DESIRE-Net: Deep Semantic Inference Network for Multi-label Few-shot Learning

Zhen Wang, Yiqun Duan, Liu Liu, Dacheng Tao
UBTECH Sydney AI Centre, School of Computer Science,
Faculty of Engineering, The University of Sydney,
Darlington, NSW 2008, Australia
zwan4121@uni.sydney.edu.au

Abstract

Few-shot learning can adapt the classification model to new labels with only a few labeled examples, which is experiencing rapid development and advancements. Previous studies mainly focus on the scenario of a single category label per example but have not solved the more challenging multi-label scenario with exponential-sized output space and low-data effectively. In this paper, we propose a semantic-aware meta-learning model, Deep Semantic Inference Network (DESIRE-Net), for multi-label few-shot learning. DESIRE-Net can learn and infer the semantic correlation between unseen labels and historical labels to quickly adapt multi-label tasks from only a few examples. Specifically, DESIRE-Net maps features into the semantic embedding space via label word vectors (learned from unsupervised text corpora) to explore and exploit the label correlation, and thus cope with the challenge on the overwhelming size of the output space. Then a novel semantic inference mechanism is designed for leveraging prior knowledge learned from historical labels, which will produce good generalization performance on new labels to alleviate the low-data problem. Finally, extensive empirical results show that the proposed method significantly outperforms the existing state-of-the-art methods on the multi-label few-shot learning tasks.

1 Introduction

Humans have an incredible ability to learn new concepts quickly from very few examples by leveraging prior knowledge and experience [1, 2]. Although current deep learning technologies outperform humans in certain primary tasks, they still rely on a tremendous number of annotated training examples, which is far from satisfactory compared to human [3, 4]. Thus, few-shot learning [5, 6] is proposed to facilitate deep learning systems to learn new concepts with very limited labeled data.

Most of the latest few-shot learning works [7, 8, 9, 10] are based on meta-learning (*learning-to-learn*) paradigm. Meta-learning is a task-level learning framework that aims to accumulate knowledge from learning a large number of tasks and generalize the knowledge to learn new tasks effectively. Previous works in this area could be categorized as *metric-based* and *gradient-based* approaches. Metric-based approaches [5, 11, 12, 13, 14] solve the "few-shot" problem by learning a feature space where category labels can be distinguished from each other based on distance metrics. Gradient-based methods [9, 10, 15, 16, 17] aim at acquiring more robust and generalized model parameters which could be adapted to given new tasks within a few optimization updates.

Nevertheless, most of the previous works only focus on the scenario where each example is associated with exclusive single label, but ignore the more actual and challenging scenario , in which each example can be simultaneously associated with multiple labels. To our best knowledge, only very

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

few papers [18, 19] have explored the multi-label few-shot learning (ML-FSL) problem, and most of them are far from satisfactory. For example, ZAG-CNN [18] focus on handling infrequent *long-tail* examples based on a text classification dataset. However, the proposed algorithm is not specially designed for ML-FSL. LaSO [19] proposes a data-augment technique to generate different label combinations, but LaSO has not applied the technique to solve ML-FSL problem.

The key challenge of multi-label few-shot learning is the huge output space, where the number of possible label sets exponentially grows with the increasing number of category labels. The huge output space simultaneously denotes a much sparser learning target and will bring difficulties in model learning for ML-FSL. Moreover, considering the fact that few-shot learning requires the model to be optimized on a relatively small dataset [20], a severe over-fitting problem will arise, which would also be aggravated by the huge output space as well.

In this paper, we propose a gradient-based meta-learning framework, DEep Semantic InfeREnce Network (DESIRE-Net), to solve the ML-FSL problem. Different from previous meta-learning approaches, we utilize word embedding vectors instead of one-hot vectors as our prediction output. The semantic embedding output naturally brings correlation to predictions, e.g., label "cake" is semantically close to label "food" naturally in word embedding space, which gives us a tighter learning object. More specifically, DESIRE-Net trains a semantic-aware *base model* that maps features into the semantic embedding space via label word vectors learned from unsupervised text corpora. That is, the semantic correlation across labels can be preserved by the base model, which helps to structure and regularize the overwhelming output space of ML-FSL. Furthermore, we propose a meta-learning framework with a **semantic inference** mechanism that can extract semantic features and exploit the correlation between novel labels and historical labels as prior knowledge to classify multiple labels only using a few examples effectively. The semantic inference mechanism has three different levels: feature-level, correlation-level, and attention-level, and two functions: training better initialization parameters of model by leveraging the knowledge learned from historical tasks; inferring the classification of novel labels according to the semantic correlation with historical labels. With the help of semantic inference, the proposed model could reach the state-of-the-art performance on ML-FSL.

The contributions of the paper are threefold. 1) We propose a meta-learning framework, DESIRE-Net, to solve the multi-label few-shot learning problem. Experimental results suggest that our model achieves state-of-the-art performance, and ablation studies validate each module's effectiveness[1]. 2) We propose a novel semantic-aware base-label classifier that embeds features into semantic space to facilitate the model to learn the semantic correlation across labels. 3) We propose a semantic inference mechanism for leveraging the meta knowledge learned from historical tasks to effectively adapt new multi-label tasks.

2 Related Work

The goal of few-shot learning is that adapting the model to predict new labels with only a few labeled examples effectively. Early works build a Bayesian model with prior knowledge learned from previous labels that can be transferred to new labels [6, 21]. Recently, meta-learning (*learning-to-learn*) [7, 8, 9, 10] framework has been applied to few-shot learning problem and exerted significant impacts by training a meta-learner with few-shot tasks sampled from training data set. Meta-learning methods for few-shot learning problem can be mainly categorized into two groups: metric-based methods and gradient-based methods.

