Variational Autoencoder 기반 의미 보존 자연어 데이터 증강 기법

- 최주환[1], 이준호[2], 진교훈[3], 장예훈[3], 장수진[3], 김영빈[3]

1 중앙대학교 전자전기공학부
2 중앙대학교 AI학과
3 중앙대학교 첨단영상대학원

Juhwan Choi

gold5230@cau.ac.kr

Intelligent Information Processing Lab.

IIPL

Index

Introduction

Importance of Data

Overfitting

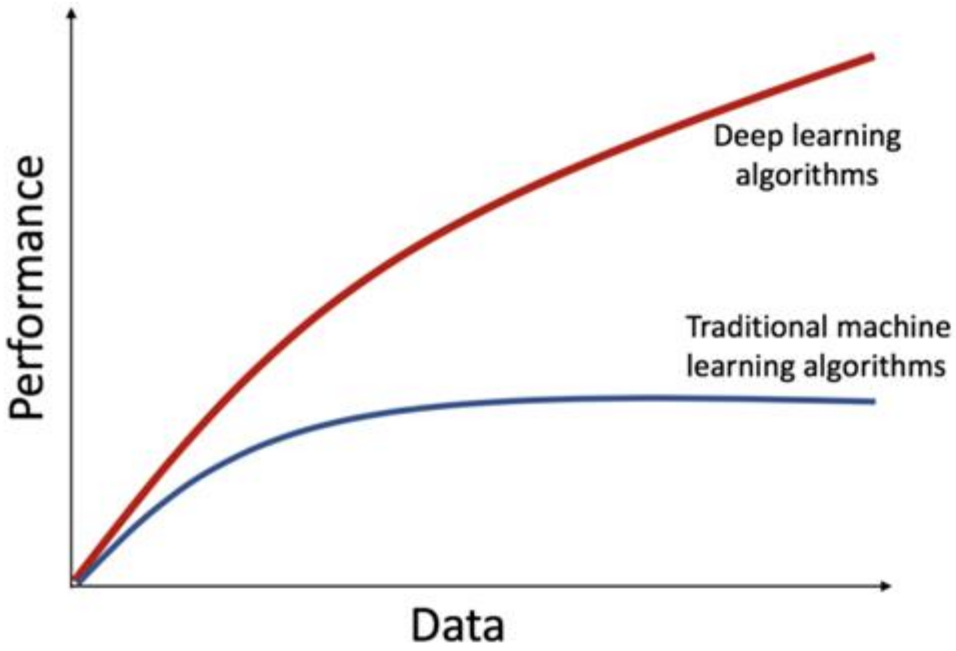
