142th Master's Thesis
Thesis Advisor : Junyeong Kim

A Study of Effective Knowledge Distillation: From Specialized Domain to Multi-Modal Application

효과적인 지식 증류에 관한 연구: 특화 도메인에서 다중 모달 응용까지

February 2025

The Graduate School
Chung-Ang University
Major in AI application of the department of AI
Sangyeon Cho

A Study of Effective Knowledge Distillation: From Specialized Domain to Multi-Modal Application

효과적인 지식 증류에 관한 연구: 특화 도메인에서 다중 모달 응용까지

Presented to the Faculties of the Chung-Ang University in

Partial Fulfillment of the Requirement of the Master's Degree

February 2025

The Graduate School
Chung-Ang University
Major in AI application of the department of AI
Sangyeon Cho

A Study of Effective Knowledge Distillation: From Specialized Domain to Multi-Modal Application

by

Sangyeon Cho

Department of AI
Chung-Ang University

Date: 17th February 2024

Bugeun Kim

Junyeong Kim

Hwanhee Lee

Thesis submitted in partial fulfillment of
the requirements for the degree of Master in the Department of
AI in the Graduate School
of Chung-Ang University

2025

ABSTRACT

A Study of Effective Knowledge Distillation: From Specialized Domain to Multi-Modal Application

Sangyeon Cho
Major in AI Application
Department of Artificial Intelligence
The Graduate School, Chung-Ang University

Knowledge Distillation (KD) is a deep learning technique that transfers knowledge from a high-performance and complex model (teacher model) to a relatively simpler and less complex model (student model). This thesis explores the effectiveness of KD in two specific contexts: domain-specific tasks, particularly in the medical field, and multi-modal applications. The first chapter of the thesis introduces research on distilling domain-specific knowledge in the field of natural language processing to improve classification performance on electronic medical record data. The second chapter addresses research on cross-modal interaction in multi-modal applications, focusing on enhancing the performance of automated audio captioning tasks through visual-audio feature distillation. By integrating findings from domain-specific and multi-modal knowledge distillation approaches, this thesis investigates the flexibility and impact of KD techniques across various domains and applications.

Keywords: *Natural Language Processing, Multi-modal Learning, Knowledge Distillation, Electronic Medical Record, Automated Audio Captioning*

Contents

List of Tables

List of Figures

1. Chapter 1: Introduction

1.1 Knowledge Distillation in Specialized Domain

The use of pre-trained language models fine-tuned to address specific downstream tasks is a common approach in natural language processing (NLP). However, acquiring domain-specific knowledge via fine-tuning is challenging. Traditional methods involve pretraining language models using vast amounts of domain-specific data before fine-tuning for particular tasks. This study investigates emergency/non-emergency classification tasks based on electronic medical record (EMR) data obtained from pediatric emergency departments (PEDs) in Korea. Our findings reveal that existing domain-specific pre-trained language models underperform compared to general language models in handling N-lingual free-text data characteristics of non-English-speaking regions. To address these limitations, we propose a domain knowledge transfer methodology that leverages knowledge distillation to infuse general language models with domain-specific knowledge via fine-tuning. This study demonstrates the effective transfer of specialized knowledge between models by defining a general language model as the student model and a domain-specific pre-trained model as the teacher model. In particular, we address the complexities of EMR data obtained from PEDs in non-English-speaking regions, such as Korea, and demonstrate that the proposed method enhances classifica-

tion performance in such contexts. The proposed methodology not only outperforms baseline models on Korean PED EMR data, but also promises broader applicability in various professional and technical domains. In future works, we intend to extend this methodology to include diverse non-English-speaking regions and address additional downstream tasks, with the aim of developing advanced model architectures using state-of-the-art KD techniques. The code is available in https://github.com/JoSangYeon/DSG-KD.

Keywords: *Bilingual medical data analysis, Emergency Room electronic health records, Code switching, knowledge distillation, multilingual language models, Natural Language Processing*

1.2 Knowledge Distillation in Multi-modal Domain

This study proposes VggCaps and Multi2Cap. VggCaps is a dataset designed to address the limitations of current Automated Audio Captioning (AAC) datasets, which typically feature short and simple captions. VggCaps includes longer, more complex captions and has been validated through comprehensive analysis and human evaluations. Multi2Cap is a framework that extends AAC tasks into a multi-modal domain. By leveraging Cross-modal Feature Distillation (CFD) and Large Language Model (LLM), Multi2Cap improves the quality and depth of the generated captions. Multi2Cap achieves state-of-the-art results on the Clotho and AudioCaps benchmarks, offering new challenges and directions for AAC research.

Keywords: *Audio Processing, Knowledge Distillation, Large Language Models, Multi-modal Learning, Natural Language Generation, Multimodality and Language Grounding to Vision-Robotics and Beyond, NLP engineering experiment, Data resources, Data analysis,*

2. Chapter 2:

DSG-KD: Knowledge Distillation from Domain-Specific to General Language Models

2.1 Introduction

With the proactive utilization of Electronic Medical Records (EMR) by healthcare institutions worldwide, a vast amount of medical information is being stored as data. [14–16] In particular, EMRs are documented in Pediatric Emergency Departments (PEDs) to capture patients' conditions at the time of visit, test results, and additional details in the form of free-text. These free-text entries are crucial components of EMRs and include essential clinical notes, such as the patient's gender, age, and vital signs. Given the importance of initial responses in PEDs, distinguishing between emergency and non-emergency patients based on EMRs recorded at the time of admission is critical. [1, 2, 26] However, owing to the nature of PED environments, clinical notes are often recorded in an unstructured and inconsistent manner, complicating the classification of emergency and non-emergency patients based on them. [3–5] This reduces the utility of the data in supporting decision-making by healthcare professionals, such as physicians. [17, 18]

In particular, EMR data obtained from PEDs in Korea, a non-English-speaking

country, are written in Korean. Non-English-speaking countries present an additional layer of complexity in classification tasks owing to the use of both English and local languages. [6, 7] In such cases, medical jargon and test results are often written in English or using English abbreviations, while the patient's condition and chief complaints are predominantly documented in the local language (Korean). [8, 9] Thus, in non-English speaking countries, the data often come in the form of N-lingual free-text, with critical medical terms presented in English and other parts present in the local language. This form of data presents significant challenges in downstream tasks and during the extraction of meaningful words.

To overcome these problems, language models have been introduced based on transformers pre-trained on specific domain data to equip them with general domain knowledge. [55, 56] In particular, KM-BERT [57], pre-trained on Korean medical data, has been proposed to perform downstream tasks in the medical domain based on fine-tuning. However, KM-BERT does not perform well in the context of data collected in non-English-speaking countries, such as Korea. As depicted in Figure 2.1, in the task of classifying emergency and non-emergency cases using EMR data obtained from Korean PEDs, general language models, such as Ko-BERT and BERT-base, outperformed domain-specific pre-trained models, such as KM-BERT and Clinical-BERT, in terms of Area Under the Receiving Operating Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC) [19, 20]. This indicates that language models pre-trained with medical domain data are vulnerable to complications caused by N-lingual free-text characteristics of EMR data. The current study was envisioned to address this issue.

In this study, we address this problem by extracting domain knowledge for training using Knowledge Distillation (KD) [63]. We define an LM pre-trained with medical domain data as the teacher model and a general LM as the student model. Then, we extract medical knowledge from the former model and transfer it to the general latter model. We identify words containing medical knowledge in the input text and design the training to enable the student model to learn the teacher model's hidden states and attention matrices for these words. Our experiments demonstrate that the proposed method achieves effective knowledge transfer between LMs, yielding the best performance in classifying emergency and non-emergency cases based on PED EMR data.

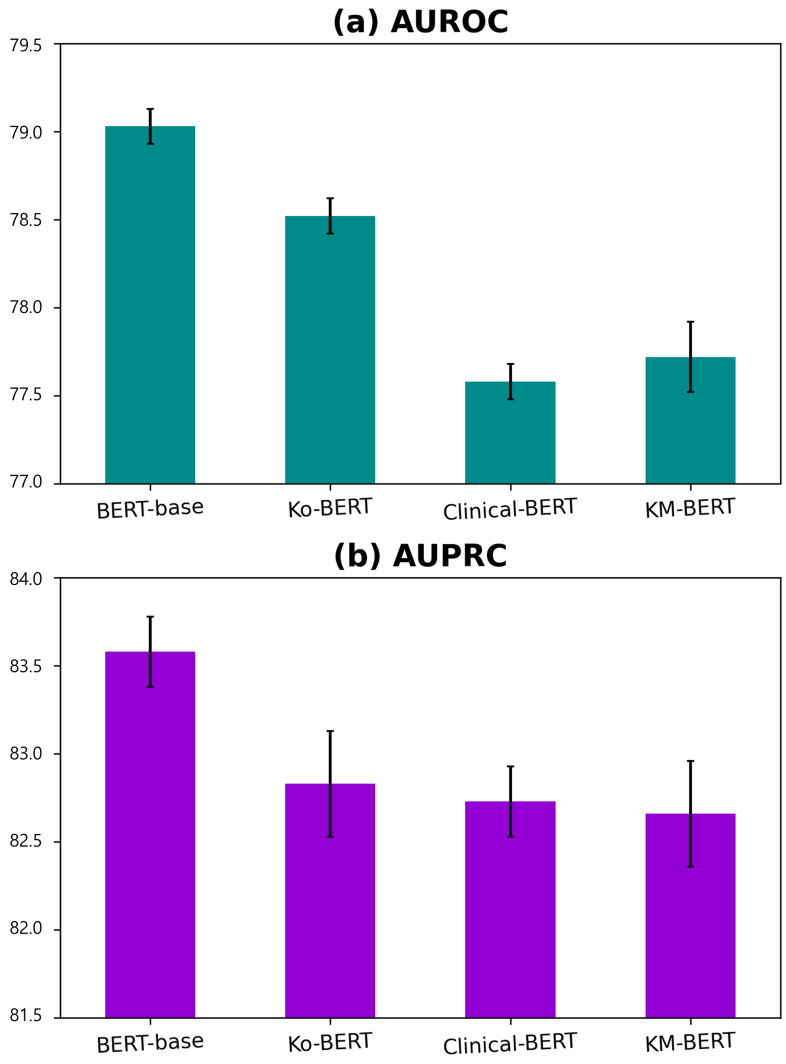


Figure 2.1: **Generalized LM performs better.** Performance of each pre-trained LM on an emergency/non-emergency classification task using EMR data from Korean PEDs in terms of (a) AUROC and (b) AUPRC.

	Total number of words	Number of Korean words(medical words)	Number of English words(medical words)	Number of other words
Train	4,599,471	1,989,252 (99,197)	1,071,523 (220,221)	1,538,696
Dev	1,152,777	499,117 (24,590)	269,161 (55,156)	384,499
Test	1,444,904	624,343 (31,372)	335,464 (68,684)	485,097
total	**7,197,152**	**3,112,712 (155,159)**	**1,676,148 (344,061)**	**2,408,292**

Table 2.1: This table shows for the **analysis of data on a word-by-word basis**, the ratio of Korean:English:Other is used to signify the language distribution. In this case, it is approximately 0.43:0.23:0.33. The proportion of medical terms in Korean is 5%, while that in English is 20%, which is very high.

2.2 Related Works

2.2.1 NLP for Clinical Notes

EMR [24, 25] represents digitally formatted data that are systematically collected and electronically stored to document patients' health information. In recent years, EMR data obtained from PEDs have garnered significant attention as the effective processing of such data is likely to enhance decision-making in clinical environments. [26–28] Various related studies have focused on the utilization, analysis, and decision-support methods of EMR data obtained from PEDs from multiple perspectives. [34–36]

Early research on EMR data processing primarily targeted English-based data [40], applying traditional NLP methods such as Bag of Words (BoW) [29–31] and Term Frequency-Inverse Document Frequency (TF-IDF) [32, 33]. However, owing to the significant increase in the amount of acquired medical data in recent years, pre-trained LMs based on transformers have emerged [37, 38], and active research

has been conducted on fine-tuning these models for various medical datasets. Notably, LMs such as KM-BERT and Clinical-BERT have been pre-trained on extensive medical data, exhibiting great potential in various NLP tasks in the medical domain. [55, 59]

However, the aforementioned LMs exhibit certain vulnerabilities in the case of N-lingual, free-text-formatted EMR data obtained from PED environments of non-English-speaking countries (see Figure 2.1). Given the importance of initial responses in PEDs, such issues are critical. [49, 50] This study describes methodologies capable of effectively learning and processing such N-lingual and free-text formatted data.

2.2.2 N-Lingual Free-text Data

The processing of N-lingual and free-text data presents various motivations and challenges for NLP and deep learning. Early research on NLP focused primarily on English-based datasets. [40] However, data obtained from diverse countries has motivated the study of N-lingual processing [41–44], leading to extensive research. This has led to the development of robust models capable of understanding the nuances of mixed languages to enhance the efficiency and accuracy of analyzing texts containing multiple languages [54]. Moreover, with the increase in online content, including social media posts, emails, news articles, and comments, the amount of unstructured free-text data has also increased. [39, 43] Consequently, numerous studies on NLP have aimed to extract meaningful insights from unstructured data and transform them into structured information. In particular, LM development in

NLP currently focuses on maximizing the understanding of free text. [53]

In non-English-speaking countries, the collection of various data types has led to the emergence of datasets that exhibit both N-lingual and free-text characteristics. [51, 52] This also constitutes a major topic of study within NLP. Indeed, research has been conducted on EMR data collected at PEDs in non-English-speaking countries, in which both N-lingual and free-text characteristics are prevalent. [60,61] Existing research has identified critical issues with the performance of traditional LMs and NLP methods on such data. Unlike previous studies, we aim to enhance our understanding of data with N-lingual and free-text characteristics using KD.

2.2.3 Knowledge Distillation

In this study, domain knowledge is transferred from the teacher model to the student model using KD. [63] Traditionally, KD is used in the NLP field to transfer knowledge from a larger teacher model to a smaller student model, with a focus on model size and lightweighting. [64, 65] Previous research approached this by having the student model mimic the teacher model's predictions using relatively simple methods. Recently, KD between transformer-based LMs has evolved to mimic the representation vectors of hidden states and attention matrices. This study, however, does not focus on model size and lightweighting, but rather on "extracting the unique knowledge possessed by the teacher model and transferring it to the student model." Although we use conventional distillation methods [64, 65], we first define the knowledge that the student model needs to learn and subsequently

extract it from the teacher model for distillation. Detailed explanations are provided in [Section 3].

By examining previous studies that applied KD from a knowledge transfer perspective rather than a lightweighting perspective, active research has been conducted on transferring knowledge from image models to language models [66, 67]; however, little research has been conducted on knowledge transfer between language models. [68] Existing studies on knowledge transfer between language models have focused on the quality of overall knowledge transfer, rather than domain-specific knowledge transfer. Our approach differs from those proposed in previous studies in that it focuses on the exchange of domain knowledge between language models in different domains.

