Query-Driven Multi-Instance Learning

Yen-Chi Hsu,[1,3] **Cheng-Yao Hong,**[1] **Ming-Sui Lee,**[3] **Tyng-Luh Liu,**[1,2]
[1]Institute of Information Science, Academia Sinica, [2]Taiwan AI Labs
[3]Department of Computer Science & Information Engineering, National Taiwan University
yenchi@iis.sinica.edu.tw, sensible@iis.sinica.edu.tw, mslee@csie.ntu.edu.tw, liutyng@iis.sinica.edu.tw

Abstract

We introduce a query-driven approach (qMIL) to multi-instance learning where the queries aim to uncover the class labels embodied in a given bag of instances. Specifically, it solves a multi-instance multi-label learning (MIML) problem with a more challenging setting than the conventional one. Each MIML bag in our formulation is annotated only with a binary label indicating whether the bag contains the instance of a certain class and the query is specified by the word2vec of a class label/name. To learn a deep-net model for qMIL, we construct a network component that achieves a generalized compatibility measure for query-visual co-embedding and yields proper instance attentions to the given query. The bag representation is then formed as the attention-weighted sum of the instances' weights, and passed to the classification layer at the end of the network. In addition, the qMIL formulation is flexible for extending the network to classify unseen class labels, leading to a new technique to solve the zero-shot MIML task through an iterative querying process. Experimental results on action classification over video clips and three MIML datasets from MNIST, CIFAR10 and Scene are provided to demonstrate the effectiveness of our method.

Introduction

Supervised learning techniques that rely on deep neural networks have made significant progress in active research fields of artificial intelligence such as classification (Simonyan and Zisserman 2014; He et al. 2016), the mainstream of computer vision applications. In solving an image classification problem, each training sample often comprises a raw image and the corresponding class/category label. However, such a classification setting may not be sufficient to satisfactorily account for real-life applications nowadays. With the rapid advances of machine learning research, it becomes feasible to simultaneously explore all the *useful* information of either an image or a batch of images. In other words, image classification is no longer restricted to the problem where an image is labeled as a single category. Among the variants of classification frameworks, *e.g.*, as illustrated in Figure 1, we aim to address the multi-instance multi-label learning (MIML) in (Zhou et al. 2012) from a novel viewpoint of *learning through queries*.

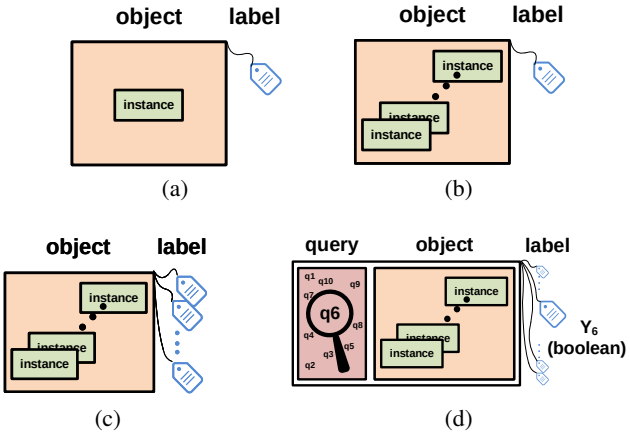

Figure 1: Variants of supervised-learning tasks: (a) Classification (b) Multi-instance learning (MIL) (c) Multi-instance multi-label learning (MIML) (d) Query-driven multi-instance learning (qMIL).

The MIML problem is characterized by that an object or a bag consists of several instances with multiple class labels. While MIMLSVM (Zhou and Zhang 2007) is proposed to deal with the problem, deep MIML in (Feng and Zhou 2017) is shown to be more effective than other traditional methods. Notably, existing supervised learning approaches for MIML are provided with the full binary label vector associated with each training bag, and thus have access to the presence of any class label in a bag. Such a learning setting requires extensive manual efforts in annotating the vast amount of training bags. In our method, a query-driven multiple instance learning (qMIL) framework is proposed to tackle MIML without specifying the full binary label vector. In fact, the qMIL formulation requires only a binary label for each bag along with the corresponding label query. The proposed method thus has two main advantages. First, it is flexible to introduce new classes into the model without the need to modify the labeling information in the existing training data and the classification layer. Second, the query mechanism enables qMIL to inherently and additionally perform zero-shot classification in a crude way.

Related Work

For the ease of discussion, we divide the literature survey of relevant techniques into three groups, namely, *multi-instance learning*, *attention mechanism*, and *zero-shot learning*.

