is an approach that identifies and extracts important sentences or phrases from the original text to generate a summary. It utilizes statistical techniques and Natural Language Processing (NLP) technology. A representative method is TF-IDF (Term Frequency-Inverse Document Frequency), which is used to extract important keywords by evaluating the significance of specific terms within a document.

Abstractive summarization creates summaries by deeply understanding the meaning of the original text and generating new sentences. This method incorporates Natural Language Generation (NLG) technology and can better convey meaning while reducing redundancy. Unlike extractive summarization, which simply takes important sentences or phrases directly from the original text, abstractive summarization generates summaries using new expressions based on the core content of the original text. This allows for not only information compression but also provides readers with clearer and more coherent information through restructuring.

The Transformer model based on deep learning is a prime example. Transformer models efficiently learn relationships between words within sentences using the Self-Attention Mechanism. Compared to traditional sequential models like RNN (Recurrent Neural Network) or LSTM (Long Short-Term Memory), it offers advantages in parallel processing capabilities and better understanding of long contexts.

Notably, models that emerged from the development of Transformer architecture, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have shown exceptional performance in various natural language processing tasks. BERT's ability to understand context bidirectionally makes it advantageous for deep text comprehension, while GPT excels at sentence generation.

Other examples of abstractive summarization models include PEGASUS and T5 (Text-To-Text Transfer Transformer). PEGASUS maximizes summarization performance by evaluating the importance of each sentence and learning through masking important sentences for generation. The T5 model redefines all natural language processing problems as text-input and text-output formats, enabling various language tasks to be processed within a unified framework.

These advanced technologies contribute to improving the accuracy of abstractive summarization and effectively summarizing various forms of text data. Therefore, abstractive summarization plays a crucial role in efficiently extracting and conveying useful information from text data.

2.2.2 Keyword and Network Analysis Techniques

Two-Mode Keyword Network Analysis is a methodology that uses two types of nodes to analyze relationships between keywords. It is particularly useful for understanding deeper network structures by simultaneously analyzing relationships between keywords and documents.
The two-mode network can be represented by the following matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \quad (1)$$

In Equation (1), a_ij represents the relationship between document i and keyword j. Rows represent documents, and columns represent keywords. Based on this matrix, the co-occurrence frequency between keywords is calculated and visualized to identify major hub keywords.
Two-mode keyword network analysis is particularly useful for identifying important topics and trends in large volumes of data. For example, Nam Hye-ju and Hong Dae-seok (2023) utilized keyword network analysis to explore research trends related to sports injuries.

2.2.3 Document Similarity Analysis

Document similarity is a methodology for measuring the similarity between two documents, primarily using techniques such as cosine similarity, Pearson correlation coefficient, and k-NN.

● Cosine Similarity
Cosine similarity is a method that measures similarity by calculating the cosine angle between two vectors. When there are two vectors A and B, the cosine similarity is calculated as follows:

Cosine Similarity(A, B) = cos(θ) = (A · B) / (||A|| ||B||)

Here, A · B is the dot product of two vectors, and ||A|| and ||B|| are the magnitudes of vectors A and B, respectively.

● Pearson Correlation Coefficient
The Pearson correlation coefficient is a method for measuring the linear correlation between two variables. The Pearson correlation coefficient between variables X and Y is calculated as follows:.

r=∑(Xi−X¯)2∑(Yi−Y¯)2∑(Xi−X¯)(Yi−Y¯)
Here, X¯\bar{X}X¯ and Y¯\bar{Y}Y¯ are the means of variables XXX and YYY, respectively.

● k-NN (k-Nearest Neighbors)
The k-NN algorithm is used to classify or predict similar data based on distance measurements between data points. This algorithm refers to the k nearest neighbors to determine the class of a new data point. Distance is primarily calculated using Euclidean Distance:.

d(A,B)=∑i=1n(Ai−Bi)2

Here, d(A,B)d(A, B)d(A,B) is the Euclidean distance between vectors AAA and BBB.

These theories contribute to efficiently summarizing and utilizing the vast amount of data provided by digital media platforms like YouTube. These studies provide deeper insights by integrating text and video data, supporting information-based decision-making.

3. Research Methods

3.1 Research Hypothesis

This study hypothesizes that analyzing only comments can identify a channel's main

characteristics and topics without watching all videos or converting them to text for analysis. To verify this hypothesis, we analyze comment data from top-viewed videos over the past year and compare it with keyword network analysis of speech-to-text (STT) converted audio data from the same videos. By comparing the similarity between keyword networks from these two datasets (comments and video text), we assess whether comment analysis alone can sufficiently summarize a YouTube channel's key information. Furthermore, we aim to validate whether this method can be utilized as an effective information summarization approach for YouTube channels.

3.2 Data Collection and Preprocessing

3.2.1 Data Collection Scope and Targets

Five YouTube channels providing content related to nutritional supplements and health supplements were selected as data collection targets. The channel selection criteria included channels with varying subscriber scales (over 1 million, over 100,000, over 10,000, and several thousand subscribers). The collected data includes videos posted during the past year on each channel and their associated comment data. Specifically, we collected channel information, video information, view counts, like counts, and comment content, with data analysis focusing on videos with the highest view counts. In particular, we analyzed the 3-5 most-viewed videos from each channel by converting their audio data to text using STT (Speech-to-Text) tools, along with their comment data. These top-viewed videos were assumed to be representative content for their respective channels, and we determined that the topics and content of these videos well reflect the overall character of the channel.

3.2.2 Data Preprocessing Process

The collected data underwent preprocessing to analyze text data characteristics. First, comment data and video text data were preprocessed separately. Comment data was collected along with video view counts, and de-identified IDs (A, B, C, D, E) were assigned to each channel for analysis. A Korean stopwords list was defined to remove unnecessary words and greetings, and a keyword dictionary related to nutritional supplements and health supplements was created for key keyword extraction.

Video text data was preprocessed in the same way as comment data after converting audio data to text using STT tools. The Korean morphological analyzer (Mecab) was used to separate text into morpheme units and remove stopwords. Based on this, key keywords were extracted from the preprocessed data and visualized through word clouds. This process provided foundational data for identifying key keywords for each channel..

3.3 Analysis Methods

3.3.1 Keyword Network Analysis Method

Keyword network analysis is the process of extracting important keywords from collected comment data and video text data and visualizing their relationships. First, using the TF-IDF (Term Frequency-Inverse Document Frequency) technique, important keywords are extracted from each text based on preprocessed text data. TF-IDF is a statistical method that evaluates how important a word is within a document, assigning higher weights to words that appear frequently in a document but less frequently in other documents. Networks are constructed based on co-occurrence frequency between extracted keywords to identify how often certain keywords appear together.

By visualizing this network, relationships between major keywords and important keywords can be intuitively understood. This allows for quick comprehension of each channel's topics and characteristics, and identification of hub keywords

(major keywords) that play important roles within the network. Keyword network analysis is useful for information compression and summarization, effectively summarizing core topics and content from vast amounts of channel information. It is particularly suitable for simultaneously summarizing and visualizing extensive video content and comment data from YouTube channels.

3.3.2 Text Similarity Analysis Method

Text similarity analysis measures the similarity between comment data and video text data generated through keyword network analysis. This method verifies whether comment analysis alone can effectively summarize the main content of videos. For text similarity analysis, cosine similarity and Pearson correlation coefficient are used. Cosine similarity measures similarity by vectorizing two texts and calculating the angle between vectors, while the Pearson correlation coefficient is a statistical indicator measuring the strength of linear relationships between two data sets. Both methods numerically express the similarity between comment and video text keyword networks, enabling evaluation of how well comment data represents video content. Higher similarity indicates that comment analysis alone can sufficiently grasp the video's topic and content, thereby proving whether the information summarization method using comment analysis can effectively summarize the main content of YouTube channels.

4. Analysis Results

4.1 Analysis Results by YouTube Channel

For each channel, keyword networks were generated using comment data and video text data, and important keywords and their relationships were visually analyzed. The 2-mode network shows relationships between keywords extracted from each data source. The main elements of network analysis are as follows::

- Nodes: Keywords extracted from comments or video text
- Links: Represent relationships between keywords, with relationship strength expressed through color and thickness
- Colors and Links: Nodes represent keywords, yellow links represent comments, green links represent video text, and blue links represent keyword relationships between the two sources.

4.1.1 Channel A

As shown in Figure 1, keywords such as 'vitamin', 'calcium', 'skin', and 'blood pressure' frequently appear in both comments and video text, indicating their importance as key topics. Thick links represent strong relationships between these keywords, and keyword clustering suggests that related topics form groups. Blue links demonstrate high correspondence between the two sources.

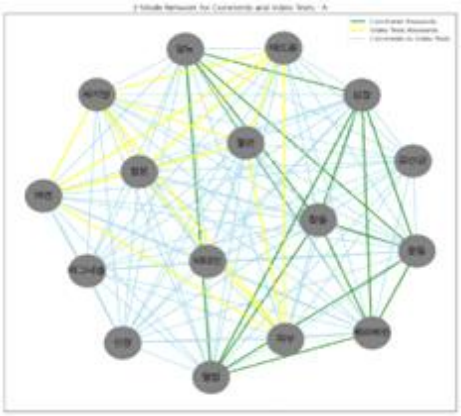


Figure 1: Keyword Network Analysis Results for Channel A

4.1.2 Channel B

As shown in Figure 2, keywords such as 'muscle', 'exercise', 'skin', and 'vitamin' are centrally positioned in the network and are treated as important topics in both sources. 'Muscle' and 'exercise' are frequently mentioned together, and there are many blue links indicating high correspondence between comments and video text.

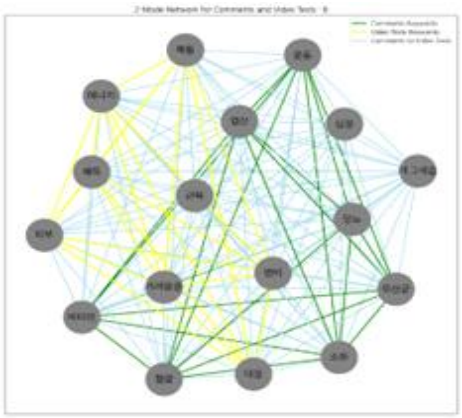


Figure 2: Keyword Network Analysis Results for Channel B

4.1.3 Channel C

As shown in Figure 3, keywords such as 'melatonin', 'magnesium', and 'energy' emerge as major discussion topics, with strong relationships appearing between them. The keyword clusters indicate that specific topics are being discussed in groups, and high similarity can be observed between comments and video text..

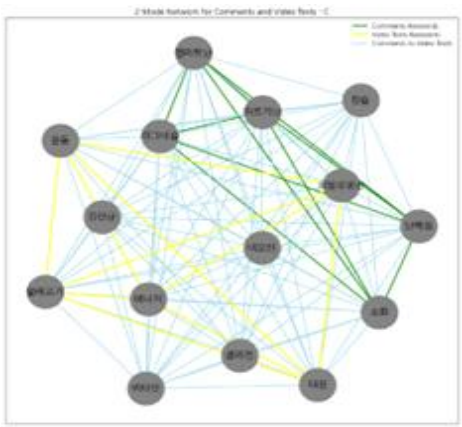


Figure 3: Keyword Network Analysis Results for Channel C

4.1.4 Channel D

As shown in Figure 4, keywords such as 'skin', 'exercise', and 'vitamin' are centrally positioned, with thick links between them indicating strong relationships. Keyword clusters have formed around terms like 'cholesterol', and there is a high degree of information correspondence between comments and video text..

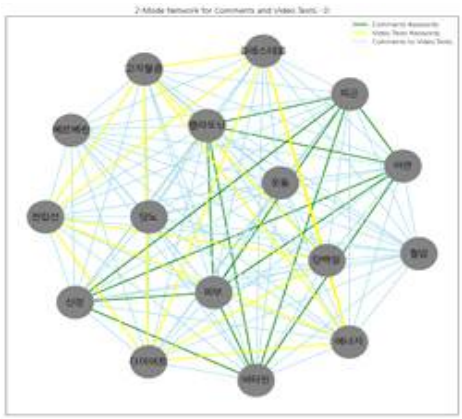


Figure 4: Keyword Network Analysis Results for Channel D

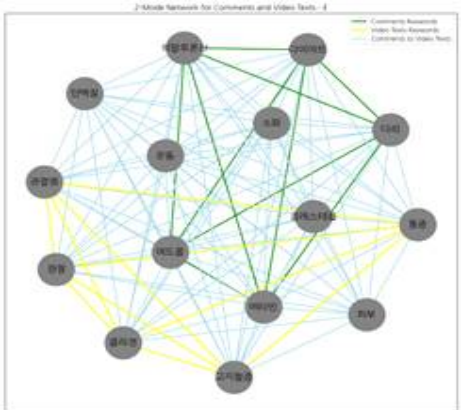
4.1.5 Channel E

As shown in Figure 5, keywords such as 'vitamin', 'exercise', and 'diet' emerge as important topics, with strong relationships appearing between them. Keyword clusters can be identified around terms like 'diet' and 'hyaluronic acid', and the abundance of blue links between the two sources suggests high similarity..

Figure 5: Keyword Network Analysis Results for Channel E

4.2 Text Similarity Analysis

The similarity between comment and video text keyword networks was measured using cosine similarity and Pearson correlation coefficient.



Channel B showed the highest similarity with a cosine similarity of 0.699, while Channel E showed the lowest at 0.424. This indicates that comment data alone can be sufficient to identify a channel's main topics.

채널구분	Cosine Similarity	Pearson Correlation	k-NN Similarity
A	0.676180	0.663575	0.521086
B	0.699733	0.688897	0.545671
C	0.614504	0.595207	0.537723
D	0.546431	0.526614	0.491970
E	0.424739	0.392427	0.472317

Table 1.: Similarity Measurement Results by Channel

4.3 Statistical Verification

The t-test results showed a significant difference with a p-Value of 0.0456. This suggests that comment data could potentially substitute for video text data, particularly in channels with high similarity scores, where comment data alone can effectively identify channel characteristics and topics..

5. Conclusion and Future Research Directions

5.1 Research Summary and Implications

This study explored methods for summarizing YouTube channel information by comparing speech-to-text (STT) data from top-viewed videos with comment data. Particular emphasis was placed on the possibility of information summarization using comments. The research findings demonstrated that keyword networks extracted from comments showed high similarity with STT data, confirming that comments can effectively summarize the main content of YouTube channels. This method can be rapidly utilized by businesses in influencer marketing and help individual subscribers quickly grasp the core content of channels.

This research has proven that comment data serves as a valid information summarization tool. Furthermore, it suggests that keyword network and text similarity analyses can be applied to various topics and channels. Future research will require multimodal data analysis integrating video, audio, and text, which could lead to the development of more sophisticated content summarization methods.

5.2 Research Limitations and Future Research Directions

This study analyzed data limited to certain YouTube channels, and research including channels with diverse topics is needed. Additionally, the heavy reliance on keyword networks and text analysis necessitates the introduction of multimodal data analysis. Future research needs to develop more comprehensive methods for analyzing and summarizing YouTube content by integrating video, audio, and text data. This can be utilized across various fields including digital marketing, education, and information delivery..

Reference

Kim, J. Y. (2023). "Enhancing Abstract Summarization through Triplet loss." Master's Thesis, Graduate School of Computer and Information Technology, Korea University, Seoul.

Lee, J. W., Kim, T. H., Shin, D. G., & Cho, W. S. (2022). Korean Information Summarization System for Summarizing National R&D Project Information. Journal of the Korea Institute of Information and Communication Engineering, Jeju.

Jung, J. I. (2016). "Improved Tag Search System through Keyword Extraction and Similarity Evaluation." Master's Thesis, Soongsil University, Seoul.

Baek, S. H., & Kim, E. J. (2021). Research Trends of Korean Elderly Exercise through

Keyword Network Analysis (1995-2019). Journal of the Korea Convergence Society, 19(1).

Nam, H. J., & Hong, D. S. (2023). A Study on Keyword Network Analysis for Exploring Research Trends Related to Sports Injuries. Korean Journal of Sports Science, 32(6), 755-766. https://doi.org/10.35159/kjss.2023.12.32.6.755

Kim, G. Y. (2019). Analysis of Tourist Destination Network Characteristics through Two-Mode Network Analysis - Focusing on Japanese Tourists Visiting Korea. Journal of Tourism Research, 31(4), 23-46. https://doi.org/10.21581/jts.2019.11.31.4.23

Yoo, J. Y., & Kim, M. K. (2019). Convergence of Korean Traditional Dance and K-pop Dance: Analysis of YouTube Comments on BTS's 'IDOL' at 2018 MMA. Journal of Korea Entertainment Industry Association, 13(8), 189-198. https://doi.org/10.21184/jkeia.2019.12.13.8.189

Lavish Mangal, Dhruv Aggarwal, Gulshan Kumar, Jai Kumar, Meenakshi Aggarwal. (2022). YouTube Transcript Summarizer. 2022 International Mobile and Embedded Technology Conference (MECON)

J. H. Shin, E. H. Kim, and M. J. Lim, “Improving the effectiveness of document extraction summary based on the amount of sentence information,” Smart media journal, Vol. 11, No. 3, pp. 31-38, Apr. 2022.

Mohamed, Fatma, and Abdulhadi Shoufan. (2024). Users’ experience with health-related content on YouTube: an exploratory study. BMC Public Health, 24:86.

Sulochana Devi, Rahul Nadar, Tejas Nichat, and Alfredprem Lucas. (2023). Abstractive Summarizer for YouTube Videos. In Proceedings of the International Conference on Applied Mathematics, Intelligent Systems and Computing Applications (ICAMIDA), vol. 105, pp. 431-438.

Kaiyang Zhou, Yu Qiao, and Tao Xiang. (2018). Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).

◐ ABOUT THE AUTHOR ◑

Hyoung-sik Park

EDUCATION

- Master of Engineering in AI Big Data
- Department of Advanced AI, Graduate School of AI
- aSSIST (Seoul School of Integrated Sciences & Technologies)

RESEARCH INTERESTS

- Artificial Intelligence
- Natural Language Processing
- Large Language Models (LLMs)
- And related fields

CONTACT E-mail: saintphs@naver.com

AI-Driven Stock Prediction and Investment Strategy Analysis for the Magnificent 7 Tech Giants

Jumin Lee

ABSTRACT

This study aims to predict stock prices and develop optimized investment strategies for the "Magnificent 7" companies using AI. By employing Linear Regression, Random Forest, XGBoost, and LSTM models, the study analyzed stock price data for Apple, Amazon, Google, Meta, Microsoft, Nvidia, and Tesla from 2015 to the first half of 2024, along with VIX index and bond yield data. The accuracy of predictions for each model was assessed, and simulations were conducted to observe which model would achieve the highest returns when applied to actual investments. The results showed varying levels of predictive accuracy and performance for each stock, leading to the conclusion that a hybrid approach combining AI and traditional investment strategies could be effective.

☞ keyword : AI, Machine Learning, Stock Prediction, Investment Strategy, Magnificent 7, LSTM, XGBoost, Random Forest, Linear Regression

1. Introdutcion

Predicting stock prices and developing investment strategies have long been focal points of research in the financial sector. The inherent volatility of the stock market presents a formidable challenge for investors, making the development of predictive methodologies crucial. Traditionally, stock price forecasting relied on economic indicators, corporate performance, and technical analyses, but these approaches often depend heavily on historical data and may fall short of capturing the complex and nonlinear market dynamics.

Meanwhile, the stock prices of the major tech giants collectively known as the "Magnificent 7" (Apple, Amazon, Google, Meta, Microsoft, Nvidia, and Tesla) exert substantial influence not only on their respective companies and the U.S. stock market but also on the global economy. Predicting the stock prices of these companies is, therefore, of critical importance for global investors from both micro and macro perspectives. This makes them an ideal target for research utilizing machine learning and artificial intelligence, further underscoring the academic and practical value of this endeavor.

The primary academic objective of this study is to evaluate the accuracy of stock price predictions using various machine learning models. For this purpose, four models were selected: Linear Regression, Random Forest, XGBoost, and Long Short-Term Memory (LSTM). By comparing the performance of these models, this study aims to identify the most effective and suitable model for forecasting specific stock prices. For example, while a simple linear regression model might perform well in certain scenarios, more complex models like XGBoost, which excel at capturing nonlinear relationships, could achieve superior performance in others. Model performance will be evaluated by comparing prediction accuracy on training and test data sets using metrics such as Root Mean Squared Error (RMSE) and R^2 scores.

The second academic objective is to optimize the performance of each model through hyperparameter tuning and feature engineering. Hyperparameter tuning is a critical process for maximizing model predictive performance, employing techniques like Grid Search to test various parameter combinations and derive the optimal settings. Feature engineering involves transforming data and creating new variables to enhance model predictive capacity. This study employed data normalization, moving average calculations, and the integration of VIX index and treasury yield data to improve data quality, thereby enabling more accurate predictions by the models.

Another intention of this study is to generate investment insights. The first investment-related objective is to simulate and assess the expected performance of investments based on predictions obtained from each machine learning model. By applying the predictions to various stocks and evaluating potential investment outcomes, this research seeks to verify the practical effectiveness and profitability of machine learning-based investment strategies.

The second investment-related objective is to derive optimal investment methodologies based on stock price predictions and simulation results for the Magnificent 7 companies. This analysis aims to determine which strategies yield the highest returns and which strategies mitigate risk while generating stable profits.

In conclusion, the purpose and significance of this

study lie in simultaneously providing academic contributions and practical investment guidance by examining the predictive capabilities of machine learning models and their applicability to real-world investment strategies. This study explores the potential of AI-driven investment strategies and seeks to verify more efficient and effective methodologies for stock market investments.

2. Theoretical Background and Previous Studies

2.1 Utilization of Machine Learning

Machine learning refers to computational algorithms designed to learn from given situations and emulate human intelligence. Recently, it has undergone rapid advancements across diverse domains. Since 2011, studies leveraging machine learning for stock market predictions have gained momentum, demonstrating its effectiveness in analyzing and predicting complex stock market patterns [1][2][3]. Machine learning algorithms can extract meaningful patterns from vast datasets, significantly improving the accuracy of stock price predictions.

2.2 Combining Statistical Methods with Machine Learning

The integration of statistical methods and machine learning techniques in stock market forecasting plays a vital role in maximizing prediction performance, with numerous systematic reviews and applications of these combined methods in stock market predictions [4]. Such studies suggest that combining traditional statistical techniques with innovative machine learning methods can enhance the performance of predictive models. For example, statistical methods are beneficial for understanding the fundamental characteristics of data, while machine learning excels at analyzing and predicting more complex patterns based on these characteristics.

2.3 Applications of Deep Learning

Deep learning, a subset of machine learning, is particularly effective at processing large-scale datasets and learning complex patterns. According to Jiang (2019), deep learning techniques have been successfully applied to stock market forecasting, with models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) demonstrating impressive performance [5]. Deep learning models are well-suited for learning hidden patterns from large datasets and making accurate predictions based on those patterns.

2.4 Practical Insights

Various studies confirm that machine learning and deep learning models are highly effective at improving the accuracy of stock market predictions [1][4]. Additionally, deep learning models have contributed significantly to analyzing complex data patterns and enhancing predictive performance [5][6].

3. Methodology

3.1 Data Collection and Preprocessing

This study utilizes daily stock price data for the Magnificent 7 companies (Apple, Amazon, Google, Meta, Microsoft, Nvidia, and Tesla) from 2015 to the present (2024). The dataset includes information such as daily closing prices, trading volumes, highs, and lows for each company. Furthermore, data on the Volatility Index (VIX) and bond yields over the same period were collected and used.

(Table 1) Types and Sources of Data Used

Type	Description	Source
Stock Prices Data	The dataset includes stock price data for seven major tech companies under study: Apple, Amazon, Google, Meta, Microsoft, Nvidia, and Tesla. Data fields include closing price (Close), trading volume (Volume), opening price (Open), highest price (High), and lowest price (Low).	Yahoo Finance
VIX Data	Volatility index data representing stock market volatility, with VIX reflecting market uncertainty and investor sentiment. It is commonly referred to as the "fear index".	CBOE (Chicago Board Options Exchange)
Bond Yield Data	Yield data of major bonds influencing the stock market. For instance, the 10-year U.S. Treasury yield significantly impacts the stock market based on economic conditions and interest rate fluctuations.	FRED (Federal Reserve Economic Data)

The collected data was integrated by stock using Python's Pandas library, with missing values handled and transformed into a format suitable for model training.

(Table 2) Sample Format of Preprocessed & Integrated Data (Apple)

Date	Close	Adj Close	Volume	Close VIX	Bond Yield	50 MA	200 MA	Volume Change

2015-10-16	27.76	25.10	156,930,400	15.05	2.04	25.39	27.27	0.04
2015-10-19	27.93	25.25	119,036,800	14.98	2.04	25.38	27.27	-0.24
2015-10-20	28.44	25.72	195,871,200	15.75	2.08	25.35	27.28	0.65
2015-10-21	28.44	25.71	167,180,800	16.70	2.04	25.35	27.29	-0.15

(Table 3) Detailed Data Items for Integrated Dataset

Tyoe	Description
Date	Trading date.
Close	Closing price of the stock for the trading day.
Adj Close	Adjusted closing price, reflecting factors such as dividends and stock splits.
Volume	Trading volume (number of shares traded on that day).
_VIX	Volatility Index (VIX) for the trading day, reflecting market uncertainty and investor sentiment.
Bond_Yield	Bond yield for the trading day (e.g., the 10-year U.S. Treasury yield, which significantly impacts the stock market based on economic conditions and interest rate fluctuations).
50_MA	50-day moving average (average closing price of the stock over the past 50 trading days).
200_MA	200-day moving average (average closing price of the stock over the past 200 trading days).
Volume_Change	Change in trading volume from the previous day (current trading day's volume minus the previous trading day's volume, indicating the extent of volume fluctuation).

3.2 Model Development and Training

Based on the literature review and the available analytical environment for this study, the following models were selected, and each was developed and trained using the appropriate methodologies:

- Linear Regression: Models the linear trends of stock prices.
- Random Forest: An ensemble technique used to capture nonlinear patterns.
- XGBoost: Optimizes prediction performance through boosting techniques.
- LSTM (Long Short-Term Memory): A recurrent neural network model designed for time series data processing.

For model training and testing, data was first normalized and transformed into a suitable format for training. Using the 'create_dataset' function, stock price data was grouped into sets of 100 days, and the model was tasked with predicting the stock price of the following day.

- Train on data from day 1 to 100 and predict day 101.
- Train on data from day 2 to 101 and predict day 102.
- Train on data from day 3 to 102 and predict day 103.
  ...
- Train on data from day n to n+99 and predict day n+100.

