ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2021 September: 27(3): 175~197 http://dx.doi.org/10.13088/jiis.2021.27.3.175

다중 레이블 분류의 정확도 향상을 위한 스킵 연결 오토인코더 기반 레이블 임베딩 방법론

김무성
국민대학교 비즈니스IT전문대학원
(kim27416@kookmin.ac.kr)

김남규
국민대학교 비즈니스IT전문대학원
(ngkim@kookmin.ac.kr)

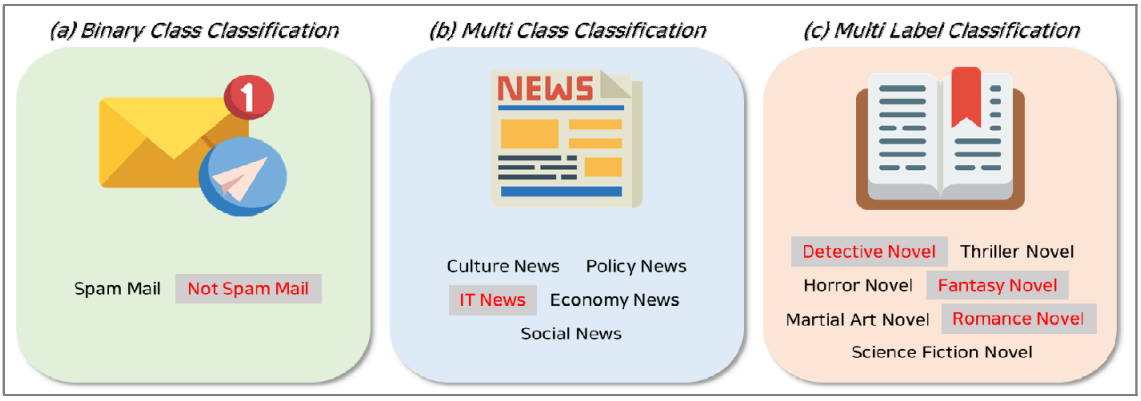
최근 딥 러닝 기술의 발전으로 뉴스, 블로그 등 다양한 문서에 포함된 텍스트 분석에 딥 러닝 기술을 활용하는 연구가 활발하게 수행되고 있다. 다양한 텍스트 분석 응용 가운데, 텍스트 분류는 학계와 업계에서 가장 많이 활용되는 대표적인 기술이다. 텍스트 분류의 활용 예로는 정답 레이블이 하나만 존재하는 이진 클래스 분류와 다중 클래스 분류, 그리고 정답 레이블이 여러 개 존재하는 다중 레이블 분류 등이 있다. 특히, 다중 레이블 분류는 여러 개의 정답 레이블이 존재한다는 특성 때문에 일반적인 분류와는 상이한 학습 방법이 요구된다. 또한, 다중 레이블 분류 문제는 레이블과 클래스의 개수가 증가할수록 예측의 난이도가 상승한다는 측면에서 데이터 과학 분야의 난제로 여겨지고 있다. 따라서 이를 해결하기 위해 다수의 레이블을 압축한 후 압축된 레이블을 예측하고, 예측된 압축 레이블을 원래 레이블로 복원하는 레이블 임베딩이 많이 활용되고 있다. 대표적으로 딥 러닝 모델인 오토인코더 기반 레이블 임베딩이 이러한 목적으로 사용되고 있지만, 이러한 기법은 클래스의 수가 무수히 많은 고차원 레이블 공간을 저차원 잠재 레이블 공간으로 압축할 때 많은 정보 손실을 야기한다는 한계가 있다. 이에 본 연구에서는 오토인코더의 인코더와 디코더 각각에 스킵 연결을 추가하여, 고차원 레이블 공간의 압축 과정에서 정보 손실을 최소화할 수 있는 레이블 임베딩 방법을 제안한다. 또한 학술연구정보서비스인 'RISS'에서 수집한 학술논문 4,675건에 대해 각 논문의 초록으로부터 해당 논문의 다중 키워드를 예측하는 실험을 수행한 결과, 제안 방법론이 기존의 일반 오토인코더 기반 레이블 임베딩 기법에 비해 정확도, 정밀도, 재현율, 그리고 F1 점수 등 모든 측면에서 우수한 성능을 나타냄을 확인하였다.

주제어 : 딥 러닝, 다중 레이블 분류, 레이블 임베딩, 오토인코더, 스킵 연결

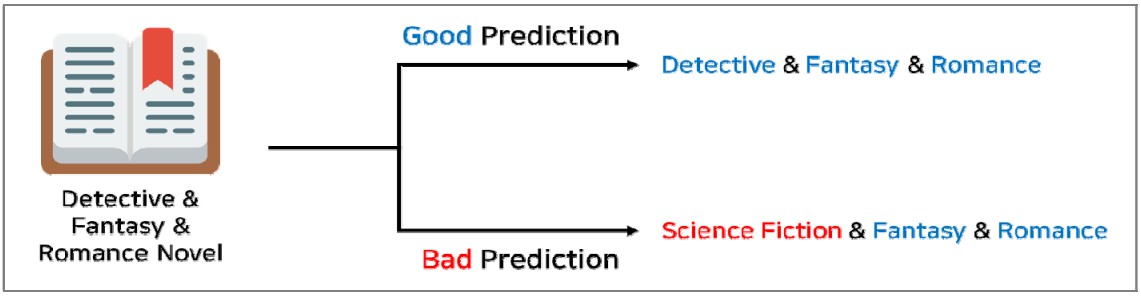
논문접수일 : 2021년 6월 9일 논문수정일 : 2021년 6월 28일 게재확정일 : 2021년 7월 19일
원고유형 : 일반논문 교신저자 : 김남규

1. 서론

최근 딥 러닝(Deep Learning) 기술이 발전함에 따라 텍스트, 이미지 등의 비정형 데이터 분석에 딥 러닝 기술을 활용하려는 시도가 증가하고 있으며, 특히 뉴스, 논문, 블로그 등 다양한 문서에 포함된 텍스트 분석에 딥 러닝 기술을 활용하는 연구가 활발하게 수행되고 있다. 딥 러닝을 활용한 텍스트 분석의 대표적인 활용 분야로는 문서를 스팸(Spam), 감성(Sentiment), 카테고리(Category) 등 사전에 정의된 특정 클래스(Class)로 분류하는 텍스트 분류(Text Classification), 텍스트에서 인명, 단체, 장소 등 사전에 정의된 개체(Entity)를 추출하는 개체명 인식(Named Entity Recognition), 많은 양의 텍스트에서 핵심 내용만 추출하거나 새롭게 구성하여 요약문으로 변환하는 텍스트 요약(Text Summarization), 챗봇(Chatbot)과 같이 질문에 대한 답변을 생성하거나 기존에 존재하



〈Figure 1〉 Classification Types

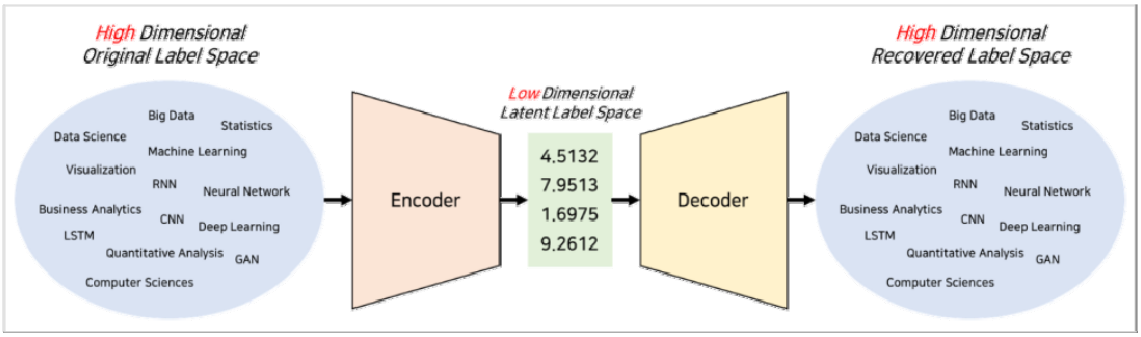


〈Figure 2〉 Multi-Label Classification Example

지 않던 새로운 시, 소설 등을 창작하는 텍스트 생성(Text Generation), 그리고 BERT와 같이 대용량의 말뭉치(Corpus)를 사용하여 단어 시퀀스(Sequence)에 대한 확률을 할당(Assign)하는 사전 학습 언어 모델(Pre-trained Language Model) 등을 들 수 있다.

다양한 텍스트 분석 응용 가운데, 학계와 업계에서 가장 많이 활용되는 대표적 기술로 텍스트 분류를 꼽을 수 있다. 텍스트 분류의 활용 예로는 정상 메일과 스팸 메일을 분류하는 이진 클래스 분류(Binary Class Classification), 뉴스 기사를 정치, 사회, 연예 등의 카테고리 중 하나의 카테고리로 분류하는 다중 클래스 분류(Multi-Class Classification), 소설을 액션, 드라마, 스릴러 등의 장르 중 하나 이상의 장르로 분류하는 다중 레이블 분류(Multi-Label Classification) 등을 들 수 있다(Figure 1). 특히 다중 레이블 분류는 현실 세계에서 자주 등장하는 문제이며, <Figure 2>의 예시와 같이 'Fantasy'와 'Romance' 장르를 올바르게 예측해도 'Science Fiction' 장르처럼 단 하나라도 레이블을 잘못 예측하면 전체적으로 예측된 장르들이 소설의 내용과 전혀 다른 장르가 되기 때문에 다중 레이블 분류는 굉장히 중요하고 어렵다.