Multi-instance Learning The MIL paradigm deals with those learning problems for which labels only exist for sets of data points. A set of data points is typically termed as a bag and each data point is considered as an instance. Following (Dietterich, Lathrop, and Lozano-Pérez 1997), a bag is said to be positive with respect to a certain binary label if at least one instance within the bag is positive. The strategy of (Chen, Bi, and Wang 2006) maps each bag into a feature space defined by the instances in the training bags via an instance similarity measure and ℓ_1-norm SVM is applied to select important features as well as construct classifiers simultaneously. In (Liu, Wu, and Zhou 2012), the authors construct nearest-neighbor graphs among instances and uncover positive instances within positively-labeled groups. The MIL formulation in (Pathak et al. 2014) is designed to learn a semantic segmentation model based on weak image-level labels. More recently, (Wang et al. 2018) employs neural networks that aim at solving the MIL problems in an end-to-end manner. An attention-based neural network model is proposed in (Ilse, Tomczak, and Welling 2018) to detect positive instances automatically. In (Dennis et al. 2018), a recurrent neural network model called MI-RNN is developed to find out the *signature*, which is linked to those positive instances in a bag. Among the aforementioned classical MIL problems, each bag has only one corresponding label. However, in many practical applications, a complex bag (such as an image), which contains various instances like pixels, may have more than one relevant label. The MIML framework of (Zhou and Zhang 2007) is established to tackle the complicated scene classification. Over the past few years, assorted algorithms, ranging from traditional, *e.g.*, SVM (Nguyen 2010; Briggs, Fern, and Raich 2012) and k-nearest neighbor (KNN) (Zhang 2010), to popular like deep neural network learning (Feng and Zhou 2017), have been proposed to address the MIML problem.

Attention Mechanism The attention mechanism has a significant impact on designing deep learning architecture to solve challenging applications in artificial intelligence, including image captioning, *e.g.*, (Xu et al. 2015; You et al. 2016), visual question answering, *e.g.*, (Lu et al. 2016), and machine translation, *e.g.*, (Luong, Pham, and Manning 2015). For solving the MIL or MIML problems, as the individual instance labels of training data are not given, the attention distribution is often learned implicitly via optimizing the bag-level objective function.

Zero-shot Learning A critical limitation of deep learning is that it often takes a massive amount of samples to train a satisfactory model, and the classifier, such as trained by cats and dogs, can only classify cats and dogs. This means that the classifier is not able to be directly applied to recognize other species. On the contrary, zero-shot learning (ZSL) refers to the learning of classifying samples of unseen categories. It implies that the training classes and the zero-shot testing classes are different. For example, the ZSL algorithm proposed in (Lampert, Nickisch, and Harmeling 2009) guides the model to classify unseen categories, empowering machines the capacity for reasoning and true intelligence.

Our Approach To establish the proposed qMIL, we first need to generate a training dataset of bags. Specifically, for each query about a certain class label, a bag of instances from randomly-selected classes are generated. If there exists at least one instance from the query class, the underlying bag is said to be positive and its binary label is set to 1. Otherwise, it is a negative bag with label 0. Notice that only the examples from the classes of interest can be included in a bag. Our setting is different from that in (Dennis et al. 2018) where a positive bag is composed of one or a few positive instances and several negative instances, which are usually noise, *i.e.*, not from any of the underlying classes of interest. In qMIL, each training sample/bag is annotated with a binary label, rather than a binary label vector over all classes as in the MIML setting. However, the proposed method still satisfactorily solves the MIML problem in that a proper bag representation for classification can be obtained by qMIL via more effectively estimating the query-adapted attention distribution over instances within a bag. We summarize the main advantages of the proposed qMIL over other existing techniques below.

1. The qMIL formulation is flexible. When new data of additional classes are included, all binary labelings of the existing training data remain the same, whereas annotating with a full label vector as in the conventional MIML needs to modify all the labeling information.
2. The qMIL network architecture is general. When additional new classes are introduced, the network architecture remains the same. It can be readily fine-tuned to classify the new classes by generating the queries of new classes and the corresponding training bags. However, with the MIML architecture, one would need to expand the classification layer to account for the new classes.
3. The qMIL framework enables zero-shot classification. When data of unseen classes are added in the testing bags, we perform iterative queries to first remove most positive instances of seen classes from a given testing bag, and then compute a more reliable attention distribution for each query of an unseen class to decide if any positive instance of an unseen class is present or not.

Our Method

The qMIL framework is developed to learn a neural network model that adapts to the underlying query and dynamically yields a proper bag representation for classification. To comprehend the main ideas, we focus on describing: 1) how to generate the training data; 2) how to establish a generalized compatibility measure to facilitate the query-visual co-embedding; 3) how to employ label-dependent regularization to yield the desirable attention distribution over bag

instances; and 4) how to use attention pooling to obtain the query-adapted bag representation for classification. Finally, we detail a handy procedure resulted from qMIL to carry out zero-shot classification via iterative queries.

The qMIL Problem

In the classical supervised learning such as multi-class classification, the aim is to train a model that predicts a target label $y \in \{1, \ldots, C\}$ for a given test sample $\mathbf{x} \in \mathbb{R}^D$, where C represents the number of classes. However, in the formulation of qMIL, each example is represented as a bag of instances, $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_{K_X}\}$, where K_X is the number of instances and could vary over bags with a pre-specified upper bound K. Notice that neither dependency nor ordering relationships are considered in generating the instances for each bag.

