arXiv:2004.01907v1 [cs.CL] 4 Apr 2020

Knowledge Guided Metric Learning for Few-Shot Text Classification

Dianbo Sui[1,2], **Yubo Chen**[1], **Binjie Mao**[1,2], **Delai Qiu**[3], **Kang Liu**[1,2], **Jun Zhao**[1,2]
[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
[3] Beijing Unisound Information Technology Co., Ltd, Beijing, 100028, China
{dianbo.sui, yubo.chen, binjie.mao, kliu, jzhao}@nlpr.ia.ac.cn
qiudelai@unisound.com

Abstract

The training of deep-learning-based text classification models relies heavily on a huge amount of annotation data, which is difficult to obtain. When the labeled data is scarce, models tend to struggle to achieve satisfactory performance. However, human beings can distinguish new categories very efficiently with few examples. This is mainly due to the fact that human beings can leverage knowledge obtained from relevant tasks. Inspired by human intelligence, we propose to introduce external knowledge into few-shot learning to imitate human knowledge. A novel parameter generator network is investigated to this end, which is able to use the external knowledge to generate relation network parameters. Metrics can be transferred among tasks when equipped with these generated parameters, so that similar tasks use similar metrics while different tasks use different metrics. Through experiments, we demonstrate that our method outperforms the state-of-the-art few-shot text classification models.

1 Introduction

The ability to quickly learn from a small number of examples is a critical feature of human intelligence. This motivates research of few-shot learning (Vinyals et al., 2016; Snell et al., 2017; Finn et al., 2017; Sung et al., 2018), which aims to classify unseen data instances (*testing* examples) into a set of new categories with few labeled samples (*support* examples) in each category. In the few-shot setting, the model is trained, when given a specific task, to produce a classifier for that specific task. Therefore, the model is exposed to different tasks during the training phase, and it is evaluated on a non-overlapping set of new tasks.

The key challenge in few-shot learning is to make full use of the limited labeled examples available in the support set to find the "right" generalizations as suggested by the task. Metric-based approaches (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018) are effective ways to address this challenge. In these approaches, examples are represented into a feature space and then predictions are made using a metric between the representations of testing examples and support examples. However, employing metric-based approaches directly in text classification faces a problem that tasks are diverse and significantly different from each other, since words that are highly informative for one task may not be relevant for other tasks (Bao et al., 2019). Therefore, a single metric is insufficient to cope with all these tasks in few-shot text classification (Yu et al., 2018).

To adapt metric learning to significantly diverse tasks, we propose a knowledge guided metric learning method. This method is inspired by the fact that human beings approach diverse tasks armed with knowledge obtained from relevant tasks (Lake et al., 2017). We use external knowledge from the knowledge base (KB) to imitate human knowledge, while the role of external knowledge has been ignored in previous methods (Yu et al., 2018; Bao et al., 2019; Geng et al., 2019). In detail, we resort to distributed representations of the KB instead of symbolic facts, since symbolic facts face the issues of poor generalization and data sparsity. Based on such KB embeddings, we investigate a novel parameter generator network (Ha et al., 2016; Jia et al., 2016) to generate task-relevant relation network parameters. With these generated parameters, the task-relevant relation network is able to apply diverse metrics to diverse tasks and ensure that similar tasks use similar metrics while different tasks use different metrics.

In summary, the major contributions of this paper are:

- Inspired by human intelligence, we present the

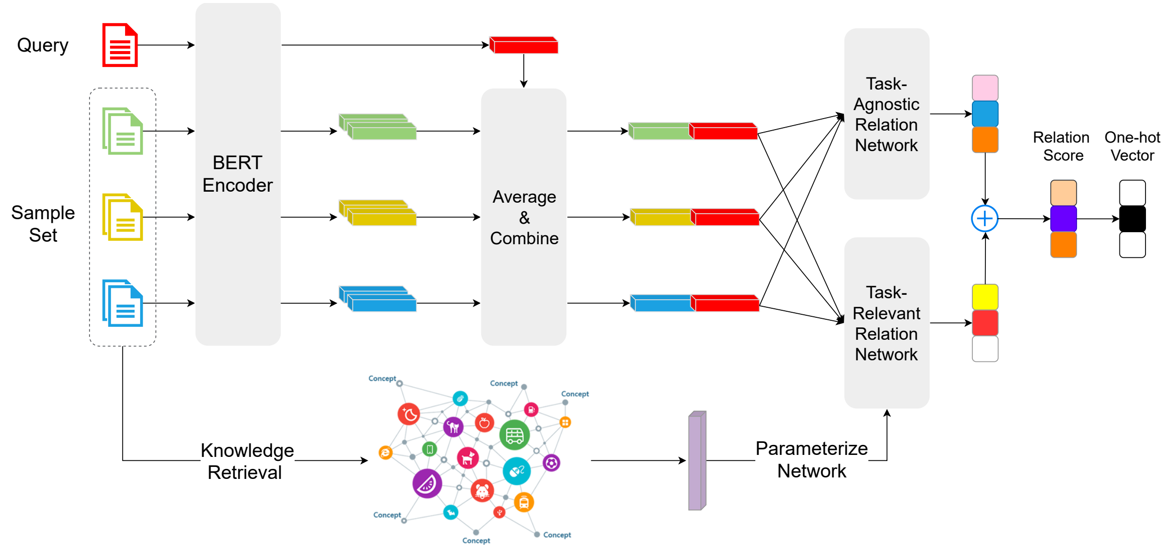

Figure 1: The main architecture for a C-way N-shot (C = 3, N = 2) problem with one query example

first approach that introduces external knowledge into few-shot learning.

- A novel parameter generator network based on external knowledge is proposed to generate diverse metrics for diverse tasks.
- The model yields promising results on the ARSC dataset of few-shot text classification.

2 Problem Definition

Few-shot classification is a task in which a classifier is learned to recognize unseen classes during training with limited labeled examples. Formally, There are three datasets: a training set, a support set, and a testing set. The support set and testing set share the same label space, but the training set has its own label space that is disjoint with support/testing set. If the support set contains N labeled examples for each of C unique classes, the target few-shot problem is called C-way N-shot. In principle, we can train a classifier with the support set only. However, such a classifier usually performs badly on the testing set due to the scarcity of labeled data. Therefore, performing meta-learning on the training set is necessary, which aims to extract transferable knowledge on the training set that will assist the classifier to classify the testing set more successfully.

In meta-learning, testing scenario is simulated during meta-training so the classifier can learn to quickly learn from a few annotations, which is called *episode* based training (Vinyals et al., 2016). In each training iteration, an episode (or task) is formed by randomly selecting C classes from the training set with N labelled samples from each of the C classes to serve as the *sample* set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{m} (m = C \times N)$, as well as a fraction of the remainder of those C classes samples to act as the *query* set $\mathcal{Q} = \{(x_i, y_i)\}_{i=1}^{n}$, where x_i is a sentence and $y_i \in \{1, ..., C\}$ is the corresponding label. This sample/query set split is designed to imitate the support/testing set when testing.

3 Methodology

3.1 Sentence Embedding Network

A pre-trained BERT (Devlin et al., 2019) encoder is used to model sentences. Given an input text $x_i = ([\text{CLS}], w_1, w_2, ..., w_T, [\text{SEP}])$ as input, the output of BERT encoder is denoted as $\mathbf{H}(x_i) \in \mathbb{R}^{(T+2)\times d_1}$, where d_1 is the output dimension of the BERT encoder. We use the first token of the sequence (classification token) as the sentence representation, which is denote as $\mathbf{h}(x_i)$.

In meta-learning, each class representation is the mean vector of the embedded sample sentences belonging to its class,

$$\mathbf{c}_z = \frac{1}{|\mathcal{S}_z|} \sum_{(x_i, y_i) \in \mathcal{S}_z} \mathbf{h}(x_i) \in \mathbb{R}^{d_1} \tag{1}$$

where $\mathcal{S}_z$ denotes the set of examples labeled with class z. Following Sung et al. (2018), we use concatenation operator to combine the query sentence representation $\mathbf{h}(x_j)$ with the class representation $\mathbf{c}_z$.

$$\mathbf{p}_{z,j} = \text{concatenation}(\mathbf{c}_z, \mathbf{h}(x_j)) \in \mathbb{R}^{2d_1} \tag{2}$$

3.2 Knowledge Guided Relation Network

This module takes sample set knowledge and combined representation (shown in Equation 2) as input, and produces a scalar in range of 0 to 1 representing the similarity between the query sentence and the class representation, which is called relation score. Compared with the original relation network (Sung et al., 2018), we decompose the relation network into two parts, task-agnostic relation network and task-relevant relation network, in order to serve two purposes. Task agnostic relation network models a basic metric function, while task-relevant relation network adapts to diverse tasks.