이러한 다중 레이블 분류는 하나의 인스턴스(Instance)가 하나 이상의 레이블을 갖는다는 특

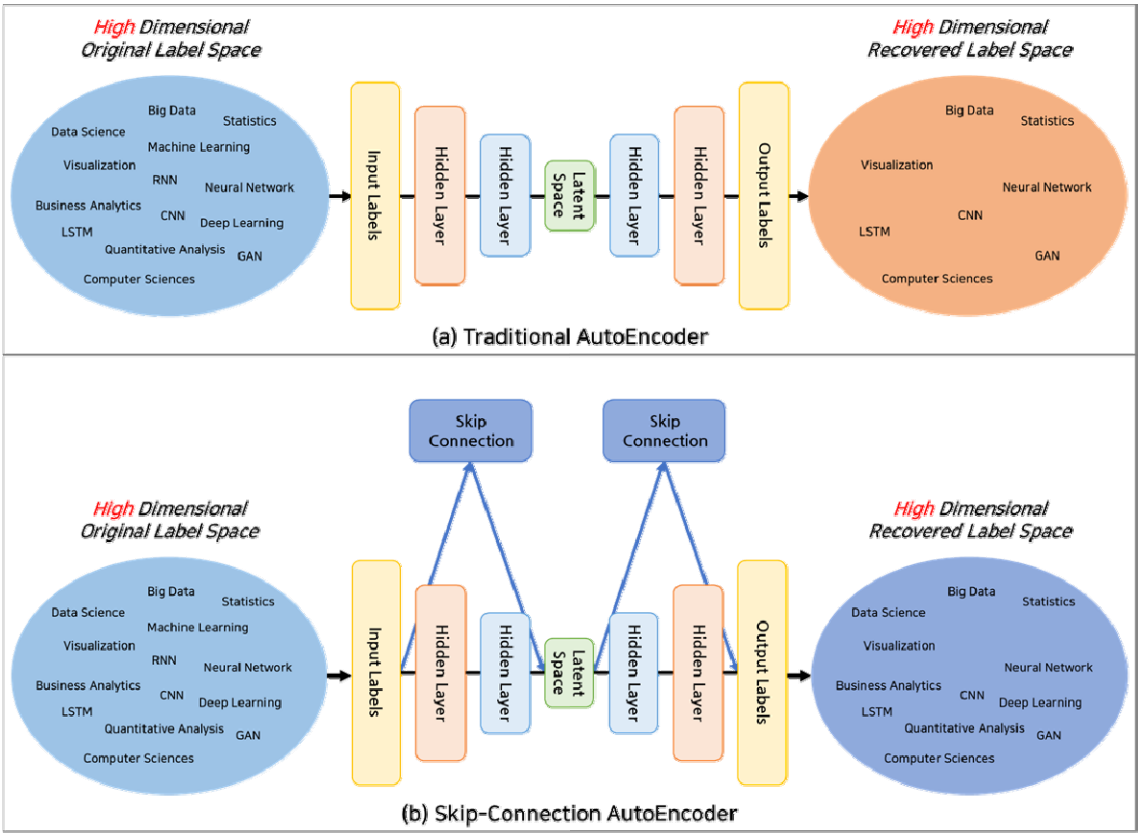


〈Figure 3〉 Label Embedding Process

성으로 인해 일반적인 분류와는 상이한 방법으로 학습이 진행된다. 구체적으로 다중 레이블 분류의 학습에는 이진 연관성(Binary Relevance), 레이블 멱집합(Label Powerset) 등의 알고리즘을 사용하여 단일 레이블(Single Label) 문제로 바꾸는 문제 변환(Problem Transformation), 그리고 기존의 알고리즘을 데이터에 맞게 바꾸는 알고리즘 적용(Algorithm Adaptations) 등의 개념이 적용된다(Dhatri and Rachana, 2018). 하지만, 다중 레이블 분류를 위한 기존의 학습 방법들은 레이블 사이의 연관성을 충분히 고려하지 못하기 때문에 성능 저하를 야기하고, 레이블과 클래스의 개수가 많아질수록 예측 난이도의 증가로 인해 이러한 성능 저하 현상은 더욱 심각하게 나타나게 된다.

이러한 한계를 극복하기 위해 레이블 공간을 압축하고 압축된 레이블을 예측하는 레이블 임베딩(Label Embedding)(Farbound and Hsuan-Tien, 2012; Jörg et al., 2012; Jörg et al., 2016; Chih-Kuan et al., 2017)에 관한 연구가 활발하게 수행되고 있다. 레이블 임베딩은 (i) 최초에 주어진 고차원의 레이블 공간(Label Space)를 저차원의 잠재 레이블 공간(Latent Label Space)으로 압축하고, (ii) 압축된 레이블을 예측하는 학습을 수행한 후, (iii) 예측된 레이블을 고차원의 원래 레이블 공간으로 복원하는 과정으로 이루어진다(Figure 3).

전통적인 레이블 임베딩 기법으로 PLST(Principal Label Space Transformation)(Farbound and Hsuan-Tien, 2012), MLC-BMaD(Multi-Label Classification via Boolean Matrix Decomposition)(Jörg et al., 2012) 등이 있으며, 최근에는 MANIAC(Jörg et al., 2016), C2AE(Chih-Kuan et al., 2017) 등 오토인코더(AutoEncoder)(Pierre and Kurt, 1989) 기반의 레이블 임베딩 기법들이 많은 주목을 받고 있다. 신경망의 일종인 오토인코더는 인코더(Encoder)와 디코더(Decoder)의 구조를 갖고 있으며, 동일한 데이터를 입력 데이터와 출력 데이터로 활용하는 자기 지도 학습(Self-Supervised Learning)을 수행한다. 인코더로 레이블을 압축하고 디코더로 복원하는 학습에서 노이즈와 은닉층을 추가하여 저차원 잠재 공간을 효과적으로 표현하며, 이를 통해 레이블 간 비선형 종속성 포착, 계산 비용 감소, 분류 성능 향상 등의 효과를 거둘 수 있다. 그럼에도 불구하고 기존의 오토인코더 기반 레이블 임베딩 기법은 클래스의 수가 매우 많은 고차원 레이블 공간을 압축할 때 정보 손실이



〈Figure 4〉 Comparison of Traditional AutoEncoderand Skip-Connection AutoEncoder

크게 발생한다는 한계를 갖는다. 원래 레이블 공간 대신 잠재레이블 공간을 타겟으로 활용하는 것이기 때문에 원래 레이블 공간의 정보가 잠재 레이블 공간에 최대한 보존되어야 하며, 잠재 레이블 공간의 정보 손실은 다중 레이블 분류의 성능 저하로 이어진다. 특히, 다중 레이블 분류는 레이블의 개수가 증가하면 이에 따라 클래스의 개수도 기하급수적으로 증가하는 특성으로 인해 멀티-핫 인코딩(Multi-Hot Encoding) 표현에서 매우 희소한(Sparse) 고차원 레이블 공간이 형성된다. 이러한 레이블 공간을 저차원으로 압축하는 레이블 임베딩을 수행하면, 의미 없는 '0' 값이 대폭 증가하게 되어 레이블의 정보가 잘 반영된 잠재 레이블 공간을 만들기 힘들다. 또한, 잠재 레이블 공간의 차원이 작아질수록 정보 손실이 크게 발생하며, 이는 잠재 레이블 공간을 다시 원래 레이블 공간으로 정확하게 복원하기가 어려워짐을 의미한다. 이러한 한계는 학습의 역전파(Backpropagation) 과정에서 발생하는 기울기 소실 문제에서 그 원인을 찾을 수 있다.

이와 같은 기울기 소실 문제를 해결하기 위해 고안된 것이 스킵 연결(Skip-Connection)이다. 스킵 연결은 층의 입력을 출력과 더해주는 구조로, 역전파 때 기울기 소실을 막음으로써 층이 깊어

져도 효율적인 학습이 가능하다. 스킵 연결은 주로 합성곱 신경망에서 이미지의 특징을 추출할 때 많이 활용되며, ResNet(Kaiming et al., 2016)에서 사용되는 구조가 대표적이다. 이러한 스킵 연결을 추가한 오토인코더 모델과 전통적인 오토인코더 모델의 구조는 <Figure 4>과 같이 비교될 수 있다. <Figure 4>는 각 모델이 n개의 클래스를 가진 데이터를 압축하는 레이블 임베딩을 수행했을 때, 잠재 레이블 공간을 얼마나 원본 레이블 공간에 가깝게 복원하는지를 비교하여 제시하고 있다. <Figure 4>의 (a) Traditional AutoEncoder는 복원된 레이블 공간이 원본 레이블 공간과 상이하며, 압축과 복원 과정에서 많은 정보가 손실되었음을 확인할 수 있다. 반면, (b) Skip-Connection AutoEncoder는 복원된 레이블 공간이 원본 레이블 공간과 유사하며, 압축과 복원 과정에서 많은 정보가 보존되었음을 확인할 수 있다. 이러한 예는 스킵 연결을 추가함으로써 궁극적으로 고차원 레이블 공간의 정보를 최대한 보존하는 저차원 잠재 레이블 공간을 도출할 수 있음을 보여준다.

이에 본 연구에서는 고차원 레이블 공간의 정보를 최대한 손실 없이 저차원 잠재 레이블 공간으로 압축하는 스킵 연결을 추가한 오토인코더 기반의 레이블 임베딩 방법론을 제안한다. 구체적으로 본 논문에서 제안하는 방법론은 (i) 데이터셋(Dataset)을 특징(Features)과 레이블(Labels)로 나누어 전처리와 멀티-핫 인코딩을 진행하고, (ii) 스킵 연결이 추가된 오토인코더로 레이블을 압축하고 복원하는 학습을 통해 잠재 레이블 공간을 도출한다. (iii) 다음으로 전처리된 특징들로 잠재 레이블 공간의 벡터들을 예측하는 회귀 학습을 수행한 후, (iv) 마지막으로 학습된 회귀 모델의 예측 값을 다시 원래 레이블 공간으로 복원하는 과정으로 구성된다.