To incorporate the query mechanism into qMIL, we have a set of C queries, $Q = \{q_1, \ldots, q_C\}$, where the query q_c inquires the existence of class label c in a bag, and is encoded with the corresponding class name/word. The proposed qMIL implicitly solves a more challenging MIML problem than the conventional one. The critical distinction is that each bag X in the training data of qMIL comes with only a single binary ground truth Y indicating the existence of at least one instance of a particular class in X, while the original MIML setting requires a full C-dimensional binary vector describing the presence of all the class labels in X. When $C = 1$, this is exactly the form of training data used for solving a binary MIL problem. For $C > 1$, we use a triplet (X, Y, q) to indicate that the bag label Y depends on the query $q \in Q$ and is defined by

$$Y = \begin{cases} 0, & \text{iff } \sum_{k=1}^{K_X} \mathbb{I}(q \equiv y_k) = 0, \\ 1, & \text{otherwise,} \end{cases} \tag{1}$$

where $y_k \in \{1, \ldots, C\}$ is the class label of the instance $\mathbf{x}_k$ in X. The notation $\mathbb{I}(q \equiv y_k)$ is an indicator function for signaling whether the query q concerns the label y_k. We emphasize that the instance-level labels y_k are not available in learning the qMIL model. They are included in (1) solely for providing an analytic form in defining the bag label Y with respect to the query q.

With (1), it is insightful to describe how the training data of qMIL are generated. Suppose we intend to work with a query subset, $Q' \subseteq Q$, and N training bags. Thus, for each query $q \in Q'$, we generate $N/|Q'|$ bags, which can be divided into two equal-numbered positive and negative subsets, denoted as $\{(X_i^+, Y_i = 1, q)\} \cup \{(X_i^-, Y_i = 0, q)\}$. The total number of instances in each bag is randomly decided with an upper bound K, and only instances with a class label in $\{1, \ldots, C\}$ are considered. These $|Q'|$ query-dependent collections of bags form the final training dataset $\mathcal{S}$ of N bags. It indicates that the training procedure considers equal number of positive and negative training bags for each $q \in Q'$, which enables focusing on learning to solve the classification task without imposing any presumed distribution on the data. In the experiments, we demonstrate that the inference performance of qMIL does not significantly vary with respect to the ratio between the numbers of positive and negative bags.

Query-adapted Attention Pooling

Although the number of instances in a qMIL bag could vary, we hereafter assume that all bags have K instances. After all, null instances can be introduced when needed. The unified bag size makes the batch training of learning the neural network, as shown in Figure 2, more convenient. Now consider an arbitrary training bag (X, Y, q), we use *word2vec* (Mikolov et al. 2013) to represent the query q as a 300-D feature vector and pass it through a two-layer MLP to obtain the query embedding $\phi(q) \in \mathbb{R}^d$. On the other hand, the image feature vector of each instance $\mathbf{x}$ is forward propagated through a three-layer MLP to yield its visual embedding which is denoted as $\psi(\mathbf{x}) \in \mathbb{R}^d$. The two mappings can be aligned to achieve query-visual co-embedding. To this end, we construct a network component $\mathcal{A}$ to function as a generalized compatibility measure for better exploring the co-embedding. Specifically, we have

$$\mathcal{A}(\phi(q), \psi(\mathbf{x})) = \sigma_2(\mathbf{w}^\intercal \sigma_1(V(\psi(\mathbf{x}) \odot \phi(q)))), \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^{L \times 1}$ and $V \in \mathbb{R}^{L \times d}$ are network parameters, $\odot$ denotes the element-wise product, and σ_1, σ_2 are activation functions. When $L = d$ and linear activation functions in (2) are used, the generalized compatibility measure $\mathcal{A}$ simply reduces to taking inner product between $\psi(\mathbf{x})$ and $\phi(q)$ if both V and $\mathbf{w}$ are fixed as the identity versions.

It follows from (2) that we can use the compatibility measure $\mathcal{A}$ to compute the unnormalized attention $\alpha_k = \mathcal{A}(\phi(q), \psi(\mathbf{x}_k))$ for each instance $\mathbf{x}_k \in X$ to a given query q. Then the attention-weighted pooling is utilized to obtain the bag representation $\mathbf{z}$ for X, which adapts to the query q as follows:

$$\mathbf{z} = \sum_{k=1}^{K} \beta_k \, \mathbf{x}_k \quad \text{and} \quad \beta_k = \frac{\exp\{\alpha_k/\tau\}}{\sum_{j=1}^{K} \exp\{\alpha_j/\tau\}}, \tag{3}$$

where τ is the temperature parameter and β_k is the normalized attention of instance $\mathbf{x}_k \in X$ to q.

Loss Function and Regularization

For each training triplet $(X, Y, q) \in \mathcal{S}$, we now know how to derive the bag's feature vector $\mathbf{z}$ according to (3) and the corresponding unnormalized attention vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\intercal$. To train the network to perform the (binary) classification task for predicting the bag label with respect to q, we need to define a proper loss function $\mathcal{L}$ to accomplish the qMIL learning. Specifically, we consider a label-dependent attention-regularized loss function:

$$\mathcal{L}(\mathcal{S}) = \mathcal{L}_1(\mathcal{S}) + \lambda \, \mathcal{L}_2(\mathcal{S}), \tag{4}$$

where λ is the weighting parameter, and the two losses for classifying each $(X, Y, q) \in \mathcal{S}$ are

$$\mathcal{L}_1(X) = Y \log \, p(X) + (1 - Y) \log \, (1 - p(X)), \tag{5}$$

$$\mathcal{L}_2(X) = Y \|\boldsymbol{\alpha}(X)\|_1 + (1 - Y)\{\mathrm{Var}(\boldsymbol{\alpha}(X))\}^{\frac{1}{2}}. \tag{6}$$

$\mathcal{L}_1$ in (5) is the cross-entropy loss and the attention regularization loss $\mathcal{L}_2$ in (6) plays a crucial role in the proposed qMIL formulation. Here we justify the form of the proposed regularization loss in (6) for the two possible cases.

