낯선 데이터를 활용한 과잉신뢰 완화 텍스트 증강 기법

• 이준호[1], 송상민[1], 최주환[2], 박주형[3], 진교훈[4], 김영빈[4]

1 중앙대학교 AI학과
2 중앙대학교 전자전기공학부
3 중앙대학교 컴퓨터공학부
4 중앙대학교 첨단영상대학원

Juhwan Choi

gold5230@cau.ac.kr

Intelligent Information Processing Lab.

IIPL

Index

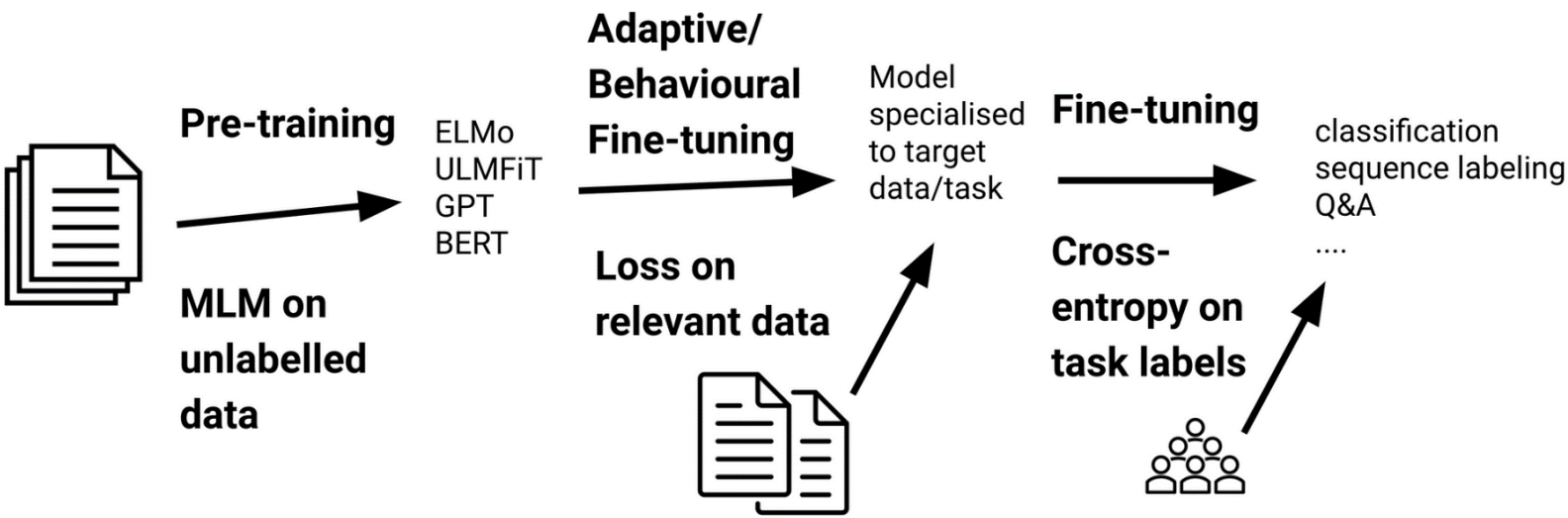
Introduction

Introduction

NLP Trends

최근 딥러닝 기반 자연어처리 연구 동향

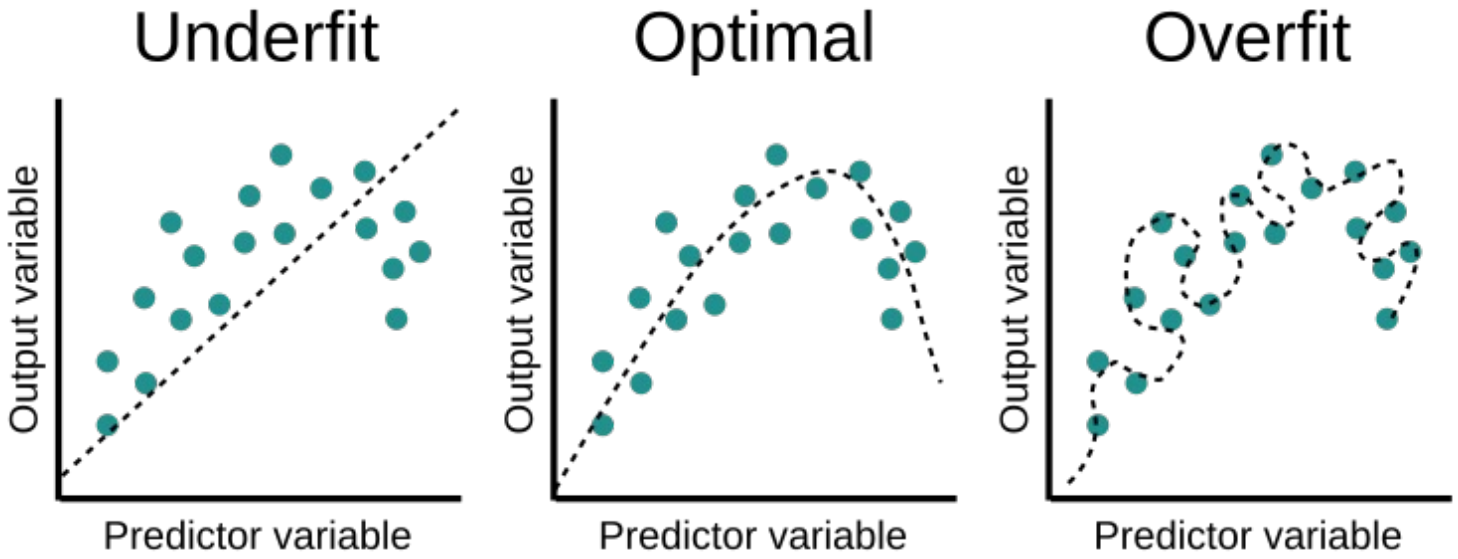
대규모 사전학습 언어모델을 원하는 작업에 미세조정



