Contextual Embeddings for Hypernym Discovery

Geonil Yun

rjsdlf45@gmail.com

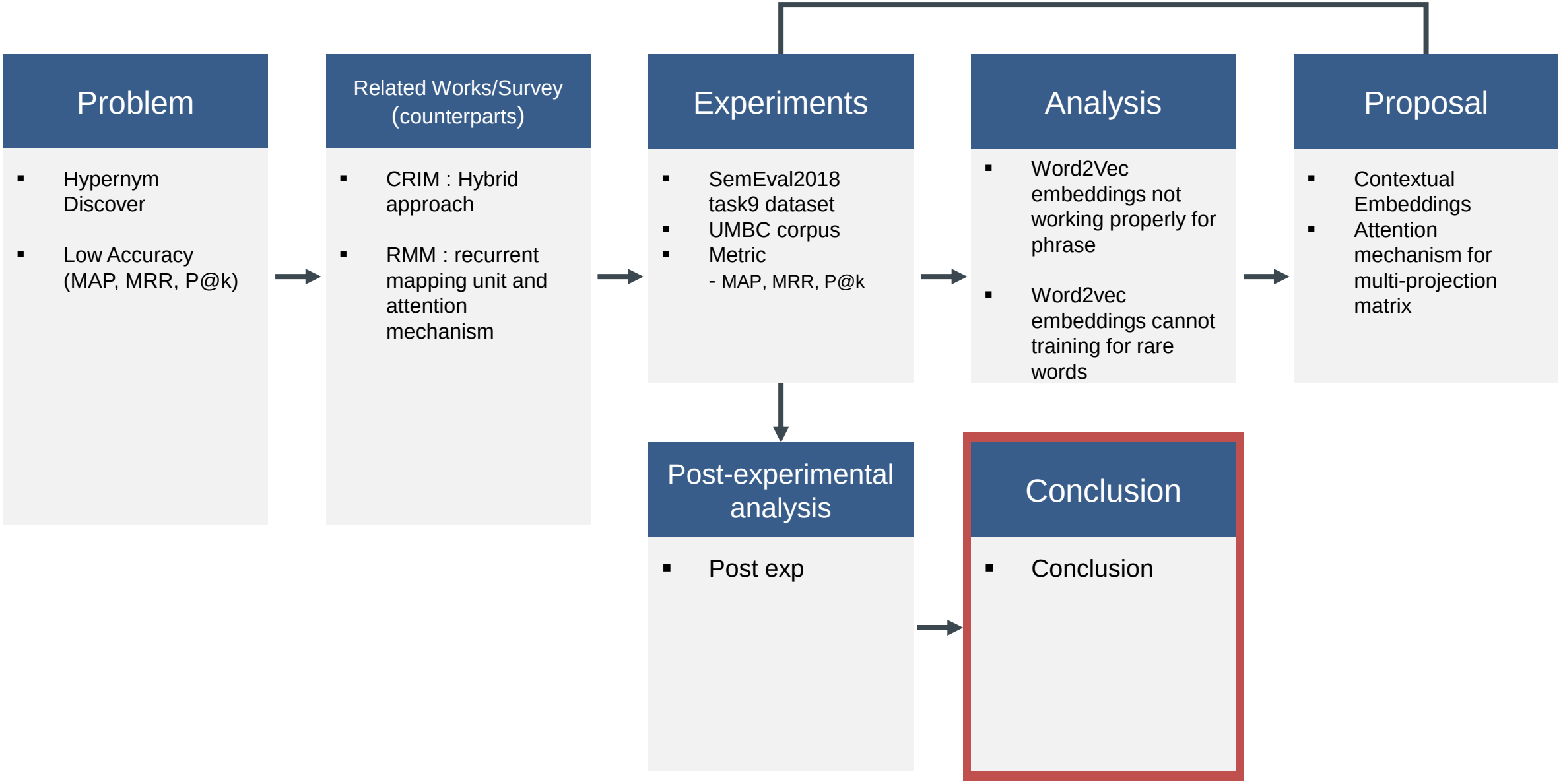
5th Oct, 2022

Contents

To Do List from Previous Seminar

- ✓ Experiments – repeat 30 times
- ✓ Experiments – using 2A, 2B subtasks

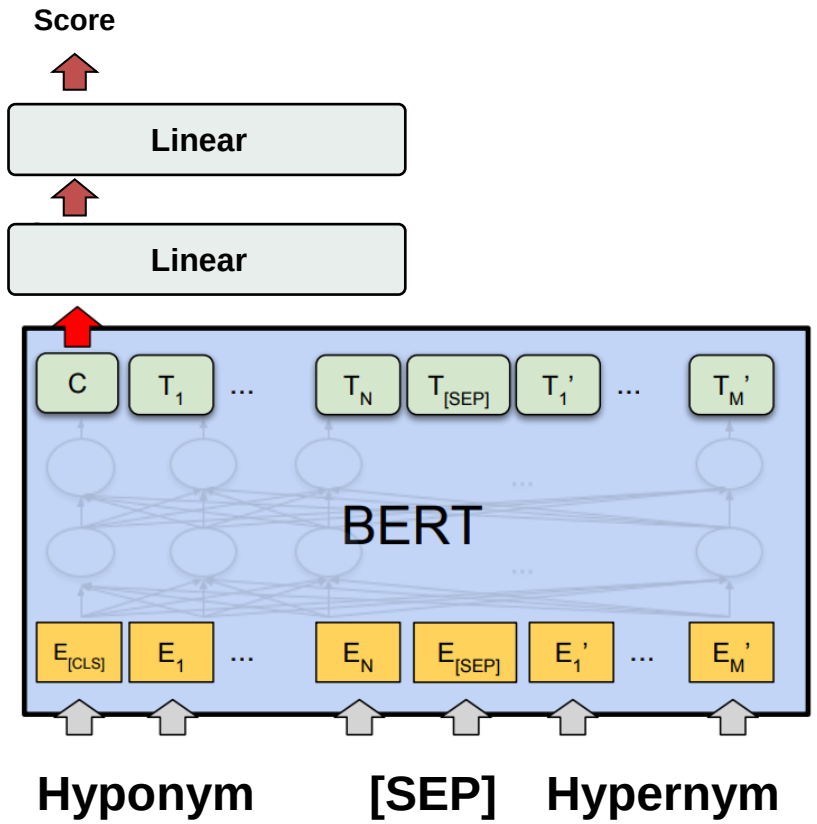
Overall Progress



Proposed Method

Previous Method

- Only use 2 linear layer for prediction
- In recent study uses projection learning with word embeddings



Proposed Method

Contextual Embeddings for Projection Learning

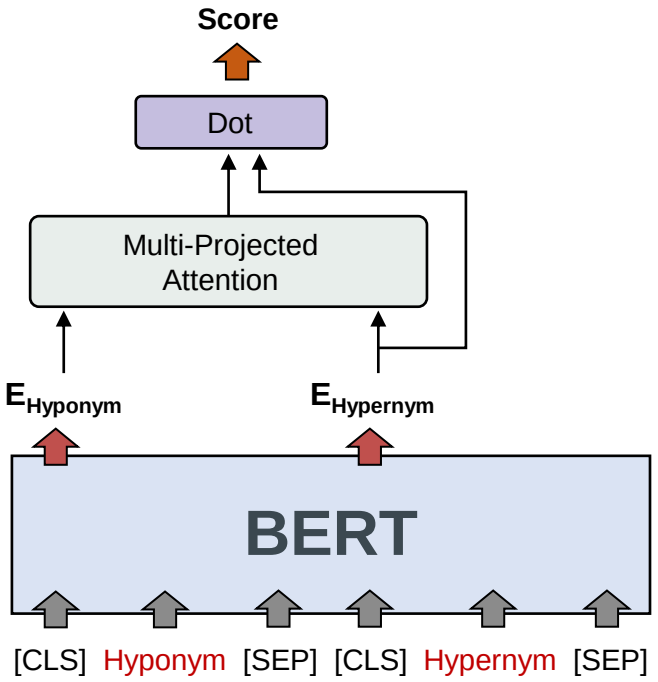


Figure 1. Proposed Method – model architecture

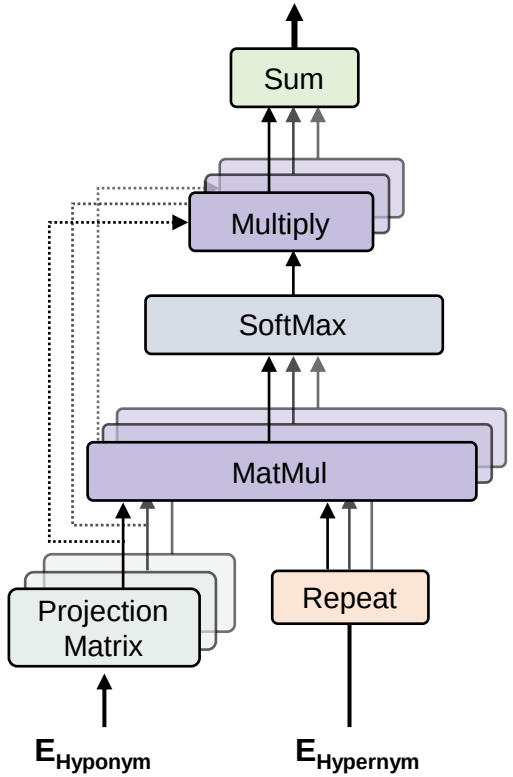


Figure 2. Multi-Projected Attention

$$P_i = \phi_i \cdot e_{\text{hyponym}}$$

$$\alpha_i = \frac{\exp\left(P_i \cdot e_{\text{hypernym}}\right)}{\sum_{k=1}^{n} \exp\left(P_k \cdot e_{\text{hypernym}}\right)}$$

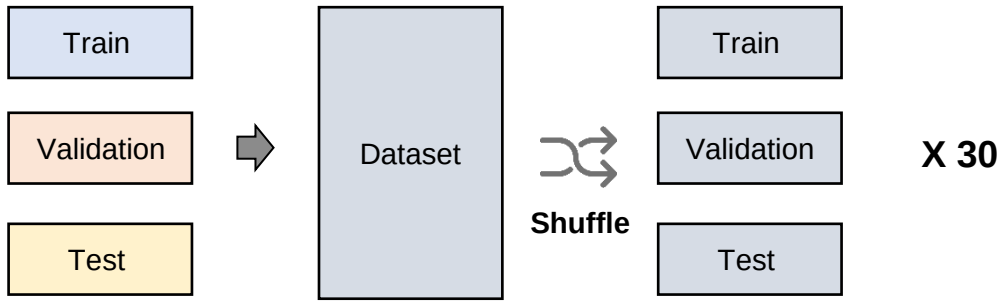
$$H = \sum_{i=1}^{n} \alpha_i \cdot P_i$$

$$s = H \cdot e_{\text{hypernym}}$$

Experimental Settings

Dataset

- The experiment was repeated 30 times with the shuffled dataset .
- We used 3 sub-tasks of SemEval2018 task9-Hypernym Discovery : 1A(general), 2A(Medical), 2B(Music)
- We employed 3 metrics for evaluation : MRR, MAP, P@k



Experimental Results

Dataset	Metric	Proposed	$CRIM_{r1}$	RMM
1A General	MRR			0.24 ± 0.18
	MAP			0.24 ± 0.10
	P@1			0.27 ± 0.17
	P@3			0.19 ± 0.10
	P@5			0.21 ± 0.10
	P@15			0.30 ± 0.11
2A Medical	MRR			0.40 ± 0.41
	MAP			0.17 ± 0.20
	P@1			0.16 ± 0.20
	P@3			0.19 ± 0.24
	P@5			0.17 ± 0.20
	P@15			0.18 ± 0.20
2B Music	MRR		22.60 ± 10.47	3.88 ± 1.46
	MAP		17.68 ± 6.96	2.90 ± 1.06
	P@1		15.64 ± 8.88	2.05 ± 1.38
	P@3		16.60 ± 7.45	2.50 ± 0.92
	P@5		17.18 ± 7.06	2.73 ± 0.91
	P@15		19.33 ± 6.69	3.41 ± 1.19

To Do List

- ✓ Experiments – repeat 30 times
  - ➢ 1A, 2A, 2B for CRIM and Proposed model
  - ➢ it will take some time
- ✓ Delete the attention mechanism from the proposed method.
  - ➢ Move to appendix
- ✓ Explain what vectors are good for hypernym discovery
  - ➢ Difference between contexture embedding and word2vec embedding

Thank you

Restoration of Homoglyph Attack on Spam Using BERT Masked Language Model

Hanyong Lee

glhy0718@gmail.com

5th Oct, 2022

Contents

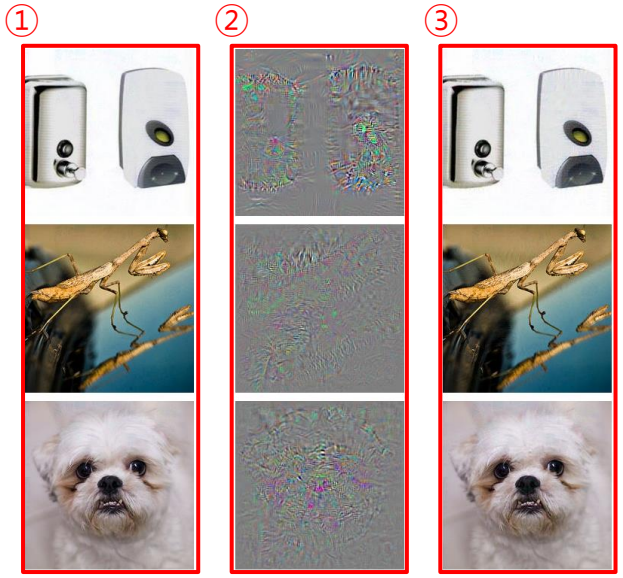
To-Do List from Previous Seminar

- ✓ Character-Level BERT → proposed method로 변경
- ✓ 12 page : text를 아래로 옮기기, MLM을 사용한 내용으로 바꾸기
- ✓ 13 page 까지 proposed method 으로 변경
- ✓ experimental setting 더 자세히 작성
  - ➢ Database 및 실험 설명
- ✓ analysis 넣기 - 왜 기존 알고리즘 (BERT)이 성능이 더 낮은지
- ✓ 통계테스트 넣기
  - ➢ 프리드만 테스트, 본페로니-둔 테스트

Introduction

NLP Attack

Adversarial examples in **Computer Vision**



① correctly predicted samples
② image difference between ① and ③
③ images predicted as “ostrich”

Adversarial examples in **Natural Language Processing**

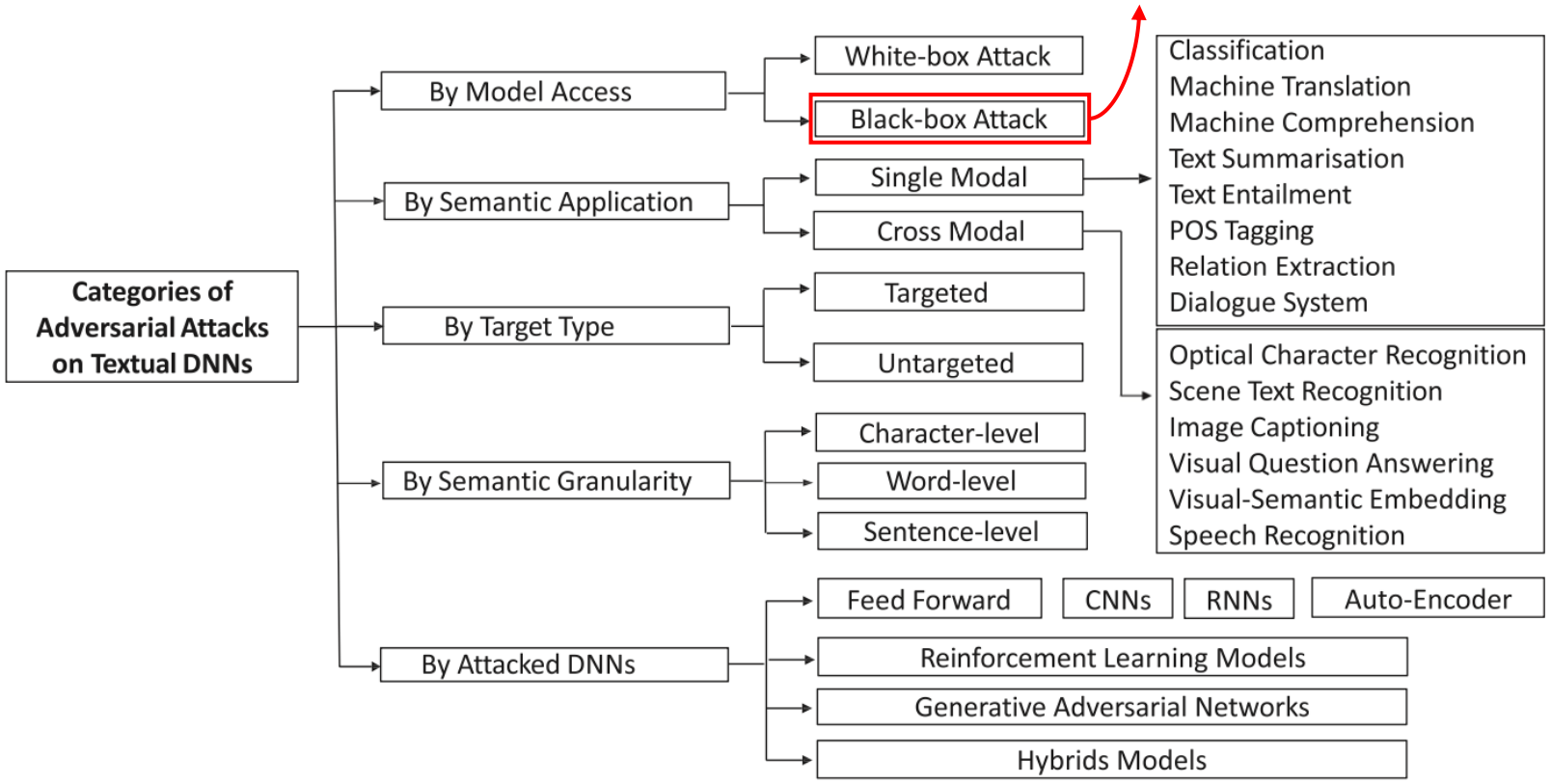
?

