SimSCLSD: A Simple Framework for Supervised Contrastive Learning of Sarcasm Detection

Taeyun Eom
Department of Electrical and Computer Engineering, Seoul Nation University

Abstract

- **Sarcasm detection** is one of the significant challenges in the Natural Language Processing (NLP).
- **Standard baselines** using pre-trained Transformer-based models, typically fine-tuned uses a Cross-Entropy (CE) Loss:
  - may be **suboptimal** for learning discriminative representations, especially for nuanced task with the subtle boundary between classes.
- We propose **SimSCLSD**, a **Sim**ple framework for **S**upervised **C**ontrastive **L**earning of **S**arcasm **D**etection:
  - Stage 1: A SupCon phase to learn a more discriminative embedding space
  - Stage 2: Standard fine-tuning phase
- Our approach creates more separable representations and achieved to superior classification performance on the FigLang 2020 dataset.

Introduction

- Sarcastic utterances are highly **dependent on the surrounding conversational context**.
- We use the FigLang 2020 Sarcasm Detection shared task datasets (Reddit & Twitter), which consists of online discourses specifically designed to benchmark this context-aware capability. [1]
- **Problem**: Standard SOTA Transformers like RoBERTa [2] fine-tuned with CE Loss [3] does not explicitly enforce a **discriminative embedding space**.
  - This can be suboptimal for nuanced tasks where the boundary between sarcastic and non-sarcastic samples is subtle.
- **Our Solution**: We first apply **Supervised Contrastive Learning** (SupCon) before fine-tuning with a standard CE loss, extending the methodology from SimSCL. [4]
  - Representations of same-label examples (e.g., two sarcastic comments) are clustered while representations of different-label examples are repelled apart.
  - Structures the discriminative embedding space.

Methods

Stage 1: Supervised Contrastive Pre-training

- **Goal**: Learn a discriminative representation space.
- **Input**: Concatenated context and response,
  `[CLS] context [SEP] ... [SEP] response [EOS]`
- **Representation**: z_i, mean-pooled over the encoder's final hidden states
- **Loss Function**: for same-label positives $P(i)$ and different-label negatives $A(i)$

$$\mathcal{L}_{\text{SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}$$

Stage 2: Classification Fine-tuning

- **Goal**: Train the final classifier.
- **Representation**: Standard [CLS] token embedding
- **Loss Function**: Standard CE Loss
- **Strategy**: Differential Learning Rates for the adapted encoder and the new classification head

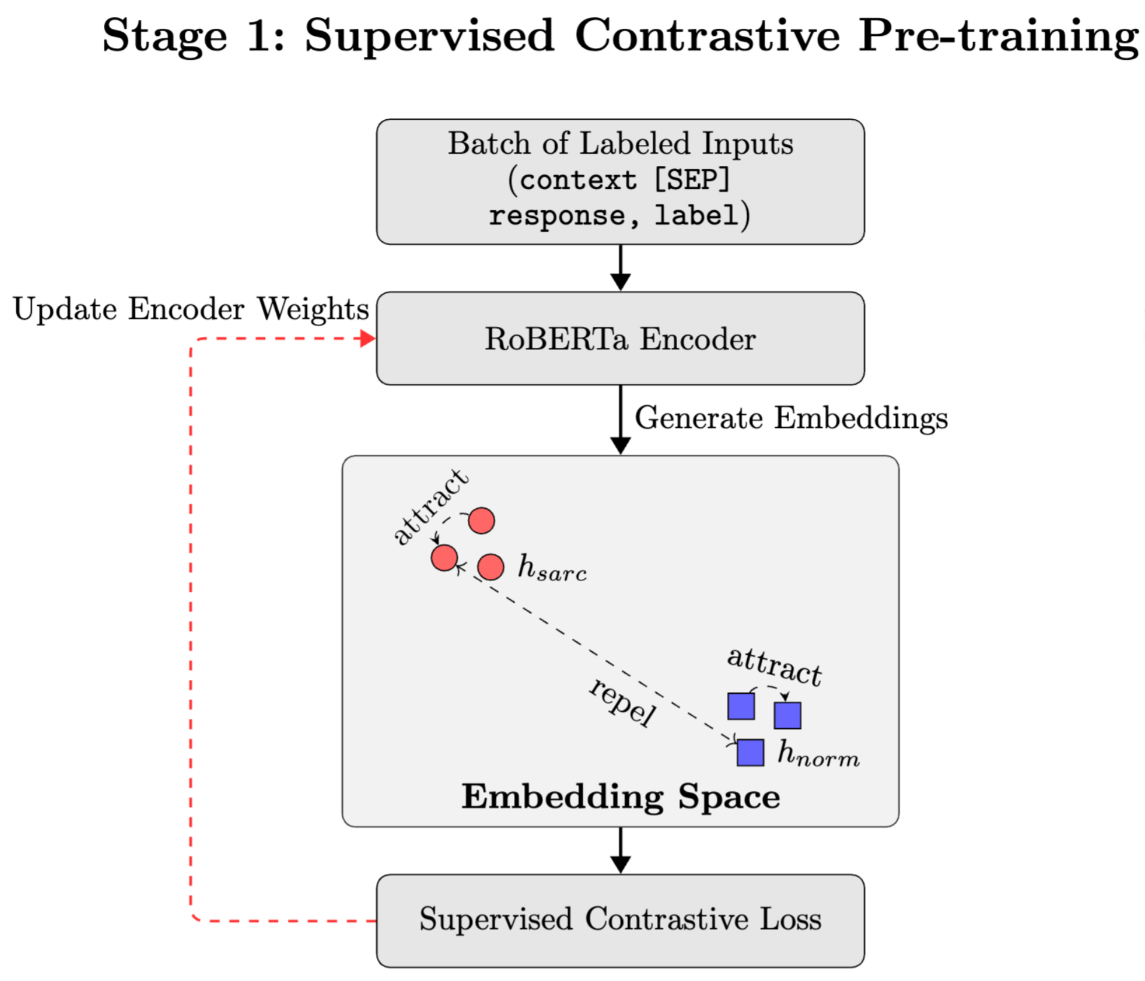
Stage 1: Supervised Contrastive Pre-training
Batch of Labeled Inputs (context [SEP] response, label)
Update Encoder Weights
RoBERTa Encoder
Generate Embeddings
attract
h_{sarc}
attract
repel
h_{norm}
Embedding Space
Supervised Contrastive Loss