Data Augmentation

NLP Data Augmentation

Importance of Data

Deep Learning의 성능에 가장 큰 영향을 미치는 요소

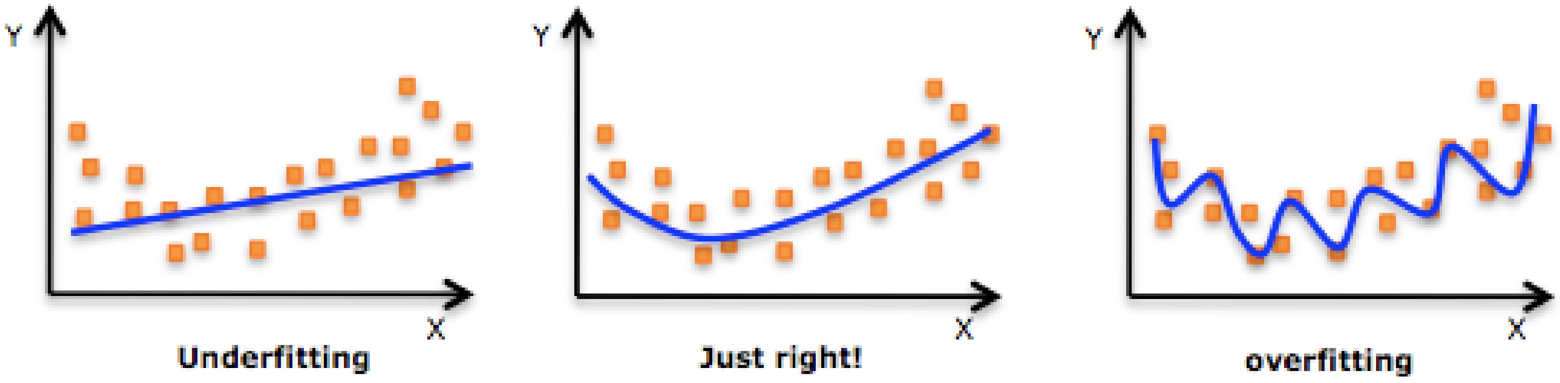
충분한 양의 데이터



Overfitting

학습에 주어진 데이터가 부족할 경우

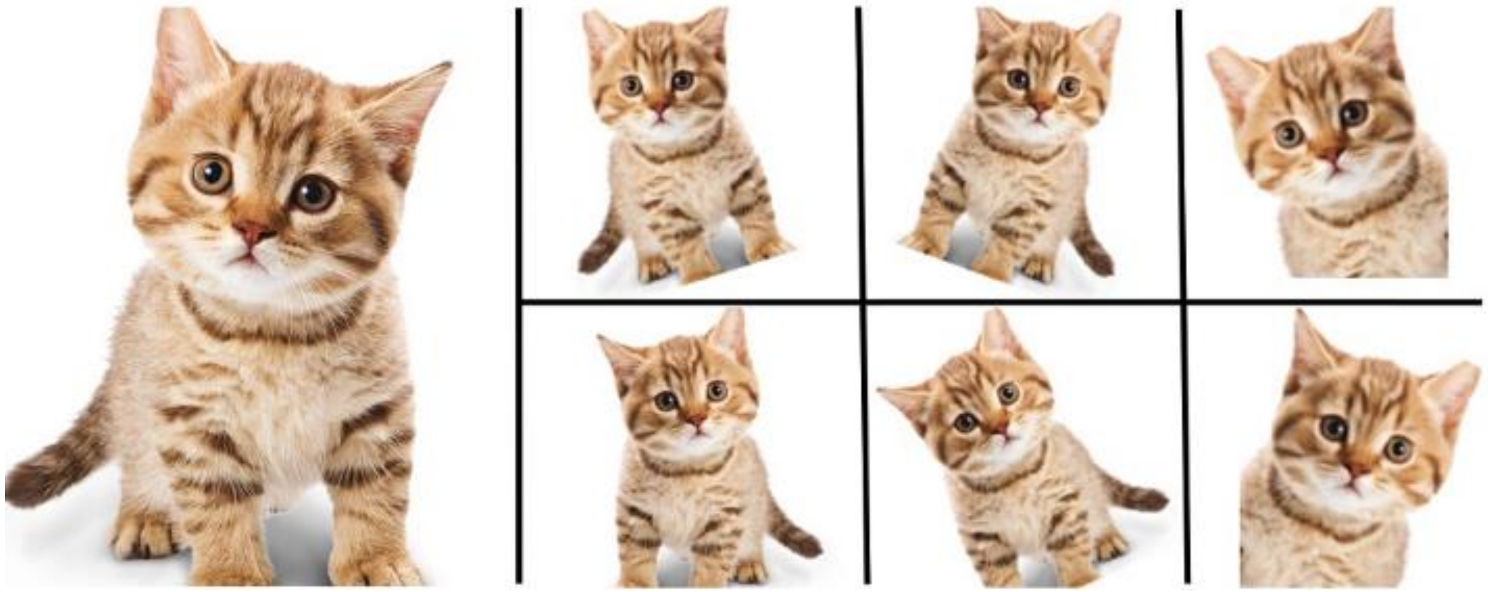
Overfitting(과적합)의 가능
성



Data Augmentation

Overfitting을 해결하기 위한 가장 일반적 방법

Data Augmentation(데이터 증가)

