Learn to Adapt for Generalized Zero-Shot Text Classification

Anonymous ACL submission

Abstract

Generalized zero-shot text classification aims to classify textual instances from both previously seen classes and incrementally emerging unseen classes. Most existing methods generalize poorly since the learned parameters are only optimal for seen classes rather than for both classes, and the parameters keep stationary in predicting procedures. To address these challenges, we propose a novel Learn to Adapt (LTA) network using a variant meta-learning framework. Specifically, LTA trains multiple meta-learners by using both seen classes and virtual unseen classes to simulate a generalized zero-shot learning (GZSL) scenario in accordance with the test time, and simultaneously learns to calibrate the class prototypes and sample representations to make the learned parameters adaptive to incoming unseen classes. We claim that the proposed model is capable of representing all prototypes and samples from both classes to a more consistent distribution in the global space. Extensive experiments on five text classification datasets show that our model outperforms several competitive previous approaches by large margins. The code and the whole datasets will be available after paper publication.

1 Introduction

Text classification plays an important role in many natural language processing (NLP) applications, such as question classification, news categorization, user intent classification and so on (Minaee et al., 2021). Although a wide variety of methods have been proved successful in supervised text classification, they often break down when applied to make predictions for incrementally emerging classes without labeled training data (Pourpanah et al., 2020). Unlike zero-shot learning (ZSL) that aims to classify unseen class instances at test time (Romera-Paredes and Torr, 2015; Wang et al., 2019), generalized zero-shot learning (GZSL), that we focus on in this work, aims to classify text samples from both previous seen and emerging novel classes. Since there is a strong bias towards seen classes (Xian et al., 2017), GZSL is a more challenging yet critical problem.

Previously methods mainly focus on transductive approaches for generalized zero-shot text classification. Rios and Kavuluru (2018) use a graph convolution network to enhance the unseen class label embeddings. Zhang et al. (2019) and Song et al. (2020) generate illusion feature embeddings for unseen classes based on side information, *i.e.*, class-level attributes or text description. More recently, Ye et al. (2020) use reinforced self-training methods to leverage unlabeled data during training stage.

With the assumption that no knowledge about unseen categories is available during the model learning phase, researchers resort to inductive approaches to handle generalized zero-shot text classification. ReCapsNet (Liu et al., 2019) uses a dimensional attention-based intent capsule network and constructs zero-shot class prototypes by similarity matrix transformation. SEG (Yan et al., 2020) exploits an outlier detection approach that can be directly applied on ReCapsNet, which discriminates the domain first, then outputs the final class label.

However, the existing methods still have two key limitations. Firstly, while the goal of these methods is to transfer beneficial knowledge for unseen classes, these models merely learn optimal parameters by minimizing the loss of instances from seen classes, regardless of explicitly calibrating the predictions on unseen classes. Therefore, domain bias towards seen classes is not fairly resolved (Vinyals et al., 2016). Secondly, although some of them take into account the inter-class relationship when constructing prototypes for unseen classes (Liu et al., 2019), the models keep static no matter what different new classes emerging in future applications. As a result, these models show

a large quality gap between instances from seen classes and from emerging unseen classes.

To address these problems, motivated by the success of meta-learning in the few-shot learning task (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018; Finn et al., 2017), we present a novel Learn-To-Adapt network (LTA) for generalized zero-shot text classification. Concretely, the proposed LTA learns class prototypes over multiple learning episodes that mimic GZSL setting explicitly during training, making the learning setting consistent with the test environment and thereby improving generalization. Then, the model notably extends its ability from two views: prototype adaption and sample adaption. In each episode, the LTA adjusts the representative prototypes of both seen classes and "fake" unseen classes, with the assumption that unseen class will help in calibrating representation of seen ones and thereby enable the model to learn the class sensitive representations. The updating for all prototypes is then used to generate a set of calibration parameters to guide the adaption of sample embeddings, which is designed to compensate for the shrinking features (Chen et al., 2018) that are ignored during training if they are not discriminating for seen classes, but could be critical for recognizing unseen classes. The refined sample embeddings are then classified based on similarity scores with all class prototypes. The same setting can be directly applied in test, where the LTA executes class prediction and adapts the learnt model rationally in an on-the-fly manner.

In summary, our contributions include: (i) We propose a novel Learn to Adapt (LTA) network for generalized zero-shot text classification which is capable of adapting incrementally between seen classes and emerging unseen classes at test time. (ii) We propose a methodology for calibrating both prototypes and samples to deduce a global representation space, efficiently avoiding over-fitting on seen classes. (iii) Experimental results on five generalized zero-shot text classification datasets show that our method outperforms previous methods with a large margin.

2 Related Work

Generalized Zero-Shot Learning The challenge of zero-shot learning (ZSL) has been the focus of attention in recent years, especially in the applications of image classification (Socher et al., 2013; Xian et al., 2017; Wang et al., 2018, 2019), intent classification (Xia et al., 2018; Liu et al., 2019; Yan et al., 2020), and question classification (Fu et al., 2018). Different from ZSL, generalized zero-shot learning (GZSL) that attempts to categorize instances from both seen and unseen classes is a more realistic condition that matches with practical applications. For example, a question classifier for question answering system has to classify not only the questions ever asked but also new questions incrementally emerging from the users.