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Introduction

NLP Attack

Adversarial examples in **Natural Language Processing**

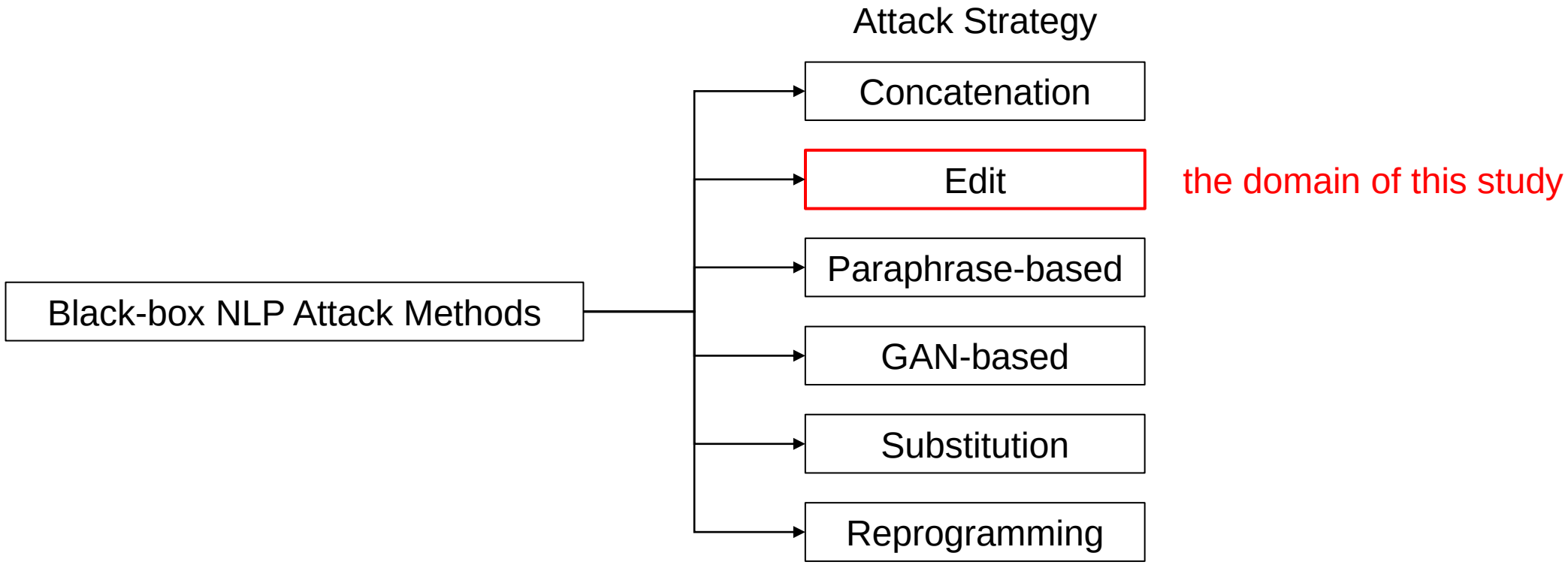


Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. ACM Transactions on Intelligent Systems and Technology (TIST), 11(3), 1-41.

Introduction

NLP Attack

Black-box NLP Attack Methods

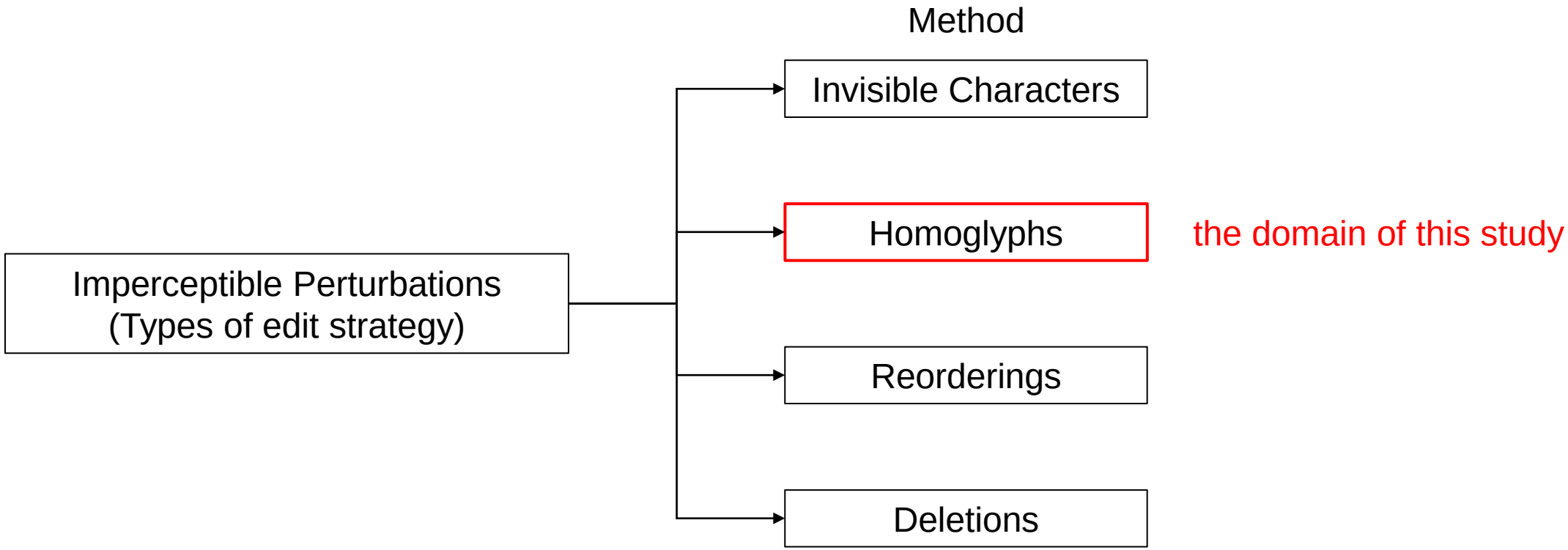


Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. ACM Transactions on Intelligent Systems and Technology (TIST), 11(3), 1-41.

Introduction

Homoglyph Attack

Imperceptible Perturbations

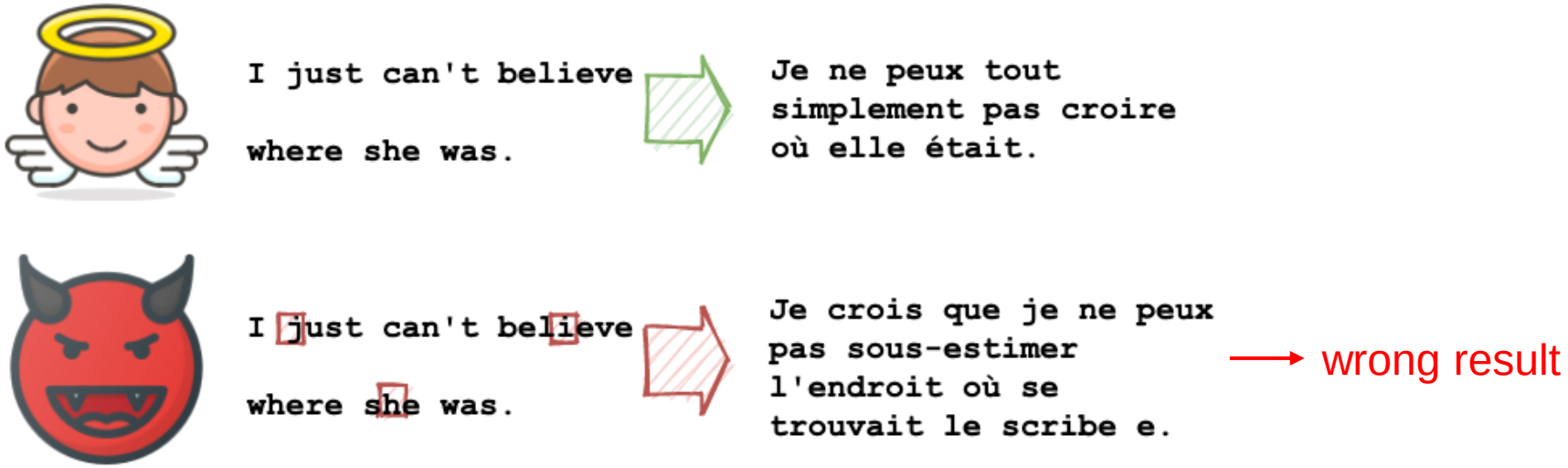


Boucher, N., Shumailov, I., Anderson, R., & Papernot, N. (2022, May). Bad characters: Imperceptible nlp attacks. In 2022 *IEEE Symposium on Security and Privacy (SP)* (pp. 1987-2004). IEEE.