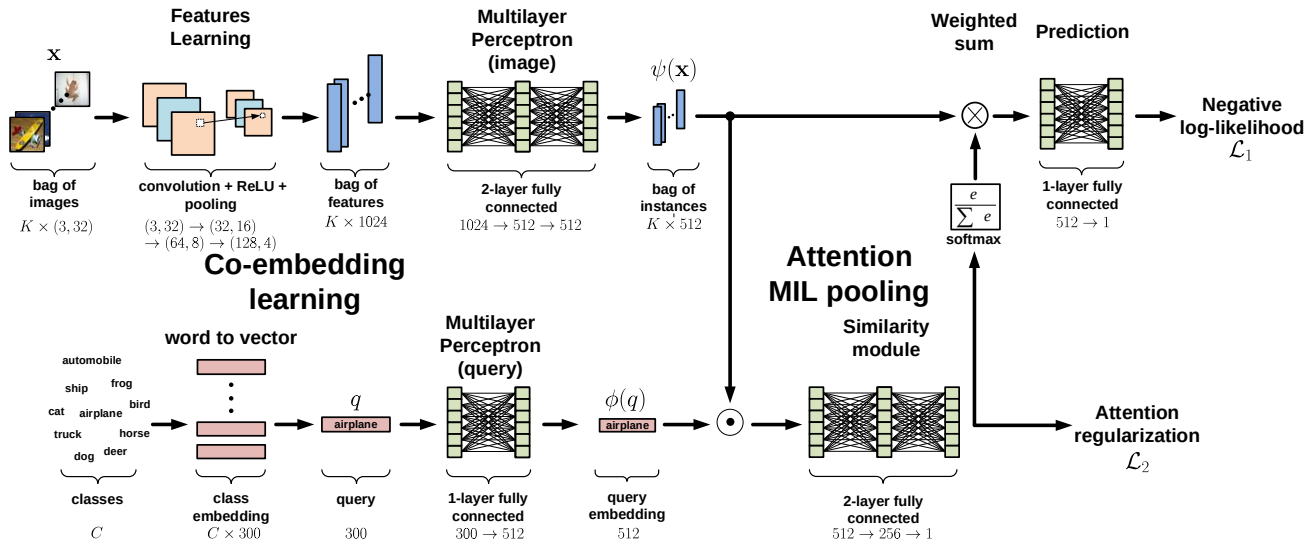

Figure 2: The proposed qMIL neural network architecture.

- When $Y = 1$, the training bag X has a positive label to q and $\mathcal{L}_2 = \|\boldsymbol{\alpha}\|_1$. The ℓ_1-norm regularization effect is to find a *sparse* distribution of the instance attention. The preference is reasonable in the case where at least one instance is relevant to the query q and the sparse prior aims to distribute most attention to the relevant instances.
- When $Y = 0$, we have $\mathcal{L}_2 = \sqrt{\mathrm{Var}(\boldsymbol{\alpha})} = \|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\|_2$. In this case all instances in the training bag X are irrelevant to the query q. The use of ℓ_2-norm thus encourages the attention to uniformly spread over all the instances.

Zero-shot Classification via Queries

Thus far we have described how to leverage with the query mechanism to implicitly solve an MIML problem with a (triplet) training dataset, where each training bag is annotated only with a single binary label. We now explain how to apply a learned qMIL model to tackle the following zero-shot scenario. Suppose that in generating testing bags, we decide to consider instances from both the seen and unseen classes. Then, inquiring an arbitrary testing bag X with a query about an unseen class would result in zero-shot classification. We use an explicit example to depict the scenario. Let `car` be a seen class and `truck` an unseen class. A testing bag X includes at least one instance of `car` and all the other instances are not `truck`. A query about `truck` for X would most likely confuse the qMIL model and yields a positive return for the false existence of a `truck` instance. The confusion is caused by that `car` and `truck` are *similar* in the space induced by *word2vec*. Thus, to tackle the resulting zero-shot classification, we consider a two-stage procedure. In stage one, we iteratively perform queries of all the seen classes to identify *strong* positive instances, and exclude them from further considerations. In stage two, now without the severe distraction from the evident instances of seen classes, qMIL can then estimate a proper attention distribution and thus refine the bag representation for zero-shot classification. Further details are provided in the experimental results.

Experimental Results

We evaluate our method mainly on the MNIST-based dataset (MNIST-BAGS) (Ilse, Tomczak, and Welling 2018) and CIFAR10-based dataset (CIFAR10-BAGS). Besides the pilot study on zero-shot classification, there are three groups of experimental results. The first set of experiments concerns a standard MIL problem where we compare qMIL with the deep MIL in (Ilse, Tomczak, and Welling 2018). In this setting, the total number of query class is just one. The second set of experiments is then extended to dealing with the MIML problem. As we have pointed out that despite using less-annotated training data, qMIL yields convincing results and shows effectiveness over the compared methods. The third set of experiments deals with a popular real-life application, action recognition. The proposed qMIL is applied to determine whether a given video clip contains a specific action to the query, where we have tested with a subset of Activity Net (Fabian Caba Heilbron and Niebles 2015).

