Fine-Tuning Pre-trained Language Model with Weak Supervision

Qikai Yang[1]

[1]Department of Computer Science , University of Illinois at Urbana-Champaign ,
201 S. Goodwin Ave, Urbana, IL 61801 , USA

Contents

Abstract

Fine-tuned pre-trained language models (LMs) have been a powerful tool to handle various problems in Natural Language Processing such as sequence classification, sentimental analysis, etc. However, it requires heavy human work on labeling and can be very time-consuming. Therefore, for this technology review, we are going to focus on studying on one of the current mainstream methods that combines weak supervision with fine-tuned pre-trained LMs - COSINE. In addition, we also want to compare COSINE with other state of the art (SOTA) methods.

Keywords: fine-tuned pre-tained language models, weak supervision, state of the art methods

1 Introduction

COSINE (Yu et al., 2021) is an algorithm that fine-tunes pre-trained LMs with only weak supervision. It is aimed to solve weakly-supervised classification problem. The input includes a bunch of unlabeled data and labeled data. And the output is a classifier that classifies each line of data into the pre-defined label class.

For the labeled data, it only make uses of weak super-vision sources such as keywords and semantic rules (Yu et al., 2021). Instead of having single weak super-vision source, it applies multiple weak super-vision sources to generate initial labels.

The high level ideas of COSINE have two innovation parts. Firstly, it divides the loss into three different parts - classification loss, contrastive loss, and confidence loss. Then each part of the loss is backpropagated separately. Secondly, the training process involves multiple iterations to make classifications for both matched and unmatched samples. The more detailed process can be seen as the figure below.

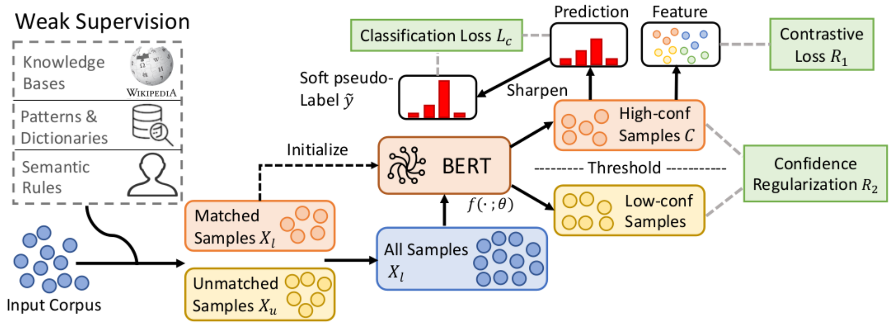
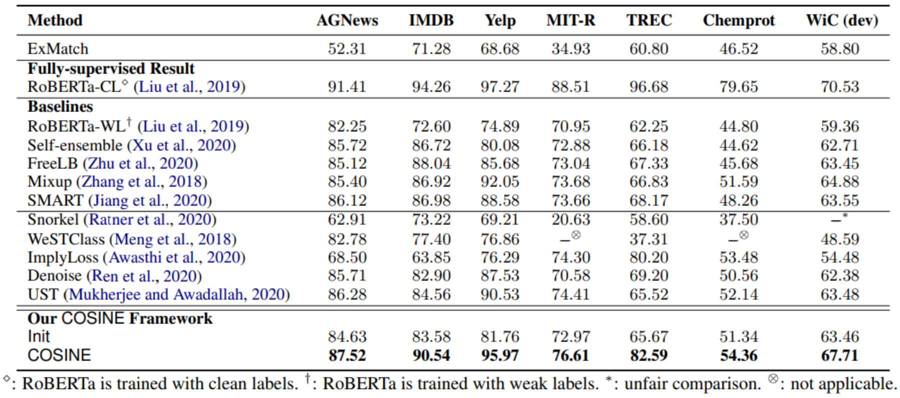


Figure 1: Training process of COSINE. Figure source - Yu et al. (2021)

As seen in the figure, the network consists of two parts - the RoBERTa (a pre-trained language model that outputs hidden representations of input samples) (Liu et al., 2019), and a task-specific classification head that outputs the prediction confidence of different classes (Yu et al., 2021). In the next section, we are going to discuss about COSINE's performance compared with other SOTA methods.

2 Comparison

In the paper (Yu et al., 2021), the experiments contain three different tasks - sequence classification, token classification, and sentence pair classification. Here we only select part of the results for our discussion.



Method	AGNews	IMDB	Yelp	MIT-R	TREC	Chemprot	WiC (dev)
ExMatch	52.31	71.28	68.68	34.93	60.80	46.52	58.80
Fully-supervised Result							
RoBERTa-CL◇ (Liu et al., 2019)	91.41	94.26	97.27	88.51	96.68	79.65	70.53
Baselines							
RoBERTa-WL† (Liu et al., 2019)	82.25	72.60	74.89	70.95	62.25	44.80	59.36
Self-ensemble (Xu et al., 2020)	85.72	86.72	80.08	72.88	66.18	44.62	62.71
FreeLB (Zhu et al., 2020)	85.12	88.04	85.68	73.04	67.33	45.68	63.45
Mixup (Zhang et al., 2018)	85.40	86.92	92.05	73.68	66.83	51.59	64.88
SMART (Jiang et al., 2020)	86.12	86.98	88.58	73.66	68.17	48.26	63.55
Snorkel (Ratner et al., 2020)	62.91	73.22	69.21	20.63	58.60	37.50	–*
WeSTClass (Meng et al., 2018)	82.78	77.40	76.86	–⊗	37.31	–⊗	48.59
ImplyLoss (Awasthi et al., 2020)	68.50	63.85	76.29	74.30	80.20	53.48	54.48
Denoise (Ren et al., 2020)	85.71	82.90	87.53	70.58	69.20	50.56	62.38
UST (Mukherjee and Awadallah, 2020)	86.28	84.56	90.53	74.41	65.52	52.14	63.48
Our COSINE Framework							
Init	84.63	83.58	81.76	72.97	65.67	51.34	63.46
COSINE	**87.52**	**90.54**	**95.97**	**76.61**	**82.59**	**54.36**	**67.71**

◇: RoBERTa is trained with clean labels. †: RoBERTa is trained with weak labels. *: unfair comparison. ⊗: not applicable.

Figure 2: Comparison results of COSINE. Figure source - Yu et al. (2021)

As seen in the table, first of all, COSINE improves the initial pre-trained models' accuracies greatly. On all of the seven datasets, the accuracy increases at least three hundred percent after applying COSINE. In addition, compared to all other SOTA methods appeared in the table, COSINE shows significantly higher accuracies. Especially for IMDB dataset, COSINE has remarkable performances. Lastly, even for the unfair comparison, COSINE doesn't show significant gap for the fully-supervised methods.

3 Summary

In summary, COSINE is an innovative algorithm that has done an impressive work on combining weak supervision and fine-tunes pre-trained LMs. It not only shows remarkable performances compared to previous SOTA methods having similar ideas, but also has comparable performance with fully-supervised methods. It also has no "corner performance" on any of the mainstream datasets that the paper is doing experiments on. It two main innovative ideas are worth further research and could be used to develop new algorithms as well.

References

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Yu, Y., S. Zuo, H. Jiang, W. Ren, T. Zhao, and C. Zhang. 2021. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach.