Dynamic Memory Induction Networks for Few-Shot Text Classification

Ruiying Geng[1], **Binhua Li**[1], **Yongbin Li**[1*], **Jian Sun**[1], **Xiaodan Zhu**[2]
[1] Alibaba Group, Beijing
[2] Ingenuity Labs Research Institute & ECE, Queen's University
{ruiying.gry,binhua.lbh,shuide.lyb,jian.sun}@alibaba-inc.com
zhu2048@gmail.com

Abstract

This paper proposes Dynamic Memory Induction Networks (DMIN) for few-shot text classification. The model utilizes dynamic routing to provide more flexibility to memory-based few-shot learning in order to better adapt the support sets, which is a critical capacity of few-shot classification models. Based on that, we further develop induction models with query information, aiming to enhance the generalization ability of meta-learning. The proposed model achieves new state-of-the-art results on the miniRCV1 and ODIC dataset, improving the best performance (accuracy) by 2∼4%. Detailed analysis is further performed to show the effectiveness of each component.

1 Introduction

Few-shot text classification, which requires models to perform classification with a limited number of training instances, is important for many applications but yet remains to be a challenging task. Early studies on few-shot learning (Salamon and Bello, 2017) employ data augmentation and regularization techniques to alleviate overfitting caused by data sparseness. More recent research leverages meta-learning (Finn et al., 2017; Zhang et al., 2018; Sun et al., 2019) to extract transferable knowledge among meta-tasks in meta episodes.

A key challenge for few-shot text classification is inducing class-level representation from support sets (Gao et al., 2019), in which key information is often lost when switching between meta-tasks. Recent solutions (Gidaris and Komodakis, 2018) leverage a memory component to maintain models' learning experience, e.g., by finding from a supervised stage the content that is similar to the unseen classes, leading to the state-of-the-art performance. However, the memory weights are static during inference and the capability of the model is still limited when adapted to new classes. Another prominent challenge is the instance-level diversity caused by various reasons (Gao et al., 2019; Geng et al., 2019), resulting in the difficulty of finding a fixed prototype for a class (Allen et al., 2019). Recent research has shown that models can benefit from query-aware methods (Gao et al., 2019).

In this paper we propose Dynamic Memory Induction Networks (DMIN) to further tackle the above challenges. DMIN utilizes dynamic routing (Sabour et al., 2017; Geng et al., 2019) to render more flexibility to memory-based few-shot learning (Gidaris and Komodakis, 2018) in order to better adapt the support sets, by leveraging the routing component's capacity in automatically adjusting the coupling coefficients during and after training. Based on that, we further develop induction models with query information to identify, among diverse instances in support sets, the sample vectors that are more relevant to the query. These two modules are jointly learned in DMIN.

The proposed model achieves new state-of-the-art results on the miniRCV1 and ODIC datasets, improving the best performance by 2∼4% accuracy. We perform detailed analysis to further show how the proposed network achieves the improvement.

2 Related Work

Few-shot learning has been studied in early work such as (Fe-Fei et al., 2003; Fei-Fei et al., 2006) and more recent work (Ba et al., 2016; Santoro et al., 2016; Munkhdalai and Yu, 2017; Ravi and Larochelle, 2016; Mishra et al., 2017; Finn et al., 2017; Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018; Allen et al., 2019). Researchers have also investigated few-shot learning in various NLP tasks (Dou et al., 2019; Wu et al., 2019; Gu et al., 2018; Chen et al., 2019; Obamuyide and Vlachos,

[*] Corresponding author.

2019; Hu et al., 2019), including text classification (Yu et al., 2018; Rios and Kavuluru, 2018; Xu et al., 2019; Geng et al., 2019; Gao et al., 2019; Ye and Ling, 2019).

Memory mechanism has shown to be very effective in many NLP tasks (Tang et al., 2016; Das et al., 2017; Madotto et al., 2018). In the few-shot learning scenario, researchers have applied memory networks to store the encoded contextual information in each meta episode (Santoro et al., 2016; Cai et al., 2018; Kaiser et al., 2017). Specifically Qi et al. (2018) and Gidaris and Komodakis (2018) build a two-stage training procedure and regard the supervisely learned class representation as a memory component.

