Meta-Information Guided Meta-Learning for Few-Shot Relation Classification

Bowen Dong[1*], **Yuan Yao**[1*], **Ruobing Xie**[2], **Tianyu Gao**[1], **Xu Han**[1], **Zhiyuan Liu**[1†], **Fen Lin**[2], **Leyu Lin**[2], **Maosong Sun**[1]
[1]Department of Computer Science and Technology
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology, China
[2]WeChat Search Application Department, Tencent, China
{dongbw18,yuan-yao18}@mails.tsinghua.edu.cn

Abstract

Few-shot classification requires classifiers to adapt to new classes with only a few training instances. State-of-the-art meta-learning approaches such as MAML learn how to initialize and fast adapt parameters from limited instances, which have shown promising results in few-shot classification. However, existing meta-learning models solely rely on implicit instance-based statistics, and thus suffer from instance unreliability and weak interpretability. To solve this problem, we propose a novel meta-information guided meta-learning (MIML) framework, where semantic concepts of classes provide strong guidance for meta-learning in both initialization and adaptation. In effect, our model can establish connections between instance-based information and semantic-based information, which enables more effective initialization and faster adaptation. Comprehensive experimental results on few-shot relation classification demonstrate the effectiveness of the proposed framework. Notably, MIML achieves comparable or superior performance to humans with only one shot on FewRel evaluation. The source code and experiment details of this paper can be obtained from https://github.com/thunlp/MIML.

1 Introduction

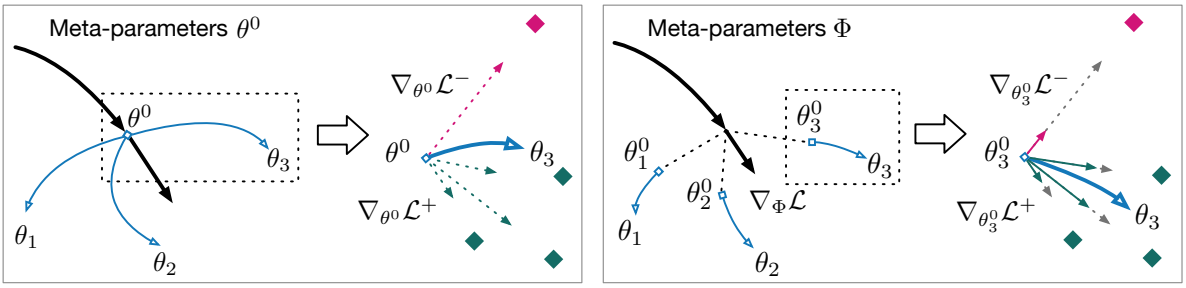
Conventional machine learning algorithms, especially neural methods, require an adequate amount of data to learn model parameters. To alleviate the heavy reliance on annotated data, few-shot learning, which aims at adapting to new tasks with only a few training examples, has drawn more and more attention. Few-shot classification is a typical few-shot learning task, which samples several new classes with a handful of training examples (i.e., support instances) and query instances, and requires models to classify these queries into given classes (Lake et al., 2011; Vinyals et al., 2016).

To grasp the patterns of new classes with limited examples, meta-learning was proposed. Inspired by human behaviors, meta-learning models focus on *learning to learn*: they learn how to better initialize parameters and fast adapt classification models from given instances. For example, MAML (Finn et al., 2017) finds the best initialization point of parameters, where it can take minimal efforts to reach the optimal points for each class. To this end, MAML adapts towards each class by gradient steps using support instances, and uses the loss of the adapted model on the query instances to optimize the initialization parameters.

However, meta-learning still has three challenges: (1) Most meta-learning methods learn how to learn (i.e., how to initialize and adapt) solely relying on instance statistics, which inevitably suffer from data sparsity and noise in low-resource scenarios, especially in text domain. (2) The approach of learning to learn, like the learning process itself, is a black-box and thus lacks interpretability. (3) Most conventional meta-learning methods are designed for few-shot classification, and cannot well handle zero-shot scenarios, where no support instances are available. In contrast, humans usually learn novel concepts with high-level descriptive definitions, instead of solely learning from several unsystematic instances. For

[*] indicates equal contribution
[†] Corresponding author: Z.Liu (liuzy@tsinghua.edu.cn)

Proceedings of the 28th International Conference on Computational Linguistics, pages 1594–1605
Barcelona, Spain (Online), December 8-13, 2020



(a) MAML (b) MIML

Figure 1: Diagram of meta-learning models. (a) MAML learns a class-agnostic representation θ^0 that can fast adapt to new classes. (b) MIML learns meta-parameters Φ to fast initialize class-aware parameter θ_i^0, and to quickly adapt to new classes using informative instances, where both phases are guided by meta-information. **Informative instances** and **noisy instances** are marked accordingly.

example, when learning a new relation *art director*, humans usually first get a rough estimation of the concept by its name and definition, and then reach a more precise understanding by concrete instances.