Overfitting

미세조정 학습 데이터가 부족할 경우

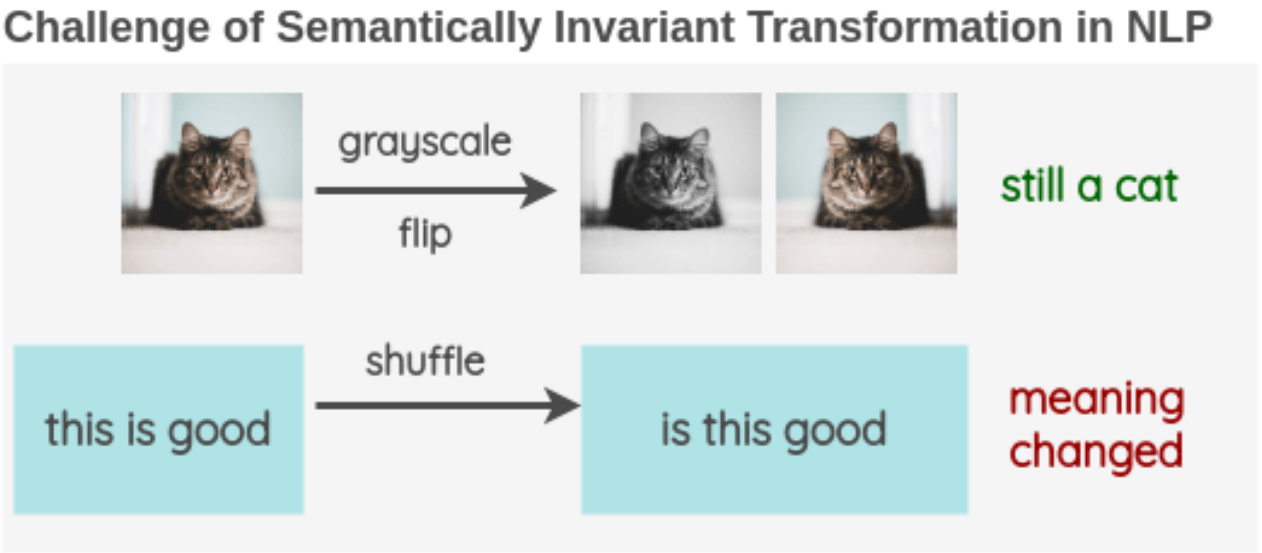
주어진 데이터에만 최적화되는 과적합 발생 가능성



Data Augmentation

과적합을 해결하기 위한 방법

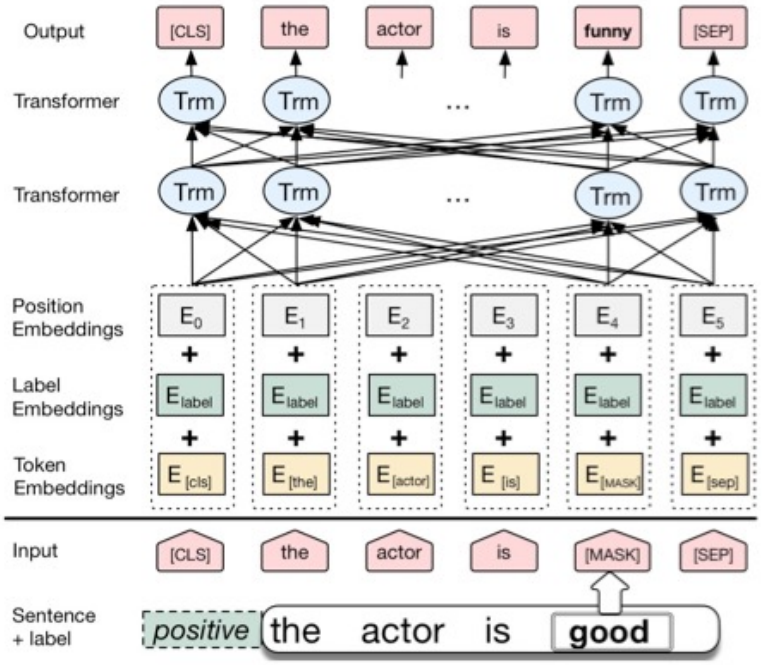
증강 시 텍스트 데이터의 특징을 고려해야 함



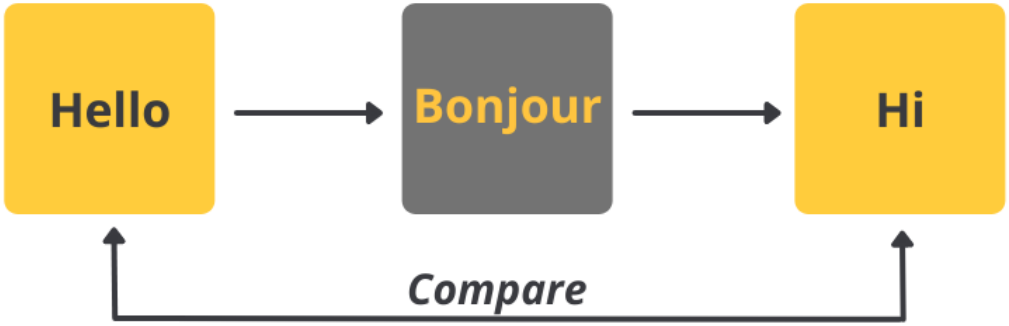
Text Augmentation: Rule-based

Operation	Sentence
원본 문장	A sad, superior human comedy played o ut on the back roads of life.
EDA: 동의어 교체	A lamentable, superior human comedy played out on the backward road of life.
EDA: 무작위 위치에 단어 삽입	A sad, superior human comedy played o ut oon funniness the back roads of life.
EDA: 문장 내 단어 순서 교체	A sad, superior human comedy played o ut on roads back the of life.
EDA: 문장 내 무작위 단어 삭제	A sad, superior human out on the roads of life.
AEDA: 문장 내 문장 부호 삽입	A sad, superior . human ! out on ? the r oads ; of life.