The generated dataset was split randomly into 80% for training and 20% for testing (using the train_test_split function) while maintaining the time-series order. The first 80% was used for training, and the remaining 20% for testing. The 80:20 ratio was adopted as it is a widely validated conventional split, ensuring a sufficiently large training set to prevent overfitting while maintaining a robust testing set to minimize the risk of inadequate model performance evaluation.

3.3 Performance Evaluation Metrics

The models' performance was evaluated by comparing prediction accuracy on the training and testing datasets using the following metrics:

- RMSE (Root Mean Squared Error): The square root of the average squared differences between predicted and actual values. Lower values indicate higher prediction accuracy.
- R^2 (R-squared): Indicates the proportion of variance in the dependent variable that is predictable from the independent variables. Values closer to 1 indicate better model explanatory power.

4. Results

4.1 Apple(APPL)

4.1.1 Model Performance Evaluation Results

- Linear Regression
  Train RMSE: 0.00721370768084068
  Train R^2: 0.9993518408562049
  Test RMSE: 0.013456124594076024
  Test R^2: 0.9978237212930056
- Random Forest
  Train RMSE: 0.003872961711382928
  Train R^2: 0.9998131681216459
  Test RMSE: 0.010092068824678632
  Test R^2: 0.9987758492002643
- XGBoost
  Train RMSE: 0.002713807236596701

Train R²: 0.9999082676402655
Test RMSE: 0.009915095578910946
Test R²: 0.99881840587347

- LSTM
  Train RMSE: 3.9092045598054934
  Train R²: 0.9954810401662878
  Test RMSE: 4.039114495183321
  Test R²: 0.9953447398420988

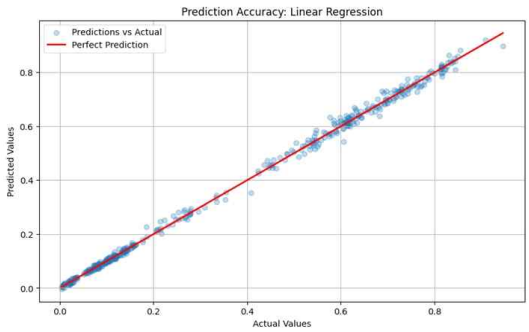
4.1.2 Investment Simulation Results

Assuming an initial seed money of $10,000 and using each model to simulate investments from January 2, 2015, to the present, the expected returns are as follows:

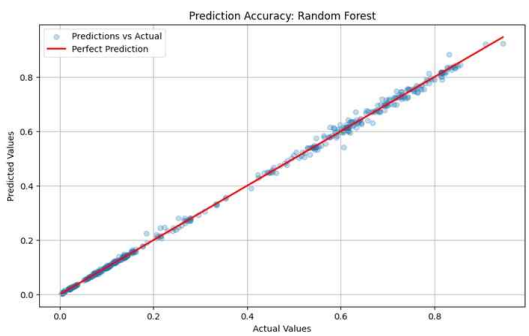
- Linear Regression: $77,702.48
- Random Forest: $74,160.32
- XGBoost: $78,953.78
- LSTM: $80,976.98
- Buy & Hold: $78,079.97
  (Hold after bulk purchase on January 2, 2015)

4.1.3 Model Performance Comparison

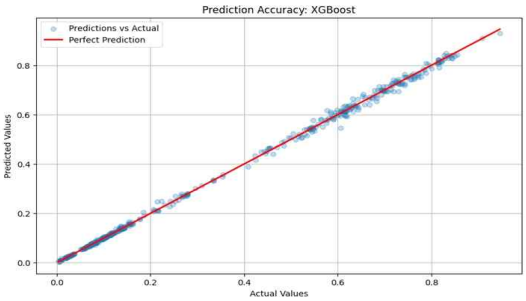
- Best Performing Model: XGBoost (based on superior RMSE and R² metrics)
- Highest Profit Model: LSTM
- Additional Notes: The specialized performance of the LSTM model for time series data may have contributed to its strong long-term profitability.



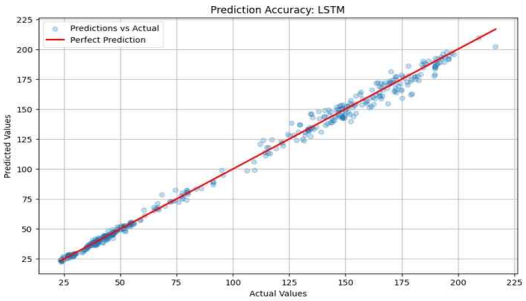
(Image 1) APPL Linear Regression Prediction Accuracy Scatter Plot



(Image 2) APPL Random Forrest Prediction Accuracy Scatter Plot



(Image 3) APPL XGBoost Prediction Accuracy Scatter Plot



(Image 4) APPL LSTM Prediction Accuracy Scatter Plot

4.2 Amazon(AMZN)

4.2.1 Model Performance Evaluation Results

- Linear Regression
  Train RMSE: 0.01024894253238069
  Train R²: 0.9984772426597099
  Test RMSE: 0.01740486189657552
  Test R²: 0.9955758516518073
- Random Forest
  Train RMSE: 0.00543436952479615
  Train R²: 0.9995718746755582
  Test RMSE: 0.014243840390853992
  Test R²: 0.9970369245178632
- XGBoost
  Train RMSE: 0.0092742528308425
  Train R²: 0.9987531033783534
  Test RMSE: 0.01467343541852455
  Test R²: 0.9968554961719535
- LSTM
  Train RMSE: 4.173629609369123
  Train R²: 0.9918381990420058
  Test RMSE: 4.268768899802902
  Test R²: 0.9913983966923228

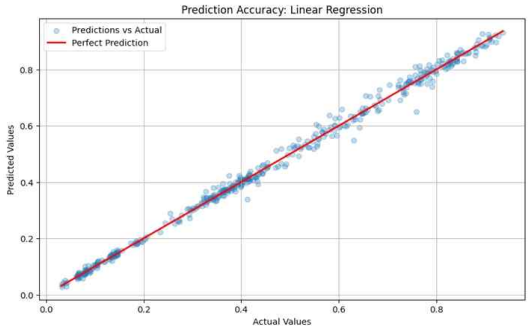
4.2.2 Investment Simulation Results

Assuming an initial seed money of $10,000 and using each model to simulate investments from January 2, 2015, to the present, the expected returns are as follows:

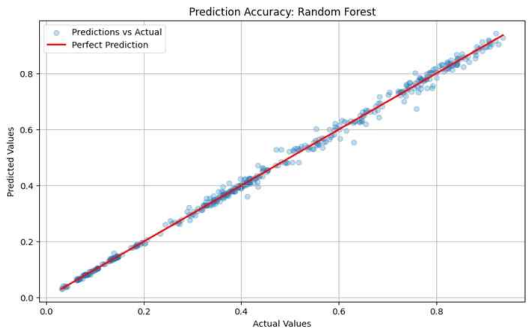
- Linear Regression: $67,315.58
- Random Forrest: $67,915.98
- XGBoost: $67,524.25
- LSTM: $72,617.99
- Buy & Hold: $69,100.85
  (Hold after bulk purchase on January 2, 2015)

4.2.3 Model Performance Comparison

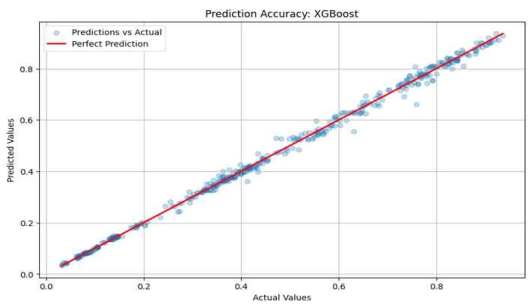
- Best Performing Model: Random Forest (based on superior RMSE and R^2 metrics)
- Highest Profit Model: LSTM
- Additional Notes: The specialized performance of the LSTM model for time series data may have contributed to its strong long-term profitability.



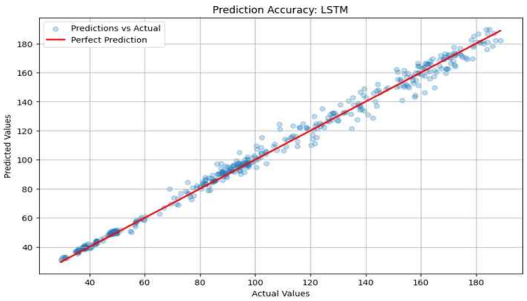
(Image 5) AMZN Linear Regression Prediction Accuracy Scatter Plot



(Image 6) AMZN Random Forrest Prediction Accuracy Scatter Plot



(Image 7) AMZN XGBoost Prediction Accuracy Scatter Plot



(Image 8) AMZN LSTM Prediction Accuracy Scatter Plot

4.3 Google(GOOG)

4.3.1 Model Performance Evaluation Results

- Linear Regression
  Train RMSE: 0.007584069124673875
  Train R^2: 0.9990256817711023
  Test RMSE: 1.306391154418243
  Test R^2: -27.65752534522947
- Random Forest
  Train RMSE: 0.00406696540091182
  Train R^2: 0.9997198199438097
  Test RMSE: 0.012584873126705991
  Test R^2: 0.9973405617786385
- XGBoost
  Train RMSE: 0.006815078229777993
  Train R^2: 0.9992132478227236
  Test RMSE: 0.012310704901908068
  Test R^2: 0.9974551741655481
- LSTM
  Train RMSE: 3.406891096143106
  Train R^2: 0.9923034011633383
  Test RMSE: 3.605051117097021
  Test R^2: 0.9914571818277564

4.3.2 Investment Simulation Results

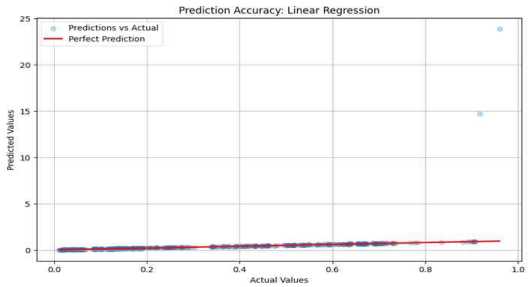
Assuming an initial seed money of $10,000 and using each model to simulate investments from January 2, 2015, to the present, the expected returns are as follows:

- Linear Regression: $57,874.78
- Random Forrest: $53,630.16
- XGBoost: $53,844.67
- LSTM: $54,501.17
- Buy & Hold: $55,720.33
  (Hold after bulk purchase on January 2, 2015)

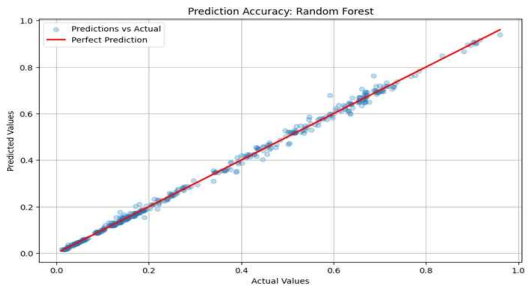
4.3.3 Model Performance Comparison

- Best Performing Model: XGBoost, Random Forest (based on superior RMSE and R^2 metrics)
- Highest Profit Model: Linear Regression
- Additional Notes: The Linear Regression model

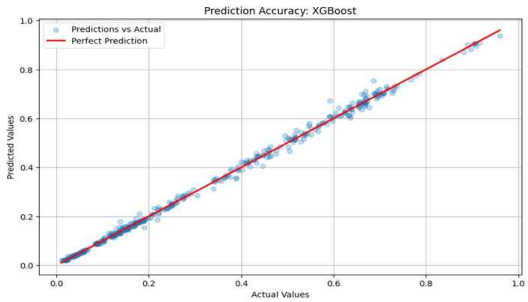
has good Train RMSE and Train R^2, but poor Test RMSE and Test R^2 indicating overfitting and particularly the low Test R^2 suggests that the model does not explain the test data well. Despite this, it shows strong performance in terms of returns.



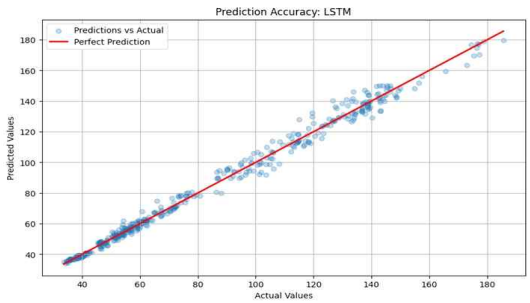
(Image 9) GOOG Linear Regression Prediction Accuracy Scatter Plot



(Image 10) GOOG Random Forrest Prediction Accuracy Scatter Plot



(Image 11) GOOG XGBoost Prediction Accuracy Scatter Plot



(Image 12) GOOG LSTM Prediction Accuracy Scatter Plot

4.4 Meta(META)

4.4.1 Model Performance Evaluation Results

- Linear Regression
  Train RMSE: 0.009598516669102537
  Train R^2: 0.997895457708489
  Test RMSE: 0.017520560791958403
  Test R^2: 0.9923682025495548
- Random Forest
  Train RMSE: 0.005150499879096363
  Train R^2: 0.999394034117066
  Test RMSE: 0.014090458652083021
  Test R^2: 0.9950639325031585
- XGBoost
  Train RMSE: 0.007878474457522076
  Train R^2: 0.9985821389464132
  Test RMSE: 0.013624728776867986
  Test R^2: 0.9953848421314374
- LSTM
  Train RMSE: 9.973018666277179
  Train R^2: 0.9888300906047836
  Test RMSE: 10.663743413230975
  Test R^2: 0.9861005639621295

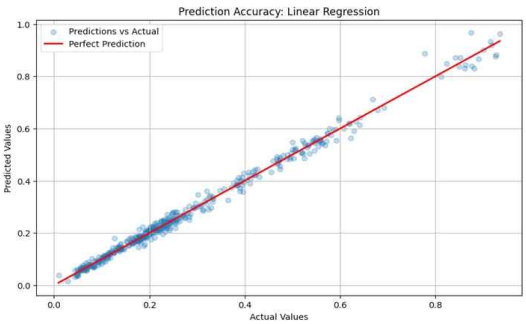
4.4.2 Investment Simulation Results

Assuming an initial seed money of $10,000 and using each model to simulate investments from January 2, 2015, to the present, the expected returns are as follows:

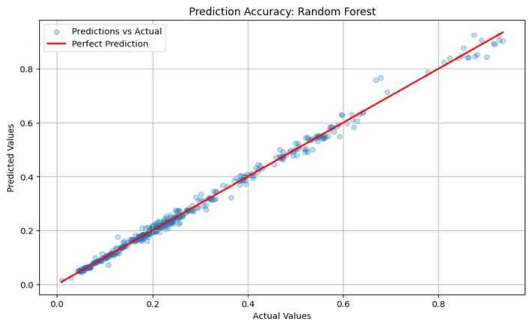
- Linear Regression: $51,615.16
- Random Forrest: $51,014.77
- XGBoost: $53,029.05
- LSTM: $51,506.32
- Buy & Hold: $51,740.82
  (Hold after bulk purchase on January 2, 2015)

4.4.3 Model Performance Comparison

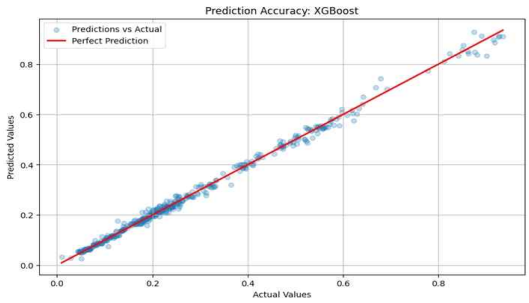
- Best Performing Model: Random Forest, XGBoost (based on superior RMSE and R^2 metrics)
- Highest Profit Model: XGBoost
- Additional Notes: Although the LSTM model is strong in time series analysis, it showed the lowest performance compared to other models and recorded the second-lowest returns.



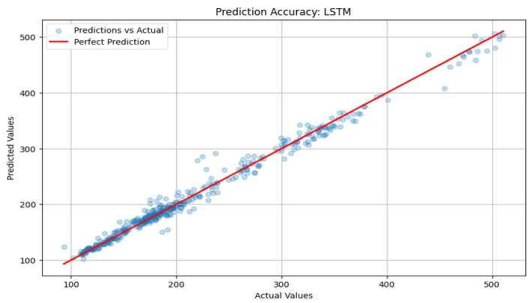
(Image 13) META Linear Regression Prediction Accuracy Scatter Plot



(Image 14) META Random Forrest Prediction Accuracy Scatter Plot



(Image 15) META XGBoost Prediction Accuracy Scatter Plot



(Image 16) META LSTM Prediction Accuracy Scatter Plot

4.5 Microsoft(MSFT)

4.5.1 Model Performance Evaluation Results

- Linear Regression
  Train RMSE: 0.006396164498521088
  Train R²: 0.9993954886056274
  Test RMSE: 0.01198268359729427
  Test R²: 0.9978577520294
- Random Forest
  Train RMSE: 0.003288831314077367
  Train R²: 0.9998401736284992
  Test RMSE: 0.010068287897096047
  Test R²: 0.9984875786230499
- XGBoost
  Train RMSE: 0.002313653802046101
  Train R²: 0.9999209026393698
  Test RMSE: 0.009884630408063724
  Test R²: 0.9985422520911019
- LSTM
  Train RMSE: 6.262766291547487
  Train R²: 0.9967201431262446
  Test RMSE: 6.6783989994057515
  Test R²: 0.9962341456385856

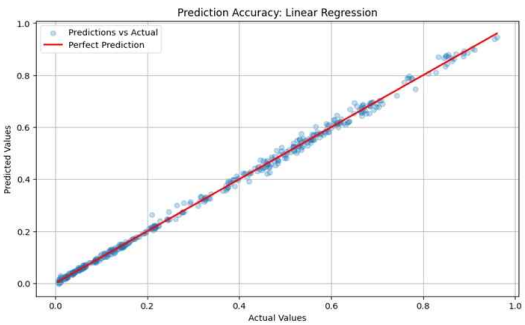
4.5.2 Investment Simulation Results

Assuming an initial seed money of $10,000 and using each model to simulate investments from January 2, 2015, to the present, the expected returns are as follows:

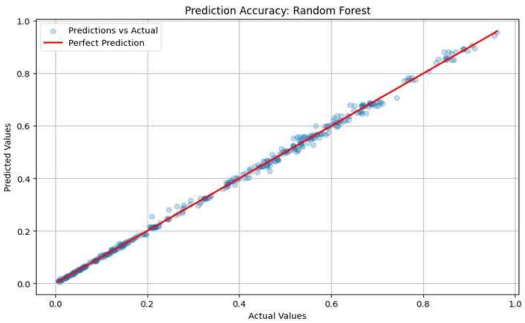
- Linear Regression: $97,156.06
- Random Forrest: $95,509.19
- XGBoost: $98,490.52
- LSTM: $94,653.07
- Buy & Hold: $96,133.45
  (Hold after bulk purchase on January 2, 2015)

4.5.3 Model Performance Comparison

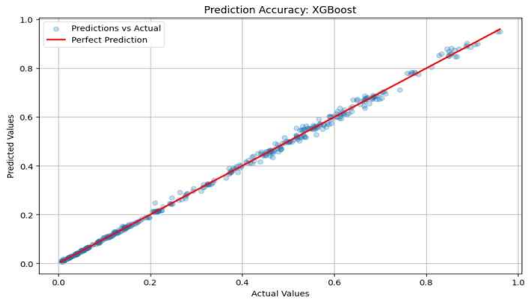
- Best Performing Model: XGBoost (based on superior RMSE and R² metrics)
- Highest Profit Model: XGBoost
- Additional Notes: Although the LSTM model is strong in time series analysis, it showed the lowest performance and returns compared to other models.



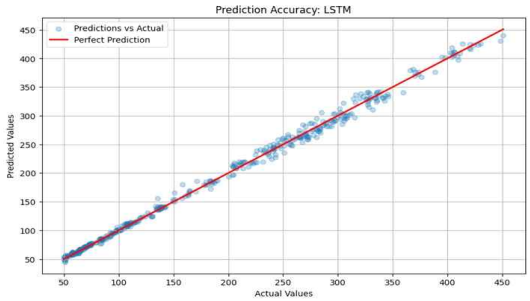
(Image 17) MSFT Linear Regression Prediction Accuracy Scatter Plot



(Image 18) MSFT Random Forrest Prediction Accuracy Scatter Plot



(Image 19) MSFT XGBoost Prediction Accuracy Scatter Plot



(Image 20) MSFT LSTM Prediction Accuracy Scatter Plot

4.6 Nvidia(NVDA)

4.6.1 Model Performance Evaluation Results

- Linear Regression
  Train RMSE: 0.003770818938547538
  Train R^2: 0.9994855066070386
  Test RMSE: 0.009463293046525049
  Test R^2: 0.9964062844613787
- Random Forest
  Train RMSE: 0.002500311935477405
  Train R^2: 0.9997737977966868
  Test RMSE: 0.005681129619470234
  Test R^2: 0.9987048248121455
- XGBoost
  Train RMSE: 0.000467221128859668
  Train R^2: 0.9999921013405159
  Test RMSE: 0.007505210468866724
  Test R^2: 0.9977396022260528
- LSTM
  Train RMSE: 1.6066052684654817
  Train R^2: 0.9948715722217412
  Test RMSE: 1.4947424804646756
  Test R^2: 0.9950767793363464

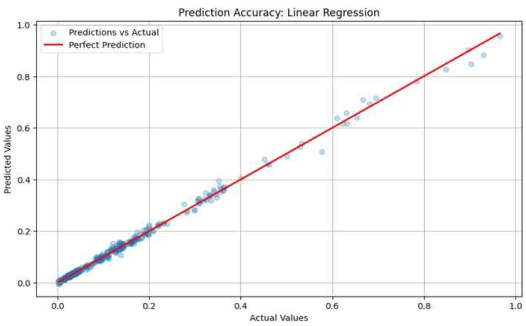
4.6.2 Investment Simulation Results

Assuming an initial seed money of $10,000 and using each model to simulate investments from January 2, 2015, to the present, the expected returns are as follows:

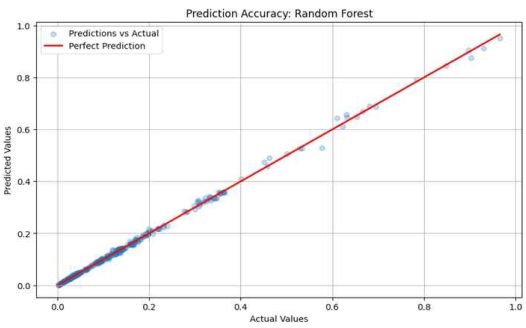
- Linear Regression: $1,995,663.85
- Random Forrest: $1,765,615.30
- XGBoost: $1,766,698.25
- LSTM: $1,553,832.96
- Buy & Hold: $1,784,637.52
  (Hold after bulk purchase on January 2, 2015)

4.6.3 Model Performance Comparison

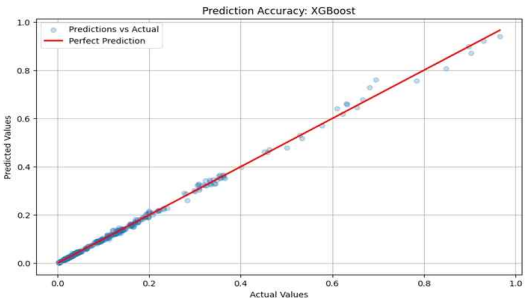
- Best Performing Model: Random Forest, XGBoost (based on superior RMSE and R^2 metrics)
- Highest Profit Model: Linear Regression
- Additional Notes: Although the LSTM model is strong in time series analysis, it showed the lowest performance and returns compared to other models.



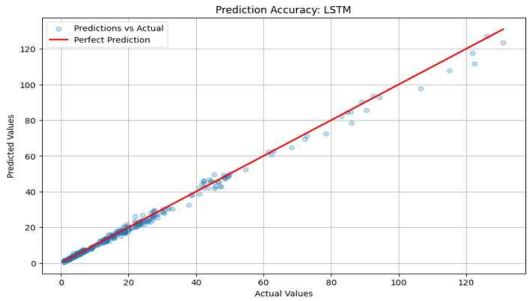
(Image 21) NVDA Linear Regression Prediction Accuracy Scatter Plot



(Image 22) NVDA Random Forrest Prediction Accuracy Scatter Plot



(Image 23) NVDA XGBoost Prediction Accuracy Scatter Plot



(Image 24) NVDA LSTM Prediction Accuracy Scatter Plot

4.7 Tesla(TSLA)

4.7.1 Model Performance Evaluation Results

- Linear Regression
  Train RMSE: 0.011075039168475045
  Train R^2: 0.998384971868137
  Test RMSE: 0.02171689158060674
  Test R^2: 0.9941871356753815
- Random Forest
  Train RMSE: 0.005993219225495276
  Train R^2: 0.9995270558281623
  Test RMSE: 0.017311966357605693
  Test R^2: 0.9963060785254125
- XGBoost
  Train RMSE: 0.000992511762286982
  Train R^2: 0.9999870293758624
  Test RMSE: 0.017603887595299603
  Test R^2: 0.9961804514585545
- LSTM
  Train RMSE: 10.010356379463106
  Train R^2: 0.9917696587410665
  Test RMSE: 10.218756656331266
  Test R^2: 0.9919717629150182

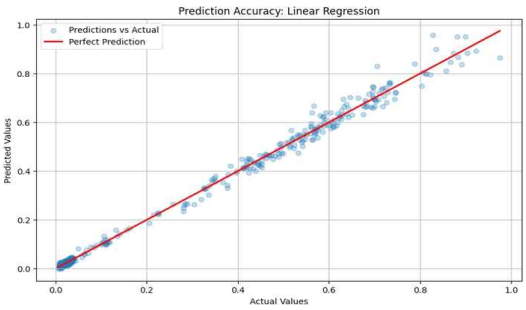
4.7.2 Investment Simulation Results

Assuming an initial seed money of $10,000 and using each model to simulate investments from January 2, 2015, to the present, the expected returns are as follows:

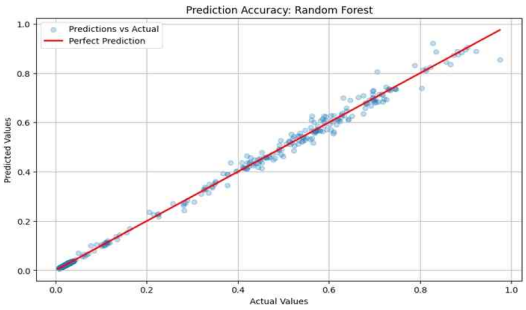
- Linear Regression: $142,466.32
- Random Forrest: $132,559.25
- XGBoost: $139,615.28
- LSTM: $141,050.34
- Buy & Hold: $138,667.90
  (Hold after bulk purchase on January 2, 2015)

4.7.3 Model Performance Comparison

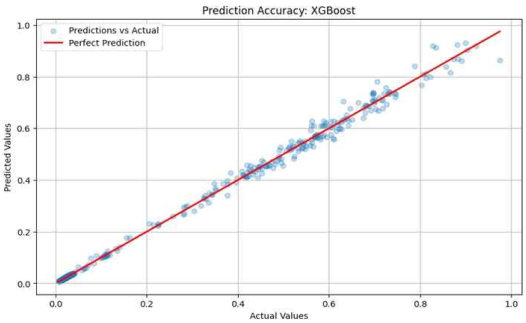
- Best Performing Model: Random Forest, XGBoost (based on superior RMSE and R^2 metrics)
- Highest Profit Model: Linear Regression
- Additional Notes: Despite its strengths in time series analysis, the LSTM model exhibited the lowest performance compared to the other models. However, it ranked second in terms of investment returns.