Learning with qMIL is advantageous, especially in creating training data. We just need to focus, in turn, on each particular category of interest, and mark whether the bag assumes the label or not. This can reduce human errors when annotating multiple labels and effectively reduce data noise. After all, in practical applications, we most likely care about only the categories we are interested in. Finally, given a novel query about an unseen class, the qMIL model is demonstrated to make reasonable predictions that are significantly better than random guesses.

Data Sampling

We follow the similar data sampling method in (Ilse, Tomczak, and Welling 2018) to create the MNIST-BAGS MIL dataset from MNIST (LeCun, Cortes, and Burges 1998) and analogously from CIFAR10 (Krizhevsky and Hinton 2009).

Table 1: Single query results on MNIST/CIFAR over ten runs of training/testing data sampling.

MNIST					CIFAR10				
Query	GatedAttnDMIL		qMIL		Query	GatedAttnDMIL		qMIL	
	accuracy	attention acc.	accuracy	attention acc.		accuracy	attention acc.	accuracy	attention acc.
0	95.4 ± 3.7	99.6 ± 1.2	$\mathbf{96.9 \pm 2.2}$	$\mathbf{99.6 \pm 1.2}$	plane	82.4 ± 1.7	82.7 ± 3.2	$\mathbf{89.9 \pm 1.7}$	$\mathbf{84.8 \pm 1.5}$
1	97.0 ± 4.1	99.6 ± 1.2	$\mathbf{98.0 \pm 2.4}$	$\mathbf{99.8 \pm 0.6}$	car	89.6 ± 1.8	95.7 ± 12.9	$\mathbf{90.7 \pm 1.4}$	$\mathbf{95.1 \pm 1.4}$
2	93.7 ± 3.6	99.6 ± 1.2	$\mathbf{95.7 \pm 2.7}$	$\mathbf{99.6 \pm 1.2}$	bird	72.4 ± 2.6	60.0 ± 22.7	$\mathbf{73.6 \pm 2.4}$	$\mathbf{69.7 \pm 9.0}$
3	93.2 ± 3.6	99.8 ± 0.6	$\mathbf{96.0 \pm 2.3}$	$\mathbf{100.0 \pm 0.0}$	cat	75.4 ± 3.0	54.1 ± 12.8	$\mathbf{76.3 \pm 2.9}$	$\mathbf{59.7 \pm 10.3}$
4	94.7 ± 2.5	99.2 ± 0.9	$\mathbf{96.5 \pm 1.3}$	$\mathbf{99.4 \pm 0.9}$	deer	71.4 ± 3.1	66.6 ± 5.9	$\mathbf{73.8 \pm 2.4}$	$\mathbf{67.6 \pm 5.6}$
5	94.0 ± 5.8	100.0 ± 0.0	$\mathbf{97.0 \pm 2.2}$	$\mathbf{100.0 \pm 0.0}$	dog	74.1 ± 2.3	62.2 ± 20.0	$\mathbf{74.3 \pm 1.8}$	$\mathbf{69.8 \pm 6.9}$
6	94.7 ± 4.1	99.00 ± 1.3	$\mathbf{97.1 \pm 2.4}$	$\mathbf{99.2 \pm 1.3}$	frog	82.2 ± 3.0	87.8 ± 1.9	$\mathbf{82.6 \pm 2.4}$	$\mathbf{88.4 \pm 2.5}$
7	94.2 ± 3.1	100.0 ± 0.0	$\mathbf{96.1 \pm 1.6}$	$\mathbf{100.0 \pm 0.0}$	horse	82.7 ± 2.9	77.8 ± 19.6	$\mathbf{82.8 \pm 1.9}$	$\mathbf{82.8 \pm 7.9}$
8	89.3 ± 6.9	99.20 ± 0.9	$\mathbf{92.1 \pm 5.9}$	$\mathbf{99.6 \pm 0.8}$	ship	87.8 ± 2.5	89.1 ± 1.8	$\mathbf{88.4 \pm 1.9}$	$\mathbf{89.8 \pm 1.4}$
9	91.3 ± 3.6	98.20 ± 1.9	$\mathbf{92.9 \pm 3.1}$	$\mathbf{98.2 \pm 1.9}$	truck	85.5 ± 1.8	90.4 ± 2.6	$\mathbf{85.9 \pm 1.6}$	$\mathbf{91.6 \pm 2.4}$

The standard MIL problem with one single query proceeds as follows. In MNIST or in CIFAR10, each of the ten categories will be chosen in turn as the one of interest, and the remaining are treated as background/noise. The instances in each bag are randomly included, and the number of instances is an integer arbitrarily sampled from the normal distribution $\mathcal{N}(10, 2)$. To speed up the training process, after data sampling and when necessary, zero images are generated to ensure that each bag has exactly K image instances. We next turn to the MIML scenario. For each image we now have multiple labels but do not indicate the specific label of each instance. (We have described how we construct such training data in establishing the qMIL problem.) There are two kinds of inference tasks for MIML. One is the classical MIML problem, and the other is ours, which is query-driven. For fair comparisons, we adopt the MIML Scene dataset (Zhou and Zhang 2007) as the benchmark and report 10-fold cross-validation results. Note that the numbers of positive bags and negative bags to a query in the MIML Scene dataset is unbalanced. The ratio between positive and negative bags is about $3:1$. The last experiment is about action recognition. In this case, a video clip can be thought of as a bag, while each frame is an instance.

