NEURAL INFORMATION PROCESSING SYSTEMS

UniT: A Unified Look at Certified Robust Training against Text Adversarial Perturbation

Muchao Ye[1] Ziyi Yin[1] Tianrong Zhang[1] Tianyu Du[2]
Jinghui Chen[1] Ting Wang[3] Fenglong Ma[1]
[1]The Pennsylvania State University, [2]Zhejiang University,
[3]Stony Brook University

PennState

ZHEJIANG UNIVERSITY 1897

Stony Brook University

Motivation

Certified Robustness for Text Classification

- Given a text data pair (X, y). $X = [w_1, \ldots, w_n]$. Suppose that f can make the correct prediction, i.e., $\arg\max_{y_i \in Y} f_{yi}(X) = y$. In the context of certified robustness, we are interested in getting a certified prediction result such that $\arg\max_{y_i \in Y} f_{yi}(X') = y$ holds for any allowed perturbed sample X' of X
- Perturbed sample X' is obtained by replacing each word w_i by its synonym (Suppose w_i has m_i synonyms)

Motivation

- Recent years have seen an urge for robust natural language processing models that can provide certified robust predictions
- The key to producing certified predictions is **certified robust training**, which introduces perturbation during training to ask the model to adapt to it
- We observe there is a structural gap in current certified robust training

