극소량 학습에서의 dev데이터 분석

Team: 코딩잡쉈어
Member: 박준우 김한솔 김현준
Professor: 한요섭 교수님
Assistant: 김영욱 조교님

Introduction & Limitations

극소량 학습이란?
기존의 대규모 데이터셋을 필요로 하는 딥러닝 모델의 한계를 극복하고자 데이터가 제한적인 상황에서도 학습을 최적화 하는 연구

본 연구는 자연어 모델의 데이터 증강기법을 적용하지 않은 극소량 학습에서 다양한 train-dev 비율을 적용하여 실험하고 그 결과를 분석하여 최적의 dev set 방법론을 추론하는 것을 목적으로 함.

AI 모델 학습시에 정확도를 높이는데 데이터의 양이 중요하지만 많은 양질의 데이터를 구하기 힘들고 시간과 비용이 많이 필요함.
그 동안 few-shot 환경에서 데이터 증강 기법을 통해 모델의 성능을 향상시키는 연구는 활발히 진행되어 왔으나 데이터 증강기법을 활용하지 않고 train-dev 데이터 분석에 대한 연구는 부족함.

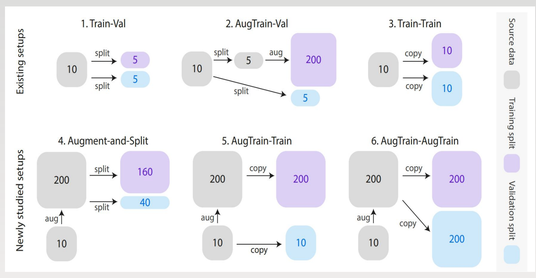
-> **한정된 양의 데이터만 있는 환경에서도 자연어 모델의 학습을 최적화하기 위해 해당 연구를 진행함.**

Experiment 1 - setup

1) 다양한 n-shot: n=10/20/50/100
2) 다양한 split 비율: 9:1/ 8:2/ … /1:9 에 대한 성능 분석

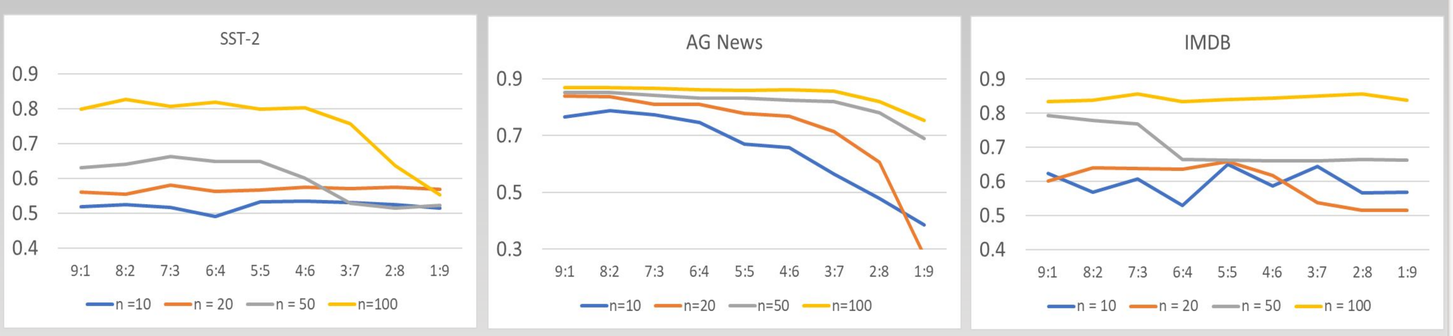
활용 모델: bert-base-uncased
dataset: SST-2, IMDB, AG News

step1) dataset 별로 class 당 n개의 sample data 추출
step2) sample data split 및 모델 학습 진행
step3) 모델의 test accuracy 측정 후 split 비율에 따른 결과 도출



Dataset	#Class	#Train	#Test	Length
AG News	4	120K	7.6K	80
SST-2	2	6.9K	1.8K	32
IMDB	2	250K	259K	256

Experiment 1 - result

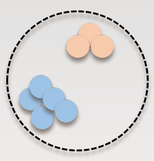


극소량인 n=10,20일 때와 n=50,100일 때가 서로 다른 경향성을 보임.
n=10, 20은 극도로 샘플이 적은 세팅인데, 해당 경우에는 데이터셋 간에 best split의 경향성이 보이지 않음.
n=50, 100는 7:3 이상에서 best split인 경향성을 보임.

