When Low Resource NLP Meets Unsupervised Language Model: Meta-Pretraining then Meta-Learning for Few-Shot Text Classification (Student Abstract)

Shumin Deng,[1,2*] **Ningyu Zhang,**[2,4*] **Zhanlin Sun,**[2,5*] **Jiaoyan Chen,**[6] **Huajun Chen**[1,2,3†]
[1]College of Computer Science and Technology, Zhejiang University
[2]Alibaba-Zhejiang University Frontier Technology Research Center Joint Lab for Knowledge Engine
[3]The First Affiliated Hospital of Zhejiang University
[4]Alibaba Group, [5]Carnegie Mellon University, [6]Oxford University
{231sm, huajunsir}@zju.edu.cn, zhanlins@andrew.cmu.edu
jiaoyan.chen@cs.ox.ac.uk, ningyu.zny@alibaba-inc.com

Abstract

Text classification tends to be difficult when data are deficient or when it is required to adapt to unseen classes. In such challenging scenarios, recent studies have often used meta-learning to simulate the few-shot task, thus negating implicit common linguistic features across tasks. This paper addresses such problems using meta-learning and unsupervised language models. Our approach is based on the insight that having a good generalization from a few examples relies on both a generic model initialization and an effective strategy for adapting this model to newly arising tasks. We show that our approach is not only simple but also produces a state-of-the-art performance on a well-studied sentiment classification dataset. It can thus be further suggested that pretraining could be a promising solution for few-shot learning of many other NLP tasks. The code and the dataset to replicate the experiments are made available at https://github.com/zxlzr/FewShotNLP.

Introduction

Deep learning (DL) has achieved great success in many fields owing to the advancements in optimization techniques, large datasets, and streamlined designs of deep neural architectures. However, DL is notorious for requiring large labeled datasets, which limits the scalability of a deep model to new classes owing to the cost of annotation. Few-shot learning generally resolves the data deficiency problem by recognizing novel classes from very few labeled examples. This limitation in the size of samples (only one or very few examples) challenges the standard fine-tuning method in DL. Early studies in this field applied data augmentation and regularization techniques to alleviate the overfitting problem caused by data scarcity but only to a limited extent. Instead, researchers have been inspired by exploration of meta-learning (Finn and et al. 2017) to leverage the distribution over similar tasks. However, existing meta-learning approaches for few-shot learning can not explicitly disentangle task-agnostic and task-specific representations, and they are not able to take advantage of the knowledge of linguistic properties via unsupervised language models.

In this paper, we raise the question that **whether it is possible to boost the performance of low-resource natural language processing with the large scale of raw corpus via unsupervised leaning**, which require us to handle both task-agnostic and task-specific representation learning. Thus we propose a **M**eta-pretraining **T**hen **M**eta-learning (**MTM**) approach motivated by the observation that meta-learning leads to learning a better parameter initialization for new tasks than multi-task learning across all tasks. The former meta-pretraining is to learn task-agnostic representations that explicitly learns a model parameter initialization for enhanced predictive performance with limited supervision. The latter meta-learning considers all classes as coming from a joint distribution and seeks to learn model parameters that can be quickly adapted via using each class's training instances to enhance predictive performance on its test set. In other words, our approach explicitly disentangles the task-agnostic and task-specific feature learning. Experimental results demonstrate that the proposed model achieves significant improvement on public benchmark datasets.

Approach

Problem Definition

Few-shot text classification (Yu and et al. 2018; Geng and et al. 2019) is a task in which a classifier must adapt new classes that are not seen in training, given only a few examples for each of these new classes. To be specific, we have a labeled training set with a set of defined classes $\mathcal{C}_{train}$. Our goal is to output classifiers on the testing set with a disjoint set of new classes $\mathcal{C}_{test}$ when only a small labeled support set is available. If the support set contains K labeled examples for each of the C unique classes, the target few-shot problem is called a C-way-K-shot problem. The sample set is usually too small to train a supervised classification model. To this end, we try to utilize meta-learning method on the training set to extract task-agnostic knowledge, which

[*]All authors contributed equally to this work.
[†]Corresponding author.

may perform better for few-shot text classification on the test set.

Training Procedure

Task-agnostic Meta Pretraining. Given all the training samples, we first utilize pretraining strategies such as BERT to learn task-agnostic contextualized features that capture linguistic properties to benefit downstream few-shot text classification tasks.

