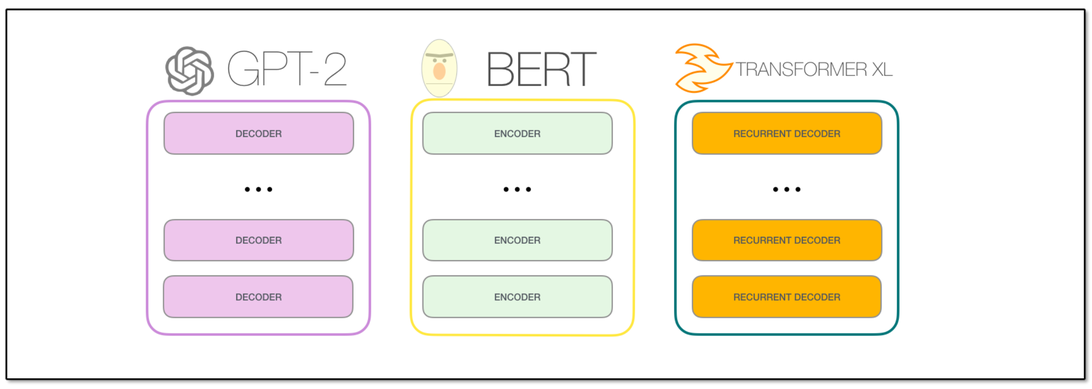
GMLP를 이용한 한국어 자연어처리 및 BERT와 정량적 비교

Cognitive Computing Lab.

이성민, 나승훈
전북대학교
cap1232@jbnu.ac.kr, nash@jbnu.ac.kr

Ⅰ. 서론

한국어 자연어처리는 대부분 Transformer 구조에 기반한 언어 모델을 사용하여 이루어지고 있다. Transformer구조는 Multi-head Attention을 통해 양방향의 문맥을 반영한 표상을 잘 얻어낼 수 있다는 것이 주요한특징이다. 하지만, 최근 Multi-Head Attention 구조 없이 MLP만으로 높은 성능을 내려는 시도들이 계속되고 있다.



Ⅱ. 한국어 GMLP + Tiny Attention 모델

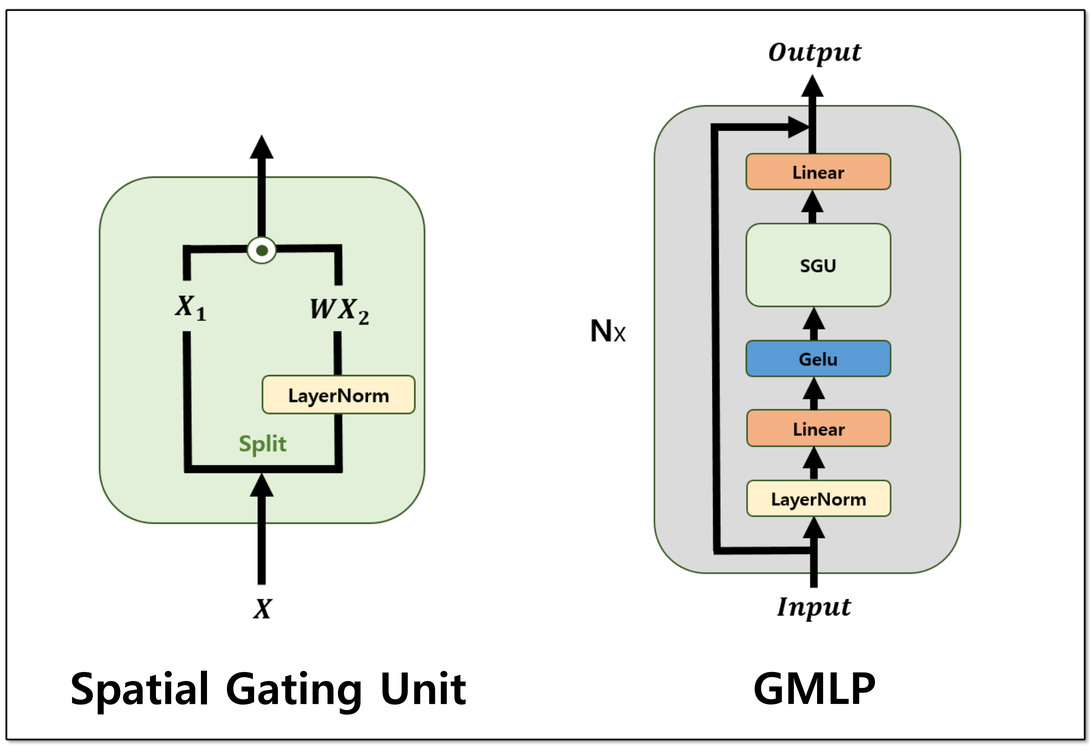
Spatial Gating Unit(SGU)

Spatial Gating Unit은 MLP와 Gating으로 구성되어 있는 간단한 구조이다. 가중치 행렬(W)을 텍스트 임베딩(X)에 곱하여 텍스트 토큰들 사이의 Spatial한 정보를 얻을 수 있다는 것이 가장 큰 특징이다.

$$SGU(X) = X_1 \odot LayerNorm(WX_2 + b)$$

GMLP

GMLP는 BERT의 Multi-Head Attention을 SGU로 바꾼 것 외에는 거의 동일하다. 한가지 다른 점은 SGU에서 Spatial한 정보를 얻기 때문에 입력에서 Position 임베딩을 필요로 하지 않는다는 점이다. 또한 Multi-Head Attention과 다르게 가중치 행렬의 크기가 사전학습시 토큰 길이에 영향을 받기 때문에 미세조정 시에는 짧아지는 길이만큼 가중치 행렬을 잘라서 사용하게 된다는 점이 있다.