Text Augmentation: Model-based



Conditional BERT



Back Translation

Contribution

간단하고 비용 효율적이면서도
문장의 의미를 훼손하지 않는
데이터 증강 방식 제안

Methodology

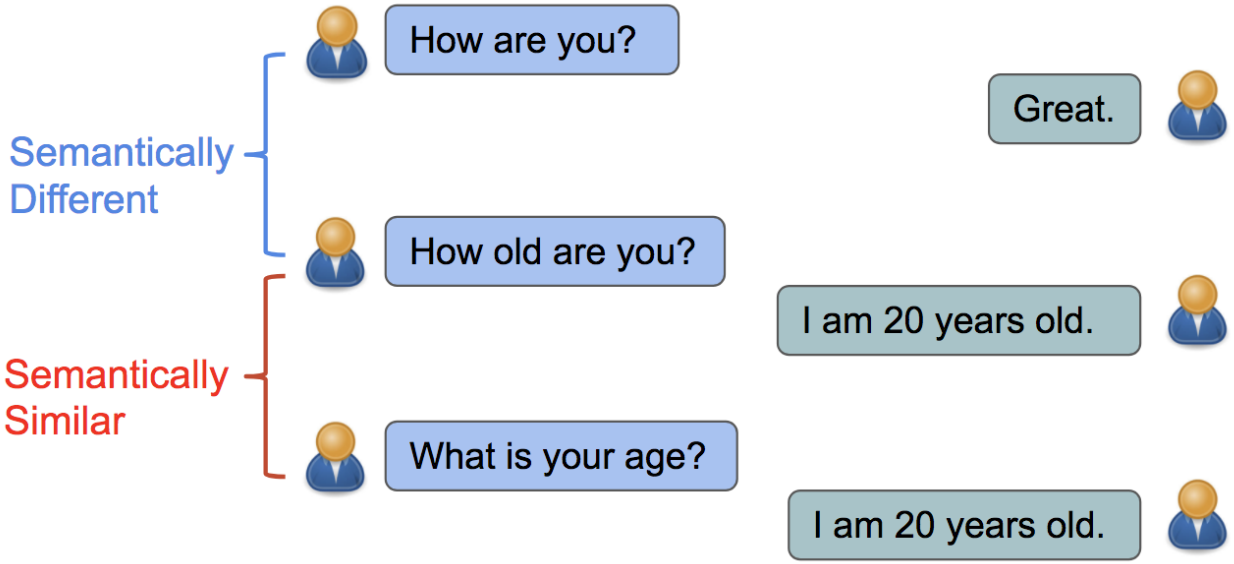
Text Similarity

Labeling

Methodology

데이터 세트 간의 유사도를 비교

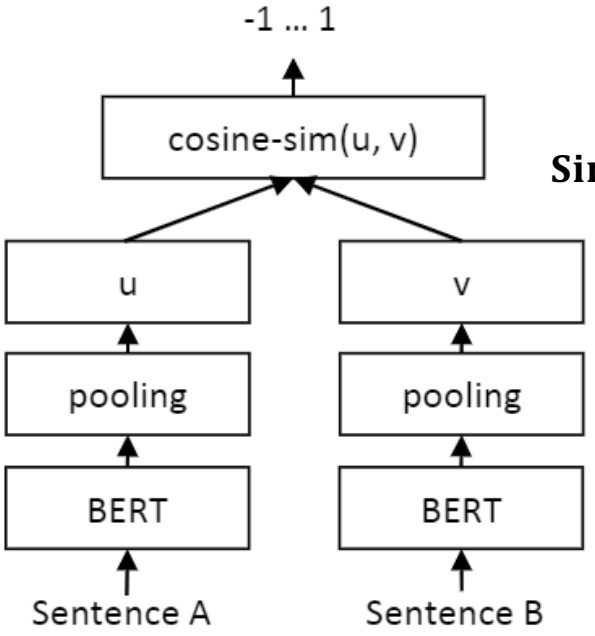
유사도가 낮은 데이터 세트를 활용



Methodology: Sentence-BERT

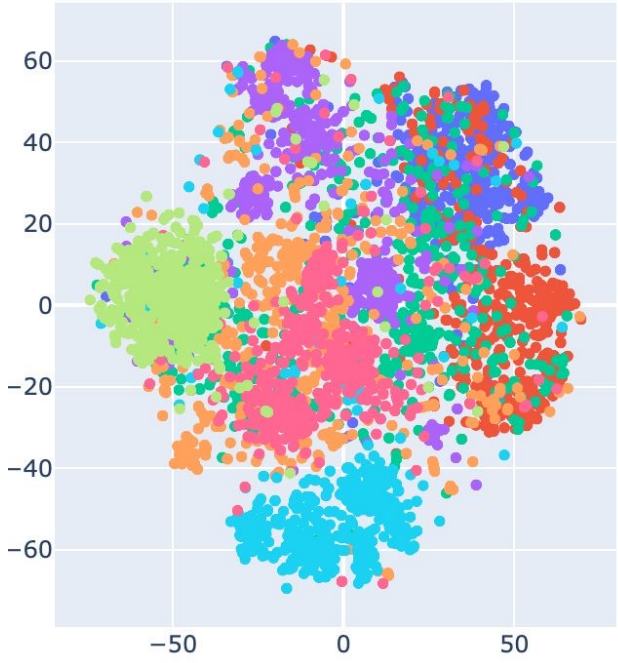
Sentence BERT(SBERT)

