CROSS-MODAL KNOWLEDGE ENHANCEMENT MECHANISM FOR FEW-SHOT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot learning problems require models to recognize novel classes with only a few supported samples. However, it remains challenging for the model to generalize novel classes with such limited samples. Driven by human behavior, researchers introduced semantic information (e.g. novel categories descriptions, class names, etc.) onto existing methods as prior knowledge to induct more precise class representations. Despite the promising performance, these methods are under one assumption that users are able to provide precise semantic information for all target categories and this is hard to be satisfied in the real scenario. To address this problem, we proposed a novel Cross-modality Knowledge Enhancement Mechanism(CKEM) to discover task-relevant information in external semantic knowledge automatically. CKEM first utilizes Cross-modality Graph Builder(CGB) to align two unitary modality information (support labeled images and external semantic knowledge) into a cross-modality knowledge graph. After that, with the message-passing mechanism, CKEM selects and transfers relevant knowledge from external semantic knowledge bank to original visual-based class representations in Knowledge Fusion Model(KFM). Through a series of experiments, we show that our method improves the existing metric-based meta-learning methods with 1% - 5% for 1-shot and 5-shot settings on both mini-ImageNet and tiered-ImageNet datasets.

1 INTRODUCTION

Generalizing new concepts from a few samples quickly is one of key signatures for human intelligence. Albeit deep learning methods have made significant progress in wide applications, such as image recognitions, object detection, etc. It remains challenging to adapt to such strict situations, where annotated samples are limited or target classes are flexible at inference stage. Unfortunately, this scenario is common in real world and has drawn many researchers' attention recently. Typically, this problem is regarded as Few-shot Learning problem Bart & Ullman (2005); Fink (2004); Fei-FeiLi et al. (2006); Lake et al. (2011).

To address this problem, one of core research interests is how to generalize novel classes with only a few labeled samples per class. Most of existing methods are under the umbrella of meta-learning mechanism Hochreiter et al. (2001), which leverages previous learning experience over tasks as prior knowledge during meta-training to improve later generation procedure at meta-testing. More precisely, the transferable prior knowledge obtained during meta-training stage can act as an inductive bias to minimize generalization error Luo et al. (2020). However, most proposed meta-learning methods merely utilize unitary modality (visual) information. Due to the limited samples for each class, generalization procedure suffers unstable problems, such as "meta-shift" problemChen et al. (2020). Naturally, recent researchers proposed many methods to introduce auxiliary information from unlabeled samples or other modality prior knowledge and achieved significant performance.

More precisely, for cross-modality methods, Xing et al. (2019) proposed Adaptive Modality Mixture Mechanism (AM3) to fuse information in two modalities by adaptively combining visual prototypes and corresponding class semantic features. Based on AM3, Schwartz et al. (2019) constructs multiple branches to make further improvements by using richer information from multiple semantic sources. To step further, Peng et al. (2019) proposed Knowledge Transfer Network (KTN) archi-

tecture, which learns the classifier of novel classes not only from visual information but also from corresponding semantic information and its class-relationship in a well-trained knowledge graph.

While promising, existing cross-modality meta-learning FSL methods mentioned above are under one strict assumption that precise semantic information for novel classes are available during inference stage. However, such semantic information is hard to acquire in practice. More concretely, for model designers, it is unmanageable to forecast novel classes chosen by users. For model users, it is also difficult and inconvenient to provide an accurate semantic information for each target classes while using the model. Furthermore, an inaccurate semantic information can even prohibit models from generating proper novel class representations and harm the performance. In this paper, we proposed Cross-modality Knowledge Enhancement Model(CKEM) to loose this restriction by empowering models the ability to index relevant information contained in auxiliary semantic knowledge bank.

To achieve this goal, CKEM first utilize Cross-modality Graph Builder(CGB) to transfer two unitary modality class information into the same metric space where similar classes are close while far in the contrary. After that, we build cross-modality knowledge graph in this space and propagate relevant information from semantic knowledge bank to target class representations in Knowledge Fusion Model(KFM) with message passing mechanism. To verify our methods, we built our methods upon metric-based meta-learning methods, which perform image classification by measuring distances between label and unlabeled samples with a non-parameterized or parameterized functions. As results shown in Table 1, our methods improve the existing metric-based methods, such as Prototype Network Snell et al. (2017), for 1% - 2% for 1-shot and 5-shot settings on both mini-ImageNet and tiered-ImageNet.