Metric-based methods [5, 11, 12, 13, 14] train a model to embed examples into a metric space where examples with same label are gathered closely, and examples with different labels are spread far away. How to define and learn the similarity among data examples in the metric space is the key for this strategy. For instance, Matching Network [5] can distinguish different labels by a weighted nearest neighbor classifier. Prototypical Network [11] can firstly train an average value of the features belonged the same label in the metric space as the prototype, and then perform nearest neighbor classification. The works [13, 14] further generalize prototypical network with a task adaptive metric [13] and a representation transfer [14]. Sung et al. [12] propose to use relation networks to learn the distance metric for the images within each task. All these metric-based methods can allocate an example to a single label through the maximum similarity (the nearest neighbor), and they can not

[1] Anonymous code and models are available on `https://github.com/DESIRE-Net/DESIRE-Net`

treat the example associated with multi-label, e.g., an image including dog, cat, and trees, etc., can be identified simultaneously.

Gradient-based methods [9, 10, 15, 16, 17, 22] aim at training parameters of models that can be well adapted to novel tasks with only a few optimization updates. Finn et al. [9] propose a model-agnostic meta-learning (MAML) framework. In order to improve the efficiency and performance of MAML on few-shot learning, many follow-up works built on top of MAML [23, 24]. Reptile [15] simply replace the second-order gradient information with the first-order gradient computation of MAML. MAML++ [17] further improve the generalized performance and stabilize the system. ATAML [16] is designed to encourage task-agnostic representation learning with attention mechanisms. LEO [22] employs a low-dimensional latent space to learn the model parameters. Current approaches mentioned above only mainly focus on multi-class (single-label) few-shot learning, but ignores the multi-label problem. Although gradient-based approaches can be easily adapted to use in multi-label problem by converting 1-hot output to n-hot output, the challenges brought by the exponential-sized output space and more serious problem such as low-data will severely restrict the performance of meta-learner.

Multi-label few-shot learning (ML-FSL) aims to adapt the model to new multi-label classification tasks with only a few labeled examples. Up to now, only a few works have been attempted to address this challenging problem. ZAG-CNN [18] aims at solving *long-tail* dataset with infrequent labels (few-shot data), but has not addressed a complete ML-FSL problem. LaSO [19] points out the difficulties in ML-FSL and then proposes a data-augment technique to generalize different label sets. However, LaSO only verifies the validity of data-augment and can not be applied to solve ML-FSL. In order to fill in the above mentioned research gaps, this paper propose a gradient-based meta-learning model, Deep Semantic Inference Network (DESIRE-Net), to solve ML-FSL effectively.

3 Methodology

In this section, after giving the mathematical definition of the proposed multi-label few-shot learning (ML-FSL) problem, the Deep Semantic Inference Network (DESIRE-Net), including a semantic-aware base-label classifier and a semantic inference mechanism is introduced in detail.

3.1 Multi-label Few-shot Learning

ML-FSL aims to learn a model that can be well adapted to novel multi-label tasks using only a few annotated examples. We formulate ML-FSL as a meta-learning problem and adopt the episode training mechanism by following the convention [5, 9, 11, 16, 17, 22]. Generally, in an N-way K-shot episode setting, each task $\mathcal{T}$ is formed by first sampling N labels from $\mathcal{D}_{meta}$ and then sampling two sets of examples associated with these labels: 1) the *support* set $\mathcal{S}_{\mathcal{T}} = \{(\boldsymbol{x}_S^i, \boldsymbol{y}_S^i)\}_{i=1}^{N\times K}$ containing K examples sampling from each of the N labels, and 2) the *query* set $\mathcal{Q}_{\mathcal{T}} = \{(\boldsymbol{x}_Q^i, \boldsymbol{y}_Q^i)\}_{i=1}^{Q}$ containing a fraction of the rest examples from the same N labels. Here, $\boldsymbol{y}^i \in \mathcal{Y} \subseteq \{0,1\}^{N\times 1}$ denotes the multi-label vector, $\boldsymbol{x}^i \in \mathcal{X} \subseteq \mathbb{R}^{d_s \times 1}$ denotes the corresponding example point, and d_s is the dimension of the example. If the j-th label is assigned to example $\boldsymbol{x}^i$, we have $y^{i,j} = 1$, otherwise $y^{i,j} = 0$. The proposed meta-learning problem is to build a mapping f_θ (known as meta-learner) from the support set and the query set examples to the query set labels.

The resulting meta-learner objective is to minimize, over all tasks, the expected loss of the prediction on examples in query set, given the support set:

$$\theta = \arg\min_{\theta} \mathop{\mathbb{E}}_{\mathcal{T}\sim\mathcal{D}_{meta}} [\![\mathcal{L}(\mathcal{T};\theta)]\!], \tag{1}$$

where

$$\mathcal{L}(\mathcal{T};\theta) = \mathop{\mathbb{E}}_{\mathcal{S}_{\mathcal{T}},\mathcal{Q}_{\mathcal{T}}} \left[\!\!\left[\sum_{q=1}^{|\mathcal{Q}_{\mathcal{T}}|} \ell\left(\boldsymbol{y}_Q^q, f_\theta(\boldsymbol{x}_Q^q, \mathcal{S}_{\mathcal{T}})\right) \right]\!\!\right], \tag{2}$$

$(\boldsymbol{x}_Q^q, \boldsymbol{y}_Q^q) \in \mathcal{Q}_{\mathcal{T}}$ and $\mathcal{S}_{\mathcal{T}}$ are, respectively, the query and support set sampled from $\mathcal{D}_{meta}$, θ are the parameters of the model, and $\ell(\cdot)$ denotes the loss function.