Motivation

- Structural gap in current certified robust training

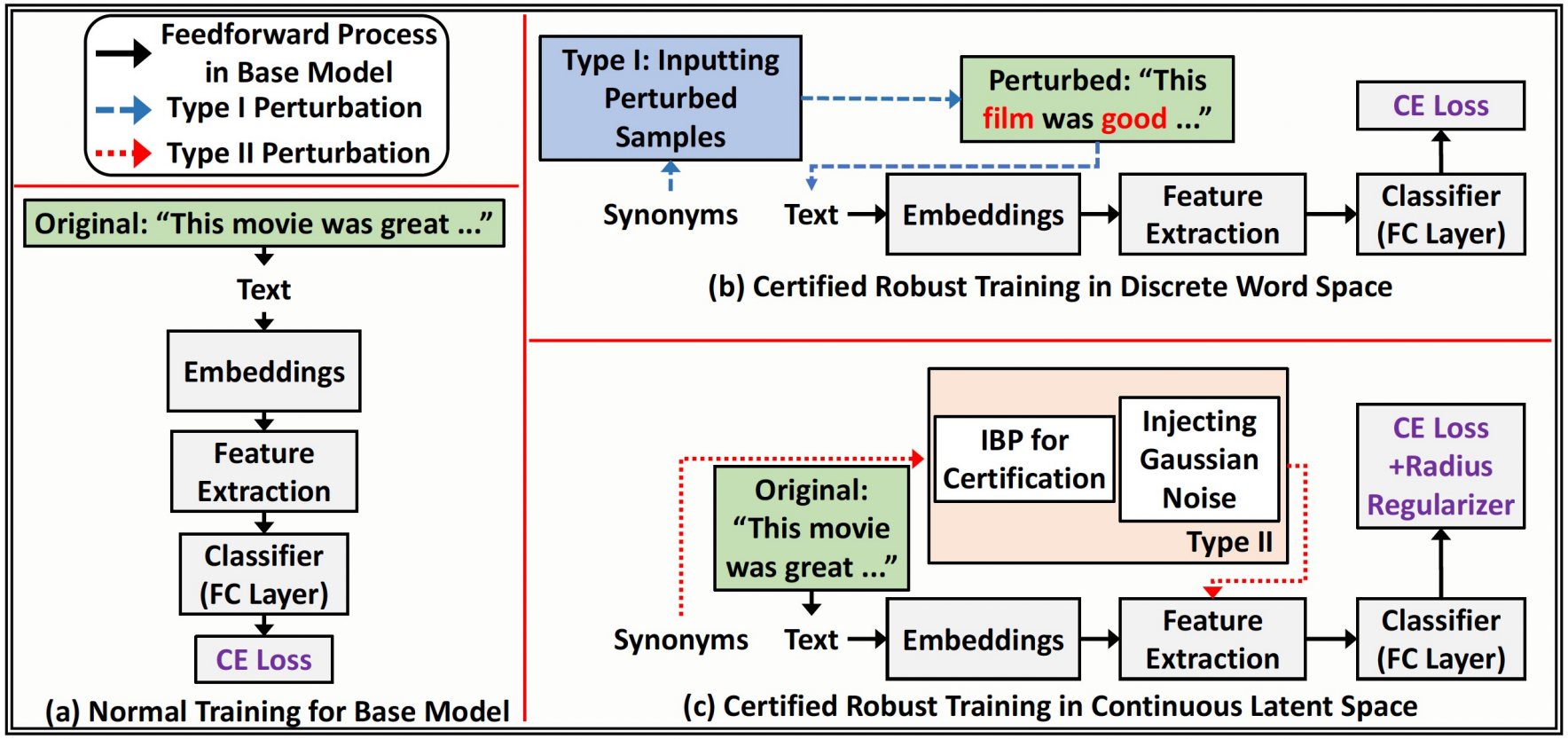

Figure 1: Given the (a) base model, (b) Type I frameworks construct the smoothed model in the discrete word space while (c) Type II frameworks construct it in the latent space with IBP. There is a need for unifying these training frameworks and improving the robustness of the base model.

Motivation

- Type II frameworks need to include an extra IBP module compared to Type I frameworks, which affects the certification because of the loose bound problem of IBP

- Research question 1: how to build a unified framework for these two types of pipelines to provide stronger certified robustness?

Motivation

- The use of cross-entropy loss lacks fine-grained robustness regularization for individual modules and consideration of the final certification target

- Research Question 2: how to design robustness regularization terms for individual modules to further improve the base model robustness

UniT

- We design a unified certified robust training framework named UniT by utilizing the embedding space as the intermediate

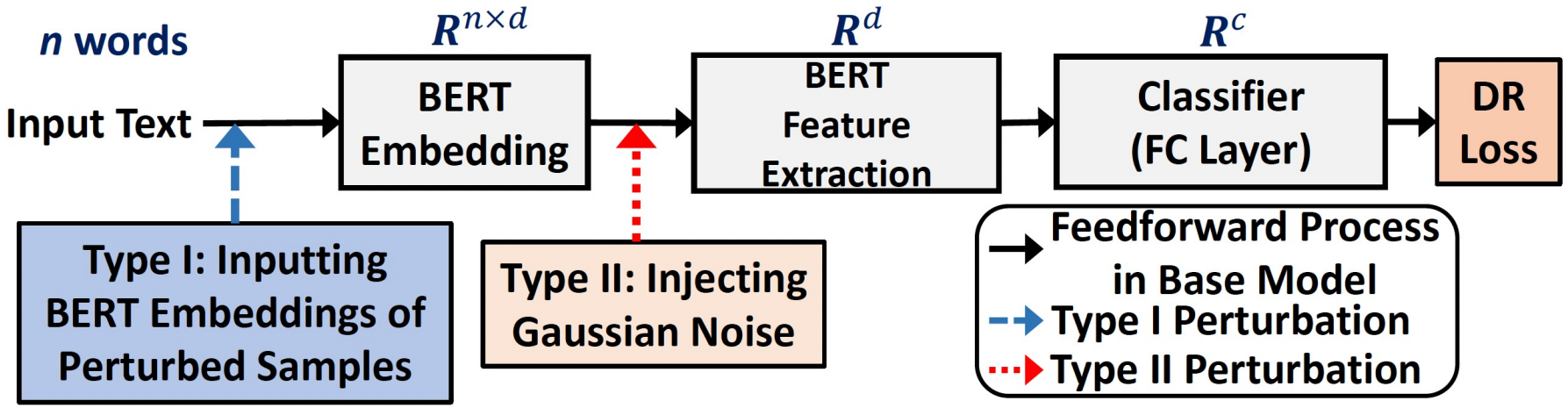

Figure 2: Given the base model, UniT unifies two frameworks by working in the embedding space. Type I training replaces original embeddings with embeddings of perturbed samples. Type II training adds Gaussian noise to original word embeddings.

UniT

Certification

- For the Type I scenario, we can obtain the certified robustness guarantee from Proposition 1 of SAFER by constructing the smoothed model based on synonym substitutions
- For the Type II scenario, we propose a new theorem (See Theorem 1 in our paper) for obtaining certified results directly in the embedding space

Mao Ye, Chengyue Gong, and Qiang Liu. Safer: A structure-free approach for certified robustness to adversarial word substitutions. In ACL 2020.