3 Dynamic Memory Induction Network

3.1 Overall Architecture

An overview of our Dynamic Memory Induction Networks (DMIN) is shown in Figure 1, which is built on the two-stage few-shot framework Gidaris and Komodakis (2018). In the supervised learning stage (upper, green subfigure), a subset of classes in training data are selected as the base sets, consisting of C_{base} number of base classes, which is used to finetune a pretrained sentence encoder and to train a classifier.

In the meta-learning stage (bottom, orange subfigure), we construct an "episode" to compute gradients and update our model in each training iteration. For a C-way K-shot problem, a training episode is formed by randomly selecting C classes from the training set and choosing K examples within each selected class to act as the support set $S = \cup_{c=1}^{C}\{x_{c,s}, y_{c,s}\}_{s=1}^{K}$. A subset of the remaining examples serve as the query set $Q = \{x_q, y_q\}_{q=1}^{L}$. Training on such episodes is conducted by feeding the support set S to the model and updating its parameters to minimize the loss in the query set Q.

3.2 Pre-trained Encoder

We expect that developing few-shot text classifier should benefit from the recent advance on pre-trained models (Peters et al., 2018; Devlin et al., 2019; Radford et al.). Unlike recent work (Geng et al., 2019), we employ BERT-base (Devlin et al., 2019) for sentence encoding , which has been used in recent few-shot learning models (Bao et al., 2019; Soares et al., 2019). The model architecture of BERT (Devlin et al., 2019) is a multi-layer bidirectional Transformer encoder based on the original Transformer model (Vaswani et al., 2017). A special classification embedding ([CLS]) is inserted as the first token and a special token ([SEP]) is added as the final token. We use the d-dimensional hidden vector output from the $[CLS]$ as the representation e of a given text x: $e = E(x|\theta)$. The pre-trained BERT model provides a powerful context-dependent sentence representation and can be used for various target tasks, and it is suitable for the few-shot text classification task (Bao et al., 2019; Soares et al., 2019).

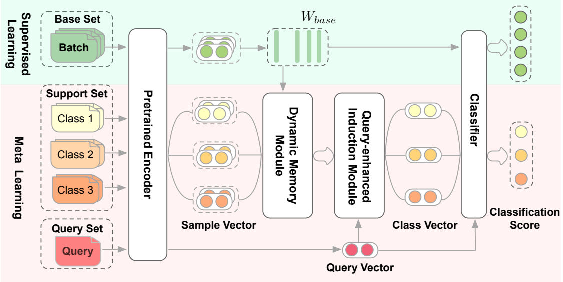


Figure 1: An overview of Dynamic Memory Induction Network with a 3-way 2-shot example.

We finetune the pre-trained BERT encoder in the supervised learning stage. For each input document x, the encoder $E(x|\theta)$ (with parameter θ) will output a vector e of d dimension. W_{base} is a matrix that maintains a class-level vector for each base class, serving as a base memory for meta-learning. Both $E(x|\theta)$ and W_{base} will be further tuned in the meta training procedure. We will show in our experiments that replacing previous models with pre-trained encoder outperforms the corresponding state-of-the-art models, and the proposed DMIN can further improve over that.

3.3 Dynamic Memory Module

At the meta-learning stage, to induce class-level representations from given support sets, we develop a dynamic memory module (DMM) based on knowledge learned from the supervised learning stage through the memory matrix W_{base}. Unlike static memory (Gidaris and Komodakis, 2018), DMM utilizes dynamic routing (Sabour et al., 2017) to render more flexibility to the memory learned from base classes to better adapt support sets. The routing component can automatically adjust the coupling coefficients during and after training, which inherently suits for the need of few-shot learning.

Specifically, the instances in the support sets are first encoded by the BERT into sample vectors $\{e_{c,s}\}_{s=1}^{K}$ and then fed to the following dynamic memory routing process.

Dynamic Memory Routing Process The algorithm of the dynamic memory routing process, denoted as DMR, is presented in Algorighm 1.

