JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 10, pp. 29-41, October 2022
https://doi.org/10.9708/jksci.2022.27.10.029

Text Classification Using Heterogeneous Knowledge Distillation

Yerin Yu*, Namgyu Kim*

*Graduate Student, Graduate School of Business IT, Kookmin University, Seoul, Korea
*Professor, Graduate School of Business IT, Kookmin University, Seoul, Korea

[Abstract]

Recently, with the development of deep learning technology, a variety of huge models with excellent performance have been devised by pre-training massive amounts of text data. However, in order for such a model to be applied to real-life services, the inference speed must be fast and the amount of computation must be low, so the technology for model compression is attracting attention. Knowledge distillation, a representative model compression, is attracting attention as it can be used in a variety of ways as a method of transferring the knowledge already learned by the teacher model to a relatively small-sized student model. However, knowledge distillation has a limitation in that it is difficult to solve problems with low similarity to previously learned data because only knowledge necessary for solving a given problem is learned in a teacher model and knowledge distillation to a student model is performed from the same point of view. Therefore, we propose a heterogeneous knowledge distillation method in which the teacher model learns a higher-level concept rather than the knowledge required for the task that the student model needs to solve, and the teacher model distills this knowledge to the student model. In addition, through classification experiments on about 18,000 documents, we confirmed that the heterogeneous knowledge distillation method showed superior performance in all aspects of learning efficiency and accuracy compared to the traditional knowledge distillation.

▸ **Key words**: Deep Learning, Knowledge Distillation, Text Classification, Model Compression

[요 약]

최근 딥 러닝 기술의 발전으로 방대한 텍스트 데이터를 사전에 학습한 우수한 성능의 거대한 모델들이 다양하게 고안되었다. 하지만 이러한 모델을 실제 서비스나 제품에 적용하기 위해서는 빠른 추론 속도와 적은 연산량이 요구되고 있으며, 이에 모델 경량화 기술에 대한 관심이 높아지고 있다. 대표적인 모델 경량화 기술인 지식증류는 교사 모델이 이미 학습한 지식을 상대적으로 작은 크기의 학생 모델에 전이시키는 방법으로 다방면에 활용 가능하여 주목받고 있지만, 당장 주어진 문제의 해결에 필요한 지식만을 배우고 동일한 관점에서만 반복적인 학습이 이루어지기 때문에 기존에 접해본 문제와 유사성이 낮은 문제에 대해서는 해결이 어렵다는 한계를 갖는다. 이에 본 연구에서는 궁극적으로 해결하고자 하는 과업에 필요한 지식이 아닌, 보다 상위 개념의 지식을 학습한 교사 모델을 통해 지식을 증류하는 이질적 지식증류 방법을 제안한다. 또한, 사이킷런 라이브러리에 내장된 20 Newsgroups의 약 18,000개 문서에 대한 분류 실험을 통해, 제안 방법론에 따른 이질적 지식증류가 기존의 일반적인 지식증류에 비해 학습 효율성과 정확도의 모든 측면에서 우수한 성능을 보임을 확인하였다.

▸ **주제어**: 딥 러닝, 지식증류, 텍스트 분류, 모델 경량화

• First Author: Yerin Yu, Corresponding Author: Namgyu Kim
*Yerin Yu (yerin1997@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
*Namgyu Kim (ngkim@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
• Received: 2022. 09. 22, Revised: 2022. 10. 24, Accepted: 2022. 10. 24.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 강력한 GPU 등 하드웨어의 발전으로 인해 복잡한 행렬 연산에 소요되는 시간이 단축되면서 딥 러닝(Deep Learning) 분야의 연구가 활발히 수행되고 있다. 딥 러닝은 인간의 뇌가 작동하는 방식과 유사한 알고리즘을 사용하는 머신 러닝(Machine Learning)의 하위 분야로서, 여러 층(Layer)을 쌓아 만든 신경망(Neural Network) 모델을 근간으로 동작한다. 딥 러닝의 발전에 힘입어 자연어 처리 분야에서 딥 러닝 기술의 활용도 크게 증가하고 있다. 딥 러닝을 활용한 자연어 처리의 주요 분야로는 문서 분류, 문서 생성, 문서 요약, 질의응답, 그리고 기계 번역 등이 있다. 이러한 응용들은 대량의 텍스트에 대한 딥 러닝 학습을 통해 모델을 생성하고, 생성된 모델을 적용하여 추론을 수행하는 방식으로 작동한다.

대량의 텍스트 데이터를 사용하는 이러한 딥 러닝 모델의 학습과 추론 과정에서는 고비용의 컴퓨팅 자원(Computing Resource)과 연산량이 요구된다. 일반적으로 딥 러닝 모델의 학습은 개발하는 모델이 목표 정확도에 도달하도록 조정을 가하며 지속적으로 갱신을 반복하는 방식으로 이루어진다. 자연어 처리를 위한 딥 러닝 모델 학습도 이와 마찬가지 방식으로 이루어지는데, 특히 자연어 처리 분야에서는 대량의 데이터에 대한 학습을 매번 진행하기 어렵다는 한계가 있다. 따라서 최근의 딥 러닝 기반 자연어 처리는 대용량의 말뭉치를 미리 학습한 사전 학습 언어 모델(Pre-trained Language Model)을 배포하고 이를 기반으로 적은 양의 데이터에 대한 추가 학습을 진행하는 방식으로 이루어진다. 최근 사전 학습 언어 모델의 정확도를 향상시키기 위한 많은 연구들이 경쟁적으로 수행되고 있으며, 일반적으로 이러한 모델의 개발에는 막대한 컴퓨팅 자원, 비용, 그리고 기간이 소요된다.

물론 모델의 정확도가 매우 중요한 평가 기준이기는 하지만, 아주 미미한 정도의 성능 향상을 위해 수십만 혹은 그 이상의 파라미터가 추가로 필요하다면 과연 이 모델이 효율성 측면에서 좋은 모델인지 고민해 볼 필요가 있다. 특히 로봇, 자율 주행 자동차, 그리고 챗봇 등 딥 러닝 기술이 적용된 실생활 서비스의 경우 정확도뿐 아니라 메모리, 전력, 통신량 등 자원이 제한된 환경 하에서 컴퓨터 비전, 음성, 자연어 등을 실시간으로 처리하여 신속히 답을 제시할 수 있는 능력이 요구되고 있다. 즉, 딥 러닝 모델을 다양한 분야에 적용하기 위해서는 모델의 정확성뿐 아니라 추론 시 요구되는 자원 및 추론 속도를 함께 고려하여 모델의 개발이 이루어져야 한다. 이러한 요구에 따라 최근에는 적은 자원을 사용하면서도 고속으로 우수한 정확도의 추론 결과를 얻기 위해 딥 러닝 모델의 경량화(Model Compression) 기술에 대한 필요성이 강조되고 있다.

모델 경량화 기술이란 모델의 정확도(Accuracy)는 기존 모델과 유사하게 유지하면서 모델의 크기와 연산량을 줄임으로써, 메모리와 에너지 측면에서 학습과 추론의 효율성을 높이는 기술이다. 모델 경량화는 대표적으로 가지치기(Pruning), 양자화(Quantization), 지식증류(Knowledge Distillation, KD) 등을 통해 구현되고 있다. 이 가운데 특히 지식증류는 작은 크기의 학생(Student) 모델이 큰 크기의 교사(Teacher) 모델에 비해 파라미터 수, 연산량, 크기는 작으면서도 매우 빠른 속도로 교사 모델과 유사한 성능을 가질 수 있으며, 도메인과 모델 구조의 제한 없이 다양한 분야에 적용이 가능하다는 점에서 가장 효율적이고 실용적인 경량화 방식으로 주목받고 있다.

지식증류는 잘 학습된 큰 모델을 교사 모델로 두고 추론 시 사용할 작은 크기의 모델을 학생 모델로 두며, 학생 모델은 교사 모델이 이미 학습한 지식을 따라가며 학습하는 대표적인 모델 경량화 방식이다. 전통적인 지식증류의 경우 일반적으로 교사와 학생은 서로 동일한 과업(Task)에 대해 학습을 수행하며, 모델의 규모나 학습 데이터의 양 측면에서만 교사와 학생이 서로 차이를 갖는다. 예를 들어 뉴스 카테고리 분류에 사용할 학생 모델을 개발하기 위해서는 방대한 양의 뉴스 데이터에 대해 카테고리 분류 학습을 마친 교사 모델이 필요하며, 객체 탐지(Object Detection)에 사용할 학생 모델을 개발하기 위해서는 객체 탐지를 학습한 교사 모델이 필요하다(Fig. 1).