Inspired by the learning process of humans, we propose a novel **Meta-Information guided Meta-Learning (MIML)** framework, as shown in Figure 1. The meta-information derives from the semantic concepts of classes, and could provide strong guidance for both parameter initialization and fast adaptation in meta-learning. Specifically, MIML integrates meta-information in two essential components, namely the meta-information guided fast initialization and fast adaptation. (1) In **meta-information guided fast initialization**, instead of using a static class-agnostic initialization point for all classes as in MAML, MIML uses meta-information to estimate dynamic class-aware initialization parameters for each class. This alleviates the reliance on support instances to reach optimal adapted parameters. (2) In **meta-information guided fast adaptation**, MIML adapts the class-aware initialization parameters with gradient steps according to the support instances, where informative support instances are selected to contribute more to the adaptation gradients with a novel meta-information based attention mechanism. By integrating high-level meta-information and concrete instances, MIML achieves superior performance on low-resource tasks. Moreover, MIML also provides better interpretability in meta-learning process.

Note that we are not the first attempt to use meta-information for low-resource classification tasks: In zero-shot learning, where there are no training examples for new classes at all, class names are used to produce semantic representations for classification (Socher et al., 2013; Frome et al., 2013; Norouzi et al., 2014). In few-shot scenarios, however, supporting examples can bring more direct supervision. In this paper, we argue that both signals are crucial to the learning process, and combining them could achieve the best results.

In experiments, the significant improvements on few-shot relation classification tasks demonstrate the effectiveness and robustness of MIML in low-resource relation classification. We show the advantage of MIML in handling noisy instances, and its potential in zero-shot classification. We also conduct comprehensive ablation study and visualization to better understand our model. In summary, our main contributions are twofold: (1) We propose a principled meta-information guided meta-learning framework for few-shot classification. To the best of our knowledge, we are the first to introduce meta-information to meta-learning for few-shot relation classification. (2) We conduct comprehensive experiments to demonstrate the effectiveness of MIML. Notably, MIML achieves human-level performance with only one shot on FewRel evaluation. We also show the robustness and interpretability of MIML, as well as its potential in zero-shot classification through experiments.

2 Preliminary

In few-shot classification, we aim to learn a model that can handle the classification task with only a few available training instances. Specifically, given a set of classes $\mathcal{C}$ from the class distribution $p(\mathcal{C})$, the

Algorithm 1 Meta-Information Guided Meta-Learning

Require: $p(\mathcal{C})$: distribution over classes
Require: β: meta learning rate
randomly initialize:
$\Phi = \{\phi_e, \phi_n, \phi_a\}$: meta-parameters
while not done **do**
  Sample batch of classes $\mathcal{C}_i \sim p(\mathcal{C})$
  Sample support instance set $\mathcal{S}$ and query instance set $\mathcal{Q}$
  for all $\mathcal{C}_i$ **do**
    Fast initialize parameters of $\mathcal{C}_i$: $\theta_i^0 = \Psi(\boldsymbol{c}_i; \phi_n)$
  for $t = 1, \ldots, T$ **do**
    Compute gradients and learning rates for fast adaptation using support instance set $\mathcal{S}$
    Compute adapted parameters with gradient descent: $\theta^{t+1} = \theta^t - \sum_{i,j} \alpha_{i,j} \nabla_{\theta^t} \mathcal{L}(f_{\theta^t, \{\phi_e, \phi_n\}}, x_j, y_j)$
  Meta-optimize using query instance set $\mathcal{Q}$: $\Phi = \Phi - \beta \nabla_{\Phi} \mathcal{L}(f_{\theta^T, \Phi}, x_j, y_j)$

model is required to first learn classifiers on the support set S, and then handle the classification task on the query set $\mathcal{Q}$, where $\mathcal{S}$ and $\mathcal{Q}$ consist of instances $\{x_j, y_j\}_{j=1}^m$ from same classes, and x_j is an instance of class y_j. Few-shot classification is usually formalized in an N way K shot setting, where $\mathcal{C}$ contains N different classes, and $\mathcal{S}$ contains K instances for each of the N classes.

Our work is inspired by MAML (Finn et al., 2017), an effective meta-learning approach to the few-shot classification problem. MAML contains two key phases: initialization and fast adaptation. Initialization aims to learn a globally shared initialization point of parameters for different classes, such that a few gradient steps of fast adaptation on the initialization parameters can produce good results on new classes. We refer readers to the paper (Finn et al., 2017) for more details about MAML.

3 Methodology

In this section, we introduce our meta-information guided meta-learning (MIML) framework. Despite the effectiveness of MAML, we observe that two assumptions in MAML limit the model capacity:

(1) In initialization, MAML assumes that the parameters of different classes can be derived from single initialization parameters from a few gradient steps. However, single initialization parameters cannot well capture the shared knowledge in different classes, especially when the number of classes is large, making it difficult to adapt the initialized parameters with a few gradient steps to reach reasonable performance.

(2) In fast adaptation, MAML assumes that different instances in support set are equally important, and thus share the same learning rate for parameter adaptation. However, instances in text are usually diverse and noisy in practice, and noisy instances can dominate the model parameters in fast adaptation to produce inferior results (Koh and Liang, 2017).