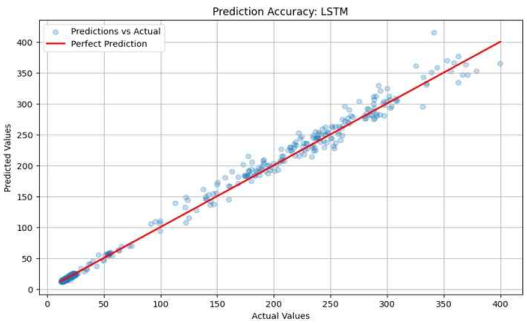
(Image 25) TSLA Linear Regression Prediction Accuracy Scatter Plot



(Image 26) TSLA Random Forrest Prediction Accuracy Scatter Plot



(Image 27) TSLA XGBoost Prediction Accuracy Scatter Plot



(Image 28) TSLA LSTM Prediction Accuracy Scatter Plot

5. Conclusion and Recommendations

5.1 Limitations of the Study

The data used in this study is limited to the period from 2015 to 2023. Although this is not a short period, it may not be sufficient to fully reflect macroeconomic cycles, market trends, and structural changes in stocks. Therefore, it may be necessary to analyze data over several decades to identify long-term patterns and cycles, which would enhance the stability and reliability of the predictive model. For instance, macroeconomic events such as economic recessions or booms can be more clearly analyzed using data spanning a longer time frame to assess their impact.

Complex machine learning models, particularly deep learning models, may face issues of overfitting, where the model becomes too tailored to the training data and loses its ability to generalize to new data. In this study, efforts were made to avoid overfitting by using methods such as cross-validation, regularization, and callbacks, and to prevent excessive complexity and hyperparameter tuning for each model. However, the results from the LSTM model suggest that further improvements may be necessary. Additionally, complex models require high computational costs and training time, which may pose challenges for real-time predictions and applications.

Moreover, the stock market is influenced by complex and diverse factors such as political events, economic policy changes, natural disasters, and technological innovations. Fully accounting for these external factors in the model is a challenging task. For example, unexpected events like a pandemic may significantly differ from the existing data patterns and affect the model's predictive performance. To integrate all these external factors, it would be necessary to incorporate real-time data feeds, news analysis, and social media data, but this would come with additional costs and technical challenges.

5.2 Recommendatinos

To enhance investment strategies using artificial intelligence, it is important to analyze the stock price patterns of each company and identify the external factors affecting them to understand the primary causes of stock price fluctuations.

- Design customized features based on the characteristics of each company and collect relevant external factor data. For example, for specific companies, technical indicators, macroeconomic indicators, and social media sentiment analysis could be included.
- Test various models and select the one most suitable for each company. For example, stocks with non-linear patterns may benefit from Random Forest or XGBoost, while time series data may be better suited for LSTM.
- Test the effectiveness of each model in actual investment scenarios through simulation.

Additionally, it is recommended to collect real-time stock price data continuously via APIs and periodically update the model by retraining it with new data every week or month, followed by deploying the retrained model for real-time predictions.

- Use data streaming platforms such as Apache Kafka or AWS Kinesis to collect and process real-time data.
- Automate the processes of data collection, preprocessing, model retraining, and deployment using tools like Apache Airflow or Jenkins.
- Use cloud services like AWS, Google Cloud, or Microsoft Azure to build scalable infrastructure.
- Establish an MLOps (Machine Learning Operations) framework for managing and operating these processes.

Using advanced ensemble models can refine predictions and yield improved results.

- Stacking: Combine different models to create a new meta-model. For instance, predictions from Random Forest, XGBoost, and LSTM can be used to train a final model to make the overall prediction.

Furthermore, integrating traditional investment risk management techniques can help reduce loss risk and improve profitability.

- Stop-loss Strategy: Automatically sell when the stock price drops below a certain level. By setting a stop-loss price based on machine learning predictions and automating its execution, significant losses can be prevented in the event of unforeseen occurrences.
- Hedging: Use derivatives such as options or futures based on model predictions to offset portfolio risks.
- Diversification: Use model predictions to invest in diverse asset classes, thus reducing risk from specific assets.

5.3 Conclusion

This study applied various machine learning models to predict stock prices and analyze investment strategies for the Magnificent 7 companies (Apple, Amazon, Google, Meta, Microsoft, Nvidia, and Tesla). The models used included Linear Regression, Random Forest, XGBoost, and LSTM, and the performance of each model was evaluated using RMSE and R^2. Further, investment simulations were conducted to validate the effectiveness of the strategies in

real-world scenarios. The conclusions drawn from the results are as follows:

First, when performing stock price predictions using four machine learning models - Linear Regression, Random Forest, XGBoost, and LSTM - the models showed varying prediction performance depending on the stock. In general, Random Forest and XGBoost demonstrated high predictive accuracy. However, in some cases, Linear Regression and LSTM also showed competitive performance.

Second, based on investment simulations using machine learning models, the traditional "buy and hold" strategy outperformed the model-based approach in certain situations. This suggests the complexity of market timing and the difficulties of real-time trading, implying that AI-based strategies do not always guarantee superior performance. However, model-based strategies have the potential to generate high returns and can be an effective investment tool depending on market conditions.

It should be noted that the relatively short data period of this study limited the ability to fully capture long-term predictions and trends. Furthermore, complex models carry the risk of overfitting, and incorporating market volatility and external factors into the model was a challenge. Future studies should include datasets over a longer time period and additional market indicators to enhance predictive performance. Additionally, ensemble learning techniques and the use of diverse unstructured data should be considered to improve the robustness and accuracy of predictions.

In conclusion, AI and machine learning prediction models are powerful tools for stock price forecasting and optimizing investment strategies. However, this study confirms that these technology-driven approaches are more effective when combined with traditional investment methodologies. This suggests that the synergy between human expertise and machine intelligence can enhance innovation and efficiency in investment management. Continuous model improvement and adaptation to dynamic market conditions are essential, and further research and efforts to overcome these limitations through a broader range of data and techniques are needed. Through such advancements, AI will increasingly play a vital role in addressing the complexities of future financial markets and will become a more sophisticated, data-driven decision-making tool for investors.

Reference

[1] Harish Kunder, B.C. Soundarya, K. Kishan Karyappa, M. Nithin, P.K. Mahesh, R. Sharath, J. Uday Kiran. (2021), Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions, Electronics, 10(21), 2717-2731.

[2] Alkhatib, K., Najadat, H., and Hmeidi, I. (2020), Machine learning approaches in stock market prediction: A systematic review, Expert Systems with Applications, 45(2), 50-60.

[3] Zhong, X., and Enke, D. (2018), Predicting the daily return direction of the stock market using hybrid machine learning algorithms, Financial Research Letters, 25, 30-40.

[4] Kumar, D., Sarangi, P.K., and Verma, R. (2021), A systematic review of stock market prediction using machine learning and statistical techniques, Journal of Financial Analytics, 12(3), 85-98.

[5] Jiang, W. (2019), Applications of deep learning in stock market prediction: recent progress, Journal of Electronic Engineering, 32(1), 15-28.

[6] 홍승환(2009), 주가 예측을 위한 CNN- LSTM 기반 모델 연구, Journal of the Korea Convergence Society, 19(5), 689-694.

[7] 김동영(2015), SNS와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형에 관한 연구, 석사학위논문, 숭실대학교 소프트웨어특성화 대학원.

◐ ABOUT THE AUTHOR ◑

Jumin Lee

LG Electronics / Professional, Global DX Tast

Aalto University Executive Master of Business Administration

Seoul School of Integrated Sciences & Technologies (aSSIST) AI·Big Data Master of Engineering

Areas of Interest : AI, Digital Transformation, AI Ethics, Data Analysis, Smart Farm and etc.

E-mail : jumin.yi@gmail.com

Study on the Effect of Data Augmentation on Image Learning

Tae-hwan Choi

ABSTRACT

This paper aims to demonstrate the effective use of data in AI research by examining the impact of data augmentation on AI learning. This study utilized a subset of the AI Hub's Ocean Deposition Garbage Image dataset. The accuracy of image classification was measured using CNN's VGG16 and VGG19. After increasing the data volume through augmentation, the accuracy was re-measured and re-compared. The results of the study indicated that accuracy improved in most cases, leading to the conclusion that data augmentation positively affects image learning.
This paper is revision of the author's master degree thesis. and This research Ocean Deposition Garbage Image used datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All data information can be accessed through 'AI-Hub (www.aihub.or.kr)'.

☞ keyword : AI, Data Augmentation, Image, CNN, VGG

1. Introduction

Artificial intelligence is considered one of the core technologies of the Fourth Industrial Revolution. AI has been described as a "machine equipped with cognitive, judgment, action, and learning functions."(Choung Hye-Uk, 2021).

AI technology is being applied and utilized in various fields such as education, art, industry, and counseling. So-yeon Park, Chang-yeop Lee, Chan-hyung Ahn(2020) stated in an article that "different institutions classify AI-related technologies in varying ways, leading to differences in both the number and scope of categories. Among the 14 institutions studied, commonly cited technological areas include visual intelligence, language intelligence, machine learning, speech intelligence, robotics, and expert systems. Among these, visual intelligence is described as a field of technology that recognizes image or video data to assess situations or processes data to generate new images or videos."

Among the technological fields, computer vision－an image recognition technology within visual intelligence－is widely utilized. In a study by Song Jae-min, Lee Sae-bom, and Park Ah-reum(2020), various cases are highlighted. For example, when a camera installed beneath a tractor captures soil images, AI analyzes the data to distinguish between weeds and crops, enabling the precise spraying of herbicide only on weeds and applying fertilizer to crops, thereby increasing crop yields. Other examples include tools that utilize images captured by drones to measure stockpile volumes and services that analyze satellite images with AI to identify the latest trends in diverse economic activities occurring worldwide.

Data is a critical element in the utilization of artificial intelligence. Mi-jeong Park(2019) stated, "Most AI applications require vast amounts of available data to learn and make intelligent decisions." Similarly, Hyung-il Koo(2018) observed that "deep learning tends to exhibit improved performance in proportion to the volume of data." Summarizing these insights, it becomes evident that a large quantity of data is essential for effectively utilizing AI technology. In particular, the application of computer vision technology necessitates extensive image data.

However, collecting and accumulating data requires substantial costs, and various challenges, such as issues with the quality of the data collected at great expense, are likely to arise.

Therefore, one approach to addressing these challenges is to use data augmentation techniques. Hee-chan Cho(2020) defined data augmentation as "a technique that increases the amount of data using various algorithms based on a small amount of initial data." By utilizing this technique, it becomes possible to train AI models with a relatively limited

quantity of data.

2. Literature Review

2.1 Data Augmentation

Soo-Young Cho(2020) defined data augmentation as “a technique that generally increases data by randomly transforming the information of images or videos into other forms.” Joo-yeong Song(2021) described it as “a method that uses subsets of the original data, or subsequences, as augmented data.” Jeong-hwa Lim(2023) defined it as “a data augmentation technique that synthesizes and generates data by maximizing the use of the various characteristics of the given data.” Sung-nam Kim(2023) stated that it involves “creating semantically identical new data by transforming the original data in various ways,” while Chan-il Park(2021) described it as “a method that randomly modifies information such as images, text, and signals to increase the data with similar yet distinct information.” Lastly, Jin-sang Lee(2023) referred to it as “a technique for increasing the amount of data by applying various methods in situations where data is scarce.”

In summary, data augmentation can be described as a method of increasing the amount of existing data by generating identical or similar data through the transformation of various types of data in specific ways. To utilize artificial intelligence effectively, it is essential to train models using collected data, which requires a large quantity of data with a high level of quality. However, in cases where data collection is challenging, data augmentation offers advantages by increasing the amount of data available for AI training.

There are various methods of data augmentation. Connor Shorten and Taghi M. Khoshgoftaar(2019) identified techniques such as geometric transformations, color space transformations, kernel filters, mixing images, random erasing, feature space augmentation, adversarial training, GAN-based augmentation, neural style transfer, and meta-learning schemes.

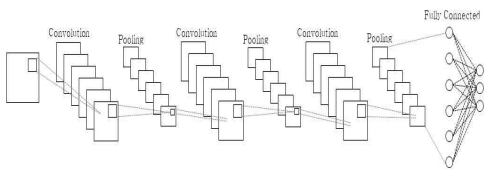
2.2 CNN(Convolution Neural Network)

2.2.1 Definition of CNN

Eun-Ju Lee(2017) defined CNN as “an optimal method for extracting high-level, abstracted features or processing texture information from images.” In-kyeong Hwang(2017) described it as “a type of deep learning that represents a feed-forward artificial neural network, which resembles the pattern connections between neurons that play a crucial role in processing visual information in animals.” Su-Bin Park(2019) defined CNN as “a type of artificial neural network particularly specialized for the field of computer vision, based on convolution operations to extract unique features of images and train these extracted features through the neural network.” Hyeon-Ho Lee(2020) described CNN as “a type of deep neural network(DNN) that uses convolution in one or more of its layers.” Kang-Hyun Park(2021) defined it as “a deep learning algorithm inspired by the human object recognition process, a learning model that adds two types of layers－convolutional layers and pooling layers－to traditional artificial neural networks.” Finally, Gyeong-Won Jang(2023) described it as “a type of DNN that mimics the structure of the human optic nerve.”

CNN is an advanced form of the Deep Neural Network(DNN). The DNN architecture consists of an input layer, multiple hidden layers, and an output layer. Due to the multiple hidden layers, the learning outcomes are enhanced compared to the Artificial Neural Network(ANN). However, in the case of large-scale image or video data, training can require a significant amount of time, and there is an issue of spatial information loss during the flattening process into a one-dimensional form.

The general structure of a CNN is shown in Figure 1 below.



(Figure 1) CNN Structure(Hyeon-Ho Lee, 2020)

CNN is “composed of three key structures: the

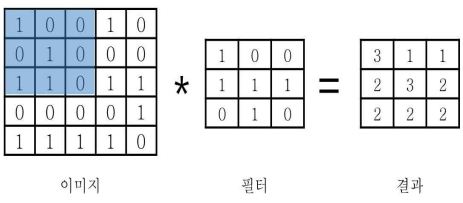
Convolution Layer, Pooling Layer, and Fully Connected Layer"(Ian Goodfellow, et al., 2016; quoted in In-kyeong Hwang, 2017). In CNNs, features are extracted in the Convolution and Pooling Layers, while classification occurs in the Fully Connected Layer.

2.2.2 Convolution Layer

In the Convolution Layer, convolution operations are performed to extract features using filters (also known as kernels). If the stride value is set to 1 during convolution, the filter moves one step at a time. For grayscale images, the channel value is 1, but for RGB images, the channel value is 3. The result of these calculations generates a feature map, with the number of feature maps depending on the number of filters used.

Kwang-Bok Hwang(2021) stated that "the position and interval settings for applying filters to input data, along with the filter size, affect the size of the output data."

The method for calculating convolution can be illustrated with an example: applying a 3×3 filter with a stride of 1 to a 5×5 original image matrix. The result of the convolution operation is shown in Figure 2 below.



(Figure 2) Example of Convolution Operation

The operation is calculated as shown in Figure 3 below, and by shifting one step at a time, the same calculation is repeated to obtain the result.

$$(1 \times 1) + (0 \times 0) + (0 \times 0) + (0 \times 1) + (1 \times 1) + (0 \times 1) + (1 \times 0) + (1 \times 1) + (0 \times 0)$$

(Figure 3) Example of Convolution Operation Method

Min Ae-ri(2020) stated that "the Convolution Layer applies filters to input data to minimize the number of shared parameters and uses nonlinear activation functions (such as ReLU or Sigmoid) to identify features in the image."

2.2.3 Pooling Layer

The Pooling Layer takes the output of the Convolution Layer as input and reduces its size or emphasizes certain parts of the data. Kwang-Bok Hwang(2021) stated that the Pooling Layer "does not contain parameters like the convolution operation and performs independent operations for each channel, generating output data with the same number of channels as the input data." As the number of filters increases, the feature maps grow, resulting in a higher-dimensional parameter count, which can lead to overfitting and increased time requirements for model size and output. Therefore, the Pooling Layer is used to reduce dimensions.

Jong-Bum Kim and Sung-kwun Oh(2015) stated that "in the Pooling Layer stage, the size is reduced by an integer factor, primarily through Max Pooling or Average Pooling."

Hyun-Su Lee(2018) noted that Max Pooling is also known as subsampling. Max Pooling involves applying the maximum value within the filter size and then moving by the stride amount to perform the calculation.

Average Pooling applies the average value within the filter size and moves by the stride amount to perform the calculation. Another method, Min Pooling, applies the minimum value within the filter size and moves by the stride amount to complete the calculation.

2.2.4 Fully Connected Layer

In-kyeong Hwang(2017) stated that "the Fully Connected Layer is similar to a convolutional layer with a filter size of (1×1). It fully connects the features extracted from the previous layers, transforming all activation values into a one-dimensional vector to enable classification."

Gyeong-Won Jang(2023) stated that "as the final layer of CNN, this stage determines the

classification. Each layer is converted into a one-dimensional array, and the final output is determined through an activation function. The size of the final output corresponds to the number of classes of the input data to be classified."

Han-Nah Kim(2020) stated that "the Fully Connected Layer functions to connect all input and output neurons, and the softmax activation function used in this layer performs multi-class classification, outputting probability values for each class."

In other words, the Fully Connected Layer is the final classification stage in the CNN architecture, where classification is performed using the softmax function.

2.3 VGG(Visual Geometry Group)

2.3.1 VGGNet

The ILSVRC (ImageNet Large Scale Visual Recognition Challenge) is one of the major conferences in computer vision, held annually.

VGG is a model that achieved outstanding results in the ILSVRC 2014. In their paper, Karen Simonyan and Andrew Zisserman(2014) developed six different VGGNet architectures to compare their performance. Among these, the structure with 16 layers is known as VGG16, while the one with 19 layers is referred to as VGG19.

Ae-Ri Min(2020) stated that "VGG16 focuses on using convolution layers with 3×3 filters with a stride of 1, instead of relying on a large number of hyperparameters, and consistently uses same-padding and max-pooling layers with 2×2 filters and a stride of 2."

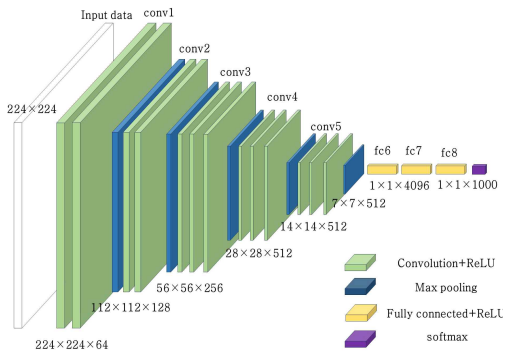
Karen Simonyan and Andrew Zisserman(2014) stated that the input image size for VGG is fixed at 224×224, and the preprocessing involves subtracting the mean RGB value from each image. In the Convolution Layers, 3×3 filters with a stride of 1 pixel are used, with padding applied to preserve resolution. The architecture includes a total of five max-pooling layers, each performed with 2×2 pixels and a stride of 2. The Fully Connected Layer consists of two layers with 4096 channels each, followed by a third layer for classification with 1000 channels, and finally a softmax layer. This structure incorporates ReLU nonlinearity in all hidden layers (Krizhevsky et al., 2012; quoted in Karen Simonyan & Andrew Zisserman, 2014).

2.3.2 VGG16

VGG16, a CNN model proposed by Karen Simonyan and Andrew Zisserman at Oxford University, is one of the renowned models submitted to ILSVRC-2014. While it has drawbacks such as requiring substantial training time and large network weights, it is relatively easy to implement (Liang Han Sheng, 2021).

Jun-Yeol Ryu, Tae-Wan Kim, Ja-Hoon Jeong(2022) stated that "despite its deep 16-layer structure, VGG16 uses 3×3 convolution filters, resulting in a relatively small number of training parameters. This provides greater nonlinearity, making it an efficient model in terms of time and accuracy."

Figure 4 illustrates the structure of VGG16, which is composed of a total of 16 layers: 13 Convolution Layers and 3 Fully Connected Layers.



(Figure 4) VGG16 Structure(Liang Han Sheng, 2021)

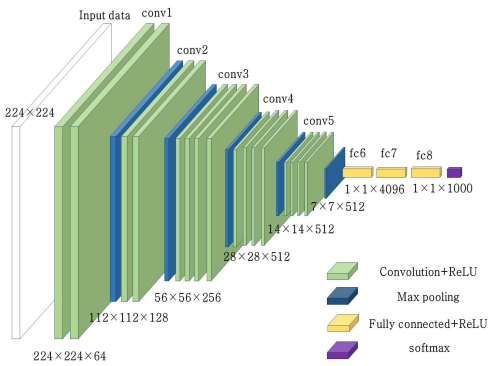
2.3.3 VGG19

(Figure 5) illustrates the structure of VGG19, which consists of a total of 19 layers: 16 Convolution Layers and 3 Fully Connected Layers.

In the study "Performance Comparison of CNN Models for Retinal Disease Diagnosis" by Hye-won Kim, Xuefeng Zhang, Yong-Soo Kim, Il-Hong

Jung(2022), experiments applying VGG16, VGG19, ResNet, and DenseNet showed that the models performed in the following order: DenseNet, ResNet, VGG19, and VGG16.

In Jeong-Hwan Lee's (2021) study, "Detection of Number and Character Areas of Car License Plates Using Deep Learning and Semantic Image Segmentation," ResNet18, ResNet50, VGG16, and VGG19 were used, with results showing performance in the order of ResNet50, ResNet18, VGG19, and VGG16. Thus, prior research indicates that VGG19 performs better than VGG16.



(Figure 5) VGG19 Structure

3. Methods

3.1 Methods

3.1.1 Research Procedure and Method

First, I explored the research topic. To create a robust model through AI learning, a large amount of data is essential. However, obtaining such extensive data proved challenging. Upon investigation, I found that data augmentation is a method that can increase the volume of existing data by utilizing the original data itself. This led me to focus on examining the effects of data augmentation on learning, aiming to provide a foundation for applying data augmentation in future research. Therefore, I decided to investigate whether data exists to support the impact of data augmentation on learning.

Next, I determined which data to use for data augmentation and found that socially relevant topics are of interest. Consequently, I identified the “Marine Submerged Waste Images” dataset available on AI Hub. This dataset includes sonar survey images and underwater photography images. The sonar survey images consist of categories such as tires, traps, nets, wood, and ropes, while the underwater images are categorized into ropes, wood, nets, tires, and traps.

The "Ocean Deposition Garbage Images" dataset from AI Hub was established with funding from the Ministry of Science and ICT and support from the National Information Society Agency of Korea. Having secured suitable data to carry out the explored research topic, I finalized the decision to conduct the research using this dataset.

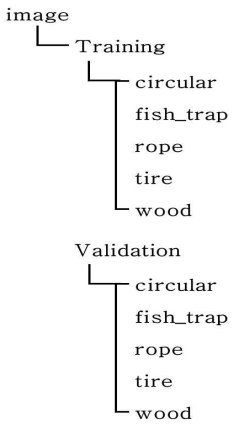
Next, I selected the research method. To examine the impact of data augmentation on images, I decided to measure the differences before and after applying data augmentation. Among deep learning methods, CNN is known for its effectiveness in image processing, so I chose to conduct the research using CNN. Specifically, I selected the VGG model due to its ease of implementation and efficiency. I planned to measure the performance of VGG16 and VGG19 three times each before and after data augmentation, setting the number of epochs—representing the training cycles of the data—to 10 for this study.

In this study, an experiment was conducted to examine the impact of data augmentation on images. The procedure involved applying CNN’s VGG16 and VGG19 models to the original image dataset obtained for the research and measuring accuracy. Next, data augmentation was applied to double the number of images, after which accuracy was measured again using the VGG16 and VGG19 models. Following the completion of implementation and execution, I planned to compare the accuracy before and after data augmentation to assess any performance differences. The data augmentation methods included image rotation, vertical and horizontal translation and transformation, brightness adjustment, flipping, zooming, and shifting.

After selecting the research method, I performed data classification to prepare the data for use. As

shown in Figure 6, images were categorized into folders by type. In this study, only the underwater photography image data were used. In the training data, I focused on underwater images of ropes, wood, nets, tires, and traps, excluding images where multiple objects appeared to be overlapping, due to concerns about quality degradation. I used a randomly selected subset of training data and validation data for the study.

Ultimately, the image data used in the study consisted of 4,618 training images: 1,010 circular, 1,331 fish_trap, 472 rope, 1,028 tire, and 777 wood images. The validation data included a total of 3,008 images: 368 circular, 978 fish_trap, 331 rope, 765 tire, and 566 wood images.



(Figure 6) Data Classification

After completing data classification, the VGG16 and VGG19 models were implemented, and the code was executed to measure the pre-selected performance metrics. Finally, the results were interpreted based on the metrics obtained from the study.

3.1.2 Research Environment

The environment in which this study was conducted is shown in Table 1 below.

NVIDIA's CUDA (Compute Unified Device Architecture) is a development environment tool for creating high-performance GPU-accelerated applications and enabling GPU acceleration.

Previously, to perform general parallel computing, programmers had to configure every detail of the GPU. To address this issue, NVIDIA enhanced the flexibility of the traditional Geforce series, transforming it from a graphics-specific architecture to one capable of supporting general-purpose programming(Ji-Hoon Jo, 2013).

(표 1) Research Environment

항 목	내 용
CPU	AMD Ryzen5 7600
Memory	DDR5-5600(2800 ㎒) 16G * 2EA
Graphics	Geforce RTX 3090 24G Driver Version : 546.33
OS	Windows 10 64bit
Python	Anaconda3 2023.09-0(64-bit) Python 3.8.18
Tensorflow	2.8.0
Tensorflow-gpu	2.4.1
cuda	CUDA 11.0
Library	cuDNN 8.0.4

cuDNN (CUDA Deep Neural Network) is a GPU-accelerated library for deep neural networks that facilitates the implementation of standardized operations such as convolution, attention, matrix multiplication, pooling, and normalization (NVIDIA Developer).