UniT

Training Loss

- We propose a decoupled regularization (DR) learning paradigm that directly conducts modular regularization to aid the CE loss

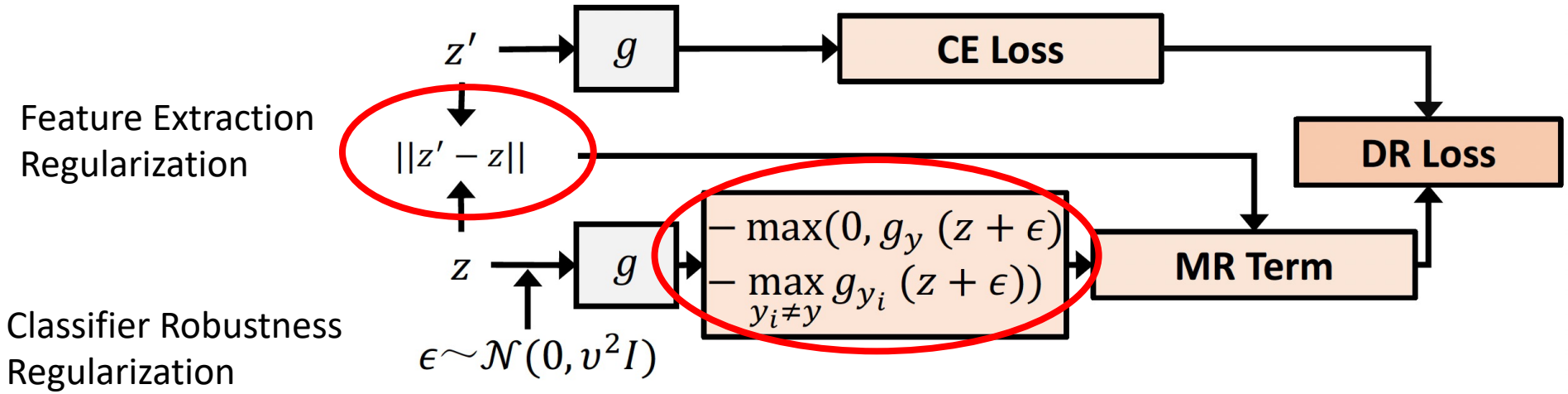

Figure 3: DR loss contains a pathway (MR term) for providing modular regularization for CE loss.

Experiment Setup

- Datasets: (1) IMDB, (2) SST2, (3) Yelp, and (4) AG
- Baselines: SAFER (Type I), CISS (Type II), PGD Loss (Adversarial Training), and TRADES Loss (Adversarial Training)
- Metric: Certified Robust Accuracy (CRA)
  - CRA = natural accuracy * certification ratio

Experiments

- Comparison on Certified Robust Accuracy

Base Model	Loss	IMDB	SST2	Yelp	AG
BERT	CE Loss (SAFER)	85.36	91.65	97.19	93.78
	PGD Loss	87.52	90.28	97.86	93.98
	TRADES Loss	86.80	90.44	97.56	93.96
	DR Loss (UniT)	**89.04**	**93.02**	**97.87**	**94.31**

Table 1: Comparison of certified robust accuracy (%) in the Type I scenario.

Method	Loss	Yelp	AG
CISS	CE	88.60	82.47
CISS	DR	89.22	82.93
UniT	DR	**91.24**	**84.32**

Table 2: Comparison of certified robust accuracy (%) in the Type II scenario.

Experiments

- Comparison on Empirical Robust Accuracy

Base Model	Loss	IMDB	SST2	Yelp	AG
BERT	PGD Loss	62.0 (87.1)	87.6 (91.5)	92.4 (96.9)	85.5 (93.7)
	TRADES Loss	57.8 (84.8)	86.3 (91.7)	91.9 (97.0)	84.4 (94.3)
	DR Loss	**72.7** (86.9)	**89.8** (93.3)	**96.9** (98.4)	**87.6** (92.9)

Table 3: Comparison of empirical robust accuracy (%) with adversarial training losses. We also show the corresponding natural accuracy indicated by the parentheses.

- We also verify the design of our unified framework and DR loss through additional analysis experiments. Please refer to our paper for further details

Conclusion

- We propose a unified certified robust training framework that can provide a stronger robustness guarantee
- Under this framework, we introduce the DR loss combining the CE loss with the modular regularization term for different modules specifically to improve the base model robustness

Thank you for listening!