Unknown 문서 탐지 – 3위

문제 정의

- 모델이 학습하지 않은 종류의 문서를 탐지하는 Task

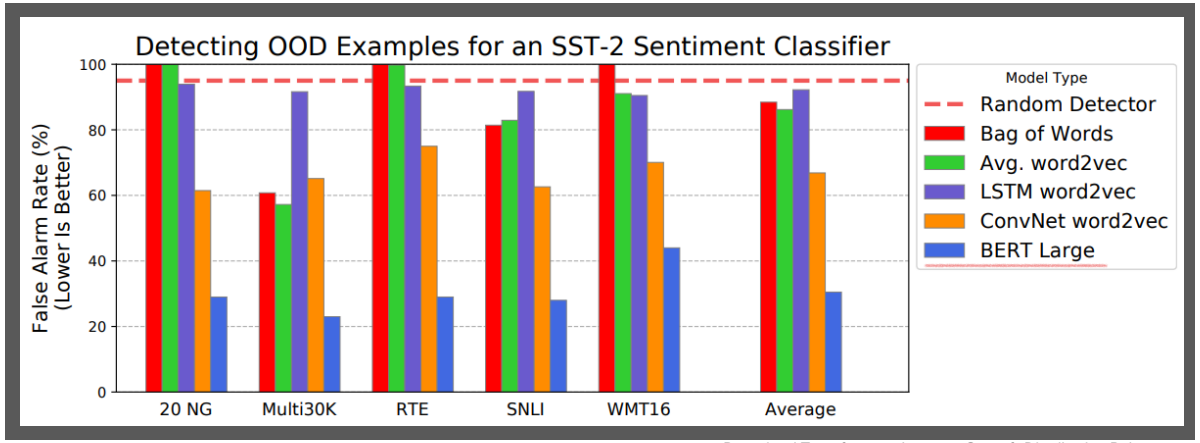
- 6개의 Class로 이루어진 Text Data를 학습
  ⇒ Test 시, 학습한 Class에 대해서는 옳게 분류, 학습하지 않은 Class에 대해서는 Unknown으로 분류
  ⇒ Multiclass Classification + Unknown Detection 문제

핵심 아이디어

1. Pretrained Transformer Model 활용
2. Diverse Loss Function
3. Preprocessing
4. Ensemble

Pretrained Transformer Model 활용

1. Pretrained Transformer Model을 활용할 경우 성능 개선



Pretrained Transformers Improve Out-of-Distribution Robustness

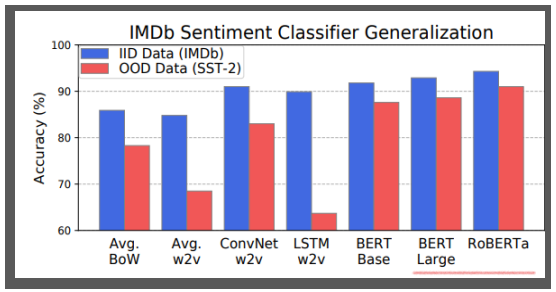
Out of Distribution (OOD) Task에서
Pretrained Transformer Model이 높은 성능을 나타냄.

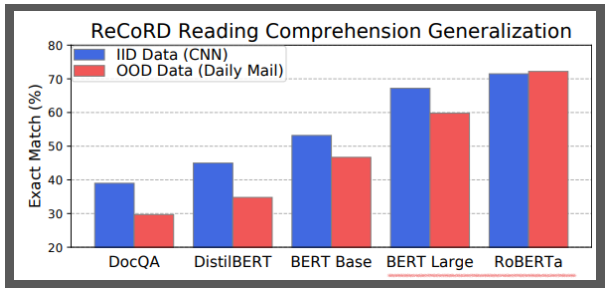
Backbone Model로서 BERT 계열 Model 채택

Baseline (0.7828) ⇒ KLUE/BERT-base (0.911)

MSP 0.85 기준, Micro F1 Score

2. 더 많은 Data로 Pretrain 할 경우 성능 개선





Pretrained Transformers Improve Out-of-Distribution Robustness

RoBERTa가 BERT Large에 비해 OOD Detection 성능이 우수 (둘의 주목할 만한 차이는 Data의 수)
⇒ 즉, 좀 더 많은 Data로 Pretrain할 경우, OOD Detection에서 우수한 성능을 보임

KPF-BERT를 Backbone Model로 선정 · KLUE: Sentence 기준 473M · KPF: 뉴스 기사 기준 약4천만

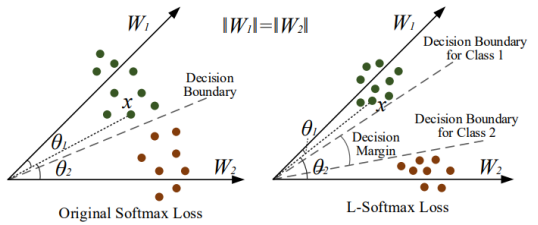
KLUE/BERT-base (0.9607) ⇒ KPF-BERT (0.9625)