본 논문의 이후 구성은 다음과 같다. 우선 다음 장인 2장에서 본 연구와 관련된 선행 연구를 소개하고, 3장에서는 본 연구에서 제안하는 방법론을 가상의 예와 함께 설명한다. 4장에서는 제안 방법론을 사용한 레이블 임베딩과 다중 레이블 분류 성능 그리고 기존의 오토인코더를 사용한 레이블 임베딩과 다중 레이블 분류 성능의 분류 실험 정확도를 비교하고, 도메인 특성에 따른 레이블 임베딩 효과와 최적의 잠재 레이블 공간 차원 수 탐색을 통해 제안 방법론의 활용성을 확인한다. 마지막으로 5장에서는 본 연구의 기여와 한계를 정리한다.

2. 관련 연구

2.1. 레이블 임베딩

레이블 임베딩은 선형 또는 비선형 임베딩을 통해 저차원 잠재 공간에 레이블을 투영하는 것을 목표로 하며, 주로 다중 레이블 분류에 사용된다. 레이블의 수가 한 개 이상인 다중 레이블 분류의 특성상 레이블의 개수가 증가하면 이에 따라 클래스의 개수도 기하급수적으로 증가하기 때문에, 일반적으로 사용되는 분류 알고리즘으로는 성능 향상이 어렵다는 한계를 갖고 있다. 이러한 한계를 극복하기 위해 고안된 것이 레이블 임베딩이며, 일반적으로 원래의 고차원 레이블 공간을 저차원 잠재 레이블 공간으로 투영하는 인코딩 과정과 투영된 잠재 레이블 공간의 벡터들을 예측하는 학습 과정, 그리고 예측된 벡터들을 원래 레이블 공간으로 복원하는 디코딩 과정으로 구성된다.

전통적인 레이블 임베딩 기법으로는 PLST (Principal Label Space Transformation)(Farbound and Hsuan-Tien, 2012), MLC-BMaD(Multi-Label Classification via Boolean Matrix Decomposition) (Jörg et al., 2012), 그리고 BML-CS(Bayesian Multi-Label Compressed Sensing)(Ashish et al., 2012) 등이 대표적이다. PLST(Farbound and Hsuan-Tien, 2012)는 원래의 레이블 행렬에 대한 특이값 분해(Singular Value Decomposition)과 선형 인코딩, 그리고 선형 디코딩을 통해 레이블 임베딩을 수행한다. 특히 고차원 레이블 공간에서 평면을 사용하여 레이블 간의 중요한 상관 관계를 포착하기 때문에, 많은 수의 레이블을 갖는 데이터셋을 분석할 때 계산 비용을 감소시킬 수 있다는 장점이 있다. MLC-BMaD(Jörg et al., 2012)는 원래 레이블의 유용한 Boolean 조합을 나타낼 수 있는 잠재 레이블을 추출한다. 구체적으로 원래 레이블 행렬은 잠재 레이블의 인수 행렬과 레이블 간의 상호 종속성을 나타내는 인수 행렬로 분해되며, 잠재 레이블의 인수 행렬을 예측하는 학습을 수행한다. 최종적으로 레이블 간 상호 종속성의 인수 행렬을 사용하여 Boolean 행렬 곱셈에 의해 레이블을 예측한다. 마지막으로 BML-CS(Ashish et al., 2012)는 임의 변환(Random Transformation)을 통해 원래 레이블 공간을 더 낮은 차원의 잠재 공간으로 투영한 다음, 투영된 벡터들을 타겟으로 하는 회귀(Regression) 학습을 진행한다. 특히, 관찰되지 않은 모든 레이블에 대한 공동 사후 분포를 제공하는 추론(Inference)을 통해, 누락된 레이블이 있는 데이터셋을 자연스럽게 처리함과 동시에 예측의 불확실성을 측정할 수 있다는 장점이 있다. 하지만 이러한 기법들은 레이블 간의 선형 관계만을 고려하거나 임의 변환으로 레이블을 압축하기 때문에, 레이블 간 비선형적인 관계를 파악하기 어려워서 원래 레이블의 정보를 충분히 담고 있는 잠재 레이블 공간을 만들지 못한다는 한계를 갖는다.

2.2. 오토인코더 및 오토인코더 기반 레이블 임베딩

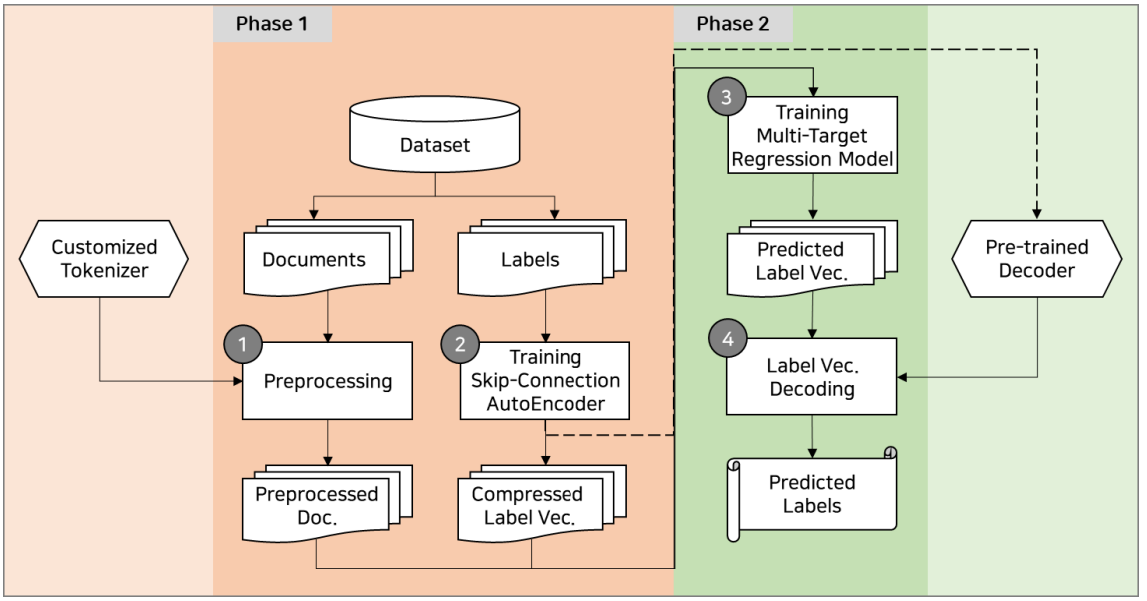
기존 레이블 임베딩 기법들의 한계를 극복하기 위해 최근에는 오토인코더 기반의 레이블 임베딩 기법들이 주목받고 있다. 오토인코더(Pierre and Kurt, 1989)는 데이터를 압축하는 인코더와 압축된 데이터를 복원하는 디코더로 구성되며, 동일한 데이터를 입력 데이터와 출력 데이터로 활용하는 자기 지도 학습(Self-Supervised Learning)을 수행한다. 이 기법은 주로 입력 데이터의 특징을 추출하거나 차원을 축소하기 위해 많이 사용되며, 입력층(Input Layer)과 병목층(Bottleneck Layer), 그리고 출력층(Output Layer)이 각각 하나씩 존재하는 오토인코더를 기본적인 구조로 사용한다. 이러한 오토인코더는 병목층에 해당하는 잠재 공간의 정보만으로 원래 정보를 복원해야 하기 때문에 잠재 공간을 잘 표현하는 것이 중요하며, 이와 같은 학습 목표로 인해 표현 학습(Representation Learning)이라고도 불린다. 이에 따라 잠재 공간을 잘 표현할 수 있는 다양한 오토인코더가 고안되었으며, 은닉층을 여러 개 쌓아 깊은 인코더와 디코더 구조를 갖는 Stacked AutoEncoder(Yoshua et al., 2006), 노이즈(Nosie)를 추가하여 입력 데이터에 대해 견고한 모델을 만드는 Denoising AutoEncoder(Pascal et al., 2008), 그리고 Stacked AutoEncoder 구조에 노이즈가 추가된 입력 데이터를 사용하는 Stacked Denoising AutoEncoder(Pascal et al., 2010)가 대표적이다.

레이블 임베딩 분야에서는 레이블의 정보를 잘 압축하고 복원할 수 있는 잠재 공간을 표현하기 위해 오토인코더가 활용되기 시작했다. 최초의 오토인코더 기반 레이블 임베딩 기법인 MANIAC(Multi-lAbel classificatioN usIng AutoenCoders) (Jörg et al., 2016)은 인코더로 원래 레이블 공간을 잠재 레이블 공간으로 압축하고, 압축된 잠재 레이블 공간을 타겟으로 하는 다중 타겟 모델(Multi Target Model)을 학습한다. 그리고 학습된 다중 타겟 모델의 예측 값을 디코더를 통해 원래 레이블 공간으로 복원하여 다중 레이블 분류 문제를 해결한다. 오토인코더 기반의 레이블 임베딩은 학습 과정에서 레이블 간의 비선형 종속성을 포착할 수 있기 때문에, 궁극적으로 다중 레이블 분류의 예측 성능을 향상시킬 수 있다는 장점이 있다.

오토인코더 기반 레이블 임베딩이 괄목할만한 성과를 보이면서 이에 대한 다양한 후속 연구가 수행되었으며, 대표적으로 레이블과 특징 정보를 함께 활용하는 특징 인식 레이블 임베딩(Feature-Aware Label Embedding) 기법인 C2AE (Canonical-Correlated Autoencoder)(Chih-Kuan et al., 2017)가 있다. 이는 깊은 정준 상관 분석(DCCA: Deep Canonical Correlation Analysis)과 오토인코더 구조를 통합하여 특징과 레이블로부터 잠재 공간을 표현한다. 학습 단계에서 데이터의 특징과 레이블이 C2AE의 입력으로, 입력으로 사용된 데이터의 레이블이 출력으로 사용된다. 이때 DCCA를 통해 데이터의 특징과 레이블을 연결한 후 오토인코더의 인코더로 공동 학습(Joint Learning)을 수행하며, 오토인코더의 디코더로 데이터의 레이블을 복원하면서 잠재 공간이 표현된다. 이러한 학습 방법을 통해 특징으로부터 바로 레이블 예측이 가능한 End-to-End 구조를 갖는다. 이 외에도 인코딩 층에 Binary Stochastic Neurons을 추가하여 볼륨 보존(Volume-Preserving) 잠재 공간을 형성하는 BSAE(Binary Stochastic AutoEncoder) (Denis et al., 2018)와, C2AE 구조에 단어 벡터 기반의 셀프 어텐션(Self-Attention)과 순위 손실(Ranking Loss)을 추가한 Rank-AE(Bingyu et al., 2019) 등 오토인코더 기반의 레이블 임베딩에 관한 연구가 활발하게 수행되고 있다.

2.3. 스킵 연결

스킵 연결은 층의 입력을 그대로 출력에 더해주는 방식으로, 층이 깊어질수록 기울기 소실이 심해지는 심층신경망의 근본적인 문제를 해결하였다. 이 기법은 주로 합성곱 신경망에서 이미지의 특징을 추출할 때 많이 활용되며, 이미지 인식 대회인 ILSVRC(ImageNet Large Scale Visual Recognition Challenge)에서 우승한 ResNet(Kaiming et al., 2016)을 통해 그 효과가 입증되었다. 이후 이미지 분류, 객체 인식, 이미지 생성, 이미지 복원 등 다양한 컴퓨터 비전 분야에서 스킵 연결 기반의 합성곱 신경망이 활용되고 있다. 또한 Seq2Seq 기반 모델의 단점이었던 고정된 문맥 벡터(Context Vector)로 인한 장기 의존성 문제를 해결하기 위해 고안된 트랜스포머(Transformer) (Ashnish et al., 2017) 역시 스킵 연결을 사용하는 등, 자연어처리 분야에서도 스킵 연결을 사용한 연구가 활발하게 수행되고 있다. 최근에는 합성곱 오토인코더(Convolutional AutoEncoder)에 인코더와 디코더를 스킵 연결하여 초해상도 이미지를 복원하는 연구(Jo et al., 2020)를 비롯하여 스킵 연결을 오토인코더 구조에 적용하려는 시도가 일부 이루어지고 있지만, 레이블 임베딩에서 정보 손실을 최소화하는 잠재 레이블 공간을 추출하기 위해 스킵 연결을 사용하는 연구는 아



〈Figure 5〉 Overall Research Process

직 부족하다.

3. 제안 방법론

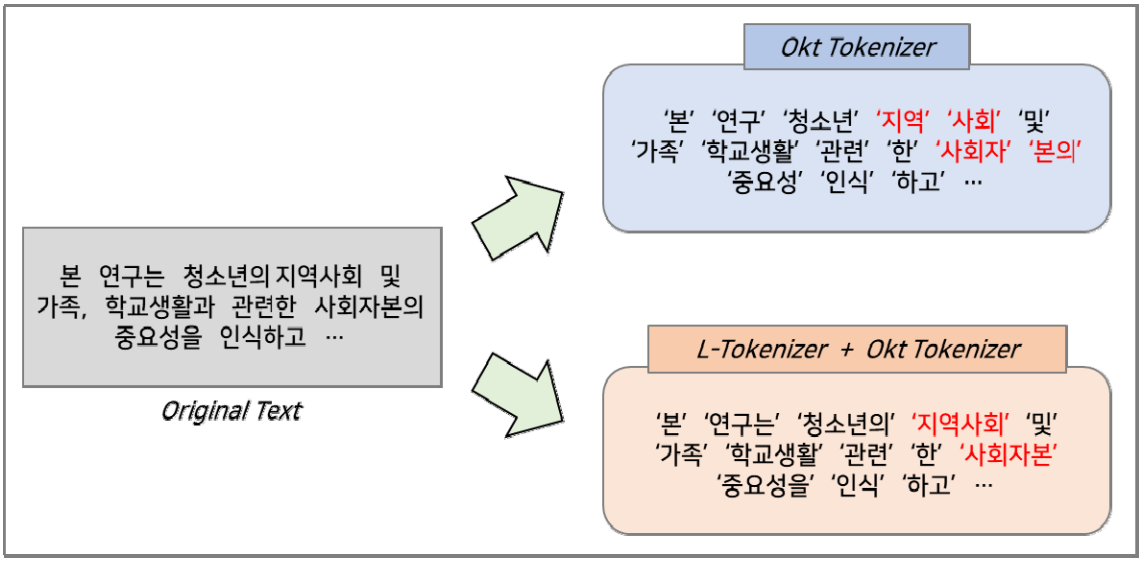
3.1. 제안 방법론 개요

본 장에서는 본 연구에서 새롭게 제안하는 스킵 연결을 추가한 오토인코더 기반의 레이블 임베딩 방법론을 소개한다. 제안 방법론의 전체적인 과정은 <Figure 5>와 같다. 제안 방법론은 먼저 문서를 전처리하고 레이블 임베딩을 통해 압축된 잠재 벡터를 도출하는 Phase 1, 그리고 문서로부터 잠재 벡터를 예측하는 다중 타겟 회귀 모델(Multi-Target Regression Model)을 학습하고 예측한 값을 원래 레이블 공간으로 복원하는 Phase 2의 두 단계로 구성된다. 구체적으로 Phase 1은 데이터셋을 문서와 레이블로 분리한 후 (1) 문서에 대해서는 형태소 분석, 불용어 제거 등의 전처리 작업을 수행하고, (2) 레이블은 스킵 연결 오토인코더를 통해 고차원 레이블 공간에서 저차원 잠재 레이블 공간으로 변환한다. 이후 Phase 2에서는 (3) 전처리된 문서로부터 저차원 잠재 레이블 공간의 벡터들을 예측하는 다중 타겟 회귀 모델을 학습하고, (4) 이후 (2)에서 학습된 스킵 연결 오토인코더의 디코더를 사용하여 다중 타겟 회귀 모델의 예측 값을 원래 레이블 공간으로 복원한다.

각 단계에 대한 구체적인 작동 원리는 다음 절부터 예시와 함께 설명하며, 실제 데이터를 적용한 제안 방법론의 성능과 평가는 4장에서 소개한다.

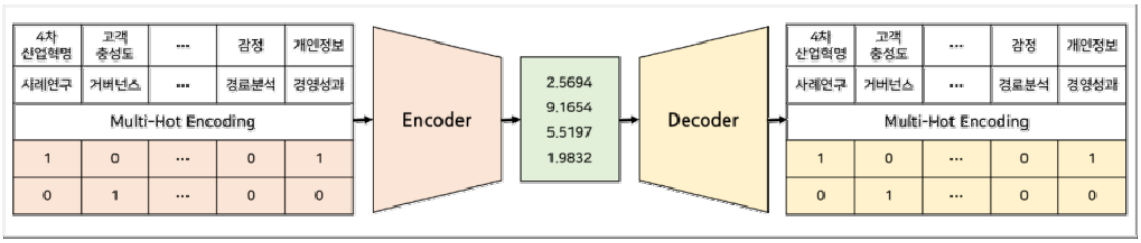
3.2. 문서 전처리

본 절에서는 <Figure 5>의 문서 전처리(단계



〈Figure 6〉 Example of Document Processing

1) 과정을 소개한다. 문서 분류 학습을 수행하기 위해서는 문서를 모델에 입력 가능한 형태로 변환해야 한다. 이를 위해 분석에 적절한 문서를 선정하는 다양한 필터링 작업과 형태소 분석, 그리고 불용어 제거를 수행하여 문서를 토큰 단위로 분리한다. 본 연구에서는 필터링 조건으로 문서의 문장 수, 일정 빈도 이상 등장한 레이블을 갖는 문서 등을 설정하며, 외국어, 특수기호, 그리고 조사 등을 불용어로 처리한다. 특히 문서 내 단어들의 의미를 최대한 보존하기 위해 두 번의 형태소 분석을 수행한다. 구체적으로는 먼저 문서를 Soynlp의 L-Tokenizer로 토큰화하고, L-Tokenizer로 도출된 토큰들로 사전을 구축한다. 이렇게 구축된 토큰 사전을 Konlpy의 Okt 사용자 사전과 병합한 후, 사전 기반 형태소 분석을 수행하는 Okt Tokenizer로 최종 토큰화 작업을 수행하였다. 본 연구에서 사용한 L-Tokenizer는 여러 개의 문서로부터 학습한 통계 정보를 활용하여 통계적 패턴 기반의 단어를 추출하는 비지도 학습 기법으로, 사전에 포함되어 있지 않은 단어의 검출 및 복합명사의 추출에 사용될 수 있다. <Figure 6>은 L-Tokenizer를 사용했을 때와 사용하지 않았을 때의 결과를 비교한 것으로 불용어 제거까지 완료된 전처리 결과의 예시이다. <Figure 6>에서 예를 들어 '지역사회', '사회자본의' 등의 표현에 대해 Okt Tokenizer만 사용하여 토큰화를 수행하는 경우 '지역', '사회', '사회자', '본의' 등의 토큰을, L-Tokenizer로 사전 토큰화를 수행한 후 Okt Tokenizer를 사용하면 '지역사회', '사회자본', '의' 등의 토큰을 추출할 수 있다. 이는 범용적으로 사용되는 사전으로 토큰화를 수행했을 때보다 도메인 문서의 통계 정보와 범용적으로 사용되는 사전을 병합한 맞춤형 사전으로 토큰화를 수행했을 때, 단어의 의미를 잘 보존하는 형태로 추출될 수 있음을 의미한다.