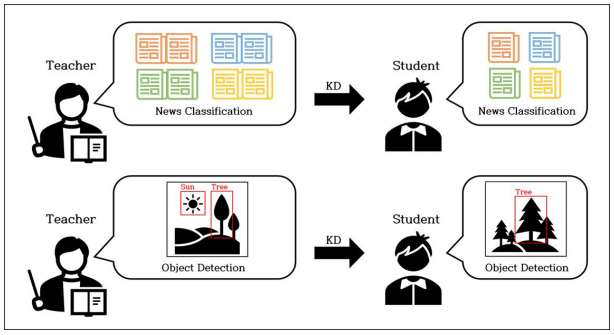

Fig. 1. Two Examples of Traditional Knowledge Distillation

이러한 전통적인 지식증류의 경우 학생은 본인이 해결해야 할 문제와 유사한 문제를 매우 많이 접해 본 교사로부터 지식을 전이받기 때문에, 교사가 학습한 문제와 유사한 문제가 주어졌을 때 해당 문제를 효율적으로 해결할 수 있다. 하지만, 당장 주어진 문제의 해결에 필요한 지식만

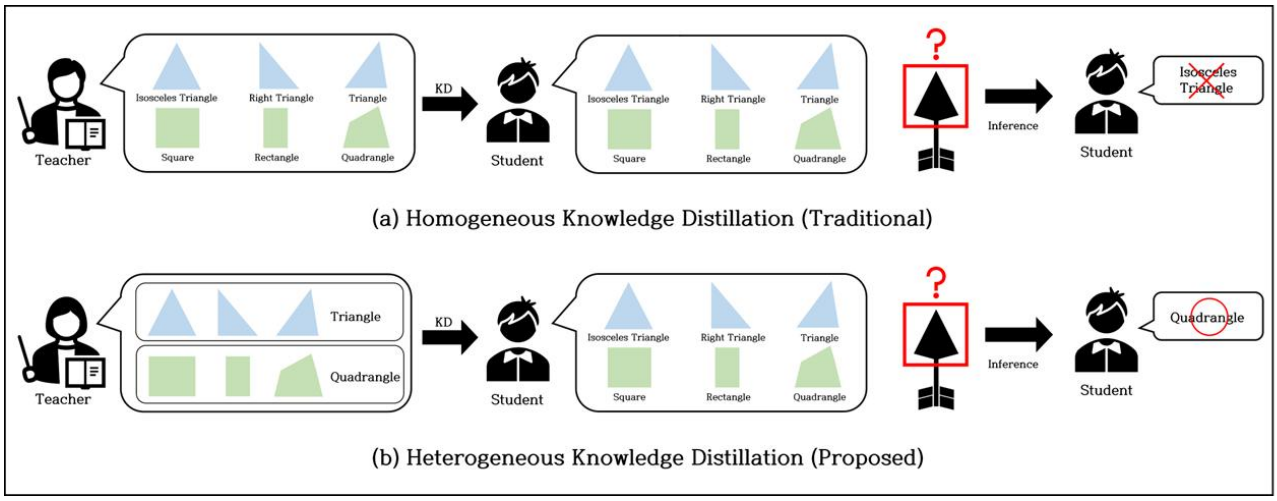

Fig. 2. Concept of the Proposed Heterogeneous Knowledge Distillation Approach

을 배웠고 동일한 관점에서만 반복해서 학습이 이루어지기 때문에, 기존에 접해본 문제와 유사성이 낮은 문제에 대해서는 해결이 어렵다는 한계를 갖는다.

이에 본 연구에서는 교사가 지식을 관통하는 개념, 즉 일반화된 원리를 발견할 수 있도록 가르치면 학생은 수많은 사실적 지식을 배울 수 있다는 교육학적 관점[1]을 반영한 지식증류 방법론을 제안한다. 즉 교사는 학생에게 당장 주어진 문제의 해결만 필요한 지식이 아닌 일반화된 개념을 가르치고, 학생은 교사로부터 전이받은 지식을 토대로 문제 해결 학습을 수행하는 방안을 제시한다(Fig. 2).

<Fig. 2>에서 학생에게 주어진 과업은 도형을 '이등변삼각형', '직각삼각형', '일반삼각형', '정사각형', '직사각형', 그리고 '일반사각형'의 6가지 중 하나로 분류하는 것이다. <Fig. 2(a)>는 전통적인 지식증류, 즉 교사와 학생이 동일한 과업을 수행하는 경우를 나타낸다. 이 경우 교사는 도형을 6가지 중 하나로 분류하는 방대한 학습을 수행하고, 해당 지식을 학생에게 전이한다. 또한 학생은 전이받은 지식을 바탕으로 동일한 과업을 추가로 수행한다. 이러한 학습을 마친 학생은 도형을 6가지로 분류하는 지식은 갖고 있지만 해당 분류의 상위 개념인 '삼각형'과 '사각형'에 대한 개념은 전혀 갖고 있지 않다. 따라서 그림에서와 같은 화살촉 모양이 추론의 입력으로 주어졌을 때, 해당 입력은 '일반사각형'으로 분류되어야 함에도 이를 '이등변삼각형'으로 잘못 분류할 가능성이 매우 높다.

한편 <Fig. 2(b)>는 본 연구에서 제안하는 지식증류 방법론, 즉 교사와 학생이 서로 다른 과업을 통해 학습을 수행하는 이질적 지식증류의 예를 보인다. 구체적으로 교사는 주어진 도형을 '삼각형'과 '사각형'으로만 분류하는 과업을 수행하여, 입력으로부터 '삼각형'과 '사각형'을 구분하기 위한 특징(Features)을 추출하고 학습한다. 다음으로 교사는 이러한 지식을 학생에게 전이하고, 학생은 '삼각형'과 '사각형'에 대한 기본 개념을 바탕으로 세부 분류를 수행하는 과업을 통해 학습을 수행한다. 이러한 학습을 마친 학생에게 화살촉 모양의 입력이 주어졌을 때, 학생은 우선 해당 이미지가 4개의 변과 4개의 각으로 이루어져 있기 때문에 '사각형'일 것이라 판단하고, 해당 범주에 속한 도형 중 가장 근접한 '일반사각형'을 정답으로 추론할 수 있을 것이라 기대한다.

본 논문의 이후 구성은 다음과 같다. 우선 다음 2장에서는 본 연구와 관련된 선행 연구를 소개하고, 3장에서는 본 연구에서 제안하는 이질적 지식증류 방법론을 소개한다. 4장에서는 제안 방법론과 기존 지식증류 방법론과의 성능을 비교하고, 마지막 5장에서는 본 연구의 기여와 한계를 정리한다.

II. Preliminaries

1. Related works

1.1 Knowledge Distillation

지식증류는 이미 학습된 큰 교사(Teacher) 모델로부터 작은 학생(Student) 모델에 필요한 지식을 전이(Transfer)하는 모델 경량화 기술로, 2015년 Hinton에 의해 처음 확립되었다[2]. 지식증류의 핵심은 교사 모델의 암흑지식(Dark Knowledge)에 있다. 암흑지식은 교사 모델의 마지막 층이 출력한 확률 분포에서, 정답이 아닌 후보들을 정답과 얼마나 유사하게 예측했는지에 대한 정보이다. 이미 학습된 교사 모델의 숨겨진 암흑지식을 학생 모

