Week1-2 과제

1. Paperswithcode에서 NLU sub task 중 하나를 선택하여 본인 블로그에 정리해보세요. 아래 3가지 항목에 대해서 정리하세요. (각 항목 고려 사항 참고)

Natural Language Inference

문제정의

전제와 가설 문장 사이의 관계를 entailment, contradiction, neutral로 분류해 전제와 가설 문장 간의 관계를 파악한다.

데이터 소개

KLUE NLI


```
[
    {
        "guid": "klue-nli-v1_train_00000",
        "genre": "NSMC",
        "premise": "힛걸 진심 최고다 그 어떤 히어로보다 멋지다",
        "hypothesis": "힛걸 진심 최고로 멋지다.",
        "gold_label": "entailment",
        "author": "entailment",
        "label2": "entailment",
        "label3": "entailment",
        "label4": "entailment",
        "label5": "entailment"
    },
    {
        "guid": "klue-nli-v1_train_00001",
        "genre": "NSMC",
        "premise": "100분간 잘껄 그래도 소닉붐땜에 2점준다",
        "hypothesis": "100분간 잤다.",
        "gold_label": "contradiction",
        "author": "contradiction",
        "label2": "contradiction",
        "label3": "contradiction",
        "label4": "neutral",
        "label5": "contradiction"
    },
    {
        "guid": "klue-nli-v1_train_00002",
        "genre": "NSMC",
        "premise": "100분간 잘껄 그래도 소닉붐땜에 2점준다",
        "hypothesis": "소닉붐이 정말 멋있었다.",
        "gold_label": "neutral",
        "author": "neutral",
        "label2": "neutral",
        "label3": "neutral",
        "label4": "neutral",
        "label5": "neutral"
    },
```


train, dev set 이 json 파일로 저장되고 test set은 공개되지 않음. KLUE 벤치마크 사이트에서 테스트 데이터 평가가 가능. 데이터의 출처는 NSMC, 위키뉴스, 에어비엔비, 정책 홍보 보도자료로 구성됨.

- train 24997문장
- dev 2999문장

SOTA 모델 소개

RoBERTAa

- BERT 논문의 실험을 복구하고 BERT의 초매개변수를 조정하고 더 긴 epoch으로 학습해 BERT보다 성능 향상

KcBERT

- 네이버 뉴스의 댓글과 대댓글을 수집해 토크나이저와 BERT 모델을 처음부터 학습(즉, BERT 구조에 데이터를 추가했다는 것)