Meta-learning Text Classification. Given the pretrained language representations, we construct episodes to compute gradients and update the model in each training iteration.

Algorithm 1 MTM Algorithm

Require: Training Datapoints $\mathcal{D} = \left\{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\right\}$

1: Construct a task T_j with training examples using a support set $\mathcal{S}_K^{(j)}$ and a test example $\mathcal{D}'_j = \left(\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\right)$
2: Randomly initialize θ
3: Pre-train $\mathcal{D}$ with unsupervised language models
4: Denote $p(\mathcal{T})$ as distribution over tasks
5: **while** not done **do**
6: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$:
7: **for** for all T_i **do**
8: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}\left(f_\theta\right)$ using $\mathcal{S}_K^{(j)}$
9: Compute adapted parameters with gradient descent: $\theta'_i = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}\left(f_\theta\right)$
10: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}\left(f_{\theta'_i}\right)$ using each $\mathcal{D}'_i$ from T_i and $\mathcal{L}_{\mathcal{T}_i}$

Experiments

Datasets and Evaluation

We use the multiple tasks with the multi-domain sentiment classification dataset ARSC[1]. This dataset comprises English reviews for 23 types of products on Amazon. For each product domain, there are three different binary classification tasks. These buckets then form $23 \times 3 = 69$ tasks in total. We select $12(4 \times 3)$ tasks from four domains as the test set, with only five examples as support set for each label in the test set. We evaluate the performance by few-shot classification accuracy following previous studies in few-shot learning (Snell, Swersky, and Zemel 2017). To evaluate the proposed model objectively with the baselines, note that for ARSC, the support set for testing is fixed by (Yu and et al. 2018); therefore, we need to run the test episode once for each of the target tasks. The mean accuracy from the 12 target tasks are compared to those of the baseline models in accordance with (Yu and et al. 2018).

Evaluation Results

The evaluation results are shown in Table 1: **MTM** is our current approach, **Match Network** (Vinyals and et al. 2016) is a few-shot learning model using metric-based attention method, **Prototypical Network** (Snell, Swersky, and Zemel 2017) is a deep matrix-based method using sample averages as class prototypes, **MAML** (Finn and et al. 2017) is a model-agnostic method that is compatible with any model trained with gradient descent and applicable to a variety of learning problems, **Relation Network** (Sung and et al. 2018) is a metric-based few-shot learning model that uses a neural network as the distance measurement and calculate class vectors by summing sample vectors in the support set, **ROBUSTTC-FSL** (Yu and et al. 2018) is an approach that combines adaptive metric methods by clustering the tasks, **Induction-Network-Routing** (Geng and et al. 2019) is a recent state-of-the-art method which learn generalized class-wise representations by combining the dynamic routing algorithm with a typical meta-learning framework. From the results shown in Table 1, we observe that our approach achieves the best results amongst all meta-learning models. Note that, our model is task-agnostic, which means it can be easily adapted to any other NLP tasks.

Model	Mean Acc
Matching Network	65.73
Prototypical Network	68.15
Relation Network	83.74
MAML	78.33
ROBUSTTC-FSL	83.12
Induction-Network-Routing	85.47
MTM	**90.01***

Table 1: Comparison of mean accuracy (%) on ARSC. * indicates $p_{value} < 0.01$ in a paired t-test (10-fold) evaluation.

Conclusion

In this study, we attempt to analyze language meta-pretraining with meta-learning for few-shot text classification. Results show that our model outperforms conventional state-of-the-art few-shot text classification models. In the future, we plan to apply our method to other NLP scenarios.

Acknowledgments

We want to express gratitude to the anonymous reviewers for their hard work and kind comments and this work is funded by NSFC 91846204, national key research program 2018YFB1402800, and Alibaba CangJingGe(Knowledge Engine) Research Plan.

References

Finn, C., and et al. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*.

Geng, R., and et al. 2019. Few-shot text classification with induction network. *In EMNLP*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *NIPS*, 4077–4087.

Sung, F., and et al. 2018. Learning to compare: Relation network for few-shot learning. In *CVPR*, 1199–1208.

Vinyals, O., and et al. 2016. Matching networks for one shot learning. In *NIPS*, 3630–3638.

Yu, M., and et al. 2018. Diverse few-shot text classification with multiple metrics. *In NAACL*.

[1] https://github.com/Gorov/DiverseFewShot_Amazon