2 PRELIMINARIES

Few-shot classification and metric-based methods. Few-shot classification problem is typically characterized as several N way (number of novel categories) and K shot (number of available samples for each class) classification tasks. Concretely, for each task/episode, models are required to generalize target samples into N novel classes with only K supported labeled samples per class. In this paper, we focused on metric-based meta-learning approach, which is one of most effective branches in this research area. Generally, a metric-based algorithm contains a feature extractor f_e, a class descriptor f_d and a metric classifier f_c. For each task/episode, the parameterized feature extractor extracts features $\mathbb{Z}_s$ and $\mathbb{Z}_q$ for support set $\mathbb{S}$ (labeled support images) and query set $\mathbb{Q}$ (target images). After that, class descriptor generalize novel classes $\mathbb{C}$ by referring to support set extracted features $\mathbb{Z}_s$. Finally, metric classifier predicts target labels $\mathbb{Y}$ for target images by comparing similarities between $\mathbb{Z}_s$, $\mathbb{C}$ and $\mathbb{Z}_q$.

Restrict cross-modal metric-based methods. The limitation for quantity of $\mathbb{S}$ restricts generalization ability in class descriptor. Motivated by human behavior, recent researchers extended classic metric-based methods to cross-modal metric-based methods by introducing class relevant semantic knowledge into class descriptor. Despite promising results, existing cross-modal metric-based methods assume that users should provide such precise information for each novel classes and this condition is hard to be satisfied in actual scenario. Thus, in this paper, we loose this restriction and formulated a restrict cross-modal few-shot setting, Task-Independent Cross-modal Few-Shot Learning setting (TIC-FSL). In TIC-FSL, we argued that model should obtain the ability to index novel-relevant information in an external knowledge bank $\mathbb{W}_{ex}$ and improve model performance by utilizing these supported informations:

$$\hat{y} = f(x; \mathbb{W}_{ex}, \theta), \tag{1}$$

where, θ notes trainable parameters in model, x notes unlabeled target sample and $\hat{y}$ notes the prediction for x. To be clarified, $\mathbb{W}_{ex}$ represents the external semantic knowledge bank, which correlation with target classes are unknown.

