김 재 현 | Jae Hyun Kim

INTRODUCTION

부스트캠프 AI Tech 과정을 진행하며 P stage의 Level이 마무리 될 때 마다 진행된 동료 피드백과 롤링페이퍼를 바탕으로 저를 소개해 드리겠습니다.

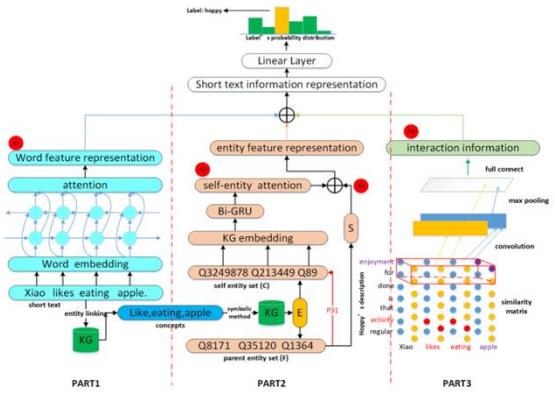
1. **빠른 성장을 이룩하는 사람 : “이미 좋은 개발자이신것 같습니다.”**
   개발 직군으로 진로를 정한지 6개월이 채 되지 않은 지금 아직 부족한 점이 많지만 동료 피드백을 통해 저를 이미 좋은 개발자라고 평가해주시는 분이 계셨습니다. 이는 제가 크게 뛰어나다기보다 다행히 팀에서 1인분은 하고 있구나 판단 할 수 있게 되는 계기가 되었습니다. 개발과 AI가 재미있어 몰입을 하여 교육과정을 진행했고 이것이 짧은 기간이지만 빠르게 성장해 팀의 일원으로써 역할을 원활히 수행할 수 있게 만들어준 원동력이 되었다고 생각합니다.
2. **공유를 좋아하는 사람 : “자료 공유 해주셔서 감사합니다!”**
   개발과 AI 공부를 진행하면서 외부 자료와 교육을 통해 성장을 할 수 있었습니다. 양질의 자료를 쉽게 구할 수 있는 이유는 바로 개발자들의 공유 문화 때문이라 생각합니다. 그래서 저도 제가 자료를 조사하면서 배우고 공부한 내용을 정리해 모두가 함께 성장하기 위해서 팀원들과 공유했습니다. 이러한 공유는 또 다른 공유를 불러왔고 결과적으로 프로젝트를 진행에 필요한 다양한 정보와 아이디어가 샘솟게 된 원동력이 되었습니다.
3. **소통과 기록을 중요시하는 사람 : “협업 요소 구성에 감사드립니다.”**
   프로젝트 진행에 있어서 가장 중요한 것은 팀원들과의 소통과 프로젝트 진행상황을 기록하는 것이라 생각합니다. 그래서 프로젝트가 시작될 때 마다 누구나 쉽게 실험 결과를 기입하고 공유할 수 있는 환경을 만들어 원활한 협력이 이뤄질 수 있도록 노력해왔습니다.