Figure 1. Overview of the Stage 1, Supervised Contrastive Pre-training.

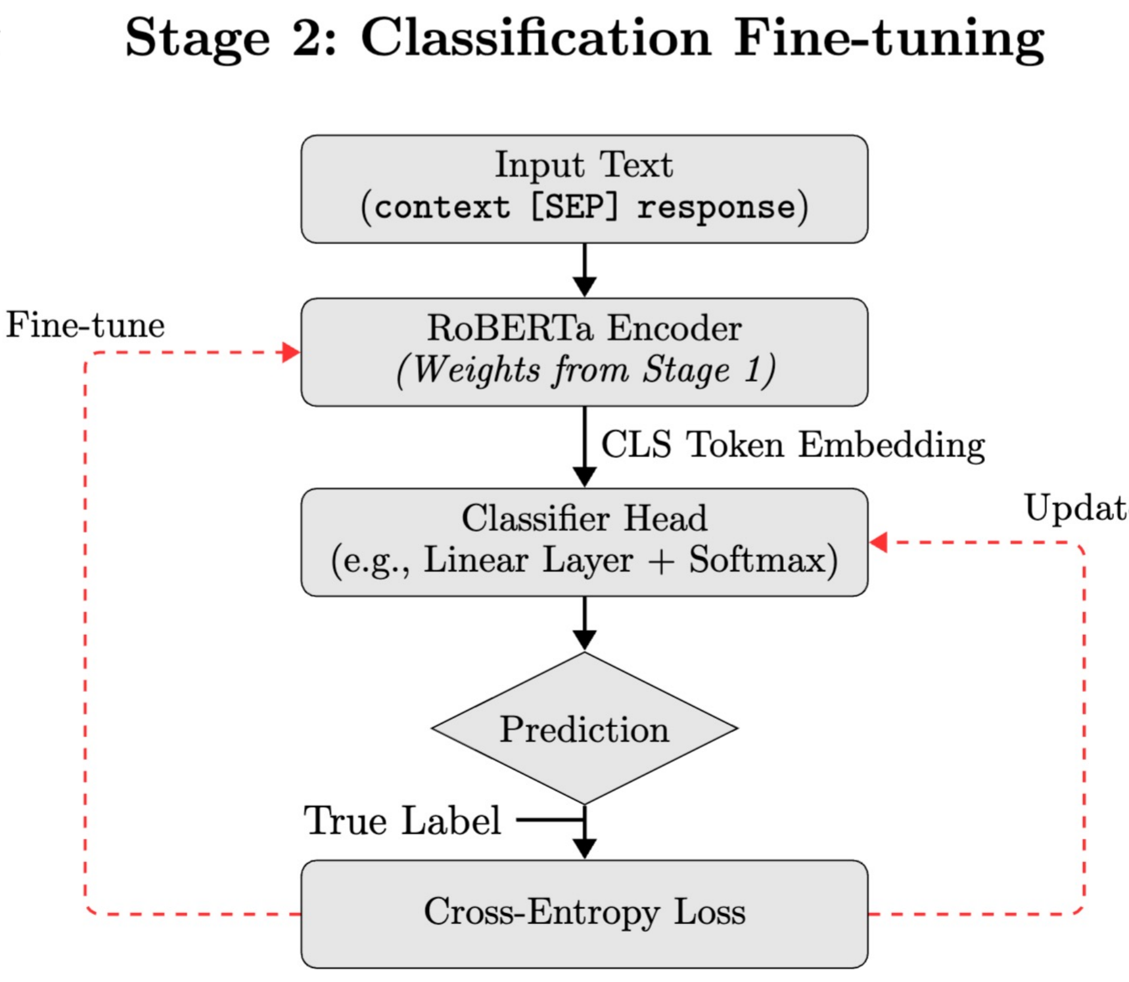
Stage 2: Classification Fine-tuning
Input Text (context [SEP] response)
Fine-tune
RoBERTa Encoder (Weights from Stage 1)
CLS Token Embedding
Classifier Head (e.g., Linear Layer + Softmax)
Update
Prediction
True Label
Cross-Entropy Loss

