Bio-BERT와 GCN을 결합한 텍스트 분류 모델

오민해, 이정우*
서울대학교, *서울대학교

minhae.oh@cml.snu.ac.kr, *junglee@snu.ac.kr

Combining Bio-BERT and GCN for text classification in clinical domain

Minhae Oh, Jungwoo Lee*
Seoul National Univ., *Seoul National Univ.

요 약

본논문에서는 텍스트 분류를 위해 생체의학분야의 텍스트 대규모 사전 학습 모델인 Bio-BERT와 GCN (Graph convolutional network)을 결합하였다. Bio-BERT는 BERT (Bidirectional Encoder Representations from Transfromers)가 전문분야 용어에 취약하다는 단점을 보강하고자 나온 모델로 생체의학분야 텍스트 분류 모델에 효과적인 결과를 낸다. 그래프 합성곱 신경망은 딥러닝 모델의 일종으로 그래프 형태의 데이터를 인풋으로 받아 여러 태스크에서 좋은 성능을 낸다. 이 두 모델을 결합하여 의료분야 텍스트 분류에 효과적인 성능을 낼 수 있는 새로운 모델을 제시한다.

Ⅰ. 서 론

텍스트 분류는 자연어처리 분야의 대표적인 과제로 감정 분석, 스팸메세지 분류 및 주제 분류 과제를 포함한다. 분류된 데이터와 분류되지 않은 데이터 모두를 사용하는 Transductive learning은 텍스트 분류 과제에서 좋은 결과를 내는 대표적인 학습방법이다. GNN (Graph neural network)은 Transductive learning의 일종으로 텍스트의 단어, 문장 또는 문단을 그래프 형태로 변환하여 학습을 진행하게 된다.

GNNs (Graph neural networks) 학습을 위해 만들어지는 그래프 데이터는 node와 edge로 구성된다. Node는 단어, 문장 또는 문단의 특징을 담은 행렬로 표현되며, edge는 node사이의 관계를 나타내는 척도로 사용된다. GNNs의 종류에는 GCN (Graph convolutional networks), GAN (Graph attention networks), graph auto-encoder, graph generative networks 등이 있다.

Bio-BERT[1]는 BERT[2] (Bidirectional Encoder Representations from Transformers)를 기반으로 한 사전 학습 언어 모델로 생체의학분야의 텍스트에 특화되도록 사전 학습된 언어모델이다. Bio-BERT는 생체의학 분야의 엔터티 인식, 관계 추출, 텍스트 분류를 포함한 다양한 태스크에서 높은 성능을 보임을 확인할 수 있다. 본연구에서는 GCN과 Bio-BERT와 결합하여 의료분야 텍스트 분류를 위한 모델인 BioBertGCN을 제시한다.

Ⅱ. 본론

1. 데이터 셋

본 논문에서는 학습을 위한 데이터 셋으로 Ohsumed 데이터 셋을 사용한다. Ohsumed 데이터 셋은 MEDLINE 데이터베이스 속한 데이터로, National Library of Medicine의 메디컬 문헌들의 초록으로 이루어져 있다. 본 연구에서는 13,929개의 심장혈관계 질병에 대한 초록

중 1개의 질병으로 분류되어 있는 데이터 7400개를 사용하였다. 총 7400개의 데이터는 3357개의 학습 데이터와 4043개의 테스트 데이터로 분리되어 사용하였다.

2. 실험방법

BioBertGCN 모델은 Bert와 GCN을 결합한 모델인 BertGCN [3] 모델을 기반으로 한 텍스트 분류 모델이다. 우선 GCN의 인풋으로 사용하도록 데이터를 그래프 형태로 변환하는 작업을 진행한다. 그 후 그래프의 node를 Bio-BERT를 사용하여 학습을 진행시킨다. 그래프의 edge는 단어와 문헌 또는 단어와 단어 사이의 등장 빈도수를 계산하는 term frequency-inverse document frequency (TF-IDF)와 positive point-wise mutual information (PPMI) 함수로 학습된다. 이후 학습된 node와 edge로 구성된 graph 형태의 데이터를 GCN의 인풋으로 넣어준 후 학습을 진행한다.

Ⅲ. 실험 결과

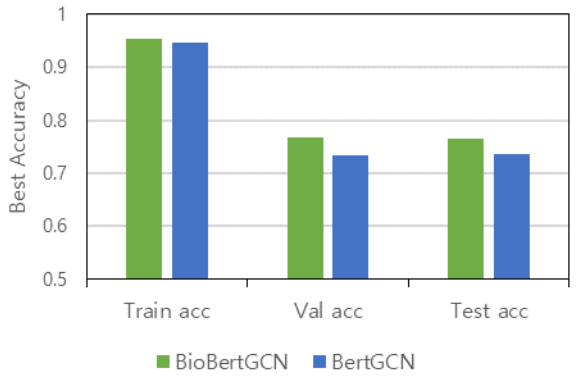

그림 1 BioBertGCN 과 BertGCN 의 결과비교

.

[그림1]은 메디컬 데이터인 Ohsumed 데이터셋을 BioBertGCN로 학습한 결과와 BertGCN로 학습한 결과를 비교하고 있다. Train accuracy, Validation accuracy, Test accuracy 모두 BioBertGCN이 더 좋은 정확도를 내고 있음을 확인할 수 있다.

Ⅳ 결론

본 논문에서는 의료분야의 텍스트 분류 태스크를 위해 Bio-Bert와 GNN을 결합한 BioBertGCN 모델을 제시하였다. 기존의 Bert와 GCN 모델 또는 그 둘을 결합한 BertGCN 모델과 비교하여 의료분야 데이터에서 학습이 더 잘 되는 것을 실험을 통해 확인하였다. 타 도메인의 텍스트 또한 도메인 특화된 사전학습 모델을 사용하면 성능을 향상시킬 수 있을 것이다.

ACKNOWLEDGMENT

This work is in part supported by National Research Foundation of Korea (NRF, 2021R1A2C2014504(30)), Institute of Information & communications Technology Planning & Evaluation (IITP- 2021-0-00106(40),IITP-2021-0-02068(40)) grant funded by the Ministry of Science and ICT (MSIT), INMAC, and BK21-plus

참 고 문 헌

[1] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding

[3] Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, Fei Wu. 2021 BertGCN : Transductive Text Classification by Combining GCN and BERT.