Improving Few-shot Text Classification via Pretrained Language Representations

Ningyu Zhang[1,2,3] **Zhanlin Sun**[1,3] **Shumin Deng**[1,3] **Jiaoyan Chen**[4] **Huajun Chen**[1,3]*
1. College of Computer Science and Technology, Zhejiang University
2. Alibaba Group
3. AZFT[†]Joint Lab for Knowledge Engine
4. Department of Computer Science, Oxford University
{3150105645,231sm,huajunsir}@zju.edu.cn
jiaoyan.chen@cs.ox.ac.uk,{ningyu.zny}@alibaba-inc.com

Abstract

Text classification tends to be difficult when the data is 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 explicit common linguistic features across tasks. Deep language representations have proven to be very effective forms of unsupervised pretraining, yielding contextualized features that capture linguistic properties and benefit downstream natural language understanding tasks. However, the effect of pretrained language representation for few-shot learning on text classification tasks is still not well understood. In this study, we design a few-shot learning model with pretrained language representations and report the empirical results. We show that our approach is not only simple but also produces 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.

1 Introduction

Deep learning (DL) has achieved great success in many fields such as computer vision, speech recognition, and machine translation (Kuang et al., 2017; Peng et al., 2018; Lin et al., 2017) thanks to the advancements in optimization techniques, larger 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. Humans, however, are readily able to learn and distinguish new classes rapidly with only a few examples. This gap between human and machine learning provides opportunities for DL development and applications.

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 (Scholkopf and Smola, 2001) 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 human learning to explore meta-learning (Finn et al., 2017) to leverage the distribution over similar tasks. Contemporary approaches to few-shot learning often decompose the training procedure into an auxiliary meta-learning phase, which includes many sub-tasks, following the principle that the testing and training conditions must match. They extract some transferable knowledge by switching the task from one mini-batch to the next. Moreover, the few-shot model is able to classify data into new classes with just a small labeled support set.

Existing approaches for few-shot learning are still plagued by problems, including imposed strong priors (Fe-Fei et al., 2003), complex gradient transfer between tasks (Munkhdalai and Yu, 2017), and fine-tuning of the target problem (Long et al., 2016). The approaches proposed by (Snell et al., 2017) and (Sung et al., 2018), which combine non-parametric methods and metric learning, may provide possible solutions to these problems. The non-parametric methods allow novel examples to be rapidly assimilated without suffering from the effects of catastrophic overfitting. Such non-parametric models only need

* Corresponding author.
† Alibaba-Zhejiang University Frontier Technology Research Center

to learn the representation of the samples and the metric measure.

Recently, a variety of techniques were proposed for training general-purpose language representation models using an enormous amount of unannotated text, such as ELMo (Peters et al., 2018) and generative pretrained transformer (GPT) (Radford et al., 2018). Pretrained models can be fine-tuned on natural language processing (NLP) tasks and have achieved significant improvements over training on task-specific annotated data. More recently, a pretraining technique named bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) was proposed and has enabled the creation of state-of-the-art models for a wide variety of NLP tasks, including question answering (SQuAD v1.1) and natural language inference, among others.

However, there have not been many efforts in exploring pretrained language representations for few-shot text classification. The technical contributions of this work are two-fold: 1) we explore the pretrained model to address the poor generalization capability of text classification, and 2) we propose a meta-learning model based on model-agnostic meta-learning (MAML) which explicitly disentangles the task-agnostic feature learning and task-specific feature learning to demonstrate that the proposed model achieves significant improvement on text classification accuracy on public benchmark datasets. To the best of our knowledge, we are the first to bridge the pretraining strategy with meta-learning methods for few-shot text classification.

2 Background: Meta-Learning

Our work is built on the recently proposed MAML framework (Finn et al., 2017), which we describe briefly here. MAML aims to learn the learners (for the tasks) and the meta-learner in the few-shot meta-learning setup (Vinyals et al., 2016; Ravi and Larochelle, 2016; Andrychowicz et al., 2016). Formally, it considers a model that is represented by a function f_θ with parameters θ. When the model adapts to a new task T_i, the model changes the parameters from one θ_i to the next, where a task contains K training examples and one or more test examples (K-shot learning). MAML updates the parameters θ_i via one or a few iterations of gradient descent based on the training examples of task T_i. For example, for one gradient update,