문장 임베딩을 도출해 코사인 유사도를 측정



$$\mathbf{D_{embedding}} = \frac{\sum \mathbf{SBERT(x_i)}}{|\mathbf{N}|}$$

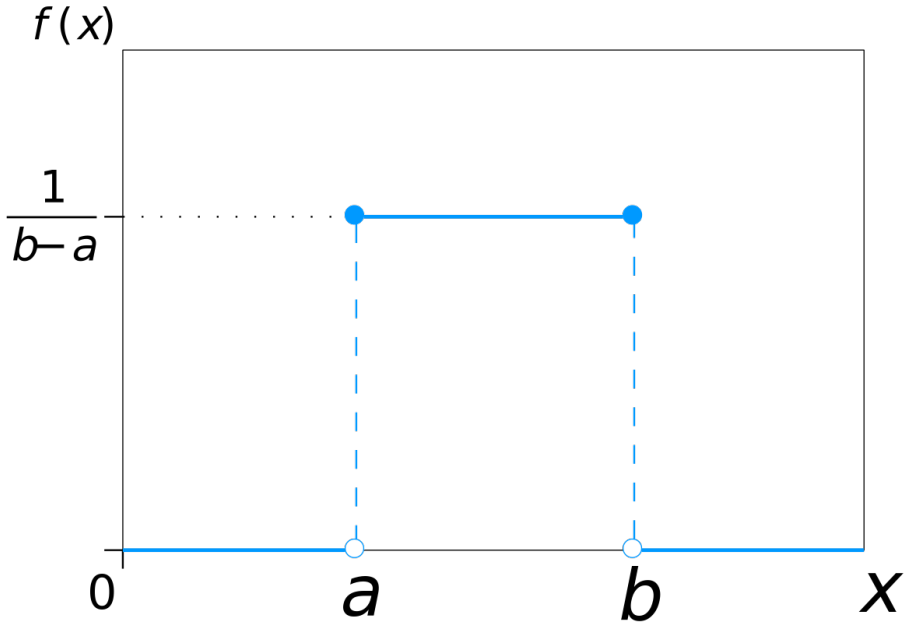
$$\mathbf{Similarity}\left(\mathbf{D^A_{emb}}, \mathbf{D^B_{emb}}\right) = \mathbf{D^A_{emb}} \cdot \mathbf{D^B_{emb}} (\mathbf{A} \neq \mathbf{B})$$



Methodology: Labeling

유사도가 낮은 데이터 세트를 학습에 추가

균일 분포 라벨값을 부여



Experiments

Experiment Results

Experiment: Task

제안하는 방법의 성능을 검증

텍스트 분류 작업에 적용

Experiment: Datasets

Korean Hate Speech Dataset

댓글 분류 작업

Categories	Example
Hate	이름도 희한하네 저놈 저거 인상 참 더럽게 생겼어 미생에서도 살인 충동 생기던데
Offensive	기안 분량좀... 다른사람나오면 노잼이라 안봄 한혜연은 왜자꾸나오는지..박나래로도 충분히 시끄러움
None	1,2화 어설펐는데 3,4화 지나서부터는 갈수록 너무 재밌던데

Datasets

NSMC: Naver Sentiment Movie Corpus

영화 리뷰 분류 작업

Categories	Example
Positive	설정이 재밌고 새로운 에피소드 내에서 메인 스토리도 차차 나오는게 재밌음
Negative	감독이 럼 먹고 영화 만들었나 보다.. 관객에게 뭘 말하는지도 모르겠고 엉망진창이다.

Datasets

KLUE Topic Classification

뉴스 주제 분류 작업

Categories	Example
Politics	국민의당 부산시당 발기인 대회 열려…26일 창당대회
Economy	작년 세탁기 분야 미국특허 LG전자·삼성전자 1~2위
Society	동탄2신도시 분양행복주택 건설에 민간 참여
Culture	울산 오후 4시 건조주의보
World	CNN 트럼프 내부고발자 상·하원 정보위 출석 동의
IT/Science	애플 퀀텀닷OLED 결합 차세대 디스플레이 특허출원
Sport	월드컵 하나은행 대표팀에 행운의 2달러 200장 선물

Experiment: Baseline Model

KCBERT를 활용

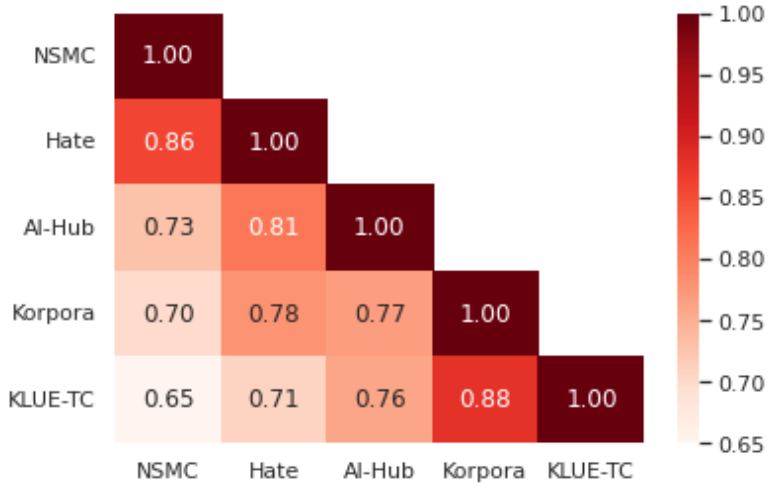
댓글 데이터를 통해 학습된 모델

	NSMC (acc)	Naver NER (F1)	PAWS (acc)	KorNLI (acc)	KorSTS (spearman)	Question Pair (acc)	KorQuaD (Dev)(EM/F1)
KcBERT-Base	89.62	84.34	66.95	74.85	75.57	93.93	60.25 / 84.39
KcBERT-Large	**90.68**	85.53	70.15	76.99	77.49	94.06	62.16 / 86.64
KoBERT	89.63	86.11	80.65	79.00	79.64	93.93	52.81 / 80.27
HanBERT	90.16	**87.31**	82.40	**80.89**	**83.33**	94.19	**78.74 / 92.02**
KoELECTRA-Base	90.21	86.87	81.90	80.85	83.21	**94.20**	61.10 / 89.59
XLM-Roberta-Base	89.49	86.26	**82.95**	79.92	79.09	93.53	64.70 / 88.94
DistilKoBERT	88.41	84.13	62.55	70.55	73.21	92.48	54.12 / 77.80