In meta-learning framework, we typically have different meta-sets $\mathcal{D}_{meta-train}$, $\mathcal{D}_{meta-val}$, and $\mathcal{D}_{meta-test}$, for meta-training, meta-validation, and meta-testing, respectively. The label sets are disjoint among these meta-sets, where the labels belonged to $\mathcal{D}_{meta-train}$ are called **base labels** and

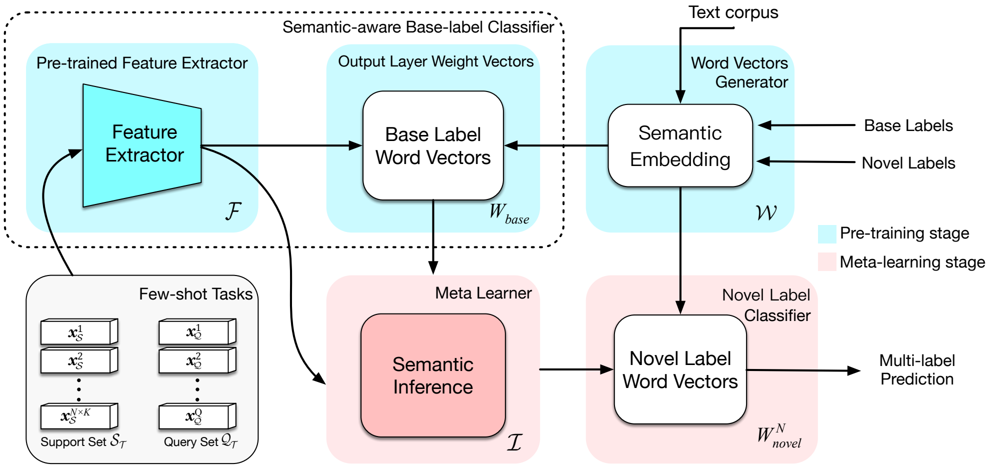

Figure 1: Deep Semantic Inference Network for multi-label few-shot learning. In the pre-training stage, we learn a semantic-aware base-label classifier with a feature extractor to embed features into the semantic space. In the meta-learning stage, we train a meta-learner with a semantic inference mechanism on different few-shot tasks to incorporate semantic knowledge. In the testing stage, we evaluate the model on novel tasks.

the labels belonged to $\mathcal{D}_{meta-test}$ are called **novel labels**. The meta-learner is trained on the tasks sampled from $\mathcal{D}_{meta-train}$ by minimizing Equation (2). We can select the hyper-parameters of the meta-learner on $\mathcal{D}_{meta-val}$ and finally evaluate generalization performance on $\mathcal{D}_{meta-test}$.

3.2 Deep Semantic Inference Network

Figure 1 shows the structure of DESIRE-Net. As a preliminary preparation, we train a word-embedding model $\mathcal{W}$ on unsupervised text corpora, e.g., Google News [25]. We employ $\mathcal{W}$ to produce label word embeddings of all labels in $\mathcal{D}_{meta-train}$ and $\mathcal{D}_{meta-test}$. Our model consists of two main components. First, we train a semantic-aware base-label classifier on $\mathcal{D}_{meta-train}$, where the output matrix of the classifier is replaced by word embedding vectors of all base labels. Thus, as a part of the classifier, the feature extractor could embed features into semantic space, which simultaneously facilitate the model to learn the semantic correlation between labels. Second, a gradient-based semantic inference mechanism is designed for leveraging the prior knowledge learned from base labels to adapt new multi-label tasks on $\mathcal{D}_{meat-test}$.

3.2.1 Semantic-aware Base-label Classifier

A key challenge of multi-label few-shot learning lies in the overwhelming size of the output space. Unlike traditional multi-class (single-label) few-shot learning that outputs one label for one example, the number of labels associated with an example in multi-label few-shot learning is uncertain, i.e., the possible number of label combinations increases exponentially as the number of labels grows. For example, for a label space with 20 category labels, the number of possible label combinations could exceed one million (i.e., 2^{20}).

To solve the challenge of exponential-sized output space, we build a semantic-aware base-label classifier within the consideration of correlations (dependencies) among labels. We train a DNN-based network on $\mathcal{D}_{meta-train}$ as the base-label classifier which consists of a feature extractor $\mathcal{F}$ and a output matrix W_{base} as shown in Figure 1. $W_{base} \in \mathbb{R}^{d\times m}$ is assigned and fixed to base-label word vectors learned from unsupervised text corpora, where d is the dimension of the word vector and m is the number of base label on $\mathcal{D}_{meta-train}$. The classifier is trained with sigmoid activation function and cross-entropy loss $\mathcal{L}_{CE}$ as shown in equations:

$$\hat{\boldsymbol{y}} = \text{sigmoid}(\mathcal{F}(\boldsymbol{x})W_{base}) \ , \tag{3}$$

$$\mathcal{L}_{CE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\sum_{j=1}^{m} \left[y^j \log(\hat{y}^j) + (1 - y^j) \log(1 - \hat{y}^j) \right] \ , \tag{4}$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}_{meta-train}$, $\hat{\boldsymbol{y}} = [\hat{y}^1, ..., \hat{y}^m]$ is the output label vector corresponding to the example $\boldsymbol{x}$, $\hat{y}^j$ is the prediction of jth label, and $\boldsymbol{y} = [y^1, ..., y^m]$ indicates the ground truth.

Through training semantic-aware base-label classifier, the features produced by feature extractor would be embedded into a semantic space that incorporates semantic similarity between labels. Specifically, the distance between features with similar semantic meaning in label-wise would be closer. In that case, the correlation across labels is preserved by the feature extractor, which naturally makes the output space of ML-FSL tighter, thus alleviate the overwhelming output space problem.

3.2.2 Meta-learner with Semantic Inference

Meta-learning methods for few-shot learning problem are to train a meta-learner that learns on a large number of different tasks to catch the generalized knowledge. We propose a novel meta-learner with a *semantic inference* mechanism, which can extract semantic features and exploit the correlation between novel labels and base labels as prior knowledge to classify multiple labels only using a few examples effectively. The semantic inference mechanism contains three different levels: feature-level, correlation-level, and attention-level. It will be described in detail as follows.

