ACTIVE DEEP MULTIPLE INSTANCE LEARNING

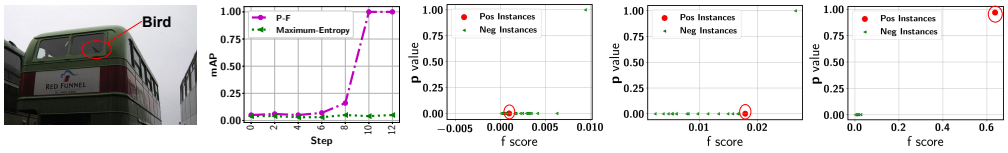
Anonymous authors
Paper under double-blind review

ABSTRACT

State-of-the-art multiple instance learning (MIL) models achieve competitive performance at the bag level. However, instance-level prediction, which is essential for many important applications, remains largely unsatisfactory. We propose a novel active deep multiple instance learning (ADMIL) model that samples a small subset of informative instances for annotation, aiming to significantly boost the instance-level prediction. A variance regularized loss function is designed to properly balance the bias and variance of instance-level predictions, aiming to effectively accommodate the highly imbalanced instance distribution in MIL and other fundamental challenges. Instead of directly minimizing the variance regularized loss that is non-convex, we optimize a distributionally robust bag level likelihood as its convex surrogate. The robust bag likelihood provides a good approximation of the variance based MIL loss with a strong theoretical guarantee. It also automatically balances bias and variance, making it effective to identify the potentially positive instances to support active sampling. The robust bag likelihood can be naturally integrated with a deep architecture to support deep model training using mini-batches of positive-negative bag pairs. Finally, a novel P-F sampling function is developed that combines a probability vector and predicted instance scores, obtained by optimizing the robust bag likelihood. By leveraging the key MIL assumption, the sampling function can explore the most challenging bags and effectively detect their positive instances for annotation, which significantly improves the instance-level prediction. Experiments conducted over multiple real-world datasets clearly demonstrate the state-of-the-art instance-level prediction achieved by the proposed ADMIL model.

1 INTRODUCTION

Multiple Instance Learning (MIL) offers an attractive weakly supervised learning paradigm, where instances are naturally organized into bags and training labels are assigned at the bag level to reduce the annotation cost (Dietterich et al., 1997; Settles et al., 2008; Li & Vasconcelos, 2015b; Sultani et al., 2018). State-of-the-art MIL models achieve competitive performance at the bag level. However, instance-level prediction, which is essential for many important applications (*e.g.,* anomaly detection from surveillance videos (Sultani et al., 2018) and medical image segmentation (Ilse et al., 2018)) remains largely unsatisfactory. In MIL, a bag is considered to be positive if at least one of the instances is positive otherwise negative (Dietterich et al., 1997; Haußmann et al., 2017). To achieve a high bag level prediction, most existing MIL models effectively leverage this key MIL assumption by focusing on the most positive instance from a positive bag that is mainly responsible for determining the bag label (Andrews et al., 2002; Kim & Torre, 2010; Sultani et al., 2018; Haußmann et al., 2017). However, they suffer from two major limitations, which lead to poor instance-level predictions. First, solely focusing on the most positive instance is sensitive to outliers, which are negative instances that look very different from other negative ones (Carbonneau et al., 2018). As a result, these instances may be wrongly assigned a high score indicating they are positive. Second, there may be multiple types (*i.e.,* multimodal) of positive instances in a single bag (*e.g.,* different types of anomalies in a surveillance video or different types of skin lesions in a dermatology image). Thus, focusing on a single most positive instance will miss other positive ones. Both cases will result in a low instance-level prediction performance. A possible solution to improve the detection of positive instances is to consider the top-k most positive instances. However, the number of positive instances may vary significantly across different bags and applying the same k to all bags may be inappropriate. Furthermore, finding an optimal k for each bag is highly challenging as it takes discrete values.



