[04/23] Wrap up Report

기술적인 도전

본인의 점수 및 순위 : LB 점수 79.9% 48등

검증(Validation) 전략

1. 주어진 dataset에서 10%를 validation dataset으로 만들어 빠르게 학습을 진행하면서 다양한 실험에 대해 결과를 확인하였다.
2. 전체 dataset에 대해 5개의 fold를 만들어 학습한 후 soft voting을 하여 최종 output을 만들었다.

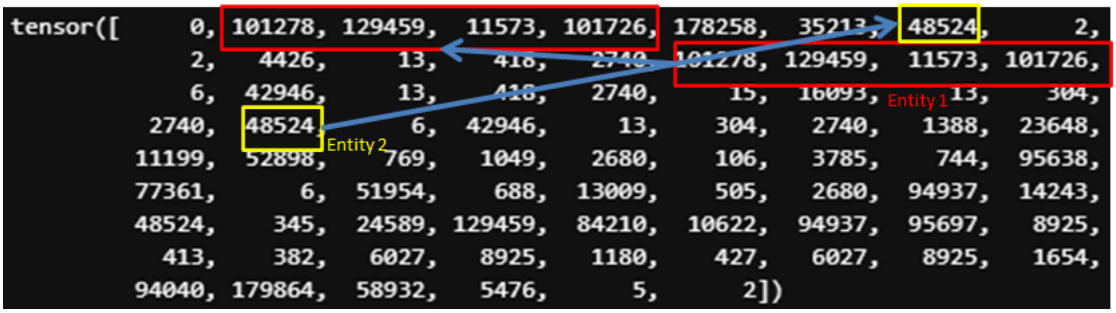
사용한 모델 아키텍쳐 및 하이퍼 파라미터

model	LB 점수	validation	Tokenizer max length	Weight decay	Learning rate	scheduler	Batch size	accumulation	loss	opti
Xlm-Roberta-large	79.9	k-fold cross validation (fold=5)	200	0.001	1e-5	None	32	1	cross entropy	Ada
R-Bert & Roberta	76.4	Validation rate : 0.1	400	0.001	1e-5	None	16	4	Label smoothing loss (label_smoothing_factor=0.5)	Ada
KoElectra-base-v3	74.4	Validation rate : 0.2	200	0.001	5e-5	None	16	1	cross entropy	Ada

기타 시도

1. Input data processing

   sentence 전체를 tokenize 하게 되면 entity에 해당하는 token들의 값이 entity만 token으로 만들어 줄 때의 값과 달라지는 경우가 있었다. 따라서 sentence를 먼저 tokenize 한 뒤 snetence 내에서 entity에 해당하는 token 값을 추출하여 multi sentence로 만들어 input을 넣어주는 방식을 사용하였다.



   그리고 sentence 내에서 entity를 잘 식별할 수 있도록 entity 단어 앞뒤로 entity token을 추가하였고, entity 단어를 구분하는데 separate token을 사용하지 않고 RELATION 단어를 사용하여 entity 단어를 구분하였다.

   유럽 축구 연맹 RELATION UEFA

   +

   <e1> 유럽 축구 연맹 </e1> (<e2> UEFA </e2>) 집행위원회는 2014년 1월 24일에 열린 회의를 통해 2017년 대회부터 UEFA U-21 축구 선수권 대회 참가국을 8개국에서 12개국으로 확대하기로 결정했다.

2. Random UNK token

   Tokenized input 중 random하게 선택된 일부 token을 unk token 혹은 mask token으로 치환하여 argumentation 효과를 볼 수 있는 방법을 사용하였다.

<s> 유럽 축구 연맹 RELATION UEFA</s></s> <e1> 유럽 축구 연맹 </e1> (<e2> UEFA </e2>) 집행위원회는 2014년 1월 24일에 열린 회의를 통해 2017년 대회부터 UEFA U-21 축구 선수권 대회 참가국을 8개국에서 12개국으로 확대하기로 결정했다.</s>

⬇

<s> 유럽 축구 연맹 RELATION UEFA</s></s><unk>e1> 유럽 축구 연맹 </<unk>1> (<e2> UEFA <unk>e2>) 집행위원회는<unk>년 1월 24<unk><unk><unk>회의를 통해 2017년 대회부터 UEFA U-21 축구 선수권 대회 참가<unk>을<unk>개국에서 12개국으로 확대하기로<unk>했다.</s>