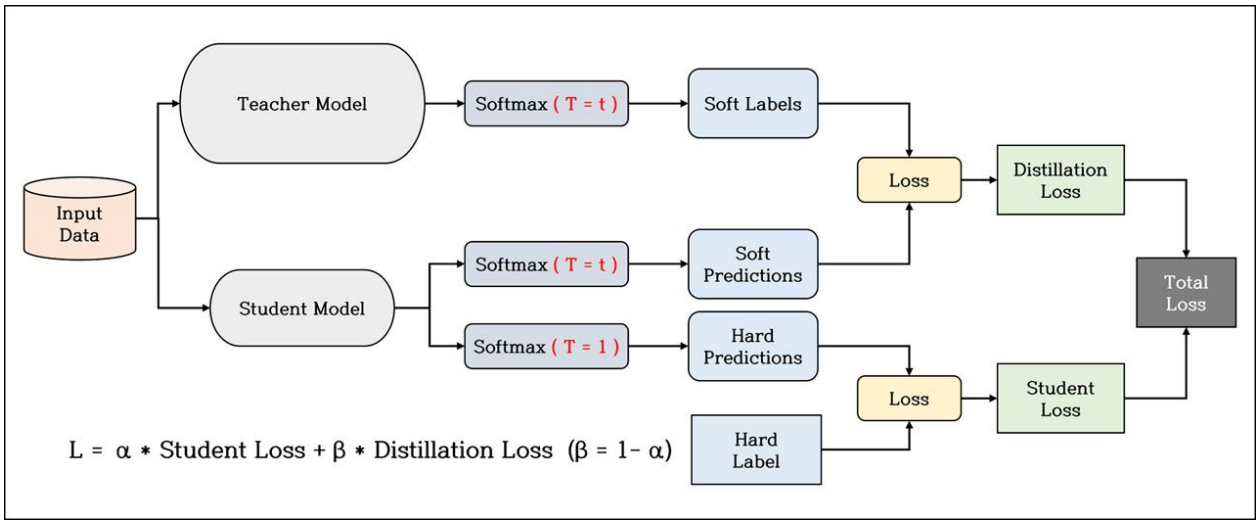

Fig. 3. Learning Process of Knowledge Distillation

델에게 전이시킴으로써, 크기는 작으면서도 교사 모델과 유사한 성능을 가진 학생 모델을 효율적으로 획득할 수 있다. 구체적으로는 교사 모델의 출력 분포와 학생 모델의 출력 분포 사이의 손실(Loss), 그리고 실제 정답과 학생 모델이 예측한 정답 사이의 손실을 가중합하여 이 손실이 최소화되도록 학습을 진행한다(Fig. 3). 이때, 교사 모델의 암흑지식을 효과적으로 전달하기 위해, 출력 분포를 평활화(Smoothing)하는 하이퍼파라미터(Hyperparameter)인 Temperature(T) 값을 사용한다. 암흑지식을 추출하는 위치 및 방법에 따라 교사 모델의 마지막 층의 확률 분포를 지식으로 간주하여 지식증류를 수행하는 Response-Based 연구[3-5], 중간 층(Intermediate Layers)의 특징(Features)을 지식으로 증류하는 Feature-Based 연구[6-8], 그리고 출력값들의 관계(Relationships)를 지식으로 증류하는 Relation-Based 연구[9-11] 등 다양한 연구가 수행되어 왔다[12].

최근 딥 러닝의 활용성 측면이 강조됨에 따라, 지식증류를 통해 모델을 경량화하기 위한 연구가 활발히 이루어지고 있다. 지금까지 모델 경량화는 주로 이미지 분야에서 다루어져 왔다. 하지만 최근에는 BERT[13]와 같이 크고 복잡한 구조로 인해 많은 컴퓨팅 자원을 요구하는 언어모델로부터 가볍고 효율적인 언어 모델을 생성하기 위해, 자연어 처리 분야에서의 모델 경량화가 시도되고 있다. 지식증류를 자연어 처리에 적용한 대표적 연구 분야로는 신경망 기계 번역 (Neural Machine Translation)[14], 텍스트 생성(Text Generation)[15], 질의응답(Question Answering)[16], 문서 검색(Document Retrieval)[17], 그리고 텍스트 분류(Text Classification)[18] 등이 있다. 이처럼 지식증류를 활용한 다양한 연구가 진행되고 있지만, 기존의 접근법은 교사 모델의 지식을 그대로 학생 모델에 전이함으로써 교사 모델이 풀지 못하거나 잘못 학습한 지식도 모두 학생 모델에 전이된다는 한계를 갖는다.

2.2 Text Classification

최근 방대한 양의 텍스트 데이터가 쏟아져 나오며, 이를 효과적으로 관리하기 위해 텍스트 데이터 분류에 대한 필요성이 급증하고 있다. 텍스트 분류(Text Classification)는 주어진 문서를 미리 정의된 클래스(Class)로 분류하는 작업을 의미한다[19]. 분류 문제는 분류하는 클래스가 두 개인 이진분류(Binary Classification)와 세 개 이상인 다중분류(Multi-class Classification)로 구분되며, 대표적인 응용 분야로는 특허 분류[20], 논문 분류[21], 스팸 메일 분류[22], 감성 분석(Sentiment Analysis), 그리고 뉴스 분류[23] 등을 들 수 있다.

방법론적인 측면에서 분류는 지도 학습(Supervised Learning)과 비지도 학습(Unsupervised Learning)으로 나뉜다. 지도 학습은 각 텍스트가 속한 클래스에 대한 정답 레이블(Label)이 미리 주어지고, 텍스트로부터 정답 레이블을 찾아내는 방식을 학습하는 방법이다. 지도 학습의 대표적인 기법으로는 나이브 베이즈 분류기(Naive Bayes Classifier), SVM(Support Vector Machine), 랜덤 포레스트(Random Forest), 의사결정나무(Decision Tree), 그리고 신경망(Neural Network) 등을 들 수 있다[24]. 비지도 학습은 데이터에 대한 정답 레이블 없이 데이터의 특징에 따라 비슷한 데이터끼리 군집화 하는 방법으로 K-평균 군집화(K-means Clustering)와 계층적 군집화(Hierarchical Clustering) 등이 있다[25].

한편 사람의 언어인 자연어를 분류하기 위해서는 비정

형 텍스트를 컴퓨터가 이해하고 처리할 수 있는 벡터 형태로 바꾸는 임베딩(Embedding) 과정이 필요하다. 대표적인 딥 러닝 기반 임베딩 방법으로는 Word2Vec[26], FastText, Glove 등이 있다. 하지만, 이러한 모델들은 데이터 부족으로 인해 텍스트 말뭉치에 없는 단어나 문장에 대한 고유한 의미를 제대로 처리하지 못하는 OOV(Out of Vocabulary) 문제를 가진다. 이에 위키피디아(Wikipedia)와 같은 대규모의 텍스트 말뭉치를 통해 일반적인 의미를 미리 학습해 둔 사전 학습 언어 모델이 등장하였으며, 대표적인 사전 학습 언어 모델로는 ELMo, BERT, ELECTRA[27] 등이 있다. 이 같은 사전 학습 언어 모델은 여러 응용 과업에 범용적으로 적용되어 뛰어난 성능을 보이고 있지만, 거대한 모델 크기, 느린 추론 속도 그리고 고비용이라는 여러 한계도 동시에 지적되고 있다. 최근에는 이러한 다양한 한계를 극복하기 위해 DistilBERT[28], TinyBERT[29], MobileBERT 등 모델을 경량화하기 위한 다방면의 연구가 이루어지고 있다. 하지만, 다양한 도메인과 여러 분야에서 널리 활용되는 텍스트 분류에 대해 세분화된 경량화 모델을 구축하기 위해서는, 수행하고자 하는 과업과 동일한 교사 모델이 각기 필요하다는 한계가 여전히 존재한다.

III. Proposed Method

1. Research Process

본 장에서는 본 논문에서 제안하는 방법론, 즉 학생과 교사가 서로 다른 과업을 통해 학습을 수행하는 이질적 지식증류 방법론의 각 단계별 구체적인 프로세스를 설명한다. 제안 방법론의 전체적인 과정은 <Fig. 4>와 같다.

<Fig. 4>는 지식증류에서 지식을 관통하는 일반적인 지식을 통해 세부적인 지식을 학습하는 과정을 나타낸다. 제안 방법론은 데이터 전처리(Preprocessing) 후 세부 지식을 아우르는 계층적(Hierarchical) 상위 지식(Generic Knowledge)을 학습한 교사 모델을 구축하는 Phase 1, 그리고 학습된 교사 모델로부터 상위 지식을 증류 받으며 구체적인 하위 지식(Specific Knowledge)도 학습하여 학생 모델을 구축하는 Phase 2의 두 단계로 구성된다. 구체적으로 Phase 1은 입력 문서에 대해 (1) 전처리 및 임베딩을 진행하고, (2) 해당 문서에 대응하는 레이블 각각에 대해 계층적 상위 레이블을 추가한 뒤, (3) 상위 레이블과 문서에 대해 분류 학습을 진행한 교사 모델을 획득한다. 이후 Phase 2에서는 (4) 학생 모델이 교사 모델의 학습 결과 분포를 따라가는 과정에서 발생하는 손실(Distillation