To address the aforementioned problems, we propose MIML to integrate meta-information into meta-learning, and provide strong guidance in both initialization and adaptation phases. The intuition behind MIML is that human learn new concepts from both high-level meta-information and concrete instances. Specifically, MIML consists of four components:

Instance Encoder. Given a sentence and the corresponding entity pair, we employ deep neural networks (with meta-parameters ϕ_e) to construct the representation of the relation between the entity pair.

Meta-Information Guided Fast Initialization. In fast initialization phase, MIML dynamically initializes the parameters for each class based on meta-information (with meta-parameters ϕ_n), which can be viewed as a rough but flexible estimation of class parameters from high-level semantics.

Meta-Information Guided Fast Adaptation. In fast adaptation phase, MIML adapts the initialized parameters according to the performance on the support set, and selects informative support instances to

contribute more to the adaptation gradients (with meta-parameters ϕ_a), which can be viewed as accurate fine-tuning from concrete instances.

Meta-Optimization. In meta-optimization phase, the meta-parameters $\Phi = \{\phi_e, \phi_n, \phi_a\}$ are optimized based on the performance of the adapted model on the query set. The framework is shown in Algorithm 1.

3.1 Instance Encoder

Given a sentence and the corresponding target entity pair (i.e., head entity and tail entity), we employ BERT model (Devlin et al., 2019) to encode the instance into contextualized representations, due to its effectiveness on a broad variety of NLP tasks. Specifically, sentences are first tokenized into word pieces (Wu et al., 2016). Inspired by Soares et al. (2019), to mark the positions of entities, we adopt four special tokens as entity markers, and insert them to the start and end of each entity. We select the representations of the start tokens of the head entity and tail entity on the top layer, and concatenate them to obtain the instance representation. The instance encoder can be formulated as follows:

$$\mathbf{x}_j = g(x_j, h, t; \phi_e), \tag{1}$$

where x_j is the sentence, h and t are head and tail entities respectively. $g(\cdot)$ is the encoder, ϕ_e is the parameters of the encoder, and $\mathbf{x}_j \in \mathbb{R}^{d_s}$ is the instance representation.

3.2 Meta-Information Guided Fast Initialization

Given a set of classes $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$ sampled from class distribution $p(\mathcal{C})$, MAML learns class-agnostic initialization that can adapt to new classes via a few gradient steps. In comparison, we utilize meta-information for class-aware initialization in a generative manner via a *meta-initializer module*.

The meta-initializer module captures meta-knowledge shared in different classes, and generates the class-aware parameters via semantic knowledge in meta-information. We initialize the parameters of each class with meta-information derived from its semantic concepts. In this work, without losing generality we utilize class names as our meta-information, i.e., relation names such as *founder of* and *birth place*. Note that it is also convenient to generate class parameters with other meta-information such as textual descriptions and hierarchical ontology. Specifically, given the name of a class $\mathcal{C}_i$, we obtain the meta-information representation $\boldsymbol{c}_i \in \mathbb{R}^{d_w}$ by the average of the word embeddings of the name. Then the parameter of the class is initialized via the meta-initializer module as follows:

$$\theta_i^0 = \Psi(\boldsymbol{c}_i; \phi_n), \tag{2}$$

where $\theta_i^0 \in \mathbb{R}^{d_s}$ is the class-aware initialization parameters for class $\mathcal{C}_i$, $\Psi(\cdot)$ is the meta-initializer, ϕ_n is the corresponding meta-parameters. In our experiments, $\Psi(\cdot)$ is implemented via a fully connected layer. Intuitively, the meta-initializer mimics the learning process of human, where we usually first get a rough but flexible estimation of a new concept based on its high-level semantics. The initialized parameter θ_i^0 can be used to measure the classification score of an instance:

$$s_{i,j} = \theta_i^{0\top} \mathbf{x}_j, \tag{3}$$

where $s_{i,j}$ is the score of x_j being an instance of $\mathcal{C}_i$. The probability $p(y = \mathcal{C}_i | x_j)$ is obtained by normalizing the score $s_{i,j}$ with a softmax layer over all classes $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$. The model after fast initialization can be denoted as $f_{\theta^0, \{\phi_e, \phi_n\}}$, where $\theta^0 = \{\theta_1^0, \theta_2^0, \ldots, \theta_N^0\}$ denotes initialized parameters.

3.3 Meta-Information Guided Fast Adaptation

In fast adaptation, like human learners, MIML fine-tunes the estimation of a new concept by concrete instances. Specifically, the initialized parameters θ^0 are adapted via gradient descent steps, according to the classification performance of instances on the support set $\mathcal{S}$. The adaptation iterates dynamically for T steps. At each time step t, the parameters θ^t are adapted as follows:

$$\theta^{t+1} = \theta^t - \sum_{i,j} \alpha_{i,j} \nabla_{\theta^t} \mathcal{L}(f_{\theta^t, \{\phi_e, \phi_n\}}, x_j, y_j), \tag{4}$$

where $\mathcal{L}$ denotes cross-entropy loss of a support instance (x_j, y_j) computed by the model $f_{\theta^t,\{\phi_e,\phi_n\}}$, and $\alpha_{i,j}$ is the learning rate of θ_i on the support instance (x_j, y_j). The parameters after T steps of adaptation are denoted as θ^T.