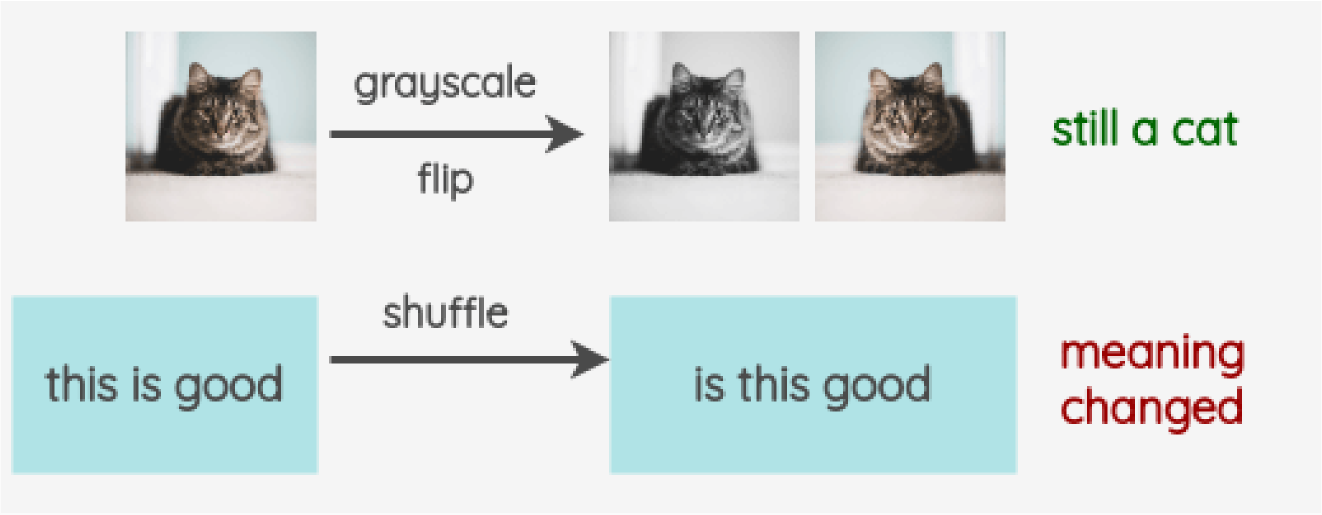


NLP Data Augmentation

Text Data Augmentation의 특징

Semantic 정보 보존이 중요

Challenge of Semantically Invariant Transformation in NLP



Related Work

Back-Translation

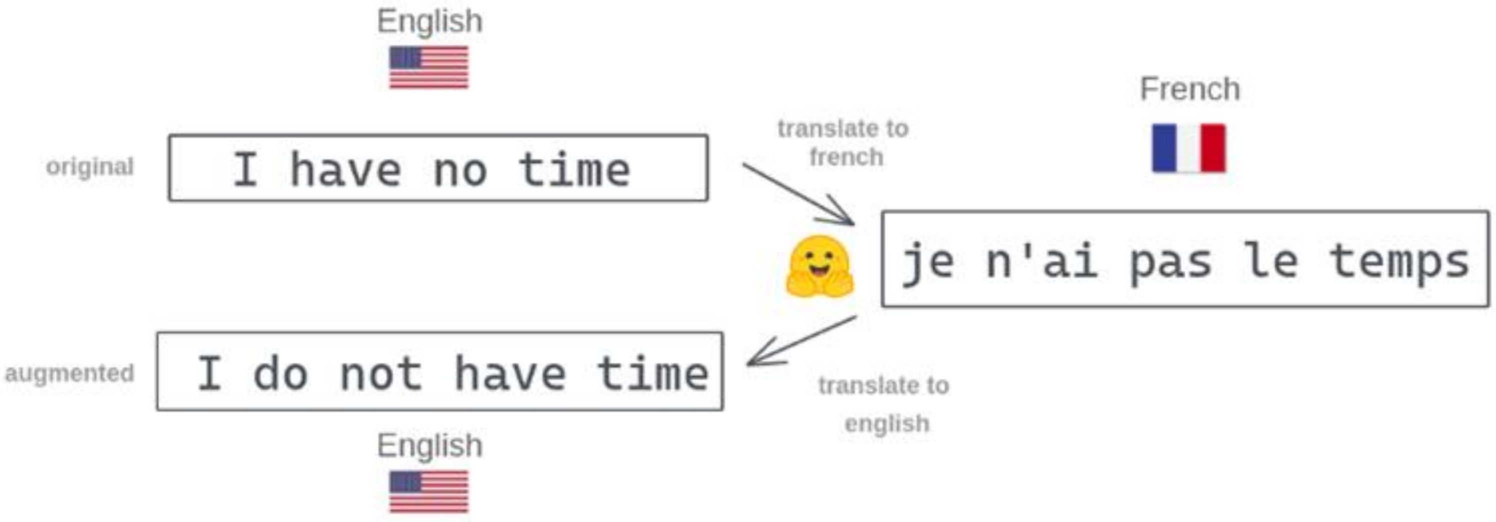
EDA: Easy Data Augmentation

Conditional BERT Contextual Augmentation

Back-Translation

다른 언어로 번역 후 다시 원래 언어로 번역