Task-Agnostic Relation Network The task-agnostic relation network uses a learned unified metric for all tasks, which is the same with the original relation network (Sung et al., 2018). With this unified metric, C task-agnostic relation scores $r_{z,j}^{agn}$ are generated for modeling the relation between one query input x_j and the class representation c_z,

$$r_{z,j}^{agn} = RN^{agn}(\mathbf{p}_{z,j}|\boldsymbol{\theta}^{agn}) \in \mathbb{R}, \quad z = 1, 2, ..., C \tag{3}$$

where RN^{agn} denotes task-agnostic relation network and $\boldsymbol{\theta}^{agn}$ are learnable parameters.

Task-Relevant Relation Network The task-relevant relation network is able to apply diverse metrics for diverse tasks armed with external knowledge. In detail, for each sample set, we retrieve a set of potentially relevant KB concepts $K(\mathcal{S})$, where each concept k_i is associated with KB embedding $\mathbf{e}_i \in \mathbb{R}^{d_2}$. (we will describe these processes in the following section). We element-wise average over these KB embeddings to form the knowledge representation of this sample set.

$$\mathbf{k}_{\mathcal{S}} = \frac{1}{|K(\mathcal{S})|} \sum_{k_i \in K(\mathcal{S})} \mathbf{e}_i \in \mathbb{R}^{d_2} \tag{4}$$

Then we use this knowledge representation to generate task-relevant relation network parameters,

$$\boldsymbol{\theta}^{rel} = \mathbf{M} \cdot \mathbf{k}_{\mathcal{S}} \in \mathbb{R}^{d_3} \tag{5}$$

where $\mathbf{M} \in \mathbb{R}^{d_3 \times d_2}$ are learnable parameters and d_3 denotes the number of parameters of the task-relevant relation network.

With these generated parameters, we use the task-relevant network to generate C task-relevant relation scores $r_{z,j}^{rel}$ for the relation between one query input x_j and the class representation c_z,

$$r_{z,j}^{rel} = RN^{rel}(\mathbf{p}_{z,j}|\boldsymbol{\theta}^{rel}) \in \mathbb{R}, \quad z = 1, 2, ..., C \tag{6}$$

where RN^{rel} denotes task-relevant relation network.

Finally, relation score is defined as:

$$r_{z,j} = Sigmoid(r_{z,j}^{agn} + r_{z,j}^{rel}) \tag{7}$$

where a sigmoid function is used to keep the score in a reasonable range. Following Sung et al. (2018), the network architecture of relation networks is two full-connected layers and mean square error (MSE) loss is used to train the model. The relation score is regressed to the ground truth: the matched pairs have similarity 1 and the mismatched pairs have similarity 0.

$$L = \sum_{z=1}^{C} \sum_{j=1}^{|\mathcal{Q}|} (r_{z,j} - \mathbf{1}(y_j == z)) \tag{8}$$

3.3 Knowledge Embedding and Retrieval

We use NELL (Carlson et al., 2010) as the KB, stored as (subject, relation, object) triples, where each triple is a fact indicating a specific relation between subject and object, e.g., (Intel, competes with, Nvidia).

Knowledge Embedding Since symbolic facts suffer from poor generalization and data sparsity, we resort to distributed representation of triples. In detail, given any tripe (s, r, o), vector embeddings of subject s, relation r and object o are learned jointly such that the validity of the triple can be measured in the real number space. We adopt the BILINEAR model (Yang et al., 2015) to measure the validity of triples:

$$f(s, r, o) = \mathbf{s}^T diag(\mathbf{r})\mathbf{o} \in \mathbb{R} \tag{9}$$

where $\mathbf{s}, \mathbf{r}, \mathbf{o} \in \mathbb{R}^{d_2}$ are the embeddings associated with s, r, o, respectively, and $diag(\mathbf{r})$ is a diagonal matrix with the main diagonal given by the relation embedding $\mathbf{r}$. To learn these vector embeddings, a margin-based ranking loss is designed, where triples in the KB are adopted to be positive and negative triplets are constructed by corrupting either subjects or objects.

Knowledge Retrieval To retrieve knowledge in KB, we first recognize entity mentions from a given passage, link the recognized entity mentions to subjects in KB by exactly string matching, and then collect the corresponding objects (concepts) as candidates. After this retrieval process, we obtain a set of potentially relevant KB concepts for each sample set, where each KB concept is associated with a KB embedding.