With a static learning rate for all instances, noisy instances can dominate the model parameters in fast adaptation (Koh and Liang, 2017), which leads to inferior performance. To select informative instances for fast adaptation in MIML, instead of using a static learning rate for all instances, the learning rate of each instance is dynamically determined by a selective attention mechanism as follows:

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_j \exp(e_{i,j})}, \tag{5}$$

where $e_{i,j}$ is the score of instance x_j for class $\mathcal{C}_i$. Intuitively, $e_{i,j}$ should be large if x_j is an informative instance of class $\mathcal{C}_i$, and thus x_j should contribute more to the adaptation of θ_i, i.e., the learning rate should be larger. The score is obtained by:

$$e_{i,j} = \mathbf{q}_i^\top \mathbf{x}_j, \tag{6}$$

where $\mathbf{q}_i \in \mathbb{R}^{d_s}$ is the query vector for class $\mathcal{C}_i$. We note that similar to classifier parameters, learning query vectors is also faced with data sparsity in few-shot classification, since only a few training instances are available for each class. Thus we estimate the query vector from meta-information via a *meta-querier module* as follows:

$$\mathbf{q}_i = \Psi(\boldsymbol{c}_i; \phi_a), \tag{7}$$

where $\Psi(\cdot)$ is implemented via a fully connected layer with meta-parameters ϕ_a.

In our experiments, we observe that the estimation of class-aware parameters (i.e., initialization parameter θ_i^0 and query vector $\mathbf{q}_i$) are prone to over-fitting, due to the limited number of classes, e.g., less than 100 in most datasets. This limits the diversity of inputs to the meta-initializer and meta-querier, which leads to complex hyper-planes in meta-information space, and hurts the generalization ability.

We address the problem by (1) regularizing class-aware parameters by L2 normalization, and (2) penalizing sharp changes in the meta-information space via *virtual adversarial training* (Miyato et al., 2017). Specifically, we normalize class-aware parameters to be of unit length in L2 norm. For virtual adversarial training, we add worst-case perturbations on the meta-information $\boldsymbol{c}_i$, such that the classification results on the query set reach the maximum changes. We measure the changes of classification results by Kullback-Leibler divergence, and penalize the changes to encourage a smooth meta-information space.

3.4 Meta-Optimization

After fast adaptation on support instances, the meta-parameters $\Phi = \{\phi_e, \phi_n, \phi_a\}$ are optimized according to the performance of the *adapted* model on the query set $\mathcal{Q}$ as follows:

$$\Phi = \Phi - \beta \nabla_\Phi \mathcal{L}(f_{\theta^T,\Phi}, x_j, y_j), \tag{8}$$

where β is the learning rate for meta-parameters. In this way, MIML learns meta-parameters that can effectively customize initialization parameters for each class, and select informative support instances for fast adaptation, so as to produce good classification results on the query set.

3.5 Implementation Details

All hyper-parameters are selected by grid-search on the development set. The class distribution $p(\mathcal{C})$ is implemented by uniform distribution. We adopt Adam (Kingma and Ba, 2015) to optimize meta-parameters. The meta learning rate β is 1 for meta-initializer and meta-querier, and $5e$-5 for instance encoder. We employ 50 dimensional GloVe (Pennington et al., 2014) for word embeddings and $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019) implemented by Wolf et al. (2019) as the instance encoder. The hidden state dimensions d_s and d_w are $1,536$ and 50 respectively. The number of adaptation steps T is 150. In virtual adversarial training, we first randomly generate a perturbation vector δ_1 for meta-information

Encoder	Model	5-way-1-shot	5-way-5-shot	10-way-1-shot	10-way-5-shot
CNN	Meta Network*	64.46 ± 0.54	80.57 ± 0.48	53.96 ± 0.56	69.23 ± 0.52
	GNN*	66.23 ± 0.75	81.28 ± 0.62	46.27 ± 0.80	64.02 ± 0.77
	SNAIL*	67.29 ± 0.26	79.40 ± 0.22	53.28 ± 0.27	68.33 ± 0.25
	Proto Network*	74.52 ± 0.07	88.40 ± 0.06	62.38 ± 0.06	80.45 ± 0.08
	MLMAN*	82.98 ± 0.20	92.66 ± 0.09	73.59 ± 0.26	87.29 ± 0.15
BERT	BERT-PAIR ♠	88.32 ± 0.64	93.22 ± 0.13	80.63 ± 0.17	87.02 ± 0.12
	MAML	87.45 ± 0.11	94.39 ± 0.13	78.91 ± 0.14	89.14 ± 0.23
	Proto Network	86.50 ± 0.14	95.01 ± 0.15	82.86 ± 0.15	91.30 ± 0.11
	MIML	$\mathbf{92.55 \pm 0.12}$	$\mathbf{96.03 \pm 0.17}$	$\mathbf{87.47 \pm 0.21}$	$\mathbf{93.22 \pm 0.22}$
-	Human*	92.22	-	85.88	-

Table 1: Main results. Accuracies (%) on few-shot relation classification on FewRel test set. Results with * and ♠ are from FewRel leaderboard and Gao et al. (2019b) respectively.

representation $\boldsymbol{c}_i$. Then the perturbation vector δ_1 is scaled such that its L2 norm is $1e$-3. We add δ_1 to $\boldsymbol{c}_i$, and compute the worst-case perturbation δ_2 based on the gradient. Finally δ_2 is scaled to $1e$-3 in L2 norm, and added to $\boldsymbol{c}_i$ to obtain the perturbed representation.