The supported CUDA version varies depending on the GPU in the development environment, and cuDNN must be installed to match the installed Python and CUDA versions, so careful attention to compatibility is required.

3.2 Evaluation Metrics

The metric defined to verify the effectiveness of this study is accuracy. Accuracy represents the number of successful predictions out of the total data, with higher accuracy indicating better predictive performance. Since accuracy is a key metric for model evaluation, it was selected as the primary measurement indicator.

Accuracy refers to the number of successful predictions out of the total data. It can be calculated using the following formula:

$$Accuravy = \frac{TP + TN}{TP + FN + FP + TN}$$

TP(True Positive) refers to cases where the actual value is true, and the predicted value is positive, meaning that the actual and predicted values match. TN(True Negative) occurs when the actual value is false, and the predicted value is negative, indicating that the actual and predicted values match. TP and TN represent successful predictions where the actual and predicted values align. FP(False Positive) is when the actual value is false, but the predicted value is positive, meaning the actual and predicted values differ. FN(False Negative) occurs when the actual value is true, but the predicted value is negative, also indicating a mismatch. FP and FN represent failed predictions where the actual and predicted values do not align.

Dividing the total data TP+FN+FP+TN by the value of TP+TN which represents correct predictions, yields the accuracy. A higher accuracy value indicates greater predictive accuracy. However, care should be taken when applying accuracy to datasets with imbalanced data, as it may not fully reflect the model's performance in such cases.

4. Results

4.1 Performance Measurement Before Applying Data Augmentation

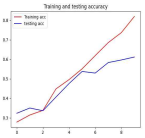
4.1.1 VGG-16

Table 2 shows the performance measurements of VGG16 before applying data augmentation. In this table, accuracy is divided into two categories: accuracy represents the accuracy on the training data, and val_accuracy represents the accuracy on the validation data. Similarly, loss denotes the loss value for the training data, and val_loss indicates the loss value for the validation data.

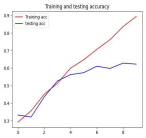
(Table 2) Performance Measurement Results of VGG16 Before Applying Data Augmentation

epoch	1		2		3	
	accuracy	val_	accuracy	val_	accuracy	val_
1	0.2735	0.3235	0.2921	0.3328	0.2739	0.3251
2	0.3224	0.3504	0.3234	0.3211	0.2880	0.3511
3	0.3344	0.3364	0.4428	0.4362	0.3031	0.3916
4	0.4169	0.4066	0.5004	0.5266	0.3961	0.3810
5	0.4904	0.4757	0.5990	0.5632	0.4542	0.4315
6	0.5493	0.5376	0.6292	0.5741	0.5379	0.4581
7	0.6177	0.5293	0.7068	0.6110	0.5645	0.5259
8	0.6743	0.5831	0.7684	0.5984	0.6385	0.6011
9	0.7491	0.5964	0.8355	0.6287	0.6955	0.6077
10	0.8201	0.6114	0.8996	0.6230	0.7753	0.6147

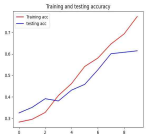
Figures 7, 8, and 9 graphically represent the results from Table 2. Observing the trends in the graphs, we can see that all values increase as epochs progress. This indicates that accuracy improves with the number of training iterations.



(Figure 7) First Measurement of VGG16 Accuracy Before Applying Data Augmentation



(Figure 8) Second Measurement of VGG16 Accuracy Before Applying Data Augmentation



(Figure 9) Third Measurement of VGG16 Accuracy Before Applying Data Augmentation

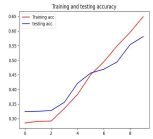
4.1.2 VGG-19

Table 3 shows the performance measurements of VGG19 before applying data augmentation, and Figures 10, 11, and 12 graphically represent these results. Observing the trends in the measured results, all three sets of results generally show an upward trend. Thus, accuracy increases with the number of training iterations.

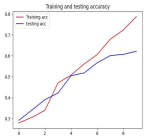
(Table 3) Performance Measurement Results of VGG19 Before Applying Data Augmentation

epoch	1		2		3	
	accuracy	val_ accuracy	accuracy	val_ accuracy	accuracy	val_ accuracy
1	0.2712	0.3245	0.2724	0.2916	0.2684	0.2543
2	0.2944	0.3251	0.2842	0.3404	0.2661	0.3251
3	0.2887	0.3281	0.3374	0.3903	0.2996	0.3022
4	0.3226	0.3561	0.4579	0.4219	0.3986	0.4146

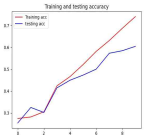
5	0.3772	0.4209	0.4896	0.5050	0.4739	0.4495
6	0.4458	0.4561	0.5462	0.5183	0.5216	0.4727
7	0.4947	0.4691	0.5969	0.5665	0.5797	0.5000
8	0.5403	0.4940	0.6789	0.6011	0.6208	0.5725
9	0.5787	0.5532	0.7189	0.6070	0.6871	0.5841
10	0.6566	0.5808	0.7941	0.6217	0.7252	0.6041



(Figure 10) First Measurement of VGG19 Accuracy Before Applying Data Augmentation



(Figure 11) Second Measurement of VGG19 Accuracy Before Applying Data Augmentation



(Figure 13) Third Measurement of VGG19 Accuracy Before Applying Data Augmentation

4.2 Performance Measurement After Applying Data Augmentation

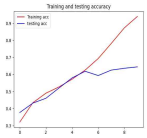
Data augmentation was applied to double the training data by adding transformations such as image rotation, vertical and horizontal shifts, and scaling. The VGG16 and VGG19 models were then applied to measure performance. For accurate assessment, performance was measured a total of three times.

4.2.1 VGG-16

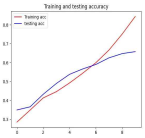
Table 4 shows the performance measurements of VGG16 after applying data augmentation, and Figures 13, 14, and 15 graphically represent these results. Observing the measurement results, all values generally increase as epochs progress, indicating that accuracy improves with the number of training iterations.

(Table 4) Performance Measurement Results of VGG16 After Applying Data Augmentation

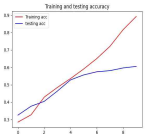
epoch	1		2		3	
	accuracy	val_ accuracy	accuracy	val_ accuracy	accuracy	val_ accuracy
1	0.2977	0.3767	0.2666	0.3494	0.2778	0.3251
2	0.4313	0.4318	0.3381	0.3660	0.2993	0.3763
3	0.4850	0.4601	0.4090	0.4325	0.4219	0.4043
4	0.5223	0.5203	0.4389	0.4897	0.4802	0.4634
5	0.5648	0.5795	0.4834	0.5379	0.5256	0.5276
6	0.6142	0.6184	0.5399	0.5665	0.5796	0.5562
7	0.6985	0.5934	0.5952	0.5901	0.6520	0.5745
8	0.7814	0.6253	0.6622	0.6253	0.7251	0.5808
9	0.8733	0.6360	0.7594	0.6476	0.8091	0.5964
10	0.9373	0.6443	0.8480	0.6582	0.8963	0.6054



(Figure 13) First Measurement of VGG16 Accuracy After Applying Data Augmentation



(Figure 14) Second Measurement of VGG16 Accuracy After Applying Data Augmentation



(Figure 15) Third Measurement of VGG16 Accuracy After Applying Data Augmentation

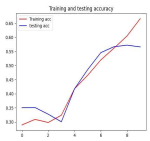
4.2.2 VGG-19

Table 5 shows the performance measurement results of VGG19 after applying data augmentation, and these results are represented graphically.

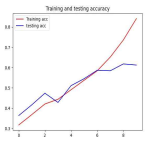
(Table 5) Performance Measurement Results of VGG19 After Applying Data Augmentation

epoch	1		2		3	
	accuracy	val_ accuracy	accuracy	val_ accuracy	accuracy	val_ accuracy
1	0.2739	0.3504	0.2908	0.3637	0.2741	0.3354
2	0.3119	0.3511	0.3672	0.4166	0.2983	0.3444
3	0.3041	0.3275	0.4200	0.4737	0.3114	0.3431
4	0.3134	0.3005	0.4469	0.4282	0.3612	0.4189
5	0.3971	0.4176	0.4824	0.5110	0.4354	0.4801
6	0.4512	0.4857	0.5325	0.5452	0.4844	0.5027
7	0.5129	0.5459	0.5838	0.5868	0.5090	0.5572
8	0.5619	0.5668	0.6511	0.5841	0.5805	0.5509
9	0.6021	0.5725	0.7373	0.6180	0.6311	0.5848
10	0.6697	0.5662	0.8475	0.6124	0.6848	0.6074

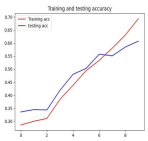
Figures 16, 17, and 18 graphically represent these results. Observing the trend in the graphs, accuracy increases with epochs for both the training and validation data, indicating that accuracy improves with the number of training iterations.



(Figure 16) First Measurement of VGG19 Accuracy After Applying Data Augmentation



(Figure 17) Second Measurement of VGG19 Accuracy After Applying Data Augmentation



(Figure 18) Third Measurement of VGG19 Accuracy After Applying Data Augmentation

5. Conclusion

5.1 Research Conclusion

This study aimed to determine the impact of data augmentation on image learning by comparing the results of CNN's VGG16 and VGG19 models before and after data augmentation.

The results, following the research sequence, are summarized as follows:

First, before applying data augmentation, accuracy increased with each epoch for both the VGG16 and VGG19 models. In the experiments with the VGG16 model, the highest accuracy at epoch 10 was 0.8996 for the training data and 0.6230 for the validation data. In the VGG19 model, the highest accuracy at epoch 10 was 0.7941 for the training data and 0.6217 for the validation data.

Second, after applying data augmentation, accuracy continued to increase with each epoch for both the VGG16 and VGG19 models. For the VGG16 model, the highest accuracy at epoch 10 was 0.9373 for the training data and 0.6582 for the validation data. For the VGG19 model, the highest accuracy at epoch 10 was 0.8475 for the training data and 0.6124 for the validation data.

(Table 6) Summary of Research Results Before and After Data Augmentation

	Before Data Augmentation	After Data Augmentation
VGG16 - Accuracy	0.8996	0.9373
VGG16 – val_Accuracy	0.6230	0.6582
VGG19 - Accuracy	0.7941	0.8475
VGG19 – val_Accuracy	0.6217	0.6124

The results show that VGG16's accuracy improved after applying data augmentation compared to before. For VGG19, the training data accuracy also increased post-augmentation, while the validation data accuracy showed a slight decrease, though this difference was minimal. Overall, most cases demonstrated an improvement in accuracy.

Consequently, it was concluded that applying data augmentation to image learning has a positive impact on model performance. Additionally, the generally lower accuracy of VGG19 compared to VGG16 appears to be related to the specific characteristics of the image data used in this study.

5.2 Limitations of Study

This study examines the impact of data augmentation on image learning, specifically using marine submerged waste images. The results may reflect the quality and characteristics of the image data applied in this context.

The study is also limited by the number of images used; thus, the results may vary depending on the amount of image data available for training. Additionally, duplicated instances of marine waste during the image classification process could have influenced the outcomes.

Since the CNN VGG model was applied to the image learning model, there are limitations in that other models, such as LeNet, AlexNet, ZFNet, GoogleNet, ResNet, or DenseNet, might yield different results.

Among the various data augmentation techniques, only specific methods were used, presenting a limitation in not applying a broader range of augmentation techniques.

Finally, image learning is heavily influenced by computing specifications, indicating potential limitations based on the research environment.

5.3 Further research

This study investigated the impact of data augmentation on image learning and concluded that

it has a positive influence. Based on these results, applying data augmentation can effectively expand limited datasets when utilizing image data in AI applications.

During this study, several areas for further in-depth research were identified, leading to the following recommendations:

First, additional research is needed to examine the impact of data augmentation on image learning using different types of image data than those employed in this study.

Second, further studies should explore the effects of applying other models, beyond those used in this research, on image learning.

Third, it would be beneficial to investigate whether data augmentation has an impact on other types of data, such as text and audio.

Fourth, conducting experiments in different research environments from those in this study is necessary to assess the extent to which results are influenced by the environment.

Finally, in this study, VGG19 was initially expected to achieve higher accuracy than VGG16. However, the actual results showed VGG19 had lower accuracy than VGG16. Therefore, further research is needed to determine which model is most suitable for specific contexts.

Reference

Hye-Uk Choung. (2021), Criminal Liability of Artificial Intelligence, *Chung-Ang Law Association,* 23. 4. 53-80.

So-yeon Park, Chang-yeop Lee, Chan-hyung Ahn. (2020), AI: Present and Future – Part 1. How is Artificial Intelligence Technology Classified?, 2econsulting, https://www.2e.co.kr/news/articleView.html?idxno=300957.

Jamin Song, Sae Bom Lee, Arum Park. (2020), A Study on the Industrial Application of Image Recognition Technology, *The Journal of the Korea Contents Association*, 20. 7. 86-96.

Mi Jeong Park. (2019), A Study on Artificial Intelligence and Data Ethics: Focusing on Health Data Used in Artificial Intelligence, *Korea J Med Ethics,* 22. 3. 255-273.

Hyung Il Koo. (2018), A Study on Artificial Intelligence and Data Ethics: Focusing on Health Data Used in Artificial Intelligence, *Korea J Med Ethics,* 22. 3. 255-273.

Hee-chan Cho. (2020), A layerd-wise data augmenting algorithm for small sampling data, Graduate School of Information Security Korea University.

Soo-Young Cho. (2020), Data augmentations to improve deep learning network model accuracy, Kwangwoon University.

Joo-Yeong Song. (2021), A Study on the Improvement of Sequential Recommender System through Data Augmentation, Graduate School of Convergence Science and Technology Seoul National University.

Jeong-Hwa Lim. (2023), A Study on ECG Augmentation to Improve Deep Learning Performance, The Graduate School Yonsei University.

Sung-Nam Kim. (2023), A study on Korean text data augmentation techniques through sentences summarization for document classfication and sentiment analysis, Kyunghee University.

Chan-Il Park. (2021), Data augmentation for sign language translation AI learning, The Graduate School Kwangwoon University.

Jin-Sang Lee. (2023), Research on Korean Sentiment Classification Model Using Data Augmentation, Korea University.

Connor Shorten., Taghi M. Khoshgoftaar. (2019), “A survey on Image Data Augmentation for Deep Learning”, *Journal of Big Data.*

Eun-Ju Lee. (2017), CNN과 RNN의 기초 및 응용 연구, *Broadcasting and media magazine,* 22. 1. 81-89.

In-Kyeong Hwang. (2017), Study on Hangul font characteristics using CNN, The Graduate School Seoul National University.

Ian Goodfellow., Yoshua Bengio., Aaron Courville. (2016), "Deep Learning", MIT Press, https://www.deeplearningbook.org/.

Su-Bin Park. (2019), Implementation of Multiple Object Detection and Recognition System using H-SSD based on VGG Network, Graduate School Don-A University.

Hyeon-Ho Lee. (2020), CNN Generalization Error Evaluation Method, The Graduate School Pusan National University.

Kang-Hyun Park. (2021), Displavement prediction of structures using Signal Image data Based CNN(SIBCNN), The Graduate School Yonsei University.

Gyeong-Won Jang. (2023), Exploring Optimization Techniques for CNN in Binary Image Classification, The Graduate Seoul National University of Science and Technology.

Kwang-Bok Hwang. (2021), A Study on the Design of Steering Controller for Autonomous Vehicle with Variable Speed Using CNN, Graduate School Gyeongnam National University of Science and Technology.

Ae-Ri Min. (2020), A Study on Transfer Learning for Image Classification of Convolutional Neural Network – Based on VGG16 Deep Convolutional Neural Network, The Graduate School Hansei University.

Jong-Bum Kim, Sung-Kwun Oh. (2015), Design of Convolutional RBFNNs Pattern Classifier for Two dimensional Face Recognition, *The Korean Institute of Electrical Engineers*, 1355-1356.

Hyun-Su Lee. (2018), A Structure of Convolutional Neural Networks for Image Contents Search, Graduate School of Chung-Ang University.

Han-Nah Kim. (2020), Sentiment Analysis and Classification using Convolutional Neural Networks, Graduate School of Soonchunhyang University.

Karen Simonyan., Andrew Zisserman. (2014), "Very Deep Convolutional Networs for Large-Scale Image Recognition", *arXiv:1409.1556*.

Alex Krizhevsky., Ilya Sutskever., Geoffrey E. Hinton. (2012), "ImageNet Classification with Deep Convolutional Neural Networks", *NIPS,* 1106-1114.

Liang Han Sheng. (2021), "VGGNet – Convolutional Network for clssification and Detection", https://medium.com/analytics-vidhya/vggnet-convolutional-network-for-classification-and-detection-3543aaf61699.

Jun-Yeol Ryu, Tae-Wan Kim, Ja-Hoon Jeong. (2022), An AI Model for classifying The State of Armored and Mechanized Weapon Systems Based on VGG16, *Journal of the Korea Academia-Industrial cooperation Society*, 23. 12. 741-748.

Hye-won Kim, Xuefeng Zhang, Yong-Soo Kim, Il-Hong Jung. (2022), Comparison of the Performance of CNN Models for Retinal Diseases Diagnosis, *Journal of Korean Institute of Intelligent Systems*, 32. 1. 51-60.

Jeong-Hwan Lee. (2021), Detection of Number and Character Area of License Plate Using Deep Learning and Semantic Image Segmentation, *Journal of the Korea Convergence Society*, 12. 1. 29-35.

Ji-Hoon Jo. (2013), An Implementation of a smart camera system using CUDA Based parallel CAM shift Algorithm, Hannam University.

◐ ABOUT THE AUTHOR ◑

Tae-hwan Choi

A Seoul School of Integrated Sciences & Technologies(aSSIST) M.S

Areas of interest : AI, Data Analysis, Cloud, Computer Vision, Natural Language Processing, Data Mining etc.

E-mail : suhosin@kakao.com

The Prioritization of Evaluation Indications for Artificial Intelligence (AI) Business Capabilities in the Financial Sector: Focusing on the Global AI Index

Ahn, Suho[1] Oh, Tae Yeon[2] Kim, Yongkyu[3]*

ABSTRACT

The financial sector is actively considering the adoption of artificial intelligence (AI) in various areas. AI technologies are being introduced not only in key areas of finance, such as marketing, product development, credit evaluation, and risk management, but also in human resources and performance management. As a result, the need to understand business capabilities to achieve successful investment outcomes, including evaluating investment effects and strategies, is increasing. From this perspective, this study aims to identify the capabilities that should be prioritized when investing in AI businesses in the financial sector, based on the Global AI Index, which is a benchmark used for evaluating AI capabilities in 62 countries worldwide. By applying the Analytic Hierarchy Process (AHP) to a group of financial sector experts, the study selected priority rankings for each specific indicator and ultimately identified the priorities for capabilities across seven areas. This study is expected to contribute to minimizing errors and preventing risks when investing in AI within the financial industry.

☞ keyword : Artificial Intelligent,Financial IT Investment,AHP, Competency Evaluation index, Global AI Index

1. Introduction

The use of artificial intelligence (AI) in the financial sector has been steadily increasing. According to a global survey conducted by NVIDIA on the adoption of AI in the financial industry, the utilization rate of AI in this sector reached 78% in 2021, significantly higher than the average across all industries. This trend is expected to maintain its growth trajectory in the future.

A similar growth trend is observed in the domestic financial AI market. The market size increased from 300 billion KRW in 2019 to 600 billion KRW in 2021, reflecting a 100.0% increase. Furthermore, it is projected to expand at an average annual growth rate of 38.2%, reaching a market size of 3.2 trillion KRW by 2026. These figures underscore the rapid adoption and growing importance of AI technologies in the financial sector, both globally and domestically.[24]

The majority of financial institutions have completed the digital transformation of their value chains, and various approaches are being considered for the integration of AI technology. In South Korea, the primary applications of AI are being explored in marketing, product development, credit rating, risk management, human resources, and performance evaluation. Additionally, the scope of machine learning is being expanded. The construction of data platforms and the implementation of machine learning algorithms serve as the foundation for process automation, enhancement of customer experiences, detection of fraudulent transactions, management of insurance contracts, assessment of creditworthiness, asset management and trading.[2] Consequently, investment in financial AI is on the rise. According to a KPMG economic research report, the number of the world's top 50 banks investing in AI-related fields has grown exponentially since 2016, driven by the acceleration of digital transformation. This growth has resulted in a substantial shift in the investment landscape, with the proportion of investments in this sector reaching 4% in 2023, a remarkable increase from just 0.2% a decade ago. Among them, venture capital investments accounted for 90.6% of the total investments, as reported in a study.[25] In Korea, there are efforts to customize existing services through prompt engineering or fine-tuning for independent development or to train models using domain-specific data. However, significant challenges arise in applying these to actual business operations.

These various attempts to adopt AI in the financial sector appear to be driven by market demands. In response, the Financial Services Commission (FSC)

introduced the "AI Guidelines for the Financial Sector" in July 2021, requiring financial institutions to develop their own capabilities for operating AI systems.

According to the guidelines, financial institutions are required to enhance organizational value and develop their own capabilities by considering AI utilization scenarios. The guidelines emphasize the need for ensuring reliability throughout the entire lifecycle of AI systems, including planning/design, evaluation/validation, deployment/operation, and monitoring, in the context of development, commercialization, and application of AI systems. Similarly, the European Central Bank (ECB) has established a dedicated unit to oversee data and AI initiatives and is working to introduce separate processes to evaluate these efforts.[26]

Such capabilities are being developed gradually and are recognized as essential for establishing evaluation criteria and fostering competencies for the internal utilization of AI. These include tasks such as operating systems to protect consumer rights, managing AI models and training data, communicating with regulatory authorities in the event of issues related to AI systems, and promoting a culture of accountability for AI within organizations.[2]

The purpose of my study is to identify the competency indicators required for AI-related projects in the financial sector and to prioritize them. Identifying and prioritizing these indicators aims to accelerate the adoption of artificial intelligence in the financial sector and strengthen the operational capabilities needed to support it effectively.

This study is expected to help financial institutions achieve effective investment outcomes in AI-related investments by identifying the key competencies they should prioritize.

The structure of the study is as follows: Chapter 2 reviews the literature on AI risks and related indices. Chapter 3 outlines the research methodology and presents the results of the Analytic Hierarchy Process (AHP). Chapter 4 provides the conclusion, implications and the limitations of this study.

2. Literature Review

This chapter examines key literature related to AI index in the financial sector. Through a review of relevant studies on index for assessing the business capabilities of AI in the financial sector, a set of baseline index was selected, and further literature was explored to define these index in detail. The literature review was conducted in two directions: Selection of Capability index: Identifying relevant index to evaluate AI capabilities in the financial sector. Definition of Detailed index: Exploring related literature to establish precise definitions of the selected index. The results of this literature review were utilized in Chapter 3 to conduct an AHP (Analytic Hierarchy Process) study aimed at determining priority rankings.

2.1 Indicator Literature Review

2.1.1 Global AI Index

Since 2019, the UK-based Tortoise Media has utilized the Global AI Index to evaluate the AI capabilities of countries. By 2023, the index had been applied to assess AI competencies across 62 countries. The Global AI Index categorizes AI capabilities into three main areas: implementation, innovation, and investment, which are further divided into subcategories such as talent, infrastructure, operating environment, research, development, government strategy, and commercial investment.

This index is recognized as a key benchmark for comparing and assessing the progress of AI technologies across nations. It provides a comprehensive analysis of various aspects of AI, including research, development, commercialization, infrastructure, and government strategies, to evaluate the AI competencies of individual countries.

The Global AI Index plays a significant role in identifying the AI capabilities of countries and comparing their global competitiveness in this rapidly evolving field.[8]

2.1.2 AI Vibrancy Tool

The Global AI Vibrancy Tool is a metric developed by Stanford University's Human-Centered AI Institute (HAI). This tool focuses on studying the impact of AI technologies on humanity and analyzing the social, economic, and ethical aspects of human-centered AI.

Based on these analyses, the tool provides recommendations for AI policies to ensure that the technology aligns with human values and serves societal interests. Designed as an advanced framework to evaluate and compare the development and adoption of artificial intelligence technologies across various countries and regions, this tool utilizes diverse index such as research, talent, infrastructure, industrial applications, policy and regulation, and social impact to assess AI capabilities.

The development of this tool involves the collaboration of experts from various universities, research institutions, and companies, who work together to establish methodologies for data collection, analysis, and evaluation. Additionally, several international organizations, government agencies and non-profit entities contribute to its development and implementation. These organizations support the initiative by shaping AI policies, providing research assistance and offering data to enhance the tool's effectiveness.

This tool was developed to comprehensively evaluate the development and adoption of AI technologies, aiming to provide objective and reliable data through the participation of various stakeholders.

Based on this foundation, the Global AI Vibrancy Tool plays a crucial role in understanding and comparing global AI trends. It provides essential insights for formulating AI strategies and is used to evaluatethe AI capabilities of various countries and regions.[9]

2.1.3 Evident AI Index

To assess the AI capabilities within the banking sector, this index evaluates AI competencies using publicly available external data. As of November 2023, the initiative is led by the UK-based fintech company 'EvidentInsight.com' and focuses on key metrics such as talent, innovation, and leadership to gauge the business capabilities of banks.

The index evaluates the AI commercialization capabilities of the top 50 global banks, in collaboration with experts from the banking, technology, and benchmarking sectors. This joint effort offers a comparative overview of the AI competencies of these leading financial institutions.[27]

2.1.4 AWS AI Adoption framework

The AWS (Amazon Web Services) framework proposes a methodology for the successful adoption of AI, taking into account the diverse organizational and cultural contexts that influence the adoption process. Cloud-based AI technology adoption

This framework presents a set of decision-making criteria for the adoption of AI technology in business contexts, with a particular focus on the outcomes and inter-organizational implications of such adoption.[10]

As a result of a literature review on related index, the Global AI Indexwas chosen as the benchmark capability indicator for evaluating AI in the financial sector (Table 1). The rationale for this selection is as follows: The financial sector is increasingly required to ensure responsible AI service operations to address risks such as data bias, discrimination, and data breaches.[28] European financial institutions are adapting to AI governance adoption by fostering collaboration between academia and industry partners. Efforts include enhancing sustainability through interactions among environment, organizations, and systems, as well as forming executive teams, securing organizational structures, and developing specialized talent.[29] Third, the domestic financial sector is also taking steps to prepare for more comprehensive management of AI competency index to support the commercialization of AI initiatives.[25] Considering these factors, along with expert opinions, the Global AI Index—a comprehensive and widely recognized indicator for evaluating AI capabilities across nations—was selected as the benchmark capability indices.

