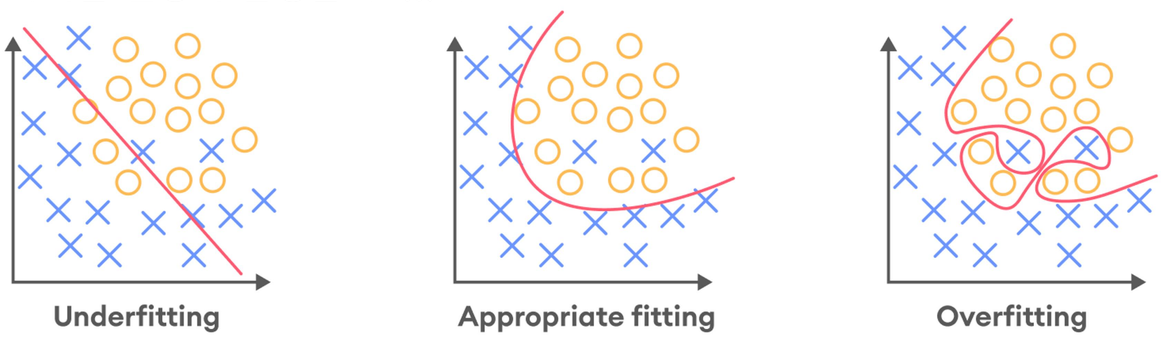
소프트 라벨을 적용한 규칙 기반 텍스트 데이터 증강 기법

최주환1, 이준호1, 송상민1, 진교훈2, *김영빈2
1중앙대학교 AI학과
2중앙대학교 첨단영상대학원
e-mail: gold5230@cau.ac.kr, jhjo32@cau.ac.kr,
s2022120859@cau.ac.kr, fhzh123@cau.ac.kr, ybkim85@cau.ac.kr

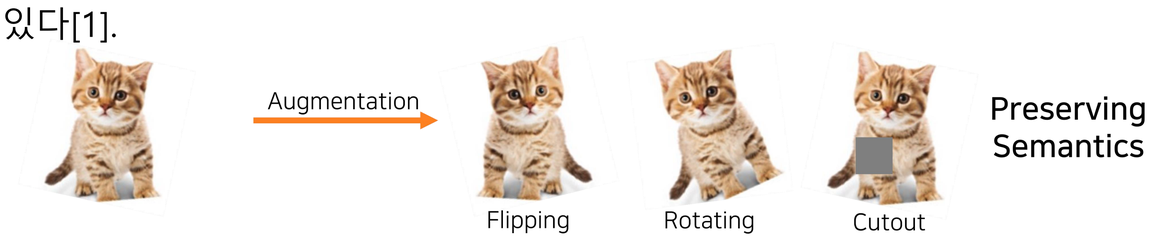
Intelligent Information Processing Lab.
IIPL

서론

데이터에 기반하여 학습을 진행하는 딥러닝 기법의 특성상 더욱 큰 규모의 모델을 학습시키기 위해서 학습 데이터 역시 더욱 많이 요구된다. 모델의 규모에 비해 학습에 사용할 수 있는 **데이터가 부족한 상태에서 학습을 진행했을 경우**, 딥러닝 모델이 주어진 데이터에 대해서 지나치게 의존하는 **과적합 현상이 발생할 수 있다**.



이미지 데이터의 경우, 주어진 이미지를 좌우로 반전시키는 등 단순한 규칙을 통해서 데이터를 증강시켜 성능 향상을 얻을 수 있음이 잘 알려져 있다[1].



This is a cat → Augmentation → Is this a cat (Semantics Changed)

그러나, 이러한 기법은 텍스트에서는 단어를 삭제하는 등의 과정을 통해 주어진 문장이 가지고 있는 본래 의미를 일부 훼손시킬 수 있다는 단점이 있다.

본 연구는 이러한 단점을 보완하고자 **증강된 데이터의 라벨 값에 대해 라벨 스무딩 기법을 적용**하여 증강된 데이터가 원본 데이터와 다른 **소프트 라벨** 값을 갖도록 한다. 이를 통해 증강 데이터에 대해서 상대적으로 약한 신호를 학습할 수 있게 되어, 결과적으로 일반화 성능을 향상시킬 수 있다.

관련 연구

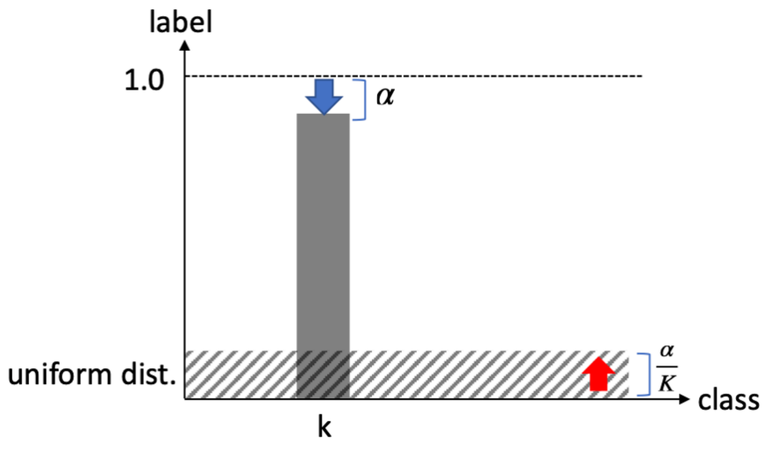
본 논문에서는 기존의 규칙 기반 텍스트 데이터 증강 기법에 소프트 라벨을 적용하여 이를 개선시키는데, 이를 위한 대상으로 가장 많이 활용되는 EDA[2]를 활용하였다.

Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD	A sad, superior human out on the roads of life.

EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks
Wei et al., EMNLP 2019

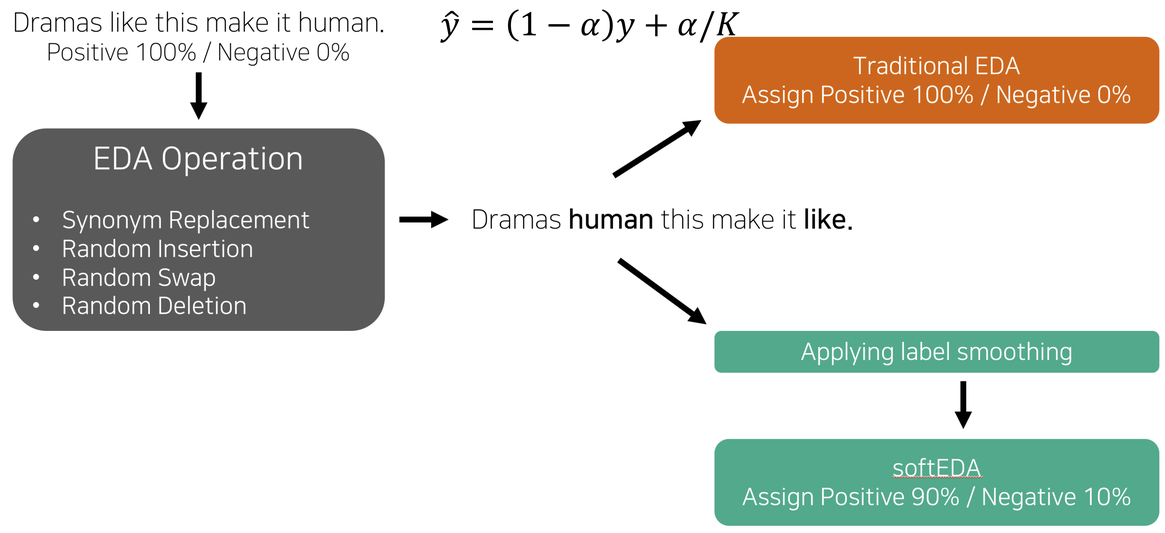
실험 내용

라벨 스무딩이란 원-핫 인코딩 형태의 라벨 값을 모든 클래스에 동일하게 값이 부여되는 균일 분포를 활용해 소프트 라벨 값으로 변환하는 방법이다.



본 연구에서는 EDA 기법이 원래 문장에서 변화를 일으키는 과정에서 원래 문장이 가지고 있는 라벨 값에 대한 의미가 손상될 수 있음을 고려하여, 증강된 데이터에 라벨 스무딩을 적용하여 원-핫 인코딩 값 대신 노이즈가 주입된 **소프트 라벨 값을 부여**한다. 소프트 라벨 값 $\hat{y}$은 아래와 같이 정의된다. 수식에 사용된 K 는 클래스의 개수이며, α 는 라벨 스무딩을 위한 하이퍼 파라미터이다.

$$\hat{y} = (1-\alpha)y + \alpha/K$$



증강된 데이터 문장	기존 기법 라벨값	제안 기법 라벨값
I really enjoyed watching the new movie. → I watching really enjoyed the new.	긍정 1 부정 0	긍정 0.9 부정 0.1
The new horror movie was an awful, boring disaster. → The boring horror was an, new movie disaster.	긍정 0 부정 1	긍정 0.1 부정 0.9

실험 결과

기법/데이터셋	SST2	CoLA	TREC
Baseline	89.74%	75.38%	95.47%
EDA	+0.71%p	-0.45%p	+0.51%p
AEDA[3]	+0.22%p	-0.34%p	-0.67%p
Proposed Method	+0.83%p	+1.50%p	+1.17%p

본 논문에서 제안하는 방법이 다른 방법과 비교했을 때 **성능을 크게 향상**시킴을 확인하였다. 또한, 제안하는 기법은 **다른 기법이 성능 하락을 보이는 경우에도 모델의 성능을 향상**시킬 수 있었다.

결론

본 논문에서는 기존의 규칙 기반 텍스트 증강 기법의 한계점을 보완하기 위해 **증강된 데이터에 라벨 스무딩을 통해 소프트 라벨을 적용하는 방법을 제안**하였다. 또한, 제안한 방법을 적용했을 때 **기존 방법과 비교해 높은 수준의 성능 향상**을 이뤄낼 수 있음을 확인하였다. 향후에는 증강 기법을 적용하는 강도에 따라 라벨 스무딩의 강도 또한 조절하는 기법에 대해서 연구해보고자 한다.

참고문헌

[1] Luke Taylor and Geoff S. Nitschke. "Improving Deep Learning with Generic Data Augmentation." 2018 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1542-1547. 2018.
[2] Jason Wei and Kai Zou. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6383-6389, 2019.
[3] Karimi Akbar, Rossi Leonardo and Prati Andrea. "AEDA: An Easier Data Augmentation Technique for Text Classification." In Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 2748-2754, 2021.