Experiment: Dataset Similarity

데이터 세트간 유사도 측정 결과

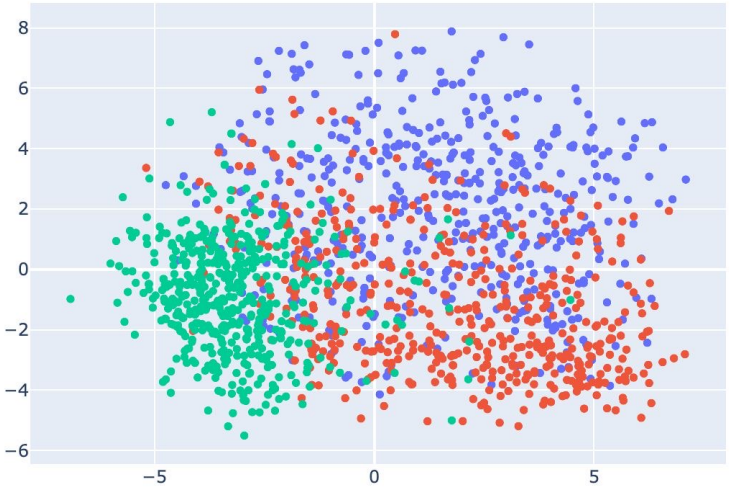
주제와 목적에 따라 유사도 차이



Experiment: Dataset Similarity

데이터 세트간 유사도 시각화 결과

유사도에 따른 시각화 결과 데이터 분포 차이



낮은 유사도

높은 유사도

Experiment: Results

각 데이터 세트에 대해 기존 기법과 비교

높은 수준의 성능 향상

기법 \ 데이터	Hate	NSMC	KLUE_TC
원본 (기준치)	64.57	88.92	85.66
라벨 스무딩	66.30	88.70	85.49
역번역	65.89	88.52	84.88
EDA	65.49	87.95	85.31
AEDA	66.16	88.45	85.43
익숙한 데이터	64.43	88.88	84.59
낯선 데이터	**68.55**	**89.04**	**85.68**

Experiment: Results

유사도가 높은 익숙한 데이터와 비교

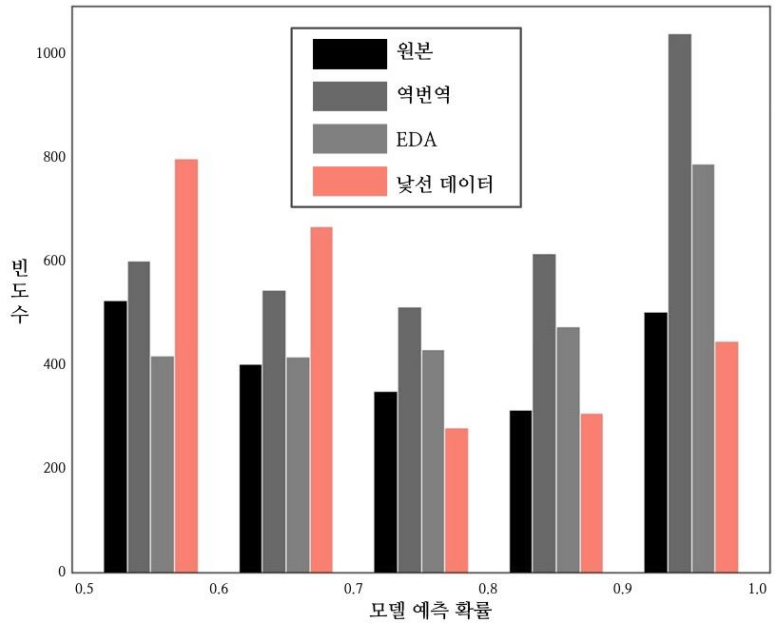
낯선 데이터 방식과 달리 성능 하락 확인

기법 \ 데이터	Hate	NSMC	KLUE_TC
원본 (기준치)	64.57	88.92	85.66
라벨 스무딩	66.30	88.70	85.49
역번역	65.89	88.52	84.88
EDA	65.49	87.95	85.31
AEDA	66.16	88.45	85.43
익숙한 데이터	64.43	88.88	84.59
낯선 데이터	**68.55**	**89.04**	**85.68**