Related Work

Homoglyph Attack

Example of an attack using homoglyph to confuse English-Franch translation task.



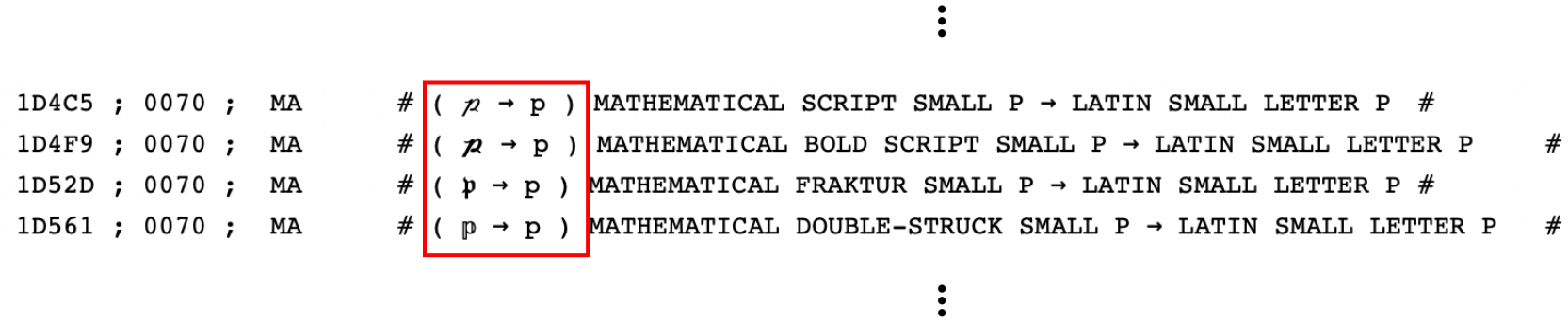
“j” is replaced with U+3F3(“j”), “i” with U+456(“i”), and “h” with U+4BB(“h”).

Boucher, N., Shumailov, I., Anderson, R., & Papernot, N. (2022, May). Bad characters: Imperceptible nlp attacks. In 2022 *IEEE Symposium on Security and Privacy (SP)* (pp. 1987-2004). IEEE.

Related Work

Defense Methods of Homoglyph Attack

Translate table for confusable characters



```
1D4C5 ; 0070 ;  MA      # ( 𝓅 → p ) MATHEMATICAL SCRIPT SMALL P → LATIN SMALL LETTER P  #
1D4F9 ; 0070 ;  MA      # ( 𝓹 → p ) MATHEMATICAL BOLD SCRIPT SMALL P → LATIN SMALL LETTER P    #
1D52D ; 0070 ;  MA      # ( 𝔭 → p ) MATHEMATICAL FRAKTUR SMALL P → LATIN SMALL LETTER P #
1D561 ; 0070 ;  MA      # ( 𝕡 → p ) MATHEMATICAL DOUBLE-STRUCK SMALL P → LATIN SMALL LETTER P   #
```

It doesn’t contain some pairs.

- μ → u
- $ → S
- 5 → S
- …

Unicode Consortium 2022, *Unicode Utilities: Confusables.* accessed 10 September 2022, <https://www.unicode.org/Public/security/15.0.0/confusables.txt>.

Related Work

Defense Methods of Homoglyph Attack

SimChar database

H. Suzuki et al. used the following formula to calculate the similarity of two bitmap images of characters.

$$\Delta = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} |I_1(i,j) - I_2(i,j)|$$

pixel of image I_1, I_2 at coordinate (i, j)

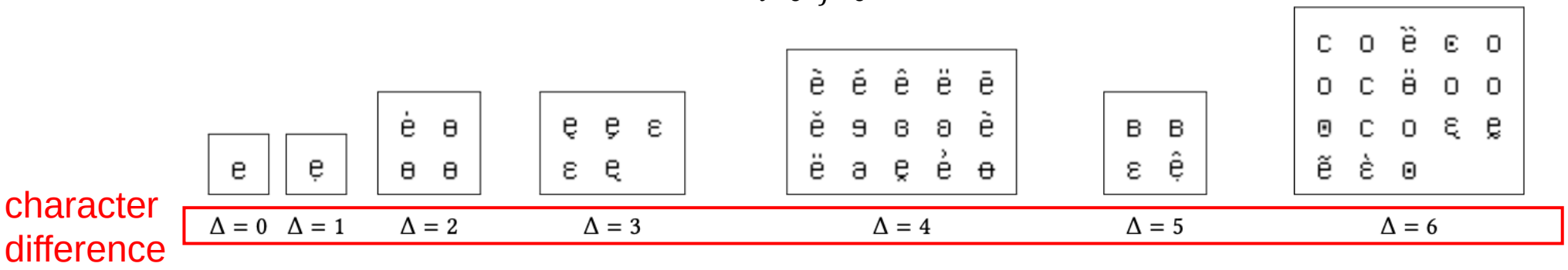


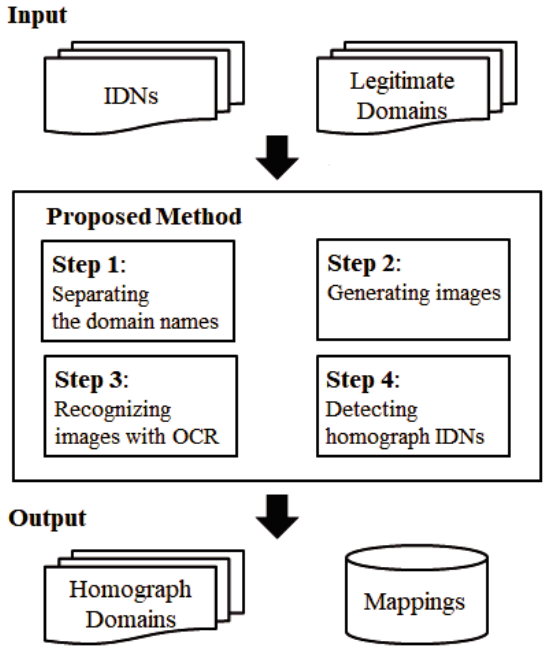
Figure 6: Basic Latin lowercase letter 'e' and characters under different values of the threshold Δ. In this work, characters with Δ ≤ 4 are detected as homoglyphs.

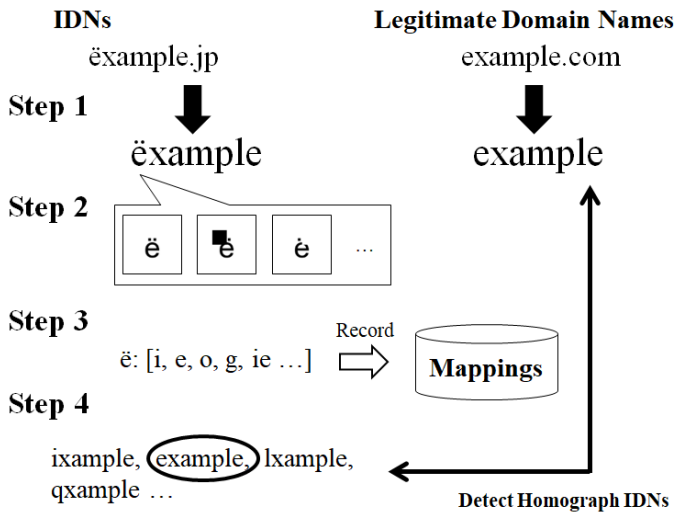
Suzuki, H., Chiba, D., Yoneya, Y., Mori, T., & Goto, S. (2019, October). ShamFinder: An automated framework for detecting IDN homographs. *In Proceedings of the Internet Measurement Conference* (pp. 449-462).

Defense Methods of Homoglyph Attack

OCR-based method

Sawabe et al. used the OCR-based Method to detect Homoglyph Attack.





Sawabe, Y., Chiba, D., Akiyama, M., & Goto, S. (2019). Detection method of homograph internationalized domain names with OCR. *Journal of Information Processing*, *27*, 536-544.