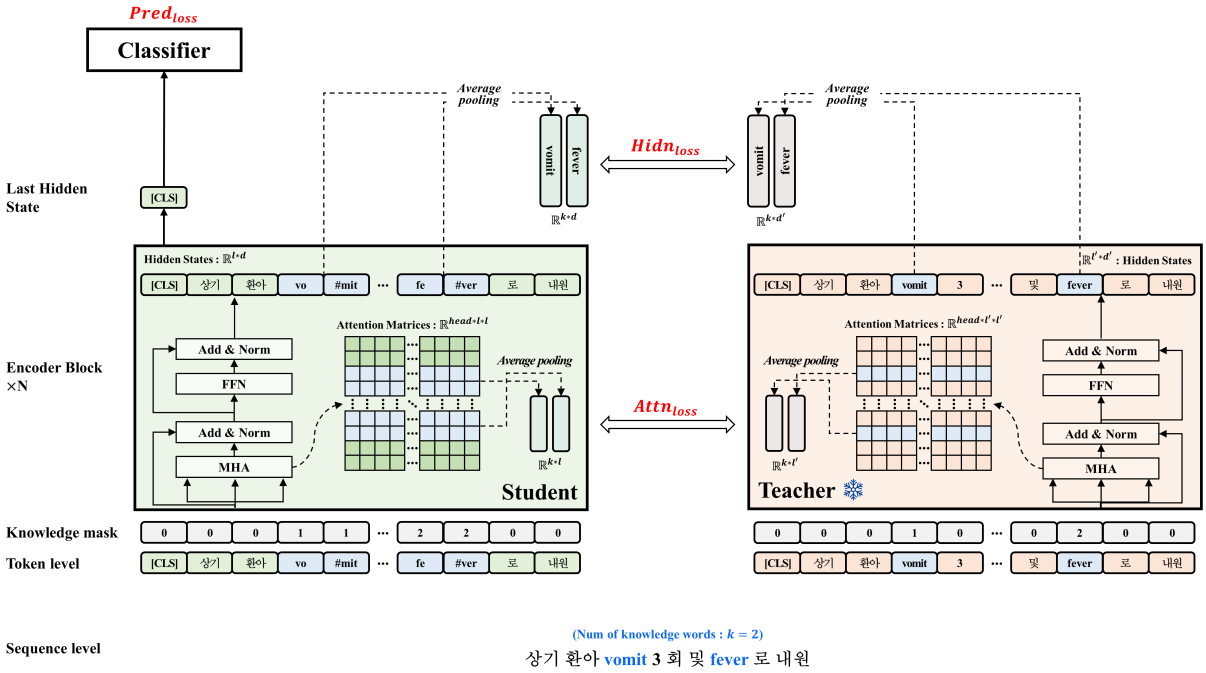


Figure 2.2: This figure illustrates **visualization of the proposed methodology**. The architecture consists of both student and teacher models based on transformers, which act as encoder blocks with multi-head attention (MHA) and feed-forward networks (FFN). The student model performs a prediction task ($\mathcal{L}_{\text{pred}}$) and receives appropriate domain knowledge from the teacher model by minimizing $\mathcal{L}_{\text{hidn}}$ and $\mathcal{L}_{\text{attn}}$. In the figure, we define "vomiting" and "fever" as domain knowledge words ($k = 2$) and perform distillation by receiving appropriate representations from the hidden states and attention matrices that arise from the teacher model. Overall, the goal of the proposed architecture is to transfer the teacher model's knowledge to the student, not only in terms of classification predictions but also throughout the model's internal representation, thereby enabling the student model to make decisions based on a deeper and more nuanced understanding of the input data.

2.3 Proposed Method: DSG-KD

In this section, we formalize the central problem considered in this study and introduce the proposed methodology.

2.3.1 Text Preprocessing & Labeling

EMR data are obtained from Korean PEDs and then analyzed and preprocessed before model training. Even though South Korea is a non-English-speaking coun-

try, English is often used alongside local languages. This is reflected in the data obtained, which are recorded in the form of bilingual free-text clinical notes. The preprocessing and analysis methods, described later, are summarized in Table 2.1.

First, the bilingual free-text clinical notes are preprocessed by performing language and symbol distinction processing and new character removal while retaining the meaning of the existing content as far as possible. Newline characters ($\backslash r$, $\backslash n$) generated during the process of recording and storing data in medical institutions are removed, and blanks are inserted between languages and symbols to separate them appropriately. This step is also performed for a method described below. Because we use a byte-pair encoding-based tokenizer, inserted blanks do not affect the input words severely [74]. In addition, because it is not possible to specify the features to be used as labels in the acquired data, active doctors are consulted to determine the label criteria using columns such as test status and medication history, as recorded in the data.

As per the preprocessed data, approximately 43% of the words are Korean, 23% are English, and 33% are symbols and other words. In addition, the number of English words in the data that are medical terms is evaluated by creating our own Korean medical dictionary by crawling the Korean Medical Search Engine and online medical dictionaries. According to Table 2.1, approximately 20% of the English words in the data are medical words, whereas only 5% of the Korean words are medical terms. This implies that almost all medical terms are more biased toward English than local languages and that medical domain knowledge is more embedded in English than in local languages.

2.3.2 DSG-KD

This section describes the domain knowledge transfer methodology proposed in this thesis in detail. The goal of this methodology is to transfer domain knowledge from a language model (teacher) pre-trained on a specific domain to a language model (student) pre-trained on general-purpose data. An overview of the framework is presented in Figure 2.2.

We define S as a generalized language model that categorizes urgent/non-urgent cases based on free-text clinical note data as input data. We train and validate the student model S. In addition, we define a domain-specific language model T that conveys domain knowledge and aids learning when S performs emergency or non-emergency classification tasks. The input free-text clinical notes are denoted by $X = \{x_i^S, x_i^T\}_{i=0}^N$. Here, x_i^S denotes the input of the S model, and x_i^T denotes the input of the T model. We also define $Y = \{y_i\}_{i=0}^N$, $Y \in \{0, 1\}$, and label it as emergency/non-emergency. Finally, we construct a dataset Data $= \{X, Y\}$ consisting of X and Y.

Definition of Domain Knowledge

To transfer domain knowledge, we must define the domain knowledge present in the input x_i. This is defined as English words in the PED's EMR data (because, as discussed previously, medical knowledge that the teacher needs to extract is almost always expressed in English, instead of the local language). To represent domain knowledge words (subwords) in the input sequence x_i, we represent the

knowledge mask $\mathbf{M}$ as follows:

$$\begin{aligned} x_i &= \{t_1, t_2, \cdots, t_{l-1}, t_l\} \in \mathbb{R}^l \\ \mathbf{M}_i &= \{m_1, m_2, \cdots, m_{l-1}, m_l\} \in \{0, 1, \cdots, k\}^l \end{aligned} \tag{2.1}$$

where the knowledge mask $\mathbf{M}$ is $\mathbf{M} \in \mathbb{R}^l$, where each element m_l of $\mathbf{M}$ represents the index of the word for which the token at that position expresses a particular piece of domain knowledge. If $m_l = 0$, the token does not contain domain knowledge; if $m_l > 0$, the input x_i contains domain knowledge. In addition, k denotes the number of domain knowledge words. The appropriate domain knowledge token d_i is extracted from the input sequence x_i using $\mathbf{M}$.

$$d_i = \{x_i[j] | m_j > 0, \forall j \in \{0, \cdots, l\}\} \tag{2.2}$$

where d_i denotes the domain knowledge tokens corresponding to each token of the input x_i with the set relation $d_i \in x_i$.

Choosing the scope of domain knowledge definition is directly linked to model performance in the proposed methodology. In the data considered in this study, knowledge words are not defined as exclusively medical terms because the boundaries between medical and non-medical terms in the language are blurred. Thus, embedding the medical background knowledge of the teacher model in the English language itself is expected to benefit student learning. In particular, defining the appropriate domain knowledge based on the available data can be adopted as an inductive bias in the proposed domain knowledge transfer methodology, which is

expected to yield better synergy.

In addition, the domain knowledge representation extracted from the teacher model must be defined. As we consider the language model of the transformer encoder series, we define the hidden states and attention matrices occurring in the encoder blocks of each layer as representations that imply domain knowledge. We distill the representation of English words observed in each input sequence x_i from the teacher to the student model.

Problem Formulation

Let S and T contain P transformer layers. Knowledge transfer is achieved by performing KD on the encoder layers of the student and teacher models. S further solves the prediction task of classifying emergency/non-emergency cases, given X. Formally, the student receives domain knowledge from the teacher and solves the prediction task by minimizing the following objective function:

$$\begin{aligned} \mathcal{L}_{total} = & \sum_{x \in X} \mathcal{L}_{pred}(S(x^S), Y) \\ & + \sum_{d \in X}^{M} \sum_{p=0}^{P} \mathcal{L}_{layer}(f_p^S(d^S), f_p^T(d^T), \lambda) \end{aligned} \tag{2.3}$$

where $\mathcal{L}_{pred}$ denotes the loss function that optimizes the classifier prediction of the student model performing the emergency/non-emergency classification task and $\mathcal{L}_{layer}$ denotes the loss function applied to a given layer of the model. Here, $f_p(\cdot)$ denotes the encoder block output in the p-th layer. The hyperparameter λ comprises $\lambda = \{\alpha, \beta\}$, where α and β are applied to $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$, respectively.

Knowledge distillation

In this section, we mathematically express the process by which a student model receives domain knowledge from a teacher model. To this end, two types of KD are performed: hidden-state and attention-matrix distillation. [64, 65]

Hidden-state Loss is expressed as follows :

$$\mathcal{L}_{hidn} = \alpha \mathsf{MSE}(\hat{H}^S, \hat{H}^T) \tag{2.4}$$

Let H_S and H_T be the hidden states of the student and teacher encoder layers, respectively. Let H^S and H^T be $H^S \in \mathbb{R}^{l \times d}$ and $H^T \in \mathbb{R}^{l' \times d'}$, respectively. (Note that $d = d'$, $l = l'$).

$\hat{H}^S$ and $\hat{H}^T$ are expressed as follows:

$$\begin{aligned} \hat{H}^S &= \sum_{i=0}^{k} \text{Avg.Pool}(H^S_{m_i^S}, m^S) \in \mathbb{R}^{k \times d} \\ \hat{H}^T &= \sum_{i=0}^{k} \text{Avg.Pool}(H^T_{m_i^T}, m^T) \in \mathbb{R}^{k \times d'} \end{aligned} \tag{2.5}$$

Let m^S be a mask marking the domain knowledge words that the student model should receive from the teacher model. m^T is a mask marking the domain knowledge words of the teacher model. $m^S \in \mathbb{R}^l$ and $m^T \in \mathbb{R}^{l'}$.

Attention Matrices Loss is expressed as follows :

$$\mathcal{L}_{attn} = \beta \mathsf{MSE}(\hat{A}^S, \hat{A}^T) \tag{2.6}$$

Let A_S and A_T be the hidden states of the encoder layer for the student and teacher models, respectively. Let A^S and A^T be $A^S \in \mathbb{R}^{h \times l \times l}$ and $A^T \in \mathbb{R}^{h \times l \times l}$, respectively (but $l = l'$).

$\hat{A}^S$ and $\hat{A}^T$ are expressed as follows:

$$\begin{aligned} \hat{A}^S &= \sum_{i=0}^{k} \text{Avg.Pool}(A^S_{m_i^S}, m^S) \in \mathbb{R}^{k \times l} \\ \hat{A}^T &= \sum_{i=0}^{k} \text{Avg.Pool}(A^T_{m_i^T}, m^T) \in \mathbb{R}^{k \times l'} \end{aligned} \tag{2.7}$$

Finally, for each encoder layer, the loss is calculated using the following formula:

$$\mathcal{L}_{layer} = \begin{cases} \alpha \mathcal{L}_{hidn} & n = 0 \\ \alpha \mathcal{L}_{hidn} + \beta \mathcal{L}_{attn} & n \neq 0 \end{cases} \tag{2.8}$$

where $n = 0$ denotes the embedding layer and $\mathcal{L}_{attn}$ is not performed because no attention matrix is generated.

Prediction Loss

The student model continues to perform and optimize the classification task.

$$\mathcal{L}_{pred} = \mathsf{CE}(S(x^S), Y) \tag{2.9}$$

where x^S denotes the input to the student model and $S(x^S)$ denotes the logits of the student model. $\mathsf{CE}(\cdot)$ denotes the cross entropy loss and Y is the label.

2.4 Experiments

2.4.1 Dataset

The proposed methodology is validated using EMR data obtained from Korean PEDs. The task considered in this study is to classify the data into emergencies and non-emergencies using binary classification.

2.4.2 Implementation Details

In this section, we introduce the experimental environment and models on which the proposed methodology is applied. The A6000 GPU device is used, with a batch size of 32, 15 epochs of training with an early termination condition at 15 to choose the weight with the best valid performance. The learning rate is set to 1e-5 for the BERT body and 1e-2 for the classifier. For reproducibility, we set the random seed to 42.

The transformer-based language models Ko-BERT and BERT-base are used as baseline student models, and KM-BERT, Clinical-BERT (C-BERT), RoBerta, Bio-RoBerta (B-RoBerta), and Bio-Multilingual-BERT (Bio-M-BERT) are used as baseline teacher models. In addition, four BoW-based machine learning models (RF, LR, XGBoost, and GB) are trained and compared.