Given a memory matrix M (here W_{base}) and sample vector $q \in R^d$, the algorithm aims to adapt the sample vector based on memory M learned in the supervised learning stage.

$$q' = DMR(M, q). \tag{1}$$

First, for each entry $m_i \in M$, the standard matrix-transformation and squash operations in dynamic routing (Sabour et al., 2017) are applied on the inputs:

$$\hat{m}_{ij} = squash(W_j m_i + b_j), \tag{2}$$

$$\hat{q}_j = squash(W_j q + b_j), \tag{3}$$

where the transformation weights W_j and bias b_j are shared across the inputs to fit the few-shot learning scenario.

We then calculate the Pearson Correlation Coefficients (PCCs) (Hunt, 1986; Yang et al., 2019) between $\hat{m}_i$ and $\hat{q}_j$.

$$p_{ij} = tanh(PCCs(\hat{m}_{ij}, \hat{q}_j)), \tag{4}$$

$$PCCs = \frac{Cov(x_1, x_2)}{\sigma_{x_1} \sigma_{x_2}}. \tag{5}$$

where the general formula of PCCs is given above for vectors x_1 and x_2. Since PCCs values are in the range of [-1, 1], they can be used to encourage or penalize the routing parameters.

The routing iteration process can now adjust coupling coefficients, denoted as d_i, with regard to the input capsules m_i, q and higher level capsules v_j.

$$d_i = softmax\left(\alpha_i\right), \tag{6}$$

$$\alpha_{ij} = \alpha_{ij} + p_{ij} \hat{m}_i v_j. \tag{7}$$

Since our goal is to develop dynamic routing mechanism over memory for few-shot learning, we add the PCCs with the routing agreements in every routing iteration as shown in Eq. 8.

$$\hat{v}_j = \sum_{i=1}^{n} (d_{ij} + p_{ij}) m_{ij}, \tag{8}$$

$$v_j = squash(\hat{v}_j). \tag{9}$$

Algorithm 1 Dynamic Memory Routing Process

```
Require: $r$, $q$ and memory $M = \{m_1, m_2, ..., m_n\}$
Ensure: $\boldsymbol{v} = v_1, v_2, ..., v_l, q'$
for all $m_i, v_j$ do
    $\hat{m}_{ij} = squash(W_j m_i + b_j)$
    $\hat{q}_j = sqush(W_j q + b_j)$
    $\alpha_{ij} = 0$
    $p_{ij} = tanh(PCCs(\hat{m}_{ij}, \hat{q}_j))$
end for
for $r$ iterations do
    $d_i = softmax\left(\alpha_i\right)$
    $\hat{v}_j = \sum_{i=1}^{n} (d_{ij} + p_{ij}) \hat{m}_{ij}$
    $v_j = squash(\hat{v}_j)$
    for all $i, j$: $\alpha_{ij} = \alpha_{i,j} + p_{ij} \hat{m}_{ij} v_j$
    for all $j$: $\hat{q}_j = \frac{\hat{q}_j + v_j}{2}$
    for all $i, j$: $p_{ij} = tanh(PCCs(\hat{m}_{ij}, \hat{q}_j))$
end for
$q' = concat[\boldsymbol{v}]$
Return $q'$
```

We update the coupling coefficients α_{ij} and p_{ij} with Eq. 6 and Eq. 7, and finally output the adapted vector q' as in Algorithm 1.

The Dynamic Memory Module (DMM) aims to use DMR to adapt sample vectors $e_{c,s}$, guided by the memory W_{base}. That is, the resulting adapted sample vector is computed with $e'_{c,s} = DMR(W_{base}, e_{c,s})$.

3.4 Query-enhanced Induction Module

After the sample vectors $\left\{e'_{c,s}\right\}_{s=1,...,K}$ are adapted and query vectors $\{e_q\}_{q=1}^{L}$ are encoded by the pretrained encoder, we now incorporate queries to build a Query-guided Induction Module (QIM). The aim is to identify, among (adapted) sample vectors of support sets, the vectors that are more relevant to the query, in order to construct class-level vectors to better classify the query. Since dynamic routing can automatically adjusts the coupling coefficients to help enhance related (e.g., similar) queries and sample vectors, and penalizes unrelated ones, QIM reuses the DMR process by treating adapted sample vectors as memory of background knowledge about novel classes, and induces class-level representation from the adapted sample vectors that are more relevant/similar to the query under concern.