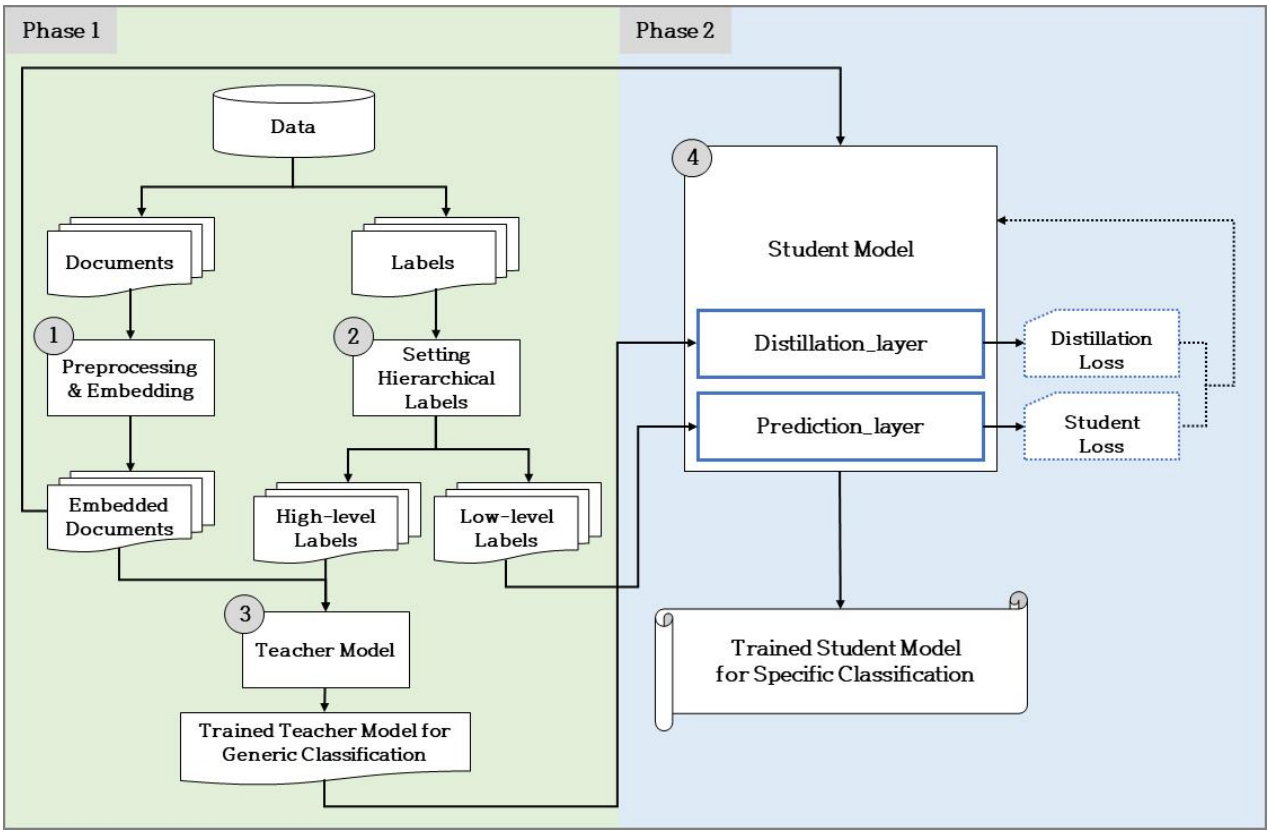

Fig. 4. Overall Research Process

Loss), 그리고 학생 모델의 예측 값과 실제 정답과의 차이에서 발생하는 손실(Student Loss)을 가중합하여, 전체 손실이 최소가 되는 방향으로 지식증류를 진행한다. 이질적 지식증류를 통해 도출되는 최종 학생 모델은, 궁극적으로 예측하고자 하는 하위 지식뿐 아니라 교사 모델이 갖는 상위 지식에 대한 관점까지 함께 학습한 결과이다.

각 과정에 대한 구체적인 내용은 본 장의 각 절에서 가상의 예시와 함께 설명하며, 실제 데이터에 대해 제안 방법론을 적용한 성능 평가 결과는 4장에서 소개한다.

2. Whole Process of Preprocessing

본 절에서는 <Fig. 4>의 단계 중 입력받은 텍스트 문서의 전처리와 임베딩(단계 1), 그리고 일반적 지식 학습을 위해 계층적 상위 레이블을 지정하는 과정(단계 2)을 소개한다. 비정형 텍스트의 경우 구조화되지 않고 일정한 규격이 없는 형태이기 때문에, 본 분석에 선행하여 분석에 적합한 형태로 데이터를 정제하는 과정인 전처리 작업이 이루어져야 한다. <Fig. 5>의 첫 번째 단계(Preprocessing)는 띄어쓰기, 오탈자 교정, 소문자 치환, 특수 문자 제거 등 기본적인 텍스트 전처리 과정을 수행한 결과의 예를 나타낸다. 이렇게 전처리 수행을 거친 텍스트는 이후 임베딩 과정을 거쳐 컴퓨터가 이해할 수 있는 벡터로 표현된다. <Fig. 5>의 두 번째 단계(Embedding)는 다양한 임베딩 기법 중 대표적인 사전 학습 언어 모델인 BERT 기반의 임베딩을 적용하여 텍스트를 벡터로 변환한 결과의 예이다.

Document	Label	H_Label
Anew "Mobile" device has been released! *(1) Preprocessing* a new mobile device has been released *(2) Embedding* [3.02313328e-01 4.60687160e-01 3.77034396e-01 ... -7.35895932e-01 1.07719079e-02 6.11038804e-01]	Mobile Mobile Mobile	 IT IT

Fig. 5. Example of Preprocessing, Embedding and High Level Label Setting

다음으로 <Fig. 4>의 단계 2에서는 각 문서에 부여된 하위 레이블에 대응하는 상위 레이블을 식별한다. 예를 들어 <Fig. 5>에 사용된 레이블이 <Fig. 6>과 같이 3개의 상위 단계(High Level) 카테고리와 6개의 하위 단계(Low Level) 카테고리의 계층적 구조로 구성되어 있다고 가정하자. 이때 <Fig. 5>의 텍스트는 하위 단계에서 Mobile 카테고리로 분류된 문서이기 때문에, <Fig. 6>의 계층적 구조에 따라 상위 레이블은 IT로 구분된다.

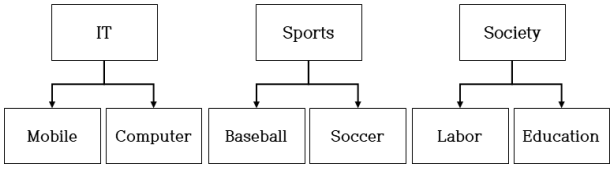

Fig. 6. Example of Hierarchical Labels

3. Teacher Model for Generic Classification

본 절에서는 일반적인 관점에 대한 분류 학습을 수행하여 교사 모델을 구축하는 과정을 소개한다. 본 방법론에서는 일반적인 관점 획득을 위해 계층적 레이블 정보를 사용했으며, 일반적으로 사용되는 DNN(Deep Neural Network) 모델을 바탕으로 큰 크기의 교사 모델에 상위 레이블에 대한 분류 학습을 진행하였다. 본 연구에서는 비교 모델과의 명확한 성능 차이를 확인하기 위해, 사전 학습 모델을 사용하지 않고 처음부터 학습하는 방식(From Scratch)으로 교사 모델을 구축하였다. <Fig. 7>은 일반적인 분류 학습을 수행하는 교사 모델의 예시를 나타낸다.

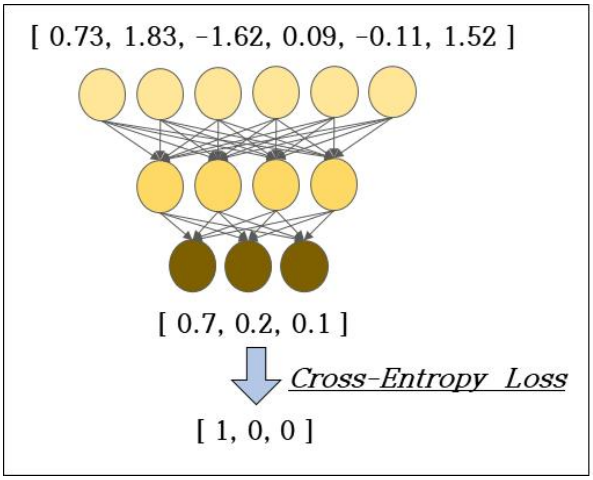

Fig. 7. Example of Teacher Training

$$(\text{식 } 1) \quad y_k = \frac{\exp(a_k)}{\sum_{i=1}^{n} \exp(a_i)}$$

$$(\text{식 } 2) \quad L_{CE} = -\sum_{i=1}^{n} T_i \log(p_i)$$

<Fig. 7>은 6차원 벡터로 주어진 입력 값을 토대로 3개의 클래스 중 하나의 정답을 예측하는 간단한 신경망의 예이다.

이때 모델은 각 클래스에 대한 예측 확률을 출력하는데, 클래스별 확률은 출력 층의 소프트맥스(Softmax) 함수를 통해 총합이 1이 되도록 0과 1 사이의 값으로 정규화 된 값이다(식 1). 이렇게 소프트맥스를 통해 출력된 확률 값이 만약 [0.7, 0.2, 0.1]이라면, 확률 값과 실제 정답 레이블인 [1, 0, 0] 사이의 교차 엔트로피 (Cross-Entropy Loss)를 계산한다. 교차 엔트로피는 실제 값과 예측 값의 차이를 계산하는 손실 함수이며 (식 2)로 계산된다. 예시의 경우 약 0.3567의 손실 값이 산출되며, 교사 모델은 이 손실 값이 작아지는 방향으로 학습을 수행한다.