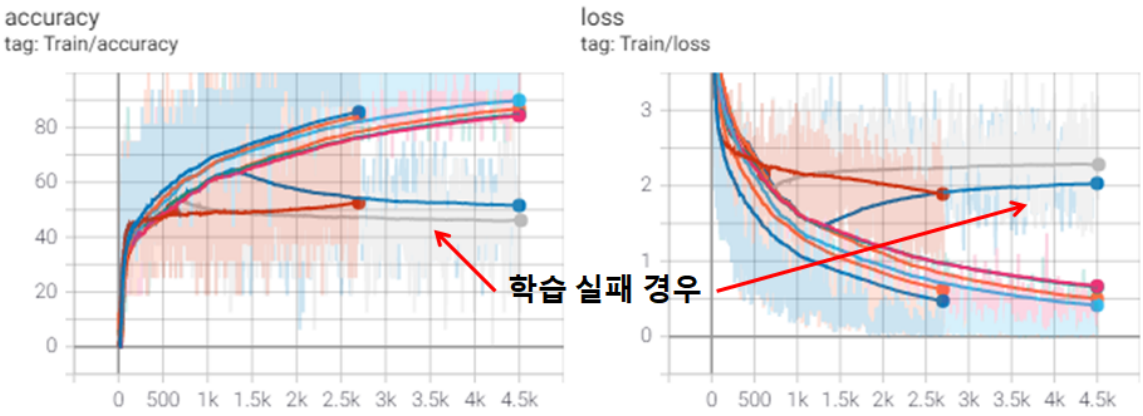
앙상블 방법

1. 5개의 fold를 나누어 각 fold의 logit에 softmax 함수를 통해 확률 값으로 만들어주고 soft voting을 하여 모델의 최종 결과를 만들었다.
2. 마지막에 리더보드 상의 상위 점수인 6개의 모델을 통해 hard voting을 한 후 그 결과를 제출하려 했으나 서버 오류로 인해 결국 마지막에 제출하지 못하여 성능 확인을 하지는 못하였다.

시도했으나 잘 되지 않았던 것들

1. UNK Token

   전체 dataset에서 unk token이 존재하지 않도록 vocab에 전부 추가해준 후 모델을 학습하였다. 약간의 성능 향상이 있었지만 학습이 안정적으로 되지 못하고 학습에 실패하는 경우가 자주 발생하여 사용할 수 없었다.



2. entity token

   모델이 원활하게 학습을 할 수 있도록 entity token을 special token으로 추가하여 활용해보려고 시도하였다.

   - 실험

     1) sentence 내의 entity 단어 앞뒤로 entity token을 추가해주는 방식

     2) sentence 내의 entity 단어를 동일한 entity token으로 치환해주는 방식

     3) sentence 내의 entity 단어를 entity 1, entity 2와 같이 각각의 token으로 치환해주는 방식

     4) entity token을 사용한 sentence만으로 학습하는 경우

     5) 문장의 길이를 표현하는 attention mask에 문장 내의 entity 단어에 해당하는 부분에 1을 더한 후 학습하는 방식

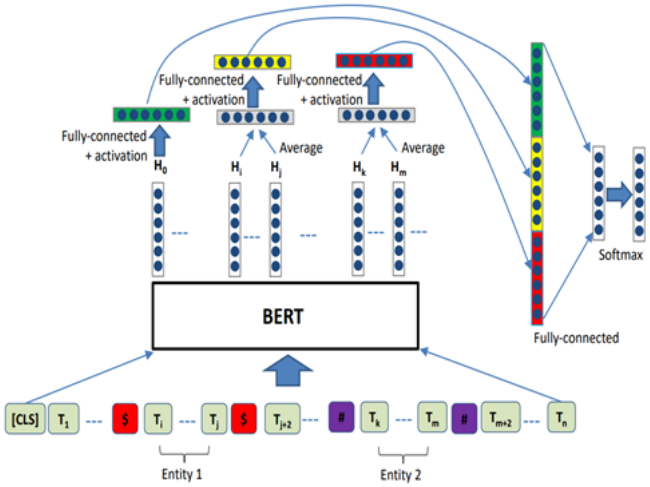
     6) 첫 번째 문장과 두 번째 문장을 구별해주는 token type ids를 entity 1, entity 2, sentence로 3개를 구별해주도록 만들어 학습하는 방식

   → 결과적으로 entity token을 사용할 경우 special token으로 추가하지 않고 사용하는 것이 가장 좋은 성능을 보였다.

3. R-BERT

   - baseline

     entity 1, entity 2 mask를 사용하여 cls token의 output만 classification에 사용하지 않고, 각 input에서 entity에 해당하는 부분의 output들도 task 수행을 위해 사용하는 concept이다.