Defense Methods of Homoglyph Attack

Spellchecker-based method

Imam, N. H et al. used a spellchecker-based method to restore the homoglyph attacks.

Table 2
Some of the adversarial text examples used in experiments.

Original	Flipping	Swapping	Deletion
Busy	Bu$y	Bsuy	Bsy
Caller	C@1ler	Claler	Cller
Reply	Rep1y	Rpely	Rply
winner	w!nner	wniner	wnner

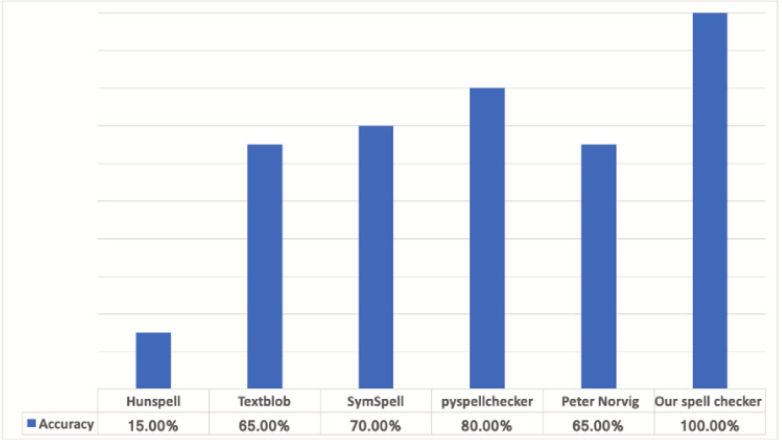
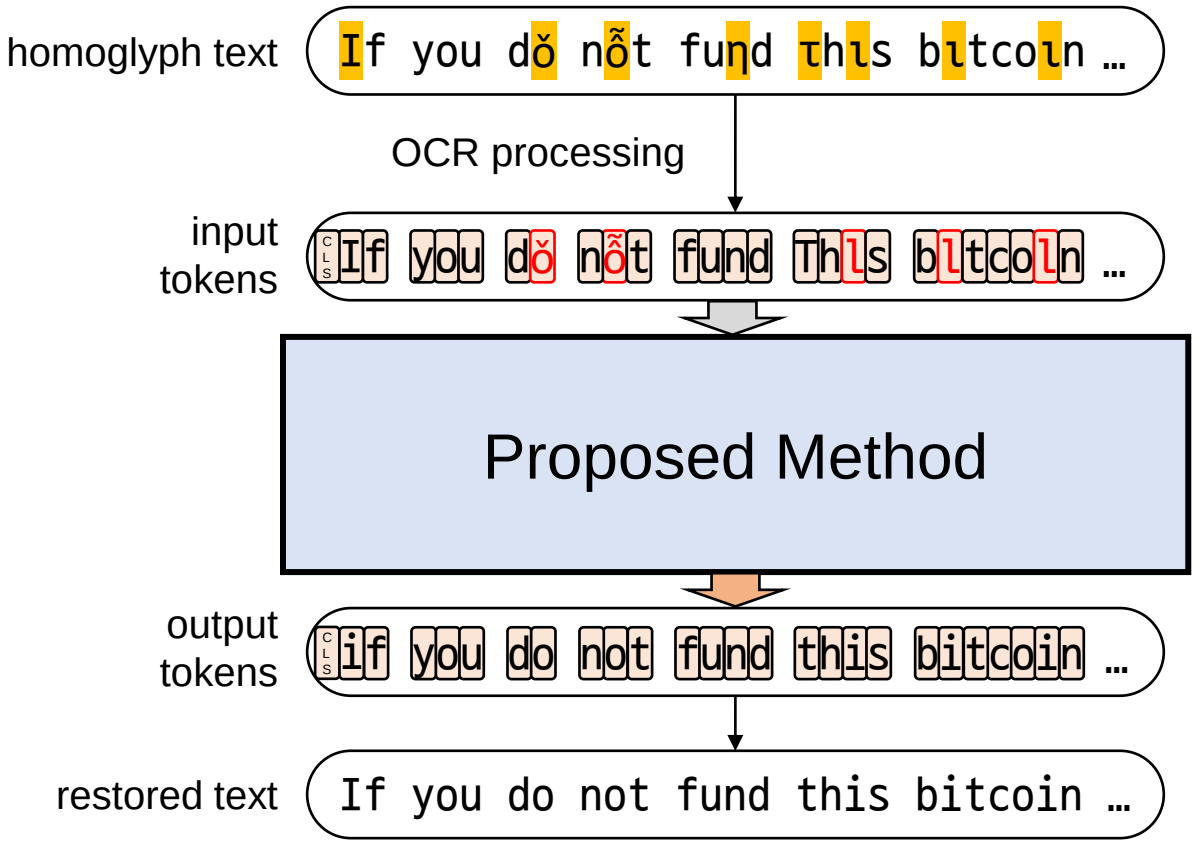


Fig. 9. Flipping.

Imam, N. H., Vassilakis, V. G., & Kolovos, D. (2022). OCR post-correction for detecting adversarial text images. *Journal of Information Security and Applications*, *66*, 103170.