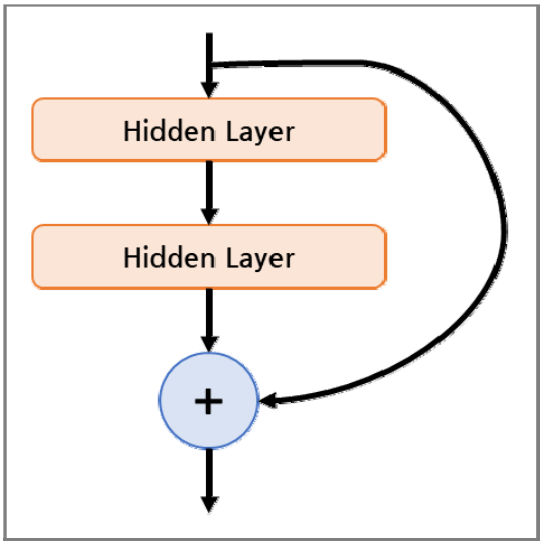


〈Figure 7〉 Example of Multi-Hot Encoding & AutoEncoder Training

3.3. 스킵 연결 오토인코더 학습

본 절에서는 <Figure 5>의 레이블 임베딩(단계 2) 과정을 소개한다. 레이블 데이터셋은 멀티-핫 인코딩을 통해 '0' 과 '1' 값으로 이루어진 고차원 레이블 공간으로 구성되며, 이러한 고차원 레이블 공간은 스킵 연결 오토인코더의 입력과 동시에 출력으로 사용된다. 본 단계에서는 스킵 연결 오토인코더를 통해 고차원 레이블 공간을 저차원 잠재 레이블 공간으로 압축하고, 저차원 잠재 레이블 공간을 다시 원래 레이블 공간으로 복원하는 레이블 임베딩 학습을 수행한다(Figure 7).

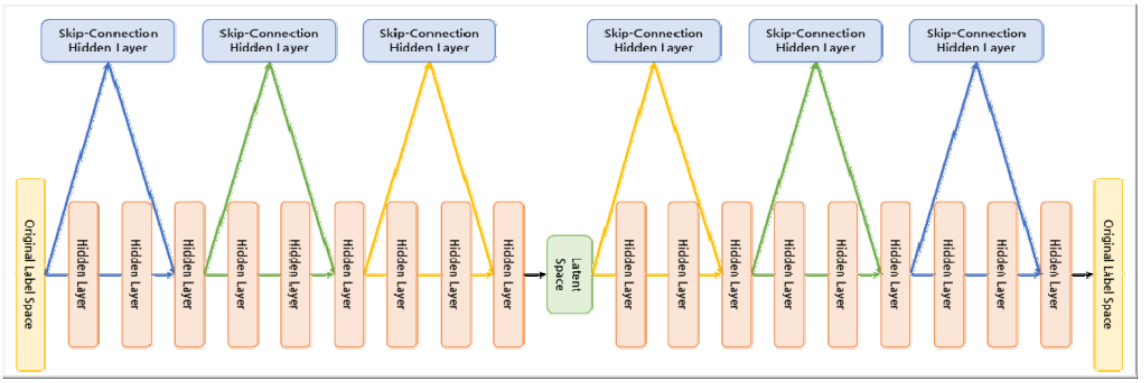
스킵 연결 오토인코더는 인코더와 디코더에 각각 3개의 스킵 연결이 추가된 형태이다. 스킵 연결은 <Figure 8>과 같이 입력을 바로 출력에 더하는 구조가 일반적이지만, 본 연구에서는 차원을 압축하고 복원하는 과정에서 은닉층 간의 차원 수를 동일하게 일치시키기 위해 스킵 연결 은닉층(Skip-Connection Hidden Layer)을 추가한 삼각 형상을 띄는 오토인코더 구조를 사용한다 (Figure 9).



〈Figure 8〉 Normal Skip-Connection

3.4. 다중 타겟 회귀 모델 학습 및 레이블 벡터 복원

본 절에서는 <Figure 5>의 다중 타겟 회귀 모델을 학습하는(단계 3) 과정과 다중 타겟 회귀 모델로부터 예측된 값을 원래 레이블 공간으로 복원하는(단계 4) 과정을 소개한다. 우선, 다중 타겟 회귀 모델 학습을 위한 데이터셋은 다음과 같다. 전처리된 문서를 입력으로, 레이블 임베딩으로 도출된 잠재 레이블 공간에 존재하는 여러 개의 연속적인 벡터를 타겟으로 설정하여 다중



〈Figure 9〉 Proposed Method Architecture

타겟 회귀 모델을 학습한다. 본 연구에서는 다중 타겟 회귀 모델로 단어 단위와 문장 단위에 어텐션 메커니즘(Attention Mechanism)(Ashnish et al., 2017)을 적용하여 학습 과정에서 문서의 중요 정보를 갖는 문장과 단어를 잘 찾아낼 수 있는 계층적 어텐션 네트워크(HAN: Hierarchical Attention Network)(Zichao et al., 2016)를 사용한다. HAN은 계층적 구조로 인해 문장으로 이루어진 문서의 벡터를 추출하는데 효과적이며, 본 연구에서는 HAN으로 추출된 문서 벡터로 잠재 레이블 공간의 벡터 값을 예측하는 다중 타겟 회귀 학습을 수행한다. 이후, (단계 2)에서 학습된 스킵 연결 오토인코더의 디코더를 사용하여 다중 타겟 회귀 모델의 예측 값, 즉 저차원 레이블을 원래 레이블 공간으로 복원한다. 최종적으로 디코더를 통해 복원된 레이블을 원래 레이블과 비교함으로써, 다중 레이블 텍스트 분류 문제에 대한 레이블 임베딩 기법의 성능을 평가할 수 있다.

4. 실험

4.1. 실험 개요

본 장에서는 3장에서 소개한 제안 방법론을 실제 데이터에 적용한 실험 결과를 분석하고, 이를 통해 제안 방법론의 성능 및 활용성을 논의한다. 실험에 사용된 데이터는 학술연구정보서비스인 'RISS'에서 수집한 논문으로, 국내 학술논문 중 2010년부터 2021년까지 한국어로 게재된 KCI 등재 혹은 KCI 우수 등재지 논문을 사용하였다. 성능 평가는 2020년 '공학 분야'에서 '많이 연구된 키워드' 상위 10개를 기반으로 수집한 공학 분야 논문을, 도메인 특성에 따른 제안 방법론의 효과를 파악하는 실험은 2020년 '전체 분야'에서 '많이 연구된 키워드' 상위 10개를 기반으로 수집한 논문을 사용하였다. 실험 환경은 Python 3.7 그리고 딥러닝 프레임워크인 Tensorflow-gpu 2.2와 Keras 2.3을 기반으로 구축하였다.

4.2. 원문 전처리 및 키워드 레이블 임베딩

우선 공학 분야의 논문에서 초록과 키워드를

〈Table 1〉 Examples of Preprocessed Abstracts

Abstracts	Tokens		
1	본	논문에서는	최신
	기계	…	평가
	모델을	개발	하였다
2	지능	정보사회의	도래
	로	…	될
	것으로	전망	된다
3	최근	빅데이터	분석과
	인공지능을	…	창출
	시스템으로	등장	하였다

〈Table 2〉 Examples of Embedded Tokens

Tokens	100 Dimension Word Embedding						
본	1.0883	-1.4512	-1.2328	…	5.6489	1.2017	-1.9599
논문에서는	0.0742	-0.1388	-0.0322	…	0.1563	0.0973	-0.0622
최신	3.1023	-1.1058	-3.4850	…	-2.7761	-6.7275	-2.7064
…	…						
모델을	0.2516	-0.4584	-0.1348	…	-0.0030	-0.3063	0.2810
개발	-0.2477	-0.1309	-0.0358	…	-0.0611	0.2857	-0.6031
하였다	-0.1864	0.0459	-0.6500	…	-0.1302	0.0433	0.0805

추출하여 전처리를 진행하였다. 본 연구에서는 5개 이상의 문장을 갖는 초록만 사용하였고, 10번 미만으로 출현한 키워드는 제거하였다. 이러한 필터링 작업을 통해 최초 수집한 24,433건의 논문 중 4,675건을 추출하였으며, 선정된 논문에 태그된 키워드는 262개가 도출되었다. 선정된 논문에 대해 L-Tokenizer로 사전 토큰화를 수행한 후, 추출된 토큰들을 기존의 Okt Tokenizer 사전

〈Table 3〉 Examples of Original Label Space

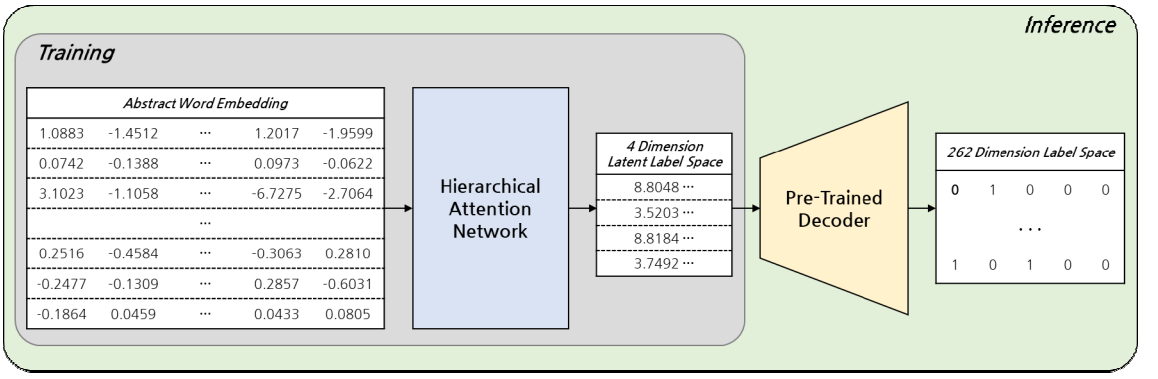
Labels						
드론	교육	디자인 씽킹	…	딥러닝	라즈베리 파이	기계학습
군집분석	과학교육	비지도 학습	…	빅데이터	사물 인터넷	사용자 경험
인공지능	블록체인	비트코인	…	인공지능 로봇	이미지	인공지능 윤리
Multi-Hot Encoding						
0	1	0	…	0	0	0
0	0	0	…	1	0	0
1	0	0	…	0	0	0

〈Table 4〉 Examples of Latent Label Space

Abstracts	4 Dimension Latent Label Space			
1	8.8048 …	3.5203 …	8.8184 …	3.7492 …
2	2.5632 …	2.5123 …	1.5131 …	2.3629 …
3	0.8704 …	3.1267 …	1.6460 …	0.8774 …

에 추가하여 논문 데이터에 특화된 사전으로 토큰화를 수행하였다. 마지막으로 외국어, 특수기호, 그리고 조사 등을 불용어로 처리하여 최종 토큰들을 선정하였다. 추가로 본 연구에서는 선정된 토큰들을 모델에 입력 가능한 형태로 변환하기 위해 Word2Vec(Tomas et al., 2013) 모델을 사용하였으며, 토큰 하나당 100차원으로 구조화하였다. <Table 1>은 임의의 초록에 대한 토큰화 결과의 예이며, <Table 2>는 <Table 1>의 1번 초록의 토큰들을 Word2Vec을 사용하여 구조화한 결과를 보여준다.

초록 전처리 이후, 문서에 태그된 262개의 키워드는 <Table 3>과 같이 멀티-핫 인코딩으로 '0'과 '1' 값을 갖는 이진 형태의 262차원 레이블 공간으로 구성된다. 이렇게 만들어진 262차원의 고차원 레이블 공간을 스킵 연결 오토인코더로 4차원의 저차원 잠재 레이블 공간으로 압축하고, 압축된 잠재 레이블 공간으로부터 원래 레이블



〈Figure 10〉 An Example of Training and Inference for Multi-Label Text Classification

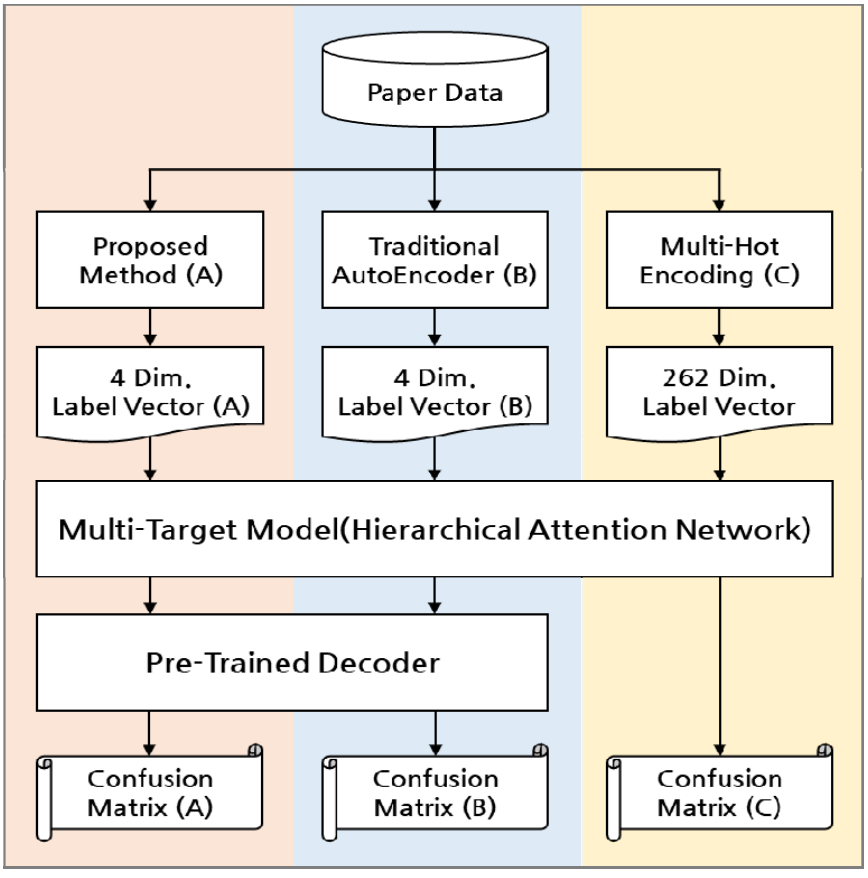
공간으로 복원하는 레이블 임베딩 학습을 수행하였다. 즉, 262차원의 레이블을 4차원으로 압축하는 레이블 임베딩 학습을 통해, 최종적으로 초록 한 건 당 압축된 4차원의 잠재 레이블 공간을 갖는 데이터셋을 구축하였다. <Table 4>는 압축된 4차원 잠재 레이블 공간의 벡터들을 보여준다.

4.3. 다중 타겟 회귀 모델 학습

다음으로 초록으로부터 잠재 레이블 공간의 벡터들을 예측하는 다중 타겟 모델 학습 실험 결과를 소개한다. 다중 타겟 모델은 계층적으로 어텐션 메커니즘이 적용되어 문서 벡터를 추출하는데 효과적인 계층적 어텐션 네트워크(HAN: Hierarchical Attention Network)(Zichao et al., 2016)를 사용하였으며, 초록으로부터 <Table 4>의 4개의 벡터를 예측하는 회귀 학습을 수행하였다. 이렇게 학습된 HAN 모델의 예측 값을 사전 학습된 스킵 연결 오토인코더의 디코더를 활용하여 원래 레이블 공간인 262차원으로 복원하였다. 마지막으로 복원된 레이블을 <Table 3>의 원래 레이블과 비교하여 다중 레이블 텍스트 분류를 수행하였다. <Figure 10>은 학습 데이터의 한 편의 초록으로부터 잠재 레이블 공간의 벡터를 예측하도록 HAN 모델을 학습한 Training 결과와 사전 학습된 디코더로 잠재 레이블 공간의 예측 값을 원래 공간의 레이블로 복원하는 Inference 결과를 나타낸다.

4.4. 분류 정확도 평가

본 절에서는 제안 방법론과 기존 오토인코더 기반 레이블 임베딩 모델, 그리고 레이블 임베딩을 사용하지 않은 모델의 분류 정확도를 분석한 결과를 소개한다. 실험의 전체 과정은 <Figure 11>과 같다. <Figure 11(a)>와 <Figure 11(b)>는 각각 제안 방법론과 전통적인 오토인코더 기반 레이블 임베딩을 통해 도출된 잠재 레이블 공간의 벡터를 타겟으로 하는 다중 타겟 모델 학습 후, 학습된 모델의 예측 값을 사전 학습된 제안 방법론의 디코더로 복원하여 다중 레이블 텍스트 분류 결과를 평가하는 과정이다. <Figure 11(c)>는 레이블 임베딩 없이 멀티-핫 인코딩을 통해 도출된 262차원의 이진 값을 가진 레이블 공간을 예측하는 방법을 나타낸다.