Assume $W_{base} = [w_{base}^1, ..., w_{base}^m] \in \mathbb{R}^{d \times m}$ is the m base label word vectors matrix, $W_{novel}^N = [w_{novel}^1, ..., w_{novel}^N] \in \mathbb{R}^{d \times N}$ is constitutive of N novel label word vectors, and a feature $\boldsymbol{z} \in \mathbb{R}^{1 \times d}$ is produced by a feature extractor $\mathcal{F}$, i.e., $\boldsymbol{z} = \mathcal{F}(\boldsymbol{x})$. Note that base labels $L_{base} \in \mathcal{D}_{meta-train}$ and novel labels $L_{novel} \in \mathcal{D}_{meta-test}$ are disjoint.

Feature-level. Since we have obtained features embedded into the semantic space as described in section 3.2.1, the output feature would naturally be closer to those with similar semantic meanings. In that case, we propose to infer our predictions by feature-level proximity between $\boldsymbol{z}$ and novel label word vectors W_{novel}^N. That is, to learn the task-specific knowledge, we build a fully connected network $\mathcal{W}_Z$ to obtain transformation from $\boldsymbol{z}$, and combine the transformation with novel label vectors. The feature-level semantic inference I_f takes the form:

$$I_f(\boldsymbol{z}) = \frac{\mathcal{W}_Z(\boldsymbol{z})}{\|\mathcal{W}_Z(\boldsymbol{z})\|_2} W_{novel}^N \ , \tag{5}$$

where $\mathcal{W}_Z(\boldsymbol{z})$ denotes a nonlinear transformation for $\boldsymbol{z}$, and $\|\cdot\|_2$ denotes l_2-norm which can eliminate the influence of the absolute magnitudes of semantic features and improve the robustness.

Correlation-level. Intuitively, we find that the semantic correlation between novel labels and base labels can help model better adapt to a novel task. For example, a feature is considered to be a "horse" in the base label space, and if we know the correlation between "horse" and a novel label "zebra", then this can help us to classify the feature correctly in the novel task. Generally, we use $[w_{base}^1, ..., w_{base}^m]^T w_{novel}^j$ to represent the correlation of the j-th novel label in terms of m base labels. In order to learn deep correlation between novel labels and base labels, we define an explicit correlation-level semantic inference mechanism I_c as:

$$I_c(\boldsymbol{z}) = \frac{\boldsymbol{z} W_{base}}{\|\boldsymbol{z} W_{base}\|_2} W_I W_{novel}^N \ , \tag{6}$$

where $W_I \in \mathbb{R}^{d \times m}$ is a learnable matrix trained on different tasks to learn deep correlation between novel and base labels. While W_I is a linear matrix, the l_2-norm can offer a non-linear operation. Since the number of base labels is different between meta-training and meta-testing (see section 3.3), l_2-norm can also limit the adverse effects of absolute magnitudes fluctuations of $\boldsymbol{z} W_{base}$.

Attention-level. Besides, we design an attention-level semantic inference to transfer the probabilistic predictions on base labels to novel labels. A basic idea of this mechanism is using output score of base labels to compute a weighted combination of novel labels embeddings in the semantic space. Given example $\boldsymbol{x}$ and base-label classifier, we can get a label prediction vector $\hat{\boldsymbol{y}} = [\hat{y}^1, ..., \hat{y}^m]$, where $\hat{y}^j$ is the probabilistic prediction for j-th label. Consequently, $\hat{y}^j w_{base}^j$ is the probabilistic weighted word vector for j-th base label. More formally, by define the convex combination of the base-label word embeddings:

$$e(\boldsymbol{x}) = \sum_{j=1}^{m} \hat{y}^j w_{base}^j = W_{base} \hat{\boldsymbol{y}} = W_{base} \text{sigmoid}(\mathcal{F}(\boldsymbol{x}) W_{base}), \tag{7}$$

Algorithm 1 Deep Semantic Inference Network: Meta Training

```
Require: Meta-train set $D_{meta-train}$
Require: Learning rates $\alpha$, $\beta$
Train feature extractor $\mathcal{F}$ on $\mathcal{D}_{meta-train}$
Initialize $\theta$ and $\boldsymbol{\gamma} = \{\gamma_f, \gamma_c, \gamma_a\}$          ▷ Initialize all parameters
while not done do
    Sample batch of tasks $\mathcal{T}_i \sim \mathcal{D}_{meta-train}$          ▷ Sample tasks for meta-training
    Let $(\mathcal{S}_{\mathcal{T}_i}, \mathcal{Q}_{\mathcal{T}_i}) = \mathcal{T}_i$          ▷ Get support set and query set
    for all $\mathcal{T}_i$ do
        $\theta'_i = \theta - \alpha \nabla_\theta \mathcal{L}(\mathcal{S}_{\mathcal{T}_i}; \{\theta, \boldsymbol{\gamma}\})$          ▷ Compute temporary parameters
    end for
    Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i} \mathcal{L}(\mathcal{Q}_{\mathcal{T}_i}; \{\theta'_i, \boldsymbol{\gamma}\})$          ▷ Update network parameters
    Update $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma} - \beta \nabla_{\boldsymbol{\gamma}} \sum_{\mathcal{T}_i} \mathcal{L}(\mathcal{Q}_{\mathcal{T}_i}; \{\theta'_i, \boldsymbol{\gamma}\})$          ▷ Update factor parameters
end while
```

we could treat $e(\boldsymbol{x})$ as a region in the semantic space. Thus, the closer novel label embedding w^j_{novel} near the region, the higher probability of $\cos(e(\boldsymbol{x}), w^j_{novel})$ will be, where we use cosine function to represent similarity. In that case, we can generate a probability prediction of $\boldsymbol{x}$ on the corresponding novel labels through this simple inference, i.e., $\cos\left(e(\boldsymbol{x})W^N_{novel}\right) = \cos\left(e(\boldsymbol{x})\left[w^1_{novel}, ..., w^N_{novel}\right]\right)$. However, the simple inference mentioned above have not taken the full utilization of task-specific knowledge. In order to improve the learning ability of the model, we further design the attention-level semantic inference mechanism based on above basic idea. Learnable weights are added to the attention framework as:

$$q(\boldsymbol{z}) = \boldsymbol{z}W_q, \quad \mathbf{K} = W^T_{base}W_K, \quad \mathbf{V} = W^T_{base}W_V, \tag{8}$$

where $W_q, W_K, W_V \in \mathbb{R}^{d\times d}$ are learnable linear matrices, $\boldsymbol{z} = \mathcal{F}(\boldsymbol{x})$. Then we can get the semantic attention value,

$$a(\boldsymbol{z}) = \frac{\text{sigmoid}\left(\frac{q(\boldsymbol{z})\mathbf{K}^T}{\tau}\right)}{\left\|\text{sigmoid}\left(\frac{q(\boldsymbol{z})\mathbf{K}^T}{\tau}\right)\right\|_1}\mathbf{V}\,, \tag{9}$$

where $\|\cdot\|_1$ is l_1-norm, and τ is a temperature. Different from the traditional attention mechanism, we use the sigmoid function to compute the probability, rather than softmax function, causing that the sum of the probabilities is not 1. Hence, we perform l_1-norm to regularize different multi-label tasks. Next, we build a network $\mathcal{W}_A$ to learn and figure out the relationship between the attention and novel label semantics. The attention-level semantic inference I_a can be formed as:

$$I_a(\boldsymbol{z}) = \mathcal{W}_A\left(\frac{a(\boldsymbol{z})}{\|a(\boldsymbol{z})\|_2}\right)W^N_{novel}\,. \tag{10}$$

Based on the above description of semantic inference at three levels, we can summarize the final meta-learning model with semantic inference: $\mathcal{I}(\boldsymbol{z}) = \gamma_f I_f(\boldsymbol{z}) + \gamma_c I_c(\boldsymbol{z}) + \gamma_a I_a(\boldsymbol{z})$, where γ_f, γ_c, and γ_a are the learnable module factors.

3.3 Meta Training and Meta Testing

In this section, we provide meta training and testing procedures of our proposed framework.

Meta Training. The training algorithm is described in Algorithm 1. We use θ to denote the parameters of the meta-learner including those of learnable matrices and networks, and use $\boldsymbol{\gamma} = \{\gamma_I, \gamma_A, \gamma_Z\}$ to denote the learnable factors of each module. It is noted that the parameters of trained feature extractor are frozen. In meta training, each N-way K-shot task $\mathcal{T}_i = \{\mathcal{S}_{\mathcal{T}_i}, \mathcal{Q}_{\mathcal{T}_i}\}$ is sampled from $\mathcal{D}_{meta-train}$.

The goal of meta-learner is to obtain good parameters for θ and $\boldsymbol{\gamma}$ that can be adapted to different tasks using a few optimization steps. To achieve this goal, gradient-based methods minimize the loss of the prediction on query set $\mathcal{Q}_{\mathcal{T}_i}$, given the support set $\mathcal{S}_{\mathcal{T}_i}$. First, we compute a *inner* loss on support set $\mathcal{S}_{\mathcal{T}_i}$ to update the parameters θ and get a temporary parameters θ'_i.

$$\theta'_i = \theta - \alpha\nabla_\theta\mathcal{L}(\mathcal{S}_{\mathcal{T}_i}; \{\theta, \boldsymbol{\gamma}\}), \tag{11}$$

Table 1: Comparing multi-label few-shot classification performance on Delicious

Method	5-way Micro-F1		10-way Micro-F1		5-way Macro-F1		10-way Macro-F1	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
pre-trained	27.8%	37.3%	17.6%	26.9%	19.1%	29.2%	12.7%	19.4%
MAML	34.1%	48.2%	28.4%	38.5%	27.9%	41.7%	22.9%	33.9%
Reptile	35.7%	51.4%	30.2%	40.4%	28.3%	42.1%	24.0%	35.1%
ATAML	43.6%	52.5%	39.5%	42.1%	36.5%	43.2%	33.8%	36.7%
MAML++	38.3%	53.1%	33.0%	43.3%	31.8%	46.0%	27.1%	38.3%
LEO	51.1%	60.7%	45.1%	51.7%	44.0%	53.5%	38.0%	45.9%
DESIRE-Net	**62.5%**	**68.8%**	**52.6%**	**57.6%**	**60.3%**	**66.9%**	**47.7%**	**52.9%**

Then, an *outer* loss is computed using θ'_i and γ on query set $\mathcal{Q}_{\mathcal{T}_i}$ to update meta-learner parameters θ_i and γ. Different from the traditional gradient-based method [9], part of parameters, factors γ, in our model can be reused across all tasks without being adapted on the support set. The novel training paradigm provides regularization of meta-learning that further improves generalization.

An interesting setting in our training is that we pick "fake" novel labels from base labels $L_{base} \in \mathcal{D}_{meta-train}$ to learn how to treat actual novel labels $L_{novel} \in \mathcal{D}_{meta-test}$. Specifically, for one training task, we sample N labels from L_{base} as "fake" novel labels, and then take these N labels out of L_{base}. Accordingly, we need to dynamically eliminate corresponding N of m vectors in the inference matrix W_I during training. Everything is trained end-to-end.

Meta Testing. Meta testing aims to test the performance of the trained DESIRE-Net for fast adaptation to novel tasks sampled from $\mathcal{D}_{meta-test}$. Given $\mathcal{T}_{novel} = \{\mathcal{S}, \mathcal{Q}\}$, we fine-tune the DESIRE-Net on $\mathcal{S}$, and test on $\mathcal{Q}$.