(a) SAMPLE BAG B_1 (b) MAP Score (c) Step 0 (d) Step 8 (e) Step 10

Figure 1: (a) Example of a challenging bag; (b) MI-AL performance on instance-level predictions; (c)-(e) Prediction scores of instances in the bag in different MI-AL steps.

The underlying reason for the less accurate instance-level prediction is due to the lack of instance labels. For positive instances that are relatively rare across bags, detecting them by only relying on bag labels is inherently challenging as the weakly supervised signal (*i.e.,* the bag label) cannot be propagated to the instance level without sufficient statistical evidence. One promising direction to tackle this fundamental challenge is to augment MIL with active learning (AL). Multiple instance AL (or MI-AL) aims to select a small number of informative instances to improve the instance level prediction in MIL. In most MIL problems, the data is highly imbalanced at the instance level, where the positive ones are much more sparse. Since the positive instances usually carry more important information, a primary goal of MI-AL is to effectively sample the positive instances from a candidate pool dominated by the negative ones. If a true positive instance can be sampled and labeled, it can help to improve the identification of other similar positive instances in the same and different bags, which will significantly improve the instance-level predictions.

However, existing MIL models may easily miss some rare positive instances (Sultani et al., 2018). They may also focus on the wrongly identified negative instances due to their sensitivity to outliers or incapability of handling multimodal bags. Thus, the true positive instances may be assigned a low prediction score, indicating that they are predicted as negative with a high confidence. As a result, commonly used uncertainty based sampling will miss these important instances. Figure 1 (a) shows a challenging bag, which is an image that contains the shadow of a bird (as the positive class). The positive instances are patches that cover (part of) the bird shadow. Figure 1 (b) shows that combining uncertainty sampling with a maximum score based MIL model (the green curve) is not able to sample effectively so that instance-level prediction remains very low over the AL process. Figure 1 (c) further confirms that the initial prediction score (F-score) of the positive instance is close to 0, making it hard to be sampled.

We propose a novel MI-AL model for effective instance sampling to significantly boost the instance-level prediction in MIL. We design an unique variance regularized MIL loss that encourages a high variance of the prediction scores to address bags with a highly imbalanced instance distribution and/or those with outliers and multimodal scenarios. Since the variance regularizer is non-convex, we propose to optimize a distributionally robust bag likelihood (DRBL), which provides a good convex approximation of the variance based loss with a strong theoretical guarantee. The DRBL automatically adjusts the impact of the bag-level variance, making it more effective to detect potentially positive instances to support active sampling. It can also be naturally integrated with a deep architecture to support deep MIL model training using mini-batches of positive-negative bag pairs. Finally, a novel P-F sampling function is developed that combines a probability vector (*i.e.,* $\mathbf{p}$) and predicted instance scores (*i.e.,* $\mathbf{f}$), obtained by optimizing the robust bag likelihood. By leveraging the key MIL assumption, the sampling function can explore the most challenging bags and effectively detect their positive instances for annotation, which significantly improves the instance-level prediction. Novel batch-mode sampling is developed to work seamlessly with the deep MIL, leading to a powerful active deep MIL (ADMIL) model to support sampling of high-dimensional data used in most MIL applications. Figure 1 (b) shows the proposed model (purple curve) that significantly improves instance predictions. Figures 1 (c)-(e) shows P-F sampling dynamically updates the probability $\mathbf{p}$ and score $\mathbf{f}$ values to effectively sample the positive instance from the highly challenging bag in a few steps.

Our main contribution includes: (i) a unique variance regularized MIL loss and its convex surrogate that address inherent MIL challenges to best support active sampling, (ii) a novel P-F sampling function to effectively explore most challenging bags with rare positive instances, (iii) mini-batch training and batch-mode active sampling to support ADMIL in broader MIL applications, and (iv) state-of-the-art instance prediction performance in MIL while maintaining low instance annotations.