$$e_c = DMR(\left\{e'_{c,s}\right\}_{s=1,...,K}, e_q). \tag{10}$$

3.5 Similarity Classifier

In the final classification stage, we then feed the novel class vector e_c and query vector e_q to the classifier discussed above in the supervised training stage and get the classification score. The standard setting for neural network classifiers is, after having extracted the feature vector $e \in R^d$, to estimate the classification probability vector p by first computing the raw classification score s_k of each category $k \in [1, K^*]$ using the dot-product operator $s_k = e^T w_k^*$, and then applying softmax operator across all the K^* classification scores. However, this type of classifiers do not fit few-shot learning due to completely novel categories. In this work, we compute the raw classification scores using a cosine similarity operator:

$$s_k = \tau \cdot cos(e, w_k^*) = \tau \cdot \overline{e}^T \overline{w}_k^*, \tag{11}$$

where $\overline{e} = \frac{e}{\|e\|}$ and $\overline{w}_k^* = \frac{w_k^*}{\|w_k^*\|}$ are l_2-normalized vectors, and τ is a learnable scalar value. After the base classifier is trained, all the feature vectors that belong to the same class must be very closely matched with the single classification weight vector of that class. So the base classification weights $W_{base} = \{w_b\}_{b=1}^{C_{base}}$ trained in the 1st stage can be seen as the base classes' feature vectors.

In the few-shot classification scenario, we feed the query vector e_q and novel class vector e_c to the classifier and get the classification scores in a unified manner.

$$s_{q,c} = \tau \cdot cos(e_q, e_c) = \tau \cdot \overline{e}_q^T \overline{e}_c. \tag{12}$$

3.6 Objective Function

In the supervised learning stage, the training objective is to minimize the cross-entropy loss on C_{base} number of base classes given an input text x and its label $\boldsymbol{y}$:

$$L_1(x, \boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_{k=1}^{C_{base}} y_k log(\hat{y}_k), \tag{13}$$

where $\boldsymbol{y}$ is one-hot representation of the ground truth label, and $\hat{\boldsymbol{y}}$ is the predicted probabilities of base classes with $\hat{y}_k = softmax(s_k)$.

In the meta-training stage, for each meta episode, given the support set S and query set $Q = \{x_q, y_q\}_{q=1}^{L}$, the training objective is to minimize the cross-entropy loss on C novel classes.

$$L_2(S, Q) = -\frac{1}{C}\sum_{c=1}^{C}\frac{1}{L}\sum_{q=1}^{L} y_q log(\hat{y}_q), \tag{14}$$

where $\hat{y}_q = softmax(s_q)$ is the predicted probabilities of C novel classes in this meta episode, with $s_q = \{s_{q,c}\}_{c=1}^{C}$ from Equation 12. We feed the support set S to the model and update its parameters to minimize the loss in the query set Q in each meta episode.

4 Experiments

4.1 Dataset and Evaluation Metrics

We evaluate our model on the miniRCV1 (Jiang et al., 2018) and ODIC dataset (Geng et al., 2019). Following previous work (Snell et al., 2017; Geng et al., 2019), we use few-shot classification accuracy as the evaluation metric. We average over 100 and 300 randomly generated meta-episodes from the testing set in miniRCV1 and ODIC, respectively. We sample 10 test texts per class in each episode for evaluation in both the 1-shot and 5-shot scenarios.

4.2 Implementation Details

We use Google pre-trained BERT-Base model as our text encoder, and fine-tune the model in the training procedure. The number of base classes C_{base} on ODIC and miniRCV1 is set to be 100 and 20, respectively. The number of DMR interaction is 3. We build episode-based meta-training models with $C = [5, 10]$ and $K = [1, 5]$ for comparison. In addition to using K sample texts as the support set, the query set has 10 query texts for each of the C sampled classes in every training episode. For example, there are $10 \times 5 + 5 \times 5 = 75$ texts in one training episode for a 5-way 5-shot experiment.