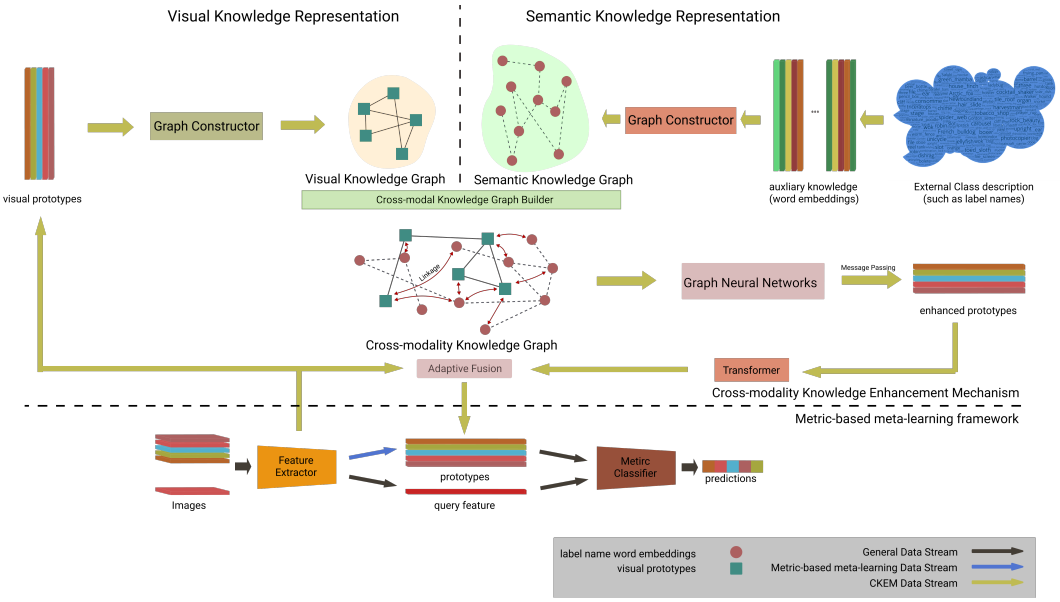

Figure 1: Overview of Cross-modality Knowledge Enhancement Mechanism

3 METHODOLOGY

3.1 OVERVIEW OF CROSS-MODALITY KNOWLEDGE ENHANCEMENT MECHANISM

In this section, we dive into details of proposed Cross-modality Knowledge Enhancement Mechanism, which is designed to enhance generalization ability for existing metric-based methods. Concretely, this mechanism explicitly enhances class representations (prototypes) by exploring class relevant semantic information in the external knowledge bank. As described in figure 1, CKEM is divided into two parts, Cross-modality Graph Builder and Knowledge Fusion Model. Cross-modality Graph Builder(CGB) construct cross-modality graph $\mathcal{G}_c$ by encoding both auxiliary and tasks-specific class representation into the same feature space. In this feature space, two modal prototypes formulate two knowledge graphs, $\mathcal{G}_t$ for semantic-based prototypes and $\mathcal{G}_v$ for visual-based prototypes, and these two graphs align to each other during training stage. After that, Knowledge Fusion Model is constructed upon this cross-modal knowledge graph to transfer auxiliary knowledge from semantic knowledge bank to visual target prototypes.

3.1.1 CROSS-MODALITY GRAPH BUILDER

Cross-modality Graph Builder aims at distilling knowledge from both two different modalities and align them in same feature space. Knowledge Graph(KG) is well-known for the ability to obtain reserve and explore its node information and relationship. Hence, we choose KG as containers to construct our cross-modality graph $\mathcal{G}_c$. Due to the heterogeneous structure of visual and semantic feature spaces Xing et al. (2019), we build parametrized adjuster for both two modality prototypes to align them during training procedure. Then, CGB constructs two unitary modality KG, $\mathcal{G}_v$ and $\mathcal{G}_t$, by regarding the adjusted prototype as vertexes separately. After that, these two unitary KGs are used to construct a cross-modal Knowledge $\mathcal{G}_c$, which plays a key role in Knowledge Fusion Model. The details of constructions are elaborated as follows:

Assuming representation of an vertex i is given by $h^i \in \mathbb{R}^D$, we define visual-based knowledge graph $\mathcal{G}_v = (H_v, A_v)$, where $H_v = \{h_v^i \mid \forall i \in [1, N]\} \in \mathbb{R}^{N \times D}$ and $A_v = \{A_v(h_v^i, h_v^j) \mid \forall i, j \in [1, N]\} \in \mathbb{R}^{N \times N}$ denote the vertex feature matrix and vertex adjacency matrix respectively. For better explanation, we first discuss about the vertex feature matrix H_v. For each episode, models are required to classify several samples into N categories, which contain only K annotated samples per class. Instead of constructing graph over instances as previous methods Liu et al. (2019); Kim et al. (2019); Zhang et al. (2019), we argued that building knowledge graphs in class-level is more

stable. Therefore, vertexes h_v^i in H_v are constructed with adjusted visual-based prototypes $\mathbb{P} = \{p_i \mid \forall i \in [1, N]\}$, as equation 2. For vertex adjacency matrix A_v, edge weight $A_v(h_v^i, h_v^j)$ denotes the similarity between h_v^i and h_v^j. Formally, similarities are gauged by the parametrized function as equation 3:

$$h_v^i = g_v(p_i; \phi_v), \tag{2}$$

$$A_v(h_v^i, h_v^j) = \sigma(W_v|h_v^i - h_v^j|/\gamma_v + b_v), \tag{3}$$

where g_v is the visual-side adjuster with trainable parameters ϕ_v and p_i is the original prototype (mean over all K labeled samples' embeddings for each class) for class i. W_v, b_v and γ_v represent learnable parameters. σ is the sigmoid function, which normalizes the output between 0 and 1.

Similar to visual-based knowledge graph $\mathcal{G}_v$, we define semantic knowledge graph $\mathcal{G}_t = \{H_t, A_t\}$, where $H_t = \{h_t^i \mid \forall i \in [1, C]\} \in \mathbb{R}^{N \times D}$ and $A_t = \{A_t(h_t^i, h^j) \mid \forall i, j \in [1, C]\} \in \mathbb{R}^{N \times N}$. Differently, $\mathcal{G}_t$ is constructed by predefined semantic embeddings provided in external semantic knowledge bank $\mathbb{W}_{ex}$. Note that C represent the number of assistant classes in $\mathbb{W}_{ex}$. Analogous to the definitions for vertexes feature matrix and edge weight in $\mathcal{G}_v$, vertexes feature matrix H_t and edge weight $A_t(h_t^i, h_t^j)$ for $\mathcal{G}_v$ are formulated as equations below:

$$h_t^i = g_v(w_i; \phi_t), \tag{4}$$

$$A_t(h_t^i, h_t^j) = \sigma(W_t|h_t^i - h_t^j|/\gamma_t + b_t), \tag{5}$$

where W_t, b_t and γ_t represent learnable parameters. w_i denotes semantic class embedding in external semantic knowledge bank $\mathbb{W}_{ex}$.