4 Experiments

In this section, we empirically evaluate MIML on few-shot relation classification. To evaluate the robustness of MIML, we conduct experiments in the presence of noisy instances. We also show the potential of MIML in zero-shot classification. Ablation study and visualization are conducted to better understand the inner mechanism of MIML.

4.1 Experiment Settings

We first introduce the experiment settings, including datasets, evaluation protocol and baselines.

Dataset. We evaluate MIML on FewRel (Han et al., 2018), a widely-used few-shot relation classification dataset. FewRel contains $70,000$ labeled sentences in 100 relations (i.e., each relation has 700 sentences). The relation annotations are first generated under distant supervision assumption (Mintz et al., 2009) by aligning Wikipedia and Wikidata (Vrandečić and Krötzsch, 2014), and then labeled by human annotators. The training set contains $44,800$ sentences in 64 relations, the valid set has $11,200$ sentences in 16 relations, and the test set has the rest $14,000$ sentences in 20 relations.

Evaluation Protocol. Following the same settings in Han et al. (2018), we consider four types of few-shot settings in evaluation, namely 5-way-1-shot, 5-way-5-shot, 10-way-1-shot and 10-way-5-shot. The N-way-K-shot setting indicates that each evaluation batch has N classes that do not appear in training set and each class has K support instances. Smaller shots or more ways imply more challenging settings. We adopt the classification accuracy of query instances as the evaluation metric.

Baseline. We compare **MIML** with strong baseline methods for few-shot classification. **Meta Network** (Munkhdalai and Yu, 2017) and **SNAIL** (Mishra et al., 2018) are classical meta-learning models that learn to fast adapt to new classes. **GNN** (Garcia and Estrach, 2018) performs message passing over instance graphs. **Prototypical Network** (Snell et al., 2017) constructs the prototypes of new classes by averaging their instance representations. **MLMAN** (Ye and Ling, 2019) obtains prototypes by a multi-level matching and aggregation network. We directly report the accuracies of these models (with CNN encoders), and **human performance** from the FewRel leaderboard.[1] We also compare with strong baselines with BERT (Devlin et al., 2019) encoders. **BERT-PAIR** (Gao et al., 2019b) measures the similarity of an instance pair using BERT. In addition, we also implement the enhanced **Prototypical Network** and **MAML** (Finn et al., 2017) with BERT encoder for fair comparisons.

[1] https://www.zhuhao.me/fewrel/

Model	Noise Rate	5-way-5-shot	10-way-5-shot	Noise Rate	5-way-5-shot	10-way-5-shot
MAML	0%	92.59 ± 0.08	85.79 ± 0.15	10%	90.81 ± 0.12	83.31 ± 0.13
Proto Network		92.62 ± 0.11	87.12 ± 0.12		91.54 ± 0.08	85.40 ± 0.18
Proto HATT		93.43 ± 0.09	89.37 ± 0.17		92.40 ± 0.13	88.19 ± 0.22
MIML		$\mathbf{95.60 \pm 0.09}$	$\mathbf{91.60 \pm 0.21}$		$\mathbf{94.82 \pm 0.08}$	$\mathbf{89.55 \pm 0.25}$
MAML	20%	88.40 ± 0.10	80.77 ± 0.13	30%	86.18 ± 0.20	78.30 ± 0.11
Proto Network		91.04 ± 0.08	83.18 ± 0.17		87.84 ± 0.12	80.28 ± 0.19
Proto HATT		91.27 ± 0.15	85.94 ± 0.29		89.62 ± 0.19	83.14 ± 0.24
MIML		$\mathbf{93.19 \pm 0.10}$	$\mathbf{87.70 \pm 0.23}$		$\mathbf{92.04 \pm 0.18}$	$\mathbf{86.19 \pm 0.27}$

Table 2: Accuracies (%) on few-shot relation classification with noise on FewRel development set.

4.2 Main Results

We report the main results in Table 1, from which we have the following observations:

(1) MIML consistently outperforms all baseline methods in four settings. Notably, MIML achieves comparable or superior performance to humans with only one shot. To the best of our knowledge, we are the first to achieve human-level performance with only one shot on FewRel without tailored pre-training for RE. The results demonstrate that MIML can effectively leverage high-level meta-information to provide strong guidance for meta-learning.

(2) The advantages of MIML are more significant in more challenging settings, i.e., with fewer shots or more ways. For example, MIML achieves 8.5 absolute accuracy improvement compared to MAML in 10-way-1-shot setting. This is because that, in comparison to static class-agnostic initialization in MAML, meta-information guided fast initialization in MIML can produce more flexible class-aware initialization, which alleviates heavy reliance on support instances. In Section 4.3, we further show the advantage of MIML when multiple shots are available in the presence of noise.