Training and Inference

In the experiments of MNIST-MIL and CIFAR10-MIL, the hyperparameters can be kept the same. This implies that the proposed attention regularization in (6) is general and not data-sensitive. In MNIST, our CNN model conforms to the LeNet architecture (LeCun, Cortes, and Burges 1998) which comprises two conv layers for MNIST, and three conv layers for CIFAR10. The learning rate is 10^{-4} at initialization and the optimization method is Adam (Kingma and Ba 2014). The weight decay is 10^{-5}, while λ in (4) is 10^{-4} for all the experiments. We fix τ in (3) as 0.5. σ_1 and σ_2 in (2) are $\tanh$ and linear mapping. For single query, the results are reported with the mean and standard deviation from ten different runs of random data sampling. For multiple queries, five random runs are instead evaluated for the sake of efficiency.

Metrics In both our model and the compared method, the output of the bag-level prediction to the MIL problem is a probability p. Thus to compute the accuracy of the bag-level prediction, the decision threshold is set as $p > 0.5$ with label $Y = 1$ and $p \leq 0.5$ with label $Y = 0$. Consider now an arbitrary bag $X = (\mathbf{x}_1, \ldots, \mathbf{x}_K)$. In both MNIST-MIL and CIFAR10-MIL, we indeed have access to the class label of each instance, *i.e.*, $(y_1, \ldots, y_K)$. The instance-level ground truth can be used to evaluate the accuracy of the predicted instance attention in each bag. We name the resulting quantity as the instance-level accuracy. The attention accuracy is evaluated as follows. Each time we predict the bag label as $Y = 1$ for a triplet (X, Y, q), we check the instance label y_{k^*} of the most *manifest* instance $\mathbf{x}_{k^*}$ where $k^* = \arg\max_k \beta_k$ from (3). If $y_{k^*} = 1$, then we have correct instance attention.

Standard MIL

In standard MIL experiments, for each single query to a specific class label we first sample 500 training bags, including 250 positive and 250 negative bags from MNIST. Analogously, another 1000 bags (500 "+" & 500 "-") are also generated for testing. The setting for CIFAR10 is the same. We compare our method with the state-of-the-art deep MIL model, denoted as GatedAttnDMIL (Ilse, Tomczak, and Welling 2018) and report the results in Table 1. The proposed qMIL achieves better performances in both bag-level accuracy and instance-level attention accuracy. In Table 2, we report the performance versus different numbers of training bags for the CIFAR10 dataset. The results are on 500 testing bags. To achieve bag-level predictions of high confidence, qMIL needs 5000 training bags (2500 "+" & 2500 "-") for a single query. Our method also achieves better results in both accuracy metrics.

MIML

In the MIML problem, we have two ways of testing. One is to make the testing data the same form by our labeling scheme on training data, and the other is the standard MIML task that a bag of instances has several labels to be predicted. Table 3 shows the performances with respect to the numbers of query classes. When excluding the use of $\mathcal{L}_2$ in (6) (shown as qMIL$^-$ in Table 3), we have trained with many different hyperparameters and report the best results. It can be observed that with the attention regularization term, $\mathcal{L}_2$,

Table 2: Single query on CIFAR10. N: total # of training bags. (**acc**: accuracy, **att**: attention)

	N bags	100	500	1000	2000	5000
GatedAttnDMIL	**acc**	55.1 ± 8.6	62.1 ± 6.7	61.2 ± 6.2	70.6 ± 4.3	82.4 ± 1.7
qMIL		$\mathbf{56.3 \pm 4.5}$	$\mathbf{62.8 \pm 3.9}$	$\mathbf{63.4 \pm 4.1}$	$\mathbf{71.8 \pm 2.8}$	$\mathbf{89.9 \pm 1.7}$
GatedAttnDMIL	**att acc**	49.2 ± 20.1	58.2 ± 13.4	66.7 ± 8.3	76.8 ± 4.4	82.7 ± 3.2
qMIL		$\mathbf{55.3 \pm 11.3}$	$\mathbf{61.2 \pm 8.1}$	$\mathbf{67.2 \pm 5.2}$	$\mathbf{78.2 \pm 2.1}$	$\mathbf{84.8 \pm 1.5}$

Table 3: Performance with respect to # of queries on CIFAR10. The notation qMIL$^-$ denotes that the regularization loss $\mathcal{L}_2$ in (6) is not used in training. For each query, we sample 5000 training bags.