Experiment: Results

과잉신뢰 현상에 대해 비교실험

기존 기법과 비교했을 때 높은 확률값으로 틀리는 빈도 감소



Conclusion

Conclusion

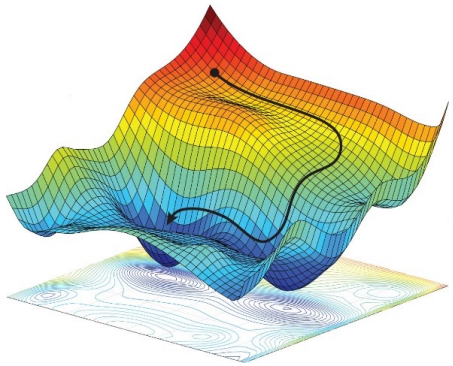
Conclusion

비용과 성능 양쪽 측면에서 효율적 데이터 증강

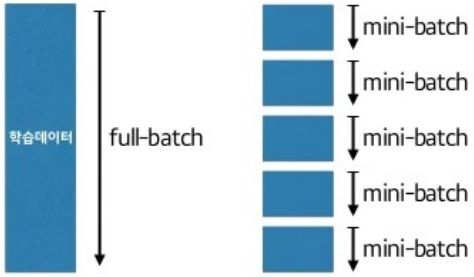
데이터 세트 유사도에 따른 낯선 데이터 활용

감사합니다.

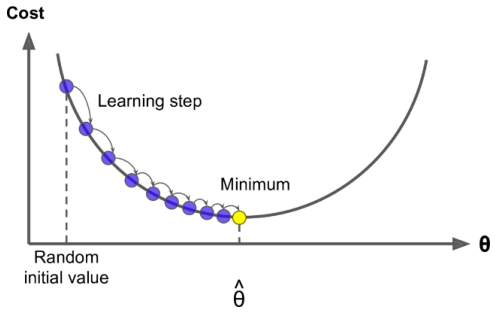
Appendix: Hyper-parameter



AdamW
Optimizer



240
Batch Size



5e-4
Learning Rate

5회 반복실험 후
평균

Appendix: Spurious Patterns

Measuring the tendency of CNNs to Learn Surface Statistical Regularities

Jason Jo
MILA, Université de Montréal
IVADO
jason.jo.research@gmail.com

Yoshua Bengio
MILA, Université de Montréal
CIFAR
yoshua.umontreal@gmail.com

Abstract

Deep CNNs are known to exhibit the following peculiarity: on the one hand they generalize extremely well to a test set, while on the other hand they are extremely sensitive to so-called adversarial perturbations. The extreme sensitivity of high performance CNNs to adversarial examples casts serious doubt that these networks are learning high level abstractions in the dataset. We are concerned with the following question: How can a deep CNN that does not learn any high level semantics of the dataset manage to generalize so well? The goal of this article is to measure the tendency of CNNs to learn surface statistical regularities of the dataset. To this end, we use Fourier filtering to construct datasets which share the exact same high level abstractions but exhibit qualitatively different surface statistical regularities. For the SVHN and CIFAR-10 datasets, we present two Fourier filtered variants: a low frequency variant and a randomly filtered variant. Each of the Fourier filtering schemes is tuned to preserve the recognizability of the objects. Our main finding is that CNNs exhibit a tendency to latch onto the Fourier image statistics of the training dataset, sometimes exhibiting up to a 28% generalization gap across the various test sets. Moreover, we observe that significantly increasing the depth of a network has a very marginal impact on closing the aforementioned generalization gap. Thus we provide quantitative evidence supporting the hypothesis that deep CNNs tend to learn surface statistical regularities in the dataset rather than higher-level abstract concepts.

1. Introduction

The generalization ability of a machine learning model can be measured by evaluating its accuracy on a withheld test set. For visual learning tasks, convolutional neural networks (CNNs) [21] have become the de facto machine learning model. These CNNs have achieved record breaking object recognition performance for the CIFAR-10 [1], SVHN [26] and ImageNet [34] datasets, at times surpassing human performance [13]. Therefore, on the one hand, very deep CNN architectures have been designed which obtain very good generalization performance. On the other hand, it has been shown that these same CNNs exhibit an extreme sensitivity to so-called *adversarial examples* [40]. These adversarial examples are perceptually quite similar to the original, "clean" image. Indeed humans are able to correctly classify the adversarial image with relative ease, whereas the CNNs predict the wrong label, usually with very high confidence. The sensitivity of high performance CNNs to adversarial examples casts serious doubt that these networks are actually learning high level abstract concepts [6, 27]. This begs the following question: *How can a network that is not learning high level abstract concepts manage to generalize so well?*

Roughly speaking, there are two ways in which a machine learning model can generalize well. The first way is the ideal way: the model is trained in a manner that captures high level abstractions in the dataset. The second way is less than ideal: the model has a tendency to overfit to superficial cues that are actually present in both the train and test datasets; thus the statistical properties of the dataset plays a key role. In this fashion, high performance generalization is possible without actually explicitly learning any high level concepts.