4.3 Results

We compare DMIN with various baselines and state-of-the-art models: BERT (Devlin et al., 2019) finetune, ATAML (Jiang et al., 2018), Rel. Net (Sung et al., 2018), Ind. Net (Geng et al., 2019), HATT (Gao et al., 2019), and LwoF (Gidaris and Komodakis, 2018). Note that we re-implement them with the BERT sentence encoder for direct comparison.

Overall Performance The accuracy and standard deviations of the models are shown in Table 1 and 2. We can see that DMIN consistently outperform all existing models and achieve new state-of-the-art results on both datasets. The differences between DMIN and all the other models are statistically significant under the one-tailed paired t-test at the 95% significance level.

Model	5-way Acc.		10-way Acc.	
	1-shot	5-shot	1-shot	5-shot
BERT	30.79±0.68	63.31±0.73	23.48±0.53	61.18±0.82
ATAML	54.05±0.14	72.79±0.27	39.48±0.23	61.74±0.36
Rel. Net	59.19±0.12	78.35±0.27	44.69±0.19	67.49±0.23
Ind. Net	60.97±0.16	80.91±0.19	46.15±0.26	69.42±0.34
HATT	60.40±0.17	79.46±0.32	47.09±0.28	68.58±0.37
LwoF	63.35±0.26	78.83±0.38	48.61±0.21	69.57±0.35
DMIN	**65.72**±0.28	**82.39**±0.24	**49.54**±0.31	**72.52**±0.25

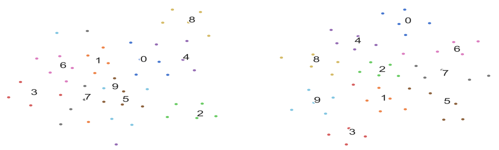
Table 1: Comparison of accuracy (%) on miniRCV1 with standard deviations.

Model	5-way Acc.		10-way Acc.	
	1-shot	5-shot	1-shot	5-shot
BERT	38.06±0.27	64.24±0.36	29.24±0.19	64.53±0.35
ATAML	79.60±0.42	88.53±0.57	63.52±0.34	77.36±0.57
Rel. Net	79.41±0.42	87.93±0.31	64.36±0.58	78.62±0.54
Ind. Net	81.28±0.26	89.67±0.28	64.53±0.38	80.48±0.25
HATT	81.57±0.47	89.27±0.58	65.75±0.61	81.53±0.56
LwoF	79.52±0.29	87.34±0.34	65.04±0.43	80.69±0.37
DMIN	**83.46**±0.36	**91.75**±0.23	**67.31**±0.25	**82.84**±0.38

Table 2: Comparison of accuracy(%) on ODIC with standard deviations.

Model	Iteration	1 Shot	5 Shot
w/o DMM	3	81.79	90.19
w/o QIM	3	82.37	90.57
DMIN	1	82.70	90.92
DMIN	2	82.95	91.18
DMIN	3	**83.46**	**91.75**

Table 3: Ablation study of accuracy (%) on ODIC in a 5-way setup.



(a) Before DMM (b) After DMM

Figure 2: Effect of the Dynamic Memory Module in a 10-way 5-shot setup.

Note that LwoF builds a two-stage training procedure with a memory module learnt from the supervised learning and used in the meta-learning stage, but the memory mechanism is static after training, while DMIN uses dynamic memory routing to automatically adjust the coupling coefficients after training to generalize to novel classes, and outperform LwoF significantly. Note also that the performance of some of the baseline models (Rel. Net and Ind. Net) reported in Table 1 and 2 is higher than that in Geng et al. (2019) since we used BERT to replace BiLSTM-based encoders. The BERT encoder improves the baseline models by a powerful context meaning representation ability, and our model can further outperform these models with a dynamic memory routing method. Even with these stronger baselines, the proposed DMIN consistently outperforms them on both dataset.

Ablation Study We analyze the effect of different components of DMIN on ODIC in Table 3. Specifically, we remove DMM and QIM, and vary the number of DMR iterations. We see that the best performance is achieved with 3 iterations. The results show the effectiveness of both the dynamic memory module and the induction module with query information.