The transformer-based models have M=12, d=768, l=512, and h=12, where M denotes the number of encoder layers, d denotes the dimension of hidden state, l denotes the maximum sequence length, and h denotes the number of multi-headers. The classification performances of the transformer models trained and measured on

		Accuracy	AUROC	AUPRC	Recall	Precision	F1 Score	Average
Ko-BERT		71.9±0.2	78.5±0.1	82.8±0.3	76.2±0.2	75.8±0.2	76.0±0.2	76.9±0.2
BERT-base		69.2±0.2	79.0±0.1	83.6±0.2	61.2±0.4	**81.7±0.2**	70.0±0.3	74.1±0.2
KM-BERT		71.0±0.1	77.7±0.2	82.7±0.3	77.9±0.2	74.0±0.3	75.9±0.2	76.5±0.2
Clinical-BERT		70.9±0.1	77.6±0.1	82.7±0.2	77.1±0.2	74.3±0.3	75.6±0.1	76.4±0.2
Bio-M-BERT		66.4±0.2	71.1±0.2	76.4±0.4	72.6±0.3	70.8±0.4	71.7±0.2	71.5±0.3
RoBerta		70.6±0.2	77.3±0.2	81.7±0.2	71.0±0.2	77.1±0.3	73.9±0.1	75.3±0.2
Bio-RoBerta		72.3±0.1	78.4±0.1	82.9±0.2	79.3±0.2	74.9±0.2	77.0±0.1	77.5±0.2
					ours			
Student	**Teacher**							
Ko-Bert								
	KM-BERT	72.6±0.4	79.4±0.4	84.1±0.4	**85.7±0.2**	72.7±0.5	**78.6±0.3**	**78.9±0.4**
	Clinical-BERT	72.6±0.2	79.4±0.2	83.8±0.2	74.6±0.3	77.8±0.2	76.2±0.2	77.4±0.2
	Bio-M-BERT	**73.4±0.2**	**80.4±0.2**	84.3±0.2	75.1±0.3	78.6±0.2	76.8±0.2	78.1±0.2
	RoBerta	73.1±0.2	80.1±0.2	84.6±0.2	77.7±0.3	76.7±0.2	77.2±0.2	78.3±0.2
	Bio-RoBerta	73.3±0.2	**80.4±0.2**	**85.0±0.2**	77.6±0.3	76.9±0.2	77.3±0.2	78.4±0.2
Bert								
	KM-BERT	70.9±0.2	77.1±0.1	82.3±0.2	79.6±0.3	73.2±0.3	76.2±0.2	76.5±0.2
	Clinical-BERT	70.8±0.2	76.7±0.2	81.6±0.3	74.4±0.3	75.4±0.2	74.9±0.2	75.6±0.2
	Bio-M-BERT	70.4±0.1	76.5±0.2	81.6±0.2	77.3±0.2	73.6±0.2	75.4±0.2	75.8±0.2
	RoBerta	70.8±0.2	76.9±0.2	82.0±0.2	84.6±0.2	71.1±0.2	77.2±0.1	77.1±0.2
	Bio-RoBerta	68.6±0.2	76.3±0.2	81.2±0.3	64.6±0.4	80.0±0.3	70.7±0.3	73.2±0.3

Table 2.2: This table shows the **performances of different models in terms of all metrics**. The mean and standard deviation values are listed. The models are validated with respect to Transformer-based models (Ko-BERT, BERT-base, KM-BERT, Clinical-BERT, Bio-M-BERT, RoBERTa, and Bio-RoBERTa). The column entitled "ours" lists the results obtained using the proposed methodology—each combination of student and teacher models is evaluated. The metrics used for evaluation are accuracy, area AUROC, AUPRC, recall, precision, and F1 Score, and the average values are provided in the Average column. In the column corresponding to each metric, the best performance is highlighted in **Bold** and the second-best performance is Underlined.

the test set are compared with that of the proposed method with all possible combinations of student-teacher models while training on the test set. In the proposed methodology, knowledge transfer and classification are performed simultaneously.

2.4.3 Results

Table 2.2 presents a comparison of all methods in different groups in terms of accuracy, AUROC, AUPRC, recall, precision, and F1 Score. In addition, the average values are listed in the Average column to facilitate the overall performance

comparison. In general, transformer-based NLP models are observed to outperform traditional NLP processes proposed in previous studies. Transformer-based language models can adapt quickly to the language patterns present in data when fine-tuned using language patterns during pretraining.

In particular, it is interesting to note from Table 2.2 that language models pre-trained with data from the medical field (KM-BERT, Clinical-BERT, etc.) perform worse than general language models (BERT-based, Ko-BERT) in terms of AUROC and AUPRC. This is because, as explained earlier, general language models perform favorably with respect to N-lingual and unstructured data while pretraining with a large amount of data, whereas language models pre-trained on the medical domain struggle on N-lingual or unstructured real data, even though they have learned medical domain information.

The bottom part of Table 2.2 presents the results obtained by training the existing language models using the proposed methodology. The student models are divided into Ko-BERT- and BERT-based cases and teacher models are taken to be pre-trained language models in the medical domain. The results are observed to be significantly better than those obtained by fine-tuning a simple language model. Moreover, the performance is even better than that of existing student and teacher models when trained alone. The aforementioned experimental results demonstrate the effectiveness of knowledge transfer between language models, leading to efficient interaction between knowledge of different language models.

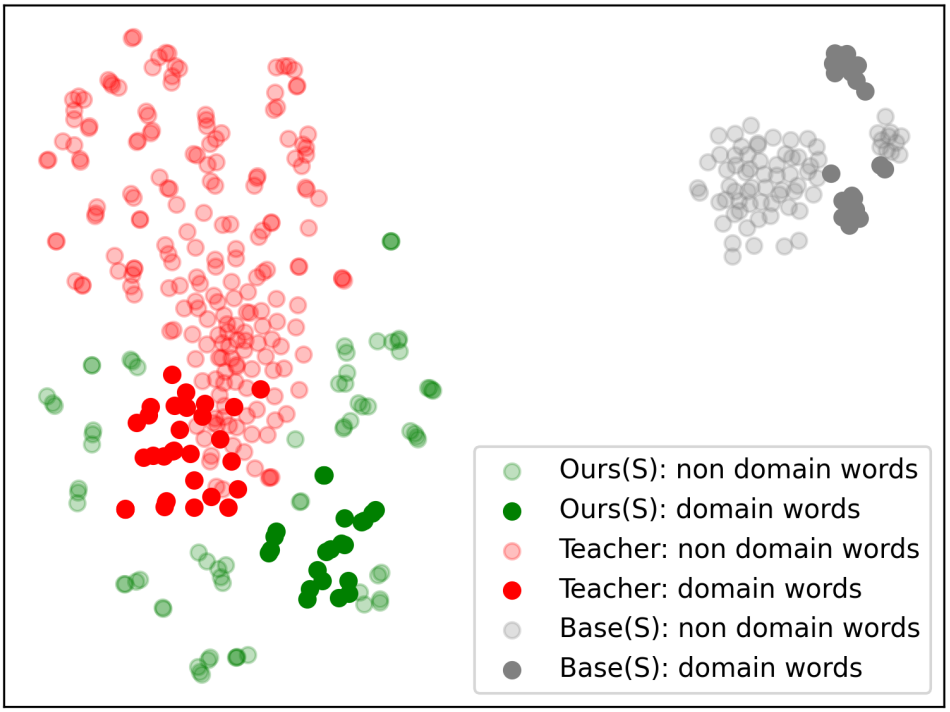


Figure 2.3: This figure illustrates **visual representation of the effectiveness** of the proposed training method. S denotes a student model, which is taken to be Ko-BERT, and the teacher is taken to be KM-BERT. The gray dots correspond to an independently trained student model, the green dots correspond to the proposed training methodology, and the red dots correspond to the spatial coordinates of the teacher model's representation

Qualitative Analysis

To verify the relationship between domain words and teacher embeddings before and after DSG-KD visually, we present a scatter plot in Figure 2.3. Random samples are obtained from the dataset and visualized. In the figure, Ko-BERT is used as the student model and KM-BERT is used as the teacher model.

First, for each color, the darker shade represents the coordinates of the embedding vectors for words that contain domain knowledge, and the lighter shade represents the coordinates of the embedding vectors for words that do not. The gray

points represent the coordinates of the embedding space when the student model is trained independently. The green points represent the coordinates of the embedding space of the student model trained using the proposed methodology. The red points represent the coordinates of the embedding space corresponding to the teacher model, which provides a good understanding of the medical domain. Visually, the embedding space of the student model trained using the proposed methodology is similar to that of the teacher model, which is rich in medical expressions. In contrast, when the student model is trained independently, it yields a relatively poor representation space and its visual representation is located far from that of the teacher model.

Case Analysis

Furthermore, to understand the impact of the proposed methodology on Domain words (in this study, medical words), we defined and measured the Medical Word Proportion Score (MWPS). The MWPS is defined as follows:

$$\text{MWPS} = \frac{\%\text{ of medical terms in English words}}{\%\text{ of English words in all words}} \tag{2.10}$$

This can be expressed mathematically as:

$$\text{MWPS} = \frac{\frac{m}{E}}{\frac{E}{A}} = \frac{m \cdot A}{E^2} \tag{2.11}$$

where m is the number of medical terms in English words, E is the total number of English words, and A is the total number of words.

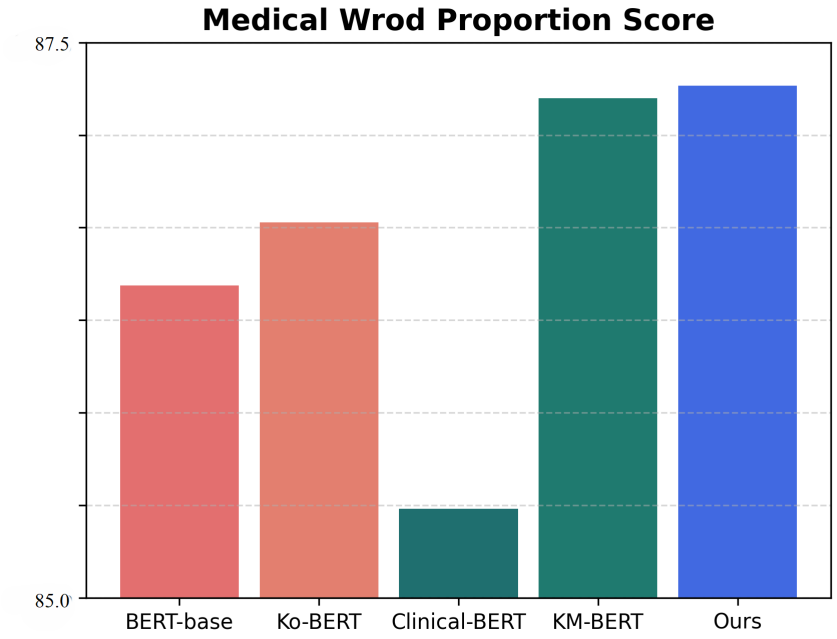


Figure 2.4: This figure shows **calculated MWPS for correctly classified cases** in the test set for each model. The proposed model (Ours) uses Ko-BERT as the student model and KM-BERT as the teacher model. By achieving the highest MWPS performance compared to all other models, our methodology demonstrates an effective understanding of medical words. Notably, the higher MWPS compared to the teacher model is particularly encouraging in terms of domain knowledge comprehension.

We measured the MWPS for correctly classified cases in the test set for each model. This allowed us to evaluate the effectiveness of the proposed methodology. The comparison of MWPS for each model is shown in Figure 2.4. From the figure, we can observe that our model achieves the highest MWPS, indicating that the DSG-KD methodology effectively learns Domain words, i.e., Domain Knowledge. Specifically, it demonstrates that medical knowledge can be transferred to a general language model and that our model shows a slightly higher understanding of medical terms than KM-BERT. In summary, this suggests that the proposed method-

$\mathcal{L}_{pred}$	$\mathcal{L}_{hidn}$	$\mathcal{L}_{attn}$	AUROC	AUPRC	F1 Score
✓	✓	✓	**79.4±0.4**	**84.1±0.4**	**78.6±0.3**
✓	-	-	78.5±0.1	82.8±0.3	76.0±0.2
✓	✓	-	79.2±0.3	83.9±0.3	78.6±0.2
✓	-	✓	78.1±0.3	83.1±0.4	67.8±0.3

Table 2.3: **Ablation study on objective function for $\mathcal{L}_{hidn}$, $\mathcal{L}_{attn}$**

ology is highly beneficial for applications in specific domains such as the medical field. Thus, it is proven that the DSG-KD methodology learns the medical knowledge embedded in medical terms and achieves superior classification performance.

2.4.4 Ablation Study

The influence of $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$ on performance is investigated by considering Equation 2.8. The influences of α and β are also studied. Two additional experiments are performed with Ko-BERT as the student model and KM-BERT as the teacher model. The performances are evaluated in terms of three metrics: AUROC, AUPRC, and F1 Score.

Influence of $\mathcal{L}_{hidn}$ & $\mathcal{L}_{attn}$ on the objective function

Table 2.3 presents the effects of $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$ on $\mathcal{L}_{total}$. The baseline model, which includes all loss components, exhibits an AUROC of 79.4±0.4, an AUPRC of 84.1±0.4, and an F1 Score of 78.6±0.3. This is used as the baseline for comparison in the ablation study, which is performed by removing $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$ individually, or in combination. The results are as follows.

α	β	AUROC	AUPRC	F1 Score
1.0	**1.0**	78.8±0.4	83.4±0.4	78.2±0.3
1.0	**0.6**	79.0±0.4	83.5±0.4	75.1±0.4
0.9	**0.5**	78.5±0.4	83.1±0.4	78.3±0.3
0.8	**0.4**	79.0±0.4	83.6±0.4	78.4±0.3
0.7	**0.3**	79.1±0.4	83.6±0.3	77.9±0.4
0.6	**0.2**	**79.4±0.4**	**84.1±0.4**	**78.6±0.3**
0.1	**0.1**	77.8±0.3	81.7±0.5	76.9±0.3

Table 2.4: **Ablation study on objective function for α and β**

Case 1: Removing both $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$: When $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$ are removed, all three performance metrics decrease, indicating that the two losses play significant roles in improving model performance. Case 2: Removing only $\mathcal{L}_{hidn}$: The absence of $\mathcal{L}_{hidn}$ also results in a decrease in performance, although to a lesser extent than in Case 1. This suggests that hidden losses contribute to model performance to a certain extent. Case 3: Removing $\mathcal{L}_{attn}$: The absence of $\mathcal{L}_{attn}$ has the largest effect on F1 Score, suggesting that this factor is particularly important for the accuracy and recall balance of the model. In contrast to Cases 2 and 3, removing both $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$ further degrades the performance, emphasizing the importance of synergistic functioning of these components to achieve high model performance.

This study demonstrates that each component plays an important role, and that prediction loss is the most important factor in maintaining the integrity of model performance. The cumulative degradation caused by removing multiple components suggests that the interactions among these components are complex and essential for high-quality model learning.

Effect of α & β on the objective function

Table 2.4 presents the impact of α and β on the performance, expressed in terms of performance. When $\alpha = 0.6$ and $\beta = 0.2$, AUROC = 79.4±0.4, AUPRC = 84.1±0.4, and F1 Score = 78.6±0.3 are observed. We systematically vary α and β to observe changes in the performance, and the results are presented below.

Case 1: Impact of equal weights ($\alpha = \beta = 1.0$): Assigning equal weights to both hyperparameters results in a slight decrease in all metrics, suggesting that the model may not need to emphasize the factors controlled by α and β equally. Case 2: Increasing the emphasis on β ($\alpha = 1.0$, $\beta = 0.6$): Increasing β relative to α slightly improves the AUROC to 79.0±0.4, indicating that aspects of the model influenced by β may be more important for predictive ability. Case 3: Gradual adjustment of α and β: A pattern is observed when proportionally adjusting β from 1.0 to 0.1 while gradually decreasing α from 1.0 to 0.1, with the highest F1 Score achieved when $\alpha = 0.8$ and $\beta = 0.4$, and the best AUROC and AUPRC at $\alpha = 0.8$ and $\beta = 0.4$. This suggests an optimal range for the hyperparameters that balances the factors controlled by the model. Case 4: Significant reduction of α ($\alpha = 0.1$, $\beta = 0.1$): Significantly reducing both α and β to 0.1 results in the lowest AUROC, a slight increase in AUPRC, and a decrease in the F1 Score, which suggests that very low values of these hyperparameters can degrade the performance of the model, especially in terms of precision and recall balance.