4.3 Robustness to Noisy Instances

Instances in real-world few-shot text classification tasks can be diverse and noisy, especially when multiple support instances are available. Previous works have shown that noisy instances can dominate the model parameters (Koh and Liang, 2017), especially for meta-learning methods where adaptation is based on gradients from instances, e.g., MAML, due to the substantially higher loss of noisy instances. To demonstrate the robustness of MIML in the presence of noise, we randomly corrupt 0%, 10%, 20%, 30% support instances, by replacing them with noisy instances randomly sampled from different relations in FewRel. In addition to Prototypical Network and MAML, we also compare MIML with hybrid attention-based prototypical networks (Gao et al., 2019a) (Proto-HATT), which uses hybrid attention to denoise for Prototypical Network. The results are shown in Table 2, from which we observe that:

(1) The performance of MAML degrades significantly when the noise rate increases, since its fast adaptation process can be dominated by noisy instances. Prototypical Network constructs the prototype with the average of all instances, and shows smaller drops in performance. The results show the disadvantage of gradient-based meta-learning models in dealing with noisy instances.

(2) MIML consistently outperforms baseline methods in different noise rates. Specifically, MIML exhibits smaller drops in performance as compared to MAML and Prototypical Network. The results show that meta-information guided fast adaptation can effectively select informative instances, which helps MIML overcome the inherent disadvantage of gradient-based meta-learning models, and achieve more robust fast adaptation in the presence of noise.

4.4 Zero-Shot Classification

In this section, we show the potential of MIML in zero-shot classification. Specifically, we remove the support instances in evaluation phase in 5-way and 10-way setting, and ask the model to classify query instances with class-aware initialization parameters. We compare MIML with strong zero-shot classification baselines. DeViSE (Frome et al., 2013) utilizes word embeddings of class names to classify

Setting	Random	DeViSE	SK4	MIML
5-way-0-shot	20.00	55.90 ± 0.09	$\mathbf{79.68 \pm 0.12}$	79.54 ± 0.06
10-way-0-shot	10.00	42.29 ± 0.08	$\mathbf{66.17 \pm 0.11}$	61.14 ± 0.10

Table 3: Experimental results of zero-shot classification on FewRel development set.

instances from unseen classes, and we implement the DeViSE model with BERT encoder. SK4 (Zhang et al., 2019) incorporates rich semantic knowledge of classes, including word embeddings, class descriptions, class hierarchy, and commonsense knowledge graphs. We report the results in Table 3, from which we observe that:

Compared to models tailored for zero-shot classification problem, MIML achieves reasonable performance. This is because that the class-aware fast initialization parameters in MIML are guided by meta-information, and thus can potentially be used to severe as classifiers without further adaptation using support instances. In summary, the results show that MIML can effectively integrate high-level meta-information and concrete instances for low-resource classification tasks, including few-shot and zero-shot classification tasks.

4.5 Ablation Study

To investigate the contribution of different components in MIML, we conduct ablation study in 10-way-5-shot setting, by removing each component, including meta-information guided fast initialization (MI) and adaptation (MA), class-aware parameter normalization (NM) and virtual adversarial training (VAT). Table 5 shows the results of ablation study.

We can observe that all components contribute to the performance of MIML. The performance drops most significantly when removing class-aware parameter normalization. This is because that estimating high-dimensional parameters in a generative manner is prone to over-fitting and also faced with high variance, which can be effectively regularized by class-aware parameter normalization. Meta-information guided fast initialization also contributes significantly to the performance, indicating the importance of class-aware initialization to meta-learning models.

Model	MAML	MIML	MIML w/o MI	MIML w/o MA	MIML w/o NM	MIML w/o VAT
Accuracy	85.79 ± 0.15	$\mathbf{91.60 \pm 0.21}$	86.43 ± 0.17	89.59 ± 0.19	84.17 ± 0.13	89.43 ± 0.09

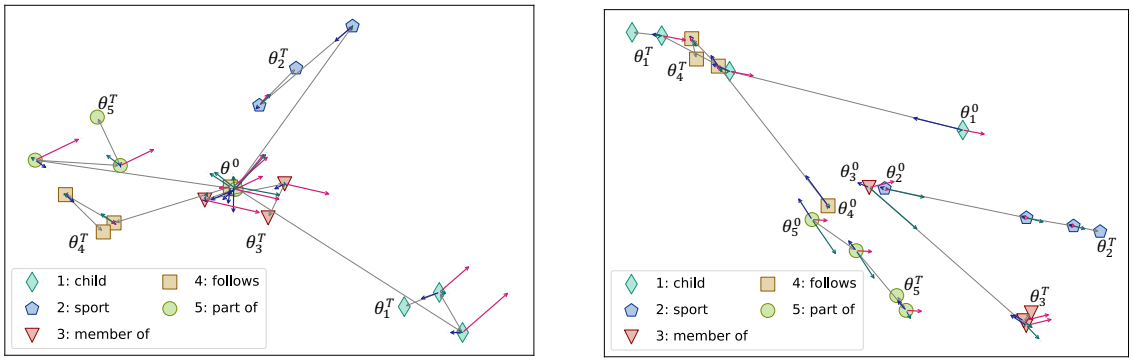
Table 4: Ablation results in 10-way-5-shot setting on FewRel development set. MI/MA: meta-information guided fast initialization/adaptation, NM: Normalization, VAT: virtual adversarial training.