4.4 Further Analysis

Figure 2 is the t-SNE visualization (Maaten and Hinton, 2008) for support sample vectors before and after DMM under a 10-way 5-shot setup on ODIC. We randomly select a support set with 50 texts (10 texts per class) from the ODIC testing set, and obtain the sample vectors before and after DMM, i.e., $\{e_{c,s}\}_{c=1,...5,s=1...10}$ and $\{e'_{c,s}\}_{c=1,...5,s=1...10}$. We can see that the support vectors produced by the DMM are better separated, demonstrating the effectiveness of DMM in leveraging the supervised learning experience to encode semantic relationships between lower level instance features and higher level class features for few-shot text classification.

5 Conclusion

We propose Dynamic Memory Induction Networks (DMIN) for few-shot text classification, which builds on external working memory with dynamic routing, leveraging the latter to track previous learning experience and the former to adapt and generalize better to support sets and hence to unseen classes. The model achieves new state-of-the-art results on the miniRCV1 and ODIC datasets. Since dynamic memory can be a learning mechanism more general than what we have used here for few-shot learning, we will investigate this type of models in other learning problems.

Acknowledgments

The authors would like to thank the organizers of ACL-2020 and the reviewers for their helpful suggestions. The research of the last author is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Kelsey Allen, Evan Shelhamer, Hanul Shin, and Joshua Tenenbaum. 2019. Infinite mixture prototypes for few-shot learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 232–241, Long Beach, California, USA. PMLR.

Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, and Catalin Ionescu. 2016. Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems*, pages 4331–4339.

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

Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. 2018. Memory matching networks for one-shot image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4080–4088.

Mingyang Chen, Wen Zhang, Wei Zhang, Qiang Chen, and Huajun Chen. 2019. Meta relational learning for few-shot link prediction in knowledge graphs. 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 4208–4217, Hong Kong, China. Association for Computational Linguistics.

Rajarshi Das, Manzil Zaheer, Siva Reddy, and Andrew McCallum. 2017. Question answering on knowledge bases and text using universal schema and memory networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–365, Vancouver, Canada. Association for Computational Linguistics.

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.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. 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 1192–1197.

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

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611.

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.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence,(AAAI-19), New York, USA*.

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.

Spyros Gidaris and Nikos Komodakis. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375.

Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631.

Ziniu Hu, Ting Chen, Kai-Wei Chang, and Yizhou Sun. 2019. Few-shot representation learning for out-of-vocabulary words. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4102–4112, Florence, Italy. Association for Computational Linguistics.

Ronald J Hunt. 1986. Percent agreement, pearson's correlation, and kappa as measures of inter-examiner reliability. *Journal of Dental Research*, 65(2):128–130.

Xiang Jiang, Mohammad Havaei, Gabriel Chartrand, Hassan Chouaib, Thomas Vincent, Andrew Jesson, Nicolas Chapados, and Stan Matwin. 2018. Attentive task-agnostic meta-learning for few-shot text classification.

Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. 2017. Learning to remember rare events. *arXiv preprint arXiv:1703.03129*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2017. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*.

Tsendsuren Munkhdalai and Hong Yu. 2017. Meta networks. *arXiv preprint arXiv:1703.00837*.

Abiola Obamuyide and Andreas Vlachos. 2019. Model-agnostic meta-learning for relation classification with limited supervision. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5873–5879, Florence, Italy. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Hang Qi, Matthew Brown, and David G Lowe. 2018. Low-shot learning with imprinted weights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5822–5830.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

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

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866.

Justin Salamon and Juan Pablo Bello. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850.

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

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412.

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.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224.

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, Tim Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638.

Jiawei Wu, Wenhan Xiong, and William Yang Wang. 2019. Learning to learn and predict: A meta-learning approach for multi-label 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 4345–4355.

Hu Xu, Bing Liu, Lei Shu, and P Yu. 2019. Open-world learning and application to product classification. In *The World Wide Web Conference*, pages 3413–3419. ACM.

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 2019. Enhancing context modeling with a query-guided capsule network for document-level translation. 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 1527–1537.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational*

Linguistics, pages 2872–2881, Florence, Italy. Association for Computational Linguistics.

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.

Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. 2018. Metagan: An adversarial approach to few-shot learning. In *Advances in Neural Information Processing Systems*, pages 2365–2374.