4. Student Model for Specific Classification

본 절에서는 <Fig. 4>의 (단계 4)에 해당하는 과정, 즉 본 연구에서 제안하는 학생 모델의 이질적인 지식증류 학습 과정을 소개한다. 학생 모델은 앞서 소개한 교사 모델과 마찬가지로 DNN 모델을 사용하며, 추론 속도를 빠르게 하기 위해 교사 모델보다 훨씬 적은 수의 파라미터를 갖는다. 본 논문에서는 최종 분류를 예측하는 층을 Prediction Layer, 그리고 교사 모델로부터 지식을 전이받는 층을 Distillation Layer라고 명명한다. 전체적인 구조는 학생 모델의 Distillation Layer에서 교사 모델의 결과 분포를 따라가며 증류 손실(Distillation Loss)을 구하고, Prediction Layer에서는 실제 정답과 학생 모델의 예측 값과의 손실(Student Loss)을 구하여 두 가지 손실의 가중합이 최소가 되는 방향으로 학습을 진행한다(Fig. 8).

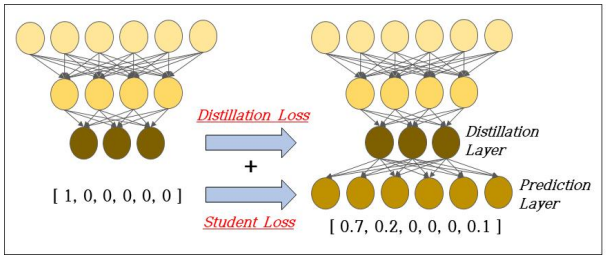

Fig. 8. Process of Heterogeneous Knowledge Distillation

전술한 바와 같이 일반적인 지식증류에는 크게 두 가지 손실이 사용된다. 학생 모델의 예측 확률과 실제 정답과의 차이를 구하는 학생 손실, 그리고 교사 모델의 확률 분포와 학생 모델의 확률 분포 사이의 차이를 구하는 증류 손실이다. 구체적으로 학생 손실은 앞서 교사 모델의 학습과 동일하게 교차 엔트로피로 계산할 수 있고, 증류 손실은 쿨백-라이블러 발산(Kullback-Leibler Divergence)을 통해 두 확률 분포가 얼마나 다른지에 대한 손실을 계산한다 (식 3).

$$(\text{식 } 3)\ D_{KL}(P\|Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)}$$

이렇게 계산된 학생 손실과 증류 손실의 가중합을 통해 지식증류의 최종 손실이 계산되고(식 4), 이 손실이 최소화되는 방향으로 증류를 수행한다.

$$(\text{식 } 4)$$
$$L = \alpha * Student\ Loss + \beta * Distillation\ Loss$$

<Fig. 9>는 기존의 전통적인 지식증류와 제안하는 이질적 지식증류의 구체적인 가상 예시이다.

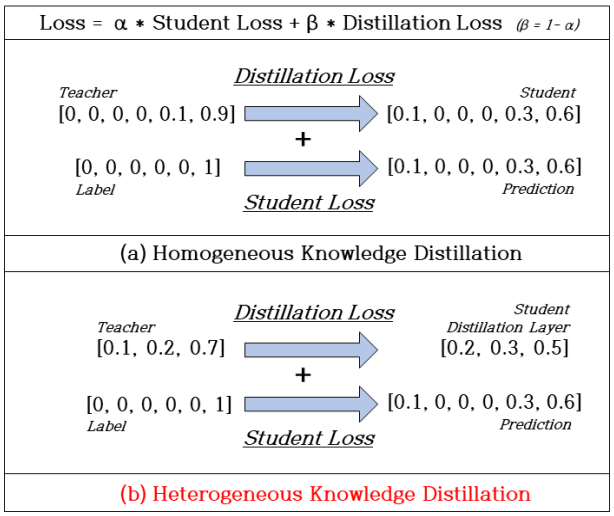

Fig. 9. Example of Calculating Loss of Knowledge Distillation

<Fig. 9>의 예에서 지식증류의 하이퍼파라미터인 Temperature 값은 1로, Student Loss의 가중치는 0.1로 가정한다. <Fig. 9(a)>의 기존 지식증류는 교사 모델과 학생 모델 모두 분류 클래스 개수가 6개인 동일한 과업을 수행하는 경우를 보인다. 반면, <Fig. 9(b)>의 이질적 지식증류는 교사와 학생 모델이 각각 3개와 6개의 분류를 수행하는 서로 다른 과업임을 나타낸다. 구체적으로 <Fig. 9(b)>는 계층적 상위 레이블에 대해 학습한 교사 모델의 예측 분포와 학생 모델의 Distillation Layer의 분포 사이의 차이를 쿨백-라이블러 발산을 통해 증류 손실로 구하고, 실제 분류를 수행하는 Prediction Layer의 확률 값과 실제 정답과의 교차 엔트로피를 통해 기존 지식증류와 동일하게 학생 손실을 구하여 이 두 손실의 가중합이 작아지는 방향으로 학습을 수행한다. 이러한 방법으로 <Fig. 9> 예시의 손실을 구하면, Distillation Loss는 0.0851, Student Loss는 0.5108이며, 이를 가중합한 최종 손실은

0.1277(= 0.9*0.0851 + 0.1*0.5108)이다.

이러한 과정을 통해 제안 방법론은 교사 모델로부터 상위 레이블에 대한 일반적인 지식을 전이받음과 동시에 목표로 하는 세부 분류 관점에 대한 학습도 수행함으로써, 다양한 관점의 지식을 포함한 경량 학생 모델을 도출할 수 있다.

IV. Experiment

1. Experiment Overview

본 절에서는 앞서 소개한 제안 방법론을 실제 데이터에 적용한 실험 결과 및 성능 분석 결과를 소개한다. 실험에는 대표적인 기계학습 라이브러리인 사이킷런(Scikit-learn)에 내장된 20 Newsgroups 데이터 셋[30]을 사용하였다. 본 데이터 셋은 총 20개의 카테고리를 가진 약 18,000개의 데이터로 구성되며, 일반적으로 각 문서의 토픽을 기준으로 컴퓨터, 레저, 과학, 정치, 종교, 그리고 기타와 같은 6개의 대분류로 구분된다. 실험 환경은 Python 3.8을 통해 구축하였으며, 구체적인 H/W 및 S/W 환경은 <Table 1>과 같다. 또한 성능 비교 실험의 전체 프로세스는 <Fig. 10>과 같다.

Table 1. Experimental Environments

HW	GPU	Tesla V100
	CPU	16 core
	Memory	160GB
SW	Python	3.8.13
	Tensorflow-gpu	2.5.0
	Pytorch	1.8.1

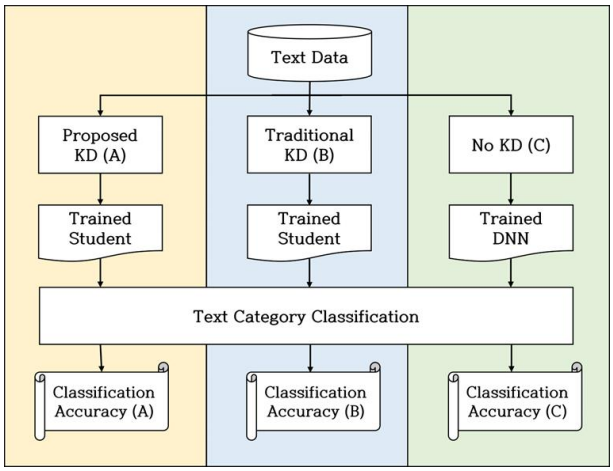

Fig. 10. Overall Process of Performance Evaluation

<Fig. 10>의 (A)는 제안 방법론을 통해 학습된 모델을 평가하는 과정으로, 상위 관점을 학습한 교사 모델로부터 지식을 전이받는 이질적 지식증류를 수행한 학생 모델의 분류 정확도(Classification Accuracy)를 측정한다. <Fig. 10>의 (B)와 (C)는 제안 방법론과의 상대적인 성능 비교를 위해 수행한 실험이다. (B)는 학생 모델이 수행하는 분류와 동일한 분류를 수행한 교사 모델로부터 지식을 전이받는 기존의 전통적인 지식증류 방법을, (C)는 지식증류 없이 단순 분류 학습을 진행한 방법을 의미한다.