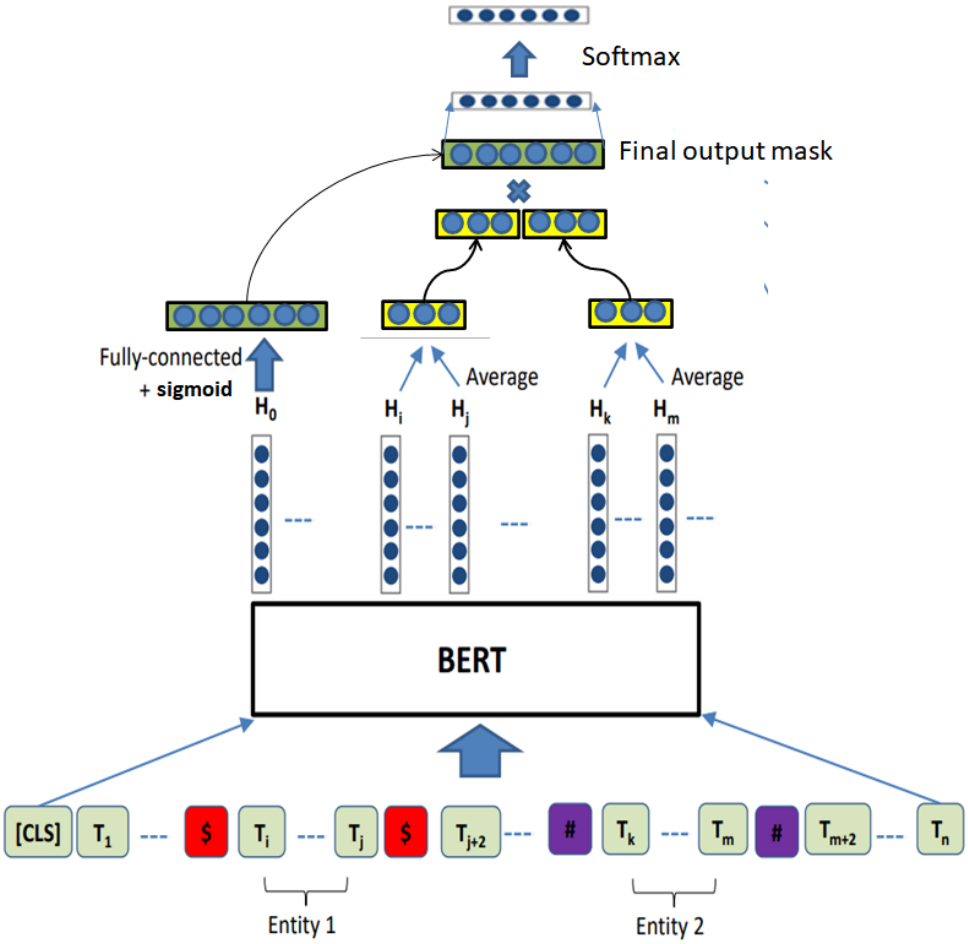


이때 sentence 내에서 각 entity가 어디에 위치하는지 알기 위해 entity 1과 entity 2 mask를 생성해주었다.

```
input ids :  <s> 유럽 축구 연맹 RELATION UEFA</s></s> <e1> 유럽 축구 연맹
</e1> (<e2> UEFA </e2> ) 집행위원회는 2014년 1월 24일에 열린 회의를 통해 2
017년 대회부터 UEFA U-21 축구 선수권 대회 참가국을 8개국에서 12개국으로 확
대하기로 결정했다.</s>
================================================================================
entity 1 masking :  <s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s> 유럽 축구
연맹<s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s>
<s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><
s><s><s><s><s><s><s><s><s><s><s><s><s>
================================================================================
entity 2 masking :  <s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s
><s><s><s><s><s><s><s><s><s><s><s> UEFA<s><s><s><s><s><s><s><s><s><s><s><s><
s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s
><s><s><s><s><s><s><s><s><s><s><s><s>
```

→ 모델을 구현한 시기가 대회 마감이 얼마 남지 않았을 때라 다양한 실험을 해볼 수 없어서 성능의 향상은 없었지만 시간만 충분했다면 잠재적인 성능을 더 끌어올릴 수 있었다고 확신한다.

- other version



Baseline의 모델의 학습 성능이 잘 나오지 않는 것의 원인으로 생각한 것이 적은 수의 data를 사용하면서 fine tuning을 하고 있는 경우인데, pretrained model과 결합된 branch에 너무 많은 parameter가 사용되었기 때문이라고 생각했다. 따라서 parameter 수를 조금이라도 줄일 수 있는 모델 구조로 바꾸었다.