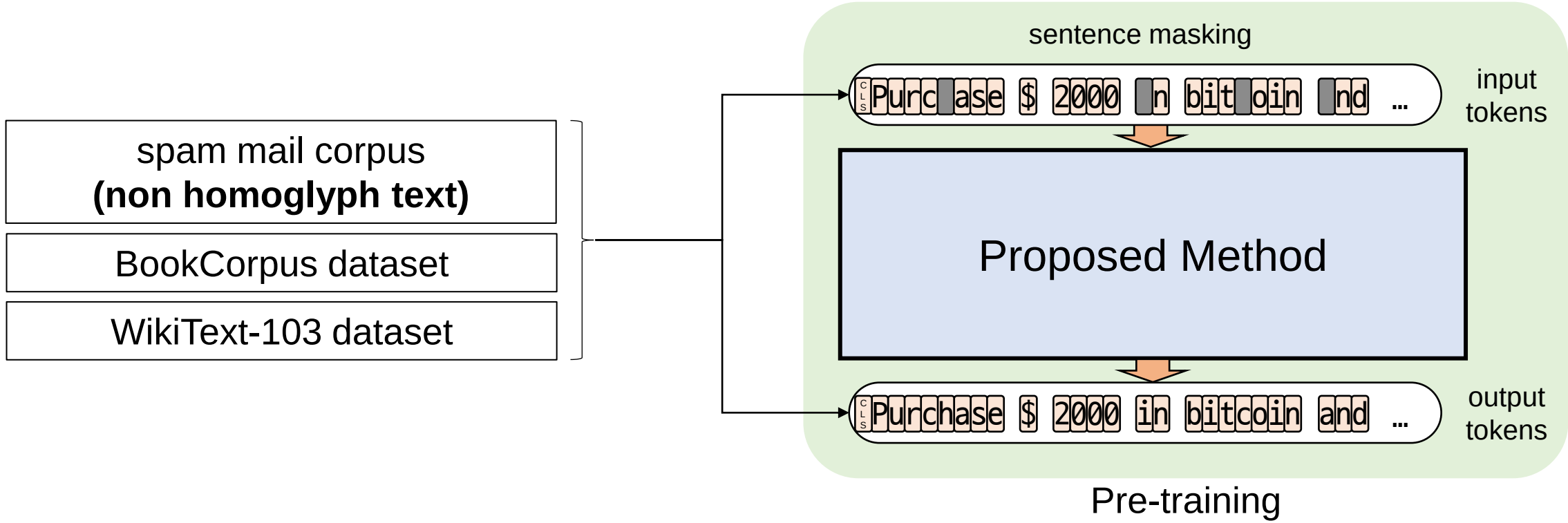
Proposed Method

Restoring Process Overview



Proposed Method

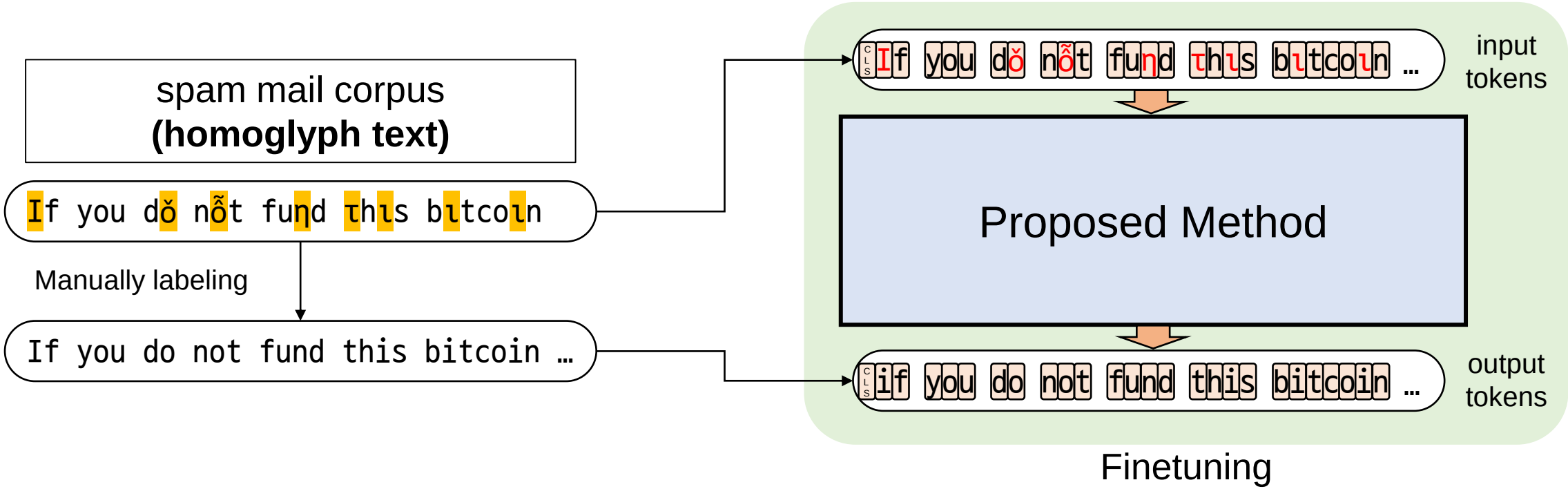
Pre-training



We pretrained the proposed model as a masked language model (MLM).

Proposed Method

Finetuning



Evaluated with a blind test (Train/Test set split ratio = 8:2 / Iteration = 50 times)

Experimental Settings

Dataset

	# of Examples		# of Tokens	
Dataset	Train Set	Validation Set	Train Set	Validation Set
WikiText-103	1,165,029	2,461	285,324,545	612,967
BookCorpus (about 2% of the total dataset)	1,480,085	3,126	79,346,399	143,939
bitcoinabuse.com dataset (non-homoglyph)	299,239	633	23,152,956	54,680
bitcoinabuse.com dataset (homoglyph)	21,342	5,336	random split	random split

Experimental Settings

Model Parameters

Parameter	Value
Input Max Sequence	512
Vocab Size	881
# of Attention Heads	12
# of Hidden Layers	12
# of Model Parameters	86,720,369

Experimental Settings

Pre-training Setting

Parameter	Value
# of Epochs	25
Train / Validation split	refer the slide 14.
Used Datasets	• WikiText-103 • BookCorpus (about 2% of the total dataset) • bitcoinabuse.com dataset (non-homoglyph sentences)
Metric	Accuracy

Experimental Settings

Finetuning Setting

Parameter	Value
# of Epochs	5
Train / Validation split	8 : 2
# of Test Iteration	50
Used Datasets	• bitcoinabuse.com dataset (homoglyph sentences)
Metric	Accuracy

Experimental Results

Finetuned Model Evaluation - Accuracy

Algorithm	Word Accuracy
Proposed	**0.9959 ± 0.0008**
BERT	0.6713 ± 0.0029 ▼
OCR	0.6518 ± 0.0028 ▼
SimChar DB	0.5512 ± 0.0046 ▼
Spell Checker	0.9087 ± 0.0011 ▼

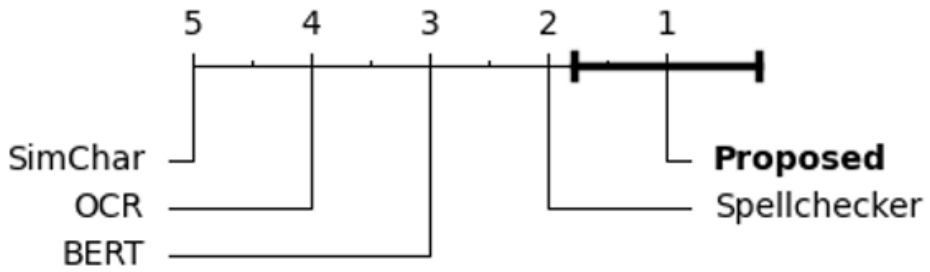
Word level accuracy and T-test with the accuracy of the proposed method
▼ means that the algorithm has a significantly lower performance than the proposed algorithm based on the t-test.
▼: $p < 0.01$

Experimental Results

Finetuned Model Evaluation – Friedman test

Statistic	p value
200.0	3.7572e-42

Finetuned Model Evaluation – Bonferroni–Dunn test (α = 0.05, CD = 0.7899)



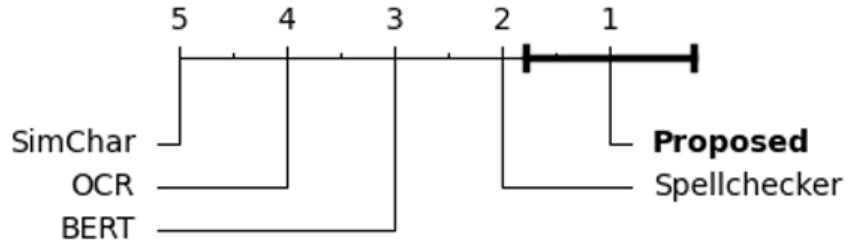
Experimental Results

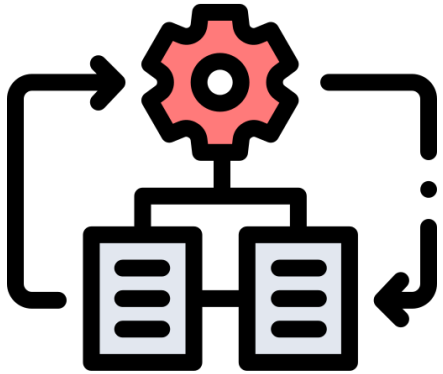
Analysis

Algorithm	PROS	CONS
Proposed	• Robust on new homoglyphs • Automated workflow	• Pre-raining the model takes a long time. • Inability when the input & output lengths are different
BERT	• Slightly robust on new homoglyphs	• Pre-raining the model takes a long time. • Word-based tokenization is not helpful for the task.
OCR	• Robust on new homoglyphs • Automated workflow	• OCR is not performing well on homoglyphs • Inability when the input & output lengths are different
SimChar DB	• Automated workflow on finding homoglyph pairs	• Not robust on new homoglyphs • Must manually add homoglyph pairs to the dictionary • Inability when the input & output lengths are different
Spell Checker	• Robust on new homoglyphs	• Must manually add words to the dictionary