MSP 1600th element 기준, Micro F1 Score

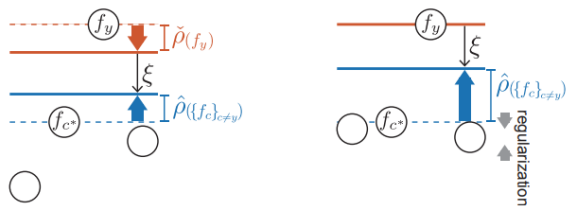
Diverse Loss Function

1. Large Margin Softmax Loss 활용

Basic Idea



+ Regularizaiton



Original LargeMargin　　Regularized LargeMargin

Large Margin Softmax Loss 활용

(Large Margin Loss? 정답 Class와 정답이 아닌 Class 사이의 Margin을 줌으로서 정답 Class를 더 확실하게 맞추도록 하는 방법론)

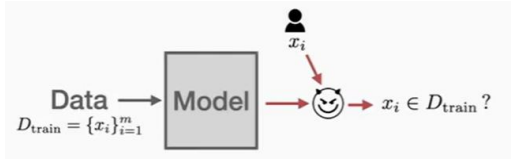
⇒ In-Distribution (ID) Dataset에 대한 Softmax Logit값이 Out-of-Distribution (OOD) Dataset에 비해 높아질 것으로 기대

Cross Entropy Loss (0.9674) ⇒ Large Margin Softmax Loss (0.9697)

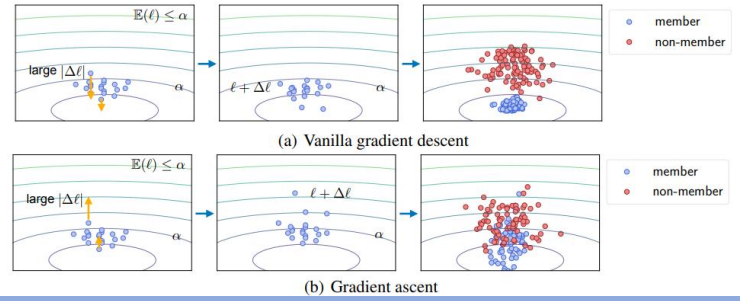
MSP 1600th element 기준, Micro F1 Score

2. Relax Loss 활용

Membership Inference Attacks



Relax Loss



Relax Loss 활용

(Relax Loss? MIAs에 대한 Defense 방법론으로서 Gradient Descent와 Ascent를 혼합하여 학습함으로서 Classification은 잘 수행하면서 Member or Non-Member를 구별하기 힘들도록 만든 기법)

⇒ ID Sample, OOD Sample 모두에서 Confidence Logit값을 완화
⇒ MSP에서의 전반적인 Threshold Logit을 완화시켜 주어 ID, OOD의 구분을 더 잘 할 것으로 기대

Cross Entropy Loss (0.9674) ⇒ Relax Loss (0.972)

MSP 1600th element 기준, Micro F1 Score

Large-Margin Softmax Loss for Convolutional Neural Networks / Large Margin In Softmax Cross-Entropy Loss
RELAXLOSS: DEFENDING MEMBERSHIP INFERENCE ATTACKS WITHOUT LOSING UTILITY

Preprocessing

1. Hanspell을 활용한 맞춤법 검사

```
original_text = "안녕하 세요, 저는 한국인입니다."
new_text = spell_check_using_hanspell([original_text])

print("Original Text:", original_text)
print("Replaced Text:", new_text)
```

```
100% 1/1 [00:00<00:00, 2.08it/s]
Original Text: 안녕하 세요, 저는 한국인입니다.
Replaced Text: ['안녕하세요, 저는 한국인입니다.']
```

- 띄어쓰기 및 맞춤법이 올바르지 않은 경우 수정
- Train Dataset 및 Test Dataset에 모두 적용

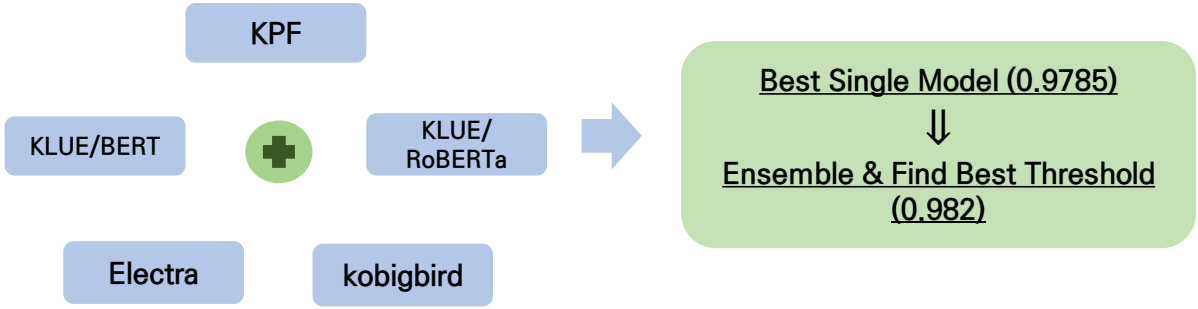
Not Preprocessed (0.972) ⇒ Preprocessed (0.9785)