PROJECT

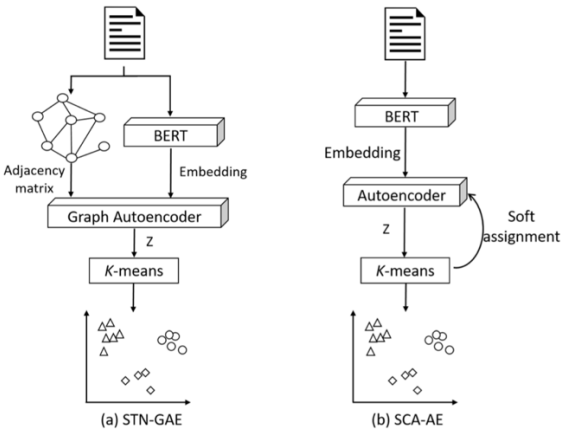
부스트캠프 AI Tech - 최종 프로젝트 : 악플 수집 서비스

주제 : 악성 댓글 문화 근절을 위한 악성 댓글 자동 수집 서비스 연구 및 제작
기간 : 2021.12.02 ~ 2021.12.21
수행 역할

1. **최종 프로젝트 아이디어 제공**
   최종프로젝트 진행에 필요한 데이터 셋을 SocialNLP@ACL에 등재된 신뢰성 있는 데이터 셋인 Korean Hate Speech 데이터 셋을 이용하기로 정했습니다. 이를 활용하기 위해 꾸준히 악성 댓글의 타겟이 되고 있는 셀럽들과 개인 방송 BJ, 유튜버 등을 대상으로 기존에 팬이나 법무팀에서 직접 수집하던 프로세스를 자동화하여 제공하자는 아이디어를 제시했습니다. 실효성 검증을 위해 여러 기획사, MCN에게 설문을 보냈고 86.7%의 긍정적 반응을 얻을 수 있었습니다.
2. **AutoML을 통해 Hyperparameter Optimization(HPO) 진행**
   CV에 맞춰져 구성되어 있던 지난 Model Optimization에 사용 했던 코드를 리팩토링하여 NLP에 맞게 구성을 바꾸었습니다. 그리고 Backbone model을 Task adapted model인 KcELECTRA 모델을 사용하였기에 모델 가중치를 유지하기 위해서 NAS를 따로 진행하지 않았고 HPO만 진행하였습니다. 이를 통해 HPO사용 전 보다 같은 데이터 셋을 사용하는 Kaggle Competition 기준 F1 Score 0.04 상승을 이뤄낼 수 있었습니다.
3. **Short Text Classification 성능 향상을 위해 논문 기반 모델 설계 진행**
   문장이 짧을 수록 긴 문장에 비해 모델이 의미를 파악하는데 힘들 것이라 생각했습니다. 그렇기에 Short Text Classification관련 논문 조사를 통해 두 가지 논문을 선정해 살펴 보았습니다.
   먼저 DNN model + Knowledge Graph + Similarity model based on CNN[그림 1] 로 구성된 모델입니다. KG(Knowledge Graph)을 이용해 text에 존재하는 인물, 고유 명사 등의 지식을 이용해 더 명확한 text 의미 파악을 하고자 했습니다. 하지만 KG 사용을 위해 조사를 중 한국어로 embedding된 KG를 찾지 못했고 reference를 찾아보니 한국어를 NMT를 통해 영어로 번역하여 KG를 적용했지만 저희가 사용하는 데이터는 연예 기사 댓글이였기에 맞춤법 파괴와 초성, 비속어로 이뤄진 데이터가 많았기에 방법을 적용하기 힘들어 다른 방법을 찾아 보았습니다.
   그래서 조사한 것이 Pre-Trained Model + AutoEncoder + K-mean Clustering[그림2] 로 구성된 모델입니다. 짧은 말뭉치들은 spaese, high-dimension, noise한 성격을 가지고 있기 때문에 효과적인 학습 방법은 Clustring이라 주장하였습니다. Pre-Trained Model을 사용하는 점에 있어서 KcELECTRA 모델을 활용하여 좋은 성능이 기대 되었습니다. 하지만 제한된 기간안에 구현이 이뤄지지 않았고 향후 구현을 완료하여 Repo에 업로드 예정입니다.



[그림 1] DNN + KG + Similarity Model



[그림 2] Pre-Trained Model + AutoEncoder + K-mean Clustering

부스트캠프 AI Tech Level3 - “초”경량 이미지 분류기

주제 : 재활용 쓰레기 데이터 셋에 대해서 이미지 분류를 수행하는 모델을 AutoML을 이용하여 성능을 유지하면서 가볍게 만드는 것이 목적
기간 : 2021.11.22 ~ 2021.12.02
수행 역할

1. **Baseline 코드 분석 후 공유**
   Optuna를 이용한 AutoML을 팀원 모두가 처음 다뤄보기 때문에 코드가 어떤식으로 흘러가는지 한 페이지씩 모듈화 해서 코드 분석한 결과를 팀원들과 공유하여 팀원 모두가 빠르게 코드를 이해하고 넘어갈 수 있도록 도왔습니다.
2. **NAS 수정 및 HPO 진행**
   Baseline 코드 분석을 바탕으로 다양하게 모듈을 바꿔가며 Optuna를 통해 Neural Architecture Search(NAS) 실행과 Hyperparameter Optimazation을 진행하여 보다 가벼우면서 성능은 유지하는 모델을 찾기 위해 노력했습니다.

부스트캠프 AI Tech Level3 - 데이터 제작 대회

주제 : 위키피디아 원시 말뭉치를 활용하여 직접 관계 추출 태스크에 쓰이는 주석 코퍼스 제작
기간 : 2021.11.08 ~ 2021.11.19
수행 역할

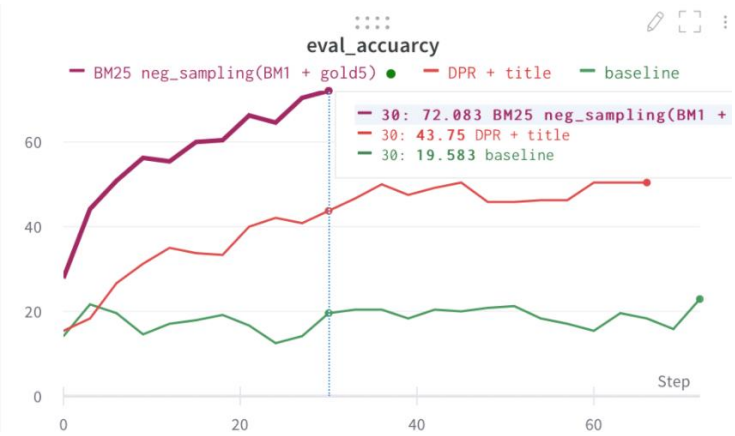
1. **데이터 Annotation 진행**
   웹 기반의 relation, entity tagging을 무료로 지원해주는 Tagtog을 사용하여 저희 팀에게 주어진 위키피디아 축구 관련 원시 말뭉치를 팀에서 지정한 Relation map에 맞게 데이터를 Annotation을 진행했습니다.
2. **가이드라인 작성**
   Annotation을 진행하는 작업자가 어떤 점에서 Tagging 작업에서 헷갈릴지 고민하며 가이드라인을 작성했습니다. 이를 과정을 설계해주신 마스터님과 멘토님들께서 좋게 봐주셔서 모든 캠퍼들 앞에서 발표를 했습니다.
   [발표 자료]