Conclusion

Contribution of the Study





1. Achieve higher performance than traditional methods such as OCR, SimChar DataBase, and spell checker.

2. Propose an automated homoglyph restoration framework by using deep learning.

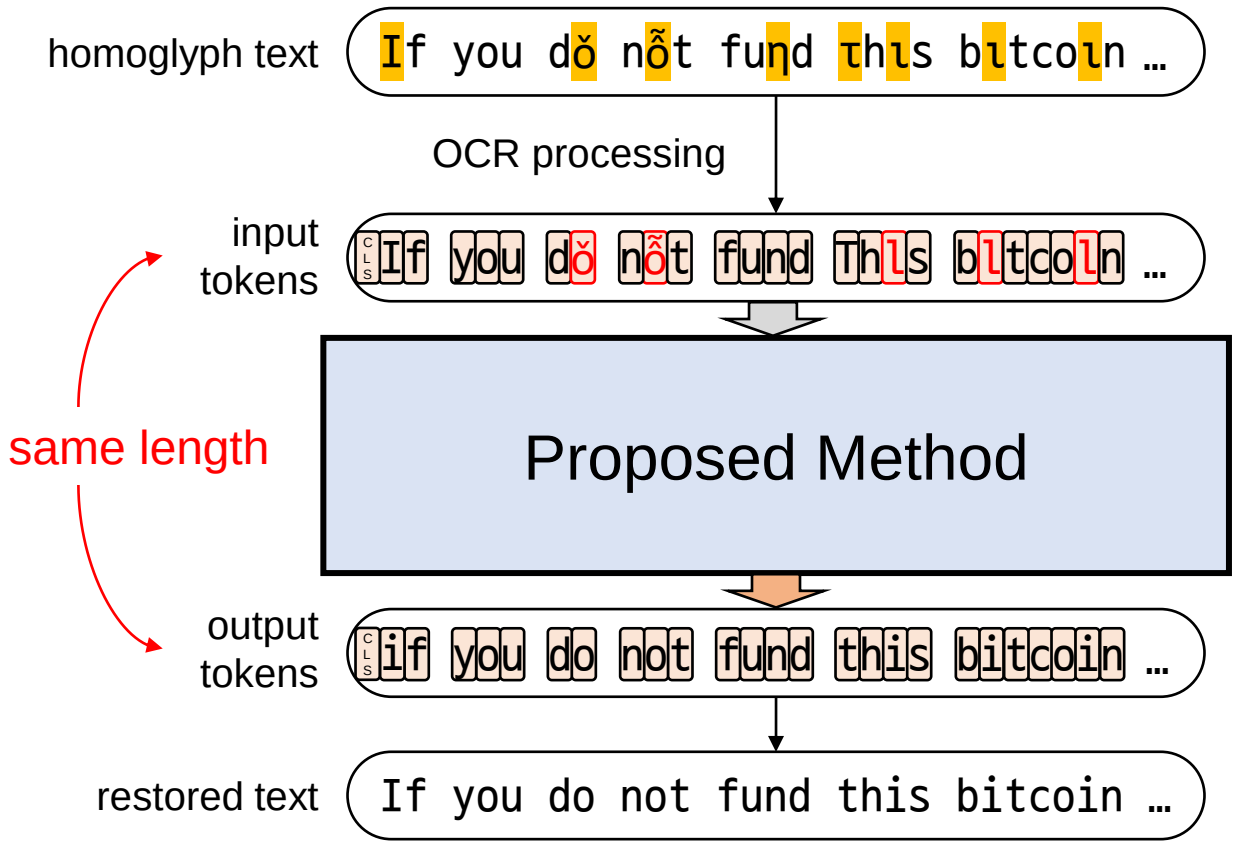
Conclusion

Limitation of Proposed Model

The proposed model is a masked language model, limited in that the input/output sentences should be the **same sequence length**.

- It can be used as a homoglyph of the alpha bet ‘m’ by concatenating the letters ‘r’ and ‘ n’. But the presented model does not solve this case.

The proposed model can be applied as a **sequence-to-sequence** model in future studie s.



References

1. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). "Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.
2. Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. ACM Transactions on Intelligent Systems and Technology (TIST), 11(3), 1-41.
3. Boucher, N., Shumailov, I., Anderson, R., & Papernot, N. (2022, May). Bad characters: Imperceptible nlp attacks. In 2022 IEEE Symposium on Security and Privacy (SP) (pp. 198 7-2004). IEEE.
4. Unicode Consortium, "Unicode Utilities: Confusables." unicode.org. https://www.unicode.org/Public/security/15.0.0/confusables.txt (accessed 10 September 2022).
5. Suzuki, H., Chiba, D., Yoneya, Y., Mori, T., & Goto, S. (2019, October). ShamFinder: An automated framework for detecting IDN homographs. In Proceedings of the Internet Mea surement Conference (pp. 449-462).
6. Sawabe, Y., Chiba, D., Akiyama, M., & Goto, S. (2019). Detection method of homograph internationalized domain names with OCR. Journal of Information Processing, 27, 536-5 44.
7. Imam, N. H., Vassilakis, V. G., & Kolovos, D. (2022). OCR post-correction for detecting adversarial text images. Journal of Information Security and Applications, 66, 103170.
8. Norvig, P., "How to write a spelling corrector." norvig.com https://norvig.com/spell-correct.html (accessed 10 September 2022).
9. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
10. Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843.
11. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching mov ies and reading books. In Proceedings of the IEEE international conference on computer vision (pp. 19-27).

To-Do List

1. 논문작성 시작 (Elsevier 양식 추천)
2. PPT 수정
    1. Introduction
        - Adversarial example on CV 제거
    2. Experimental Settings
        - Dataset 표에 Pre-train/Finetuning dataset 표시
    3. Experimental Results
        - Freidman / Dunn test 코드 재검토
        - Slide 21 표에서 proposed row 제거하고 Related Works Section으로 옮기기
    4. Analysis
        - Proposed와 BERT 비교하는 내용 추가
            - output 차이, Proposed의 성능이 좋은 이유, BERT의 한계점 등의 관점에서
    5. Conclusion
        - Contribution
            - Automated 관련 내용 제거
            - Char-BERT 제안에 대한 내용 추가

Thank you

Appendix

A. Proposed Model Information

Character-level BERT

- Number of parameters: 86,720,369
- Max length of input tokens: 512
- Vocab size: 881

Appendix

B. Pre-trained Model Evaluation

Evaluated with the **whole test sentences**

Character level BERT is only pretrained with the **WikiText-103, BookCorpus & spam mail datasets**.

Algorithm	Word Accuracy
Proposed (not finetuned)	**0.9516 ± 0.1289**
Normal BERT	0.6713 ± 0.2679 ▼
OCR	0.6519 ± 0.2976 ▼
SimChar DB	0.5506 ± 0.3627 ▼
Spell Checker	0.9090 ± 0.1620 ▼

Word level accuracy and T-test with the accuracy of the proposed method

▼ means that the algorithm has a significantly lower performance than the proposed algorithm based on the t-test.

▼: $p < 0.01$

Appendix

C. Test Output Examples of Pretrained Model

Example 1

- input sentence L Need ɗ y0UR TOTɑı ɑTTeNTıOɴ fOR The c0MıNg TweNTY-fOUR hRS,
- output sentence : i need your total attention for the coming twenty - four hrs,

Example 2

- input sentence : ɑftêr thät, i hȧvë ṡtȧrtèd trȧcкiṅġ ẏȯur ṅtérпët ȧctívítìêṡ.
- output sentence : after that, i have started tracking yo ur internet activities.

Appendix

C. Test Output Examples of Pretrained Model

Example 3

- input sentence : do ṅȯt trẏ tȯ coṅtáct the ṗolḭćê àɴd ȯtɳêr sëcurity servḭces.
- output sentence : don't try to contact the police and other security services.

Example 4

- input sentence : yøuř åccøuňt wås åttåckěd.
- output sentence : yogr account was attacked.

Data Imputation for missing value

Hyejin Baek

bhj4208@gmail.com

5th Oct, 2022

Contents

1. 논문
2. 실험

논문

A novel to optimize multiple imputation algorithm for missing data using evolution methods (Biomedical Signal Processing and Control, 2022)

- Genetic algorithm + multiple imputation method 제시
- HCC dataset 사용
- Experiment result >

Table 2
Sample of Comparison between our optimization method results and (NEq. and KNN results).

y *val*	$\hat{y}$ *NEq.*	$\hat{y}$ *KNN*	$\hat{y}$ *after* 1st *Optimization*	$\hat{y}$ *after* 2nd *Optimization*
0.9	1.272	1.040	0.767	0.866
0.9	1.470	16.100	0.852	0.921
5.8	6.946	0.880	6.178	6.113
7.1	7.525	3.160	6.806	6.984
2.4	2.760	8.640	2.265	2.460
1.8	2.695	0.840	2.222	2.209
20	20.636	0.920	19.775	20.103
1	1.773	0.660	0.953	1.012
0.8	1.110	1.040	0.584	0.714
0.9	1.453	1.940	0.755	0.887
0.7	0.923	3.800	0.906	0.894
0.9	1.366	9.760	0.684	0.834
1	1.374	1.700	1.172	1.083
0.7	1.394	1.700	1.065	1.038
8.9	9.294	1.800	8.683	8.905
1.3	1.103	3.860	1.164	1.203
0.6	1.044	9.020	0.834	0.769
1.5	1.930	0.740	1.917	1.774
11.1	13.194	1.040	10.651	10.933
0.8	1.484	1.920	1.026	1.008
0.8	0.314	1.460	0.615	0.621
3.3	2.758	6.480	3.447	3.407

- Conclusion

→ 논문의 저자가 제시한 method로 새로운 데이터셋을 이용하여 다양한 결과 도출 필요

실험

Zero/Mean imputation 으로 XGBoost 진행

Dataset : UCI(Breast-cancer) dataset

1. Zero imputation

	0%	20%	40%	60%	80%
Accuracy	0.95	0.9714	0.9786	0.9857	0.9929
Precision	0.9268	1.0000	0.9524	0.9286	0.8333
Recall	0.9048	0.8750	0.9091	0.9286	1.0000
F1	0.91	0.9333	0.9302	0.9286	0.9091
ROC_AUC	0.9883	0.9954	0.9969	0.9972	1.0000

2. Mean imputation

	20%	40%	60%	80%
Accuracy	0.9714	0.9714	0.9857	0.9929
Precision	1.0000	0.9091	0.9286	0.8333
Recall	0.8750	0.9091	0.9286	1.0000
F1	0.9333	0.9091	0.9286	0.9091
ROC_AUC	0.9954	0.9592	0.9510	1.0000

진행 예정

- Tabular data/Missing value/Data Imputation 관련된 추가적인 논문 연구 진행 예정
- 실험
    1. 성능 지표 결과 추가 분석
    2. Imputation 기법 추가하여 XGBoost, TabNet 테스트 진행 예정

