2022 관광데이터 AI 경진대회

Team hyy

Method Description

- 트레이닝과 테스트 코드는 깃허브에 있음 (https://github.com/YiyuHong/dacon_tour2022)
- 텍스트 정보만 사용함. 50% 확률로 training data에 text augmentation을 진행하며, text augmentation은 단순히 전체 문장의 임의의 위치의 25~50%의 연속적인 characte를 지우는 것 임.
- KLUE-RoBERTa-Large pre-trained language model를 사용함 (https://github.com/KLUE-benchmark/KLUE)
- Cat1, Cat2, Cat3 클래스를 모두 예측 함. 단 모델의 마지막 layer의 feature로부터 모두 예측하는 것이 아니라, Cat1은 마지막 5번째 layer의 feature; Cat2는 마지막 3번째 layer의 feature; Cat3은 마지막 layer의 feature로 부터 예측함. Feature중 첫번째 [cls]토큰에 해당하는 word embedding을 두개의 transformer encoder layer와 하나의 Linear layer를 붙여 각 category class를 예측함.
- Test 단계에서 최종 예측 값은 Cat1, Cat2, Cat3의 예측 값을 더하여 산출함. 구체적으로 Cat1과 Cat2의 클래스는 모두 해당되는 하위 Cat3 클래스가 있으므로 매칭되는 클래스끼리 예측 확률 값을 더하는 방식으로 최종 예측 값을 얻음.
- 5fold와 seed를 달리하여 똑같은 모델을 다수 훈련하여 앙상블을 하여 1개 결과파일을 제출하였고, 추가로 knowledge distillation을 적용하여 1개 결과파일을 제출하였음.

Training Detail

- Mixed precision training is used (24GB GPU is required)
- Batch size = 16
- Gradient Accumulation = 4
- Max word length = 256
- Max Epoch = 31
- Radam optimizer (learning rate= 0.00003)
- StepLR scheduler (step_size= 10 epoch, gamma=0.5)
- Cross Entropy Loss function (weighted for Cat1 Cat2 and Cat3)
  - Total loss = Cat1_loss*0.05 + Cat2_loss*0.15 + Cat3_loss*0.8
- Five-fold cross validation setting to select best model for each fold depend on best validation score
- Run with five different seed, so totally 5*5=25 models result ensemble recorded public leaderboard score: 0.86543, private leaderboard score: 0.85923

Knowledge Distillation

- 25개 model의 Cat1, Cat2, Cat3의 예측값을 앙상블하여 test data에 대한 정답 값(soft label)으로 간주하고 train data와 섞어서 새로운 모델을 트레이닝 함
- 매 batch에 train data 와 test data를 1:1 비율로 넣음
- Loss function: (weighted for Cat1 Cat2 and Cat3 and distillation data)
  - Total loss = (Train_Cat1_loss+Test_Cat1_Loss*0.3)*0.05 + (Train_Cat2_loss+Test_Cat2_Loss*0.3)*0.15 + (Train_Cat3_loss+Test_Cat3_Loss*0.3)*0.8
- Other settings are same
- Five-fold cross validation setting to select best model for each fold depend on best validation score
- 5 model ensemble results recorded public leaderboard score: 0.86447, private leaderboard score: 0.86012

Tried but no improvement

- Heavy text augmentation, word swap, word deletion.
- Test time augmentation,
- Predict all cat1, cat2, cat3 from features of last layer.
- transformer encoder layer 2개 뒤에 MLP 2층 3층을 붙여 cat1, cat2, cat3 예측
- 개수 적은 class의 샘플을 upsampling하여 트레이닝