There are two key issues that GZSL has to address: (1) how to incrementally learn beneficial knowledge for unseen classes from seen ones, and (2) how to tackle the domain bias caused by the extremely imbalanced data of seen and unseen domains.

To alleviate the first issue, some of the earliest works on ZSL attempt to learn a matching model between instance embedding and class prototype embeddings represented by extra information including class-level attribute, text description, or their combinations (Frome et al., 2013; Jinseok Nam, 2016; Zhu et al., 2019; Xia et al., 2018). In a similar vein, other methods (Wang et al., 2018; Rios and Kavuluru, 2018; Si et al., 2020) also investigate the semantic relationship between the side information for obtaining better prototype representation.

The key problem of the second issue is that the model is trained with data from the seen classes and the parameters are actually optimized on seen domain, thus they are not aware of unseen classes. Assuming the extra information about unseen classes is available, another prominent approach attempts to use generative models to generate virtual samples or features for unseen domains (Xian et al., 2018; Schönfeld et al., 2019; Zhang et al., 2019; Song et al., 2020). By using synthesized samples, the generative approaches can convert GZSL problem to the conventional supervised learning problem where biases towards seen classes are largely alleviated. Nevertheless, these models are trained using data from seen classes and fails to incrementally adapt to emerging new classes. Additionally, studies also extend to exploit the unlabeled data for unseen classes (Xian et al., 2019; Rahman et al., 2019; Ye et al., 2020).

However, these models assume that they have access to the extra information about the unseen classes, which is not very realistic since often neither the test data nor their label descriptions is

available at the training phrase (as supposed in this work). In contrast, our model can involve all classes (seen and unseen) jointly during inference, essentially it is trained towards continuous generalization for new classes, hence it is capable to adapt to incoming new class dynamically.

Episode-Based Training in GZSL Our approach is primarily based on episodic training/meta-learning that has been widely used in few-shot learning (FSL) (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018). The primitive goal of episodic training is to quickly learn a meta-task from a small number of sampled classes and supporting sets. A particular advantage of episodic learning is that, by constructing meta-tasks, the setting of training is consistent with that of test, which is essential for classification problems.

Studies extend to exploit episodic training in the "generalized" settings. Gidaris and Komodakis (2018); Ye et al. (2019); Shi et al. (2019) utilize weight generators or relationships to update representative prototypes in generalized FSL (GFSL). Yu et al. (2020) use a generative network to generate unseen prototypes in GZSL. These methods only consider the prototype adaptation while the sample embeddings are still static whatever the unseen classes are. On the contrary, Bao et al. (2020) uses distributional signatures to update sample embeddings in GFSL. Considering that distributional signatures can be equal for two different tasks, our method uses a novel semantic update extractor to update samples following the prototype adaptation rather than statistical information.

A compelling property of our method is that it tackles knowledge transferring and domain bias simultaneously in an adaptive episodic training framework by adapting both prototypes and sample embeddings, and draws a fast adaption to the novel classes without the cost of dramatic damage in discriminating the seen classes.

3 Methodology

3.1 Problem Definition

Formally, let $\mathcal{Y}^s = \{y_1^s, ..., y_{C^s}^s\}$ and $\mathcal{Y}^u = \{y_1^u, ..., y_{C^u}^u\}$ denote C^s seen classes and C^u unseen classes respectively, and $\mathcal{Y} = \mathcal{Y}^s \cup \mathcal{Y}^u$ denote the global label space with $\mathcal{Y}^s \cap \mathcal{Y}^u = \emptyset$. Suppose we have a collection of training samples $\mathcal{D}^s = \{(x_i^s, y_i^s, a_i^s)\}_{i=1}^M$, that consists of M samples from C^s seen classes, where $x_i^s \in \mathcal{X}^s$ represents i-th text utterance, y_i^s is and a_i^s are its one-hot class label and class-level textual description, respectively. At the test time, provided with a class description set $\mathcal{A}^u = \{a_j^u\}_{j=1}^{C^u}$ for unseen classes, the GZSL task is to classify the test instance into either a seen or an unseen class.

3.2 Overview

Encoder An textual input x with T words is encoded by a BERT (Devlin et al., 2018) into a sequence of hidden vectors $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_T] \in \mathbb{R}^{T \times d_h}$, where d_h is the hidden dimension. The text embedding $f(x) \in \mathbb{R}^{d_h}$ is then obtained by averaging over the T hidden vectors.