Finally, after constructing two unitary KGs, CGB connects them to construct a cross-modal knowledge graph $\mathcal{G}_c$, which plays a key role in Knowledge Fusion Model. Moreover, CGB utilizes non-parameterized function to create intra-adjacent matrix to connect two sub-graph instead of using parameterized ones. By doing this, models are able to learn the ability to align class embeddings from different modalities during training stage.

More precisely, cross-modal knowledge graph $\mathcal{G}_c$ is formulated as $\mathcal{G}_c = \{H_c, A_c\}$, where $H_c = (H_v; H_t) \in \mathbb{R}^{(N+C \times D)}$ and $A_c = (A_v, A_s; A_s^t, A_t) \in \mathbb{R}^{(N+C) \times (N+C)}$. We denote $A_s = \{A_s(h_v^i, h_t^j) \mid \forall i \in [1, N], \forall j \in [1, C]\} \in \mathbb{R}^{N \times C}$ as its intra-adjacent matrix and the link weight $A_s(h_v^i, h_t^j)$ is calculated by applying softmax over Euclidean distances between h_v^i and $\{h_t^j | \forall j \in [1, C]\}$ as following:

$$A_s(h_v^i, h_t^j) = \frac{exp(-||h_v^i - h_t^j||_2^2)}{\sum_{k'=1}^{N} exp(-||h_v^i - h_t^k||_2^2)} \tag{6}$$

3.1.2 KNOWLEDGE FUSION MODEL

In section 3.1.1, we discuss the construction for the cross-modality knowledge graph $\mathcal{G}_c$. In this section, we are about to focus on the knowledge fusion progress. Among the knowledge fusion progress, Knowledge Fusion model is required to explore target relevant semantic information in auxiliary modality and fuse them with visual-based adjusted prototypes accoutered in each episode. With help of the cross-modality knowledge graph $\mathcal{G}_c$, we are able to propagate relevant semantic knowledge from semantic knowledge graph G_t to the prototype-based visual knowledge graph $\mathcal{G}_v$ by building a Graph Neural Networks (GNN) upon it. In this work, following the *message-passing* framework Gilmer et al. (2017), GNN is formulated as:

$$H_i^{(l+1)} = MP(A_c, H_i^{(l)}; W^{(l)}), \tag{7}$$

where $MP(\cdot)$ is the message passing function and has several possible implementations Hamilton et al. (2017); Kipf & Welling (2017); Velickovic et al. (2018), $H_i^{(l)}$ is vertexes feature matrix which is regarded as input of the l th layer of GNN and $W^{(l)}$ denotes a learnable weight matrix in lth layer. The input of whole GNN is formulated as $H^{(0)} = H_c$. After stacking a GNN layers, we obtain

the knowledge mix-up prototypes for N enhanced class embeddings as the output of the ath layer, which is denoted as $H_f = \{h_f^j \mid j \in [1, N]\}$.

After that, we regard these enhanced prototypes H_f as complement materials to original prototypes set $\mathbb{P}$. Moreover, knowledge transfer process are applied in the cross-modality feature space, which is different with the one for original prototypes. Hence, we apply a parametrized transformer $g_{trans}(\cdot; \phi_{trains})$ to align enhanced prototypes in H_f with the original ones and perform an adaptive combination between them as equation 9.

$$\lambda_i = \frac{1}{1 + exp(-f_\lambda\ (g_{trans}(h_f^i; \phi_{trans})))} \tag{8}$$

$$p_i{}' = (1 - \lambda_i)\ p_i + \lambda_i\ g_{trans}(h_f^i; \phi_{trans}), \tag{9}$$

where f_λ is the adaptive mixing network with learnable parameters ϕ_λ. λ_i denotes the mixing coefficient for class i to balance between original prototype and the complement one.

3.2 Object Function and Training Procedure

Cross-modality Knowledge Enhance Mechanism performs as a plug-and-play mechanism to improve existing metric-based meta-learning methods by exploring and transferring task-relevant semantic information onto corresponding prototypes. In this section, we introduce the details of its object function and training procedure.