In conclusion, the hyperparameters α and β play important roles in the performance of the Ko-BERT + KM-BERT model. For each data environment, an optimal parameter range yields the best performance. This emphasizes the importance

of fine-tuning hyperparameters to satisfy specific performance goals during model training.

2.5 Conclusion and Future Work

This study addresses the need for effective domain knowledge transfer in pre-trained language models and provides a promising approach for bridging the gap between general-purpose language understanding and domain-specific expertise. As the demand for domain expertise in NLP continues to increase, this study makes a timely and valuable contribution to improving the ability of language models to utilize specific and nuanced domain knowledge.

To this end, we introduce a novel methodology to extract domain knowledge from domain-specific pre-trained language models and transfer the knowledge to more generalized language models. We find that domain-specific pre-trained language models are vulnerable to bilingual data in non-English-speaking countries such as Korea, and demonstrate that by transferring domain knowledge to generalized language models, the proposed methodology is robust in N-lingual environments and exhibits effective classification performance in specific domains.

In this study, a highest average metric of 789±0.4 is achieved by transferring medical knowledge from the teacher model (KM-BERT) to the student model (Ko-BERT) using Korean emergency room EMR data. This demonstrates that the proposed framework can be used as a tool to assist doctors and healthcare workers in decision-making by helping them overcome N-lingual challenges inherent in EMR data recorded in non-English speaking countries. Beyond the medical field, the methodology can also be applied in various professional and technical fields, opening up potential avenues for its use as a decision aid in many forms.

In future works, we intend to extend and improve this methodology by utilizing EMR data obtained from diverse non-English-speaking countries, or by considering downstream tasks from various technical and professional fields in specific non-English-speaking countries. We also aim to develop an advanced model architecture by applying the latest KD techniques to deep learning methods.

3. Chapter 3:

Multi2Cap: Improving Automated Audio Captioning with Cross-modal Feature Distillation and Large Language Model

3.1 Introduction

Automated Audio Captioning (AAC) [75] is the task of generating natural language descriptions from audio content, and it is becoming an increasingly important task in the field of artificial intelligence. Unlike Automatic Speech Recognition (ASR) [83], which converts speech to text, AAC must comprehensively understand and describe not only linguistic elements but also non-verbal audio signals (e.g., environmental noise, animal sounds, music, etc.). Although AAC research is a relatively recent field, it is receiving growing attention due to the increasing demand for complex utilization of audio data, such as audio interaction and retrieval [85, 86]. Driven by this demand, AAC is actively ongoing, with new technologies continuously being developed to efficiently process and describe diverse audio information [88, 126].

However, existing AAC researches [88, 130] have primarily been limited to

the audio-text modality, leaving the potential of multi-modal learning underexplored. Real-world experiences are composed of multi-modalities that combine various types of information [76, 77], such as visual and auditory, and humans understand and explain the world through multiple senses. Moreover, research in artificial intelligence that reflects the real world through multi-modal approaches has been gaining attention [79], and many studies are actively exploring the integration of different modalities to complement each other and improve performance across various fields and tasks [78]. In this context, AAC task require new insights from a broader multi-modal perspective. Utilizing additional modalities in AAC tasks will play a crucial role in handling more complex auditory situations and providing richer descriptions.

Furthermore, current AAC datasets tend to describe audio content in a short and concise manner [103, 105]. In fact, we identified that the existing datasets are represented by an average of 9 words, and the lexical diversity of them are not abundant, which is illustrated in table 3.1, figure 3.2 and 3.3. This can lead to overfitting [80] during training and, limit the ability to provide rich and detailed descriptions of audio content, restricting the task's scalability. In the real world, audio is not described merely as sound [81, 82]; it can be explained with complex elements, including the surrounding environment, the source of the sound, and its context. Therefore, it suggests that AAC tasks need to evolve to express this complex information and provide longer descriptions.

Based on the above motivations, we construct and propose VggCaps, a multi-modal audio captioning dataset that incorporates visual elements. VggCaps is de-

signed to include longer and more complex captions compared to existing datasets. Specifically, we design the prompt to LLMs for incorporate both audio and image data simultaneously, enabling the model to capture the complex relationships between the two modalities more precisely. These captions focus on using visual information complementarily to provide more accurate and enriched descriptions of the audio's meaning. Additionally, to verify the validity and robustness of the dataset, we conduct human evaluations and measured human performance on the dataset.

Additionally, we propose the Multi2Cap framework, a multi-modal approach that extends the existing AAC tasks by incorporating visual information. The goal is to improve the quality and depth of audio captions using image-audio feature distillation and LLMs. Specifically, we apply the Cross-modal Feature Distillation (CFD) technique, which uses image modality complementarily to learn the interactions between audio and images. The audio encoder learns the unique features of the image, maximizing the interaction between the two modalities. This allows the generation of richer audio descriptions based on more comprehensive information. In this study, we utilize VggCaps as pre-training data to train Multi2Cap, and report model performance across various benchmarks.

Main contributions of this study are as follows:

- **We attempt a multi-modal expansion of AAC**. By utilizing image modality complementarily, we perform the AAC task and present a new paradigm in AAC research. This multi-modal approach broadens the boundaries of the traditional AAC task and provides new motivation for future research.

- **We propose a novel multi-modal audio captioning dataset with added visual information**. Using LLMs as a guide, we generate captions by supplementing visual information based on audio content, and these captions are longer and contain more diverse vocabulary than existing datasets. This leads to the creation of a dataset that requires more intricate and high-level descriptions compared to traditional audio captioning datasets.
- **We propose the Multi2Cap framework**, which enhances the understanding of audio-visual interactions for AAC tasks through Cross-modal Feature Distillation. Additionally, through various experiments, we analyze the impact of adding extra modalities to AAC tasks.

3.2 Related works

3.2.1 Automated Audio Captioning

Automated Audio Captioning (AAC) is a task that generates natural language descriptions from audio content, with serious research beginning around 2017. AAC can often be confused with Automatic Speech Recognition(ASR), but the two tasks are fundamentally different [83]. ASR focuses on converting speech from audio into text, aiming to transcribe the content of the audio literally. In contrast, AAC describes not only speech but also surrounding non-verbal sounds. In other words, AAC requires a comprehensive audio understanding ability to interpret the context of the sounds and describe them [84]. Therefore, AAC demands a high-level ability to understand various sounds and their backgrounds, requiring a more in-depth processing of audio data. Due to these characteristics, AAC presents new challenges and unique motivations.

AAC research has primarily developed around neural network models with an encoder-decoder structure. Initially, the approach of combining Convolutional Neural Networks(CNN) and Recurrent Neural Networks(RNN) [75, 85] was commonly used for ACC tasks. More recently, research utilizing pre-trained Transformer models for both the encoder and decoder has gained attention [86]. This developmental process has contributed to generating more sophisticated audio descriptions. Furthermore, the emergence of Large Language Models (LLMs) has brought about a significant breakthrough in the field of artificial intelligence, and has also opened new avenues in AAC research [87]. Leveraging the capabilities of

LLMs, AAC research has seen substantial performance improvements [88]. However, existing AAC research [130] has been limited to the modality of Audio-Text. To address this issue, this study proposes a new approach to the multi-modal expansion of AAC tasks.

3.2.2 Feature Distillation

Knowledge Distillation (KD) [63] is a deep learning technique that distills knowledge from a complex and high-performance model to a relatively simpler and less complex model. The goal of traditional knowledge distillation is model compression, aiming to reduce computation and memory usage while maintaining the performance of the teacher model. On the other hand, recent studies have focused on effectively distilling the unique knowledge of the teacher model [89, 90]. Particularly, research has been actively conducted on improving the richness and precision of knowledge transfer between different modalities from a multi-modal perspective [91,92], not limited to a specific task or domain. Specifically, in the context of Contrastive Learning [93, 94], KD is performed by comparing features across modalities in the same feature space. This technique [145] trains the model so that similar inputs are positioned closer in the feature space, while dissimilar inputs are pushed farther apart. This approach enhances the interaction between modalities and helps align learned representations more consistently and effectively.

In light of this, this study proposes Multi2Cap, which applies the Cross-modal Feature Distillation (CFD) technique, focusing on aligning the interaction between audio and image within the feature space by utilizing image modality as a supple-

mentary component.

3.2.3 Large Language Models

Large Language Model(LLMs) are deep learning-based models that learn from large-scale text data to understand the structure and meaning of language, enabling the generation of natural text. LLMs use billions of parameters to learn from vast amounts of data, demonstrating outstanding performance in various natural language processing(NLP) tasks [97] such as text generation, translation, summarization, and question answering. Representative LLMs include GPT-4 [99] and Gemini [101], and recently, models like LLaMA [121], which can be used openly for AI research, have emerged, leading to active research utilizing LLMs across various fields of artificial intelligence [95, 96]. In other words, LLMs are currently playing an innovative role in research and applications throughout the field of artificial intelligence [98].

In this study, to overcome the limitations of existing Automated Audio Captioning(AAC) datasets, we utilize LLMs to construct an audio captioning dataset with additionally aligned image modality. Existing AAC datasets tend to include relatively simple captions [103–105], and it is evident that such simple descriptions can easily achieve high performance when utilizing LLMs. In this context, this study emphasizes the need for the evolution of AAC tasks by generating longer and more complex vocabulary-inclusive captions. Therefore, we designed a framework to generate audio captions that supplement visual information, guided by LLMs. Furthermore, by going beyond simple descriptions to include more diverse vocab-

ulary and complex contextual information, this study presents a new challenge that advances traditional AAC tasks.

3.3 Proposed Dataset: VggCaps

In this section, we propose and describe the VggCaps dataset for AAC tasks. We explain the pipeline for its construction and post-processing, as well as the multi-faceted analysis of the dataset and the human evaluation procedure. Additionally, this dataset should not be distributed in the real world for any purpose other than research.

We utilize VggSound [102] data to construct the LLM-Guided Audio Captioning Dataset. VggSound is a large-scale audio-visual dataset, designed to enable learning from the sounds occurring in videos. VggSound includes various types of sounds along with the video scenes associated with them, and it has been widely used in a variety of research [106, 107]. However, since VggSound is an audio-visual dataset, it does not contain captions (descriptions) mapped to each audio. Therefore, we utilized a Large Language Model (LLM) to generate appropriate captions.

3.3.1 Data Processing

The VggSound dataset contains only category information for audio and video, but no captions. To generate appropriate captions from such a dataset, we utilized a Large Language Model(LLM) to create captions mapped to the audio-visual data. The processing is conducted through the pipeline shown in figure 3.1.

As input for generating captions, we extract an appropriate snapshot (image) from the video. Then, we extract an audio clip of 10 seconds, consisting of 5 seconds before and after the snapshot. We convert the extracted audio into a Mel-

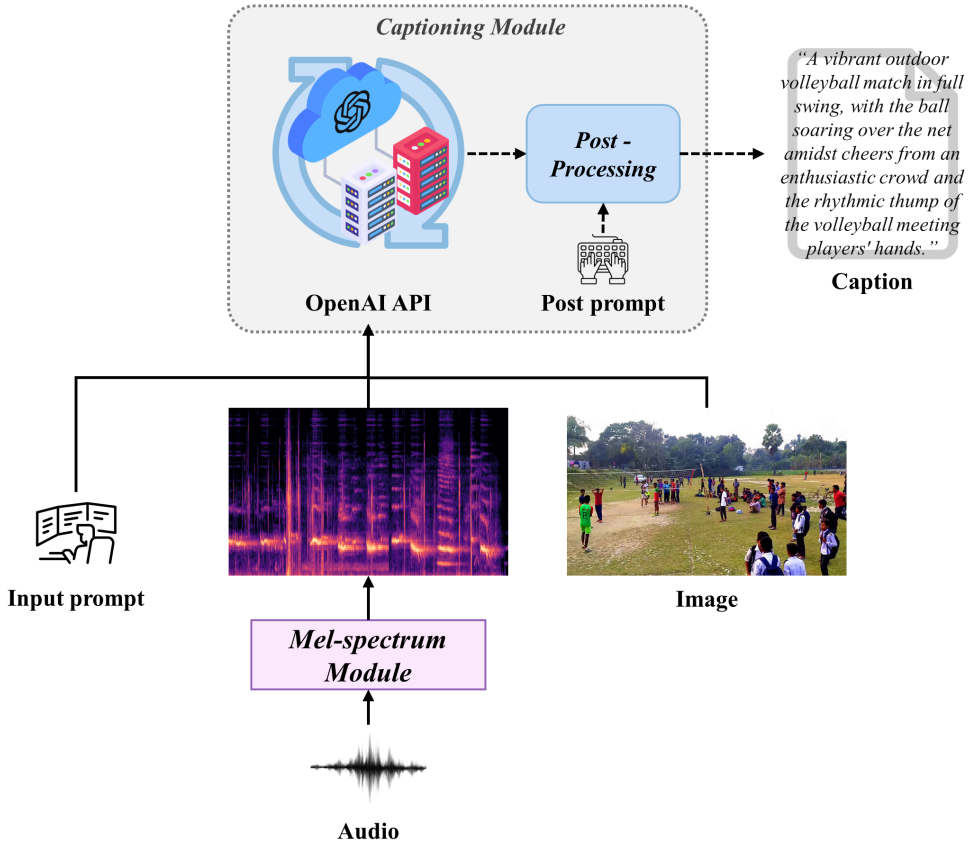


Figure 3.1: **Pipeline of VggCaps Data Processing**

spectrum [108] and input it into the LLM along with the video snapshot. Additionally, we provide the input with a prompt to generate suitable captions, and finally feed it into the LLM. The raw captions generated this way undergo a Post-Processing step. This step refines the captions by removing exaggerated or unnecessary expressions and polishing them into cleaner sentences. The final captions used in this study are those generated after Post-Processing. The Captioning Module in figure 3.1 utilizes GPT-4o [100].

In this study, we obtained a total of 173,494 VggCaps data samples. This

Dataset	num. of row	num. of audio	avg(std). audio length(s)	num. of caption	avg(std). caption length	additional modal
AudioCaps [105]	57,188	51,308	10.0 (0.6)	57,188	9.0 (4.3)	Image(Potentially)
Clotho [104]	29,645	5,929	22.5 (4.3)	29,645	11.3 (2.8)	X
WavCaps [103]	403,050	403,050	67.6 (-)	403,050	7.8 (-)	X
VggCaps (ours)	**173,494**	**173,494**	**10.0 (0.1)**	**173,494**	**21.1 (5.3)**	**Image**

Table 3.1: **Statistics of Datasets** - We statistically compare the existing AAC dataset with VggCaps. VggCaps includes longer captions and additional modalities compared to the existing datasets.

dataset is used as the pre-training dataset for the Multi2Cap framework discussed later. Additionally, 1% of the total dataset, randomly selected, was used as a test subset for validation and performance reporting. Further details and analysis of the dataset are provided in section 3.3.2.