Training In the training stage, we introduce an episodic learning paradigm, which trains the model by simulating multiple generalized zero-shot text classification tasks on seen classes. Following the principle that train and test conditions must match (Vinyals et al., 2016) and recent studies on "generalized" setting (Gidaris and Komodakis, 2018; Shi et al., 2019; Ye et al., 2019; Bao et al., 2020; Verma et al., 2020; Yu et al., 2020), each episode involves an N^s-way K-shot learning task for seen classes, denoted as $\mathcal{D}_i^s = \{(x_j^s, y_j^s, a_j^s)\}_{j=1}^{N^s \times K}$ with K labelled instances for each of N^s classes randomly sampled from the seen data $\mathcal{D}^s$ in the i-th episode, and a N^u-way K-shot learning task for "fake" unseen classes, denoted as $\mathcal{D}_i^u = \{(x_j^u, y_j^u, a_k^u)\}_{k=1}^{N^u}$ which is also from $\mathcal{D}^s$, with $N^s + N^u \leq C^s$. More precisely, let $\mathcal{Y}_i^s$ and $\mathcal{Y}_i^u$ denote the sampled seen class space and sampled "fake" unseen class space respectively, with $\mathcal{Y}_i^s \subset \mathcal{Y}^s$, $\mathcal{Y}_i^u \subset \mathcal{Y}^u$, and $\mathcal{Y}_i^s \cap \mathcal{Y}_i^u = \emptyset$. For a new query instance x, the generalized zero-shot learning model performs

$$\hat{y} = \arg\max_{y \in \{\mathcal{Y}_i^s \cup \mathcal{Y}_i^u\}} p(y|x, \mathcal{D}_i^s, \mathcal{D}_i^u) \quad (1)$$

The model is designed to maintain a globally joint class prototype space as well as dynamic adaption to unseen classes with zero labeled instances, whose detailed implement is described as follows.

3.3 Prototype Adaptation

The proposed LTA network first introduces a learnable look-up table $\boldsymbol{S} \in \mathbb{R}^{C^s \times d_h}$ from which to extract the seen prototypes $\boldsymbol{S}_i \in \mathbb{R}^{N^s \times d_h}$ on demand. The $\boldsymbol{S}$ is initialized using the supervised classifier by reducing the error on the training samples from the random initialization. The virtual unseen prototypes $\boldsymbol{U}_i$ is produced by the BERT encoder

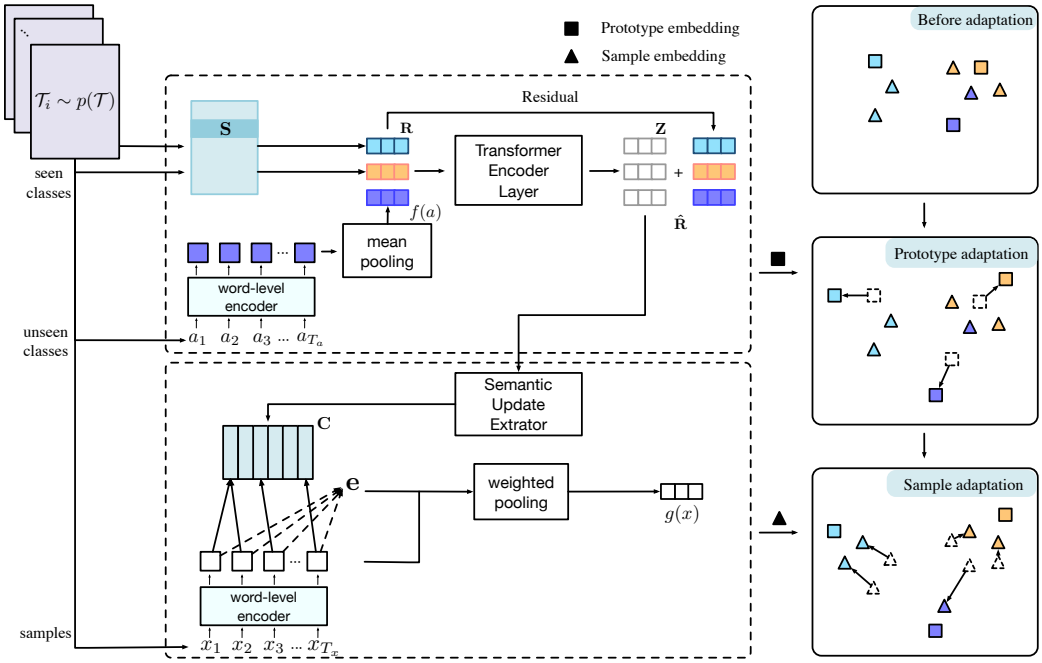


Figure 1: Illustration of the proposed LTA framework. The right part demonstrates the prototype adaption and sample adaption, in which ■ and ▲ respectively denote prototypes and samples, solid border and dotted border represent before and after adaption, respectively.

$f(\cdot)$ using their corresponding class descriptions: $\mathbf{U}_i = [f(a_y)]_{y \in \mathcal{Y}_i^u} \in \mathbb{R}^{N^u \times d_h}$.