4 Experiments

In this section, we conduct experiments to illustrate that our proposed method could outperform the previous state-of-the-art methods. Ablation studies are also conducted to give corresponding analysis to our framework.

4.1 Experimental Setup

We conduct experiments on a widely used multi-label dataset, Delicious[2], which has 983 labels and 16105 examples. We removed 8 punctuation labels without semantics and split the dataset into three parts where labels are mutually disjoint: $\mathcal{D}_{meta-train}$ including 600 labels, $\mathcal{D}_{meta-val}$ including 175 labels, $\mathcal{D}_{meta-test}$ including 200 labels. For an N-way K-shot setting, each task is sampled with N labels, and each label includes K support examples and 15 query examples. Note that due to the label distribution of multi-label data [26], several labels may correspond to more examples than other labels, however, the comparison on different methods are fair because of the same setting. We use GloVe [25] to generate the word vectors for the category labels as the semantic embeddings. The GloVe model is trained with large unsupervised text corpora.

Baselines. We compare the proposed model with five well-established or state-of-the-art few-shot learning algorithms: **MAML** [9], **Reptile** [15], **ATAML** [16], **MAML++** [17], **LEO** [22]. **MAML** is a groundbreaking gradient-based meta-learning method. **Reptiles** propose a *shortest descent* method to further improve efficiency and performance. **ATAML** introduces attention mechanism into meta-learning to learn task-agnostic representation. **MAML++** is the state-of-the-art meta-learning framework for few-shot learning problem, which employ multi-step loss optimization to improve the generalization performance. **LEO** applies pre-trained representations on a low-dimensional latent space instead of the original high-dimensional parameter space, which achieves current state-of-the-art classification performance. Since metric-based meta-learning methods [5, 11, 12] can only output a single label of the nearest neighbor, it can not be competent for multi-label task with an uncertain number of labels.

Evaluation Metrics. The predictive accuracy of multi-label classification task can be evaluated in different ways. Our proposed method was evaluated on Micro-F1, Macro-F1, and AUC.

[2] `http://mulan.sourceforge.net/`

Table 2: Ablation study of framework components

Ablation Study	5-way Micro-F1		10-way Micro-F1		5-way AUC		10-way AUC	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
DESIRE-Net$\backslash l_2$-norm	55.1%	63.0%	47.4%	51.3%	84.0%	88.5%	83.8%	86.1%
DESIRE-Net$\backslash\mathcal{F}$	46.9%	53.7%	32.8%	40.4%	80.1%	85.8%	78.6%	83.5%
DESIRE-Net$\backslash\mathcal{I}$	39.4%	52.9%	30.9%	41.7%	77.6%	84.3%	77.2%	83.8%
DESIRE-Net$\backslash I_f$	60.7%	65.3%	50.8%	54.9%	87.4%	90.1%	87.7%	89.1%
DESIRE-Net$\backslash I_c$	61.3%	66.4%	51.5%	56.5%	88.0%	90.7%	88.1%	89.4%
DESIRE-Net$\backslash I_a$	60.2%	63.1%	50.5%	55.8%	87.3%	90.2%	87.3%	88.1%
DESIRE-Net	**62.5%**	**68.8%**	**52.6%**	**57.6%**	**89.2%**	**91.5%**	**88.9%**	**90.7%**

Due to the limited space, more details about the experimental setup can be found in Appendix.

4.2 Results and Discussion

The multi-label few-shot classification performance for DESIRE-Net and other baselines are show in Table 1. We evaluate 5-way 1-shot, 5-way 5-shot, 10-way 1-shot and 10-way 5-shot learning on Micro-F1 and Macro-F1, respectively. The results show that DESIRE-Net significantly outperforms other methods and get the state-of-the-art performance on multi-label few-shot learning tasks. Specifically, DESIRE-Net performs better than MAML with an improvement up about 25%, and better than LEO with an improvement up about 10%. LEO also starts from a pre-trained model to extract features and achieve decent performance, but ignores the correlation between labels. DESIRE-Net with the ability of semantic inference can explore and exploit the label correlation, and thus facilitate to classify multiple labels using only a few training examples. We can find that DESIRE-Net has larger improvement in 1-shot task than 5-shot task. This is because DESIRE-Net incorporates the label semantics to effectively overcome the challenge of insufficient examples. More detailed experimental results are given in Appendix.

4.3 Ablation Study

To assess the effects of the proposed different components, we perform extensive ablation studies, with detailed result in Table 2. We first evaluate the the impact of l_2-norm in the semantic inference. If remove the l_2 normalization (denoted as DESIRE-Net$\backslash l_2$-norm), the performance of the model will be degraded. This is because the l_2 normalization can limit the adverse effects of absolute value fluctuations in semantic space, then the semantic similarity can be better represented. DESIRE-Net$\backslash\mathcal{F}$ indicates that the model uses a random initial network instead of the feature extractor $\mathcal{F}$, which show the effectiveness of $\mathcal{F}$ embedding examples to the semantic space. The semantic inference mechanism is the key of our model. If we remove it from DESIRE-Net (denoted as DESIRE-Net$\backslash\mathcal{I}$), DESIRE-Net is actually reduced to a MAML algorithm with semantic embedding and produces a large performance degradation. It confirms that the semantic inference mechanism helps boost the performance of multi-label few-shot learning. Specifically, we also assess the effects of the three different levels in the semantic inference mechanism. We denote the model without feature-level, correlation-level, and attention-level semantic inference as DESIRE-Net$\backslash I_f$, DESIRE-Net$\backslash I_c$, and DESIRE-Net$\backslash I_a$ respectively. It is showed that the semantic inference of each level improves the performance of DESIRE-Net, and thus can prove their effectiveness.

5 Conclusion

In this paper, we propose a deep semantic inference network (DESIRE-Net) for multi-label few-shot learning, which can quickly adapt multi-label classification tasks using only a few labeled examples. Evaluation metrics illustrate that DESIRE-Net could achieve averaging 10% higher than previous state-of-the-art methods by introducing our proposed semantic inference mechanism. Further experimental results demonstrate the significance of leveraging knowledge learned from the semantic correlation between features to facilitate the multi-label few-shot tasks.