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}\left(f_\theta\right),$$

where the step size α is a hyperparameter; $\mathcal{L}_{\mathcal{T}_i}\left(f_\theta\right)$ is a loss function that evaluates the error between the prediction $f_\theta\left(\mathbf{x}^{(j)}\right)$ and target $\mathbf{y}^{(j)}$, where $\mathbf{x}^{(j)}$, $\mathbf{y}^{(j)}$ are input–output pairs sampled from the training examples of task T_i. Model parameters θ are trained to optimize the performance of $f_{\theta_i'}$ on the unseen test examples from T_i across tasks. The meta-objective is as follows:

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}\left(f_{\theta_i'}\right) = \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}\left(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}\left(f_\theta\right)}\right)$$

The goal of MAML is to optimize the model parameters θ such that the model can learn to adapt to new tasks with parameters via a few gradient steps on the training examples of the new tasks. The model is improved by considering how the test errors on the unseen test data from T_i change with respect to the parameters. The meta-objective across tasks is optimized using stochastic gradient descent (SGD). The model parameters θ are updated as follows:

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\tau_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}\left(f_{\theta_i'}\right)$$

3 Approach

3.1 Problem definition

Few-shot classification is a task in which a classifier must adapt and accommodate new classes that are not seen in training, given only a few examples of each of these new classes. We have a large labeled training set with a set of defined classes C_{train}. However, after training, our ultimate goal is to produce classifiers on the testing set with a disjoint set of new classes C_{test} for which only a small labeled support set will be 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. Usually, K is a too small sample set to train a supervised classification model. Therefore, we aim to perform meta-learning on the training set in order to extract transferrable knowledge that will allow us to perform better few-shot learning on the support set to classify the test set more successfully.

3.2 Training Procedure

The training procedure of our approach consists of two parts.

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

Episode-based Meta Training. Given the pretrained language representations, we construct episodes to compute gradients and update the model in each training iteration with MAML.

3.3 Language Representation Pretraining

While the pretraining tasks have been designed with particular downstream tasks in mind (Felbo et al., 2017), we focus on those training tasks that seek to induce *universal* representations suitable for downstream few-shot learning tasks. We utilize BERT (Devlin et al., 2018) as a recent study (Peters et al., 2019) has shown its potential to achieve state-of-the-art performance when fine-tuned in NLP tasks. BERT combines both word and sentence representations (via masked language model and next sentence prediction objectives) in a single very large pretrained transformer (Vaswani et al., 2017). It is adapted to both word- and sentence-level tasks with task-specific layers. We feed the sentence representation into a softmax layer for text classification based on (Devlin et al., 2018).

3.4 Episode-Based Meta Training

Given the pretrained language representations, we construct episodes to compute the gradients and update our model in each training iteration. The training episode is formed by randomly selecting a subset of classes from the training set, then choosing a subset of examples within each selected class to act as the support set S with a subset of the remaining examples to serve as the query set Q. Training with such episodes is achieved by feeding the support set S to the model and updating its parameters to minimize the loss in the query set Q. We call this strategy as episode-based meta training. The adaptation of meta-learning using the MAML framework with pretrained language representations is summarized in Algorithm 1, called **P-MAML**. The use of episodes makes the training procedure more faithful to the test environment, thereby improving generalization. It is worth noting that there are exponentially many possible meta tasks to train the model with, thus making it difficult to overfit.

Algorithm 1 P-MAML Algorithm

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

Construct a task T_j with training examples using a support set $\mathcal{S}_K^{(j)}$ and a test example $\mathcal{D}'_j = \left(\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\right)$
Randomly initialize θ
Pre-train $\mathcal{D}$ with BERT
Denote $p(\mathcal{T})$ as distribution over tasks
while not done **do**
    Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$:
    for all T_i **do**
        Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}\left(f_\theta\right)$ using $\mathcal{S}_K^{(j)}$
        Compute adapted parameters with gradient descent: $\theta'_i = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}\left(f_\theta\right)$
    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 $\mathcal{L}_{\mathcal{T}_i}$

4 Experiments

4.1 Datasets and Evaluation

We use the multiple tasks with the multi-domain sentiment classification (Blitzer et al., 2007) 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. Following (Yu et al., 2018), we select 12 (4 $\times$ 3) tasks from four domains (i.e., Books, DVDs, Electronics, and Kitchen) as the test set, with only five examples as support set for each label in the test set. We thus create 5-shot learning models on this dataset. We evaluate the performance by few-shot classification accuracy following previous studies in few-shot learning (Snell et al., 2017; Sung et al., 2018). To evaluate the proposed model objectively with the baselines, note that for ARSC, the support set for testing is fixed by (Yu 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 is compared to those of the baseline models in accordance with (Yu et al., 2018). We use pretrained BERT-Base[2] for the ARSC dataset. All model parameters are updated by backpropagation using Adam with a learning rate of 0.01. We regularize our network using dropout with a rate of 0.3 tuned using the development set.