2.2 Index Definition

2.2.1 Implementation

When seeking investments for the commercialization of artificial intelligence, the foremost critical factor is the extent to which an organization possesses skilled personnel and has well-established structures to nurture and manage them. It is essential to cultivate internal teams capable of mastering AI technologies or to

effectively leverage external expertise. Organizations must ensure they retain skilled professionals by implementing regular training programs, conducting evaluations, and facilitating role transitions based on performance and evolving business needs.[8] In the financial sector, as investments in digital transformation accelerate, workforce restructuring and organizational changes are intensifying, resulting in shifts in personnel composition. Automated algorithm-based trading introduces new types of risks, while competition and collaboration between financial and non-financial industries are becoming increasingly dynamic.

In the financial sector, investments in digital transformation are accelerating, leading to intensified workforce restructuring and organizational changes, which are beginning to alter personnel composition. Automated algorithm-based trading introduces new risks, while competition and collaboration between financial and non-financial industries are becoming increasingly dynamic.

Services such as robo-advisors provide customer support without the need for specialized personnel. However, internal organizational limitations often hinder their widespread adoption. To enhance the utilization of artificial intelligence, high-quality datasets are essential. Yet, regulatory restrictions and limitations on data usage exacerbate the shortage of skilled professionals capable of effectively leveraging these resources.

Moreover, there is an even greater scarcity of personnel proficient in handling big data or working with AI algorithms, a challenge that is particularly severe for regional banks compared to major commercial banks.[2]

Secondly, infrastructure serves as a critical criterion, as establishing an adequate infrastructure environment is essential for promoting development. In financial operations, the primary areas where AI is utilized include enhancing call center efficiency, streamlining internal workflows, automating responses, improving customer interactions, managing investments and assets, evaluating credit, and ensuring compliance.

When considering the process of identifying areas for improvement and driving performance outcomes through investments, there is a concern that overly optimistic expectations or improper utilization may result in excessively high infrastructure investment costs. Therefore, a balanced approach is necessary to ensure that investments are both efficient and aligned with achievable outcomes.

Thirdly, it is essential to determine whether there is a well-defined process for workforce transition in response to changes in the work environment related to artificial intelligence. This includes assessing whether HR and organizational structures support such transitions and whether short-term and long-term programs are in place to cultivate talent and facilitate workforce adjustments. It also involves examining the existence of specific KPIs (Key Performance index)for workforce transitions or processes that enable these transitions effectively.

In particular, applying AI to financial sector operations requires an understanding of appropriate regulations and the ability to translate them into actionable and compliant workflows. Regulatory compliance and financial servicesare key requirements, making it necessary to consider specific legal frameworks alongside technological development. Policymakers in the financial industry must comprehend both technical and regulatory aspects to develop and expand services that align with these considerations.

An approach that fails to comply with regulations could expose critical financial systems to significant risks and uncertainties. Conversely, excessive regulation may stifle innovation and hinder the potential for AI technology adoption in the financial sector. Striking a balance between these diverse interests and considerations is essential, necessitating a careful integration of multiple perspectives and existing domain expertise to effectively promote the adoption of AI technologies.

It is crucial to ensure that all stakeholders in the financial sector possess a consistent understanding of AI and its applications. This allows for efficient integration of AI into their respective functions. Continuous innovation is required to bridge knowledge gaps and ensure that AI technologies are implemented effectively, driving both compliance and progress in the financial industry.[14]

2.2.2 Innovation

When making investment decisions in artificial intelligence, it is important to distinguish between

research and development as key innovative factors.

Research involves identifying and nurturing future technologies while establishing a foundation for strategic planning and investment decisions. To achieve this, it is essential to have specialized organizations, such as dedicated research laboratories, to carry out research activities. Currently, most financial organizations focus on applying artificial intelligence (AI) technologies, with an emphasis on the following areas: targeting customers for marketing purposes, analyzing customer profiles, enhancing customer engagement through chatbots, conducting credit checks, evaluations, and rating adjustments in investment transactions, and detecting anomalies for risk management.

To support these applications and advance future innovations, research labs need to focus on developing future technologies. These labs can explore and expand AI capabilities in fields such as Real-time predictive analytics, Identifying patterns and predicting outcomes in real-time.

Machine learningand deep learning are advancing the capabilities of AI models. Vision technologies for video analysis utilize AI to provide video-based insights. Graph analysis employs AI for data visualization and to illustrate complex graphical relationships. Natural Language Processing (NLP) involves the development of systems capable of understanding and generating human language.

By investing in these research areas, organizations can establish a robust foundation for future AI applications, ensuring continuous innovation and maintaining competitiveness in the financial sector.[15]

Development focuses on activities and proposals aimed at applying and implementing new AI technologies in real-world operations. This field promotes the adoption of innovative technologies and identifies effective applications within business contexts. The primary emphasis is on creating applied services and conducting development activities that facilitate their practical use.

A key element in development is data. Creating proprietary AI models can be challenging, making it essential to effectively utilize data while adhering to regulatory compliance. Preparing for development requires identifying applicable services and ensuring that they are built on a solid data foundation.

Development efforts should be approached from both short-term and long-term perspectives, with strategies tailored to address immediate needs as well as future goals. Additionally, fostering an organizational culture that adapts to and supports development practices is crucial. Encouraging teams to embrace and align with a development-focused culture will be essential for the successful integration of AI technologies.[17] In particular, Article 15, Paragraph 1 of the Financial Electronic Supervision Regulations stipulates that information processing systems located within data centers, as well as terminals that directly access these systems for purposes such as operation, development, and security, must be physically separated from external communication networks, such as the internet. This network separation policy presents significant challenges in utilizing data effectively, highlighting a real obstacle for AI development and application in the financial sector.[18]

Such a research environment highlights the need for clear guidelines on the role it plays within the realms of social responsibility, ethical conduct, and sustainability. The ethical necessity of responsibly utilizing sensitive data in AI applications is widely acknowledged, emphasizing that research and implementation must be conducted with greater care to support the universal management principle of sustainability.

Specific issues such as algorithmic bias, facial recognition, and monitoring decision-making processes on critical matters are recognized as challenges that must be addressed to develop ethically sound applications. By adhering to ethical standards, companies not only comply with these recommendations but also build higher levels of trust and gain a competitive edge, ensuring that sustainable management practices are maintained.[19]

2.2.3 Investment

In implementing artificial intelligence, AI-related strategies or plans, along with an adequate level of investment, are critically important.

Regarding AI-driven business models, Peter Weilland Stephanie Woerner, professors at the MIT Sloan School of Management, introduced the concept of Digital

Business Models in the AI era. Their framework emphasizes the strategic integration of AI technologies into business operations to enable transformation and value creation in a digital landscape. This concept serves as a guide for aligning investment decisions with AI strategies to maximize business impact.[20] The Digital Business Model combines three key elements—content, customer experience, and platform—to create customer value. In this context, Content (Data) encompasses not only digitalized products but also information about these products. Data is recognized as the most critical resource in the AI era.

A platform refers to the foundation for delivering services, serving as a space where customers can analyze and utilize data, thereby creating a pathway for generating new value.

Customer experience focuses on how customers interact with the content and platform, shaping their overall perception and engagement.

To implement and deliver these three elements, significant investment in infrastructure, skilled personnel, and a supportive organizational culture is essential. In the AI business landscape, the platform plays a pivotal role as the venue for creating and delivering value. For example, AI technologies such as image processing, voice recognition, natural language processing, and recommendation systems are utilized to develop services that meet customer demands and provide tailored products.

Online reviews and customer experiences generated on such platforms become sources of competitive advantage, enabling the business model to focus on value creation and profitability. This process demonstrates how businesses can leverage big data, select appropriate platforms, determine pricing models, and seek innovative ways to materialize profits.

These activities not only enhance the competitive edge but also contribute to building a new business ecosystem, where strategic investments in AI technologies, data utilization, and platform development are indispensable for sustainable success.[21]

According to a survey conducted by the Korea Institute of Finance targeting domestic commercial banks, the key areas of AI adoption in the financial sector include "credit evaluation and loan screening"and "risk monitoring," with utilization rates reaching 87.5%. The adoption of chatbots was also significant, at 75%.

However, the survey identified several challenges hindering AI implementation. Data-related issues, Cited by 75% of respondents. Shortage of skilled professionals, Highlighted by 62.5%. Regulatory constraints, Identified as a challenge by 75%. These findings underline the need for strategic investments in data infrastructure, workforce development, and regulatory reforms to enhance AI adoption in the financial sector.

In the areas of credit evaluation and loan screening, AI recommendations are being implemented the most rapidly in South Korea. AI is widely utilized for tasks such as calculating credit scores, approving interest rates, adjusting credit limits, and performing error analysis.

In the field of risk monitoring, AI plays a significant role in applications such as anti-money laundering, credit limit adjustments, and error analysis. For chatbots, their primary use cases are in customer and employee interactions, where they are extensively deployed to enhance responsiveness and streamline communication processes. These applications demonstrate the growing role of AI in optimizing operational efficiency and decision-making in the financial sector.[1]

In conclusion, a literature review was conducted to identify key index for securing business capabilities in the financial sector (Table 1). Existing studies primarily focus on index research for evaluating artificial intelligence, frameworks for its application, and policy research addressing overall governance or cases of AI implementation in the financial sector.

However, the review highlights a significant gap: there is a lack of detailed definitions for specific capability indices related to AI-driven business initiatives in the financial sector, as well as concrete studies that prioritize these indices or provide case studies based on them. This underscores the necessity for further research to effectively establish and apply such indices.

3. Research Method and AHP Analysis

This chapter outlines the process of constructing the research model and methodology to identify the priorities of AI business capabilities in the financial

(Table 1) Global Index Definition of Key Items

Lv1	Lv2	Definitions	Ref.
Implementaion	Operational Workforce	An operational team to manage AI technologies, supported by skilled personnel, along with the provision of relevant training programs and the availability of qualified staff.	[2]
	Infrastructure	Possession of AI-related infrastructureand large-scale data storage systems, status of internal/external data utilization, and the establishment of AI-related information systems(e.g., GPU, foundational models)	[3]
	Environment	Plans for AI-related task transitions, short-term and long-term workforce development programs to facilitate role transitions, establishment of dedicated KPIs for AI-based task transitions, and possession of procedures for task transitions.	[14]
Innovation	Research	Possession of a dedicated research lab for future technology exploration, supported by specialized research personnel, including dedicated executives or team members.	[19]
	Development	Status of adoption of new AI technologies and related activities, including the introduction or commercialization readiness of technologies such as Foundational Models (FM), Prompt Engineering, Large Language Models (LLM), and Retrieval-Augmented Generation (RAG).	[17]
Investment	Strategy	Establishing strategies based on data, platforms, and customer experience, which are the core components of the business model, through AI strategies or planning.	[21]
	Commercialization	Determining the scale of investment in AI and establishing investment criteria to enable new business opportunities.	[1]

sector.

3.1 Research Method

3.1.1 Construction of the Research Model

To evaluate the importance and priorities of assessment index for investments in AI business initiatives in the financial sector, an AHP analysis was conducted using an Excel VBA-based program.

AHP (Analytic Hierarchy Process) is a multi-criteria decision-making (MCDM) methodology developed by T. Saaty in the early 1970s. It is widely recognized for its ability to handle qualitative MCDM problems. AHP represents decision-making problems hierarchically, clustering evaluation elements into homogeneous groups and organizing them into multiple levels. Based on the judgments of decision-makers, AHP assigns priorities to alternatives and synthesizes the results at each level to reach a final decision. Given its systematic structure, AHP was deemed suitable for determining the priority of assessment index in this analysis.

AHP operates on a set of principles that define the scope of decision-making problems and analyze the factors that influence them. The methodology is grounded in the mathematical structure of eigenvectors and matrices, which are used to calculate weights for each factor. Through pairwise comparisons, AHP evaluates the relative importance of two factors or alternatives, calculating their priority scores and providing a robust framework for decision-making.[22] AHP can be applied regardless of whether the data used to evaluate alternatives is quantitative or qualitative. It is particularly valuable as it quantitatively analyzes the qualitative opinions of expert groups involved in the evaluation. This capability makes AHP highly versatile and widely recognized for its usefulness in various decision-making scenarios.

This study aims to identify the key investment factors that carry the highest weights and priorities influencing the decision to adopt artificial intelligence (AI).

The research model was developed based on the Global AI Index, dividing it into upper-level and lower-level index within a hierarchical structure. The definitions of these index were established through an extensive review of relevant literature.

3.1.2 Sample Selection and Data Collection Method

For this study, a survey was conducted with 62 experts in the financial industry who are either currently involved in AI-related investments or actively engaged in specific tasks related to AI projects. The occupational characteristics of the participants are as follows:

Banking and Finance (including Fintech): 30 participants (29.6%). Capital/Loans: 4 participants (3.8%); Securities/Credit Cards: 6 participants (6.4%); Insurance: 22 participants (21.9%). In terms of general characteristics, the gender distribution is as follows: Male: 44 participants (71.0%); Female: 18 participants (29.0%).

Age Distribution: 20s: 3 participants (4.8%), 30s: 19 participants (30.6%), 40s: 24 participants (38.7%), 50s: 11 participants (17.7%), 60s: 5 participants (8.1%). This demographic data offers a comprehensive overview of the participants, reflecting a diverse range of roles and sectors within the financial industry.

The survey was conducted online from June 10 to June 19. 2024, involving a total of 62 participants. Respondents provided their opinions through pairwise comparisons of major categories (Lv1) and subcategories (Lv2). The survey included explanatory materials about the pairwise comparison method and descriptions of the evaluation criteria to ensure clarity and consistency.

The 9-point scale proposed by the AHP methodology was used for the pairwise comparisons of the evaluation criteria, enabling participants to assess the relative importance of each item effectively.[22]

When the number of comparison targets exceeds three, the Consistency Ratio (CR) is calculated to verify the logical consistency of the survey results. This is done by dividing the Consistency Index (CI) by the Random Index (RI).

In AHP, the pairwise comparison matrix A is a square matrix of size $\eta \times \eta$ where each element a_{ij} indicates the relative importance of i compared to j. The maximum eigenvalue(λ max) is derived from the matrix using the characteristic equation $\det(A-\lambda I)=0$, and the largest eigenvalue is used to calculate the CI using the following formula. Subsequently, the CR is calculated by dividing the CI by the RI (the predefined random index corresponding to the size of the matrix). According to standard practice, a CR value of 0.1 or less is considered acceptable for reasonable evaluations. However, prior studies suggest that CR values below 0.2 may also be acceptable. For this study, data with CR values exceeding 0.2 were excluded to maintain logical consistency. As a result, the analysis focused on the responses from 19 participants with CR values below 0.2, ensuring the reliability and consistency of the evaluation results.

$$CI = \frac{\lambda_{\max} - n}{n-1} \quad , \quad CR = \frac{CI}{RI} \qquad (1)$$

The Consistency Ratio (CR) for the higher-level factors was calculated as 0.154, which is below the threshold of 20%. This indicates that the logical consistency of the respondents' evaluations was sufficiently ensured.[23]

3.2 AHP Analysis

3.2.1 Importance and Priorities of Higher-Level Factors (Lv1)

The AHP analysis results indicate(Table 2) that among the higher-level factors (Lv1) for evaluating AI investments in the financial sector, Investment was identified as the most critical factor with a weight of 0.451. This was followed by Innovation with a weight of 0.300, and Implementation with a weight of 0.249. The findings highlight that the financial sector prioritizes strategic planning and investment direction when adopting AI technologies. This indicates that investment decisions are made based on well-defined strategies and investment plans.

The second most important factor, Innovation, reflects the emphasis on whether the organization possesses

(Table 2) AHP Comprehensive Analysis

Lv1	Weights	Lv2	Weights	Total Weights	Rankings
Implementaion	0.249	Operational Workforce	0.382	0.095	5
		Infrastructure	0.253	0.063	7
		Environment	0.366	0.091	6
Innovation	0.300	Research	0.499	0.135	4
		Development	0.551	0.165	3
Investment	0.451	Strategy	0.461	0.208	2
		Commercializatio n	0.539	0.243	1

specialized research labs or facilities for AI research and development. The results suggest that once the investment direction is determined, efforts are directed toward advancing future research and service development.

Finally, Implementation was ranked as the lowest priority among the factors. This category evaluates whether organizations are assembling teams with skilled personnel in artificial intelligence (AI) and investing in the necessary infrastructure. The relatively low priority assigned to this factor suggests that, at present, financial institutions are more focused on investment planning and organizational development rather than on the immediate recruitment of skilled personnel for Implementation.

These results demonstrate that the financial sector's approach to AI adoption emphasizes investment and innovation as foundational elements, with Implementation regarded as a subsequent step in the process. This reflects a focus on long-term planning and development prior to
operational deployment.

3.2.2 Importance and Priorities of Implementation Factors

The Consistency Ratio (CR) for Implementation factors was calculated as 0.196, which is below the threshold of 20%. This indicates that the logical consistency of respondents' evaluations was adequately ensured. Among the Implementation factors, operational workforce emerged as the most critical factor with a weight of 0.382, followed by infrastructure with 0.366, and environment with 0.253.

The highest priority among Implementation factors is assigned to the operational workforce. This underscores the importance of organizing skilled teams to support AI technologies, providing training to develop relevant expertise, and ensuring the availability of specialists. Various initiatives are actively underway to cultivate and retain professionals with the necessary skill sets to successfully implement AI initiatives. Infrastructure was identified as the second most critical factor. This includes the presence of systems such as GPUs, foundational models (FM), and data storage facilities required to support AI-related services. The availability of robust infrastructure is deemed a crucial prerequisite for successful AI implementation. The environment which pertains to an organization’s ability to facilitate job transitions and implement relevant procedures for AI-related workflows, ranked the lowest in priority. This index that preparations for transitioning tasks and processes to align with AI technologies are still in the early stages within most financial institutions.

The analysis reveals that the financial sector places the highest emphasis on building and maintaining a skilled operational workforce and establishing a strong infrastructure to support AI initiatives. However, creating an organizational environment that enables seamless job transitions and process adaptation for AI-related tasks is still at an early stage. These findings suggest a need for further focus on developing procedural frameworks and fostering adaptability within organizations to fully realize the potential of AI in the financial sector.

3.2.3 Importance and Priorities of Innovation

Factors

An analysis of the innovation factors in AI investment for the financial sector revealed that development is a more critical factor, with an importance weight of 0.551, compared to research, which has a weight of 0.449.

In the context of innovation, the factors were categorized into research and development, focusing on advancing AI technologies and preparing for future tasks.

Research involves securing specialized research facilities and skilled personnel to explore and develop new AI technologies. However, research was ranked lower compared to development, indicating that its priority is considered less immediate.

Development encompasses activities aimed at enhancing ongoing services and adopting new technologies. It emphasizes the introduction of new services and the establishment of robust development environments. Development was identified as more important than research, reflecting its immediate relevance to the current implementation and application of AI technologies.

The findings emphasize that within the realm of innovation, development activities are more significant than exploratory research. This highlights the financial sector's emphasis on implementing practical advancements in AI capabilities to address current business needs while research is viewed as a longer-term priority.

3.2.4 Importance and Priorities of Investment Factors

The analysis of investment factors in AI business initiatives within the financial sector revealed that commercialization is the most important factor with a weight of 0.539, compared to strategy, which has a weight of 0.461. Commercialization has been identified as the most critical investment factor, underscoring the priority placed on establishing a solid investment framework and developing comprehensive plans for the commercialization of AI technologies in the financial sector. The significant emphasis on commercialization underscores its crucial role in ensuring that AI technologies are effectively integrated into business operations and yield measurable results. Strategy ranked as the second most important factor, highlighting the need for well-defined strategic plans to support and guide the implementation of AI business initiatives. Strategic planning is essential for aligning AI investments with organizational goals and ensuring their effective Implementation. The results indicate that commercialization efforts, supported by precise investment planning, are the top priority in AI business initiatives for the financial sector. Strategic planning, while critical, follows as a secondary priority, serving as a foundational framework for achieving successful commercialization. These findings indicate the financial sector's focus on transforming AI technologies into tangible business value through structured investment and strategy development.

3.2.5 Comprehensive Importance and Priorities of Evaluation index for AI Business Capabilities in the Financial Sector

The analysis of the comprehensive importance and priorities of capability index for AI business in the financial sector revealed that, among investment index, commercialization was evaluated as the most critical factor, with a weight of 0.243. This was followed by strategy at 0.208, developmentat 0.165, research at 0.135, operational workforce at 0.095, infrastructure at 0.091, and environment at 0.063.

Overall, the findings indicate a focus on the investment domain, with significant attention to evaluating strategies and formulating detailed Implementation plans. These efforts can be interpreted as part of an organizational improvement process aimed at supporting research and development as well as business operations.

This trend indicates that AI-based businesses in the financial sector are still in their early stages, with gradual progress anticipated toward more robust research, development, and operational capabilities in the future. This progression reflects a maturing landscape in which foundational investments and strategies will ultimately evolve into long-term operational and technological advancements.

3.3 Discussion

Current executives in the financial sector emphasize the necessity for increased investments in hardware and software to advance artificial intelligence (AI) initiatives. According to the NVIDIA Financial AI Analysis Report, only 34% of respondents agreed with the statement, "Our company invests adequately in AI." In contrast, 28% were neutral or uncertain, while 39% believed that their company's investment in AI was inadequate. Nevertheless, 97% of respondents indicated that they plan to increase their spending on AI infrastructure this year.[30]

In a similar survey, 56% of customers identified AI investment as a primary focus area for the financial industry. [25] The findings support the results of this study, where investment ranked as the highest priority among AI capability index in the financial sector. Strengthening investment capabilities highlights the importance of aligning investment priorities with market trends and maximizing investment efficiency.

As significant investments are anticipated, it is essential to establish guidelines for business investments and prepare decision-making criteriaduring the review process. This preparation will ensure that investments are well-targeted and strategically aligned to achieve maximum impact.

Global top 50 banks are increasingly investing in and acquiring AI and technology-focused companies through venture and partner investments. To accelerate digital transformation initiatives, they are also expanding their partnerships with universities/academic institutionsand suppliers.

For example, Morgan Stanleyhas commercialized the AI @ Morgan Stanley Assistantservice by integrating OpenAI, while 29 bankshave formed partnerships with universities to accelerate research activities. Additionally, 12 banksare operating dedicated AI code repositories on GitHubto foster innovation and collaboration.[27]

4. Conclusion

This study aimed to identify priority areas for enhancing AI business capabilities within the financial sector. This need arises from evolving market conditions that necessitate responsibility and accountability in AI adoption, driven by trends such as the increasing utilization of AI, modificationsin existing workflows, rising investments and the introduction of financial AI guidelines. Acknowledging these dynamics, the study sought to establish the priorities essential for developing AI business capabilities.

The main index were categorized into three primary factors: Implementaion, Innovation, and Investment, with detailed subcategories defined through a literature review. Implementaion includes talent, environment, and infrastructure; Innovation encompasses research and development. Investment focuses on strategy and commercialization.

The priority analysis revealed the following order of importance: Commercialization > Strategy > Development > Research > Operational Workforce > Infrastructure > Environment. Among these, commercialization was identified as the most critical capability.

The study offers a framework for organizations in the financial sector to evaluate their existing capabilities and identify the necessary steps to develop and enhance their AI business capabilities. Even organizations lacking specific plans or capabilities for AI commercialization can utilize the capability index to assess their current status and determine the appropriate sequence of actions for improvement.

The significance of this research lies in its ability to facilitate discussions on concrete preparations for the future through self-assessment. By focusing on the prioritized key capabilities, fostering necessary skills, and pursuing gradual development, it is expected to accelerate business outcomes and maximize the effectiveness of investments.

According to guidelines in the financial sector, individual financial institutions are required to develop their own capabilities for utilizing artificial intelligence (AI) with a strong emphasis on the importance of formulating effective strategies. Reports indicate that companies prioritizing enterprise-wide AI strategies have achieved over a 20% improvement in Earnings Before Interest, Taxes, and Amortization (EBITA) compared to those that have not. This finding underscores that organizations that identify and nurture the core capability index necessary for strategy formulation are better positioned to gain a competitive advantage in an AI-driven business landscape.

Lastly, the limitations of this study are as follows: this study relied on the Global AI Index as foundation for designing survey questions and conducting research on capability index for artificial intelligencein the financial sector. While this approach provides valuable insights, it underscores the necessity for future research to represent the domestic financial sector and develop more localized index tailored to the unique characteristics of the Korean financial industry. Additionally, there is a need for domain-specific analysis to explore granular index based on specific areas within the financial sector. As expertise in the field continues to advance, it will also be important to conduct research on index that focus on specialized groups, such as prompt engineers and model trainers. This suggests a broader scope for future research on financial AI capability index.

Reference

[1] B. J. Kim, S. B. Yun, M. O. Kim, S. H. Chun," A Case Study on the Introduction and Use of Artificial Intelligence in the Financial Sector", Industrial Promotion Institute,Vol.8,No.2,pp.21-27,2023

[2] J. H. Seo, " The Utilization of Artificial Intelligence in the Financial Industry and Policy Challenges", Korea Development Institute, 2022.

[3] S. J. Kim, "A Review and Implications of Cases of Introducing Artificial Intelligence into Payment and Settlement Systems of Financial Institution", Journal of Payment and Settlement, Vol. 14, No.1, 2022, pp.89-116, 2022.

[4] "http://www.technologyreview.com/2019/11/11/131983“

[5] S. Agarwal, S. Alok, P. , Ghosh, S. Gupta,“Financial Inclusion and Alternate Credit Scoring for the Millennials: Role of Big Data and Machine Learning in Fintech,”, Indian school of business, working paper,2020.

[6] K. H. Lee, " The Impact of IT Project Size and Types on IT Investment Decision Criteria", Journal of Information Technology Application and Management, Vol.12, No.1, pp.191-211, 2005.

[7] N. Wunderlich, R. Beck, “Artificial intelligence for the financial services industry: What challenges organizations to succeed,”, IEEE Computer Society, 2019.

[8] “https://www.tortoisemedia.com”

[9] "https://hai.stanford.edu/tools"

[10] "https://aws.amazon.com/cloud-adoption-framework"

[11] Kaushal, S., Mishra, D. , “Strategic Implications of AI in Contemporary Business and Society,”,IEEE, 2nd International Conference on, pp.1469-1474,2024.

[12] Attkan, A.,“Cyber-physical security for IoT networks: a comprehensive review on traditional, blockchain and artificial intelligence based key-security,”, Complex & Intelligent Systems,Vol.8,No.4,pp.3559-3591,2022.

[13] “https://fintechnews.sg/17150/fintech/

[14] Truby, J., Brown, R., Dahdal, A., ”Banking on AI: mandating a proactive approach to AI regulation in the financial sector,“, Law and Financial Markets Review, Vol.14, No.2, pp.110-120, 2020.