Then the joint prototype matrix $\boldsymbol{R}$ is obtained by concatenating $\boldsymbol{S}_i$ and $\boldsymbol{U}_i$, $\mathbf{R} = [\mathbf{S}_i, \mathbf{U}_i] \in \mathbb{R}^{C^s \times d_h}$ with $\mathbf{r}_j$ as the j-th prototype. Then $\boldsymbol{R}$ is fed into an inter-class Transformer encoder (Vaswani et al., 2017) to explicitly model the updates for the representations of both seen prototypes and novel prototypes:

$$\mathbf{Z} = \text{TransformerEncoder}(\mathbf{R}) \tag{2}$$

$$\hat{\mathbf{R}} = \mathbf{R} + \mathbf{Z} \tag{3}$$

where $\mathbf{Z} \in \mathbb{R}^{C^s \times d_h}$ highlights the adjustment after mutual reflections, and the updated prototypes $\hat{\mathbf{R}} \in \mathbb{R}^{C^s \times d_h}$ is regarded as the calibrated representative prototypes of both seen and unseen categories, with $\hat{\mathbf{r}}_j$ as the adjusted j-th prototype. The self-attentions used in Transformer is agile to capture the inter-class relationship of seen and unseen classes and thereby it is beneficial to derive globally discriminative prototypes. The prototypes simultaneously update both seen and unseen classes, which enables the model to represent and discriminate the newly incoming categories in an on-the-fly manner.

3.4 Sample Adaptation

As been discussed in (Chen et al., 2018), the zero-shot learning tasks are prone to produce semantics loss, where some features would be discarded during training if they are not discriminating for seen classes, but critical for recognizing unseen classes. We observe that the similar problem is exacerbated in GZSL task due to the extreme unbalance between seen and unseen classes. We tackle this problem by introducing sample adaption following the trajectories of prototypes adaption. In concrete, we apply a semantic update extractor via attention mechanism to capture synchronous updating of the prototypes:

$$\mathbf{F} = \mathbf{Z}\mathbf{W}_1 \tag{4}$$

$$\mathbf{A} = \text{Softmax}(\mathbf{W}_3 \text{ReLU}(\mathbf{W}_2 \mathbf{F}^T)) \tag{5}$$

$$\mathbf{C} = \mathbf{A}\mathbf{F} \tag{6}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_h \times d_h}, \mathbf{W}_2 \in \mathbb{R}^{d_a \times d_h}, \mathbf{W}_3 \in \mathbb{R}^{d_r \times d_a}$ are trainable parameters, $\mathbf{A}$ denotes the attention weight matrix and $\mathbf{C} \in \mathbb{R}^{d_r \times d_h}$ extracts different semantic components with $\mathbf{c}_j$ is the j-th semantic components. To offset the semantic loss mentioned above, we compare the attention score for each $\mathbf{h}_t$ to get most related semantic adjustment and reconstruct the contribution of each feature:

$$e_t = \text{Softmax}(\beta \max_j (\frac{\mathbf{h}_t \mathbf{c}_j}{\|\mathbf{h}_t\| \|\mathbf{c}_j\|})) \tag{7}$$

$$g(x) = \sum_{t=1}^{T} e_t \mathbf{h}_t \tag{8}$$

where the self-attention weight $e_t \in \mathbb{R}^T$ is used to re-weight the t-th word of sample x to be classified, and β is a scalar to control the differentiation of attention scores. In this way, In this way, the different attention weight discriminate the importance of words rather than averaging them.

One notable reason of choosing of the above feature-level calibration is that, in classification task, the encoder is trained to produce feature embeddings that collapses to its ground-truth prototype, therefore the adjustment of feature embedding should cater to the adjustment of a reliable global prototype space. In addition, since this calibration is applied after the encoding, it reduces the complicated parameter tuning for a massive encoder (*e.g.*, BERT), which elegantly helps the GZSL task to fast adapt to the incoming test instances.

3.5 Loss function

With the adapted prototypes $\hat{\mathbf{R}}$ and the adapted sample $g(x)$, a Softmax classifier is used with cosine similarity:

$$p\left(\hat{y} = y \mid x\right) = \frac{\exp(s(g(x), \hat{\mathbf{r}}_y))}{\sum_{\hat{y}} \exp(s(g(x), \hat{\mathbf{r}}_{\hat{y}}))} \tag{9}$$

where $s(a, b) = \frac{\tau \cdot ab}{\|a\|\|b\|}$ is cosine similarity with temperature τ. Finally the model is trained the minimize the losses across all episodes:

$$\mathcal{L} = -\frac{1}{N} \sum_{i} \mathcal{L}_i \tag{10}$$

where $\mathcal{L}_i$ is the loss of the i-th episode:

$$\mathcal{L}_i = \frac{1}{M} \sum_{(x,y,a) \in \mathcal{D}_i^s \cup \mathcal{D}_i^u} \log p\left(\hat{y} = y \mid x\right) \tag{11}$$

The training process of LTA is summarized in Algorithm 1.

Algorithm 1: LTA training algorithm.