We train CKEM with base method with episode training mechanism and update all trainable parameters Θ for the entire framework including the backbone for each episode independently. Motivated by Yao et al. (2020), to stabilize the training procedure, we additionally construct two auto encoder branches for adjusted visual prototypes set H_v and the mix-up prototypes set H_f to regularize the model. More precisely, we choice L1 loss between the input feature x and output of the auto encoder as our reconstruction loss $\mathcal{L}_r(x)$, as equation 10.

$$\mathcal{L}_r(x) = ||x - AE_{dec}(AE_{enc}(x; \phi_{enc}); \phi_{dec})|| \tag{10}$$

where x denotes the input of the auto encoders, AE_{enc} and AE_{dec}. ϕ_{enc} and ϕ_{dec} denote the learnable parameters of encoder and decoder in auto-encoder. These reconstruction losses over H_v and H_f are added to the original object function $\mathcal{L}_{base}$ of base metric-based method to formulate the whole object function. As shown in equation 11, after obtaining the whole object function, we update learnable parameters Φ of both CKEM and base method using Stochastic Gradient Descent mechanism to minimize the total loss. Formally:

$$\min_{\Theta} \mathcal{L}_{all} = \min_{\Theta} \mathcal{L}_{base}\ +\ \mu_1 \sum_{h_v \in H_v} \mathcal{L}_r(h_v)\ +\ \mu_2 \sum_{h_f \in H_f} \mathcal{L}_r(h_f), \tag{11}$$

where Φ denotes the learnable parameters for both CKEM framework and the base metric-based FSL method. μ_1 and μ_2 are introduced to balance the importance of these three terms.

4 Related Work

With rapid development of few-shot learning methods Wang et al. (2019); Zhang et al. (2019); Inoue & Shinoda (2018); Dong et al. (2018), researchers delve into several few-shot tasks, image recognition task, image segmentation task, text classification, etc. In this paper, we focus on one fundamental problem in computer vision area, image recognition problem. We roughly break related few-shot image recognition methods into two branches, visual unitary modality and classic cross modality few-shot meta-learning methods.

4.1 Visual-based Few-shot Learning

Recently, researchers have made significant progress in few-shot learning area. Among these methods, meta-learning played an dominant role during this progress. Our proposed approaches are also

located in this branch. Meta-learning methods can roughly divide into two parts, Initialization-based methods and Metric-based methods.

Initialization-based meta-learning methods aim at obtaining task-specific classifier parameters for novel classes with only a few annotated samples supplied. One way to achieve this goal is to regard meta-learner as an optimizer which gathers gradient flows from different tasks to refine parameter of the model. By doing this, models can generalize well to novel tasks with only a few fine-tuning updates. One of the base methods under this branch is Model-Agnostic Meta-Learning(MAML) framework Finn et al. (2017); Nichol et al. (2018); Zintgraf et al. (2018); Mishra et al. (2018). Build upon this base method, many follow-up approaches were proposed to improve its performance. Kim et al. (2018) and Finn et al. (2018) propose a probabilistic extension to MAML by training with variational approximation. Conditional Class-Aware Meta-Learning (CAML) Zintgraf et al. (2018) conditionally transforms embeddings based on a metric space trained with prototypical networks to capture inter-class dependencies. On the other hand, many methods conduct meta-generator to hallucinate parameters of classifier to classify samples to novel categories Rusu et al. (2019); Zhou et al. (2019). Latent embedding optimization (LEO) Rusu et al. (2019); Qiao et al. (2018) use a few updates on a low data regime to train models in a high dimensional parameter space, from which to decode classifier parameters. Similar to LEO, during training stage, Visual Analogy Graph Embedded Regression (VAGER) Zhou et al. (2019) learns a linear mapping function to generate classification parameters, which are applied to new class embeddings through their visual analogy with base classes. And Qiao et al. (2018) adapt a pretrained neural network to novel categories by directly predicting the parameters from the activations.