TO DO LIST

- Dataset 20-25개 이상 사용하여 실험 진행
- TabNet, NLP(unknown token) 관련 search
- KNN imputation method 이용하여 실험 진행

Thank you

Gastrointestinal track Polyp Segmentation

Goeun Kim

goeun678@gmail.com

5th Oct, 2022

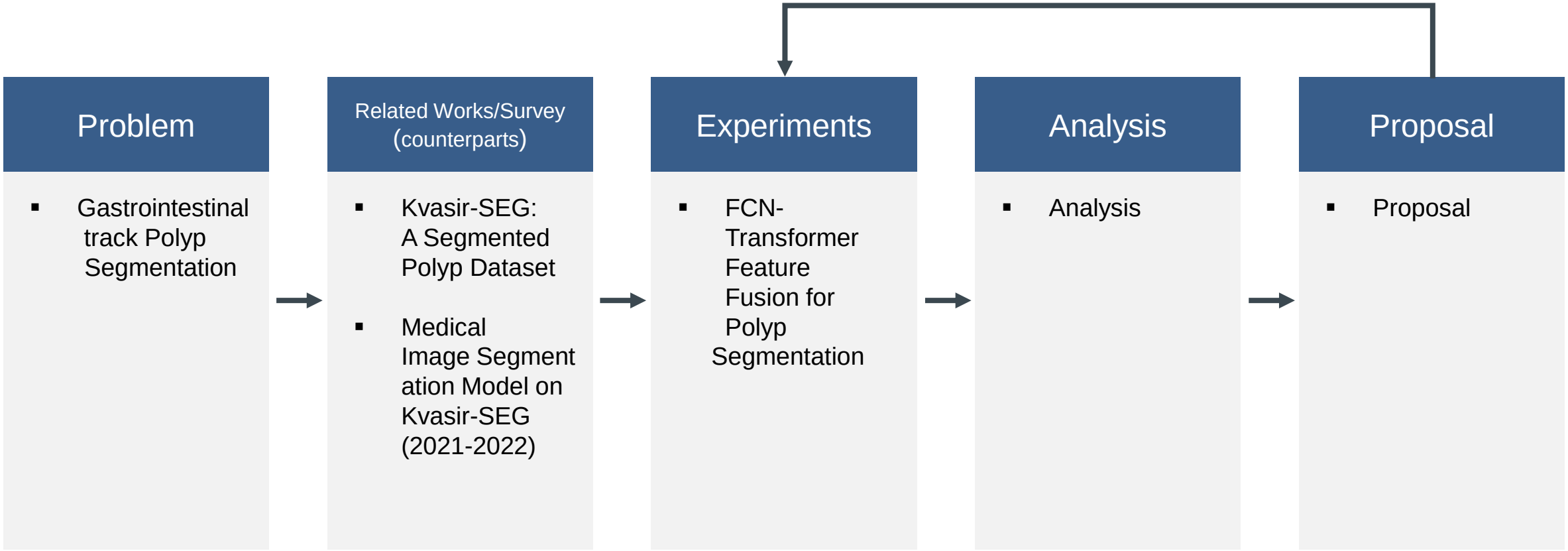
Contents

To Do List from Previous Seminar

- ✓ To do list 1
  - ➢ GI tract polyp segmentation Public dataset 조사

- ✓ To do list 2
  - ➢ GI tract polyp segmentation DL model 조사 및 code 분석

Overall Progress



Introduction

Gastrointestinal track Polyp Segmentation (GI tract Polyp Segmentation)

GI track Polyp : 위창자관(위장관) 용종

➢ **딥러닝의 필요성**

기존 검사법	**Deep Learning**			**효과**		
Colonoscopy > 대장암 위험	Segmentation model	정확도	& 속도	-> polyp 검출율	-> 검출 간과율	-

➢ **데이터셋**

Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, ETIS-LARIBPOLYPDB, Synapse multi-organ CT, MoNuSeg, 2018 Data Science Bowl, Automatic Cardiac Diagnosis Challenge(ACDC), GlaS, Medical Segmentation Decathlon

Related Work

Kvasir-SEG: A Segmented Polyp Dataset

Kvasir SEG

Segmented Polyp Dataset for Computer Aided Gastrointestinal Disease Detection.

- 위장관 용종 이미지와 해당 segmentation mask의 개방형 액세스 데이터 세트
- Contribution of the dataset

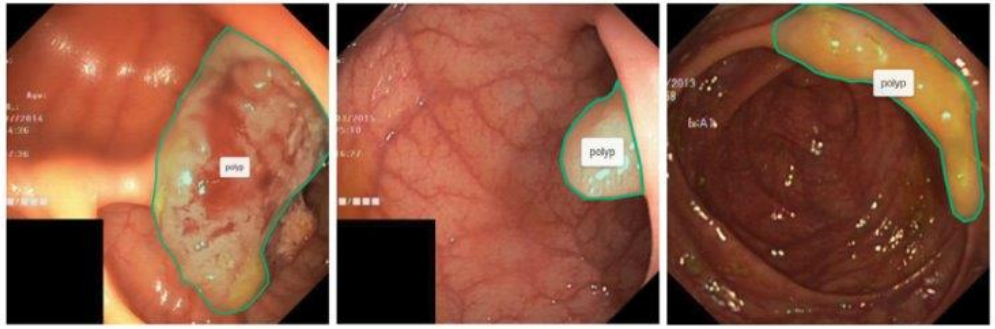


Fig. 1: Example frames from the Kvasir dataset where we additionally have marked the polyp tissue with green outlines.

Proposed Method

FCN-Transformer Feature Fusion for Polyp Segmentation

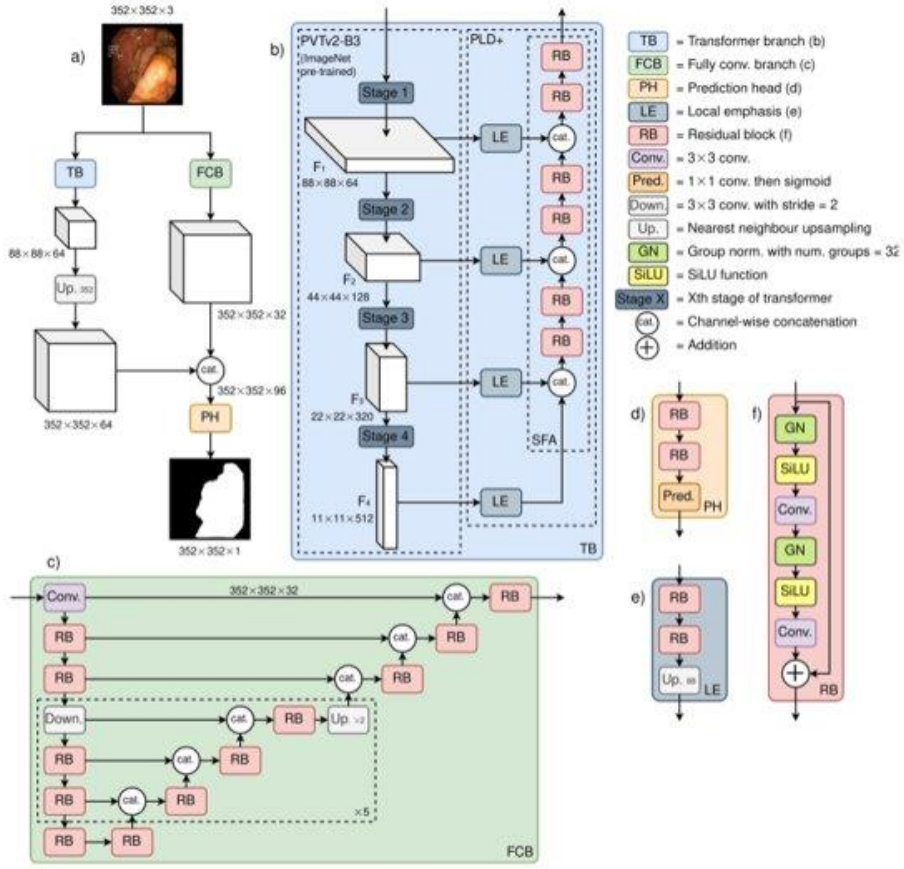


Fig. 1. The architectures of a) FCBFormer, b) the transformer branch (TB), c) the fully convolutional branch (FCB), d) the prediction head (PH), e) the improved local emphasis (LE) module, f) the residual block (RB).

To Do List

- ✓ To do list 1
  - ➢ Kvasir-SEG 데이터 셋으로 FCN-Transformer Feature Fusion for Polyp Segmentation 모델 학습 및 논문 리뷰

- ✓ To do list 2
  - ➢ CVC-ClinicDB, CVC-ColonDB, ETIS-LARIBPOLYPDB 등의 위장계 polyp 검출용 데이터를 사용하여 FCN model학습-> 결과 비교

Thank you

Appendix