In Section 2 we discuss the generalization ability of a machine learning model and its relation to the surface statistical regularities of the dataset. In particular, by drawing on computer vision literature on the statistics of natural images, we believe that it is possible for natural image train and test datasets to share *many superficial cues*. This leads us to postulate our main hypothesis: ***the current incarnation of deep neural networks has a tendency to learn surface statistical regularities in the dataset.*** In Section 3 we discuss related work.

To test our hypothesis, we will quantitatively measure this tendency. To this end, for a dataset X it is sufficient to construct a perturbation map F:

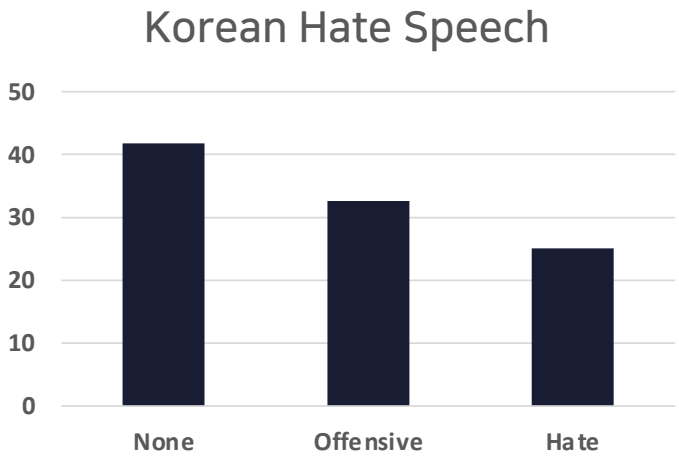
$$F : X \mapsto X', \tag{1}$$

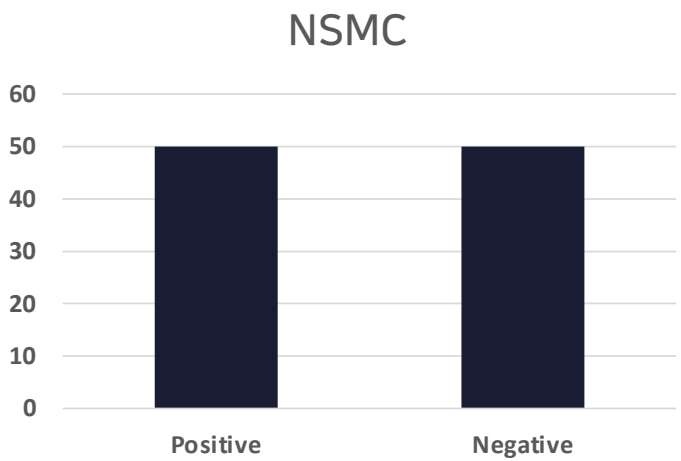
which satisfies the following properties:

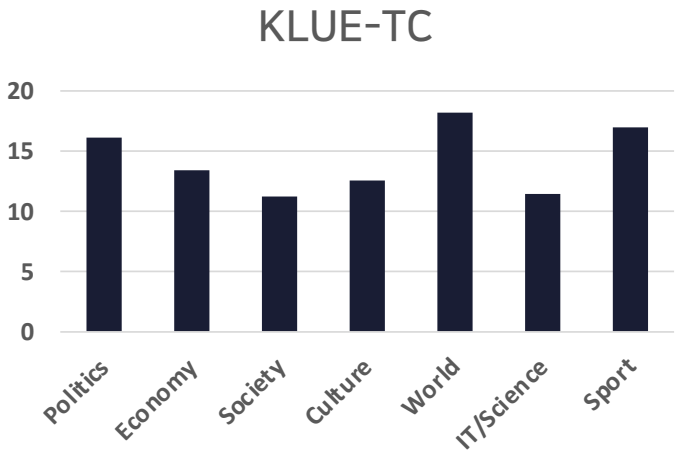
Jason Jo and Yoshua Bengio
ArXiv Preprint 2017

"Clues in the image's background to recognize foreground objects even when that seems both unnecessary and somehow wrong: the beach is not what makes a seagull a seagull."

Appendix: Dataset Statistics







Appendix: Future Work

본 연구에서는 균일한 라벨값을 가정

학습 데이터 분포를 고려한 라벨값 부여