Metric-based approaches address the few-shot image recognition problem by *learning to compare*. To achieve this goal, these methods are applied to episode training mechanism to generalize distinguish representations which have close intra-class distances and far inter-class distances. Among these metric-based methods Koch et al. (2015); Sung et al. (2018); Vinyals et al. (2016); Snell et al. (2017); Allen et al. (2019); Oreshkin et al. (2018), Prototypical Network (PN) is famous for its simplicity and effectiveness, which affect many followers to extend this approach. Allen et al. (2019) allow each class to be represented by multiple prototypes to improve the representation power of PN. TADAMOreshkin et al. (2018) use a context-conditioned embedding network to produce prototypes that are aware of the other classes. In this paper, similarly our proposed method is also designed to enhance these prototypes. To utilize the relationship to assist the model, some researchers Liu et al. (2019); Kim et al. (2019); Zhang et al. (2019); Luo et al. (2020) also introduce Graph Neural Networks to propagate label information in the graph and transductively classify samples in the query set.

4.2 Cross-modal Few-shot learning Methods

Above mentioned approaches are rely solely on visual features for few-shot classification. However, due to the random sample strategy and limited quantity of annotated samples, merely using visual samples to produce class representation leads instable generalization procedure, such as meta-shift problems. Chen et al. (2019a) And semantic information contained in descriptions over categories can alleviate this problems. In order to utilize the abundant text information during the training and inference stage, researchers set about combining text information with existing visual information. Driven by zero-shot learning methods Schönfeld et al. (2018); Tsai et al. (2017); Frome et al. (2013), Xing et al. (2019) proposed Adaptive Modality Mixture Mechanism (AM3) to fuse information from two modalities by adaptively combining visual-based prototypes and corresponding semantic features. Based on AM3, Schwartz et al. (2019) construct multiple branches to make further improvements by using richer semantics and multiple semantic sources. To utilize inter-class relationship, Peng et al. (2019) proposed Knowledge Transfer Network (KTN) architecture, which learns the classifier of novel classes not only from visual information but also from corresponding text class information and its class-relationship in a well-trained knowledge graph.

Albeit the promising results, these methods are under one strict assumptions that users are acquired to manually provide class relationship between auxiliary semantic knowledge and target categories. In this paper, we loose this constrain and build our proposed approaches in a more challenge but practical cross-modality settings, TIC-FSL problem.

Table 1: Overall performance over unitary modality few-shot learning methods

	mini-ImageNet		tiered-ImageNet	
Methods	1-shot	5-shot	1-shot	5-shot
Prototypical Network Snell et al. (2017)	49.42% +- 0.78%	68.20% +- 0.66%	53.31% +- 0.89%	72.69% +- 0.74%
Relation Network Sung et al. (2018)	50.40% +- 0.80%	65.30% +- 0.70%	55.00% +- 1.00%	69.30% +- 0.80%
MAML Finn et al. (2017)	48.70% +- 1.84%	63.10% +- 0.92%	58.90% +- 1.90%	71.50% +- 1.00%
REPTILE Nichol et al. (2018)	49.97% +- 0.32%	65.99% +- 0.58%	62.95% +- 0.03%	71.03% +- 0.22%
TADAM Oreshkin et al. (2018)	58.50% +- 0.30%	76.70% +- 0.30%	62.13% +- 0.31%	81.92% +- 0.30%
MetaOptNet Lee et al. (2019)	62.64% +- 0.61%	78.63% +- 0.46%	65.99% +- 0.72%	81.56% +- 0.53%
SNAIL Mishra et al. (2018)	55.71% +- 0.99%	68.88% +- 0.92%	-	-
LEO Rusu et al. (2019)	61.76% +- 0.08%	77.59% +- 0.12%	66.33% +- 0.05%	81.44% +- 0.09%
ProtoNet (normalize)	61.93% +- 0.74%	77.90% +- 0.35%	64.06% +- 0.27%	78.03% +- 0.19%
ProtoNet (normalize) + CKEM	**63.29% +- 0.71%**	**80.12% +- 0.22%**	**66.69% +- 0.75%**	**83.04% +- 0.61%**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

To verify the efficiency of cross-modality knowledge enhancement mechanism, we conduct main experiments with the most popular few-shot image recognition datasets: mini-ImageNet Cai et al. (2018) and tiered-ImageNet Ren et al. (2018). *Mini-ImageNet* dataset consists of a subset of 100 classes from the ImageNet dataset Russakovsky et al. (2015) and contains 600 images for each class. The dataset was first proposed by Cai et al. (2018), but most recent researchers use the follow-up settings provided by Ravi & Larochelle (2017) which is composed of randomly selected 64 base, 16 validation and 20 novel classes. *Tiered-ImageNet* dataset is also consists of a subset of ImageNet, but in a larger scale. It contains 608 classes from 34 super-categories, which are then split to 20, 6, 8 super-categories, to select 351, 97, 160 classes as training, validation and test set respectively.