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

[2] We use BERT-BASE for computation efficiency and maybe BERT-LARGE will gain more improvement.

4.2 Evaluation Results

To evaluate the performance of our model, we compared it with various baseline models. The evaluation results are shown in Table 1: **P-MAML** is our current approach, **Match Network** (Vinyals et al., 2016) is a few-shot learning model using metric-based attention method, **Prototypical Network** (Snell et al., 2017) is a deep matrix-based method using sample averages as class prototypes, **MAML** (Finn et al., 2017) is an MAML method that is compatible with any model trained with gradient descent and applicable to a variety of learning problems, **Relation Network** (Sung 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 et al., 2018) is an approach that combines adaptive metric methods by clustering the tasks, **Induction-Network-Routing** (Geng et al., 2019) is a recent state-of-the-art method which learn generalized classwise 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. Compared with ROBUSTTC-FSL and Induction-Network-Routing, which adopt several metric methods and dynamic routing algorithms, our approach still provides more advantages. We believe the performance of our model can be further improved by adopting additional mechanisms like adaptive metrics, which will be part of our future work. Note that, our approach is very simple and independent of the encoder choices, and can, therefore, be easily adapted to fit other encoder architectures for sophisticated 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	85.63
P-MAML	**86.65***

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

4.3 Ablation Study

To analyze the contributions and effects of language representation pretraining in our approach, we perform ablation tests. **GloVe** is the method with pretrained GloVe (Pennington et al., 2014) word embeddings; **w/o pretrain** is our method without pre-trained embeddings (random initialization). From the evaluation results in Table 2, we observe the performance drop significantly without pretraining, which proves the effectiveness of explicit common linguistic features learning. We also notice that our model with GloVe does not achieve good performance even compared with the random initialization, which indicates that the poor generalization capability for few-shot text classification.

Model	Mean Acc
P-MAML	**86.65**
GloVe	78.50
w/o pretrain	79.81

Table 2: Results of ablation study.

4.4 Discussions

It should be noted that human beings are intelligent to leverage learned knowledge about the world in understanding language. (Cook and Newson, 2014) think human beings have a universal grammar, and our daily language system is only a formal expression of this universal grammar. In other words, there are deep structures related to concepts and superficial structures related to speech and symbols in a language. Moreover, neuroscience research has proposed a prominent idea that language processing may offer such a principle that the brain contains partially separate systems for processing syntax and semantics. The part of the prefrontal cortex responsible for language production, called Brocas area, is thought to be important for parsing syntactic information and applying selective attention to help a separate comprehension system interpret the semantics (Russin et al., 2019). Our idea for few-shot learning in NLP is somewhat similar to this assumption as the pretraining stage may learn common syntax information across tasks, and the meta-learning stage may learn semantic knowledge, which is task specific.

4.5 Conclusion

In this study, we attempt to analyze language representation pretraining for few-shot text classification empirically. We combine the MAML algorithm with the pretraining strategy to disentangle the task-agnostic and task-specific representation learning. 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.

References

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. 2016. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, pages 3981–3989.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Domain adaptation for sentiment classification. In *45th Annv. Meeting of the Assoc. Computational Linguistics (ACL07)*.

Vivian Cook and Mark Newson. 2014. *Chomsky's universal grammar*. John Wiley & Sons.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Fe-Fei et al. 2003. A bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1134–1141. IEEE.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org.

Ruiying Geng, Binhua Li, Yongbin Li, Yuxiao Ye, Ping Jian, and Jian Sun. 2019. Few-shot text classification with induction network. *arXiv preprint arXiv:1902.10482*.

Shaohui Kuang, Junhui Li, António Branco, Weihua Luo, and Deyi Xiong. 2017. Attention focusing for neural machine translation by bridging source and target embeddings. *arXiv preprint arXiv:1711.05380*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2016. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144.

Tsendsuren Munkhdalai and Hong Yu. 2017. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2554–2563. JMLR. org.

Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuanjing Huang. 2018. Cross-domain sentiment classification with target domain specific information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2505–2513.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Matthew Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI.

Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning.

Jake Russin, Jason Jo, and Randall C O'Reilly. 2019. Compositional generalization in a deep seq2seq model by separating syntax and semantics. *arXiv preprint arXiv:1904.09708*.

Bernhard Scholkopf and Alexander J Smola. 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. *arXiv preprint arXiv:1805.07513*.