2개의 번역 모델 학습이 필요



Understanding Back-Translation at Scale
Edunov et al., EMNLP 2018

EDA: Easy Data Augmentation

무작위로 문장의 단어를 선택

문장의 의미가 훼손될 가능성

동의어로 교체

무작위 위치에 단어 삽입

순서를 교체

단어를 삭제

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.

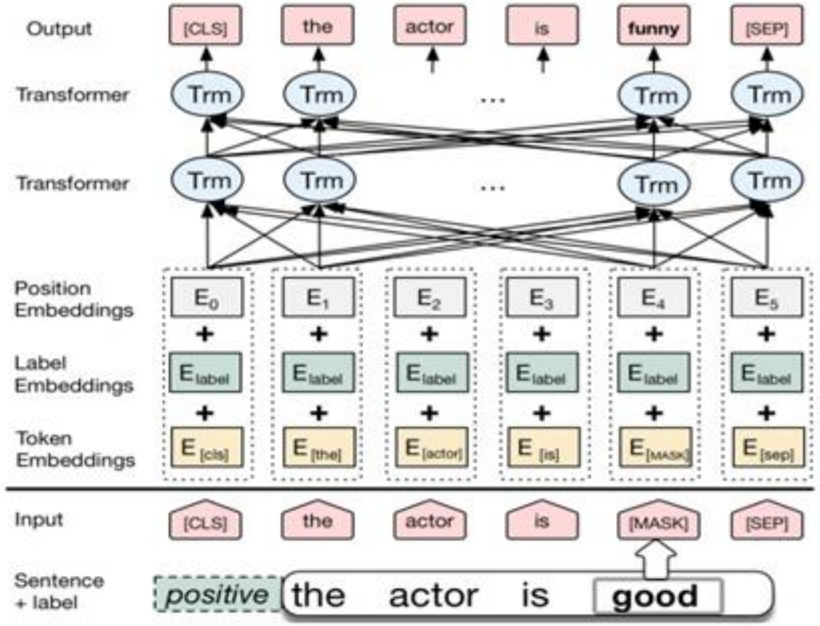
EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks

Wei et al., EMNLP 2019

Conditional BERT Contextual

Pretrained Language Model을 활용

Fine-tuning이 필요



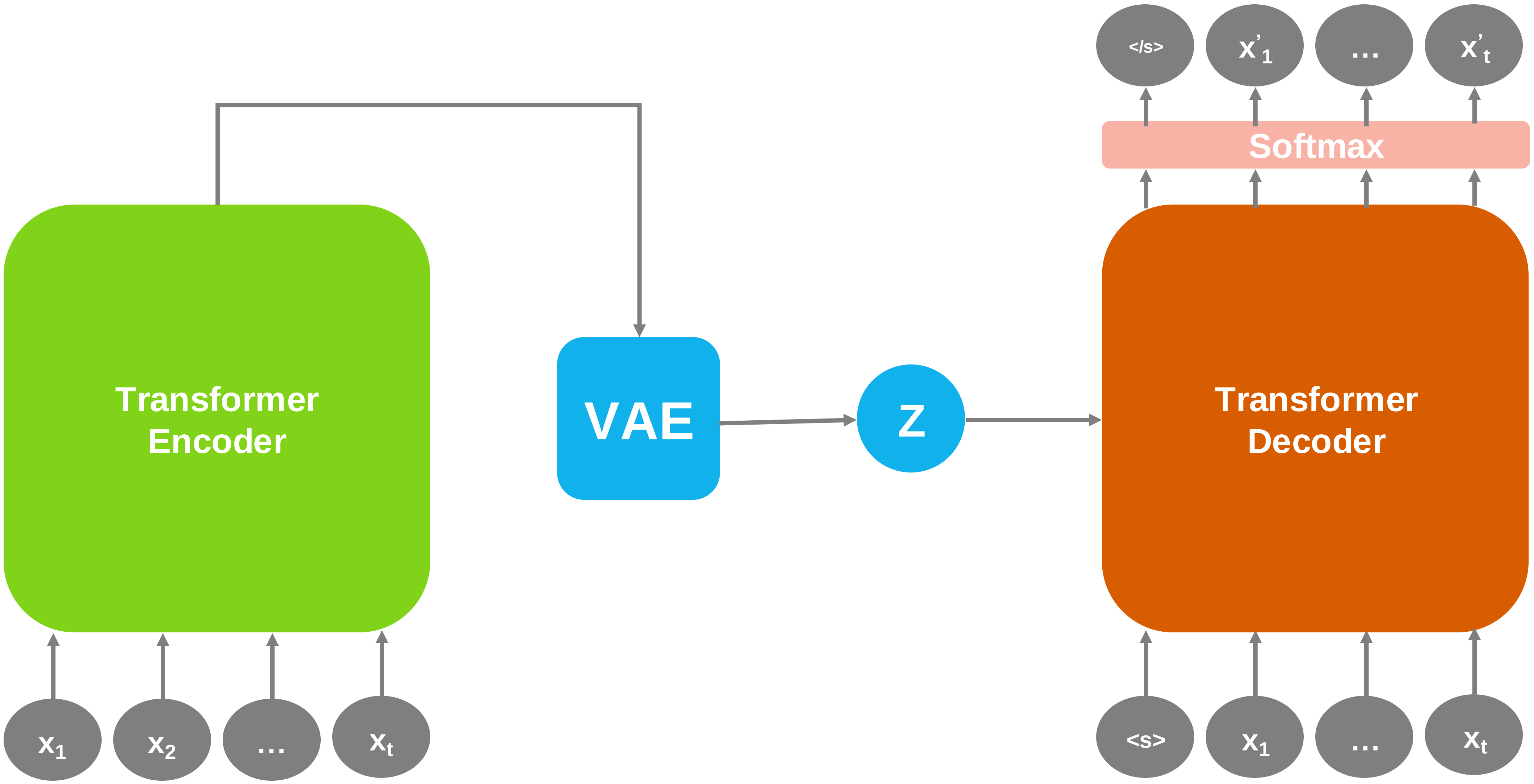
Conditional BERT Contextual Augmentation
Wu et al., arXiv:1812.06705