3.3.2 Dataset Analysis

The analysis of the constructed data focuses on comparing the generated captions with existing AAC datasets. Table 3.1 provides a statistical comparison between our dataset and the existing datasets. There are two notable differences. First, the caption length is about twice as long as that of the existing datasets, which is a result of using LLM to describe the content with richer vocabulary. Second, Snapshot images extracted from the video were included as an additional modality. This allows the VggCaps dataset to enable multi-modal research in AAC tasks.

The second analysis evaluates how the constructed VggCaps dataset uses more diverse vocabulary and describes the content in a more complex manner compared to the existing datasets. The analysis focuses on readability level and lexical diversity. The results of the analysis are provided in figure 3.2, 3.2. First, readability

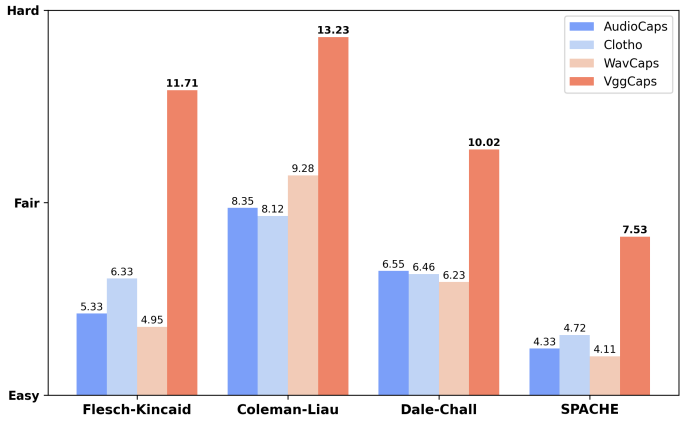


Figure 3.2: **Readability Level Comparison by Datasets**

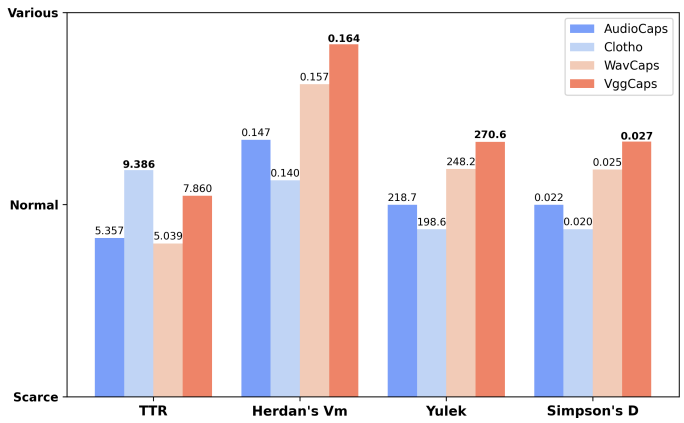


Figure 3.3: **Lexical Diversity Comparison by Datasets**

level (figure 3.2) is evaluated using four metrics: Flesch-Kincaid [110], Coleman-Liau [111], Dale-Chall [112], and SPACHE [113]. These metrics indicate that the lower the score, the easier the text is to read, whereas higher scores indicate the need for deeper understanding. In all metrics, the VggCaps shows a higher level compared to the existing datasets. Lexical diversity (figure 3.3) is analyzed using four metrics: Type-Token Ratio(TTR) [114], Herdan's VM [115], Yulek [116], and Simpson's D [117]. Higher values for these metrics indicate the use of more di-

Figure 3.4: **Wordcloud in VggCaps - VERB**

verse vocabulary. In the figure, each metric is normalized to a scale from 0 to 10, with the actual values before normalization displayed. The analysis confirms that the constructed dataset uses a more diverse vocabulary compared to the existing datasets.

Finally, figure 3.4, 3.5 shows the word cloud of the captions in the constructed dataset. For verbs, it can be observed that more linguistically sophisticated expressions such as “fill” and “reverberate” are used, rather than simple expressions like “hear” and “sound.” Additionally, for nouns, not only words with auditory meanings but also words with spatial or visual meanings are included.

3.3.3 Human Evaluation/Performance

To verify the validity and robustness of the constructed dataset, we conducted an experiment to perform human evaluation on a subset of the VggCaps dataset and derive human performance. For this purpose, 100 samples were randomly selected from the test subset of the VggCaps, and we recruited 18 evaluators who

Figure 3.5: **Wordcloud in VggCaps - NOUN**

volunteered to participate. Among them, 10 evaluators were responsible for the human evaluation of the captions, while the remaining 8 were tasked with human performance.

The purpose of the human evaluation was to assess how accurately the captions of the VggCaps describe the audio content. The evaluators were provided with audio samples and asked to evaluate how accurately each caption described the corresponding audio. For this, they were instructed to assign a score between 1 and 5 based on the Mean Opinion Score(MOS) [109] method. A score of 1 indicates that the caption does not describe the audio accurately at all, while a score of 5 indicates that the caption perfectly describes the audio. The evaluation results measured an MOS score of 4.1 ± 0.09. This suggests that the captions of VggCaps are generally accurate and reliable, with a high level of agreement among the evaluators. The distribution of MOS scores is visually presented in Figure 3.6. Additionally, the evaluators checked whether the evaluation data contained any sensitive information and agreed that it did not.

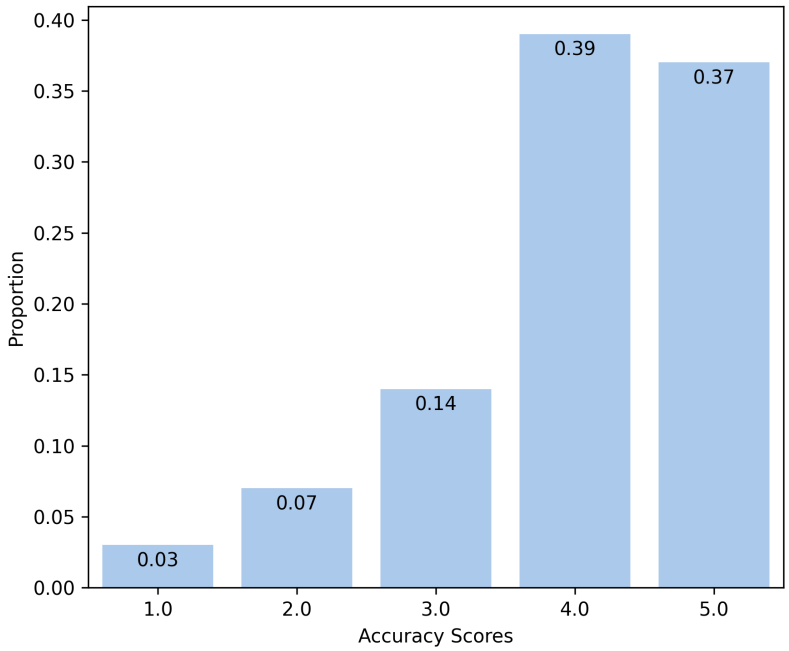


Figure 3.6: **Mean Opinion Score (MOS)**: This table shows the distribution of MOS for VggCaps calculated through human evaluation. It indicates that the vast majority of samples have appropriate captions.

Human performance experiment was conducted in two stages. In the first stage, the evaluators were asked to generate captions for the audio content provided to them. In the second stage, aligned video snapshots were provided as supplementary material along with the audio, and the evaluators were asked to generate captions based on this information. This aimed to evaluate how humans perform in single-modality versus multi-modality situations. In other words, it allowed us to assess the impact of providing multi-modal information on caption generation and compare how performance changes when evaluators utilize multi-modal information. The specific results and analysis are further detailed in the section 3.5.2 and table 3.2.

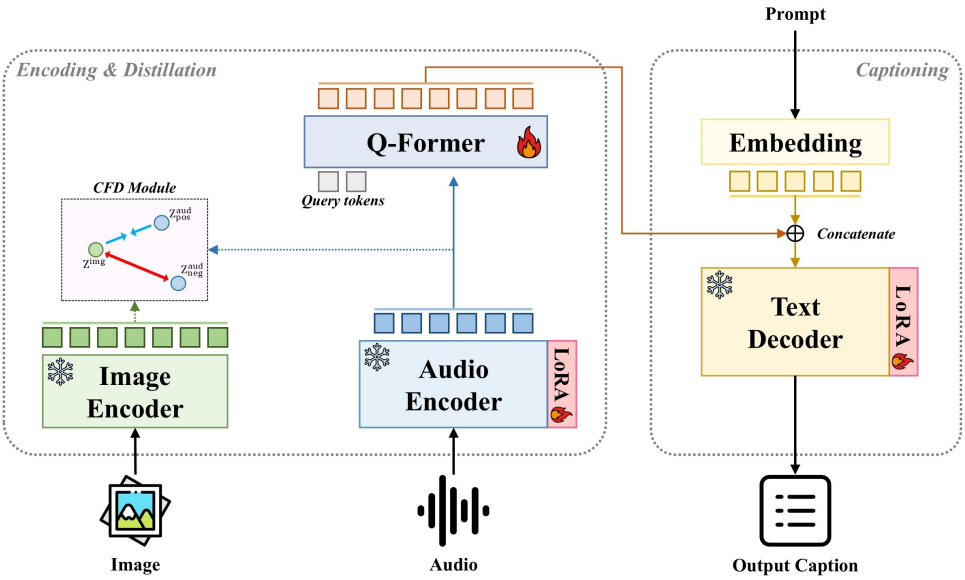


Figure 3.7: **Overview of Multi2Cap Pre-training Flow & Cross-modal Feature Distillation (CFD) Module**

3.4 Multi2Cap

In this section, we elaborate our proposed Multi2Cap. The operation of the Multi2Cap framework is illustrated in figure 3.7.

3.4.1 Creating Caption

Multi2Cap is structured similarly to previous mainstream methods that combine audio encoding and text decoding. The audio encoder and text decoder are composed of CED [119] and LLaMA [121], respectively, based on previous research, and each is fine-tuned using LoRA [118]. Q-Former [120] is constructed between the audio encoder and text decoder, and both the audio embedding and prompt are input together into the text decoder. The audio encoder, Q-Former, and text decoder are trained to optimize the cross-entropy loss $\mathcal{L}_{cap}$ as shown in the

equation below. $\mathcal{L}_{cap}$ is expressed as follows:

$$\mathcal{L}_{\text{cap}} = \texttt{CrossEntropy}(\text{origin}, \text{output}) \tag{3.1}$$

where, origin refers to the ground truth caption, and output refers to the caption generated by the model.

3.4.2 Cross-modal Feature Distillation

An image encoder is additionally constructed as the content-teacher of the audio encoder to learn image representations. For the image encoder, ResNet50 [122] and ViT-Large [123] are used, and the weights are fixed. In the proposed framework, the image encoder, as the content-teacher of the audio encoder, performs Cross-modal Feature Distillation (CFD) with the audio encoder. CFD is optimized through contrastive learning [145] between the image representation Z^{img} extracted from the image encoder and the audio representations Z^{aud}_{pos} and Z^{aud}_{neg} extracted from the audio encoder. The image feature (Z^{img}) is trained to be closer to Z^{aud}_{pos} in the feature space ($\in \mathbb{R}^{dim}$), while it is trained to be farther from Z^{aud}_{neg} in the feature space. The objective function of the CFD module, $\mathcal{L}_{cfd}$, is expressed as follows:

$$\mathcal{L}_{\text{cfd}} = \texttt{TripletMarginLoss}(\text{anc}, \text{pos}, \text{neg}) \tag{3.2}$$

where anc refers to Z^{img}, and pos and neg refer to Z^{aud}_{pos} and Z^{aud}_{neg}, respectively. Additionally, the margin is set to a default value of 1.0.

3.4.3 Objective of Multi2Cap

The ultimate goal of Multi2Cap is to effectively combine audio encoding, Cross-modal Feature Distillation(CFD), and text decoding to generate high-quality captions about audio content. In the pre-training phase, Multi2Cap optimizes the following objective function:

$$\mathcal{L}_{total} = \mathcal{L}_{cap} + \lambda \mathcal{L}_{cfd} \tag{3.3}$$

where, $\mathcal{L}_{cap}$ refers to the cross-entropy loss between the ground truth caption and the model-generated caption, ensuring that the generated captions are accurate and contextually appropriate. On the other hand, $\mathcal{L}_{cfd}$ ensures that the CFD module effectively learns the interaction between the image encoder and the audio encoder. where, λ is set to a default value of 0.5. In Section 3.5, we analyze the effect of λ on model performance and compare different types of loss functions for $\mathcal{L}_{cfd}$ in the CFD module.

3.5 Experiments

3.5.1 Experimental Setup

Datasets. the AAC model is evaluated using two datasets. First, Clotho [104] includes approximately 6,000 audio clips, each ranging from 15 to 30 seconds in length, and five captions are provided for each clip. AudioCaps [105] consists of approximately 50,000 audio clips, each 10 seconds long, with one caption provided in the training set and five captions in the validation and test sets. The primary evaluation dataset is the Clotho dataset, while the AudioCaps dataset is additionally included in the experiments to enhance reliability.

Performance metrics. In this study, we use various metrics, including BLEU(BL-1–4) [131], ROUGE-L(RG-L) [132], METEOR(ME) [133], CIDEr(CD) [134], SPICE(SP) [135], SPIDEr(SD) [136], SPIDEr-FL(SD-F) [137], Sentence-BERT(SB) [138], and FENSE(FS) [139] to evaluate model performance. BLEU, ROUGE-L, and METEOR assess lexical similarity based on N-grams, while CIDEr measures lexical similarity using TF-IDF weighting with reference sentences. SPICE evaluates semantic similarity by considering objects, relationships, and attributes. SPIDEr balances lexical and semantic evaluation by averaging CIDEr and SPICE, and SPIDEr-FL and FENSE further assess fluency and grammatical correctness. Sentence-BERT measures semantic similarity through cosine similarity between sentence embeddings, providing a comprehensive analysis of model performance.

Audio Contents Augmentation. in this study, four audio augmentation techniques are used to compensate for the insufficient data compared to WavCaps [103].

		BL-1	BL-2	BL-3	BL-4	RG-L	ME	CD	SP	SD	SD-F	SB	FS
w/o image	Human	46.1	22.9	**14.6**	**8.4**	29.6	13.5	15.4	7.5	11.4	11.4	48.6	48.6
	LOAE [88]	33.0	18.3	10.9	7.0	29.1	12.4	47.6	14.2	30.9	27.6	61.8	55.3
w/ image	Human	**49.2**	**27.0**	13.6	7.5	**32.9**	**14.8**	19.4	7.5	13.4	13.4	48.9	48.9
	Multi2Cap$_{\text{small}}$	33.8	18.9	11.5	7.5	29.5	13.7	52.1	15.4	33.7	30.8	62.6	57.7
	Multi2Cap$_{\text{base}}$	34.1	19.2	11.6	7.6	29.9	13.9	**55.0**	**15.5**	**35.2**	**33.7**	**63.5**	**59.7**

Table 3.2: **Performance comparisons on VggCaps**: This table shows the performance of Multi2Cap on the VggCaps dataset. Each column represents an evaluation metric, and the abbreviations for the metrics are mentioned in section 3.5.1. Multi2Cap is compared with LOAE [88], which utilized the same LLM. The performance shows superior results across all metrics. Best performance for each metric is in **Bold**, and the second-best is Underlined.