Broader Impact

Our work aims at providing a more accurate recognition system for multi-label classification with very few learning samples. We think it may benefit the society that deep learning systems could be applied to more actual daily life scenarios. However, we do have an increasing concern that more advanced few-shot learning systems may lead to any entities to conduct censorship efficiently. In that case, we think that the accurate recognition system should be used under a reliable regulatory system.

References

[1] E. M. Markman. Categorization and naming in children: Problems of induction. *The MIT Press series in learning, development, and conceptual change*, 1991.

[2] Lauren A Schmidt. Meaning and compositionality as statistical induction of categories and constraints. Phd thesis, Massachusetts Institute of Technology, 2009.

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[4] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85 – 117, 2015.

[5] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29*, pages 3630–3638. 2016.

[6] Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[7] Jurgen Schmidhuber. Evolutionary principles in self-referential learning. on learning now to learn: The meta-meta-meta...-hook. Diploma thesis, Technische Universitat Munchen, Germany, 1987.

[8] S. Thrun. Lifelong learning algorithms. In *Learning to Learn*, pages 181–209, 1998.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 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, 2017. PMLR.

[10] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *In International Conference on Learning Representations (ICLR)*, 2017.

[11] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30*, pages 4077–4087. 2017.

[12] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O. Pinheiro. Adaptive cross-modal few-shot learning. In *Advances in Neural Information Processing Systems 32*, pages 4847–4857. 2019.

[14] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. A new meta-baseline for few-shot learning. *ArXiv*, abs/2003.04390, 2020.

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

[16] Xiang Jiang, Mohammad Havaei, Gabriel Chartrand, Hassan Chouaib, Thomas Vincent, Andrew Jesson, Nicolas Chapados, and Stan Matwin. On the importance of attention in meta-learning for few-shot text classification. *ArXiv*, abs/1806.00852, 2018.

[17] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your MAML. In *International Conference on Learning Representations*, 2019.

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

[19] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M. Bronstein. Laso: Label-set operations networks for multi-label few-shot learning. In *CVPR*, 2019.

[20] Ruibing Hou, Hong Chang, Bingpeng MA, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems 32*, pages 4003–4014. 2019.

[21] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[22] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019.

[23] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 9537–9548, 2018.

[24] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 7343–7353, 2018.

[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[26] M. Zhang and Z. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

Supplementary Material

In this supplementary material, we provide detailed experimental settings and results. Appendix A presents implementation details and definition of evaluation metrics. Appendix B presents additional empirical evaluation for multi-label few-shot learning. Appendix C makes further analyses of the proposed model and experimental results.

A Experimental Settings

A.1 Implementation Details

Pytorch[3] is used to implement the proposed algorithm and to conduct all the experiments. All the computations are performed on a 64-Bit Linux workstation with 10-core Intel Core CPU i7-6850K 3.60GHz processor, 256 GB memory, and 4 Nvidia GTX 1080 Ti GPUs. For meta-training stage, we use Adam optimizer with a fixed learning rate 0.001, weight decay 10^{-6}, $\beta_1 = 0.9$ and $\beta_2 = 0.99$. We train models 100 epochs, where each epoch contains 1,000 tasks randomly sampled from $\mathcal{D}_{meta-train}$. For the meta-testing stage, we test models on 1,000 novel tasks randomly sampled from $\mathcal{D}_{meta-test}$ to get average results. The semantic embedding model, GloVe [25], generates 300-dimension word vectors for the category labels[4]. For the architecture of DESIRE-Net, the feature extractor $\mathcal{F}$ has 3 layers of fully connected layers, and the meta-learner $\mathcal{I}$ consists of different networks with several connected layers. The temperature hyperparameter (in Equation (9)) is set in the range [0.1, 10], and dropout rate in the networks is set in the range [0.1, 0.5]. All hyperparameters are cross-validated in $\mathcal{D}_{meta-val}$ and fixed afterward in all experiments. We also provide the source code for reference.

A.2 Evaluation Metrics and Settings

Given a query set $\mathcal{Q}_{\mathcal{T}}$ sampled from meta-test dataset $\mathcal{D}_{meta-test}$ denoted by $\mathcal{Q}_{\mathcal{T}} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_q, \boldsymbol{y}_q)\}$, where $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^{d_s \times 1}$ is a real vector representing an input feature (instance) and $\boldsymbol{y}_i \in \mathcal{Y} \subseteq \{0,1\}^{N \times 1}$ is the corresponding output label vector ($i \in [n]$, defined as $i \in \{1, ..., n\}$). Moreover, $y_i^j = 1$ if the j-th label is assigned to the instance $\boldsymbol{x}_i$ and $y_i^j = 0$ otherwise. For notational simplicity, we use $Y_{i\cdot}^{+}$ to denote the index set of associated (non-associated) labels of $\boldsymbol{y}_i$. Formally, $Y_{i\cdot}^{+} = \{j | y_i^j = 1\}$ and $Y_{i\cdot}^{-} = \{j | y_i^j = 0\}$. With respect to j-th column of label matrix, $Y_{\cdot j}^{+} = \{i | y_i^j = 1\}$ denotes the index set of associated instances of the j-th label and $Y_{\cdot j}^{-} = \{i | y_i^j = 0\}$ denotes the set of non-associated instances similarly. We use $| \cdot |$ to represent the cardinality of a set.

Table 3: Definitions of multi-label performance measures.