2. Results of Preprocessing and Hierarchical Label Setting

본 절에서는 실험 데이터에 대해 전처리를 수행하고, 계층적 레이블을 획득한 과정과 결과를 소개한다. 각 데이터에는 0부터 19까지 총 20개 중 하나의 하위 레이블과, A부터 F까지 총 6개 중 하나의 상위 레이블이 부착되어 있다. <Table 2>는 각 하위 레이블(L)의 주제에 따라 상위 레이블(H)이 구분된 계층적 레이블 구조를 나타낸다.

Table 2. Hierarchical Label Structure

H	L	H	L	H	L	H	L
A	1	B	6	D	11	E	16
	2	C	7		12		17
	3		8		13		18
	4		9		14	F	0
	5		10	F	15		19

우선, 사이킷런에 내장된 20 Newsgroups 데이터는 훈련용(Training)과 평가용(Test) 데이터로 구분되어 있으며, 본 실험에서는 훈련용 데이터 중 일부를 분리하여 검증용(Validation)으로 사용하였다. 구체적으로 띄어쓰기, 오탈자 교정, 소문자 치환, 특수 문자 제거, Null 값 및 중복 제거 등의 전처리 과정을 거친 후, 최종적으로 훈련용 8,765개, 검증용 2,192개, 그리고 평가용 7,274개의 데이터를 실험에 사용하였다. 전처리된 문서는 BERT 모델을 통해 768차원 벡터로 임베딩 되었으며, 레이블의 계층적 구조(Table 2)에 따라 상위 레이블(H_Label)을 식별하였다. <Fig. 11>은 전처리, 상위 레이블 식별 그리고 임베딩 과정을 모두 마친 결과의 예이다.

Document	Preprocessing	Embedding	Label	H_ Label
I must say that I have been a customer of Midwest Micro for over 4 years now …	i must say that i have been a customer of midwest micro for over 4 years now …	[-2.28075236e-01 1.07601464e+00 … 1.55390489e+00 -4.42596436e-01]	3	A
There have been a few postings in the past on alleged pathological …	there have been a few postings in the past on alleged pathological esp neurological …	[-5.28818071e-01 7.10034549e-01 … -8.33536237e-02 -4.01775807e-01]	13	D
For an expansion team? I'm pretty sure I would go with the rings, as long as …	for an expansion team im pretty sure i would go with the rings as long as …	[3.99977833e-01 1.08548677e+00 … 5.69219410e-01 1.41983330e-02]	9	C

Fig. 11. Example of Preprocessed Data

3. Training Teacher for KD

본 절에서는 지식증류에 필요한 교사 모델을 획득한 실험 과정과 성능을 소개한다. 본 실험에서는 제안 방법론의 교사 모델과 비교 모델의 교사 모델로 총 두 개의 교사 모델을 <Table 3>과 같이 구축하였다.

Table 3. Configuration of Two Teacher Models

	Task	# of Param.	Accuracy
Proposed KD (A)	High-level Classification (6 Labels)	1,045,378	0.7822
Traditional KD (B)	Low-level Classification (20 Labels)	1,046,288	0.5980
Common Setting	Batch Size = 128, Epoch = 100 Optimizer = Adam, Early Stopping = True		

먼저, 기존의 전통적인 지식증류의 교사 모델은 학생 모델이 풀고자 하는 과업과 같은 20개의 카테고리를 분류하는 학습을 진행하였다. 1,046,288개의 파라미터를 가진 DNN 모델이며, 학습 후 20개의 분류에 대한 성능은 0.5980으로 나타났다. 한편, 제안하는 이질적 지식증류의 교사 모델은 학생 모델이 풀고자 하는 과업과 서로 다른 과업인 상위 계층 6개의 카테고리를 분류하는 학습을 진행하였다. 1,045,378개의 파라미터를 가진 DNN 모델이며, 학습 후 6개의 분류에 대한 성능은 0.7822로 나타났다. 또한 두 교사 모델은 최종 분류를 수행하는 마지막 레이어의 출력 수만 다르고, 이외의 모든 구조와 학습 조건은 동일하게 구성하여 비슷한 크기의 모델로 구축하였다. 구체적으로, 하이퍼파라미터로 배치 크기(Batch Size)는 128, 에폭(Epoch)은 100, 옵티마이저(Optimizer)는 아담(Adam)으로 지정하였으며, 과적합(Overfitting)을 방지하기 위해 조기 종료(Early Stopping)를 사용하였다.

4. Knowledge Distillation

본 절에서는 20개의 카테고리를 분류하는 경량화된 학생 모델을 구축하기 위한 지식증류 과정을 소개한다. <Fig. 12>는 본 실험에서 구축한 이질적 지식증류의 실제 구조를 나타낸다. 학생 모델은 교사 모델보다 작은 DNN 모델로, 교사 모델에 비해 약 165배가량 적은 6,346개의 파라미터를 갖는다. 20개의 카테고리 분류에 대한 예측을 수행하는 마지막 층인 Prediction Layer는 출력 뉴런의 수를 하위 카테고리의 수인 20으로, 교사 모델의 분포를 따라가며 상위 관점을 전이받는 Distillation Layer는 상위 단계의 레이블 개수인 6으로 출력 크기를 구성하였다.

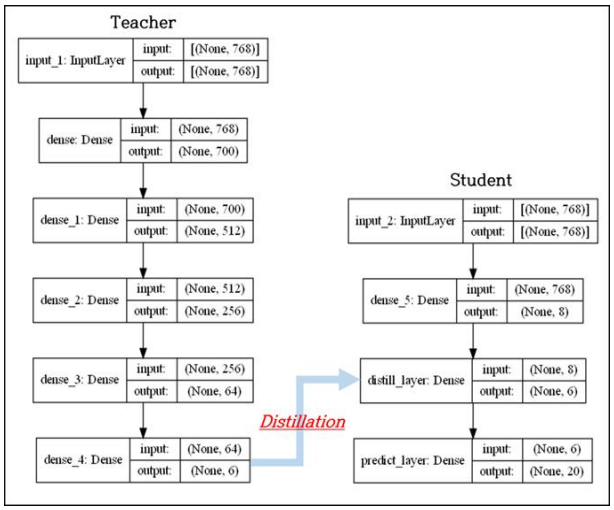

Fig. 12. Architecture of Heterogeneous KD

한편, 지식증류의 하이퍼파라미터인 Temperature(T) 값과 손실 가중치 값 알파(α)는 교사 모델과 학생 모델의 구성, 크기 차이, 데이터에 따라 최적의 값이 달라진다. 본 실험에서는 최적의 하이퍼파라미터를 찾기 위해 Temperature 값과 손실 가중치 값을 바꿔가며 최적의 하이퍼파라미터 조합을 비교하였으며, 이외의 학습에 필요한 하이퍼파라미터는 교사 모델과 동일하게 지정하였다.

5. Performance Evaluation

본 절에서는 본 연구에서 제안하는 이질적 지식증류의 성능을 비교 모델들과 함께 평가한 결과를 소개한다. 우선 <Fig. 10>의 단순 분류 DNN 모델(C)을 통해 지식증류 없는 작은 크기의 모델의 성능을 확인하고, 전통적인 지식증류(B)를 통해 이 성능이 얼마나 상승하는지 확인하였다. 이후, 전통적인 지식증류(B)와 이질적 지식증류(A)의 성능을 비교하여 제안 모델의 우수성을 확인하였다. 성능 비교에는 분류 정확도와 함께 매크로 평균 F1(Macro

Average F1-Score)을 사용하였다. 매크로 평균 F1은 클래스별 점수에 가중치를 두지 않고, 각 클래스별 F1-Score의 평균을 취하여 산출한다. 본 실험에 사용된 데이터 셋은 20개의 카테고리가 각각 유사한 수의 데이터로 구성되었기 때문에, 분류 정확도와 함께 매크로 평균 F1을 사용하여 모델의 성능을 검증하였다. 또한, 실험을 통해 도출한 최적의 하이퍼파라미터 조합인 알파와 Temperature 값은 (A) 모델이 각각 0.3과 28, (B) 모델이 각각 0.2와 23이며, 이를 적용한 본 실험의 세 모델의 성능 평가 결과는 <Table 4>와 <Fig. 13>에 나타나있다.