The techniques employed are *Adding white noise*, *Shifting*, *Stretching*, and *Flipping*. Each technique dynamically adjusts its application and order during training to maximize diversity. The operation of each technique is described in Appendix 3.7.2.

Implementation Details. in this study, we use two versions of our model: Multi2Cap$_{\text{small}}$ and Multi2Cap$_{\text{base}}$. Both model uses CED as the audio encoder, Llama2-7B as the text decoder. Then, ResNet is used as the Multi2Cap$_{\text{small}}$ and ViT-Large as the Multi2Cap$_{\text{base}}$ for the image encoder. The models were trained using the AdamW optimizer [140], with a learning rate of 5e-5 for the pre-training phase and 1e-4 for the fine-tuning phase. In pre-training, a batch size of 320 was used with 15 epochs and 2 warm-up epochs, while in fine-tuning, a batch size of 384 was used, and training was conducted for 30 epochs. Additional implementation details are provided in Appendix 3.7.1.

3.5.2 Overall Performance Comparison

Performance of VggCaps

The pre-training performance of the proposed method is compared with previous research [88] and the human performance conducted by ourselves. The results are presented in table 3.2. The most noticeable finding is that both humans and AI models show improved performance when the Image Modality is added. Additionally, Human evaluators performed well on relatively simple n-gram-based performance metrics, indicating that the VggCaps dataset is clearly and naturally structured for humans to generate captions. On the other hand, Multi2Cap demonstrated outstanding performance on more complex metrics based on semantic relationships, such as CIDEr (CD), SPIDEr-FL (SD-F), Sentence-BERT (SB), and FENSE (FS). This is the result of effectively learning the potential representations between audio and images through Cross-modal Feature Distillation (CFD), suggesting that the interaction between the two modalities played a crucial role in caption generation. In particular, the significant improvement in Multi2Cap's performance over previous research in CIDEr and SPIDEr (SD) demonstrates that the addition of image modality played an important role in enhancing the semantic performance of audio captioning. In conclusion, it can be confirmed that the VggCaps dataset is well-constructed for caption generation, and that the Multi2Cap framework has significantly contributed to improving semantic performance in audio captioning tasks.

	ME	CD	SP	SD	SD-F
ASR Whisper [124]	17.2	41.4	12.3	26.9	26.7
ConvNeXt [125]	19.3	48.6	14.2	31.4	31.4
BEATs [126]	19.5	50.5	**14.9**	32.7	32.7
LOAE [88]	19.7	<u>51.3</u>	14.7	<u>33.0</u>	<u>33.0</u>
EnCLAP++ [127]	**19.9**	48.0	<u>14.8</u>	31.4	31.4
Ours					
Multi2Cap$_{\text{small}}$	19.2	48.4	13.9	31.1	31.1
Multi2Cap$_{\text{base}}$	<u>19.7</u>	**53.0**	14.4	**33.7**	**33.6**

Table 3.3: **Performance Comparison on Clotho**: This table shows the comparison of the fine-tuning results of pre-trained Multi2Cap on Clotho (AAC benchmark dataset) with the performance of previous studies. It can be seen that Multi2Cap achieved state-of-the-art performance in most metrics. Best performance for each metric is in **Bold**, and the second-best is <u>Underlined</u>.

	ME	CD	SP	SD	SD-F
Human [105]	28.8	91.3	21.6	-	-
EnCLAP [128]	25.5	80.3	18.8	49.5	-
LOAE [88]	26.7	81.6	<u>19.3</u>	50.5	<u>50.4</u>
AutoCap [129]	25.3	<u>83.2</u>	18.2	50.7	-
EnCLAP++ [130]	**26.9**	82.3	**19.7**	**51.0**	-
Ours					
Multi2Cap$_{\text{small}}$	25.0	79.3	17.2	48.2	48.1
Multi2Cap$_{\text{base}}$	25.9	**83.7**	18.2	<u>50.9</u>	**50.9**

Table 3.4: **Performance Comparison on AudioCaps**: This table shows the comparison of the fine-tuning results of pre-trained Multi2Cap on AudioCaps (AAC benchmark dataset) with the performance of previous studies. It can be seen that Multi2Cap achieved state-of-the-art performance in most metrics. Best performance for each metric is in **Bold**, and the second-best is <u>Underlined</u>.

Performance of Benchmark

The proposed method is compared with previous top models on a per-dataset basis. The comparison results are presented in table 3.3, 3.4. In the case of the Clotho dataset (table 3.3), despite the absolute shortage of pre-training data compared to studies with similar settings [88], it shows superior performance across

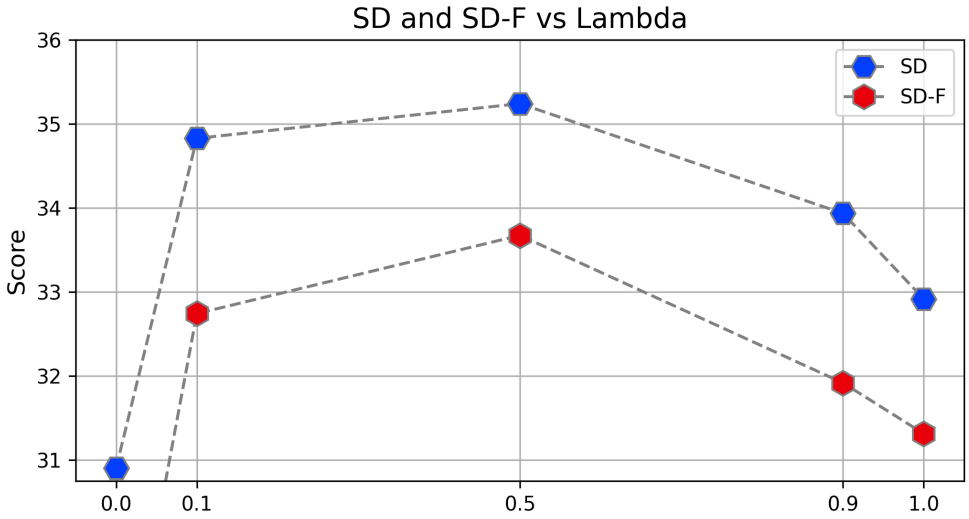


Figure 3.8: **Comparisons with Different λ in CFD**

most metrics. Additionally, for the AudioCaps dataset (table 3.4), it demonstrates performance that is either similar to or better than previous models. In particular, when compared to previous research [130], it shows improved performance in CIDEr(CD).

These results indicate that Multi2Cap effectively captures and utilizes the potential interactions between the audio and visual modalities that previous models could not. The impact of introducing a multi-modal perspective is particularly evident in specific metrics like CIDEr(CD), enabling the generation of more detailed and contextually rich audio captions. Through these results, we propose the potential for further development of multi-modal learning frameworks for audio captioning in the future.

	Smooth L1	MSE	InfoNCE	Cosine	Triplet
ME	12.6	13.8	13.7	13.8	**13.9**
CD	50.5	51.8	52.7	52.8	**55.0**
SP	14.3	15.3	15.2	15.3	**15.5**
SD	32.4	33.6	34.0	34.1	**35.2**
SD-F	30.2	31.6	31.9	31.8	**33.7**
SB	62.8	62.8	63.0	63.1	**63.5**
FS	58.0	58.6	58.8	58.9	**59.7**

Table 3.5: **Comparisons CFD by Objective Function Type**

3.5.3 Ablation Study

Comparisons $\mathcal{L}_{cfd}$ by λ difference

The pre-training performance according to λ in the objective function $\mathcal{L}_{cfd}$ is compared. The comparison results are presented in figure 3.8. The results are based on the SPIDEr(SD) and SPIDEr-FL(SD-F) performance of the model, with λ varying from 0 to 1 during training. Other fixed settings include $\mathcal{L}_{cfd}$ based on `TripletMarginLoss`$(\cdot)$, with experiments conducted on Multi2Cap$_{\text{base}}$. The baseline shows significant performance improvement with a relatively small weight value (0.1) at $\lambda = 0$. This demonstrates that the addition of another modality can benefit AAC tasks, as is the case in other tasks. However, when the weight is set to the maximum value (1.0), the learning effect is diluted, and the influence diminishes.

Comparisons $\mathcal{L}_{cfd}$ by objective function

The results based on the type of objective function for $\mathcal{L}_{cfd}$ are provided in table 3.5. With fixed settings, $\lambda = 0.5$ in the Multi2Cap$_{\text{base}}$ model. In the case of

Smooth L1 [141] and MSE [142], which are regression-based loss functions, they are very simple functions aimed at minimizing the difference between features. As a result, they show relatively low performance improvement. On the other hand, in the case of contrastive learning-based methods like InfoNCE [143], Cosine Similarity [144], and Triplet [145], they seem to capture potential interactions more effectively, as they can extract additional information from the comparison target.

3.6 Conclusion and Future Work

In this study, we propose a new framework called Multi2Cap to overcome the limitations of Automated Audio Captioning(AAC) tasks that relied on a single modality. Multi2Cap applies the Cross-modal Feature Distillation(CFD) technique to effectively capture the potential interactions between audio and visual information, significantly improving the richness and complexity of the audio captions. The experimental results confirm that Multi2Cap, which adopts a multi-modal approach, outperforms traditional single-modality-based AAC models. Additionally, human evaluations and various experiments have demonstrated that the inclusion of visual information makes descriptions of audio content more accurate and clearer. This suggests that the multi-modal approach is a key factor in improving performance in AAC tasks. Overall, Multi2Cap experimentally proves that expanding multi-modal learning can lead to richer and more accurate results in AAC tasks.

Moreover, to address the limitations of short and concise captions in existing datasets, we constructed a new dataset called VggCaps. This dataset leverages an LLM as a guide to supplement audio information with visual information, providing longer and more diverse descriptions with contextually rich captions. Experiments show that VggCaps surpasses existing datasets in terms of lexical diversity and complexity of expression, and its validity and robustness were confirmed through human evaluations. Overall, VggCaps demonstrates the potential to pose higher-level challenges in AAC tasks.

In conclusion, Multi2Cap presents a new paradigm in AAC research by effec-

tively utilizing multi-modal learning and emphasizes the importance of a multi-modal approach in future AAC tasks. This study demonstrates the potential to build more sophisticated and rich AI systems by effectively utilizing multi-modal data beyond a single modality, which will serve as an important milestone in the future development of multi-modal AI research.

3.7 Appendix

3.7.1 Additional Details

Pre-training Implementation Details

For reproducibility, the implementation details used in pre-training are presented in table 3.6. Based on Multi2Cap$_{\text{base}}$, the AdamW optimizer is used, with the base learning rate set to 5×10^{-5} and the weight decay set to 1×10^{-6} to prevent overfitting. The batch size is 320, and training is conducted for a total of 15 epochs, with the first 2 epochs set as a warm-up phase to stabilize initial training. The β parameters of the Adam optimizer are set to $(0.9, 0.999)$. The sampling rate of the audio input is fixed at 16,000Hz, and four audio augmentation techniques—*AddWhiteNoise*, *Shifting*, *Stretching*, and *Flipping*—are applied. The visual information is processed using ResNet50 and ViT-Large image encoders, with the image resolution set to 224×224 pixels. Additionally, the *RandomResizedCrop* technique is used for image augmentation.

Fine-tuning Implementation Details

The implementation details for the fine-tuning phase of the Multi2Cap model on benchmark datasets are presented in table 3.7. Based on Multi2Cap$_{\text{base}}$, the AdamW optimizer is used. The base learning rate is set to 1×10^{-4}, and the weight decay is set to 1×10^{-6}. The β parameters of the Adam optimizer are specified as $(0.9, 0.999)$. The batch size is 384, and training is conducted for a total of 30 epochs. Of these, the first 2 epochs are used as a warm-up phase to stabilize the

Hyper-parameters	Value
Optimizer	AdamW
Base learning rate	5×10^{-5}
Weight decay	1×10^{-6}
Adam β	(0.9, 0.999)
Batch size	320
Training epochs	15
Warmup epochs	2
Audio sample rate	16000
Audio Augmentation	AddWhiteNoise
	Shifting
	Stretching
	Flipping
Image Teacher	ResNet50, ViT-Large
Image resolution	224×224
Image augmentation	RandomResizedCrop

Table 3.6: Default Pre-training Setting

model during the initial stages of training. The audio input is processed at a sampling rate of 16,000Hz, and four audio augmentation techniques—*AddWhiteNoise*, *Shifting*, *Stretching*, and *Flipping*—are applied.

3.7.2 Additional Experiments

Overall Comparisons by λ Difference

The results of the pre-training performance for each λ of $\mathcal{L}_{cfd}$, as presented in figure 3.8, are summarized based on all metrics in table 3.8. When $\lambda = 0.5$, the best performance was observed across all metrics. Specifically, a score of 55.0 in CIDEr(CD), 15.5 in SPICE(SP), and 35.2 in SPIDEr(SD) was recorded, which is significantly higher compared to other λ values. Additionally, the highest perfor-

Hyper-parameters	Value
Optimizer	AdamW
Base learning rate	1×10^{-4}
Weight decay	1×10^{-6}
Adam β	(0.9, 0.999)
Batch size	384
Training epochs	30
Warmup epochs	2
Audio sample rate	16,000
Audio Augmentation	AddWhiteNoise
	Shifting
	Stretching
	Flipping

Table 3.7: Default fine-tuning setting

λ	ME	CD	SP	SD	SD-F	SB	FS
0	12.4	47.6	14.2	30.9	27.6	61.8	55.3
0.1	13.7	54.4	15.3	34.8	32.7	63.1	59.1
0.5	**13.9**	**55.0**	**15.5**	**35.2**	**33.7**	**63.5**	**59.7**
0.9	13.8	52.4	15.4	33.9	31.9	63.0	59.1
1	13.7	50.6	15.2	32.9	31.3	63.0	59.9

Table 3.8: Performance comparison by lambda (all metrics)

mance was also achieved in the SD-F and Sentence-BERT(SB) metrics with scores of 33.7 and 63.5, respectively, while the FENSE(FS) also showed excellent results with a score of 59.7. Therefore, it can be confirmed that $\lambda = 0.5$ is the most suitable value for deriving the optimal performance of the Multi2Cap model.