[15] Alsheiabni, S., Cheung, Y., Messom, C., “Factors inhibiting the adoption of artificial intelligence at organizational-level: A preliminary investigation,”, Faculty of Information Technology, 25th Americas Conference on Information Systems,2019

[16] "https://www.mckinsey.com/industries/"

[17] Bedu, P., Fritzsche, A.,“Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption,”, Journal of Enterprise Information

Management, Vol.35, No,2, pp.530-549,2022.
[18] "https://www.law.go.kr"
[19] Alsheiabni, S., Cheung, Y., Messom, C., “Factors inhibiting the adoption of artificial intelligence at. organizational-level: A preliminary. investigation,”, Faculty of Information Technology, 25th Americas. Conference on Information Systems,2019
[20] P. Weill, S. L. Woerner, ”What's Your Digital Business Model?: Six Questions to Help You Build the Next-Generation Enterprise,“, MIT center for information system research,2018.
[21] J. K. Bea,"A Study on the Framework Application Method in Business Artificial Intelligence", Logos Management Review, Vol.20, No.1, pp.87-102, 2022.
[22] T. L. Saaty(1980), “The analytic Hierarchy process (AHP) for decision making,”, McGraw-Hill International Book Co..
[23] K. H. Kim, S. Y. Baek, D. H. Kim, H. Y. Kim,M. Park, S. W. Kim, " Investigation on the Prioritization of Factors in Assessment Criteria for Smart Housing Services Using the Analytic Hierarchy Process(AHP) Method", Korea Institute of Ecological Architecture and Environment, Vol.23, No.5, pp.33-42, 2023.
[24] D. S. Hong, “Financial AI Market Outlook and Current Applications”, korea Credit Information Services, No.1, 2022.
[25] KPMG, “Finance Embraced by AI: The Booster of Innovation”, 2024
[26] “https://www.eiopa.europa.eu/browse”
[27] Evident Insigth,“Evident AI Index”, 2023
[28] “https://www.fsc.go.kr/”
[29] “https://ai-governance.eu/about-aiga/”
[30] “https://resources.nvidia.com/en-us-2024-fsi-survey/ai-financial-services”

◐ ABOUT THE AUTHOR ◑

Suho Ahn

Mr. Ahn is a Ph.D. candidate in Applied Economics at the Graduate School of Hanyang University. From 2004 to 2013, he worked at Korea Telecom, followed by roles at Samsung Electronics from 2013 to 2020, where he contributed to various digital transformation initiatives in the Wireless Business Division and the Business Innovation Team. Subsequently, he served as a digital transformation consultant at IBM. Currently, he is engaged in consulting for digital transformation at Amazon AWS.

His research interest includes financial digital transformation, digital marketing, eCommerce, AI/Big Data, and Cloud technologies.

E-mail : suhoahn@amazon.com

Tae Yeon Oh

Prof. Oh earned a Ph.D. in Sports Management from Seoul National University. From 2018 to 2022, he served as a professor at the University of Mississippi. Currently, he is a professor in the AI/Big Data program at the Seoul School of Integrated Sciences & Technologies (aSSIST). His research interests include sports marketing, sports media, and big data analytics

Yongkyu Kim

Prof. Kim earned a Ph.D. in Economics from Columbia University in 1992. From 1992 to 2001, he was a researcher at the Korea Information Society Development Institute (KISDI). He is currently a professor in the Department of Economics at Hanyang University ERICA Campus. His research focuses on regulatory policies in the information and communications industry.

Reducing Training Time and Enhancing Efficiency in Large Language Models: Ternary Output Quantization(TOQ)

Do-hoon Lee

ABSTRACT

This study examines the background of recent advancements in artificial intelligence technology and the need for large-scale language models (LLM) and multimodal models (LMM), proposing the Ternary Output Quantization (TOQ) technique to reduce computational complexity and shorten training times for these models. TOQ simplifies the final outputs of models to three discrete values (-1, 0, +1), significantly reducing data processing requirements and enhancing the efficiency of model training and inference processes. Experimental results show that applying TOQ to BERT and DistilBERT models significantly reduces training times and improves both training and validation loss compared to traditional models. This paper presents the potential for wide application of TOQ across various forms of AI models, anticipating that it will facilitate easier training and deployment of models, especially in resource-constrained environments.

☞ keyword : AI, Ternary Output Quantization, LLM, Learning Efficiency, Data Processing Minimization

1. Introduction

The field of artificial intelligence has seen rapid advancements since the introduction of the Transformer model (Vaswani, et al., 2017). In particular, large language models (LLMs) and large multimodal models (LMMs), built on top of Transformers, have emerged, solving many previously intractable problems. These models, however, require massive computational resources to achieve their impressive performance (Vaswani et al., 2017; Brown et al., 2020; Kaplan et al., 2020; Narayanan, et al., 2021).

Large-scale models typically involve billions of paramet ers for training and inference, necessitating high-performan ce hardware, which implies high associated costs. These ch allenges become especially critical in resource-constrained environments. To address these issues, techniques such as Low-Rank Adaptation (LoRA) and Retrieval-Augmented G eneration (RAG) have been proposed. These methods aim to reduce the training burden by modifying only parts of the model or integrating external knowledge sources (Hu, et al., 2021; Lewis, et al., 2020). However, these approach es still struggle with the time and resource demands associ ated with large-scale data processing. In cases where suffic ient data and computational resources are available, fully tr aining a large model from scratch often achieves better per formance for specific tasks (Narayanan, et al., 2021).

This study hypothesizes that reducing the computational complexity by converting the complex and diverse continuous output variables of the model into three discrete values (-1, 0, +1) can improve model performance while maintaining similar accuracy across epochs. To this end, we propose a novel technique called Ternary Output Quantization (TOQ), which aims to reduce computational overhead by quantizing the model's output. TOQ simplifies the output to three distinct values (-1, 0, +1), thereby improving processing speed while maintaining or enhancing model performance. The introduction of TOQ is particularly beneficial in terms of energy efficiency and processing speed, making AI models more feasible for use in environments with limited resources.

The objective of this research is to reduce the computational cost involved in training and inference of large language models through the application of TOQ. Specifically, we aim to experimentally validate how TOQ enhances model training efficiency and inference speed, and evaluate its impact on generalization and overall performance. We also explore the applicability of TOQ by integrating it with models like BERT and DistilBERT to assess its utility in real-world natural language processing (NLP) tasks. The findings are expected to contribute significantly to improving model performance and reducing training time in NLP models, particularly facilitating the training and deployment of large-scale models in resource-limited environments.

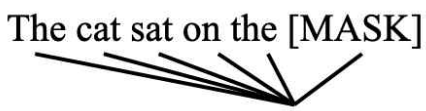
2. Theoretical Background and Related Work

2.1 Theoretical Background

2.1.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a powerful language representation model widely applied in natural language processing. It employs two main pre-training techniques, namely Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), to learn deep bidirectional representations.

Masked Language Model (MLM) is one of BERT's core pre-training mechanisms, where parts of the text are randomly masked, and the model is trained to predict these masked tokens based on the context. During this process, the model learns the ability to infer masked words from their surrounding context. For instance, in the sentence “The cat sat on the [MASK],” the model must predict the masked word using clues from “cat,” “sat,” “on,” and “the.” This mechanism helps BERT understand the bidirectional context of text. By predicting masked tokens, the model gains a deep understanding of the context, thereby learning richer language representations.



(Image 1) MLM

Next Sentence Prediction (NSP) is another important pre-training mechanism in BERT, focusing on determining whether two given sentences are sequentially related. For example, given the sentences “I drank coffee. [SEP] It was very strong,” BERT must decide whether these sentences logically follow one another. In contrast, given the sentences “I drank coffee. [SEP] She traveled to America,” BERT should recognize that the sentences are unrelated. Through NSP, BERT learns logical continuity and semantic connections between sentences, enhancing its understanding of overall document structure.

$$Attention(Q,\ K,\ V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

According to Equation (1), the attention mechanism calculates the relationships between input tokens using Query (Q), Key (K), and Value (V). This helps determine which tokens in the sentence are most related to each other (where $d_k$represents the dimension of the key vector). Q represents the question, K represents the target of that question, and Q matched with K returns the value V. Essentially, the degree of similarity between Q and K is represented as a probability, which is then multiplied by V to generate the final attention score. This information is used to generate new token representations.

BERT takes tokenized text as input and converts each token into embedding vectors. These vectors are combined with positional encodings to include information about the order of tokens in the sentence. Then, the input passes through multiple layers of Transformer encoders, which use multi-head self-attention to learn how each token relates to others in the sequence.

The training of BERT involves pre-training on large amounts of unlabelled text, followed by fine-tuning for specific NLP tasks. During pre-training, MLM is used to mask random tokens and predict them. At the same time, NSP is used to predict if two given sentences are sequential. This bidirectional pre-training allows BERT to understand the context from both directions, providing richer understanding compared to unidirectional language models.

2.1.2 DistilBERT

DistilBERT is a lightweight version of BERT that reduces model size while retaining most of BERT's language understanding capabilities and improving inference speed. This model leverages a technique known as Knowledge Distillation to effectively learn during the pre-training stage, using significantly fewer computational resources compared to the original BERT model, while still delivering excellent performance.

$$L_{ce} = \Sigma_i t_i * \log(s_i) \qquad (2)$$

Knowledge Distillation is a technique that transfers the knowledge of a larger model (teacher model) to a smaller model (student model). In Equation (2), t_i represents the probability from the teacher model, and s_i represents the probability from the student model. The loss function helps the student model to accurately replicate the output

distribution of the teacher model. During this process, the student model learns to mimic the soft probability distribution produced by the teacher model. The prediction probability distribution from the teacher model can be adjusted using a parameter called Softmax temperature. The Softmax temperature is used to smooth out the model's output, where a higher value produces a smoother distribution (with probabilities more evenly spread), and a lower value makes it sharper (with a higher probability for specific classes).

DistilBERT retains the core BERT architecture but reduces its size by removing token type embeddings and the pooler and by halving the number of layers, thus reducing the model size by 40%. This reduced number of layers greatly enhances the model's computational efficiency. Additionally, DistilBERT is initialized using intermediate layers from the teacher model, enabling it to effectively leverage the context learned by the teacher model.

DistilBERT's training employs a combination of three loss functions: Masked Language Modeling Loss (L_{MLM}), Distillation Loss (L_{CE}), and Cosine Embedding Loss (L_{COS}). These three loss functions help the model to effectively imitate not only the linguistic characteristics of the text but also the teacher model's performance.

Masked Language Modeling Loss (L_{MLM}) is a technique originally used in BERT pre-training, where parts of sentences are masked, and the model is trained to predict the masked words. For instance, in the sentence, "He ate [MASK] this morning," the model learns to correctly predict the masked word as 'apple,' 'banana,' or other plausible options.

Distillation Loss(L_{CE}) is a loss function that helps the student model to mimic the output distribution from the teacher model. The student model aims to predict each class's Softmax probability as closely as possible to the teacher model. For example, if the teacher model predicts a "positive" sentiment with a probability of 0.7, and the student model predicts the same sentiment with a probability of 0.5, the training process will minimize this difference.

Cosine Embedding Loss (L_{COS}) ensures that the embedding vectors of the student and teacher models point in similar directions. The goal is to keep the embedding spaces of the two models close to each other by minimizing the angle between the vectors. For instance, if the embedding vector of the teacher model points in a particular direction, the student's embedding vector is adjusted to point as closely as possible in that direction.

Through this combination of loss functions, DistilBERT can effectively reduce model size and computational cost without significantly sacrificing the teacher model's performance.

2.2 Previous Studies

2.2.1 TTQ

TTQ (Trained Ternary Quantization) reduces model size and enhances computational efficiency by representing neural network weights using only three values (-W, 0, +W). Each layer employs two scaling coefficients to quantize weights into ternary values, with these coefficients being optimized during the training process. For each weight w, the quantized weight w_t takes one of three possible values:

$$w_t \begin{cases} +W_p & \text{if } w > \Delta \\ 0 & \text{if } |w| \leq \Delta \\ -W_n & \text{if } w < -\Delta \end{cases} \quad (3)$$

In Equation (3), W_p and W_n denote scaling coefficients for positive and negative weights, respectively, while Δ represents the threshold. For example, if a neural network layer's weight vector is given as [-1.5, 0.2, 0, -0.3, 1.2] and the threshold Δ is set to 0.5, the quantized weights become [$-W_n$, 0, 0, 0, W_p]. If W_p = 1 and W_n = 1, the final weights are quantized to [-1, 0, 0, 0, 1]. During training, the model retains full precision weights w while updating W_p and W_n by computing gradients.

2.2.2 FGQ

FGQ (Fine-Grained Quantization) is a method that converts traditional models into ternary format. This technique transforms the model weights into ternary values while minimizing classification accuracy loss, achieving top performance on the ImageNet dataset without additional training. FGQ reduces computational complexity by finding an optimal scaling factor α using the Frobenius norm to minimize the error between the ternarized weights and the original weights.

$$\| A \|_F = \sqrt{\Sigma_{i,j} a_{i,j}^2} \quad (4)$$

$$\alpha^*, \beta^* = \text{argmin}_{\alpha \geq 0, \beta > 0} E(\alpha, \beta),\ E(\alpha, \beta) = \| W - \alpha \widehat{W} \|_F^2 \quad (5)$$

In Equation (4), $\| \cdot \|_F$ represents the Frobenius norm, calculated as the square root of the sum of the squares of a matrix's elements. This is equivalent to the Euclidean norm when matrix A's elements are flattened into a vector.

In Equation (5), α(scaling factor) is used during the ternarization of each weight in W. This factor scales the original weights to minimize the difference from the ternary weights $\widehat{W}$, where $\widehat{W}$ is the ternarized form of the original weights W, converting each element to -1, 0, or +1. The conversion criteria are determined primarily by a threshold β. For example, if the weight vector is W =[0.5, -1.5, 0.3, 2.1] and the threshold β=1, ternarization proceeds as follows: |0.5|<β → 0, |-1.5|>β → -1, |0.3|<β → 0, |2.1|>β → +1. Thus, $\widehat{W}$ = [0, -1, 0, 1]. Next, the scaling factor α is calculated as a parameter determined automatically during model optimization. Finally, the Frobenius norm is calculated and to simplify calculations and facilitate optimization (e.g., for easier differentiation), the squared Frobenius norm (Equation 6) is often used.

$$\| W - \alpha \widehat{W} \|_F^2 = \| (0.5, -1.5, 0.3, 2.1) - \alpha(0, -1, 0, 1) \|^2 \quad (6)$$

This value depends on α and methods such as differentiation are employed to find the optimal α. This process allows ternarized weights to closely resemble the original weights, minimizing overall performance degradation in the neural network.

2.3 Differences from Previous Studies

TTQ is a technique that reduces the model size and improves computational efficiency by representing neural network weights with only three values: -W, 0, and +W. It quantizes the weights of each layer into ternary values and optimizes them during training. However, TTQ limits the dynamic range of weights and lacks flexibility in specific weight values, which can lead to performance degradation. This restriction makes it challenging to apply TTQ to various neural network structures or language models.

FGQ is a method that minimizes performance loss by converting a trained model into ternary form. It optimizes a scaling factor using the Frobenius norm to minimize the error between original weights and ternarized weights. However, FGQ has limitations, as it requires a complex optimization process and may not be suitable for real-time applications. Additionally, FGQ processes the weights of the entire model in the same manner, which may fail to fully capture the unique characteristics of specific layers.

Unlike these conventional quantization methods, TOQ applies ternary quantization only to the output data, reducing computational complexity while enabling the model to focus on essential information. This approach simplifies the output data into three states(-1, 0, +1) to enhance computational efficiency and create a structure suitable for real-time processing. Consequently, TOQ addresses the issue of limited dynamic range in previous methods and can be flexibly applied to various neural network architectures. Furthermore, TOQ efficiently processes information in the final layer, reducing computational load and enhancing processing speed. By focusing on the output layer rather than the internal model layers, TOQ maximizes performance gains and resource savings in real-world applications. This approach suggests that TOQ can overcome the limitations of previous studies and make a significant contribution to improving the training and efficiency of large-scale language models, especially in resource-constrained environments.

3. Principle of TOQ Implementation

3.1 Basic Principle of TOQ

TOQ simplifies the continuous output values of the model into three discrete states (-1, 0, +1). The quantization is performed for each neural network output z_i based on the following conditions. Equation (7) presents an example of Ternary Quantization when the threshold is set to 0.5.

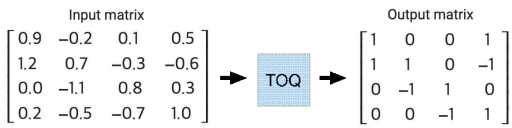
$$Q(z_i) = \begin{cases} -1 & \text{if } z_i < -0.5 \\ 0 & \text{if } -0.5 \leq z_i \leq 0.5 \\ +1 & \text{if } z_i > 0.5 \end{cases} \quad (7)$$

In this context, $Q(z_i)$ represents the quantized output value. This simplified output value guides the

model to focus more attention on certain parts during the weight update process. For instance, -1 and 1 indicate that the model should assign a larger or smaller weight to the corresponding output, prompting the model to concentrate on important information, discard unnecessary information, and diminish the impact of negative information. Unlike traditional continuous weight adjustments, this discrete quantization approach enhances computational efficiency, helping to maintain model performance, particularly in resource-constrained environments.

3.2 Weight Adjustment Mechanism

In a model where TOQ is applied, the quantized output values are used to adjust the weights during the training process. A value of 1 signifies that the feature is considered more important in the learning process, leading to positive reinforcement by increasing its weight. A value of 0 suggests that the feature does not significantly contribute to the weight update and can be neutral or disregarded. A value of -1 indicates that the feature may have a negative impact, prompting the model to reduce its influence. In the context of classification tasks, this ternary output affects how the model adjusts its internal weights, which in turn influences the way the model makes predictions. This mechanism allows the model to learn more efficiently and to focus on critical information, enhancing its performance and prioritizing essential features.



(Image 2) Ternary Output Quantization

4. Experiment and Results

4.1 Study

4.1.1 Research Environment and Data

This study was conducted on a MacBook Pro equipped with Apple's M1 Max chip (10-core CPU, 24-core GPU, 16-core Neural Engine, and 32GB unified memory). The software used was Jupyter Notebook from the Anaconda distribution.

The research data used was the 2023 newspaper corpus provided by the National Institute of the Korean Language. This dataset encompasses various contemporary Korean usage examples and reflects real-world language usage patterns. For training data, the seventh file in the "Newspaper Regional Comprehensive Media Article Raw Corpus" (NLRW2300000007.json) was used, and for validation data, the second file (NLRW2300000002.json) was utilized.

4.1.2 Text Preprocessing and Model

For preprocessing Korean data, we utilized the Okt object from the Konlpy library. Konlpy is a natural language processing library specifically designed for Korean, which is particularly useful for complex morphological analysis and part-of-speech tagging in Korean. During the preprocessing stage, special characters were removed from the Korean text, and only necessary nouns were extracted for model training.

BERT, developed by Google, is a model designed to process input data bidirectionally, enabling it to understand context more effectively. This model shows high performance, especially when handling large-scale language data, and is widely used in various natural language processing tasks. DistilBERT is a lightweight version of BERT that reduces the model size while improving training speed, maintaining performance close to the original model.

In this study, we developed a custom model based on the well-established BERT and DistilBERT models, incorporating the TOQ module. The TOQ module is applied to the output layer of the model, maximizing training efficiency by quantizing the output data.

4.1.3 Experimental Design and Evaluation Method

The main task used in this study is MLM. MLM randomly masks certain words in a sentence and evaluates the model's language comprehension by asking it to predict the missing words. MLM plays a critical role in enhancing the language model's prediction capability and assessing the appropriateness of words within a given context.

In this study, the effectiveness of TOQ was quantitatively evaluated by comparing models with and without the TOQ module (bert-base-uncased and

distilbert) on the MLM task. The evaluation metrics include Training Loss, Validation Loss, and Training Speed. Training Loss indicates how well the model is learning from the training data. By comparing the Training Loss between the TOQ-applied model and the baseline model, we can assess the effectiveness of TOQ. Validation Loss measures how well the model generalizes to the validation data, helping us analyze the impact of TOQ's quantization on the model's generalization capability. Lastly, Training Speed is assessed by measuring the time taken for model training, which shows the impact of TOQ on training efficiency. This metric is particularly significant in environments with limited computing resources.

4.2 Experimental Results

The experimental results of this study demonstrate that TOQ improves the learning and generalization capabilities of both BERT and DistilBERT models. For instance, the DistilBERT model with TOQ applied shows improved Training Loss and Validation Loss compared to the baseline model, while significantly reducing training time. This provides considerable advantages for training and deploying language models, particularly in environments with limited computational resources.

(Table 1) Results of BERT and DistilBERT Experiments

Model	Epochs	Batch size	Gradient accumulation steps	Learning rate	Training Loss	Validation Loss	Train runtime(sec)
BERT	0.5	16	4	3e-3(0.003)	3.5572	3.5414	11,865
BERT+TOQ	0.5	16	4	3e-3(0.003)	3.6289	3.7412	8,310
BERT	3	16	4	3e-3(0.003)	3.5359	3.5379	81,021
BERT+TOQ	3	16	4	3e-3(0.003)	3.3941	3.5932	56,156
DistilBERT	3	16	4	3e-3(0.003)	0.1046	0.1123	57,812
DistilBERT +TOQ	3	16	4	3e-3(0.003)	0.0998	0.1039	24,326
DistilBERT	3	8	4	3e-3(0.003)	0.1045	0.1123	52,599
DistilBERT +TOQ	3	8	4	3e-3(0.003)	0.0987	0.1039	42,588

In the case of 0.5 epochs, despite the short training duration, TOQ demonstrated some effectiveness in reducing training time. During 0.5 epochs, the average training loss of the BERT model was 3.5572, which increased to 3.6289 after applying TOQ, suggesting that the model experienced slight challenges during the early training phase. However, after completing 3 epochs of training, the training loss of the TOQ-applied BERT model improved to 3.3941 compared to the original model, and the validation loss was also comparable to that of the original model. This indicates that while TOQ may initially pose difficulties for the model during early training, it enhances the model's learning ability over the long term. Ultimately, TOQ proves its potential as an efficient language model training method by enhancing model performance and reducing training time. One of the primary advantages of TOQ is its ability to facilitate fast training and efficient deployment, especially in resource-constrained environments.

Through empirical validation, this study has shown that TOQ optimizes the balance between computational efficiency and model performance, enabling practical applications of large-scale language models. This finding is particularly significant in resource-limited environments, as TOQ not only improves model learning and generalization capabilities but also reduces model size and computational demands, greatly enhancing usability in real-world scenarios. In conclusion, this study demonstrates that the TOQ method is an effective way to reduce training time and improve model performance in the training and validation processes of language processing models. It is anticipated to make a valuable contribution to the development and application of large-scale language models in the future.

5. Conclusion and Future Research Directions

In this study, we applied the Ternary Output Quantization (TOQ) module to BERT and DistilBERT models, demonstrating its potential to reduce training time and enhance model performance. Experimental results indicated that TOQ can effectively reduce loss rates while improving training speed, particularly showcasing its potential to enhance model efficiency in environments with limited computational resources. These findings may be especially valuable in fields such as cloud computing and real-time language processing, where performance improvement is crucial.

However, there are two limitations to this study. First, the current TOQ module is primarily limited to specific types of language models, necessitating further evaluation of its applicability and effectiveness across diverse model types. Second, there is a lack of research on how TOQ impacts model interpretability, which poses challenges in fully understanding the model's decision-making processes. Future research should focus on validating the generalizability of the TOQ module by applying it to various types of neural network models. Studies will explore the effects of TOQ on generative models like GPT, multimodal models such as CLIP, and image processing models like CNNs. This will enable a more comprehensive evaluation of how TOQ can improve performance in a range of real-world applications.

In conclusion, this study demonstrates that the TOQ technique is an effective method for reducing training time and enhancing model performance in language processing models, with promising implications for the future development and application of large-scale language models.

Reference

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017), Attention is All You Need, *Advances in Neural Information Processing Systems, 30.*

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020), Language models are few-shot learners, *arXiv preprint arXiv:2005.14165.*

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D. (2020), Scaling laws for neural language models, *arXiv preprint arXiv:2001.08361.*

Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanty, V., Vainbrand, D., Kashinkunti, P., Bernauer, J., Catanzaro, B., Phanishayee, A. and Zaharia, M. (2021), Efficient large-scale language model training on GPU clusters, *arXiv preprint arXiv:2104.04473.*

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. (2021), LoRA: Low-Rank Adaptation of Large Language Models, *arXiv preprint arXiv:2106.09685.*

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D., (2020), Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, *arXiv preprint arXiv:2005.11401.*

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint arXiv:1810.04805*

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019), DistilBERT: A distilled version of BERT: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108*

Mellempudi, N., Kundu, A., Mudigere, D., Das, D., Kaul, B. and Dubey, P. (2017), “Ternary Neural Networks with Fine-Grained Quantization”, *arXiv preprint arXiv:1705.01462v3*

Zhu, C., Han, S., Mao, H. and Dally, W. J. (2016), “TRAINED TERNARY QUANTIZATION”, *arXiv preprint arXiv:1612.01064v3.*

Joshua K. Cage, 임선집, 채호창(2023), “101가지 문제로 배우는 딥러닝 허깅페이스 트랜스포머 with 파이토치,” 32-90.

◐ ABOUT THE AUTHOR ◑

Do-hoon Lee

Enrolled in the Master's in AI·Big Data program at the Graduate School of AI, Seoul School of Integrated Sciences & Technologies (aSSIST)

Currently working at the Youth Counseling and Welfare Center, Gyeongsangnam-do Youth Support Foundation

Areas of interest : AI, Data Analysis etc.

E-mail : ehgnsslekd@naver.com

A Study on the LLM-Based Software Architecture for Multi Robot Control

Sangho Choi

ABSTRACT

This paper examines the latest developments in the convergence of robotics and AI, with a particular focus on analyzing key cases of robot control research using large language models (LLMs). Recent AI technologies such as computer vision, neural networks, reinforcement learning, and LLMs have significantly improved robots' visual recognition, motion control, perception, and decision-making abilities. This study presents an LLM-based robot control architecture that can be applied to small, medium-sized, and home/educational robots without requiring large-scale robot data.

This architecture proposes a 'self-describing' method that automatically generates system prompts from robot control programs, allowing non-experts to easily control robots. Additionally, it offers various LLM model selection options to reduce API usage costs and presents a multi-agent architecture for future robot functionality expansion and control of multiple robots. The feasibility of the proposed method is verified through robots tests using a robot simulator. As robot popularization expands beyond manufacturing sites into everyday life, this research is expected to lower the barriers to applying AI technology to small and medium-sized robots.