4.6 Visualization

In addition to the improvements in performance, the meta-information guided meta-learning process in MIML can also provide better interpretability in few-shot classification problems. To give a more intuitive picture and show the interpretability of MIML, we visualize the workflow of MIML in the presence of 20% noise in 5-way-5-shot setting, and compare it with MAML. Specifically, we visualize the initialization representations and adaptation steps using principal component analysis (Jackson, 2005). From Figure 2, we have the following observations:

(1) In comparison to MAML, the initialization parameters in MIML reflect the semantic similarity between classes. For example, the initialization point of relation *sport* is close to *member of*, and far from *child*. This is achieved by the semantic guidance from high-level meta-information.

(2) The fast adaptation of MAML is highly influenced by noisy instances, and exhibits high variance in adaptation trajectories. In comparison, noisy instances in MAML are assigned with smaller learning rates by the proposed attention mechanism (not shown in figure), and thus produce smaller noisy gradient steps, which results in more stable adaptation trajectories.



(a) MAML (b) MIML

Figure 2: Visualization of initialization and adaptation process of meta-learning models, in 5-way-5-shot setting with 20% noise. At each iteration, the adaptation gradients for a class parameter θ_i come from three parts: informative instances from class $\mathcal{C}_i$ (marked in **green** arrows), noisy instance for class $\mathcal{C}_i$ (marked in **red** arrows), and instances for other classes (marked in **blue** arrows).[1] Best viewed in color.

5 Related Work

Few-Shot Learning. Few-shot learning aims to grasp new tasks with only a handful of training data. There are mainly two lines of approaches for few-shot learning:

(1) **Metric-Learning** methods learn an embedding space that can well measure the similarities between instances. Koch et al. (2015; Vinyals et al. (2016) use vector distance functions to measure the similarities of examples, while Sung et al. (2018; Garcia and Estrach (2018) use neural networks to learn the metrics. Besides, Snell et al. (2017) propose to calculate prototypes of each few-shot class for classification. Specifically targeting few-shot relation classification, Gao et al. (2019a) introduce a hybrid attention mechanism to alleviate noise data problems. Ye and Ling (2019; Soares et al. (2019; Gao et al. (2019b; Sui et al. (2020) utilize local feature comparison to further improve few-shot performance.

(2) **Meta-Learning** models, on the other hand, transfer the experience about how to "learn" a new class from the training set to the test domain. One way of meta-learning is to use recurrent networks to grasp the meta knowledge and predict the updated parameters in a black-box manner (Ravi and Larochelle, 2017; Munkhdalai and Yu, 2017; Mishra et al., 2018). Another direction is to learn how to better initialize parameters for new classes (Finn et al., 2017; Finn et al., 2018) or apply faster adaptation (Bertinetto et al., 2018; Zintgraf et al., 2019; Rajeswaran et al., 2019) through meta-training. Our work is mainly based on MAML (Finn et al., 2017), about which we have given a brief introduction in Section 1. Many efforts have been devoted to improving MAML. In addition to initialization parameters, Li et al. (2017) propose to also meta-learn adaptation learning rate from implicit instance statistics. Rusu et al. (2018) learns a data-dependent representation of model parameters for initialization, and performs gradient-based meta-learning in the low-dimensional space. Yao et al. (2019) clusters relevant tasks and initialize the tasks within the same cluster with the same parameters. In comparison, MIML integrates meta-information into meta-learning, which provides strong guidance in both initialization and adaptation.

Zero-Shot Learning. Zero-shot learning focuses on grasping new tasks with no training data, which usually takes meta-information, such as names or descriptions to learn new classes. There are many efforts for zero-shot learning in the cross-modal scenario, where class names serve as meta-information for images (Socher et al., 2013; Frome et al., 2013; Norouzi et al., 2014). The general idea of these approaches is to align the semantic spaces of images and their names.

Existing meta-learning approaches provide an efficient framework for transfer learning and fast adaptation, while zero-shot models prove the effectiveness of meta-information. To the best of our knowledge, MIML is the first attempt to combine meta-information with meta-learning for few-shot classification.

[1]For clearer visualization, we only show 3 adaptation iterations. We show the average gradients from 4 informative instances, and the average gradients from the other 4 classes. Thus the lengths of these gradients are 4× longer than shown.

6 Conclusion and Future Work

In this work, we propose a meta-information guided meta-learning framework (MIML) for few-shot relation classification. We conduct comprehensive experiments and achieve human-level performance in few-shot relation classification with only one shot. In addition, we show the advantage and interpretability of MIML in handling noisy instances, and its potential in zero-shot classification.

We plan to explore the following directions as our future work: (1) We will explore more meta-information for meta-learning, such as class descriptions and knowledge graphs. (2) We will develop more sophisticated models to capture the fine-grained interactions between the high-level meta-information and concrete instances, to better guide meta-learning for few-shot classification problem.