Comparison About Audio Content Augmentation

Additionally, the results based on whether audio content augmentation was applied are examined. The results are presented in table 3.9. The augmentation tech-

	w/o augment	w/ augment
B-1	**34.1**	**34.1**
B-2	19.1	**19.2**
B-3	11.5	**11.6**
B-4	7.5	**7.6**
RG-L	29.6	**29.9**
ME	12.7	**13.9**
CD	51.8	**55.0**
SP	14.2	**15.5**
SD	33.0	**35.2**
SD-F	30.9	**33.7**
SB	62.6	**63.5**
FS	58.2	**59.7**

Table 3.9: Comparison about Audio Content Augmentation

niques used for pre-training in Multi2Cap are *AddWhiteNoise*, *Shifting*, *Stretching*, and *Flipping*.

Adding white noise is the simplest method, which involves adding random noise to the audio signal. This technique helps improve the model's generalization performance in environments with various noise by applying slight noise to the input data.

Shifting shifts the start point of the audio by a certain amount of time forward or backward. This contributes to increasing the model's robustness to temporal shifts in the data.

Stretching changes the playback speed of the audio while maintaining the pitch. This provides the model with generalization capabilities to handle audio data at different speeds.

Flipping inverts the phase of the audio waveform. While this results in no perceptible change to the human ear, it alters the mathematical structure of the signal,

providing additional data diversity.

These four augmentation techniques are dynamically set in terms of whether to apply them during training and in what order, to maximize diversity during learning. This can be expressed in the following formula. Although augmentation generally contributes positively to performance improvement, it does not show consistent performance gains across all evaluation metrics. Specifically, in BLEU scores, the impact of augmentation on performance is minimal or nearly nonexistent, whereas clear performance improvements can be observed in metrics such as CIDEr(CD) and SPICE(SP). This suggests that audio content augmentation techniques do not significantly affect simple n-gram-based performance but have a positive effect on semantic consistency and the generation of sophisticated captions.

3.7.3 VggCaps

Prompt Templates

In figure 3.1, the input-prompt and post-prompt are introduced. The input-prompt is provided in listing 3.1. The input-prompt requests the creation of an initial caption using both audio and images. Additionally, rules are applied to set conditions for generating the raw-caption. Furthermore, the post-prompt is provided in listing 3.2. The post-prompt modifies the raw-caption into more general and descriptive sentences. Similarly, rules are applied to impose generation conditions, and examples are provided to ensure that refined captions are generated.

```
The two images are a video snapshot and an audio
   spectrum, categorized into {}. Write a caption that
    includes the visual and auditory elements that
   these images suggest :

[Rules]
1. Focus on audio-visual elements
2. Write with a rich vocabulary
3. Exclude anything other than captions

>>> Caption :
```

Listing 3.1: Input Prompt Template for VggCaps

Examples

Table 3.10 shows samples from the constructed VggCaps.

```
Revise the sentence below to a more general and
   descriptive sentence, and one that is appropriate
   for captioning images and audio:

[Rules]
1. Focus on the background sound element
2. Exclude hashtags (like "#peoplemarching")
3. Avoid captions that read like film titles.(like "A
   Joyful Heart's Muffled Cough-terlude ~~~")
4. Write with a rich vocabulary
5. Exclude anything other than captions

[example]
1. A sea of determined faces marches in unison, their
   boots thundering in harmony across the ground,
   echoing the orchestrated rhythm of military
   discipline and resolve.
2. Amidst the whispering reeds, the pheasant's crowing
    breaks the silence of the meadow, a resonant call
   that echoes the wild heart of nature.
3. Explosive bursts light up the night, resonating
   with the jubilant cracking and booming of
   celebratory fireworks.


>>> Caption :
```

Listing 3.2: Post Prompt Template for VggCaps

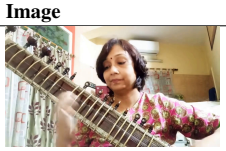
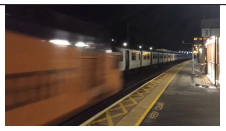
ID	Image	Caption	Category
1v5mmZoJJ50		Her fingers dance gracefully on the sitar strings, weaving a tapestry of sound that resonates with the serene and soulful essence of the music.	playing sitar
5IuRzJRrRpQ		The joyful bleats of sheep mingle with the soft rustling of grass and the occasional bark of an energetic dog, bringing a lively atmosphere to the green pastures.	sheep bleating
0fTwdhslb6E		The thunderous crack of the ball against the wall reverberates as two players immerse themselves in the rhythm of their squash game.	playing squash
1tPjBLXRHqM		The lively hum of the festival is accompanied by the drummer's rhythmic beats, their sticks creating a pulsating rhythm that resonates through the crowd.	playing drum kit
2zJiY9Mqhtc		A canopy alive with song as the harmonious tweets and chirps of birds enliven the surroundings, weaving a vibrant tapestry of nature's own symphony.	bird chirping, tweeting
3ymE2QOPRCA		The invigorating sounds of a volleyball match fill the air, blending cheers with the sharp slap of the ball.	playing volleyball
1mpFmBJ3nv0		The thunderous crescendo of a train horn slices through the stillness of the night, a wild call that reverberates along the tracks.	train horning
1t3sNHA0Vd4		In the cacophony of urban sounds, the police car's siren pierces the night air, signaling urgency and command.	police car (siren)
P0Mzdxr6F58I		In the stillness of the night, a lone frog's persistent ribbit pierces the quiet, adding a rhythm to the tranquil scene.	cattle mooing

Table 3.10: Examples of VggCaps

4. Chapter 4: Conclusion

In this thesis, we explored the effectiveness of Knowledge Distillation (KD) in two significant areas: domain-specific applications, particularly in the medical field, and multi-modal tasks such as Automated Audio Captioning (AAC). By investigating these two distinct use cases, we demonstrated the versatility and robustness of KD as a deep learning technique for transferring specialized knowledge from teacher models to student models, both within specific domains and across different modalities.

The first part of this study focused on Domain-Specific to General Knowledge Distillation (DSG-KD), particularly in the context of natural language processing (NLP) using Electronic Medical Record (EMR) data from Pediatric Emergency Departments (PEDs). Through our experiments, we demonstrated that domain-specific pre-trained models often struggle with N-lingual and free-text data complexities. However, by using KD to transfer domain-specific knowledge from specialized models to generalized models, we significantly improved classification performance. This demonstrates the potential for KD to enhance real-world tasks, such as medical decision-making, by allowing general-purpose models to leverage domain-specific expertise.

In the second part of the thesis, we expanded the scope of KD to multi-modal applications by proposing the Multi2Cap framework. This framework extended the

conventional Automated Audio Captioning task by introducing cross-modal feature distillation, leveraging both visual and auditory data to generate richer and more detailed audio captions. Our proposed Multi2Cap approach achieved state-of-the-art results, demonstrating that KD can effectively capture the interaction between different modalities, leading to enhanced performance in multi-modal tasks. The findings of this thesis provide valuable insights into how KD can be applied flexibly across different domains and applications. In the domain-specific context, KD enables more efficient and accurate model performance in specialized fields like healthcare. In multi-modal applications, it opens up new possibilities for integrating diverse data types to improve task performance.

Looking forward, several promising directions for future research arise from this work. First, extending the DSG-KD methodology to more diverse non-English-speaking regions and incorporating a wider range of downstream tasks will further validate its generalizability. Additionally, exploring new model architectures and more advanced KD techniques could lead to even more robust domain-specific and multi-modal learning frameworks. As multi-modal research continues to grow, the integration of KD techniques into these frameworks will remain a critical area of investigation, driving innovation in artificial intelligence across diverse fields.

This thesis has shown that effective knowledge transfer through KD, whether within specialized domains or across multiple modalities, can significantly enhance model capabilities, offering new pathways for both academic research and practical applications.

REFERENCES

[1] Lundberg, S. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

[2] Lambert, V., Matthews, A., MacDonell, R., & Fitzsimons, J. (2017). Paediatric early warning systems for detecting and responding to clinical deterioration in children: a systematic review. BMJ open, 7(3), e014497.

[3] Harrison, C. J., & Sidey-Gibbons, C. J. (2021). Machine learning in medicine: a practical introduction to natural language processing. BMC medical research methodology, 21(1), 158.

[4] Zhang, Y., Cai, T., Yu, S., Cho, K., Hong, C., Sun, J., ... & Liao, K. P. (2019). High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nature protocols, 14(12), 3426-3444.

[5] Castro, V. M., Minnier, J., Murphy, S. N., Kohane, I., Churchill, S. E., Gainer, V., ... & International Cohort Collection for Bipolar Disorder Consortium. (2015). Validation of electronic health record phenotyping of bipolar disorder cases and controls. American Journal of Psychiatry, 172(4), 363-372.

[6] Gianfrancesco, M. A., & Goldstein, N. D. (2021). A narrative review on the validity of electronic health record-based research in epidemiology. BMC medical research methodology, 21(1), 234.

[7] Barbazza, E., Allin, S., Byrnes, M., Foebel, A. D., Khan, T., Sidhom, P., ... & Kringos, D. S. (2021). The current and potential uses of Electronic Medical Record (EMR) data for primary health care performance measurement in the Canadian context: a qualitative analysis. BMC health services research, 21, 1-11.

[8] Park, Y. T., & Han, D. (2017). Current status of electronic medical record systems in hospitals and clinics in Korea. Healthcare informatics research, 23(3), 189-198.

[9] Shinozaki, A. (2020). Electronic medical records and machine learning in approaches to drug development. In Artificial intelligence in Oncology drug discovery and development. IntechOpen.

[10] Raza, S., & Schwartz, B. (2023). Constructing a disease database and using natural language processing to capture and standardize free text clinical information. Scientific Reports, 13(1), 8591.

[11] Crema, C., Attardi, G., Sartiano, D., & Redolfi, A. (2022). Natural language processing in clinical neuroscience and psychiatry: A review. Frontiers in Psychiatry, 13, 946387.

[12] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[13] Kim, K. (2020). Pretrained language models for Korean. GitHub. https://github.com/kiyoungkim1/LMkor

[14] Menachemi, N., & Collum, T. H. (2011). Benefits and drawbacks of electronic health record systems. Risk management and healthcare policy, 47-55.

[15] Hartswood, M., Procter, R., Rouncefield, M., & Slack, R. (2003). Making a case in medical work: implications for the electronic medical record. Computer Supported Cooperative Work (CSCW), 12, 241-266.

[16] Williams, F., & Boren, S. A. (2008). The role of the electronic medical record (EMR) in care delivery development in developing countries: a systematic review. Informatics in primary care, 16(2).

[17] Adnan, K., Akbar, R., Khor, S. W., & Ali, A. B. A. (2020). Role and challenges of unstructured big data in healthcare. Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019, Volume 1, 301-323.

[18] Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., & Godtliebsen, F. (2021). Challenges and opportunities beyond structured data in analysis of electronic health records. Wiley Interdisciplinary Reviews: Computational Statistics, 13(6), e1549.

[19] Lasko, T. A., Bhagwat, J. G., Zou, K. H., & Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. Journal of biomedical informatics, 38(5), 404-415.

[20] Flach, P., & Kull, M. (2015). Precision-recall-gain curves: PR analysis done right. Advances in neural information processing systems, 28.

[21] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

[22] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.

[23] Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019, June). Transfer learning in natural language processing. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials (pp. 15-18).

[24] Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural language processing of clinical notes on chronic diseases: systematic review. JMIR medical informatics, 7(2), e12239.

[25] Ben-Assuli, O., Shabtai, I., Leshno, M., & Hill, S. (2014). EHR in emergency rooms: exploring the effect of key information components on main complaints. Journal of medical systems, 38, 1-8.

[26] Hong, W. S., Haimovich, A. D., & Taylor, R. A. (2018). Predicting hospital admission at emergency department triage using machine learning. PloS one, 13(7), e0201016.

[27] Anderson, J. E., & Chang, D. C. (2015). Using electronic health records for surgical quality improvement in the era of big data. JAMA surgery, 150(1), 24-29.

[28] Kirubarajan, A., Taher, A., Khan, S., & Masood, S. (2020). Artificial intelligence in emergency medicine: a scoping review. Journal of the American College of Emergency Physicians Open, 1(6), 1691-1702.

[29] Zuccon, G., Wagholikar, A., Nguyen, A., Butt, L., Chu, K., Martin, S., & Greenslade, J. (2013). Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed ct ontology. AMIA Summits on Translational Science Proceedings, 300-304.

[30] Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., ... & Shah, N. H. (2014). Text mining for adverse drug events: the promise, challenges, and state of the art. Drug safety, 37, 777-790.

[31] Delahanty, R. J., Alvarez, J., Flynn, L. M., Sherwin, R. L., & Jones, S. S. (2019). Development and evaluation of a machine learning model for the early

identification of patients at risk for sepsis. Annals of emergency medicine, 73(4), 334-344.

[32] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), 11-21.

[33] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

[34] Weng, W. H., Wagholikar, K. B., McCray, A. T., Szolovits, P., & Chueh, H. C. (2017). Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC medical informatics and decision making, 17, 1-13.

[35] Spasic, I., & Nenadic, G. (2020). Clinical text data in machine learning: systematic review. JMIR medical informatics, 8(3), e17984.

[36] Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Mohi Ud Din, M. (2020). Machine learning based approaches for detecting COVID-19 using clinical text data. International Journal of Information Technology, 12, 731-739.

[37] Medrouk, L., & Pappa, A. (2017). Deep learning model for sentiment analysis in multi-lingual corpus. In Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part I 24 (pp. 205-212). Springer International Publishing.

[38] Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., ... & Xu, H. (2020). Deep learning in clinical natural language processing: a methodical review. Journal of the American Medical Informatics Association, 27(3), 457-470.

[39] Baker, C. (2011). Foundations of bilingual education and bilingualism. Multilingual matters.

[40] Aman, S., & Szpakowicz, S. (2007, September). Identifying expressions of emotion in text. In International Conference on Text, Speech and Dialogue (pp. 196-205). Berlin, Heidelberg: Springer Berlin Heidelberg.

[41] Park, J. E. (1990). Korean/English intrasentential code-switching: Matrix language assignment and linguistic constraints. University of Illinois at Urbana-Champaign.

[42] Miwa, N. I. S. H. I. M. U. R. A. (1985). Intrasentential codeswitching in Japanese and English. Ph. D dissertation. University of Pennsylvania.

[43] Ahn, J., La Ferle, C., & Lee, D. (2017). Language and advertising effectiveness: Code-switching in the Korean marketplace. International Journal of Advertising, 36(3), 477-495.

[44] Amazouz, D., Adda-Decker, M., & Lamel, L. (2017, August). Addressing code-switching in French/Algerian Arabic speech. In Interspeech 2017 (pp. 62-66).

[45] Segura-Bedmar, I., Colón-Ruíz, C., Tejedor-Alonso, M. Á., & Moro-Moro, M. (2018). Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. Journal of biomedical informatics, 87, 50-59.