Measure	Formulation	Note
macro-F1	$\textit{macro-F1}(H) = \frac{1}{m} \sum_{j=1}^{m} \frac{2 \sum_{i=1}^{n} y_{ij} h_{ij}}{\sum_{i=1}^{n} y_{ij} + \sum_{i=1}^{n} h_{ij}}$	F-measure averaging on each label.
micro-F1	$\textit{micro-F1}(H) = \frac{2 \sum_{j=1}^{m} \sum_{i=1}^{n} y_{ij} h_{ij}}{\sum_{j=1}^{m} \sum_{i=1}^{n} y_{ij} + \sum_{j=1}^{m} \sum_{i=1}^{n} h_{ij}}$	F-measure averaging on the prediction matrix.
AUC	$\textit{micro-AUC}(F) = \frac{\lvert\mathcal{S}_{\text{micro}}\rvert}{(\sum_{i=1}^{n} \lvert Y_{i\cdot}^{+}\rvert) \cdot (\sum_{i=1}^{n} \lvert Y_{i\cdot}^{-}\rvert)}$ $\mathcal{S}_{\text{micro}} = \{(a,b,i,j) \lvert (a,b) \in Y_{\cdot i}^{+} \times Y_{\cdot j}^{-},\ f_i(\boldsymbol{x}_a) \geq f_j(\boldsymbol{x}_b)\}$	AUC averaging on prediction matrix. $\mathcal{S}_{\text{micro}}$ is the set of correct quadruples.

Table 3 summarizes three popular multi-label evaluation metrics used in this paper, which can be divided into bipartition-based metrics, i.e., macro-F1 and micro-F1, and a ranking-based metric, i.e., AUC [26]. We assume that $H : \mathbb{R}^d \to \{0,1\}^m$ is the multi-label classifier and predicts which labels an instance is associated with. H can be decomposed as $\{h^1, ..., h^m\}$ and $h^j(\boldsymbol{x}_i)$ represents the

[3] `https://pytorch.org/`

[4] `https://nlp.stanford.edu/projects/glove/`

prediction of y_i^j. The results of H can be evaluated by bipartition-based metrics. $F : \mathbb{R}^d \rightarrow \mathbb{R}^m$ is the multi-label predictor, whose predicted value could be regarded as the confidence of association. $F = \{f^1, ..., f^m\}$ and $f^j(\boldsymbol{x}_i)$ denotes the predicted value of y_i^j, which can be evaluated by ranking-based metrics. H can be induced from F by thresholding techniques $t(\cdot)$. For example, $h^j(\boldsymbol{x}_i) = \mathbb{1}\{f^j(\boldsymbol{x}_i) > t(\boldsymbol{x}_i)\}$, where we use $\mathbb{1}\{event\}$ to denote the indicator function for $event$. In the experiment, we simply use 0.5 as the threshold for the output of all models. The evaluation metrics implementation is based on scikit-learn tools (`https://scikit-learn.org/stable/`).

B Additional Results

Table 4 shows the average performance of multi-label few-shot classification for DESIRE-Net and baselines. The results consistently shows that our model outperforms the baselines on AUC metric for both 1-shot and 5-shot, 5-way and 20-way. DESIRE-Net can extract semantic features and exploit the correlation between novel labels and base labels as prior knowledge, therefore, it can achieve better results in dealing with the problem of multi-label few-shot learning.

Table 4: Comparing multi-label few-shot classification performance on Delicious

Method	5-way AUC		10-way AUC	
	1-shot	5-shot	1-shot	5-shot
pre-trained	67.5%	75.3%	73.0%	81.7%
MAML [9]	75.9%	81.9%	79.0%	82.3%
Reptile [15]	76.1%	81.6%	79.2%	84.2%
ATAML [16]	83.1%	82.9%	83.0%	85.1%
MAML++ [17]	81.3%	84.0%	80.6%	84.4%
LEO [22]	85.8%	87.2%	84.8%	86.8%
DESIRE-Net	**89.2%**	**91.5%**	**88.9%**	**90.7%**

C Additional Analyses

Influence of semantic correlation. Based on the results in Table 1 and Table 2, it can be shown that label semantic correlation is a key for multi-label few-shot learning. We leverage label semantic correlation in both semantic-aware feature extractor $\mathcal{F}$ and meta-learner with semantic inference mechanism $\mathcal{I}$. Exploiting label correlation can facilitate the multi-label learning process to cope with the challenge of the overwhelming size of output space. For instance, if an image has been annotated with label *whale*, the probability of the image being associated with labels *ocean* and *seaweed* would be high, and the image is unlikely to be labeled as *grassland* and *lion*. Moreover, semantic embedding (learned from large unsupervised text corpora) can serve as prior knowledge and context to supplement the label correlation. The proposed DESIRE-NET incorporates the label correlation not only from the learning process but also from label semantic embedding, which has achieved great success in multi-label few-shot learning. Table 2 shows that If we remove feature extractor $\mathcal{F}$ or semantic inference $\mathcal{I}$ (denoted as DESIRE-Net$\backslash\mathcal{F}$ or DESIRE-Net$\backslash\mathcal{I}$, respectively), DESIRE-Net will produce a significant performance degradation.

Influence of l_2-norm. l_2-norm is a technique that is often used to provide regularities for deep neural networks. However, in our meta-leaner, l_2-norm plays a more critical role. We employ l_2-norm in three different levels of semantic inference. In feature-level inference (Equation 5), l_2-norm is used to eliminate the influence of the absolute magnitudes of semantic features and improve the robustness. In correlation-level inference (Equation 6), l_2-norm not only offers nonlinear operation for feature transformation but also limits the adverse effects of absolute magnitudes fluctuations of $\boldsymbol{z}W_{base}$. Since the number of base labels is different between meta-training and meta-testing (described in Section 3.3), by using l_2-norm, the absolute value of the feature transformation in meta-training and meta-testing can be kept consistent, which is important for the convergence of the model. In attention-level inference (Equation 10), l_2-norm provides the regularity of the attention value. Table 2 demonstrates that if remove the l_2 normalization (denoted as DESIRE-Net$\backslash l_2$-norm), the performance of the model will be degraded.