2 RELATED WORK

Multiple Instance Learning (MIL). Existing supervised learning models have been leveraged to tackle MIL problems, including SVM (Andrews et al., 2002), boosting (Xu & Frank, 2004), graph-based models (Zhou et al., 2009), conditional random field (Deselaers & Ferrari, 2010) and Gaussian Processes (Haußmann et al., 2017; Kim & Torre, 2010). Other approaches try to relax the MIL assumption, which allows positive instances in a negative bag to handle noisy bags (Li & Vasconcelos, 2015a). As MIL is commonly applied to problems, such as video anomaly detection and image segmentation that involve high dimensional data, deep neural networks have become a popular choice for training MIL models (Ilse et al., 2018; Sultani et al., 2018). Despite the significant progress made so far, most existing models focus on improving the bag-level predictions. As a result, instance-level performance still falls short in meeting the high standard in critical applications (Sultani et al., 2018; Ilse et al., 2018; Haußmann et al., 2017). The proposed ADMIL model aims to fill out this critical gap by augmenting MIL with novel active sampling strategies to significantly boost instance predictions using limited labeled instances to maintain a low annotation cost.

Active Learning (AL). Uncertainty and margin based measures are commonly leveraged in existing AL models to achieve efficient data sampling (Tong & Koller, 2001; Roy & McCallum, 2001; Holub et al., 2008; Culotta & McCallum, 2005; Rajan et al., 2008; Joshi et al., 2009). Distributionally robust optimization has also been adopted in multi-class AL to address sampling bias and imbalanced data distribution (Zhu et al., 2019). Deep learning (DL) models are good candidates for AL because of their high-dimensional data processing and automatic feature extraction capability. Existing models mainly target at improving uncertainty quantification of the network for reliable sampling (Wang et al., 2016; Gal & Ghahramani, 2015; 2016; Kendall et al., 2015; Leibig et al., 2017). Batch-mode sampling is commonly used in active DL to avoid frequent model re-training. It focuses on constructing representative batches to avoid redundant information given by similar instances (Kirsch et al., 2019; Sener & Savarese, 2017; Ash et al., 2019). AL in the MIL setting has been rarely investigated. One exception is the MI logistic model and its three uncertainty measures to simultaneously consider both instance and bag level uncertainty (Settles et al., 2007). However, uncertainty sampling is ineffective to explore challenging bags, where all instances are confidently predicted as negative. In addition, the original model is a simple linear model, which does not provide sufficient capacity for high-dimensional data. There is no systematic way to support batch-mode sampling, either. An AL framework is developed for MIL tasks in (Yuan et al., 2021), however, sampling is conducted at the bag level (*i.e.,* choosing bags instead of instances). Thus, it is essentially a multi-label AL model, aiming to improve the bag-level predictions with fewer annotated bags. This is fundamentally different from the design goal of ADMIL.

3 ACTIVE DEEP MULTIPLE INSTANCE LEARNING

Let $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ denote a set of instances associated with each bag $\mathcal{B}$, where each $\mathbf{x}_i \in \mathbb{R}^D$ is a D-dimensional feature vector. Let $t_{\mathcal{B}} \in \{+1, -1\}$ indicate the bag type. All symbols are summarized by Table 1 in Appendix A. Following the standard MIL assumption discussed earlier, active sampling will focus on instances in the positive bags as all instances in a negative bag are negative. We also allow the number of instances to vary from one bag to another.