For Task-Independent Cross-modal Few-Shot Learning setting, we provided word embeddings of categories label as external knowledge bank. Note that, as described in Section 2, the relationship between target novel classes and auxiliary semantic information is unknown. Specifically, we use GloVe Pennington et al. (2014) algorithm to generate label semantic representations as previous works Xing et al. (2019). To evaluate our methods, we applied CKEM upon most famous metric-based method, Prototypical Network and compared the results with latest existing methods in Table 1. As results shown, combined with CKEM, our methods exceeded current uni-modal supervised few-shot learning methods. Detail results and analysis are discussed in section 5.3.

5.2 IMPLEMENTATION DETAIL

In this section, we discuss about details of implementations for CKEM and baseline. For both datasets, following recent worksGidaris & Komodakis (2017); Oreshkin et al. (2018); Xing et al. (2019), we conduct ResNet-12 as backbone for our method and pretrained backbone on base classes before meta-training. The configuration of pretraining the backbone are similar but slice different for tiered-ImageNet and mini-ImageNet. For both of them, we use the SGD optimizer with initial learning rate of 0.1 and momentum of 0.9 and set decay factor, weight decay to 0.1, 0.0005 respectively. For tiered-ImageNet, we set batch size to 512, max epoch to 120 and learning rate are decayed at 40th and 80th epoch. For mini-ImageNet, we set batch size to 128, max epoch to 100 and learning rate are decayed at 90th. During the meta-training stage, we also use SGD optimizer with same hyper parameters (set fixed learning rate to 0.001, momentum to 0.9 and weight decay to 0.0005) for both ProtoNet+CKEM and ProtoNet. Note that, we do not apply special sample technique during meta-training process to enhance the methods.

For CKEM. We adapt one layer GCN Gidaris & Komodakis (2017) with tanh activation as the implementation of GNN in equation 7. For auto-encoders, we choose two layers of MLPs with ReLU activations for both encoder and decoder. Then for $\mathcal{L}_{all}$, we set 1 to both μ_1 and μ_2. For baselines, we also enhanced the original Prototypical Networks with a deeper pretrained ResNet-12, which is also used with CKEM. Note that to adapt PN with pretrained backbone, we perform normalization over all extracted sample features. We note *ProtoNet (normalize)* to represent this enhanced baseline. Following evaluation settings in Chen et al. (2019b), we evaluate our methods

with the mean of 200 classification accuracies on randomly generated test episodes as well as the 95% confidence intervals.

5.3 Experiment Results and Analysis

To verify the effectiveness of our method, we applied CKEM onto classic unimodal metric-based meta-learning method, Prototypical Network, and compared the results with existing most latest methods. As shown in Table 1, reproduced ProtoNet(normalize) achieve 61.93% and 77.59% accuracies for 1-shot and 5shot settings for mini-ImageNet. It also achieve 64.06% and 78.03% accuracies for tiered-ImageNet. After combining with CKEM, our method enhance this baseline for both two datasets. Precisely, CKEM improve ProtoNet (normalize) on both mini-ImageNet and tiered ImageNet.(1.36% and 2.63% for 1-shot, 2.22% and 5.01% for 5-shot). CKEM performance better in 5-shot rather than 1-shot settings. We argued that the reason for this phenomena is that quality of query mechanism is based on original prototypes and the prototypes formulated by 5 samples are more accurate than those with only 1 sample.

After that, we also compared our methods with with report results of many existing methods for both unitary modality methods including Prototypical Network Snell et al. (2017), Relation Network Sung et al. (2018), MAML Finn et al. (2017), REPTILE Nichol et al. (2018), TADAM Oreshkin et al. (2018), MetaOptNet Lee et al. (2019), SNAIL Mishra et al. (2018) and LEO Rusu et al. (2019). As results shown, combining with CKEM, our method improve the classic method to exceed report results of most latest methods.

6 Conclusions

In this paper, we introduce a more challenge but practical cross-modality few-shot learning problem, TIC-FSL, where relationship of classes among auxiliary knowledge and target classes is unknown. To address this problem, we proposed Cross-modality Knowledge Enhancement Mechanism as plug-and-play module upon existing metric-based meta-learning methods. More concretely, CKEM utilizes Cross-modality Graph Builder to align and represent two modalities information in a cross-modality knowledge graph. After that, Knowledge Fusion Model transfers information from external semantic knowledge bank to original prototypes via GNN with message passing mechanism. To evaluate the performance of CKEM, the proposed method is applied to existing metric-based meta-learning methods and achieves comparable results on both mini-ImageNet and tiered-ImageNet for both 1-shot and 5-shot supervised few-shot image recognition settings.