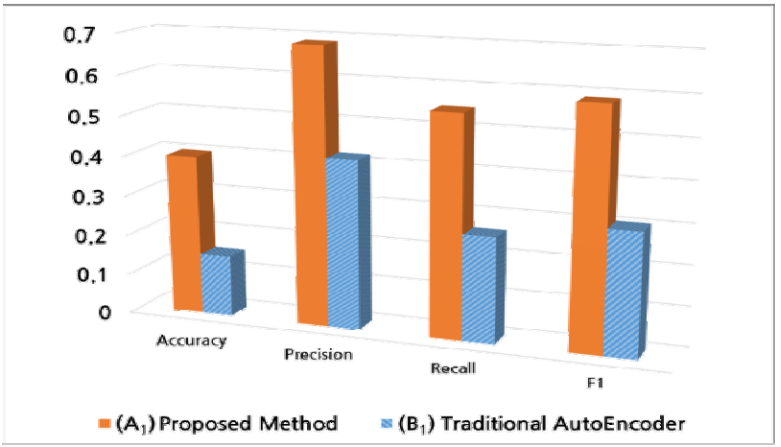


〈Figure 11〉 Overall Process of Performance Evaluation

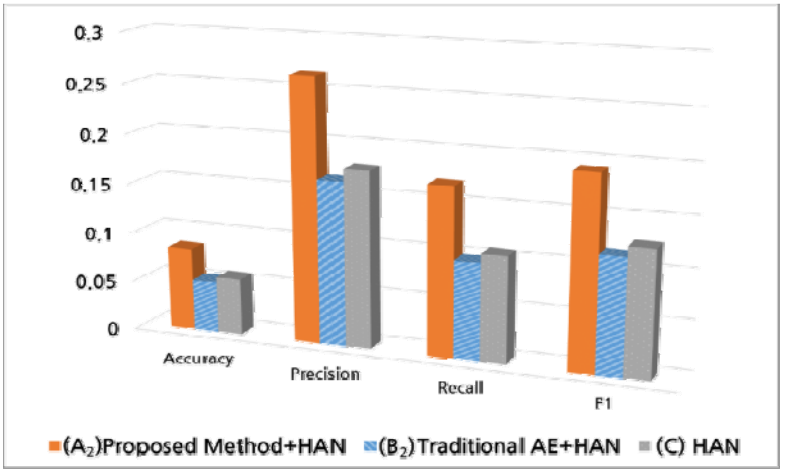
실험을 위해 전체 4,675건의 논문을 학습용 2,805건과 검증용 935건, 그리고 테스트용 935건으로 분할하였으며, 레이블 임베딩과 다중 레이블 텍스트 분류에 대한 평가 지표로는 혼동 행렬(Confusion Matrix)을 사용하였다. 실험 결과는 <Table 5>에 요약되어 있으며, 레이블 임베딩에 대한 성능 평가와 다중 레이블 텍스트 분류에 대한 성능 평가는 각각 <Figure 12>와 <Figure 13>에 나타나있다. <Figure 12>는 제안 방법론에 따른 레이블 임베딩과 기존의 오토인코더 기반 레이블 임베딩의 성능을 나타낸 것으로, 정확도와 정밀도, 재현율 그리고 F1 점수 모두 제안 방법론이 우수한 성능을 보임을 확인할 수 있다. 또한, <Figure 13>은 제안 방법론에 따른 레이블 임베딩, 기존의 오토인코더 기반 레이블 임베딩, 그리고 레이블 임베딩을 적용하지 않은 세 가지 모델로 다중 레이블 텍스트 분류를 수행한 성능을 나타낸 것으로, 역시 정확도, 정밀도, 재현율 그리고 F1 점수 모두 제안 방법론이 가장 우수한 성능을 나타냄을 확인하였다.

〈Table 5〉 Results of Performance

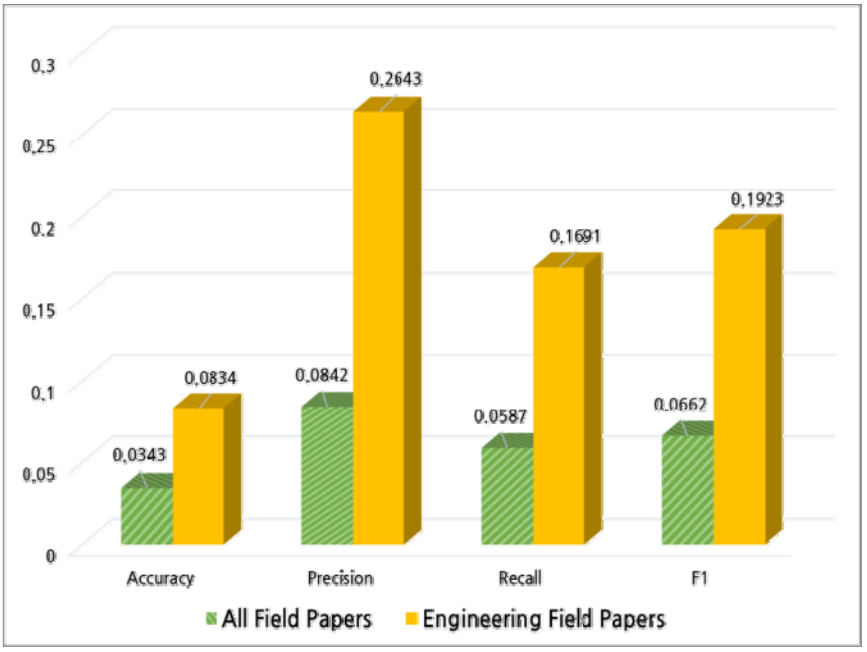
	(A_1) Proposed Method	(A_2) Proposed Method + HAN	(B_1) Traditional AutoEncoder	(B_2) Traditional AutoEncoder + HAN	(C) Multi-Hot Encoding + HAN
Accuracy	0.3989	0.0834	0.1519	0.0513	0.0567
Precision	0.6874	0.2643	0.4193	0.1647	0.1775
Recall	0.5480	0.1691	0.2612	0.0974	0.1064
F1	0.5887	0.1923	0.3050	0.1160	0.1263



〈Figure 12〉 Performance of Label Embedding



〈Figure 13〉 Performance of Multi-Label Text Classification



〈Figure 14〉 Performance of Multi-Label Text Classification Regarding Domain Characteristics

4.5. 제안 방법론의 활용성

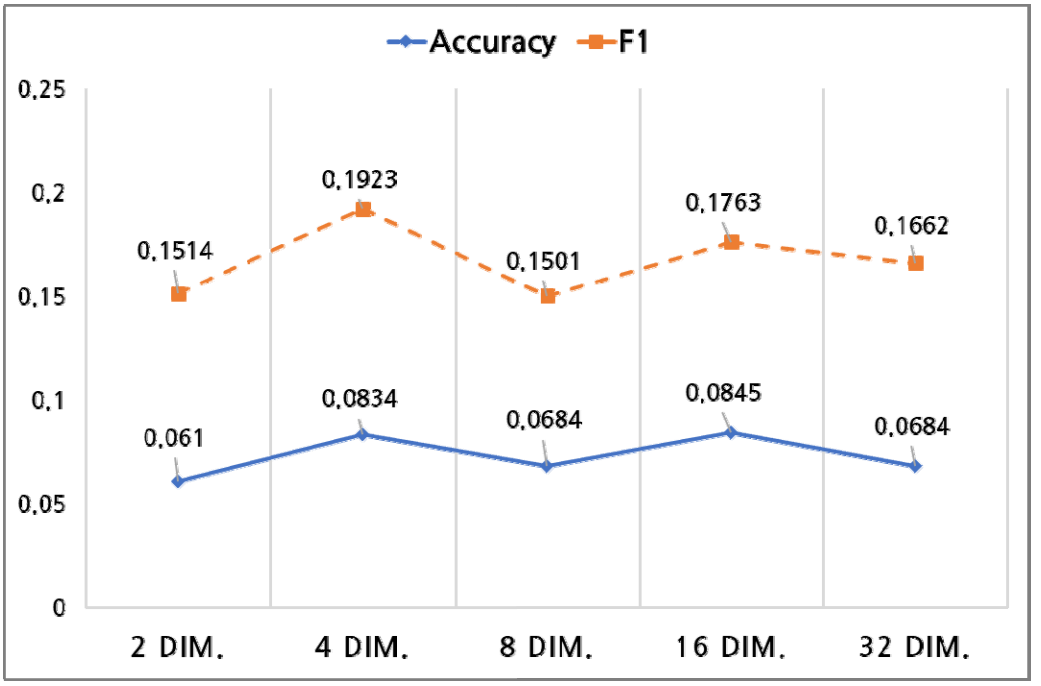
4.5.1. 도메인 특성에 따른 성능 비교

본 절에서는 다른 방법들에 비해 우수한 성능을 보인 제안 방법론을 다중 레이블 텍스트 분류에 적용할 때, 적용 대상 도메인의 특성이 분류 성능에 어떠한 영향을 주는지 분석한 실험 결과를 소개한다. 이러한 목적의 실험을 위해 두 가지 실험 세트, 즉 유사한 레이블들을 갖는 단일 분야 논문으로 구성된 실험 세트와 다양한 레이블들을 갖는 복합 분야 논문으로 구성된 실험 세트를 구성하였다. 단일 분야 분류 성능의 경우 앞 절의 성능 평가에서 확인한 공학 분야 논문에 대한 성능 평가 결과를 그대로 사용하였다. 또한 복합 분야 분류 성능 평가를 위해 전체 분야의 논문을 새로 수집하여 앞에서와 동일한 방식의 분류 실험을 수행하였다. <Figure 14>는 제안 방법론으로 공학 분야와 전체 분야 논문에 대해 다중 레이블 텍스트 분류를 수행한 결과를 보인다.

실험 결과 전체 분야에 비해 공학 분야, 즉 특정 분야 세트에서의 분류 성능이 모든 측면에서 우수하게 나타남을 확인할 수 있으며, 이는 제안 방법론을 유사한 도메인을 갖는 데이터의 다중 레이블 텍스트 분류에 적용했을 때 가장 우수한 성능을 나타내는 것으로 해석될 수 있다.

4.5.2. 잠재 레이블 공간 최적 차원 수 탐색

본 부절에서는 압축한 결과 차원의 수가 분류 성능에 미치는 영향을 분석한 추가 실험 결과를 소개한다. 최적의 차원 수 탐색은 어떤 차원을



〈Figure 15〉 Accuracy and F1-Score in Accordance with Number of Latent Label Space Dimensions

선택할 것인지 결정하기 위해 진행한 실험으로, 제안 방법론을 수행하는데 있어 보조적인 지표이다. 실험 결과 차원 수의 증가에 따라 정확도가 꾸준히 증가하거나 감소하지는 않았으며, 본 실험에서는 제안 방법론을 4차원으로 압축하고 복원하는 학습을 수행했을 때의 분류 성능이 가장 우수한 것으로 나타났다(Figure 15).