MSP 1600th element 기준, Micro F1 Score

https://github.com/ssut/py-hanspell

Ensemble

1. 다양한 Backbone Model 활용
2. Logit Ensemble 수행 (Soft Ensemble)
3. Find Best Threshold



HyperParameter Settings

HyperParameters	
Batch Size	32
Max Length	512
Learning Rate	2e-5
Epochs	2
Optimizer	AdamW
Scheduler	Decrease Linearly
Seed	42

Best Single Model 기준

Optimizer & Scheduler	
Betas & Eps	Default
Weight decay	0.01
Warmup steps	100
Training steps	Total steps

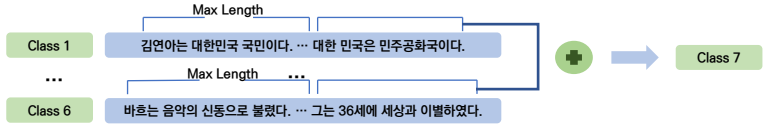
Relax Loss	
Alpha	0.01
Ascent Step	300

Appendix

1. 시도하였으나 유의미한 성능 향상으로 이어지지 않음

Another Class 활용 (7th Class)

- 활용하지 않은 Data를 조합하여 7th class 제작
- Test 시에 7th로 분류할 경우 Unknown 판정

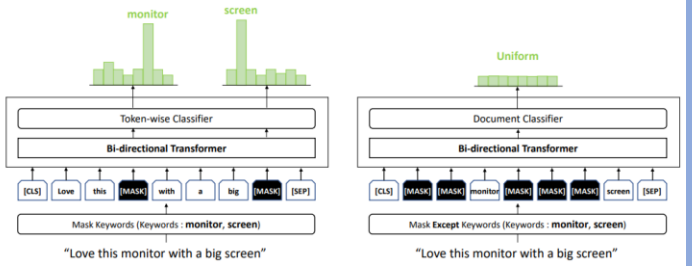


Diverse Preprocessing

- 중복 문장 제거
- 개인 정보 삭제
- 특수문자 및 기호 삭제 등

MASKER

- Attention or TF-IDF을 활용하여 가장 유의한 Token 추출
- 유의한 Token을 Masking한 후 Token-wise Classification Loss를 추가로 활용

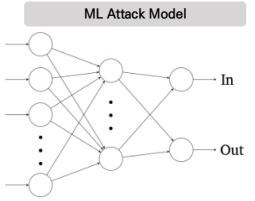


Membership Inference Attacks 방법론 응용

- Stratified KFold를 사용하여 Dataset 분배 후 각 Fold 별 Model을 Shadow Model로서 활용
- K개의 Shadow Model에서 추출된 Logit값들을 ML Model에 입력으로 주어 Member or Non-Member Classification 수행

Stratified KFold Shadow Model

Confidence levels	Label
0.89, …, 0.012	In
0.86, …, 0.013	Out
.	.
.	.
.	.
0.98, …, 0.001	In



An Effective Baseline for Robustness to Distributional Shift / DEEP ANOMALY DETECTION WITH OUTLIER EXPOSURE
MASKER: Masked Keyword Regularization for Reliable Text Classification

Reference

Pretrained Transformers Improve Out-of-Distribution Robustness (https://aclanthology.org/2020.acl-main.244.pdf)
KLUE: Korean Language Understanding Evaluation (https://arxiv.org/pdf/2105.09680v4.pdf)
KPF-BERT (https://github.com/KPFBERT/kpfbert/blob/main/BERT-MediaNavi.pdf)

Large-Margin Softmax Loss for Convolutional Neural Networks (https://arxiv.org/pdf/1612.02295.pdf)
Large Margin In Softmax Cross-Entropy Loss (https://staff.aist.go.jp/takumi.kobayashi/publication/2019/BMVC2019.pdf)

Membership Inference Attacks Against Machine Learning Models (https://arxiv.org/pdf/1610.05820.pdf)
Demystifying the Membership Inference Attack(https://medium.com/disaitek/demystifying-the-membership-inference-attack-e33e510a0c39)
RELAXLOSS: DEFENDING MEMBERSHIP INFERENCE ATTACKS WITHOUT LOSING UTILITY (https://arxiv.org/pdf/2207.05801.pdf)
[ICLR 2022 spotlight] RelaxLoss: Defending Membership Inference Attacks without Losing Utility (https://www.youtube.com/watch?v=Iyu0gNC3oYE&t=202s)

py-hanspell (https://github.com/ssut/py-hanspell)

An Effective Baseline for Robustness to Distributional Shift (https://arxiv.org/pdf/2105.07107.pdf)
DEEP ANOMALY DETECTION WITH OUTLIER EXPOSURE (https://arxiv.org/pdf/1812.04606.pdf)
MASKER: Masked Keyword Regularization for Reliable Text Classification (https://arxiv.org/abs/2012.09392)