☞ keyword : AI, Large Language Models, robot, control, sofrtware architectdure, few-shot learning, prompt

1. Introduction

Since the introduction of Unimate, the first industrial robot, at General Motors in 1961, robotics has driven innovative changes across various sectors, maximizing automation and efficiency. Recently, the convergence of artificial intelligence (AI) and robotics has further expanded its potential and effectiveness. Robots' capabilities have been significantly enhanced through the integration of multiple AI technologies: computer vision and neural networks have improved visual recognition capabilities, reinforcement learning has advanced motion control, and Large Language Models (LLMs) have enhanced cognitive and decision-making abilities.

AI technologies have brought transformative improvements to robots' three fundamental capabilities: perception, planning, and action. From a perceptual perspective, advancements in computer vision and sensor technologies have enabled more precise environmental recognition. In terms of planning capabilities, AI technologies, particularly reinforcement learning, have enabled efficient planning of complex tasks. The action control domain has seen substantial improvements in motion accuracy through sophisticated algorithms and real-time feedback systems.

The combination of LLMs exceptional capabilities and familiar chatbot interfaces has opened new frontiers in AI robotics research. LLMs, such as OpenAI's GPT-3, with their remarkable natural language processing capabilities, are being extensively studied in the domain of robot control. These advancements in LLMs have facilitated more intuitive and natural human-robot interaction, enabling simpler execution of complex task instructions and enhanced environmental understanding (Brown et al., 2020; Huang et al., 2022).

This paper examines a critical aspect of robot control: prompt generation methodology. Conventional approaches necessitated manual system prompt updates with each functional addition. To overcome this constraint, this study proposes an automated prompt generation methodology that leverages Python's decorator functionality to derive prompts directly from robot function definitions. Additionally, the research presents an architecture that accommodates multiple LLM engines and

introduces a framework for discriminative command execution across multiple entities (agents), including multi-robot systems.

The primary objective of this study is to enhance the efficiency and accuracy of LLM-based robot control while demonstrating its potential across various application domains. To achieve this, this paper proposes a methodology for validating and optimizing robot control outcomes using the PyBullet simulator. This approach offers the advantage of testing diverse scenarios while minimizing risks in real-world environments.

Furthermore, this study presents the design and implementation methodology for an efficient LLM-based robot control system. The organization of this paper comprises the following sections: theoretical foundations of robot control systems and LLMs; methodological framework for implementing LLM-based robot control systems; experimental evaluation and analysis of the proposed system's performance; and conclusions, including key findings and directions for future research.

This research contributes to the field by presenting a practical methodological framework for the design and implementation of LLM-based robot control systems, advancing robots' capabilities in perception, planning, and control. Furthermore, this study investigates the innovative potential applications emerging from the integration of robotics and AI, establishing groundwork for the development of more sophisticated and adaptable robotic systems.

2. Theoretical Background and Related Work

2.1 Overview of Robot Control Systems

Robot control systems serve as fundamental components for executing precise robotic operations through the acquisition and processing of sensor data to generate appropriate control commands. These systems are primarily classified into open-loop and closed-loop architectures.

Open-loop control systems generate outputs based on input signals without feedback mechanisms, providing architectural simplicity but limited environmental adaptability. Conversely, closed-loop control systems employ feedback loops for continuous operational adjustment to achieve desired states, exemplified by PID control, adaptive control, and predictive control methodologies.

The evolution of robot control technologies spans multiple technological epochs. While initial implementations predominantly employed open-loop control architectures, the advent of closed-loop control technologies in the 1960s-70s facilitated complex task execution. The 1980s witnessed the emergence of sophisticated control algorithms enabled by microprocessor advancement, while the 1990s marked the integration of artificial intelligence and machine learning, significantly enhancing autonomous capabilities and adaptive responses. Contemporary research emphasis has shifted toward deep learning for visual perception, reinforcement learning for autonomous control systems, and natural language processing for human-robot interaction interfaces (Javier et al., 2018; Liu et al., 2021). Additionally, research into collaborative control methodologies incorporating distributed control systems and multi-agent architectures has intensified, facilitating the expansion of robotic applications across diverse sectors, including industrial, medical, and service domains.

2.2 LLMs: Progress and Applications

LLMs constitute a breakthrough technology in Natural Language Processing (NLP), emerging from accelerated developments in artificial intelligence. LLM advancement has substantially enhanced capabilities in text comprehension and generation, demonstrating significant impact across diverse application domains.

The trajectory of artificial intelligence

commenced in the 1950s. Initial AI research originated from explorations into machine-based problem-solving through human-analogous cognitive processes. The term ‘artificial intelligence' was formally introduced at the 1956 Dartmouth Conference, leading to subsequent developments in expert systems and rule-based architectures (Rawas, 2024). The 1980s marked the advancement of neural network theory, establishing machine learning as the predominant paradigm in AI research.

The 1990s and early 2000s witnessed the ascendance of data mining and statistical learning methodologies, succeeded by the proliferation of deep learning in the 2010s. Deep learning advancement, leveraging extensive datasets and enhanced computational capabilities, achieved unprecedented performance in diverse domains, encompassing speech recognition, image classification, and natural language processing.

LLMs emerged as a response to fundamental limitations in natural language processing within this evolutionary trajectory of artificial intelligence. Initial NLP implementations relied predominantly on rule-based methodologies or statistical models trained on constrained datasets, exhibiting significant limitations in processing complex linguistic contextual relationships.

Natural Language Processing reached a pivotal turning point in June 2018 with OpenAI's introduction of GPT (Generative Pre-trained Transformer) (Radford et al., 2018). Subsequently, in October of the same year, Google introduced BERT (Bidirectional Encoder Representations from Transformers) (Delvin et al., 2019). BERT, based on the transformer architecture, demonstrated superior performance across various NLP tasks through its capability to comprehend context bidirectionally. The continuous evolution of OpenAI's GPT series further expanded the potential of LLMs. Notably, GPT-3, released in 2020, demonstrated human-comparable text generation capabilities through its 175 billion parameters.

LLMs exhibit several distinctive architectural characteristics. First, they utilize extensive pre-training on large-scale datasets, facilitating the acquisition of linguistic patterns and structural regularities through comprehensive textual corpus analysis. Second, they leverage parallel processing capabilities inherent to the transformer architecture, enabling efficient processing of extended contextual sequences and substantially accelerating learning and inference processes through parallel computational mechanisms. Third, they demonstrate zero-shot and few-shot learning proficiency, enabling task execution through general language understanding capabilities without task-specific fine-tuning procedures.

Contemporary advancements in LLMs are characterized by concurrent progression in model architecture scalability and performance metrics, coupled with application domain diversification. Multiple large-scale language models have emerged, including OpenAI's GPT-4, Google's PaLM (Pathways Language Model), and Meta's LLaMA, exhibiting superior performance across traditional NLP tasks, creative content generation, computational assistance, and complex problem resolution frameworks (Zhao et al., 2023).

Additionally, commercial implementations of LLMs are experiencing rapid proliferation. The integration of LLMs across diverse operational domains—encompassing conversational interfaces, machine translation systems, content generation platforms, intelligent virtual assistants, and automated customer service solutions—has yielded substantial improvements in operational efficiency and productivity metrics. Significantly, LLMs serve as fundamental enablers in the delivery of personalized user experience frameworks.

A significant new trend is emerging in recent LLM developments: the advancement of sLLMs (small Language Models) and On-Device AI. sLLMs represent technology

that maintains performance while reducing model size or optimizing for specific tasks, increasing their utility in mobile devices and edge computing environments. Concurrent developments in On-Device AI technology provide benefits including enhanced privacy protection and reduced network latency. These technologies are expanding the application scope of LLMs and are expected to play crucial roles in systems requiring real-time processing, such as robot control systems and mobile applications (Yi et al., 2023). The progression of sLLMs and On-Device AI facilitates more efficient and widespread implementation of LLM technologies.

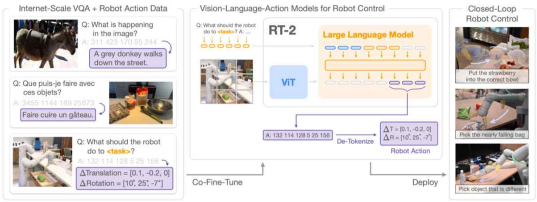
In conclusion, the advancement of Large Language Models continues to reveal unprecedented possibilities in artificial intelligence, with projected significant implications across NLP and diverse application domains. This research investigates the operational efficiency and practical utility of distributed multi-agent architectures through the implementation of LLM characteristics in robotic control systems. The integration of LLM's natural language processing capabilities with robotic systems is anticipated to facilitate the development of more sophisticated and adaptable architectural frameworks.

2.3 Multi-modal Approaches in Robot Control

The convergence of robotics with various domains of AI continues to accelerate, with multimodal approaches significantly enhancing robotic intelligence and behavior. Multimodal robot control enables more sophisticated and autonomous robotic operations through simultaneous processing of various data modalities, including vision, audio, and text.

2.3.1 Robot-Vision Integration

The evolution of computer vision technologies is fundamentally transforming robotic environmental perception and interaction paradigms. Google's recently introduced RT-2 (Vision-Language-Action Models) demonstrates novel methodologies in robotic control through the synthesis of vision and language models. RT-2 facilitates enhanced operational intuition and behavioral flexibility by integrating visual data streams with language model architectures, enabling sophisticated robot-environment interactions (Brohan et al., 2023).



(Figure 1) Robot Control using Vision-Language Model Capabilities

RT-2 demonstrates three fundamental architectural characteristics. First, the integration of vision and language modalities enables robotic systems to parse and comprehend visual data through natural language frameworks, facilitating enhanced interpretation of complex visual environments and precise execution of user-specified directives. Second, RT-2 implements a transformer-based Vision-Language Model (VLM) architecture for concurrent processing of visual and linguistic data streams, utilizing this integrated analysis for action determination. Third, the model exhibits zero-shot generalization capabilities through extensive training on heterogeneous datasets, enabling adaptive responses to novel operational contexts and environmental conditions not explicitly encountered during the training phase.

2.3.2 Robot-LLMs Integraion

LLMs enable enhanced human-robot interaction paradigms through advanced natural language processing capabilities. LLMs facilitate robot comprehension and execution of complex directives by leveraging linguistic understanding and generation mechanisms derived from

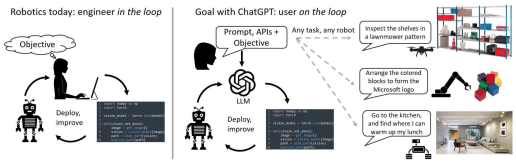
textual corpus analysis. This functionality, when integrated with visual perception systems, enables more sophisticated and adaptable robotic behavioral patterns.

Robotic control systems incorporating LLMs have emerged as a significant research focus. ProgPrompt represents an innovative implementation utilizing large language models for generating robotic task planning sequences (Singh et al., 2022). This research framework enables the derivation of robotic task plans from natural language specifications, facilitating user-directed complex task execution. While ProgPrompt demonstrates substantial improvements in task planning efficiency and adaptability, inherent linguistic ambiguities present limitations, potentially resulting in sub-optimal plan generation and execution precision.

SayCan study presented a methodology for robots to comprehend natural language commands and translate them into executable actions (Ahn et al., 2022). This research enhanced robots' ability to understand environmentally feasible actions and convert natural language commands into corresponding behaviors. While limitations remain in precise action determination within complex environments considering all possibilities, the study's significance lies in its advancement of robots' capabilities to accurately interpret and execute user commands.

Code as Policies study proposed a methodology for generating robot control programs using language models (Liang et al., 2023). This research demonstrated LLMs' capability to comprehend programming languages and generate robot control code, validating their potential to produce code solutions for complex robotic control problems. However, the study indicated limitations regarding the execution efficiency and operational stability of generated code in real-world environments.

Research has explored the potential applications of large language models like ChatGPT in robot control systems (Vemprala et al., 2023). This study established design principles focusing on natural language interaction, situational adaptability, and safety assurance, demonstrating LLMs' effectiveness in robot control through their capabilities in natural language comprehension, situational awareness, and adaptive learning. While the research validated LLMs' ability to accurately and naturally interpret user commands, limitations persist regarding comprehensive understanding and adaptation across all operational scenarios.



(Figure 2) Robot control with chatGPT

2.3.3 Multimodal Robot Control

Multimodal robotic control systems demonstrate significant advantages over unimodal approaches. First, they enhance environmental perception capabilities through the integration of heterogeneous data modalities. For instance, the synthesis of visual and linguistic data streams enables more precise command interpretation based on visual perception frameworks. Second, multimodal data integration provides diverse analytical perspectives in decision-making processes, facilitating more sophisticated and adaptable behavioral responses. Third, the system enhances operational utility through more naturalistic and intuitive human-robot interaction paradigms.

Contemporary research in multimodal robotic control systems is advancing toward optimization of autonomous capabilities and environmental adaptability through the synthesis of diverse artificial intelligence technologies. For instance, intensive investigation is being conducted on the integration of LLMs with reinforcement learning frameworks to facilitate complex task acquisition and execution.

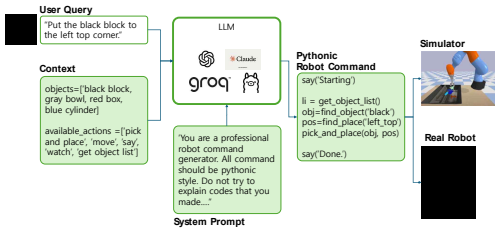
Furthermore, technological developments incorporating vision systems with LLMs are enabling real-time environmental analysis and appropriate action execution based on natural language command interpretation.

In conclusion, multimodal robotic control architectures are substantially advancing robotic cognitive capabilities and behavioral competencies, with LLM-based control systems playing an instrumental role in enabling autonomous operation within complex operational environments. These research initiatives are facilitating the expansion of robotic implementation domains and enhancing deployment feasibility across diverse industrial sectors. Future research trajectories are anticipated to focus on addressing the constraints of current multimodal methodologies and developing more robust and computationally efficient robotic control architectures.

3. Research Methodology

3.1 Design of LLM-based Control Module

For robots to comprehend operators' commands delivered in natural language, several LLM application techniques exist. This study implements few-shot learning prompting techniques as illustrated in Figure 3. Few-shot learning enables rapid learning and generalization from a limited number of examples, proving particularly valuable in robotic control where pre-training for all possible scenarios is infeasible. Combined with LLMs' robust generalization capabilities, few-shot learning enables more flexible and adaptive robot operation, which is particularly crucial when deploying robots in diverse and unpredictable real-world scenarios.



(Figure 3) Conceptual Framework of Robot Command Generation using LLM

Within the control module architecture, an LLM Selector component, engineered for LLM selection and management utilizing LangChain, implements chaining operations through LCEL (LangChain Expression Language). This architectural framework enhances system flexibility and extensibility through dynamic integration of decomposed components: prompt structures, LLM engine implementations, and output parsing mechanisms.

Furthermore, for automated prompt generation in robotic command synthesis, the system implements a methodology leveraging Python's decorator functionality to programmatically extract Agent class methods and dynamically integrate them into few-shot learning prompt structures. The '@include_action' decorator specification facilitates automatic integration of method annotations into prompt compositions, thereby enhancing architectural flexibility during functional expansion implementations.

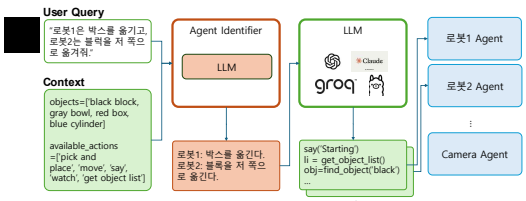
```python
# Custom Decrorator
# 이 데코레이터가 선언된 메서드만 프롬프트(Context)에 자동으로 포함된다.
# LLM은 Context에 정의된 메서드들만을 로봇명령어 생성에 사용할 수 있다.
def include_action(func):
    func._include_action = True
    return func
```

```python
    @include_action
    @debug_codeExecution
    def get_camera_image(self, showImage=True):
        """
            # Take some pictures
            say(f'OK - taking some pictures')
            get_camera_image(showImage=True)
        """
        # Compute view and projection matrices
        view_matrix = p.computeViewMatrix(
            self.cameraPostion, [0, 0, 0], [0, 0, 1])
        projection_matrix = p.computeProjectionMatrixFOV(
            self.fov, self.aspect, self.near, self.far)

        # Capture the image
        images = p.getCameraImage(
            self.width, self.height, view_matrix, projection_matrix,
            shadow=True, renderer=p.ER_BULLET_HARDWARE_OPENGL)
        # images[0] : width
        # images[1] : height
        # images[2] : RGB pixel data
        # images[3] : depth data
        # images[4] : segmentation mask
```

(Figure 4) Automated Prompt Generation using Decorator Implementation

The system implements a mechanism enabling LLM to identify individual agents in multi-robot scenarios. For example, when given a user command "Robot1, move the box, and Robot2, move the block over there.", the LLM engine redefines commands in an "{agent}:{action}" format, allowing each agent to execute its designated task. This implementation enables effective control in complex multi-robot environments.



(Figure 5) Identification of Multi Agent

3.2 Robot Simulator

Given the proximity of robotic operational environments to human spaces, simulation-based motion validation is essential prior to physical deployment. This study conducted comparative analysis of various robot simulators, ultimately selecting PyBullet for implementation. PyBullet offers advantages including straightforward installation, support for diverse robotic models, robust physics engine capabilities, open-source accessibility, real-time simulation functionality, and an active user community.

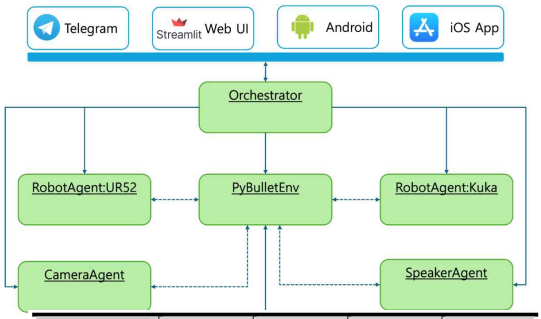
(Table 1) Types of Robot Simulators

Functions	PyBullet	Gazebo	MuJoCo	Webots
Language	Python	C++, Python	Python, C	C++, Python, Matlab, Java
Usage	High	Moderate	Moderate	Moderate
Accuracy	Moderate	High	High	High
Format	URDF, SDF, MJCF	URDF, SDF	MJCF	URDF, VRML
Real Time	Yes	Yes	Yes	Yes
Open Source	Yes	Yes	No	Yes
Performance	Moderate	High	Very High	High
Community	High	High	Moderate	Moderate

The PyBullet simulation environment served as an instrumental platform for performance validation and optimization of the proposed LLM-based robotic control architecture, facilitating the establishment of operational safety protocols and efficiency metrics prior to physical system implementation.

3.3 System Architecture

The system architecture proposed in this investigation is structured as a collaborative multi-agent framework, facilitating coordinated task execution among multiple robotic agents and supplementary agent entities. The fundamental architectural components encompass the Orchestrator for system coordination, PyBulletEnv for environmental simulation, and diverse agent implementations for task execution.



(Figure 6) Overall System Structure

The 'Orchestrator' module functions as the central control hub of the system. It manages communications between external interfaces and internal agents, processing user natural language commands for internal agent distribution and returning results to external interfaces. Additionally, it generates and manages the PyBullet-based main simulation environment (PyBulletEnv).

'PyBulletEnv' serves as a PyBullet-based simulation environment that simulates the operations of various robots and objects. It executes simulations based on commands received from the Orchestrator and interacts with the objectManager to generate and

manage objects within the simulation.

Various agents perform specific roles within the system. The RobotAgent controls robots such as UR5e and Kuka, utilizing the LLM engine to convert natural language commands into executable robot commands. The CameraAgent and SpeakerAgent are responsible for collecting visual information from the simulation environment and managing audio output, respectively.

Individual agents execute domain-specific functionalities while the Orchestrator facilitates comprehensive system integration. PyBulletEnv provides high-fidelity physical simulation capabilities, enabling pre-deployment validation and operational optimization of robotic movements.

The implemented architectural framework facilitates natural language interface for complex robotic control directives, with the system architecture enabling efficient command processing for safe and precise robotic operation execution. Additionally, the distributed architectural paradigm ensures enhanced system scalability and operational flexibility, establishing platform-agnostic compatibility across diverse robotic systems and environmental contexts.

4. Experimental Results

4.1 Comparative Analysis of LLM Code Generation Performance

While LLMs conventionally generate code accompanied by comprehensive explanations in response to natural language queries, robotic control applications require exclusively executable code output. Consequently, system prompts are implemented to explicitly define LLM operational parameters and constraints. To validate the efficacy of robotic command code generation, computational latency and code generation accuracy are evaluated across multiple LLM implementations using four standardized natural language test cases, as illustrated in Figure 7.

```
SYSTEM_PROMPT = """
You are a professional software programmer.
When given a prompt, write concise Python code that satisfies the user's request.
Follow these guidelines:
1. Provide the simplest possible solution.
2. Only include the code, without any explanations or comments.
3. Always wrap your code in triple backticks with 'python' specified.
4. Do not repeat the question or given prompt.
"""

QUESTION = [
"# Find the average of all numbers in a list called 'scores'.",
"# Count how many times the number 7 appears in a list named 'sequence'.",
"# Concatenate two strings, 'first_name' and 'last_name', with a space in between.",
"# Check if a given string 'word' is a palindrome (reads the same backwards as forwards).",]
```

(Figure 7) Questions for LLMs' Code geration Evaluation

Performance analysis indicates that OpenAI (model: gpt-3.5-turbo-0125) achieves optimal accuracy (100%) while maintaining comparatively efficient computational latency. Groq exhibits minimal execution latency but demonstrates suboptimal accuracy metrics. LLAMA (model: Codellama:7b) maintains moderate computational efficiency but exhibits reduced accuracy performance (50%), while Claude (model: claude-3-opus-20240229) demonstrates maximal accuracy (100%) but requires substantially increased processing duration.

(Table 2) LLM Execution Time(s)

LLM	Question (1)	Question (2)	Question (3)	Question (4)
OpenAI	0.59	0.87	0.91	1.41
Groq	0.39	0.49	0.35	0.41
LLAMA	4.75	1.24	1.92	1.39
Claude	2.80	1.93	5.98	3.61

(Table 3) LLM Accuracy Summary(%)

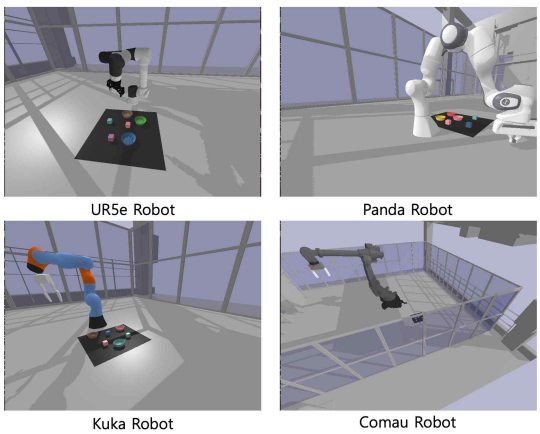
LLM	Question (1)	Question (2)	Question (3)	Question (4)
OpenAI	100	100	100	100
Groq	100	75	75	50
Codellama:7b	50	100	50	50
Claude	100	100	100	100

Although open-source models currently exhibit lower accuracy metrics compared to proprietary implementations, their accelerated development trajectory and enhancement rate necessitate periodic performance assessments. This research implements OpenAI's model as the primary LLM engine for robotic control applications, selected from among the highest-accuracy

performers (OpenAI and Claude), based on its superior computational efficiency metrics.

4.2 Robot control Experiments

To validate LLM control implementation across heterogeneous robotic platforms, a test environment was configured with four distinct robotic systems, incorporating both mechanical gripper and vacuum-based end effector configurations.



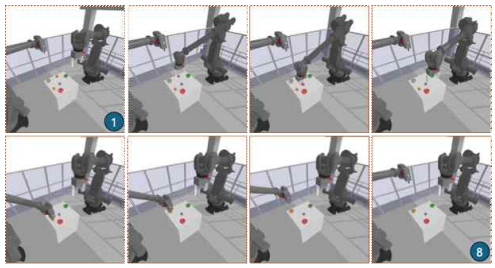
(Figure 8) Robots for simulation

The system uses Telegram as the command transmission protocol interface. User-initiated Telegram string inputs are processed through the Orchestrator, where natural language instructions undergo LLM-based analysis and transformation into pythonic robot command structures, culminating in physical task execution by the robotic system.

The experimental validation protocol was executed in two phases: initial evaluation of natural language command interpretation for fundamental operations on a single robotic platform, followed by sequential operation validation across multiple robotic systems. The implemented natural language test commands were as follows:

"I want robot1 to put the red block on the same-colored bowl, and robot2 will move the orange block on the gray bowl."

Operational verification demonstrated successful command differentiation and execution by both robotic platforms (robot1 and robot2), with movement sequences conforming to operator-intended trajectories and task specifications. (https://youtu.be/Gg9zhdg3zmg)



(Figure 9) Sequential Control of multi robots

5. Conclusion and Future Research Directions

This research aimed to design and implement a multi-agent architecture for LLM-based robot control, validating its performance through diverse experimental evaluations. Primary research achievements include establishing an environment for selecting and utilizing various LLM engines through LLM Selector design. System flexibility and scalability were significantly enhanced by implementing LangChain to separate and dynamically link prompts, LLM engines, and output parsers.

Furthermore, an automated prompt generation method utilizing Python decorators enabled automatic system prompt updates with each functionality addition, substantially improving user and developer convenience. Additionally, the implementation of a multi-agent identification mechanism enabled precise command recognition and execution by individual agents in multi-robot environments.

Finally, system practicality and stability were validated through experiments utilizing the PyBullet simulator. Experimental results

demonstrated high accuracy and efficiency, with superior code generation capabilities and robot control performance achieved through diverse LLM model implementations.

Primary limitations and future research directions of this study include insufficient validation across diverse environments, necessitating more rigorous verification of real-world applicability. Additionally, further research is required for precise control in complex command scenarios and exceptional situations due to inherent LLM model limitations.

System scalability presents significant challenges, requiring solutions for network latency, resource management, and synchronization issues that may arise during large-scale deployment. Limitations in automated prompt generation necessitate additional research for accurate prompt creation in cases involving complex functions and exception handling logic. Finally, improvements in user interface and interaction are required, with enhanced UI/UX needed to facilitate more intuitive and accessible system operation for end users.

Future research necessitates exploration of diverse methodologies to address these limitations and enhance system practicality. Key research challenges include implementation of sLLM and On-Device AI in real environments, ensuring operational safety regarding inter-robot motion interference, advancement of LLM models for enhanced environmental affordance and grounding, strengthening system scalability, improving automated prompt generation accuracy, and enhancing user interface design.

This research has presented practical methodologies for designing and implementing LLM-based robot control systems, contributing to the advancement of robotic perception, planning, and control capabilities. Future investigations are expected to enable the development of more sophisticated robotic control systems.