Experiment 1 - analysis

result 1의 문제
split 비율마다 실험 시 data를 random하게 분배함. 나눴을 때 train과 validation이 너무 다른 성질의 샘플로 구성되어 있다면, validation set으로 측정하는 성능이 별 의미가 없을 수 있음.

가능한 방법론
유사한 data들을 train/dev에 나누어 분배.
즉, 비슷한 data들이 하나의 set에 집중되지 않도록 분배하였을 때 더 좋은 성능을 기대할 수 있음.



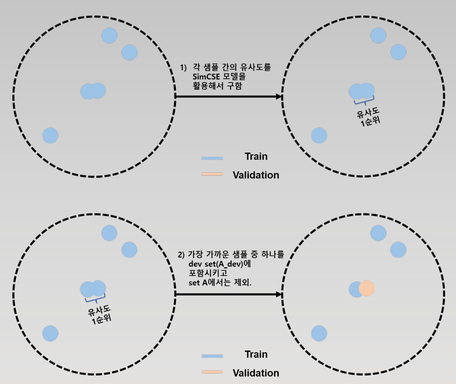
dev set이 train set을 대표하지 못하는 경우

dev set이 train set을 대표할 수 있는 경우

Apporach

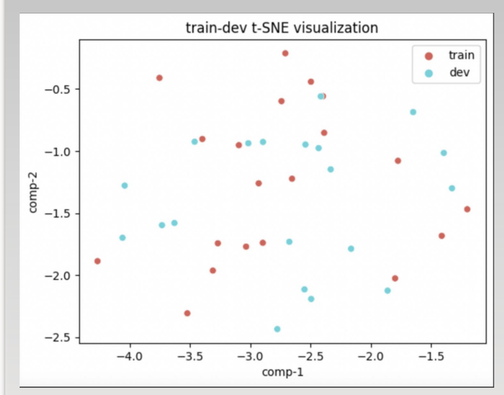
Goal
Dev set이 보다 정확한 일반화 성능을 측정할 수 있어야 함.
Dev set이 Train set을 대표할 수 있어야 함.

1차 실험에서의 best split 비율에서 random보다 더 잘 분배
1) SimCSE를 활용해 sample간의 유사도를 측정
2) 가장 유사한 샘플 pair를 분리하는 식으로 dev set 구성

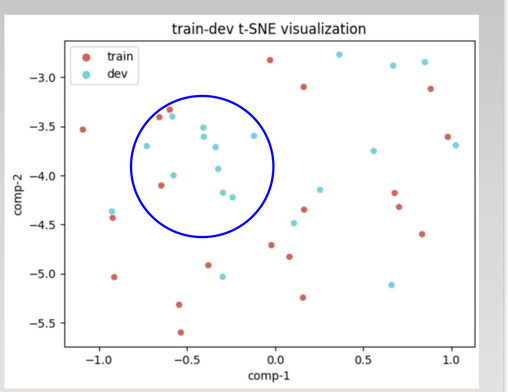


Hypothesis
유사한 sample을 분리하여 dev set이 train set 을 잘 대표할수록 일반화 성능이 잘 측정되고 최종 성능도 높을 것이다.

Experiment 2 - analysis (t-SNE)



Random



SImCSE

SimCSE의 데이터가 뭉쳐있는 것을 확인할 수 있음(다양성 부족) -> Train set이 Dev set을 대표하지 못함 -> **성능 저하의 원인 추정**

Experiment 2 - result

SimCSE를 활용하였을 때 성능이 좋아질 것이라는 가설과 다르게
data를 SimCSE 활용한 것이 random하게 split 했을 때보다 성능이 떨어짐.

이와 같은 결과가 발생한 예상할 수 있는 이유로는
1) train set에 남아있는 가장 가까운 sample에 overfitting 되는 식으로 학습이 되었을 가능성

2) sample들이 몰려 있는 경우, 계속해서 비슷한 sample을 validation set에 할당함으로 인한 data 다양성 부족

따라서 cluster를 활용하는 방식을 통해서 다양하면서도, overfitting을 방지할 수 있는 split 방법론을 further study로 고려하는 것이 필요함.

IMDB	SimCSE	Random	Seed
n = 10/ 5:5	0.59721	0.64905	7
n = 20 / 5:5	0.61802	0.65768	7
n = 50 / 9 : 1	0.75579	0.79209	7
n = 100 / 7 : 3	0.83632	0.83887	1