Method

Model Structure

Variational Autoencoder

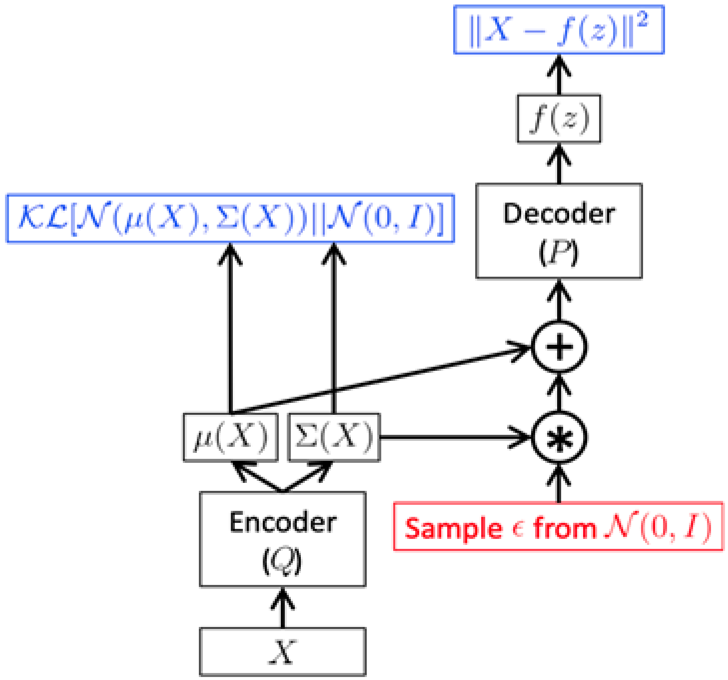
Model Structure



Variational Autoencoder

Latent Vector를 추출하여 원래 입력을 복원

Semantic을 포함한 정보를 추
추



Experiments

Text Classification

Comparison with EDA

Text Classification

제안하는 방법의 성능을 검증

텍스트 분류 작업에 적용

Datasets

	IMDB	Yelp_5	ProsCons	MR
Subject	Movie Review	Business Review	Product Review	Movie Review
Number of Sentences	50,000	650,000	39,419	9,594
Number of Classes	Binary (Pos / Neg)	5-Classes	Binary (Pros / Cons)	Binary (Pos / Neg)

Comparison with EDA

	IMDB	Yelp_5	ProsCons	MR
Baseline	91.95%	65.52%	93.65%	84.05%
EDA	90.98% (-0.97%p)	67.92% (+2.40%p)	94.21% (+0.56%p)	84.18% (+0.13%p)
Proposed Model	94.39% (+2.44%p)	70.65% (+5.13%p)	95.16% (+1.51%p)	84.27% (+0.22%p)

Comparison with EDA

	Text
Original	I **loved** this movie since I was **7** and I saw it on the opening day. It was so touching and beautiful. I strongly recommend seeing **for all**. It's a movie to watch with your family by far.
EDA	I this movie since I was and I saw it on the opening day. It was so touching and beautiful. I strongly recommend seeing ***disastor***. It's some movie to watch with your family by far.
Proposed Model	I **loved** this movie since I was **9** and I saw it on the opening day. It was so touching and beautiful. I recommend seeing **for all**. It's a movie to watch with your family by far.