	# queries	1	3	5	7	10
qMIL$^-$	**acc**	82.4 ± 1.7	81.22 ± 1.8	71.23 ± 3.4	65.66 ± 4.6	78.33 ± 2.3
qMIL		$\mathbf{89.9 \pm 1.7}$	$\mathbf{81.77 \pm 1.4}$	$\mathbf{79.45 \pm 2.7}$	$\mathbf{82.09 \pm 2.1}$	$\mathbf{86.14 \pm 1.3}$
qMIL$^-$	**att acc**	82.7 ± 3.2	65.52 ± 9.9	53.21 ± 10.37	45.66 ± 20.3	70.64 ± 5.4
qMIL		$\mathbf{84.8 \pm 1.5}$	$\mathbf{87.22 \pm 1.1}$	$\mathbf{83.30 \pm 1.3}$	$\mathbf{86.01 \pm 1.2}$	$\mathbf{89.18 \pm 1.0}$

Table 4: 10-fold cross validation on MIML Scene dataset.

	accuracy
deep MIML (Feng and Zhou 2017)	89.45 ± 1.22
qMIL	$\mathbf{90.20 \pm 0.96}$

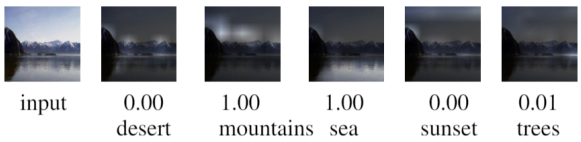

Figure 3: From column 2 to column 6: Each includes an attention heatmap and its bag-level probability, while the input image is shown in the first column.

learning the model becomes easier and more stable during training. (Further details about the regularization effect with $\mathcal{L}_2$ can be found in the supplementary material.)

We have also tested according to the standard MIML task by evaluating the model with each query for a given bag. Table 4 and Figure 3 include the results of the MIML task on the MIML Scene dataset and the comparison with the deep MIML (Feng and Zhou 2017) which is shown to outperform MIML SVM, MIML KNN, MIML RBF and MIML Boost (Zhou et al. 2012). We adopt a pre-trained ResNet50 (He et al. 2016) and re-implement the deep MIML by following the details described in the paper. The resulting deep MIML architecture consists of the pre-trained ResNet50, 2D sub-concept layer for multiple instances, and max pooling twice to yield the multi-label prediction. It is trained from scratch and learned end-to-end.

To better capture the effect of attention regularization, we investigate how the attention weights of a bag vary with respect to different queries of a class label. Table 5 shows the bag-level prediction of probability p and the attention weight distribution according to each query at testing.

MIML for Video Applications

The proposed qMIL can be readily applied to deal with video-related applications. Particularly, we explore the problem involving the Activity Net (Fabian Caba Heilbron and Niebles 2015) and convert the problem into our formulation described in the proposed qMIL. Following (Wang et al. 2016), each snippet comprises 16 consecutive frames, and a video clip can thus be represented as a sequence of snippets. Under such a setting, a video clip is a bag and each snippet is an instance, while its bag label is defined with respect to the query. In our experiment, we consider those video clips related to the following three action classes, namely, shot put, discus throw, and tumbling. Figure 4 shows the result of the proposed qMIL approach to action recognition.

Zero-shot Scenarios

We also test qMIL for zero-shot classification on CIFAR10. Specifically, we train the proposed qMIL with seven seen classes and test on the remaining three unseen classes. Each bag in the training data is randomly composed of instances from the seven seen classes, and the testing data are formed based on two kinds of sampling methods. The fist scenario is that the testing bags are sampled only from the three unseen classes, and the other is sampled from all of the ten classes (seen & unseen). For the latter case, the learned qMIL is carried out with the help of iterative queries as described in **Zero-shot Classification via Queries**. The experimental results of zero-shot classification are shown in Table 6 and Figure 5. We remark that the zero-shot scenario is essentially different from the conventional formulation. Therefore, it is not appropriate to directly compare it with other specific zero-shot learning techniques, which are cast in a very different way. The application demonstrates that the advantages and flexibility of the proposed qMIL formulation over

Table 5: Given a testing bag (13 instances), the instance attention weights vary w.r.t. different queries.

														p
plane	0.04	0.07	0.02	0.01	0.02	0.01	0.03	0.05	0.01	0.12	**0.55**	0.05	0.01	0.99
car	0.00	0.00	0.01	**0.81**	0.01	0.00	0.01	0.01	0.00	0.11	0.01	0.01	0.00	0.99
bird	0.03	0.07	0.03	0.01	0.02	0.03	0.02	**0.54**	0.03	0.01	0.02	0.03	0.15	0.98
cat	0.07	**0.24**	0.19	0.01	0.05	0.11	0.02	0.04	0.08	0.01	0.02	0.02	0.16	0.96
deer	0.15	0.09	0.08	0.01	0.03	0.04	**0.33**	0.07	0.06	0.03	0.03	0.03	0.05	0.01
dog	0.04	0.09	0.29	0.01	0.01	0.08	0.01	0.02	**0.37**	0.01	0.01	0.01	0.06	0.99
frog	0.07	0.08	0.11	0.11	0.04	**0.14**	0.05	0.06	0.07	0.07	0.03	0.04	0.12	0.01
horse	0.06	0.03	0.05	0.01	0.01	0.02	**0.68**	0.02	0.06	0.02	0.01	0.02	0.02	0.96
ship	0.01	0.02	0.00	0.01	**0.53**	0.01	0.00	0.01	0.00	0.01	0.04	0.36	0.01	0.99
truck	0.05	0.05	0.07	0.13	0.05	0.03	0.08	0.05	0.04	**0.26**	0.08	0.09	0.02	0.01

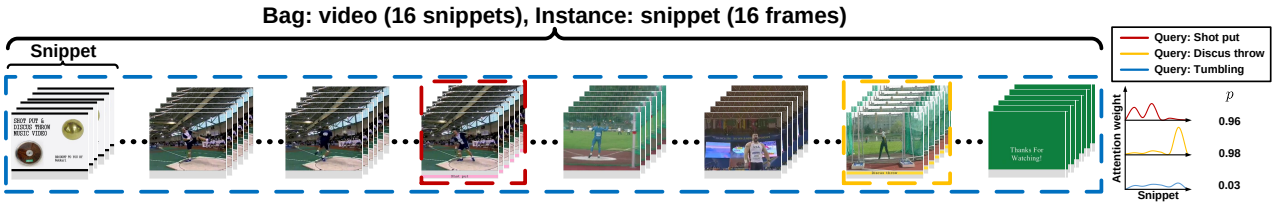

Figure 4: qMIL for action recognition. Each video clip comprises 16 snippets. Three different queries are chosen for testing. p is the bag-level probability prediction for supporting a query.