Input: distribution over tasks $p(\mathcal{T})$, class set $\mathcal{Y}^s$
Output: learned model parameters

```
while not done do
    Randomly sample a meta GZSL task T_i ~ p(T) with seen meta-test D_i^s and unseen meta-test D_i^u.
    Get adapted prototypes R̂ by Eq 2~3.
    Get semantic components C by Eq 4~6.
    for all D_i^s ∪ D_i^u do
        Get adapted sample embeddings by Eq 7~8.
    end
    Update model by Eq 9 and Eq 11.
end
```

4 Experiments

4.1 Datasets

Intent Classification Datasets. We collect four intent classification datasets. (1) **SNIPS-SLU** (Coucke et al., 2018), a widely used benchmark for English GZSL intent detection with 5 seen intents and 2 unseen intents. (2) **SMP-18** (Zhang et al., 2017), a Chinese dialogue corpus for user intent detection with 24 seen intents and 6 unseen intents. (3) **ATIS** (Hemphill et al., 1990), an English airline travel domain dataset, from which we extract 17 intents with at least 5 samples, and split them into 12 seen intents and 5 unseen intents. (4) **CLINC** (Larson et al., 2019) is a recently published intent detection dataset includes 22,500 in-scope queries covering 150 intent classes from 10 domains. We randomly split the 150 intents into 120 seen intents and 30 unseen intents.

Question Classification Dataset. In order to draw a comprehensive analysis of the proposed method, we construct a question classification task from the Quora Question Pairs dataset [1], which is aimed to identify duplicate questions. We collect questions with at least 5 duplicate samples into classes. In each class, we choose the question with minimum words as the label description, which is widely used in real-world question-answering systems (Sakata et al., 2019). Table 1 summarizes all datasets statistics. It is worth to note that intents in ATIS are highly unbalanced with *flight* accounts for about 87% of training data.

Dataset Settings. Following (Siddique et al., 2021), we use random seen/unseen classes for 10 runs instead of manual selection used in (Yan et al., 2020), which leads to more fair results because every class could be unseen class. We randomly take

[1] www.kaggle.com/c/quora-question-pairs

Table 1: Dataset statistics. "FS" indicates "few-shot", "BAL" indicates "balance", "IBAL" indicates "imbalance". The "avg #samples" indicates the average number of samples per class.

Dataset	#classes		#samples		sent	type
	seen	unseen	total	avg	len	
SNIPS	5	2	13802	1384	9.10	BAL
SMP	24	6	2460	60	4.83	FS
ATIS	12	5	4972	245	11.44	IBAL
Clinc	120	30	22500	105	8.23	BAL
Quora	1360	340	17394	7	10.46	FS

70% samples of each seen class as the training set, and the remaining 30% samples of each seen class as the seen test and take all the samples of unseen classes as the unseen test. All the textual labels of the same class are regarded as the description for this class.

4.2 Baseline Methods

To validate the benefits of the proposed LTA, we compare against with other approaches in three aspects:

Supervised Learning Methods. To show the performances on seen classes with supervised learning instead of ZSL/GZSL setting, we use (1) **BiLSTM** (Schuster and Paliwal, 1997) and (2) **BERT** (Devlin et al., 2018) as the encoder with a linear softmax classifier

Metric Learning Methods. Metric-based embedding methods are commonly used as baselines for ZSL/GZSL. Thus we introduce three different metric learning methods: (1) **EucSoftmax**: We adapt (Snell et al., 2017) that uses squared Euclidean distance as the metric and softmax classifies; (2) **Zero-shot DNN**: We adapt (Kumar et al., 2017) that uses squared Euclidean distance and triplet loss to maintain a margin for different classes. We choose the label embedding(prototype) as the anchor and the closest sample as negative sample in each triplet tuple; (3) **CosT** (Gidaris and Komodakis, 2018) refers to Cosine Distance with temperature scalar $s(a, b) = \tau cos(a, b)$ where τ is a learnable temperature scalar to dynamically control the peakiness of the probability distribution generated by the Softmax.

SOTA Methods. We also compare our model with two recent state-of-the-art (SOTA) methods: (1) **ReCapsNet** (Liu et al., 2019) uses a dimensional attention-based intent capsule network and a matrix transformation method for ZSL/GZSL. (2) **SEG** (Yan et al., 2020) is an outlier detection approach that can be directly applied on ReCapsNet. SEG acts as a domain discriminator which first determines whether a test sample belongs to seen classes or unseen classes and then classifies in their own domain. RIDE (Siddique et al., 2021) is not considered because they use outer knowledge not available in our settings and the data they used is limited within the intent detection task due that the labels in their method need to be a combination of "action" and "object".

4.3 Experimental Setup

Evaluation Metrics. We basically use accuracy (Acc) to estimate the performances on seen and unseen test sets. Besides, we adopt Macro-F1 (F1) rather than Micro-F1 to better evaluate the performances on imbalanced and few-shot datasets, because Macro-F1 gives the same weight of F1 scores for each class. For overall assessments, we adopt the widely used Harmonic Mean (HM) of Acc and F1 on Seen and Unseen, because the overall Acc and F1 scores are influenced by the ratio of seen and unseen test set sizes.

Implementation Details. We use pretrained BERT-base encoder with $d_h = 768$ on intent classification datasets and BiLSTM with 128 hidden size each direction on Quora dataset as basic encoder. The scalars of our model is set to be $\tau = 10.0, \beta = 10.0, d_a = d_h$, which is trained via Adam (Kingma and Ba, 2015) optimizer, with learning rates 10^{-5} for BERT, 10^{-4} for BiLSTM and 10^{-3} for the others. During training, we set $K = 5$ and $C^{u_i} = [2, 2, 2, 10, 20]$, $d_r = [4, 16, 32, 64, 64]$ for SNIPS, SMP, ATIS, CLINC and Quora datasets, respectively. The learnable $\mathbf{R}$ is initialized from the prototypes trained from metric learning methods, and is used as our basic baseline. We also conduct an ablation study to investigate the effectiveness of each proposed component. As depicted in Table 2 and Table 3, "w / o Init" refers to the model that randomly initialize $\mathbf{R}$. "w / o SA" refers to the model that only uses prototype adaptation without "Sample Adaptation". "w / o" means none of the adaptation steps is applied.

4.4 Results

The results on four intent datasets and Quora dataset are given in Table 2 and Table 3. Our

Table 2: Results (in %) on four intent benchmarks. The Top1 results of GZSL methods are highlighted in bold and underline for Top2 results, the same below.

Model	SNIPS-NLU						SMP-18					
	Seen		Unseen		HM		Seen		Unseen		HM	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Bi-LSTM	98.23	98.23	0.00	0.00	0.00	0.00	93.65	93.43	0.00	0.00	0.00	0.00
BERT	98.91	98.91	0.00	0.00	0.00	0.00	95.28	94.87	0.00	0.00	0.00	0.00
EucSoftmax	81.09	65.50	45.89	58.21	58.61	61.64	89.84	87.85	76.65	77.51	82.72	82.36
Zero-shot DNN	81.09	65.28	45.91	58.53	58.63	61.72	**90.97**	87.67	75.38	77.32	82.44	82.17
CosT	**91.68**	**75.76**	47.73	62.84	62.77	68.70	90.65	88.41	72.59	73.89	80.62	80.50
ReCapsNet	**96.26**	67.70	11.57	18.45	20.66	29.00	76.32	74.92	20.56	15.09	32.39	25.10
+ SEG	92.11	73.08	50.29	62.33	65.06	67.28	67.10	67.39	36.65	32.84	47.70	44.16
LTA (Ours)	74.05	74.11	**90.09**	**84.22**	**81.28**	**78.84**	89.84	**90.79**	79.19	75.20	84.18	82.26
w / o Init	82.57	75.22	64.36	71.63	72.34	73.87	89.03	87.23	**80.71**	**81.74**	**84.67**	**84.40**
w / o SA	67.31	70.56	84.70	77.51	75.01	73.87	84.52	81.40	75.89	74.40	79.97	77.75
w / o A	75.26	71.82	83.85	80.77	79.33	76.03	84.35	86.93	76.90	73.54	80.72	80.50
Model	**ATIS**						**CLINC**					
	Seen		Unseen		HM		Seen		Unseen		HM	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Bi-LSTM	93.24	79.51	0.00	0.00	0.00	0.00	92.07	92.06	0.00	0.00	0.00	0.00
BERT	97.18	93.71	0.00	0.00	0.00	0.00	97.37	97.37	0.00	0.00	0.00	0.00
EucSoftmax	67.67	16.11	7.78	5.50	13.96	8.20	96.02	87.07	58.02	66.00	72.33	75.08
Zero-shot DNN	63.56	23.12	8.05	12.02	14.29	15.82	95.31	86.65	58.49	65.89	72.49	74.68
CosT	**98.02**	59.55	46.04	45.21	62.66	51.40	**96.31**	87.33	62.73	70.28	75.98	77.89
ReCapsNet	86.19	23.88	12.80	4.89	22.32	8.12	88.53	69.83	4.24	3.33	8.10	6.36
+ SEG	**93.75**	40.90	14.78	6.36	25.53	11.01	81.04	78.89	9.07	5.44	16.31	10.18
LTA (Ours)	96.28	63.13	66.09	**55.02**	78.38	**58.80**	92.22	87.57	73.18	75.74	81.60	81.23
w / o Init	89.96	47.48	**69.79**	52.14	**78.60**	49.70	93.07	**88.19**	**73.80**	**77.54**	**82.32**	**82.52**
w / o SA	90.20	51.74	66.23	47.24	76.38	49.38	92.46	87.30	69.27	73.26	79.20	79.67
w / o A	94.94	**63.25**	57.52	49.19	71.64	55.34	93.81	88.12	70.11	74.58	80.25	80.79

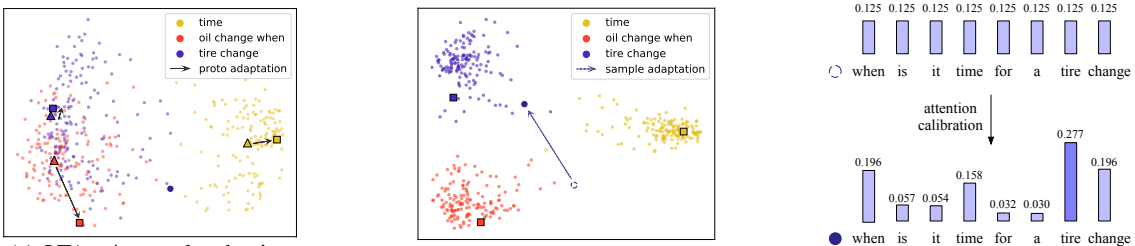
Table 3: Results (in %) on Quora question classification dataset.

Model	Seen		Unseen		HM	
	Acc	F1	Acc	F1	Acc	F1
BiLSTM	71.70	69.04	0.00	0.00	0.00	0.00
EucSoftmax	79.88	74.42	56.85	62.39	66.43	67.88
Zero-shot DNN	72.52	67.42	48.68	53.27	58.26	59.52
CosT	**88.50**	81.39	62.21	73.55	73.06	77.27
LTA (Ours)	84.69	**83.56**	74.83	**76.93**	**79.45**	**80.11**
w / o Init	82.11	81.99	**75.49**	76.53	78.66	79.17
w / o SA	84.95	82.79	73.56	76.67	78.84	79.62
w / o A	84.21	82.40	72.50	75.23	77.92	78.65

proposed methods achieve the overall best performances compared to baselines.

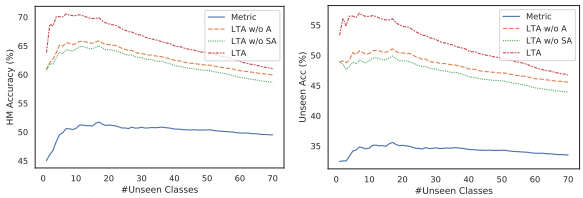
Detailed and interesting observations can also be derived from the results: (1) Supervised Metric-Learning methods as the basic baselines, achieve comparable results on *Seen Test* for all datasets. However, it suffers from the domain bias problem and the performance drops with a large margin on *Unseen Test*, where the task is complex due to the imbalanced and few-shot scenarios. (2) The performances on SNIPS-NLU and SMP-18 of ReCapsNet and SEG are worse than those in their original paper although we use the open-source code, this is because we random split the test unseen classes which makes it more challenging. Besides, these methods fail to recognize unseen samples well on datasets with large scale of categories, yielding worse 0% Acc and F1 on Quora. The most likely reason is that ReCapsNet uses label embedding similarities to construct unseen prototypes in capsule network, which imposes a non-trivial computational and memory burden. (3) Our method shows its privilege for all datasets. In particular, with the help of continuous adapting ability, it observes smaller gap between seen and unseen domain, which proves the adaptation on testing phase effectively works. Although the performance on seen domain drops sightly, the proposed LTA outperforms the competitive metric-learning baselines by 9.54% HM Acc and 12.90% HM F1 average on the whole datasets, indicating that our model fairly balances the seen and unseen classes.

Ablation Study. To better understand the contribution of each component of our method, we explore three variants of LTA. We can observe that LTA with both prototype adaptation and sample adaptation outperforms those without adaptations in all cases. Generally "LTA w / o SA" with only prototype adaptation achieves better performance compared to "LTA w / o A". The "LTA w / o Init" has relatively stable performances.



(a) LTA w/o sample adaption (b) LTA (c) example

Figure 2: PCA plots of encoded unseen sample representations(·) and prototype representations(■) from (a) LTA w/o sample adaption model and (b) full LTA model with sample adaptation (c) is an unseen example with sample-level raw attention and adapted attention. ▲ denotes the raw prototype before adaptation. ○ and ● respectively denote the example representations before and after sample adaptation.



(a) HM Acc. (b) Unseen Acc.

Figure 3: The performance with different numbers of unseen classes on CLINC dataset.

4.5 Results on Emerging Unseen Classes

As the partition of seen / unseen classes is fixed in previous experiments, in order to study the robustness of the proposed adaption method, we conduct the experiment across unseen class sets of different scales. Specifically, we select 70 classes as seen classes and 10 classes as validating unseen classes. The testing unseen classes are randomly sampled from the remaining 70 classes and each experiment is repeated 50 times with different sampling sets for a more stable result. Figure 3 (a) shows the HM accuracy on all classes as the number of the unseen classes grows. We can see that our LTA model outperforms the metric baseline and ablation models in all cases, where the well performance is mainly attributed to the improvements on unseen classes as shown in Figure 3 (b). These results suggest that our adaptation method is robust and effective for adapting to new classes as well as improving the overall performance on all classes.

4.6 Visualization

To demonstrate how our adaptation method works, we further visualize the encoded representation via PCA in Figure 2. When there is no unseen class, seen classes (yellow and red) is discriminative enough. But when the new class "tire change" (purple) comes, it is ambiguous with class "oil change when" (red). We observe that the seen and unseen class prototypes are updated to be far away from each other after prototype adaptation as shown in (a), which eases the domain bias problem. However, the performance is unsatisfactory since the sample representations are still not discriminative no matter how prototype updates. As we can see, with the sample adaptation as shown in (b), the sample representations are independently clustered by the adapted prototype and easy to to be distinguished.

To further study how the sample adaptation works, we select an representative case "when is it time for a tire chance" and show its attention weights used as calibration parameters in (c). The case is still misclassified after the prototype adaptation due to the common word "time" and "change" also appear in seen classes. After the sample adaptation, however, it can be seen that the word "tire" which is a key word for classifying, get the most attention while the other confusing words are not. This result suggests that calibrating the attention weights is useful for acquiring a prototype aware representation which helps the sample adaptation.

5 Conclusion

This paper proposed a novel adaptive meta-learning network for generalized zero-shot text classification. The model was trained under a consistent setting with testing. In particular, it efficiently alleviated the bias towards seen classes by utilizing both prototype adaptation and sample adaptation. Experiments on five text classification datasets validated that our model achieved compelling results on both seen classes and unseen classes, meanwhile

was capable of fast adapting to new classes.

References

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *International Conference on Learning Representations*.

Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. 2018. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *CVPR*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.

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

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 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia. PMLR.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, MarcAurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, volume 26, pages 2121–2129. Curran Associates, Inc.

Hao Fu, Caixia Yuan, Xiaojie Wang, Zhijie Sang, Shuo Hu, and Yuanyuan Shi. 2018. Zero-shot question classification using synthetic samples. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 714–718. IEEE.

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.

Charles T. Hemphill, John Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Proceedings DARPA Speech and Natural Language Workshop*, pages 96–101, Hidden Valley, PA. Morgan Kaufmann.

Johannes Fürnkranz Jinseok Nam, Eneldo Loza Mencía. 2016. All-in text: Learning document, label, and word representations jointly. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Anjishnu Kumar, Pavankumar Reddy Muddireddy, Markus Dreyer, and Björn Hoffmeister. 2017. Zero-shot learning across heterogeneous overlapping domains. In *INTERSPEECH*, pages 2914–2918.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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)*.

Han Liu, Xiaotong Zhang, Lu Fan, Xuandi Fu, Qimai Li, Xiao-Ming Wu, and Albert YS Lam. 2019. Reconstructing capsule networks for zero-shot intent 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 4801–4811.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40.

Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, and Xi-Zhao Wang. 2020. A review of generalized zero-shot learning methods.

Shafin Rahman, Salman H. Khan, and Nick Barnes. 2019. Transductive learning for zero-shot object detection. In *ICCV*, pages 6081–6090. IEEE.

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

Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2152–2161, Lille, France. PMLR.

Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. FAQ retrieval using query-question similarity and bert-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1113–1116. ACM.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. 2019. Generalized zero- and few-shot learning via aligned variational autoencoders. In *CVPR*, pages 8247–8255. Computer Vision Foundation / IEEE.

Xiahan Shi, Leonard Salewski, Martin Schiegg, Zeynep Akata, and Max Welling. 2019. Relational generalized few-shot learning. *arXiv preprint arXiv:1907.09557*.

Qingyi Si, Yuanxin Liu, Peng Fu, Jiangnan Li, Zheng Lin, and Weiping Wang. 2020. Learning disentangled intent representations for zero-shot intent detection.

AB Siddique, Fuad Jamour, Luxun Xu, and Vagelis Hristidis. 2021. Generalized zero-shot intent detection via commonsense knowledge. *arXiv preprint arXiv:2102.02925*.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, volume 30, pages 4077–4087. Curran Associates, Inc.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26:935–943.

Congzheng Song, Shanghang Zhang, Najmeh Sadoughi, Pengtao Xie, and Eric Xing. 2020. Generalized Zero-Shot Text Classification for ICD Coding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4018–4024, Yokohama, Japan. International Joint Conferences on Artificial Intelligence Organization.

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*.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, page 5998–6008. Curran Associates, Inc.

Vinay Kumar Verma, Dhanajit Brahma, and Piyush Rai. 2020. Meta-learning for generalized zero-shot learning. In *AAAI*, pages 6062–6069.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, volume 29, pages 3630–3638. Curran Associates, Inc.

Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2).

Xiaolong Wang, Yufei Ye, and Abhinav Gupta. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3090–3099. Association for Computational Linguistics.

Yongqin Xian, Christoph Lampert, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP.

Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. 2018. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551. IEEE Computer Society.

Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. 2019. f-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR Workshops*, pages 46–49. Computer Vision Foundation / IEEE.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert Y.S. Lam. 2020. Unknown intent detection using Gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1050–1060, Online. Association for Computational Linguistics.

Han-Jia Ye, Hexiang Hu, and De-Chuan Zhan. 2019. Learning adaptive classifiers synthesis for generalized few-shot learning. *CoRR*, abs/1906.02944.

Zhiquan Ye, Yuxia Geng, Jiaoyan Chen, Jingmin Chen, Xiaoxiao Xu, SuHang Zheng, Feng Wang, Jun Zhang, and Huajun Chen. 2020. Zero-shot text classification via reinforced self-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3014–3024, Online. Association for Computational Linguistics.

Yunlong Yu, Zhong Ji, Jungong Han, and Zhongfei Zhang. 2020. Episode-Based Prototype Generating Network for Zero-Shot Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14032–14041, Seattle, WA, USA. IEEE.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long Papers)*, Minneapolis, USA. Association for Computational Linguistics.

Weinan Zhang, Zhigang Chen, Wanxiang Che, Guoping Hu, and Ting Liu. 2017. The first evaluation of chinese human-computer dialogue technology. *CoRR*, abs/1709.10217.

Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. 2019. Generalized zero-shot recognition based on visually semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.