7 Acknowledgement

This work is funded by the NSFC/DFG Collaborative Research Centre SFB/TRR169 "Crossmodal Learning" II and Beijing Academy of Artificial Intelligence (BAAI).

References

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2019. Few-shot text classification with distributional signatures. In *Proceedings of ICLR*.

Luca Bertinetto, Joao F Henriques, Philip Torr, and Andrea Vedaldi. 2018. Meta-learning with differentiable closed-form solvers. In *Proceedings of ICLR*.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2018. A closer look at few-shot classification. In *Proceedings of ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of ICML*, pages 1126–1135.

Chelsea Finn, Kelvin Xu, and Sergey Levine. 2018. Probabilistic model-agnostic meta-learning. In *Proceedings of NeurIPS*, pages 9516–9527.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Proceedings of NIPS*, pages 2121–2129.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019a. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of AAAI*, pages 6407–6414.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019b. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of EMNLP-IJCNLP*, pages 6250–6255.

Victor Garcia and Joan Bruna Estrach. 2018. Few-shot learning with graph neural networks. In *Proceedings of ICLR*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of EMNLP*, pages 4803–4809.

J Edward Jackson. 2005. *A user's guide to principal components*, volume 587. John Wiley & Sons.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *Proceedings of the Workshop of ICML*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of ICML*, pages 1885–1894.

Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. 2011. One shot learning of simple visual concepts. In *Proceedings of CogSci*.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *JMLR*, 5:361–397.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. 2017. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL*, pages 1003–1011.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2018. A simple neural attentive meta-learner. In *Proceedings of ICLR*.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *Proceedings of ICLR*.

Tsendsuren Munkhdalai and Hong Yu. 2017. Meta networks. In *Proceedings of ICML*, pages 2554–2563.

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. 2014. Zero-shot learning by convex combination of semantic embeddings. In *In proceedings of ICLR*.

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

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. 2019. Meta-learning with implicit gradients. In *Proceedings of NeurIPS*, pages 113–124.

Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *Proceedings of ICLR*.

Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. 2018. Meta-learning with latent embedding optimization. In *Proceedings of ICLR*.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of NIPS*, pages 4077–4087.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of ACL*, pages 2895–2905.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *In Proceedings of NIPS*, pages 935–943.

Dianbo Sui, Yubo Chen, Binjie Mao, Delai Qiu, Kang Liu, and Jun Zhao. 2020. Knowledge guided metric learning for few-shot text classification. *arXiv preprint*.

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 CVPR*.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Proceedings of NIPS*, pages 3630–3638.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint*.

Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. 2019. Hierarchically structured meta-learning. In *Proceedings of ICML*, pages 7045–7054.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. In *Proceedings of ACL*, pages 2872–2881.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. In *Proceedings of NAACL*, pages 1031–1040.

Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. 2019. Fast context adaptation via meta-learning. In *Proceedings of ICML*, pages 7693–7702.

A Few-Shot Text Classification

In addition to relation classification, MIML can also potentially be applied to other few-shot classification tasks. We perform experiments on a text classification dataset RCV1 (Lewis et al., 2004), which contains Reuters newswire articles under different topics. Following Bao et al. (2019), we use a subset of RCV1 with 740 articles in 37 topics for training and 680 articles in 34 topics for validation. We compare with baselines from Bao et al. (2019), where models are treated as a combination of text representation and learning algorithm:

(1) **Text Representations. AVG** calculates the average embeddings of the words as representations. **IDF** weights the word embeddings by inverse frequency. **CNN** represents text by the outputs after a one-dimensional convolution layer and a max-pooling layer. **DS** (Distributed Signature) uses attention scores learned by a meta-learning framework to weight word embeddings (Bao et al., 2019). In implementing **BERT** encoder, we obtain the input by the concatenation of the first 60 and the last 40 tokens of the article for better efficiency.

(2) **Learning Algorithms.** In addition to **PROTO** and **MAML**, we compare with another three learning algorithms. **NN** finds the nearest neighbor of Euclidean distance. **FT** first pre-trains a classifier using all training examples, and then fine-tunes on the support set (Chen et al., 2018). **DS-ML** estimates the attention score over word embeddings via a meta-learning framework (Bao et al., 2019).

The results are shown in Table 5, from which we observe that MIML achieves competitive performance on few-shot text classification, demonstrating its effectiveness. We leave exploring the potential of MIML in other few-shot classification tasks as future work.

Alg.	Rep.	5-way-1-shot	10-way-1-shot
NN	AVG	43.76	60.84
	IDF	41.96	58.27
FT	CNN	40.33	62.34
PROTO	AVG	28.48	31.22
	IDF	32.14	35.63
	CNN	28.43	29.33
	BERT	39.64	48.66
MAML	AVG	39.98	50.69
	IDF	42.58	54.14
	CNN	39.03	51.15
	BERT	56.41	72.58
DS-ML	DS	54.15	75.38
MIML	BERT	**57.75**	**80.46**

Table 5: Accuracies (%) of few-shot text classification on RCV1 validation set. Results without BERT encoders are from Bao et al. (2019).