3.1 VARIANCE REGULARIZATION FOR EFFECTIVE MI ACTIVE SAMPLING

Let $\mathbf{x}_i^+$ (or $\mathbf{x}_j^-$) be the i^{th} (or j^{th}) instance in a positive bag $\mathcal{B}_{pos}$ (or a negative bag $\mathcal{B}_{neg}$). Following the MIL assumption, a commonly used loss function for training deep MIL models is to make the maximum prediction score of instances from a positive bag to be higher than a negative bag (Sultani et al., 2018):

$$\mathcal{L}^{MS}(\mathcal{B}_{pos}, \mathcal{B}_{neg}) = \left\{1 - \max_{i \in \mathcal{B}_{pos}} [f(\mathbf{x}_i^+; \mathbf{w})] + \max_{j \in \mathcal{B}_{neg}} [f(\mathbf{x}_j^-; \mathbf{w})]\right\}_+ \tag{1}$$

where $f(\mathbf{x}; \mathbf{w}) \in [0, 1]$ is the prediction score of instance $\mathbf{x}$ provided by a deep neural network parameterized by $\mathbf{w}$ and $[a]_+ = \max\{0, a\}$. We omit $\mathbf{w}$ from $f(\mathbf{x}; \mathbf{w})$ in its future reference to keep the notation uncluttered. The above objective function aims to maximize the gap between the maximum prediction score of instances from a positive bag and maximum score from a negative bag. Model training can be performed by sampling pairs of positive and negative bags $(\mathcal{B}_{pos}, \mathcal{B}_{neg})$, using their bag-level labels to evaluate the loss, and performing back-propagation. The maximum score based MIL (referred to as MS-MIL) models are designed primarily for bag label prediction as it aims

to identify a single most positive instance from a positive bag and maximizes its prediction score. In this way, it fully leverages the MIL assumption (*i.e.,* at least one positive instance in $\mathcal{B}_{pos}$) and the weakly supervised signal (*i.e.,* bag-level label).

As discussed earlier, MS-MIL and it top-k extensions suffers from key limitations that impact their instance-level prediction performance. Meanwhile, they provide inadequate support to sample the most informative instances to enhance the instance predictions. Inspired by the recent advances in learning theory to automatically balance bias and variance in risk minimization (Duchi & Namkoong, 2019), we propose a novel variance regularized MIL loss function to capture the inherent characteristics of MIL, aiming to collectively address highly imbalanced instance distribution, existence of outliers, and multimodal scenarios. As a result, minimizing the new MIL loss can effectively improve the prediction scores of the positive instances, making them easier to be sampled for annotation by the proposed sampling function. In particular, the variance regularized loss introduces *two novel changes* to (1), which are formalized below:

$$\mathcal{L}^{VAR}(\mathcal{B}_{pos}, \mathcal{B}_{neg}) = \left\{ 1 - \left[\frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i^+) + C \sqrt{\frac{\text{Var}_n[f(X^+)]}{n}} \right] + \max_{j \in \mathcal{B}_{neg}} \left[f(\mathbf{x}_j^-) \right] \right\}_+ \tag{2}$$

where $\forall i \in [1, n], \mathbf{x}_i^+ \in \mathcal{B}_{pos}$, n is the size of $\mathcal{B}_{pos}$, Var_n is the empirical variance of $f(X^+)$ with X^+ being a random variable representing an instance from a positive bag, and C is a hyperparameter to balance the mean score and the variance.

The first key change is to use the mean score to replace the maximum score in (1), which avoids the model to only focus on the most positive instance in a bag to make it robust to outliers and multimodal scenarios. Since positive bags are guaranteed to include positive instances and instances in a negative bag are all negative, it is desirable that the mean score for a positive bag should be high. Maximizing the mean score in a positive bag using a complex model (*e.g.,* a deep neural network) could effectively reduce the training loss (by reducing the bias) in estimating the bag-level labels. However, using the mean score alone is problematic as most instances in a positive bag are usually negative in a typical MIL setting. As a result, such a low bias model will lead to a very high false positive rate, which negatively impacts the overall instance-level prediction. The proposed loss function addresses this issue through the novel variance term, which effectively handles the highly imbalanced instance distribution. With only a small number of instances being truly positive, the empirical variance Var_n for the bag should be high due to the large deviation of a small number of high scores from the majority of low scores. It is worth to note that the variance term in (2) plays a distinct role than risk minimization in standard supervised learning, where it is minimized to control the estimation error. In contrast, the variance in (2) is encouraged to be large to allow a small set of instances in a bag to be positive, aiming to precisely capture the imbalanced distribution. To our best knowledge, this is the first bias-variance formulation in the MIL setting. Conducting MI-AL using variance regularization still faces two remaining challenges. First, its effectiveness hinges on an optimal balance between the mean score and the empirical variance, which is controlled by the hyperparameter C. Similar to the standard supervised learning, there lacks a systematic way of setting such a hyperparameter to achieve an optimal trade-off. Second, the variance term is non-convex with multiple local minima (Duchi & Namkoong, 2019), which makes model training much more difficult and time-consuming. Thus, it is not suitable for real-time interactions to support active sampling.