Table 4. Performance Comparison

	(A) Proposed Method	(B) Traditional KD	(C) Simple DNN
Accuracy	0.5861	0.5837	0.5781
Macro F1	0.5668	0.5643	0.5623

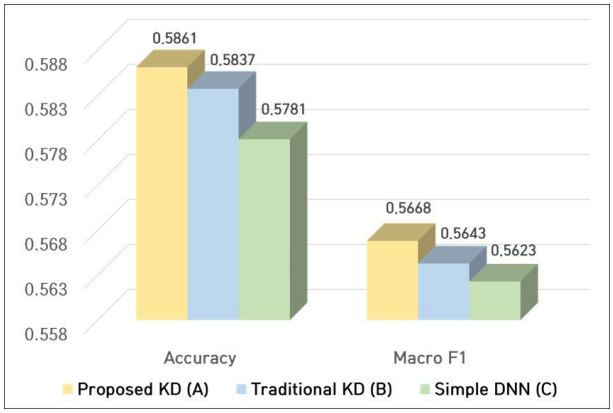

Fig. 13. Results of Performance Comparison

실험 결과 제안 모델 (A)가 정확도와 매크로 평균 F1의 모든 측면에서 가장 우수한 성능을 보이는 것으로 나타났다. 특히, (A) 모델은 해결하고자 하는 과업과 동일한 관점에 대한 지식증류를 수행한 (B) 모델보다도 우수한 성능을 보임으로써, 동질적 관점을 증류한 기존 지식증류보다 이질적 관점을 증류한 제안 방법론이 효과적임을 확인하였다.

또한, 본 연구에서는 학생 모델이 교사 모델과 유사한 성능에 도달하는 속도를 통해 지식증류 학습의 효율성을 평가하기 위해, 조기종료를 해제하고 각 모델의 에폭별 분류 정확도를 확인하였다. <Table 5>와 <Table 6> 그리고 <Fig. 14>는 세 가지 모델의 에폭별 정확도와 매크로 평균 F1을 비교한 결과이다.

Table 5. Comparison of Accuracy in Each Epoch

Epoch	(A)	(B)	(C)
1	**0.2861**	0.2843	0.2850
2	**0.3940**	0.3933	0.3939
3	0.4523	0.4537	**0.4552**
4	0.5021	**0.5025**	0.5005
5	**0.4999**	0.4963	0.4975
6	**0.5349**	0.5323	0.5315
7	**0.5423**	0.5404	0.5410
8	**0.5465**	0.5451	0.5454
9	**0.5645**	0.5608	0.5610
10	**0.5637**	0.5623	0.5627
11	**0.5654**	0.5619	0.5610
12	**0.5621**	0.5594	0.5587
13	0.5652	**0.5653**	0.5642
14	0.5659	**0.5676**	0.5653
15	**0.5763**	0.5723	0.5712
16	**0.5767**	0.5742	0.5731
17	**0.5792**	0.5781	0.5777
18	0.5789	**0.5811**	0.5810
19	0.5861	0.5837	0.5845
20	0.5784	**0.5803**	0.5795
21	**0.5771**	0.5737	0.5744
22	**0.5797**	0.5782	0.5781

Table 6. Comparison of Macro F1 in Each Epoch

Epoch	(A)	(B)	(C)
1	**0.2386**	0.2379	0.2383
	. . .		
7	**0.5176**	0.5156	0.5163
8	**0.5236**	0.5231	0.5234
9	**0.5429**	0.5392	0.5397
10	**0.5391**	0.5374	0.5381
11	**0.5442**	0.5411	0.5400
12	**0.5406**	0.5385	0.5378
13	0.5408	**0.5415**	0.5404
14	0.5437	**0.5457**	0.5438
15	**0.5547**	0.5507	0.5501
16	**0.5515**	0.5493	0.5487
17	0.5595	**0.5600**	0.5598
18	0.5593	0.5617	**0.5622**
19	0.5668	0.5643	0.5658
20	0.5585	**0.5608**	0.5598
21	**0.5547**	0.5512	0.5520
22	**0.5641**	0.5627	0.5623

<Fig. 14>는 <Table 5>와 <Table 6>의 수치를 한 그래프에 나타내고, 정확도에 도달했을 때의 수치를 표시한 결과이다. <Table 5>와 <Table 6>에서는 세 모델들 중 에폭별로 가장 높은 수치를 굵게 표시했으며, 각 모델이 도달한 최종 성능은 색상으로 구분하였다. 실험 결과 제안하는 이질적 지식증류 (A)는 대부분의 에폭별 결과에서 비교 모델 (B)와 (C)보다 우수한 성능을 보임을 확인하였다. 즉, 교사 모델로부터 전이받는 이질적 관점을 통해 학습 초기부터 기존 지식증류 방법보다 빠르게 구체적인 문제해결 능력을 학습할 수 있음을 확인하였다. 또한 이를 통해 정

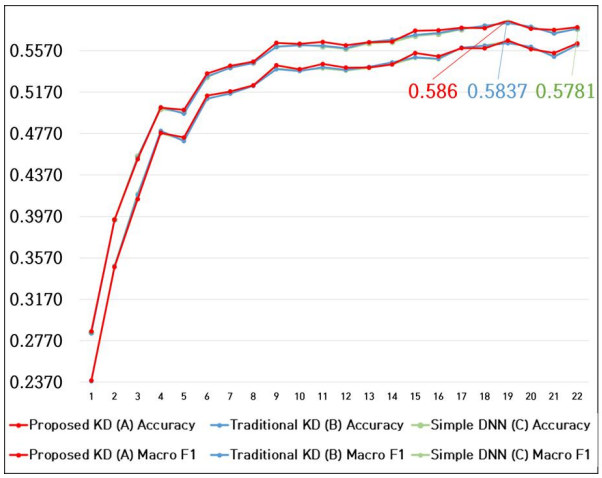

Fig. 14. Accuracy and Macro F1 of Three Models in Each Epoch

확도 목표가 주어진 상황에서의 학습 효율성도 향상시킬 수 있는데, 이는 <Fig. 15>를 통해 확인할 수 있다.

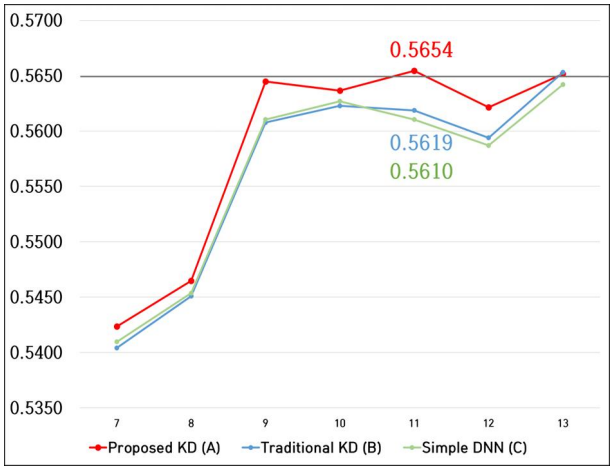

Fig. 15. Epochs Needed to Reach Target Performance

<Fig. 15>는 목표 정확도가 주어진 상황에서의 학습 예를 나타낸다. 예를 들어 정확도 0.565의 성능을 목표로 하는 학생 모델을 구축하고자 할 때, 모델 (A)는 11에폭, (B)는 13에폭, (C)는 그 이상의 학습이 필요함을 알 수 있다. 즉, 연산량이 제한된 환경에서 학생 모델을 구축하는 경우에도, 제안 방법론이 학습 초기부터 서로 다른 관점을 바탕으로 구체적인 학습을 수행하여 기존 지식증류의 방법보다 빠르게 목표 성능에 도달함을 알 수 있다.

이상의 실험 결과를 요약한 <Table 4>를 통해 본 연구에서 제안한 지식증류 방법론이 기존 지식증류 대비 정확도 0.24%p, 매크로 평균 F1 0.25%p의 향상을 달성하였으며, <Table 5>를 통해 정확도 성능 목표치에 더 빨리 도달하는 효율성을 보임을 확인하였다. 즉, 해결하고자 하는 과업 이외의 상위 관점을 전이받는 이질적 지식증류가 학습의 효율성 및 분류 정확도 측면에서 기존의 동질적 지식증류에 비해 우수한 성능을 나타냄을 확인하였다.

V. Conclusions

최근 딥 러닝 모델을 실제 서비스나 제품에 적용하기 위해 모델의 크기와 연산량을 줄이는 경량화 기술의 필요성이 대두되고 있다. 특히, 방대한 크기의 교사 모델과 유사한 성능을 가지는 경량화된 학생 모델을 획득하기 위한 지식증류 기반의 경량화 연구가 활발히 이루어지고 있다. 본 연구에서는 교사 모델과 학생 모델이 동일한 과업을 수행하는 동질적 지식증류의 한계를 개선하고자, 학생 모델의 과업과 상이한 과업을 수행하고 이에 대한 지식을 증류하는 이질적 지식증류 방식을 제안하였다. 또한, 제안 방법론을 사용하여 텍스트 데이터에 대한 카테고리 분류 실험을 수행한 결과, 제안 방법론이 학습 효율성과 정확성 측면 모두에서 단순 분류 학습 모델 및 기존 지식증류 모델에 비해 우수한 성능을 보임을 확인하였다.