한국어 GMLP + Tiny Attention 모델

GMLP만으로도 괜찮은 성능을 낼 수 있지만 Attention을 추가하면 더 나은 성능을 얻을 수 있으므로 사전학습에 작은 Attention 층을 추가하여 SGU의 기능을 강화하여 사전학습을 진행했다. 데이터셋은 위키피디아, 모두의 말뭉치 뉴스 데이터를 이용했고, 토크나이저는 형태소+자소단위 BPE 토크나이저를 사용하였다. 모델의 레이어 개수는 36, 히든사이즈는 512, 최대 문서 길이는 512, 어텐션 차원은 64으로 설정하였다. 학습은 FULL-SENTENCES, 동적 마스킹, NSP를 제외한 MLM으로 진행하였으며, 학습률 1e-4, 옵티마이저 AdamW, 56 배치사이즈로 90만 스텝 사전학습을 진행하였다.

Ⅲ. GMLP를 이용한 한국어 자연어처리

한국어 감성분석

감성분석은 입력 텍스트가 어떤 감정을 가지고 있는지 분류하는 텍스트 분류 테스크이다. 데이터셋은 네이버 영화 평점 데이터셋을 사용하였다. 긍정, 부정 레이블 분류를 위해 출력 임베딩 중 첫번째 임베딩에 레이어를 추가하여 학습을 진행하였다.

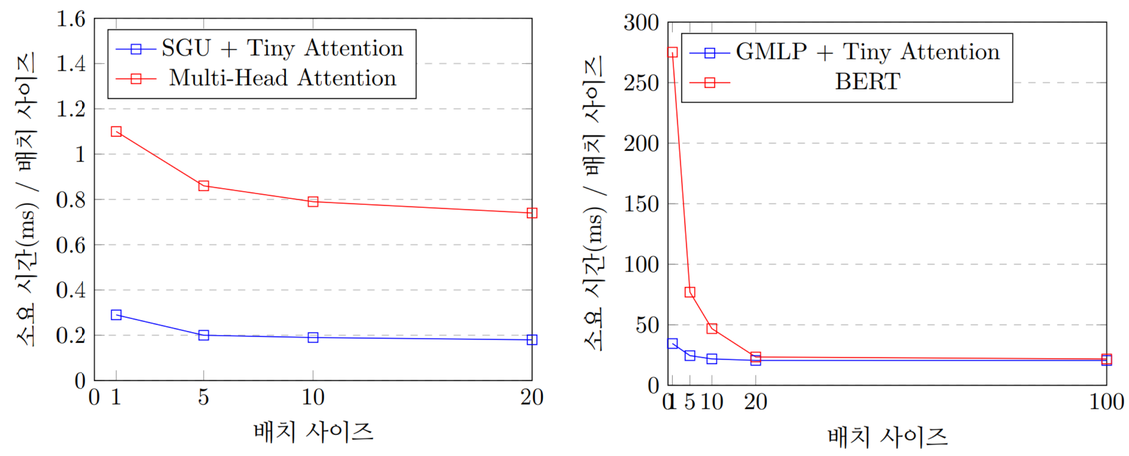
모델	Accuracy
LSTM	79.79
BERT(형태소-태그)	86.57
BERT(Multilingual)	87.43
GMLP + Tiny Attention	**87.70**
RoBERTa	89.88

한국어 개체명인식

개체명인식은 입력 텍스트에서 인물, 시간, 장소 등 의미를 가지고 있는 개체명을 인식하는 시퀀스 레이블링 테스크이다. 데이터셋은 네이버 NER 데이터셋을 사용하였다. 출력 임베딩에 토큰레이블 수를 고려한 레이어를 추가하여 학습을 진행하였다.

모델	F1 Score
CNN-BiLSTM-CRF	74.57
BERT(Multilingual)	84.20
GMLP + Tiny Attention	**85.82**
KoBERT	86.11
RoBERTa	87.58

Ⅳ. GMLP와 BERT의 추론 속도 비교 실험



GMLP + Tiny Attention과 BERT 추론 속도를 비교했을 때, 배치사이즈가 20보다 작을 때는 GMLP + Tiny Attention이 약 1에서 6배 가량 빨랐고, 그보다 클 경우에는 속도 차이가 미미해 지는 것을 알 수 있다. 이에 따라 GMLP가 적은 배치사이즈에서는 추론 속도가 BERT보다 빠른 것을 알 수 있다.

Ⅴ. 결론 및 향후 연구

GMLP가 기존 Transformer Encoder의 Multi-head Attention 없이 SGU와 작은 Attention 만으로도 BERT와 견줄만한 성능을 보일 수 있음을 확인할 수 있었다. 또한 배치사이즈를 20보다 작게 하여 추론 시 BERT 보다 추론 속도가 빠른 것을 실험을 통해 확인할 수 있었다. 하지만, 미세조정 성능 평가 결과를 보면 RoBERTa, KoBERT에 못 미치는 걸로 보아 SGU가 Multi-Head Attention을 완전히 대체하기는 힘들어 보인다. 따라서 향후에는 지금껏 많은 모델들에서 잘 작동해왔던 Multi-Head Attention을 사용하지 않는 것보다 잘 활용하면서 SGU의 장점을 더해 새로운 구조의 모델을 구성하는 연구를 진행할 예정이다.