우선 전체 output에서 entity mask를 통해 구해지는 2개의 vector는 sentence 내에서 entity에 대한 포괄적인 정보를 가지고 있다고 생각했고 cls token을 통해 나온 output은 현재 classification을 수행하기 위해 필요한 정보를 가지고 있다고 생각했다. 따라서 cls token에서 나온 output을 통해 포괄적인 정보를 담고 있는 vector에서 task에 필요한 정보 만을 가져올 수 있는 mask를 생성해주면 이 학습 방식의 이점은 어느 정도 유지하면서 적은 parameter로 학습을 할 수 있지 않을까 하는 생각에 시도해보았다.

→ 하지만 시간이 부족하여 정밀하게 성능 검증을 하진 못하였고, baseline의 경우와 비슷한 성능을 보였다.

학습과정에서의 교훈

1. 학습과 관련하여 개인과 동료로서 얻은 교훈

   Model을 만들고, training, inference를 진행하고, 결과가 잘 나오든 잘 나오지 않든 왜 그런지 이유를 파악하고 문제를 차근차근 해결해나가는 과정이 너무 재미있다. 그리고 이런 과정을 즐기는 다른 동료들과 내가 미처 생각하지 못한 부분들에 대해 같이 discussion 할 수 있다는 점이 너무 기분 좋게 만들었다.

   특히, stage 1이 끝나며 남은 stage를 같이 할 동료들과 팀을 만들어 한 주에 최소 2번 피어세션과 별개로 회의를 진행하였고, 이 과정을 통해 정말 많은 부분에 대해 고려할 수 있었고 생각할 수 있는 범위와 깊이를 크게 늘릴 수 있었다고 생각한다.

   그 중 가장 인상 깊었던 것은 태양님께서 내신 아이디어인 entity 주변의 일정 범위의 문장만 학습에 사용하는 방법과 수지님께서 내신 아이디어인 entity token만 추가해주는 것이 아니라 attention mask에 entity에 해당하는 부분에 값을 더해주는 방법은 혼자였다면 절대 생각하지 못했을 아이디어라고 생각한다. 결과보다는 이러한 방식으로 문제를 해결할 수도 있다고 생각할 수 있게 만들어주셔서 정말 감사하게 생각한다.

2. 피어세션을 진행하며 좋았던 부분과 동료로부터 배운 부분

   피어세션에서도 정말 유익한 시간을 보낼 수 있었다. 특히 송광원 캠퍼님과 이태환 캠퍼님이 문제를 해결하기 위해 생각하는 방식은 전혀 색다르게 느껴져서 처음엔 공감하기 어려웠지만 논리적으로 잘 설명해주셔서 굉장히 감명 받았다.

특히 송광원 캠퍼님께서 validation dataset을 구분하기 위해 굉장히 많은 시간을 들이고 고민을 하였는데 이 부분에서 처음엔 왜 그렇게 많은 시간을 들이는지 이해하지 못했지만 오늘 마지막 피어세션에서 결과를 듣고 모델을 만들고, 학습하는데 있어 굉장히 중요한 부분이라는 것을 깨달을 수 있었다.

마주한 한계와 도전 숙제

1. 아쉬웠던 점들

   우선 제공된 baseline의 code가 아닌 pytorch로 새로 구현한 코드를 사용하여 학습을 하였다. 하지만 동일한 hyper parameter를 사용하여 학습을 하는데도 huggingface에서 제공하는 trainer 객체를 사용한 학습 결과보다 좋지 못한 성능을 보였고, 아직 원인을 파악하지 못한 점이 너무 아쉽다.

   그리고 초반 3일 정도는 NLP task에 대한 기초 지식이 너무 부족하여 공부밖에 할 수 없었던 점이 너무 아쉽다. 빠르게 대회에 참여하여 다양한 시도를 해보고 싶은 욕심이 있었지만 알고 있는 지식이 부족하였다.

2. 한계/교훈을 바탕으로 다음 스테이지에서 새롭게 시도해볼 것

   다음 stage는 개인이 아닌 팀으로 진행되는 점에서 절대 팀에 피해가 되지 않도록 코드도 좀 더 꼼꼼하게 살펴보며 실수 없이 작성하려 한다. 그리고 굉장히 운이 좋게도 능력 있는 팀원들과 함께하게 되어 뒤쳐지지 않도록 더 많이 공부하고 노력할 것이다. 딱히 지금보다 더 나아지기 위해 더 많은 공부를 하는 것 외에 어떤 노력을 해야 할지도 사실 잘 모르겠다.