Reference

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020), Language models are few-shot learners, *Advances in Neural Information Processing Systems, 33, 1877-1901.*

Huang, W., Abbeel, P., Pathak, D., & Mordatch, I. (2022), Language models as zero-shot planners: extracting actionable knowledge for embodied agents. Proceedings of the 39th International Conference on Machine Learning, 9383-9407.

Javier Ruiz-del-Solar, et al. (2018), A Survey on Deep Learning Methods for Robot Vision, *arXiv.* *https://arxiv.org/abs/1803.10862.*

Liu, R., et al. (2021), Deep Reinforcement Learning for the Control of Robotic Manipulation: A Focused Mini-Review, *arXiv:2102.04148.*

Rawas, S. (2024), AI: the Future of Humanity. Discover Artificial Intelligence. 4(25), *Springer*

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018), Improving language understanding by generative pre-training. *OpenAI Technical Report.*

Devlin, J., et al. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805.*

Zhao, W. X. and Zhou, K., et al. (2023), A survey of Large Language Models. *arXiv:2303.18223*

Yi, L., Guo, L., Zhou, A., Wang, S., Xu, M. (2023), EdgeMoE: Fast on-device inference of MoE-based large language models, *arXiv:2308.14352.*

Signh,I., Blukis, V., et al. (2023), ProgPrompt: Generating Situated Robot Task Plans using Large Language Models, *In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), 11686-11693.*

Ahn, M., Brohan, A., Brown, N., et al. (2022), Do as I can, not as I say: grounding language in robotic affordances, *arXiv:2204.01691.*

Liang, Y., Cabi, S., et al. (2023), Code as Policies: Language Model Programs for Embodied Control, *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6045-6056.*

Vemprala, S., Bonatti, R., Bucker, A., Kapoor, A. (2023), ChatGPT for Robotics: Design Principles and Model Abilities, *arXiv:2306.17582.*

◐ ABOUT THE AUTHOR ◑

Sangho Choi
Enrolled in the Master's in AI·Big Data program at the Graduate School of AI, Seoul School of Integrated Sciences & Technologies (aSSIST)
Areas of interest: AI, robot, LLM, Computer Vision, etc.
E-mail : sangho.0529@gmail.com

Articles of Association for “AI Business Academy”

Chapter 1 General Provisions

Article 1 (Title) This corporation shall be called the “AI Business Academy” (AIBA for short , hereinafter referred to as “the Academy ”) .

Article 2 (Purpose) This academy aims to promote the development of Korean management studies through the convergence and integration of AI and management . Furthermore, it aims to contribute to the development of future Korean academy and the establishment of future strategies for companies .

Article 3 (Location of Office) The office of this academy shall be located under the Industrial Policy Research Institute (located at 7th floor , 203 Sinchon-ro, Seodaemun-gu, Seoul) , and regional headquarters may be established in other regions . A small number of administrative staff may be employed to operate the office .

Article 4 (Business)

1. In order to achieve the purpose of Article 2, this academy conducts the following business.
(1) Research related to theory and practice for grafting and integrating AI and Business
(2) Conducting research services commissioned by the government or companies at the level of industry-academia-government cooperation
(3) Holding regular academic seminars, policy discussions, and research presentations
(4) Publication of academic journals and research books
(5) Exchange with foreign similar academic societies
(6) Other projects necessary to achieve the purpose of this academy
2. When this academy carries out its business objectives, the benefits it provides to the beneficiaries shall be provided free of charge . However , in unavoidable cases, it may obtain prior approval from the competent authority and have the beneficiaries bear a portion of the cost .
3. The benefits provided through the performance of the objectives of this academy shall not discriminate based on the beneficiary’s gender, place of birth, school attended, place of employment, occupation, or social status .

Chapter 2 Members

Article 5 (Types of Members) Members of this academy are individuals and groups that agree with the purpose of Article 2 , and include regular members (annual members and lifetime members) , associate members , and institutional members .

Article 6 (Regular Member) Regular members are individuals who fall under each of the following categories and have completed the membership process .

1. Full-time faculty members who teach or research AI or business-related subjects at universities , colleges, and graduate schools.
2. or part-time faculty or doctoral program or have obtained a degree in AI or business administration
3. Researchers at accredited research institutes and employees of AI- related companies
4. Other individuals recognized by the Board of Directors as having equivalent qualifications.

Article 7 (Classification of regular members) Regular members are classified into annual members and lifelong members . Annual members are those who have paid the annual membership fee , and the annual membership qualification is one year from the date of registration . Annual members are renewed by paying the annual membership fee . Lifelong members are those who have paid the lifetime membership fee .

Article 8 (Associate Member) Associate members are students enrolled in an undergraduate program in AI or business administration or individuals who have completed the membership process .

Article 9 (Institutional Member) Companies and organizations that agree with the purpose of this academy may become institutional members upon resolution of the board of directors. Membership is in principle for one year, but may be renewed .

Article 10 (Rights and Obligations) Members of this academy have the following rights and obligations .

1. All members can attend the general meeting and participate in discussions , and participate in the academy's business such as research presentations .
2. Regular members have the right to vote and be elected , but institutional members do not have the right to vote or be elected .
3. Members have the obligation to cooperate in the operation of the academy, including paying membership fees and attending academic conferences .

Article 11 (Suspension and loss of membership) Members of this academy will have their membership suspended or lost in the following cases :

1. If the membership fee is not paid , membership will be suspended and the rights under Article 9, Paragraphs 1 and 2 will be suspended until the membership fee is paid in full .
2. Members who damage the reputation of the Academy by violating the purpose of the Academy or the code of ethics established by the Academy shall lose their membership if the Board of Directors resolves to expel them .
3. The Code of Ethics is separately established by the Board of Directors and approved by the

general meeting .

Article 12 (Withdrawal of membership) Members of this academy may withdraw at will .

Chapter 3 Executive Officers

Article 13 (Types of Executive Officers) The Executive officers of this academy shall be a president , vice president , directors and auditors , and the composition of the officers shall be as follows . However , regional headquarters shall be established in major regions at home and abroad , and the head of the regional headquarters shall be the vice president .

1. 1 Chairman
2. About 20 vice presidents
3. The number of directors is 5 to 50 people.
4. 2 auditors

Article 14 (Term of executive officers) The term of executive officers shall be two years, but they may be reappointed .

Article 15 (Appointment and election of executives)

1. The chairman and auditors are elected at the general meeting .
2. The Vice-Chairman and Directors are recommended by the Chairman and appointed by the Chairman with the approval of the Board of Directors . However, the first Vice-Chairman and Directors are appointed by the Chairman and approved by the General Meeting .
3. When an officer other than the chairman is unable to perform his/her duties during his/her term of office, the board of directors shall elect a replacement , and the term of office of the officer elected through the replacement shall be the remaining term of office of his/her predecessor .

Article 16 (Duties of Executive Officers)

1. The president represents the academy and supervises its business .
2. The director attends the board of directors meeting and deliberates on matters related to the business of this academy , and serves as the board of directors or chairman .

from Handle delegated tasks .

3. The auditor performs the following duties :

(1) Auditing the financial status and business of this academy.

(2) Auditing matters related to the operation of the board of directors

(3) If any irregularities or injustices are discovered in the audit results of (1) and (2), the Board of Directors or Requesting the General Assembly to correct the situation

(4) When necessary for reporting under (3) , requesting the convocation of a general meeting or of directors meeting.

Article 17 (Acting Chairman) When the Chairman is unable to perform his/her duties due to an accident or vacancy, the Vice Chairman designated by the Board of Directors shall act in his/her stead for the remainder of the term .

Article 18 (Dismissal of Executive Officers) If an executive officer loses membership qualifications pursuant to Article 11 or commits any of the following acts, he or she may be dismissed upon resolution of the Board of Directors .

1. Unfair acts such as serious dereliction of duty that run counter to the purpose of this academy.
2. Any act that causes a dispute between executives and is judged to make it impossible for the executives to perform their duties.
3. Any other acts that unfairly interfere with the work of this academy.

Chapter 4 General meeting

Article 19 (Matters to be resolved at the general meeting) The general meeting shall resolve the following matters :

1. Appointment and dismissal of the chairman and auditors
2. Endorsement of the first executive team
3. Changes to the Articles of Academy
4. Approval of settlement and business report
5. Provisions on the rights and obligations of members
6. Voting on matters proposed by the Chairman and Board of Directors
7. Other important matters

Article 20 (Convening of the General Meeting)

1. The general meeting is divided into a regular general meeting and a special general meeting .
2. The regular general meeting is held in February every year , and the extraordinary general meeting is convened by the chairman in the following cases :

(1) When the Chairman deems it necessary
(2) When there is a resolution by the board of directors
(3) When more than one-fifth of the members request in writing, stating the reason for holding the meeting.
(4) When the auditor requests a meeting pursuant to the provisions of Article 18, Paragraph 4.

Article 21 (Quorum for General Meeting Resolutions)

1. The general meeting shall be opened with members present , and attendance may be delegated by scanning an electronic document of a written or signed document.
2. The resolutions of the general meeting shall be decided by a majority vote of the members present, excluding those present in writing . However , In case of a tie, the chairman decides .
3. If the Steering Committee determines that it is difficult to convene an offline general meeting,

an online general meeting may be held instead.

Article 22 (Reasons for exclusion from general meeting resolution) If the chairman or member falls under any of the following, he or she may not participate in the resolution .
1. Matters concerning oneself in the appointment and dismissal of executives
2. Matters involving the receipt of money or property, and in which the interests of the chairperson or member conflict with those of the academy.

Chapter 5 Board of Directors

Article 23 (Matters to be resolved by the Board of Directors) The Board of Directors resolves and agrees on the following matters :
1. Convening of a temporary general meeting
2. Resolution of agenda to be submitted to the general meeting
3. Matters delegated by the general meeting
4. Approval of executives appointed by the Chairman
5. Approval of business plan and budget
6. Approval and amendment of the composition and operating regulations of various committees.
7. Important changes to the journal's editorial policy
8. Matters concerning investment and fund management
9. Resolution on membership of institutional members
10. Resolution on expulsion of members
11. Determination of membership fees
12. Other Matters

Article 24 (Delegation of Board of Directors' Resolutions) Routine matters among the Board of Directors' resolutions may be delegated to the Management Committee through a Board of Directors resolution .

Article 25 (Board of Directors) The Board of Directors is composed of a chairman , vice chairman , directors , and auditors .

Article 26 (Convening of the Board of Directors) The Board of Directors shall be convened by the Chairman, who shall also serve as its chairman, in the following cases :
1. When the Chairman deems it necessary
2. When more than half of the members of the board of directors, excluding the chairman, request a meeting by stating the purpose of the meeting.

Article 27 (Quorum for Board of Directors Resolution)
1. The board of directors shall open with the attendance of a majority of its members, and attendance may be delegated in writing .

2. The Board of Directors' resolutions are decided by a majority vote of the members present, excluding those present in writing . However , In case of a tie, the chairman decides .
3. If the Management Committee determines that it is impossible to convene a Board of Directors meeting due to scheduling issues that require a Board of Directors resolution,
If a decision is made, the vote can be made online or in writing .

Article 28 (Reasons for exclusion from board of directors' resolution) A board of directors member may not participate in the resolution if any of the following applies .
1. Matters concerning oneself in deciding on the loss of membership qualifications
2. Matters involving the receipt of money or property and in which the interests of the member and the academy conflict.

Chapter 6 Various Committees

Article 29 (Management Committee) An Management Committee is established to ensure smooth decision-making and efficient execution of the Academy's business .
1. The Management Committee deliberates and decides on matters delegated by the Board of Directors in accordance with Article 24 of these Articles of Incorporation .
2. The Steering Committee is composed of the Chairman, Vice Chairman , Directors , and Secretariat members , and is composed of 10 or so members.
Appoint one person as chairman .
3. The operating committee chairman convenes the operating committee as needed .
4. The Steering Committee shall open with the attendance of a majority of its members , and resolutions shall be passed with the approval of a majority of the members present.

Article 30 (Editorial Committee) An editorial committee shall be established to publish the journal of this academy , and regulations regarding the editorial committee shall be separately determined by resolution of the board of directors .

Article 31 (Fund Management Committee) A Fund Management Committee shall be established to manage the funds of this Academy , and the regulations regarding this shall be separately determined by resolution of the Board of Directors .

Article 32 (Report on Committee Activities) The chairperson of each committee shall report the results of its activities to the board of directors .

Article 33 (Secretariat) A secretariat may be established to manage the affairs of this academy , and the regulations of the secretariat shall be established through a resolution of the board of directors .

Chapter 7 Assets and Accounting

Article 34 (Composition of assets)

The assets of this academy consist of the following items :

1. Donation
2. Membership fee
3. Income from basic assets
4. Other income

Article 35 (Types of Assets)

1. The assets of this academy are divided into two types: basic assets and general assets .
2. The basic assets listed in each clause below cannot be disposed of or provided as collateral . However , when there is a reasonable cause, a portion of them may be disposed of or provided as collateral with the approval of the Board of Directors and the authorization of the Minister of Strategy and Finance .

(1) Property contributed and designated as basic property

(2) Property that the board of directors has decided to use as basic property

3. Ordinary assets consist of assets other than basic assets .

Article 36 (Assessment and collection of membership fees) The method of levying and collecting membership fees shall be determined by the Board of Directors .

Article 37 (Expenses)

The expenses of this academy are paid from general assets .

Article 38 (Management of Assets)

1. The assets of this academy shall be managed by the chairman , and the method of management shall be determined by the board of directors .
2. This academy must open an Internet homepage and disclose the annual amount of donations raised and the results of their use through it .

Article 39 (Disposal of Surplus)

If there is a surplus at the end of the fiscal year, all or part of it is incorporated into capital or carried over to the next fiscal year, as determined by the Board of Directors .

Article 40 (Approval of Budget and Settlement of Accounts)

1. The budget for each fiscal year of this academy must be resolved by the Board of Directors and approved by the General meeting prior to the start of the fiscal year .
2. The chairman has the authority to execute the budget and business accounting .
3. The Chairman shall prepare a general accounting settlement report and a fund management settlement report within two months after the end of the fiscal year, attach an audit opinion, and report to the general meeting for approval .

Article 41 (Executive Compensation) As a rule , no compensation is paid to executives .

Article 42 (Fiscal Year)
fiscal year of this academy runs from September 1 to August 31 of the following year .

Chapter 8 Amendment of Articles of Academy and Dissolution

Article 43 (Amendment of Articles of Academy)
These Articles of Association may be amended with the approval of the Board of directors by a majority vote of at least three- quarters of the members present .

Article 44 (Dissolution)
This academy may be dissolved with the consent of more than two -thirds of its members .

“AI Business Academy” Research Ethics Regulations

Chapter 1 General Provisions

Article 1 (Purpose) The purpose of these research ethics regulations (hereinafter referred to as “ these regulations ”) is to present basic principles and directions for the members of the “AI Business Academy” (hereinafter referred to as the “ the Academy ”) to promote research ethics and truthfulness, create a sound research atmosphere , and contribute to academic and social development .

Article 2 (Scope and Scope)

1. These regulations apply to all members of the academy and stakeholders directly or indirectly related to research activities .
2. Except in cases where there are other special provisions related to the establishment of research ethics and verification of research authenticity, these regulations shall apply .

Article 3 (Scope of misconduct) The misconduct set forth in these regulations includes the following acts that may damage the reputation of not only individuals but also the academy as a member of the academy .

1. Forgery : Here, “ Forgery ” refers to the act of falsely creating non-existent data or research results .
2. Modulation : Here, “ Modulation ” refers to an act of distorting research content or results by intentionally changing or deleting research data .
3. Plagiarism : Here, “ plagiarism ” refers to the act of stealing another person’s ideas , research content , research results, etc. without clearly indicating the source .
4. Double publication of papers : Here, “ Double publication of papers” refers to the act of publishing the same data, analysis methods, and analysis results in two or more academic journals .
5. Unfair authorship of a paper : Here, “ unfair authorship ” refers to an act of not granting authorship of a paper to a person who has made scientific or technical contributions or contributions to the research content or results without just cause , or granting authorship of a paper to a person who has not contributed as an expression of gratitude or courtesy .
6. Duplicate research : Here, “ Duplicate research ” refers to an act of conducting two or more research projects with the same content and publishing the same research results .
7. Public False Statement : Here, “ Public false statement ” refers to an act of making false statements about one’s academic background , career , qualifications , research achievements , results , etc.
8. Refers to an act of intentionally obstructing an investigation into suspicions of misconduct by oneself or another person or causing harm to the whistleblower .
9. Refers to any act that seriously deviates from the scope normally tolerated by the Academy

of Information Systems in relation to other research .

Chapter 2 Basic Ethics for Researchers

Article 4 (Researcher's Morality) Researchers must conduct research in an honest and correct manner, not in a socially acceptable or ethically wrong manner, and present the results . Accordingly, the ethics that researchers must observe in relation to research are as follows . 1. Using others' ideas, research content, or research results (hereinafter " research content ") without clearly indicating them is clear plagiarism , and researchers must indicate the source as follows :

(1) When citing someone else's research content without re-editing (direct quotation) , the author's name , year , and page number must be indicated in the text with quotation marks ("") , and the complete source must be indicated in the references section .

(2) When editing and using the research content of others, the author's name and year must be indicated in the text , and the full source must be indicated in the references section .

(3) When requoting content quoted by others, you must indicate " requotation " along with the quoter's name and year .

2. When conducting research, researchers must recognize acts such as falsification or alteration of data as serious crimes and make efforts to prevent such misconduct from occurring .

3. It must not duplicate the previously published (or under review) research papers as if they were new research . Therefore, papers published in academic journals must be original work .

4. Those who have contributed directly or indirectly to the research must be listed as co-authors in the research paper , and those who have not contributed cannot be listed as authors . In addition, when conducting joint research , the person who has contributed the most to the research is the first author , and the corresponding author is limited to the person who is in charge of all correspondence between the authors and the academic academy and who has overseen the research or an equivalent person . However , in cases where co-authors cannot be listed, such as in cases of administrative , financial, or research support, this should be indicated in a footnote or acknowledgement .

5. It must truthfully write about their academic background , career , qualifications , research achievements , results , etc.

6. In relation to other research, one must not engage in any conduct that goes beyond the scope generally accepted in academia .

Article 5 (Social Responsibility) Researchers must be deeply aware of the impact their research will have on academy and, as professionals, must fulfill their responsibilities and obligations regarding the results of their research .

Article 6 (Respect for Others) Researchers must not harm or infringe upon the rights of others and must respect intellectual property rights . In addition , when dealing with others in relation to research, they must be treated equally regardless of race , gender , nationality ,

disability , etc.

Chapter 3 Roles and Responsibilities of the Research Ethics Committee

Article 7 (Composition and term of office of the committee)

1. The Research Ethics Committee (hereinafter referred to as the " Committee ") shall be composed of 10 or fewer members, including the chairperson .
2. The chairman of the committee concurrently serves as the editor-in-chief of this academy .
3. The committee member concurrently serves as an editorial committee member of this academy .
4. The term of office for the chairperson and members shall be one year, but they may be reappointed .
5. The secretary of the committee shall be one of the vice-presidents of this academy , and the secretary shall not have voting rights .

Article 8 (Functions of the Committee) The Committee deliberates on the following matters :

1. Matters concerning fraudulent acts such as forgery , falsification , and plagiarism related to academic papers
2. Matters concerning research ethics raised regarding research plans , research results, etc. related to the academic academy
3. Investigation into misconduct related to the academic academy
4. Complaints raised regarding the truthfulness of research related to academic societies or journals
5. Matters concerning research ethics and education in research and development
6. Matters concerning the processing and follow-up measures of the results of research integrity verification
7. Matters to prevent misconduct that may occur during research
8. Other matters regarding research ethics raised by the president or committee chairperson

Article 9 (Convening , review , and resolution of the committee)

1. When a report of research misconduct is received, the chairperson convenes a committee to investigate and deliberate .
2. The committee is established with the attendance of a majority of its members , and resolutions are passed with the approval of a majority of the members present .
3. When the chairperson deems the agenda for deliberation to be minor, he/she may replace it with a written deliberation .
4. During the investigation process, the committee may request the attendance of the informant , the person under investigation , witnesses , and reference persons when necessary .

Article 10 (Duties and protection of rights of whistleblower)

1. " Whistleblower " refers to a person who has reported to this committee facts or related

evidence of an act of misconduct .

2. The informant may report in any possible way , including verbally , in writing , by phone , or via e - mail . In principle, reports must be made under one's real name . However , even if the report is anonymous, if the report includes the name of the research project , the title of the paper , and specific details and evidence of the misconduct in writing or via e-mail, and if it is found to be true, it will be processed in the same way as a report made under one's real name .

3. Under no circumstances will the identity of the informant be directly or indirectly exposed , and the informant's name will not be included in the investigation results report for the purpose of protecting the informant unless absolutely necessary .

4. A whistleblower who reported something even though he or she knew or could have known that the information was false is not eligible for protection .

5. The informant has an obligation to respond to the committee's request to appear .

Article 11 (Duties and protection of rights of the subject of investigation)

1. " Subject of investigation " refers to a person who has become the subject of an investigation into an unethical act through a report or knowledge thereof, or a person who has become the subject of an investigation due to being presumed to have participated in an unethical act during the course of the investigation . Reference persons or witnesses during the course of the investigation are not included .

2. The person under investigation has an obligation to faithfully respond to any requests made by the committee during the investigation process . In addition, the person under investigation is guaranteed the right and opportunity to express opinions , raise objections, and defend himself/herself during the investigation process .

3. The committee must take care to ensure that the reputation or rights of the person under investigation are not violated until the verification of whether or not there was any misconduct is complete , and must endeavor to restore the reputation of the person under investigation who is found not to have been involved in any misconduct .

4. The subject of the investigation must faithfully comply with the request for submission of evidence requested by the committee .

Article 12 (Confidentiality) All matters related to the investigation, including reporting , investigation , deliberation , resolution, and recommendation measures , shall be kept confidential . Persons directly or indirectly participating in the investigation and related committee members shall not disclose any information obtained during the investigation and performance of duties . If the need for reasonable disclosure arises, it may be disclosed after a resolution by the committee .

Article 13 (Notification and Objection)

1. The results of the judgment on misconduct must be notified to the reporter and the person under investigation .

2. If there is a reasonable basis for an objection filed by the reporter or the subject of the investigation against the results of the decision, the objection may be filed within 30 days from the date of notification , and a reinvestigation may be conducted if necessary .

Chapter 4 Research Ethics Committee Verification Procedures and Standards

Article 14 (Preliminary Investigation)

1. A preliminary investigation refers to a procedure to determine whether or not there is a need to investigate a suspicion of misconduct , and must be initiated within 30 days from the date of receipt of the report . The preliminary investigation is decided by a preliminary investigation committee consisting of a chairperson , vice chairperson , two members , and a secretary .
2. If the subject of the investigation admits to all facts of the misconduct as a result of the preliminary investigation , a decision can be made immediately without going through the main investigation procedure , and if it is determined that there is a possibility of significant damage to the evidence, the chairperson can take measures to preserve the evidence prior to the formation of the committee .
3. If it is decided not to conduct a main investigation in the preliminary investigation , the specific reason for this will be notified to the reporter within 10 days from the date of decision . However , no notification will be given in the case of anonymous reports .

Article 15 (Main Investigation)

1. “ Main investigation ” refers to the procedure to verify the facts of an act of misconduct , and must be conducted by forming a committee in accordance with the provisions of Article 7, Paragraph 1 .
2. The entire investigation schedule from the start to the decision must be completed within 6 months . However , if it is determined that the investigation cannot be conducted within this period, the reason for this may be notified to the president of the academic academy and the investigation period may be extended .
3. Main investigation If it is determined that fair progress is difficult due to a conflict of interest between a member of the committee and the subject of the investigation, the president of the academy may restrict the participation of the member in the relevant matter . However , the chairperson and vice chairperson are exceptions .
4. If the informant or the person under investigation raises a legitimate objection regarding the avoidance of a specific member, the committee may decide to accept it .

Article 16 (Recording and reporting of investigation results)

1. Records related to misconduct must be kept in document form at the academy for 5 years .
2. The committee must report the results , contents , and decisions of the investigation to the president of the academy and the board of directors .
3. The record of the judgment contents includes the following items .

(1) Report contents

(2) Allegations of misconduct and related papers or reports that are the subject of the investigation
(3) Objections or arguments of the informant and the subject of the investigation
(4) Judgment results and related supporting materials and witnesses
(5) List of judges

Article 17 (Post - judgment measures)

1. After the committee completes the review, it decides on the type of disciplinary action .
2. The types of disciplinary action are as follows depending on the severity of the misconduct , but may be applied multiple times .
(1) Expulsion
(2) Suspension of membership
(3) Restrictions on publication in academic journals (AI Management Journal)
(4) Official apology at the conference
(5) Notification of misconduct regarding related papers in academic journals or newsletters
3. If the investigation results confirm that there was no misconduct, the committee may take appropriate follow-up measures to restore the reputation of the person under investigation .

Article 18 (Administrative Matters)

1. Matters not specified in these regulations shall be decided and implemented by the committee in accordance with social norms .
2. The revision of the research ethics regulations shall follow the revision procedures of this academy .
3. For the smooth operation and activities of the committee, the academy Necessary administrative and financial support must be provided .

AI Journal of Business
Call for Papers

The AI Business Academy, a leader in the AI era, is recruiting papers to be published in Volume 2, in March 2025, AI Journal of Business.
Now, all companies and businesses cannot grow and develop without utilizing AI and big data. In addition to the existing AI models representing artificial intelligence. The AI Business Academy would like to invite excellent papers that present various cases of utilizing AI and big data to innovate business and create new opportunities, as well as AI models that are utilized and models of AI strategies that serve as the basis.
We hope for active participation from many people.

1. Recruitment areas: Various management strategies and business innovations based on AI and big data, engineering improvements, and creative fields that combine management and engineering.

2. Recruitment deadline: January 31, 2025

3. Journal publication schedule: Reviewed by the end of February 2025, then published in March

4. Thesis Format: **Principle of submission as English version of paper** . It is much more advantageous and meaningful to write a paper in English in order to share and attract more interest from AI researchers around the world and to be cited in foreign papers. If you request it to the contact information below, a paper format file will be distributed. The basic format is the same as the format of this published journal (volume 1).

5. Submission : Academy website or editorial committee email: jhchang@assist.ac.kr

6. Contact: AI Business Academy Office (**aiba2023@naver.com)** or jhchang@assist.ac.kr

AI Journal of Business

Vol. 1, No. 1

Issued on November , 2024

Publisher : Dongsung Cho

Editor : Joongho Chang

Publisher : AI Business Academy

7 floor , 203 Sinchon-ro, Seodaemun-gu, Seoul, Korea

Tel) **02-360-0775**

ISSN 3058-6992