5. 결론

다수의 레이블을 갖는 다중 레이블 분류는 하나의 레이블만 갖는 이진 분류 혹은 다중 클래스 분류보다 현실 세계에서 자주 등장하는 문제이다. 이러한 다중 레이블 분류는 일반적인 분류 문제와는 상이한 방식의 학습을 요구하며, 특히 레이블과 클래스의 개수가 많아질수록 예측 난이도가 급격하게 증가한다는 점에서 분류의 난제 중 하나로 인식되고 있다. 이러한 어려움을 극복하기 위해 다수의 레이블을 압축하여 압축된 레이블을 타겟으로 하는 분류 모델을 학습한 후, 학습된 모델을 통해 추론한 예측 값을 원래의 레이블로 복원하는 레이블 임베딩 기법이 주목을 받고 있다.

본 연구에서는 고차원의 레이블 공간 정보를 충분히 반영한 저차원 잠재 레이블 공간을 형성하기 위해, 각각의 인코더와 디코더에 스킵 연결을 추가한 오토인코더 기반 레이블 임베딩 방법론을 새롭게 제안하였다. 또한, 실제 논문 초록 및 키워드를 사용하여 실험을 수행한 결과, 제안

방법론에 따른 레이블 임베딩 모델이 기존의 레이블 임베딩 모델에 비해 분류 정확도 측면에서 우수한 성능을 보임을 확인하였다. 제안 방법론은 기울기 소실 문제를 해결하기 위해 사용되어 왔던 스킵 연결 구조를 오토인코더에 추가하여 레이블 임베딩 과정에서 발생하는 정보 손실을 해결하였기 때문에, 전통적인 오토인코더 기반 레이블 임베딩에 비해 우수한 성능을 나타낸다.

본 연구는 인코더와 디코더 각각에 스킵 연결을 추가한 오토인코더 구조를 제안하고, 이를 레이블 임베딩에 활용하는 방안을 새롭게 제안했다는 점에서 학술적 기여를 인정받을 수 있을 것이다. 또한, 다중 레이블 분류 문제를 해결하기 위한 레이블 임베딩에 관한 연구가 부족한 국내에서, 국내 데이터를 사용하여 다중 레이블 분류의 성능 향상을 보였다는 점은 실무적 측면의 기여로 인정받을 수 있을 것이다. 하지만 레이블 임베딩을 활용한 다중 레이블 분류는 레이블 임베딩과 분류기가 각각 학습되기 때문에 End-to-End 학습이 어렵다는 단점이 있다. 또한 전통적인 오토인코더 기반 레이블 임베딩은 고차원 레이블 공간을 압축하고 복원할 때, 정보 손실이 크게 발생한다는 한계가 있다. 본 논문에서는 이러한 한계를 극복하기 위해 스킵 연결을 추가한 오토인코더 기반 레이블 임베딩 방법론을 제안하였지만, 제안 방법론을 텍스트 데이터의 분류에 적용한 실험 결과만을 분석하였다. 추후 연구에서는 정형 데이터를 포함한 다양한 유형의 데이터에 대한 제안 방법론의 성능 평가가 반드시 다루어져야 한다.

참고문헌(References)

Ashish, K., P. Jain, and R. Viswanathan, “Multilabel Classification using Bayesian Compressed Sensing,” Advances in Neural Information Processing Systems 25, 2012.

Ashish, V., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” arXiv:1706.03762, (2017).

Bingyu, W., L. Chen, W. Sun, K. Qin, K. Li, and H. Zhou, “Ranking-Based Autoencoder for Extreme Multi-label Classification,” arXiv: 1904.05937, (2019).

Chih-Kuan, Y., W-C. Wu, W-J. Ko, and Y-C. F. Wang, “Learning Deep Latent Space for Multi-Label Classification,” Thirty-First AAAI Conference on Artificial Intelligence, Vol.31, No.1(2017).

Denis, L., A. Aussema, and M. Gasse, “On the use of binary stochastic autoencoders for multi-label classification under the zero-one loss,” Procedia Computer Science, Vol.144, (2018), 71~80.

Farbound, T. and H-T. Lin, “Multilabel Classification with Principal Label Space Transformation,” Neural Computation Vol.24, No.9(2012), 2508~2542.

Ganda, D. and R. Buch, “A Survey on Multi Label Classification,” Recent Trends in Programming Languages, Vol.5, No.1(2018), 19~23.

Jo, I. S., Y. H. Kang, D. B. Choi, and Y. B. Park, “Clustering Performance Analysis of Autoencoder with Skip Connection,” KIPS Transactions on Software and Data Engineering, Vol.9, No.12 (2020), 403~410.

Jörg, W., A. Tyukin, and S. Kramer, "A Nonlinear Label Compression and Transformation Method for Multi-label Classification Using Autoencoders," Advances in Knowledge Discovery and Data Mining, (2016), 328~340.

Jörg, W., B. Pfahringer, and S. Kramer, "Multi-label classification using boolean matrix decomposition," Proceedings of the 27th Annual ACM Symposium on Applied Computing, (2012), 179~186

Kaiming, H., X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2016), 770~778.

Pascal, V., H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," Proceedings of the 25th international conference on Machine learning, (2008), 1096~1103.

Pascal, V., H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," Journal of Machine Learning Research, Vol.11, (2010), 3371~3408.

Pierre, B. and K. Hornik, "Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima," Neural Networks, Vol.2, (1989), 53~58.

Tomas M., K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781, (2013).

Yoshua, B., P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-Wise Training of Deep Networks," Advances in Neural Information Processing Systems 19, 2007.

Zichao Y., D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (2016), 1480~1489.

Abstract

Label Embedding for Improving Classification Accuracy UsingAutoEncoderwithSkip-Connections

Museong Kim* · Namgyu Kim**

Recently, with the development of deep learning technology, research on unstructured data analysis is being actively conducted, and it is showing remarkable results in various fields such as classification, summary, and generation. Among various text analysis fields, text classification is the most widely used technology in academia and industry. Text classification includes binary class classification with one label among two classes, multi-class classification with one label among several classes, and multi-label classification with multiple labels among several classes. In particular, multi-label classification requires a different training method from binary class classification and multi-class classification because of the characteristic of having multiple labels. In addition, since the number of labels to be predicted increases as the number of labels and classes increases, there is a limitation in that performance improvement is difficult due to an increase in prediction difficulty.

To overcome these limitations, (i) compressing the initially given high-dimensional label space into a low-dimensional latent label space, (ii) after performing training to predict the compressed label, (iii) restoring the predicted label to the high-dimensional original label space, research on label embedding is being actively conducted. Typical label embedding techniques include Principal Label Space Transformation (PLST), Multi-Label Classification via Boolean Matrix Decomposition (MLC-BMaD), and Bayesian Multi-Label Compressed Sensing (BML-CS). However, since these techniques consider only the linear relationship between labels or compress the labels by random transformation, it is difficult to understand the non-linear relationship between labels, so there is a limitation in that it is not possible to create a latent label space sufficiently containing the information of the original label. Recently, there have been increasing attempts to improve performance by applying deep learning technology to label embedding. Label embedding using an autoencoder, a deep learning model that is effective for data compression and restoration, is

* Graduate School of Business IT, Kookmin University
** Corresponding author: Namgyu Kim
Graduate School of Business IT, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Korea
Tel: +82-2-910-5425, Fax: +82-2-910-4017, E-mail: ngkim@kookmin.ac.kr

representative. However, the traditional autoencoder-based label embedding has a limitation in that a large amount of information loss occurs when compressing a high-dimensional label space having a myriad of classes into a low-dimensional latent label space.

This can be found in the gradient loss problem that occurs in the backpropagation process of learning. To solve this problem, skip connection was devised, and by adding the input of the layer to the output to prevent gradient loss during backpropagation, efficient learning is possible even when the layer is deep. Skip connection is mainly used for image feature extraction in convolutional neural networks, but studies using skip connection in autoencoder or label embedding process are still lacking.

Therefore, in this study, we propose an autoencoder-based label embedding methodology in which skip connections are added to each of the encoder and decoder to form a low-dimensional latent label space that reflects the information of the high-dimensional label space well. In addition, the proposed methodology was applied to actual paper keywords to derive the high-dimensional keyword label space and the low-dimensional latent label space. Using this, we conducted an experiment to predict the compressed keyword vector existing in the latent label space from the paper abstract and to evaluate the multi-label classification by restoring the predicted keyword vector back to the original label space.

As a result, the accuracy, precision, recall, and F1 score used as performance indicators showed far superior performance in multi-label classification based on the proposed methodology compared to traditional multi-label classification methods. This can be seen that the low-dimensional latent label space derived through the proposed methodology well reflected the information of the high-dimensional label space, which ultimately led to the improvement of the performance of the multi-label classification itself. In addition, the utility of the proposed methodology was identified by comparing the performance of the proposed methodology according to the domain characteristics and the number of dimensions of the latent label space.

Key Words : Deep Learning, Multi-Label Classification, Label Embedding, AutoEncoder, Skip-Connection

Received : June 9, 2021 Revised : June 28, 2021 Accepted : July 19, 2021
Corresponding Author : Namgyu Kim

저 자 소 개

김 무 성
현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학 중이며, 국민대학교 경영정보학부에서 학사 학위를 취득하였다. 한국지능정보시스템학회 학술대회 최우수 논문상, 데이콘 인공지능 비트 트레이더 경진대회, 교통데이터 활용 공모전 한국교통연구원장상, 포스트코로나 AI 챌린지 KT사장상 등을 수상하였다. 주요 관심분야는 자연어처리, 이미지처리, 멀티모달 학습, 딥 러닝 등이다.

김 남 규
현재 국민대학교 비즈니스IT전문대학원장 및 경영정보학부 교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국지능정보시스템학회 부회장, 한국정보기술응용학회 부회장, 한국경영학회 상임이사, 한국경영정보학회 이사, 한국인터넷정보학회 이사를 역임하였다. 주요 관심분야는 텍스트 마이닝, 딥러닝, 데이터 모델링 등이다.