3.2 Distributionally Robust Bag Likelihood

To address the challenges as outlined above, we propose to formulate a distributionally robust bag level likelihood (DRBL) as a convex surrogate of the variance regularized loss in (2). By extending the distributionally robust optimization framework developed for risk minimization in supervised learning (Namkoong & Duchi, 2017; Duchi & Namkoong, 2019), we theoretically prove the equivalence between DRBL and variance regularization with high probability. Being convex, DRBL is easier to optimize that facilitates MIL model training to support fast active sampling. Furthermore, by setting a proper uncertainty set as introduced next, we show that the parameter C is directly obtained when optimizing the robust bag likelihood, where the instance distribution in the bag is constrained by the uncertainty set. As a result, it achieves automatic trade-off between the mean prediction score and the variance.

We first introduce a probability vector $\mathbf{p} = (p_1, ..., p_n)^\top$, where $\sum_i p_i = 1, p_i \geq 0, \forall i \in \{1, ..., n\}$ and let p_i denote the probability that instance $\mathbf{x}_i^+ \in \mathcal{B}_{pos}$ can represent the bag. We further introduce

a binary indicator vector $\mathbf{z} = (z_1, ..., z_n)^\top$, where $p(z_i = 1) = p_i$. Let Y be a binary random variable that denotes the bag label. Conditioning on all the instances in the bag, the (conditional) bag likelihood for bag $\mathcal{B}_{pos}$ is given by $p(Y = 1|\mathbf{z}, \mathbf{f}) = \prod_i f(\mathbf{x}_i^+)^{z_i}$, where $\mathbf{f} = (f(\mathbf{x}_1^+), ..., f(\mathbf{x}_n^+))^\top$. By integrating out the indicator variables, we have the marginal bag likelihood as $p(Y = 1|\mathbf{p}, \mathbf{f}) = \sum_i p_i f(\mathbf{x}_i^+)$. Instead of letting a single most positive instance to determine the bag label, where $p(y = 1|\mathbf{p}, \mathbf{f}) = f(\mathbf{x}_k^+)$ with $k = \arg\max_i f(\mathbf{x}_i^+)$, which is equivalent to MS-MIL, or assigning equal probability to each instance (*i.e.,* $p_i = 1/n$), which is equivalent to the mean score, we introduce an uncertainty set $\mathcal{P}_n$ that allows $\mathbf{p}$ to deviate from a uniform distribution to some extent:

$$\mathcal{P}_n := \left\{ \mathbf{p} \in \mathbb{R}^n, \mathbf{p}^\top \mathbb{1} = 1, 0 \leq \mathbf{p}, D_f \left(\mathbf{p} || \frac{\mathbb{1}}{n} \right) \leq \frac{\lambda}{n} \right\} \tag{3}$$