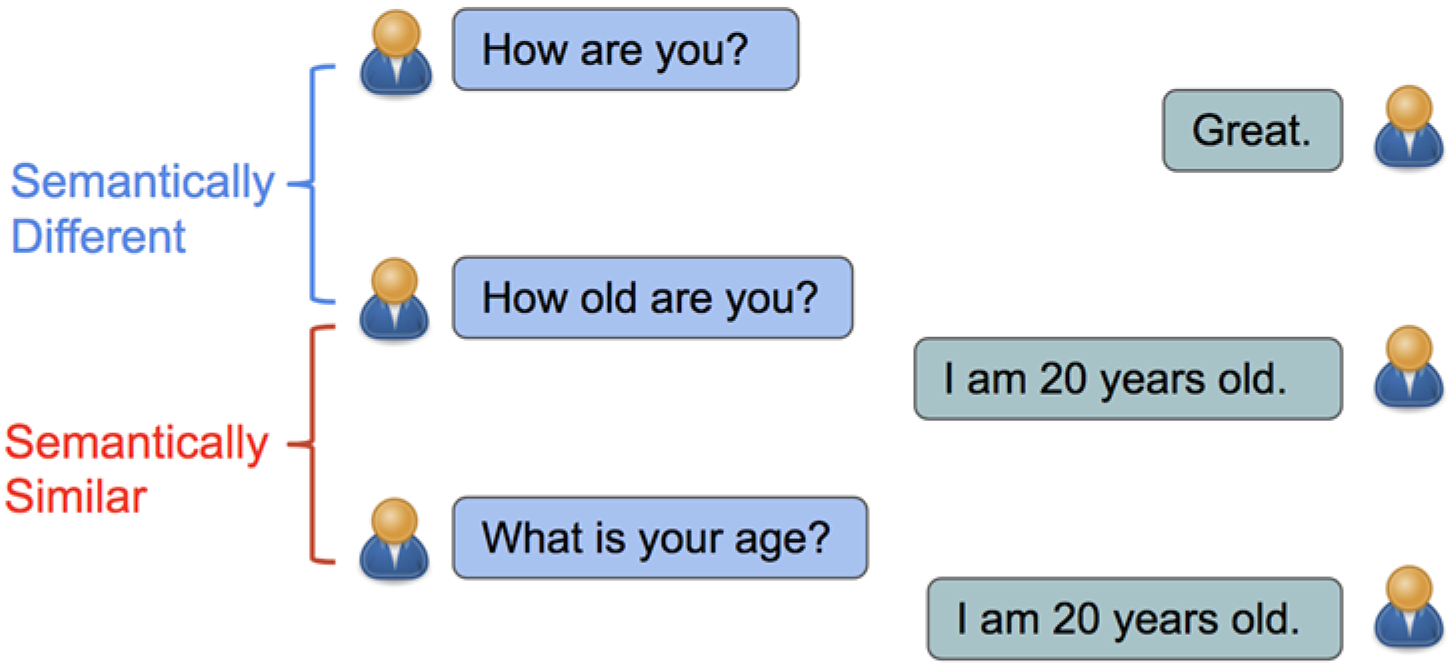
Conclusion

Contribution

Future work

Contribution

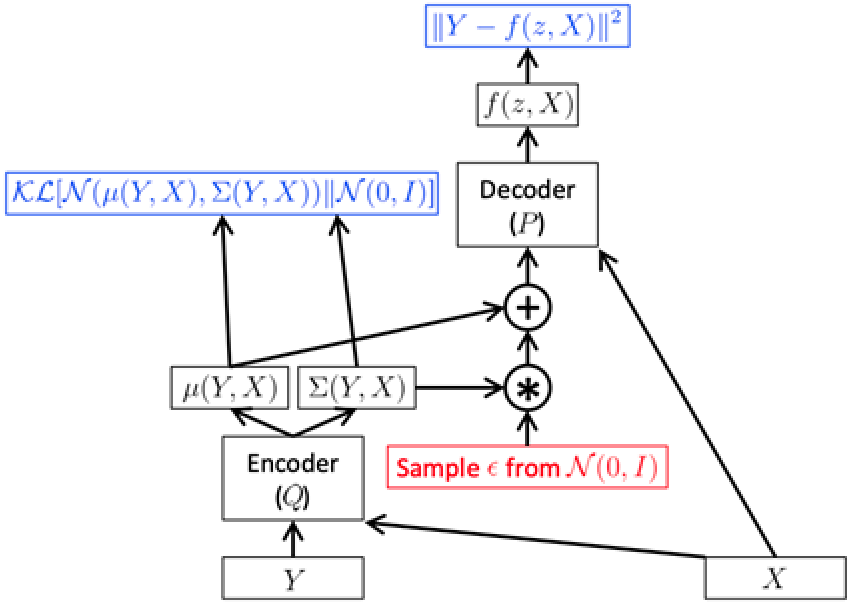
Semantic 정보를 보존하는 Text Data Augmentation



Future work

Conditional Variational Autoencoder (CVAE)

Label 정보를 직접 주입



감사합니다.

Intelligent
Information
Processing
Lab.

IIPL

Q&A

Intelligent
Information
Processing
Lab.

IIPL