Table 6: Zero-shot testing accuracy with seven seen classes and three unseen classes. Test data are sampled from seen+unseen (ten classes) or from unseen (seven classes). IQP denotes the iterative query process.

	horse	ship	truck	total
seen & unseen	58.80	62.20	59.60	60.20
seen & unseen (IQP)	57.80	64.20	63.00	61.67
unseen	66.66	72.00	66.33	68.33

conventional MIL frameworks.

Conclusions

From the viewpoint of problem reduction, the proposed qMIL framework indeed can be considered as decomposing MIML into a series of query-driven MIL sub-tasks. The reduction yields advantages in two different aspects. First, annotating each training bag requires a single binary label, rather than a binary label vector. It also has the flexibility to expand the training dataset to include data of new classes without the need to modify the labeling information in the existing training bags. Second, the reduced sub-tasks can all be cast as query-driven MIL, and thus can be addressed in a unified neural network architecture. By focusing on solving the reduced MIML problem, we are able to establish a query-visual co-embedding with the label-adapted regularization in (6) and represent a given MIL bag with a proper representation for more effective classification. Our future work will focus on improving the qMIL attention mechanism and expanding its application aspect in image/video processing.

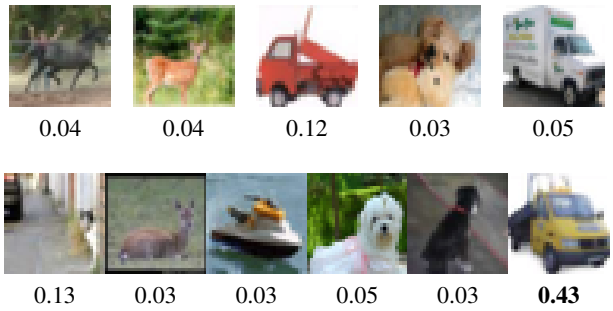

Figure 5: The "truck" class is not in the training data. Given the query of unseen "truck", qMIL with IQP will pay more attention to the "truck" image in a bag and the bag-level probability is $p = 0.96$. The numbers are the attention weights.

Acknowledgements. This work was supported in part by the MOST, Taiwan under Grant 108-2634-F-001-007. We are grateful to the *National Center for High-performance Computing* for providing computational resources and facilities.

References

Briggs, F.; Fern, X. Z.; and Raich, R. 2012. Rank-loss support instance machines for miml instance annotation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 534–542. ACM.

Chen, Y.; Bi, J.; and Wang, J. Z. 2006. Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(12):1931–1947.

Dennis, D.; Pabbaraju, C.; Simhadri, H. V.; and Jain, P. 2018. Multiple instance learning for efficient sequential data classification on resource-constrained devices. In *Advances in Neural Information Processing Systems*, 10976–10987.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Pérez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* 89(1-2):31–71.

Fabian Caba Heilbron, Victor Escorcia, B. G., and Niebles, J. C. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 961–970.

Feng, J., and Zhou, Z.-H. 2017. Deep miml network. In *Thirty-First AAAI Conference on Artificial Intelligence*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In Dy, J., and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 2127–2136.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 951–958. IEEE.

LeCun, Y.; Cortes, C.; and Burges, C. J. 1998. The mnist database of handwritten digits, 1998. *URL http://yann. lecun. com/exdb/mnist* 10:34.

Liu, G.; Wu, J.; and Zhou, Z.-H. 2012. Key instance detection in multi-instance learning. In *Asian Conference on Machine Learning*, 253–268.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, 289–297.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Nguyen, N. 2010. A new svm approach to multi-instance multi-label learning. In *2010 IEEE International Conference on Data Mining*, 384–392. IEEE.

Pathak, D.; Shelhamer, E.; Long, J.; and Darrell, T. 2014. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, 20–36. Springer.

Wang, X.; Yan, Y.; Tang, P.; Bai, X.; and Liu, W. 2018. Revisiting multiple instance neural networks. *Pattern Recognition* 74:15–24.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4651–4659.

Zhang, M.-L. 2010. A k-nearest neighbor based multi-instance multi-label learning algorithm. In *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, volume 2, 207–212. IEEE.

Zhou, Z.-H., and Zhang, M.-L. 2007. Multi-instance multi-label learning with application to scene classification. In *Advances in neural information processing systems*, 1609–1616.

Zhou, Z.-H.; Zhang, M.-L.; Huang, S.-J.; and Li, Y.-F. 2012. Multi-instance multi-label learning. *Artificial Intelligence* 176(1):2291–2320.