Figure 2. Overview of the Stage 2, Classification Fine-tuning with CE Loss.

Experiments & Results

Datasets and Setup

- **Dataset**: FigLang 2020 Shared Task (Reddit, Twitter)
- **Model**: RoBERTa-base
- **Baseline (Ablation)**: Fine-tuned with CE loss for 10 epochs (with early stopping)
- **Ours (SimSCLSD)**: SupCon for 20 epochs + CE for 10 epochs (with early stopping)
- **Hyperparameters**: Batch = 64, Temp (τ) = 0.2, Dropout = 0.5 (R) / 0.1 (T), SupCon LR = $5 * 10^{-5}$, FT Encoder LR = $1 * 10^{-6}$ (R) / $5 * 10^{-7}$ (T), FT Classifier LR = $5 * 10^{-5}$ (R) / $3 * 10^{-5}$ (T)

Classification Results

Dataset	Model	Precision	Recall	Macro F1
Reddit	Baseline	0.5244	0.5228	0.5148
	SimSCLSD	0.6166	0.6139	0.6116
Twitter	Baseline	0.7047	0.6861	0.6788
	SinSCLSD	0.7477	0.7461	0.7457

Table 1. Comparison of test set classification performance (Macro-averaged) between the Baseline (standard fine-tuning) and our proposed SimSCLSD framework.

- **9.7%p** improvement in F1-score Reddit test dataset
- **6.7%p** improvement in F1-score Twitter test dataset

Conclusion

- We proposed **SimSCLSD**, a simple and effective 2-stage framework for context-aware sarcasm detection.
  - By first adapting a RoBERTa encoder with a supervised contrastive loss, we create a **more structured and discriminative feature space**.
- This methodology provides a **superior initialization for the final classification** fine-tuning, leading to performance gains on the Reddit(**9.7%p F1 increase**) and Twitter(**6.7%p F1 increase**) datasets.
- **Key Takeaway**: Separating the optimization of the feature representation from the classification is a highly effective strategy for nuanced, context-dependent tasks, boundary.
- **Future work:** Apply the framework to larger language models or extend this methodology to other complex, context-dependent NLP tasks(e.g. irony, stance, or sentiment detection).

References

[1] D. Ghosh, A. Vajpayee, and S. Muresan, "A Report on the 2020 Sarcasm Detection Shared Task," in Proc. FigLang, 2020.
[2] Y. Liu, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint, 2019.
[3] K. Pant and T. Dadu, "Sarcasm Detection using Context Separators in Online Discourse," in Proc. FigLang, 2020.
[4] Y. Moukafih, et al., "SimSCL: A Simple fully-Supervised Contrastive Learning Framework for Text Representation," arXiv preprint, 2022.
[5] P. Khosla, et al., "Supervised Contrastive Learning," in Proc. NeurIPS, 2020.
[6] J. Devlin, et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL, 2019.