본 연구는 지식을 관통하는 일반적인 관점에 대한 지식을 통해 오히려 세부적인 지식도 더욱 잘 학습할 수 있다는 교육학적 관점을 배경으로 이질적 지식증류를 새롭게 제안했다는 점에서 학술적 기여를 인정받을 수 있을 것이다. 또한, 수행하고자 하는 과업에 따라 매번 다른 교사 모델이 필요한 기존 지식증류의 한계를 극복하여, 다양한 과업에 대해 이미 학습된 여러 모델을 이질적 지식증류의 교사 모델로 활용할 수 있다는 측면은 본 연구의 실무적 기여로 인정받을 수 있을 것이다.

다만 본 연구에서 제안한 이질적 지식증류를 수행하기 위해서는 교사와 학생 모델의 학습에 서로 다른 레벨의 정답이 사용되어야 한다. 즉, 분석 대상 데이터 자체가 이미 계층적 레이블을 갖고 있다면 제안 방법론을 그대로 적용할 수 있지만, 그렇지 않은 경우라면 제안 방법론을 적용하기 위해 새로운 계층을 직접 정의해야 한다는 부담이 존재한다. 또한, 본 연구에서는 새롭게 제안한 이질적 지식증류의 성능을 평가하기 위해 텍스트 데이터에 대한 분류 실험 한 가지만을 수행하였는데, 이는 본 연구의 실험 측면의 한계로 지적될 수 있다. 특히 제안 방법론을 통해 성능을 향상시킬 수 있음은 확인하였으나 그 향상 폭이 충분히 크게 나타나지 않았으므로, 향후 연구에서는 텍스트 데이터를 사용하는 여러 과업에 대한 다양한 실험을 통해 제안 방법론의 성능을 더욱 향상시키고 그 효과를 엄밀하게

평가할 필요가 있다. 더 나아가 실시간 추론의 수요가 많은 이미지와 영상 분야의 모델에도 제안 방법론을 적용하여 아이디어의 확장 가능성을 확인할 필요가 있다.

REFERENCES

[1] H. L. Erickson, L. A. Lanning, and R. French, "Concept-Based Curriculum and Instruction for the Thinking Classroom," 2nd Edition, Corwin, 2017. DOI: 10.4135/9781506355382

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," arXiv:1503.02531, Mar, 2015. DOI: 10.48550/arXiv.1503.02531

[3] J. Kim, S. Park, and N. Kwak, "Paraphrasing Complex Network: Network Compression via Factor Transfer," Advances in neural information processing systems 31, 2018. DOI: 10.48550/arXiv.1802.04977

[4] J. Ba and R. Caruana, "Do deep nets really need to be deep?," Advances in neural information processing systems 27, 2014. DOI: 10.48550/arXiv.1312.6184

[5] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved Knowledge Distillation via Teacher Assistant," Proceedings of the AAAI conference on artificial intelligence, Vol. 34, No. 04, pp. 5191-5198, April, 2020. DOI: 10.1609/aaai.v34i04.5963

[6] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FITNETS: HINTS FOR THIN DEEP NETS," arXiv:1412.65504, Mar, 2015. DOI: 10.48550/arXiv.1412.6550

[7] Z. Huang and N. Wang, "Like What You Like: Knowledge Distill via Neuron Selectivity Transfer," arXiv:1707.01219, Dec, 2017. DOI: 10.48550/arXiv.1707.01219

[8] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons," Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, pp. 3779-3787, July, 2019. DOI: 10.1609/aaai.v33i01.33013779

[9] J. Yim, D. Joo, J. Bae and J. Kim, "A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4133-4141, 2017. DOI: 10.1109/cvpr.2017.754

[10] S. Lee and B. Song, "Graph-based Knowledge Distillation by Multi-head Attention Network," arXiv:1907.02226, Jul, 2019. DOI: 10.48550/arXiv.1907.02226

[11] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, and Y. Duan, "Knowledge distillation via instance relationship graph," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7096-7104, 2019. DOI: 10.1109/cvpr.2019.00726

[12] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," International Journal of Computer Vision 129.6, pp. 1789-1819, Mar, 2021. DOI: 10.1007/s11263-021-01453-z

[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, May, 2019. DOI: 10.48550/arXiv.1810.04805

[14] S. Hahn and H. Choi, "Self-Knowledge Distillation in Natural Language Processing," arXiv:1908.01851, Aug, 2019. DOI: 10.48550/arXiv.1908.01851

[15] Y. C. Chen, Z. Gan, Y. Cheng, J. Liu, and J. Liu, "Distilling Knowledge Learned in BERT for Text Generation," arXiv:1911.03829, Jul, 2020. DOI: 10.48550/arXiv.1911.03829

[16] S. Arora, M. M. Khapra, and H. G. Ramaswamy, "On Knowledge Distillation from Complex Networks for Response Prediction," 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. 3813-3822, June, 2019. DOI: 10.18653/v1/N19-1382

[17] S. Shakeri, A. Sethy, and C. Cheng, "Knowledge Distillation in Document Retrieval," arXiv:1911.11065, Nov, 2019. DOI: 10.48550/arXiv.1911.11065

[18] S. Zhang, L. Jiang, and J. Tan, "Cross-domain knowledge distillation for text classification," Neurocomputing, Vol. 509, pp. 11-202022, Oct, 2022. DOI: 10.1016/j.neucom.2022.08.061

[19] N. Kin, D. Lee, H. Choi, and W. X. S. Wong, "Investigations on Techniques and Applications of Text Analytics," Journal of Korean Institute of Communications and Information Sciences, Vol. 42, No. 2, pp. 471-492, Feb, 2017. DOI: 10.7840/kics.2017.42.2.471

[20] S. Kim and S. Kim, "Recursive Oversampling Method for Improving Classification Performance of Class Unbalanced Data in Patent Document Automatic Classification," Journal of The Institute of Electronics and Information Engineers, Vol. 58, No. 4, April, 2021. DOI: 10.5573/ieie.2021.58.4.43

[21] B. Dipto and J. Gil, "Research Paper Classification Scheme based on Word Embedding," Proceedings of the Korea Information Processing Society Conference, Vol. 28, No. 2, Nov, 2021. DOI: 10.3745/PKIPS.y2021m11a.494

[22] H. Son, S. Choe, C. Moon, and J. MIn, "Rule-based filtering and deep learning LSTM e-mail spam classification," Proceedings of the Korean Information Science Society Conference, pp. 105-107, 2021.

[23] S. Ji, J. Moon, H. Kim, and E. Hwang, "A Twitter News-Classification Scheme Using Semantic Enrichment of Word Features," Journal of KIISE, Vol. 45, No. 10, pp. 1045-1055, Oct, 2018. DOI: 10.5626/JOK.2018.45.10.1045

[24] W. X. S. Wong, Y. Hyun, and N. Kim, "Improving the Accuracy of Document Classification by Learning Heterogeneity," Journal of Intelligence and Information Systems, Vol. 24, No. 3, Sep, 2018. DOI: 10.13088/jiis.2018.24.3.021

[25] S. U. Park, "Analysis of the Status of Natural Language Processing Technology Based on Deep Learning," Korean Journal of BigData, Vol. 6, Aug, 2021. DOI: 10.36498/kbigdt.2021.6.1.63

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781, Sep, 2013. DOI: 10.48550/arXiv.1301.3781

[27] K. Clark, M. T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," arXiv:2003.10555, Mar, 2020. DOI: 10.48550/arXiv.2003.10555

[28] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv:1910.01108, Mar, 2020. DOI: 10.48550/arXiv.1910.01108

[29] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for Natural Language Understanding," arXiv:1909.10351, Oct, 2020. DOI: 10.48550/arXiv.1909.10351

[30] K. Lang, "NewsWeeder: Learning to Filter Netnews," Proceedings of the Twelfth International Conference on Machine Learning, pp. 331-339, 1995. DOI: 10.1016/B978-1-55860-377-6.50048-7

Authors

Yerin Yu received the B.A. degree in Management Information Systems from Kookmin University in 2022 and currently enrolled in Graduate School of Business IT, Kookmin University.

She is interested in deep learning, data modeling, and model compression.

Namgyu Kim received the B.S. in Computer Engineering from Seoul National University in 1998, M.S. and Ph.D. degrees in Management Engineering from KAIST, Korea, in 2000 and 2007, respectively.

Dr. Kim joined the faculty of the School of Management Information Systems at Kookmin University, Seoul, Korea, in 2007. He served as the Dean of the Graduate School of Business IT at Kookmin University and is currently a professor at the Business IT. He is interested in deep learning, text mining, and data modeling.