References

Kelsey R. Allen, Evan Shelhamer, Hanul Shin, and Joshua B. Tenenbaum. Infinite mixture prototypes for few-shot learning. In *ICML*, 2019.

Evgeniy Bart and Shimon Ullman. Cross-generalization: learning novel classes from a single example by feature replacement. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:672–679 vol. 1, 2005.

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

Da Chen, Feng Mao, Ming-Li Song, Yuan He, Xiang Wu, Jinqiao Wang, Wenbin Li, Yongliang Yang, and Hui Xue. Class regularization: Improve few-shot image classification by reducing meta shift. *ArXiv*, abs/1912.08395, 2019a.

Da Chen, Yongliang Yang, Zunlei Feng, Xiang Wu, Mingli Song, Wenbin Li, Yuan He, Hui Xue, and Feng Mao. Semantic regularization: Improve few-shot image classification by reducing meta shift, 2020.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019b.

Xuanyi Dong, Linchao Zhu, De Zhang, Yi Yang, and Fei Wu. Fast parameter adaptation for few-shot image captioning and visual question answering. *Proceedings of the 26th ACM international conference on Multimedia*, 2018.

Fei-FeiLi, FergusRob, and PeronaPietro. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

Michael Fink. Object classification from a single example utilizing class relevance metrics. In *NIPS*, 2004.

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

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *NeurIPS*, 2018.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *ICLR*, 2017.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. *ArXiv*, abs/1704.01212, 2017.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, 2017.

Sepp Hochreiter, A. Steven Younger, and Peter R. Conwell. Learning to learn using gradient descent. In *ICANN*, 2001.

Nakamasa Inoue and Koichi Shinoda. Few-shot adaptation for multimedia semantic indexing. *Proceedings of the 26th ACM international conference on Multimedia*, 2018.

Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D. Yoo. Edge-labeling graph neural network for few-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Taesup Kim, Jaesik Yoon, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. *ArXiv*, abs/1806.03836, 2018.

Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ArXiv*, abs/1609.02907, 2017.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. One shot learning of simple visual concepts. *Cognitive Science*, 33, 2011.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10649–10657, 2019.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sungju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *International Conference on Learning Representations*, 2019.

Yadan Luo, Zi Huang, Zheng Zhang, Ziwei Wang, Mahsa Baktashmotlagh, and Yang Yang. Learning from the past: Continual meta-learning with bayesian graph neural networks. In *AAAI 2020*, 2020.

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

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *ArXiv*, abs/1803.02999, 2018.

Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 721–731, 2018.

Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image recognition with knowledge transfer. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 441–449, 2019.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L. Yuille. Few-shot image recognition by predicting parameters from activations. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7229–7238, 2018.

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

Mengye Ren, Sachin Ravi, Eleni Triantafillou, Jake Snell, Kevin Swersky, Josh B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations*, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019.

Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8239–8247, 2018.

Eli Schwartz, Leonid Karlinsky, Rogério Schmidt Feris, Raja Giryes, and Alex M. Bronstein. Baby steps towards few-shot learning with multiple semantics. *ArXiv*, abs/1906.01905, 2019.

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

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

Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3591–3600, 2017.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *ArXiv*, abs/1710.10903, 2018.

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

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 2019.

Chen Xing, Negar Rostamzadeh, Boris N. Oreshkin, and Pedro H. O. Pinheiro. Adaptive cross-modal few-shot learning. In *NeurIPS*, 2019.

Huaxiu Yao, Xian Wu, Zhiqiang Tao, Yaliang Li, Bolin Ding, Rui rui Li, and Zhenhui Li. Automated relational meta-learning. *ArXiv*, abs/2001.00745, 2020.

Chenrui Zhang, Xiaoqing Lyu, and Zhi Tang. Tgg: Transferable graph generation for zero-shot and few-shot learning. *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, and Qi Tian. Learning to learn image classifiers with visual analogy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11497–11506, 2019.

Luisa M. Zintgraf, Kyriacos Shiarlis, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Caml: Fast context adaptation via meta-learning. *ArXiv*, abs/1810.03642, 2018.