4 Experiment

4.1 Dataset

To make a fair comparison with previous methods, our model is evaluated on widely used ARSC (Blitzer et al., 2007) dataset. 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 form 69 tasks in total. Following previous works, we select 12 tasks from four domains (Books, DVDs, Electronics, and Kitchen) as testing set, with only five examples as support set for each label in the testing set. According to meta-training setting, we create 5-shot learning models on the dataset.

4.2 Implementation Details

In our experiments, we use hugginface's implementation[1] of BERT (base version) and initialize parameters of the BERT encoding layer with pre-trained models officially released by Google[2]. To represent knowledge in NELL (Carlson et al., 2010), BILINEAR model (Yang et al., 2015) is implemented with the open-source framework OpenKE (Han et al., 2018) to obtain the embedding of entities and relations. The size of embeddings of entities and relations is set to 100. To train our model, We use Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.00001.

4.3 Experiment Results

Baseline We compare our method to the following baselines: (1) **Match Network** (Vinyals et al., 2016) is a metric-based attention method for few-shot learning; (2) **Prototypical Network** (Snell et al., 2017) is a deep metric-based method using sample averages as class prototypes; (3) **Relation Network** (Sung et al., 2018) is a metric-based method that uses BERT as embedding module and uses two full-connected layers as metric function;(4) **MAML** (Finn et al., 2017) is an optimization-based method through learning to learn with gradients; (5) **ROBUSTTC-FSL** (Yu et al., 2018) is an approach that combines adaptive metric methods by clustering the tasks; (6) **Induction Network** (Geng et al., 2019) is a metric-based method by using dynamic routing to learn class-wise representations. (7) **P-MAML** (Zhang et al., 2019) is the current SOTA method that combine the MAML algorithm with BERT.

Model	Mean Acc
Matching Network	65.73
Prototypical Network	68.15
Relation Network	86.09
MAML	78.33
ROBUSTTC-FSL	83.12
Induction Network	85.63
P-MAML	86.65
Ours	**87.93**

Table 1: Comparison of mean accuracy (%) on ARSC dataset

Analysis Experiment results on ARSC are presented in Table 1. We observe that our method achieves the best results amongst all meta-learning models. Compared with P-MAML, our model not only achieve better performance, but also does exempt from requiring backpropagation to update parameters during testing. Both Induction Network and Relation Network use a single metric to measure the similarity. Compared with these methods, we attribute the improvements of our model to the fact that our model can adapt to diverse tasks with diverse metrics. Compared with ROBUSTTC-FSL, our model leverages knowledge to get implicit task clusters and is trained in an end-to-end manner, which can mitigate error propagation.

4.4 Ablation and Replacement Studies

To analyze the contributions and effects of external knowledge in our approach, we perform some ablation and replacement studies, which are shown in Table 2. **Ablation** means that we delete the task-relevant relation network and the model is reduced to the original relation network. We observe that ablation degrades performance. In order to exclude the factor of reduction in the number of parameters, we conduct a replacement experiment. **Replacement** means that we replace the task-relevant relation network with a task-agnostic relation network. We find out that increasing the number of parameters can slightly improve performance, but there is still a big gap between our model.

According to the results gained from ablation and replacement experiments, we conclude that the effectiveness of our model is credited with introducing external knowledge rather than increasing the number of model parameters.

[1] https://huggingface.co/pytorch_transformers/

[2] https://github.com/google-research/bert

Model	Mean Acc
Ours	87.93
Ablation	86.09
Replacement	86.40

Table 2: Ablation and replacement studies of our model on ARSC dataset

5 Conclusion

Inspired by human intelligence, we introduce external knowledge into few-shot learning. A parameter generator network is investigated to this end, which can use external knowledge to generate relation network parameters. With these parameters, the relation network can handle diverse tasks with diverse metric. Through various experiments, we demonstrate that our model outperforms the current SOTA few-shot text classification models.

References

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2019. Few-shot text classification with distributional signatures. *arXiv preprint arXiv:1908.06039*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3895–3904, Hong Kong, China. Association for Computational Linguistics.

David Ha, Andrew Dai, and Quoc V Le. 2016. Hypernetworks. *arXiv preprint arXiv:1609.09106*.

Xu Han, Shulin Cao, Lv Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. Openke: An open toolkit for knowledge embedding. In *Proceedings of EMNLP*.

Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. 2016. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, pages 667–675.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

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.

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.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *International Conference on Learning Representations*.

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. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1206–1215.

Ningyu Zhang, Zhanlin Sun, Shumin Deng, Jiaoyan Chen, and Huajun Chen. 2019. Improving few-shot text classification via pretrained language representations. *arXiv preprint arXiv:1908.08788v1*.