부스트캠프 AI Tech Level2 - Open-Domain Question Answering

주제 : Question가 들어오면 Wikipedia에서 질문과 관련된 문서를 찾아주는 ‘retriever단계와 관련된 문서를 읽고 적절한 답변을 찾아주는 ‘reader’ 단계의 결합을 통해 Answer를 찾는 것이 목표
기간 : 2021.10.12 ~2021.11.04
수행 역할

1. **논문에 기반한 Dense Passage Retrieval 구현**
   Baseline에 DPR(Dense Passage Retrieval)이 간단하게 구현이 되어있었지만 성능이 좋지 않았습니다. 생각보다 낮은 성능을 보여주자 코드에 이상이 있는지 확인하기 위해 DPR을 제시한 논문을 찾아보며 코드와 negative sampling을 하는 과정에서 다른 점을 찾아 수정을 하였고 이전보다 성능이 확실히 오르는 것을 확인[그림 3] 할 수 있었습니다. 이후 앙상블을 진행할 때도 Sparse만 단일로 앙상블 한 결과보다 DPR과 함께 앙상블 한 모델이 성

능이 가장 잘나오는 결과를 보였습니다.



[그림 3] 논문에 기반한 DPR과 Baseline 비교 (자주색 : 논문 기반 DPR)

2. **팀 Wandb 프로젝트 생성**

   긴 프로젝트 기간인 만큼 많은 모델링 실험이 많았고 각자 실험한 모델의 성능을 쉽게 비교하기 위해서 팀 Wandb를 구축했습니다. 이를 통해 효과적으로 모델 성능 비교와 함께 각 모델의 Hyperparameter의 로그도 쉽게 비교가 가능해졌고 이후 프로젝트에도 계속 적용하게 되었습니다.

부스트캠프 AI Tech Level2 - 문장 내 개체간 관계 추출

주제 : 문장, 단어에 대한 정보(Entity)를 통해, 문장 속에서 단어 사이의 관계를 추론하는 모델 학습

기간 : 2021.09.27 ~ 2021.10.07

수행 역할

1. **AEDA(An Easeier Data Augmentation) 적용**

   주어진 데이터가 Imbalance가 심했기 때문에 Data Augmentation을 진행하기 위해 AEDA를 적용했습니다. 문장이 주어지면 문장의 일정비율에 첨자를 넣으며 Augmentation을 진행했습니다. 하지만 적은 데이터에 과도한 증강을 시도하다보니 overfitting이 일어나 큰 효과를 주지 못했습니다. 비율을 줄이고 모든 label에 균일하게 적용시 약간의 성능 향상이 있었는데 이 결과로 AEDA는 데이터 증강보다는 적절히 사용할 때 모델을 robust 하게 만들어주는 효과가 있다고 결론을 지었습니다.

2. **모델 구조와 아이디어 파악을 위한 논문 리뷰**

   지난 마스크 착용 상태 분류 프로젝트 진행 시 모델의 구조와 아이디어는 알지 못한채 단순히 SOTA 모델이라 소개되어 있는 Pretrained model들을 불러와 사용한 점이 아쉬워 NLP의 대표적인 Pretrained Model BERT, RoBERTa, ELECTRA 의 논문을 찾아 읽어보고 발표하는 세미나를 열었습니다.

부스트캠프 AI Tech Level1 - 마스크 착용 상태 분류

주제 : 카메라로 찍은 사람 얼굴 이미지만으로 이 사람이 마스크를 올바르게 쓰고 있는지 성별과 나이대까지 추정하는 모델 학습

기간 : 2021.08.22 ~ 2021.09.02

수행 역할 :

1. **Cutmix와 Albumentation을 적용한 augmentation 진행**

   모델의 Generalization과 feature 표현을 잘 잡아내기 위해 Cutmix와 Albumentation 을 통해 augmentation을 진행 했습니다.

2. **TIMM 라이브러리를 이용해 Pretrained model 적용**

   Paperswithcode 에 소개되어 있는 SOTA 모델에서 제한된 하드웨어에서 돌릴 수 있는 적절한 모델을 찾아 TIMM(PyTorch Image Models)에서 pretained model을 불러와 fine-tuning을 진행했습니다.

3. **Weighted Random Sampler 적용**

   데이터에서 연령 label 불균형이 심하여 label이 적은 고연령 데이터를 더 자주 불러오고 그때마다 다른 Augmentation을 진행하여 upsampling 효과를 줄 수 있도록 Weighted Random Sampler 적용을 시도했습니다.