[46] Rivas, R., Montazeri, N., Le, N. X., & Hristidis, V. (2018). Automatic classification of online doctor reviews: evaluation of text classifier algorithms. Journal of medical Internet research, 20(11), e11141.

[47] Alzoubi, H., Ramzan, N., Alzubi, R., & Mesbahi, E. (2018, August). An automated system for identifying alcohol use status from clinical text. In 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE) (pp. 41-46). IEEE.

[48] Reddy, B. K., & Delen, D. (2018). Predicting hospital readmission for lupus patients: an RNN-LSTM-based deep-learning methodology. Computers in biology and medicine, 101, 199-209.

[49] Fuchs, S., Terry, M., Adelgais, K., Bokholdt, M., Brice, J., Brown, K. M., ... & Marx, M. (2016). Definitions and assessment approaches for emergency medical services for children. Pediatrics, 138(6).

[50] Tohira, H., Finn, J., Ball, S., Brink, D., & Buzzacott, P. (2022). Machine learning and natural language processing to identify falls in electronic patient care records from ambulance attendances. Informatics for Health and Social Care, 47(4), 403-413.

[51] Chen, M., Shi, W., Zhou, B., & Roth, D. (2020). Cross-lingual entity alignment with incidental supervision. arXiv preprint arXiv:2005.00171.

[52] Jin, H., Li, C., Zhang, J., Hou, L., Li, J., & Zhang, P. (2019). XLORE2: large-scale cross-lingual knowledge graph construction and application. Data Intelligence, 1(1), 77-98.

[53] Lee, J. (2020). Kcbert: Korean comments bert. In Annual Conference on Human and Language Technology (pp. 437-440). Human and Language Technology.

[54] Pires, T. (2019). How multilingual is multilingual BERT. arXiv preprint arXiv:1906.01502.

[55] Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine, 4(1), 86.

[56] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

[57] Kim, Y., Kim, J. H., Lee, J. M., Jang, M. J., Yum, Y. J., Kim, S., ... & Song, S. (2022). A pre-trained BERT for Korean medical natural language processing. Scientific Reports, 12(1), 13847.

[58] Gao, W., Zheng, X., & Zhao, S. (2021, April). Named entity recognition method of Chinese EMR based on BERT-BiLSTM-CRF. In Journal of physics: conference series (Vol. 1848, No. 1, p. 012083). IOP Publishing.

[59] Yu, X., Hu, W., Lu, S., Sun, X., & Yuan, Z. (2019, August). BioBERT based named entity recognition in electronic medical record. In 2019 10th international conference on information technology in medicine and education (ITME) (pp. 49-52). IEEE.

[60] Bae, Y. S., Kim, K. H., Kim, H. K., Choi, S. W., Ko, T., Seo, H. H., ... & Jeon, H. (2021). Keyword extraction algorithm for classifying smoking status from unstructured bilingual electronic health records based on natural language processing. Applied Sciences, 11(19), 8812.

[61] Park, J., You, S. C., Jeong, E., Weng, C., Park, D., Roh, J., ... & Park, R. W. (2021). A framework (SOCRATex) for hierarchical annotation of unstructured electronic health records and integration into a standardized medical database: development and usability study. JMIR medical informatics, 9(3), e23983.

[62] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

[63] Hinton, G. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

[64] Sanh, V. (2019). DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv preprint arXiv:1910.01108.

[65] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., ... & Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351.

[66] Qin, D., Bu, J. J., Liu, Z., Shen, X., Zhou, S., Gu, J. J., ... & Dai, H. F. (2021). Efficient medical image segmentation based on knowledge distillation. IEEE Transactions on Medical Imaging, 40(12), 3820-3831.

[67] Zhao, L., Qian, X., Guo, Y., Song, J., Hou, J., & Gong, J. (2023). MSKD: Structured knowledge distillation for efficient medical image segmentation. Computers in Biology and Medicine, 164, 107284.

[68] Nateras, L. G., Dernoncourt, F., & Nguyen, T. (2023, July). Hybrid knowledge transfer for improved cross-lingual event detection via hierarchical sample selection. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5414-5427).

[69] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[70] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[71] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[72] Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[73] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964.

[74] Berglund, M., & van der Merwe, B. (2023). Formalizing BPE Tokenization. arXiv preprint arXiv:2309.08715.

[75] Drossos, K., Adavanne, S., & Virtanen, T. (2017, October). Automated audio captioning with recurrent neural networks. In 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (pp. 374-378). IEEE.

[76] Marian, V., Hayakawa, S., & Schroeder, S. R. (2021). Cross-modal interaction between auditory and visual input impacts memory retrieval. Frontiers in Neuroscience, 15, 661477.

[77] Blattner, M. M., & Glinert, E. P. (1996). Multimodal integration. IEEE multimedia, 3(4), 14-24.

[78] Li, L., Chen, G., Shi, H., Xiao, J., & Chen, L. (2024). A Survey on Multimodal Benchmarks: In the Era of Large AI Models. arXiv preprint arXiv:2409.18142.

[79] Munikoti, S., Stewart, I., Horawalavithana, S., Kvinge, H., Emerson, T., Thompson, S. E., & Pazdernik, K. (2024). Generalist Multimodal AI: A Review of Architectures, Challenges and Opportunities. arXiv preprint arXiv:2406.05496.

[80] Eldan, R., & Li, Y. (2023). Tinystories: How small can language models be and still speak coherent english?. arXiv preprint arXiv:2305.07759.

[81] Heittola, T., Mesaros, A., Virtanen, T., & Eronen, A. (2011). Sound event detection and context recognition. In CHiME 2011 Workshop on Machine Listening in Multisource Environments.

[82] Lee, Y., Yeon, I., Nam, J., & Chung, J. S. (2024, April). VoiceLDM: Text-to-Speech with Environmental Context. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 12566-12571). IEEE.

[83] Benesty, J., Chen, J., & Huang, Y. (2008). Automatic speech recognition: A deep learning approach.

[84] Narisetty, C., Tsunoo, E., Chang, X., Kashiwagi, Y., Hentschel, M., & Watanabe, S. (2022, May). Joint speech recognition and audio captioning. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7892-7896). IEEE.

[85] Mei, X., Liu, X., Plumbley, M. D., & Wang, W. (2022). Automated audio captioning: An overview of recent progress and new challenges. EURASIP journal on audio, speech, and music processing, 2022(1), 26.

[86] Xu, X., Xie, Z., Wu, M., & Yu, K. (2023). Beyond the status quo: A contemporary survey of advances and challenges in audio captioning. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[87] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[88] Liu, J., Li, G., Zhang, J., Dinkel, H., Wang, Y., Yan, Z., ... & Wang, B. (2024). Enhancing Automated Audio Captioning via Large Language Models with Optimized Audio Encoding. arXiv preprint arXiv:2406.13275.

[89] Niu, W., Wang, Y., Cai, G., & Hou, H. (2024). Efficient and Robust Knowledge Distillation from A Stronger Teacher Based on Correlation Matching. arXiv preprint arXiv:2410.06561.

[90] Gou, J., Xiong, X., Yu, B., Du, L., Zhan, Y., & Tao, D. (2023). Multi-target knowledge distillation via student self-reflection. International Journal of Computer Vision, 131(7), 1857-1874.

[91] Ienco, D., & Dantas, C. F. (2024). Discom-kd: Cross-modal knowledge distillation via disentanglement representation and adversarial learning. arXiv preprint arXiv:2408.07080.

[92] Huo, F., Xu, W., Guo, J., Wang, H., & Guo, S. (2024). C2KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16006-16015).

[93] Sang, Z., Funakoshi, K., & Okumura, M. (2024). Contrastive Knowledge Distillation for Robust Multimodal Sentiment Analysis. arXiv preprint arXiv:2410.08692.

[94] Wu, H., Xiao, L., Zhang, X., & Miao, Y. (2024). Aligning in a Compact Space: Contrastive Knowledge Distillation between Heterogeneous Architectures. arXiv preprint arXiv:2405.18524.

[95] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2023). A survey on multimodal large language models. arXiv preprint arXiv:2306.13549.

[96] Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., ... & Qiao, Y. (2023). Sphinx: The joint mixing of weights, tasks, and visual embeddings for multimodal large language models. arXiv preprint arXiv:2311.07575.

[97] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

[98] Fei, H., Yao, Y., Zhang, Z., Liu, F., Zhang, A., & Chua, T. S. (2024, May). From Multimodal LLM to Human-level AI: Modality, Instruction, Reasoning, Efficiency and Beyond. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries (pp. 1-8).

[99] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

[100] Shahriar, S., Lund, B. D., Mannuru, N. R., Arshad, M. A., Hayawi, K., Bevara, R. V. K., ... & Batool, L. (2024). Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. Applied Sciences, 14(17), 7782.

[101] Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J. B., ... & Mustafa, B. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.

[102] Chen, H., Xie, W., Vedaldi, A., & Zisserman, A. (2020, May). Vggsound: A large-scale audio-visual dataset. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 721-725). IEEE.

[103] Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., ... & Wang, W. (2024). Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[104] Drossos, K., Lipping, S., & Virtanen, T. (2020, May). Clotho: An audio captioning dataset. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 736-740). IEEE.

[105] Kim, C. D., Kim, B., Lee, H., & Kim, G. (2019, June). Audiocaps: Generating captions for audios in the wild. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:

Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 119-132).

[106] Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., & Zisserman, A. (2021). Localizing visual sounds the hard way. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16867-16876).

[107] Mo, S., & Morgado, P. (2023, July). A unified audio-visual learning framework for localization, separation, and recognition. In International Conference on Machine Learning (pp. 25006-25017). PMLR.

[108] Hannun, A. (2014). Deep Speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.

[109] Streijl, R. C., Winkler, S., & Hands, D. S. (2016). Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. Multimedia Systems, 22(2), 213-227.

[110] Flesch, R. (1948). A new readability yardstick. Journal of applied psychology, 32(3), 221.

[111] Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60(2), 283.

[112] Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. Educational research bulletin, 37-54.

[113] Spache, G. (1953). A new readability formula for primary-grade reading materials. The Elementary School Journal, 53(7), 410-413.

[114] Templin, M. C. (1957). Certain language skills in children; their development and interrelationships.

[115] Herdan, G. (1964). Quantitative linguistics or generative grammar?.

[116] Yule, C. U. (2014). The statistical study of literary vocabulary. Cambridge University Press.

[117] Simpson, E. H. (1949). Measurement of Diversity. Nature, 163.

[118] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[119] Dinkel, Heinrich, et al. ”CED: Consistent ensemble distillation for audio tagging.” ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.

[120] Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning (pp. 19730-19742). PMLR.

[121] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[122] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[123] Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[124] Kadlčík, M., Hájek, A., Kieslich, J., & Winiecki, R. (2023). A whisper transformer for audio captioning trained with synthetic captions and transfer learning. arXiv preprint arXiv:2305.09690.

[125] Labbé, E., Pellegrini, T., & Pinquier, J. (2023, May). IRIT-UPS DCASE 2023 audio captioning and retrieval system. In Proc. Conf. Detection Classification Acoust. Scenes Events Challenge (pp. 1-5).

[126] Wu, S. L., Chang, X., Wichern, G., Jung, J. W., Germain, F., Le Roux, J., & Watanabe, S. (2024, April). Improving audio captioning models with fine-grained audio features, text embedding supervision, and llm mix-up augmentation. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 316-320). IEEE.

[127] Kim, J., Jung, J., Jeon, M., Woo, S. H., & Lee, J. EXPANDING ON ENCLAP WITH AUXILIARY RETRIEVAL MODEL FOR AUTOMATED AUDIO CAPTIONING.

[128] Kim, J., Jung, J., Lee, J., & Woo, S. H. (2024, April). Enclap: Combining neural audio codec and audio-text joint embedding for automated audio captioning. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6735-6739). IEEE.

[129] Haji-Ali, M., Menapace, W., Siarohin, A., Balakrishnan, G., Tulyakov, S., & Ordonez, V. (2024). Taming Data and Transformers for Audio Generation. arXiv preprint arXiv:2406.19388.

[130] Kim, J., Jeon, M., Jung, J., Woo, S. H., & Lee, J. (2024). EnCLAP++: Analyzing the EnCLAP Framework for Optimizing Automated Audio Captioning Performance. arXiv preprint arXiv:2409.01201.

[131] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

[132] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

[133] Denkowski, M., & Lavie, A. (2014, June). Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the ninth workshop on statistical machine translation (pp. 376-380).

[134] Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).

[135] Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14 (pp. 382-398). Springer International Publishing.

[136] Liu, S., Zhu, Z., Ye, N., Guadarrama, S., & Murphy, K. (2017). Improved image captioning via policy gradient optimization of spider. In Proceedings of the IEEE international conference on computer vision (pp. 873-881).

[137] Labbe, E., Pellegrini, T., & Pinquier, J. (2022). SPIDEr-FL: An extension of SPIDEr for evaluating fluency and linguistic diversity. Retrieved from https://hal.archives-ouvertes.fr/hal-03810396

[138] Reimers, N. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084.

[139] Zhou, Z., Zhang, Z., Xu, X., Xie, Z., Wu, M., & Zhu, K. Q. (2022, May). Can audio captions be evaluated with image caption metrics?. In ICASSP 2022-

2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 981-985). IEEE.

[140] Loshchilov, I. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

[141] Huber, P. J. (1992). Robust estimation of a location parameter. In Breakthroughs in statistics: Methodology and distribution (pp. 492-518). New York, NY: Springer New York.

[142] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

[143] Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

[144] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[145] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).

[146] Torch Contributors. (2017). Torchvision.transforms.RandomResizedCrop. Retrieved from https://pytorch.org/vision/stable/transforms.html

국 문 초 록

효과적인 지식 증류에 관한 연구: 특화 도메인에서 다중 모달 응용까지

조 상 연
AI학과 AI응용 전공
중앙대학교 대학원

Knowledge Distillation(KD)는 고성능의 복잡한 모델(교사모델)에서 비교적 간단하고 덜 복잡한 모델(학생 모델)로 지식을 전이하는 심층학습 기법이다. 본 논문은 KD의 효과를 두 가지 맥락, 즉 도메인 특화 작업(특히 의료 분야)와 다중 모달 응용에서 구체적으로 탐구한다. 첫번째 장에서는 전자의료기록 데이터의 분류 성능을 개선하기 위해 자연어 처리 분야에서 의료 도메인의 지식을 증류하는 연구를 소개한다. 두 번째 장에서는 다중 모달 분야에서 모달리티간의 상호작용을 증류하는 연구를 다루며, 시각-청각 특징 증류를 통해 자동 오디오 캡셔닝 작업의 성능 개선을 위한 연구를 소개한다. 본 논문은 도메인 특화 및 다중 멀티모달 지식 증류 접근법에서 얻은 연구 결과를 통합하여, 다양한 도메인 및 응용 분야에서 지식 증류 기술의 유연성과 영향을 탐구한다.

핵심어: *자연어처리, 멀티모달 학습, 지식 증류, 전자의료기록, 자동 오디오 캡션*