where $D_f(\mathbf{p}||\mathbf{q})$ is the f-divergence between two distributions $\mathbf{p}$ and $\mathbf{q}$, $\mathbb{1}$ is a n-dimensional unit vector, and λ controls the extent that $\mathbf{p}$ can deviate from a uniform vector, which essentially corresponds to the imbalanced instance distribution in the bag. Note that $\mathcal{P}_n$ only specifies a neighborhood that $\mathbf{p}$ may deviate from a uniform distribution. Since $\mathcal{P}_n$ is a convex set, an optimal $\mathbf{p}$ can be easily computed for each specific bag by optimizing the robust bag likelihood according to its specific imbalanced instance distribution. This is fundamentally more advantageous than a top-k approach, where k is discrete and hard to optimize. Next, we show that the optimal robust bag likelihood is equivalent to the variance regularized mean prediction score with high probability, which allows us to define a new MIL loss based on DRBL.

Theorem 1. *Let X^+ be a random variable representing an instance from a positive bag, $f(X^+) \in [0, 1]$ is the score assigned to an instance, $\sigma^2 = Var[f(X^+)]$ and $Var_n[f(X^+)]$ denote the population and sample variance of $f(X^+)$, respectively, and D_f takes the form of χ^2-divergence. For a fixed λ and with $n \geq \max(2, \frac{\lambda}{\sigma^2} \max(8\sigma, 44))$, we have*

$$\max_{\mathbf{p} \in \mathcal{P}_n} \sum_{i=1}^n p_i f(\mathbf{x}_i^+) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i^+) + \sqrt{\frac{\lambda Var_n[f(X^+)]}{n}} \tag{4}$$

with probability at least $1 - \exp\left(-\frac{7n\sigma^2}{20}\right)$, where $\mathcal{P}_n$ is an uncertainty set defined by (3)*.*

It is worth to note that given the highly imbalanced positive instances in a typical MIL setting, the true variance σ^2 should be high. For a bag with a decent size, it guarantees the equivalence in (4) with high probability. Furthermore, maximizing the robust bag likelihood given on the l.h.s. of (4) assigns $C = \sqrt{\lambda}$, which automatically adjusts the impact of variance based on the uncertainty set. In Theorem 1, D_f takes the form of χ^2-divergence between two distributions. Theorem 2 in Appendix C further generalizes this result to the KL-divergence. Detailed proofs are given in Appendix C. Leveraging the key result from Theorem 1, we formulate a DRBL-based MIL loss,

$$\mathcal{L}^{DRBL}(\mathcal{B}_{pos}, \mathcal{B}_{neg}) = \left\{ 1 - \max_{\mathbf{p} \in \mathcal{P}_n} \left[\sum_{i=1}^n p_i f(\mathbf{x}_i^+) \right] + \max_{j \in \mathcal{B}_{neg}} \left[f(\mathbf{x}_j^-) \right] \right\}_+ \tag{5}$$

The DRBL loss offers a very intuitive interpretation on the newly introduced probability vector $\mathbf{p}$. Since it can deviate from the uniform distribution as specified by the uncertainty set $\mathcal{P}_n$, each entry p_i essentially corresponds to the contribution (or weight) of $\mathbf{x}_i^+$ to the bag likelihood (being positive). As a result, to maximize the robust bag likelihood, instances with a higher prediction score should receive a higher weight. Meanwhile, constrained by $\mathcal{P}_n$, multiple instances will contribute to the bag likelihood with a sizable weight as $\mathbf{p}$ cannot deviate too much from being uniform. Hence, their prediction scores will simultaneously be brought up by the model. This makes DRBL robust to the outlier and multimodal cases as it increases the chance for the true positive instances or multiple types of true positive instances to be assigned a high prediction score. This provides fundamental support to the proposed P-F active sampling function that combines the probability vector $\mathbf{p}$ and the prediction score $\mathbf{f}$ in a novel way to choose the most informative instances in a bag for annotation.

3.3 P-F Active Sampling

Since we have the prediction score $f(\mathbf{x}_i^+) \in [0, 1]$, it can be naturally interpreted as the probability of instance $\mathbf{x}_i^+$ being positive. As a result, a straightforward way to perform uncertainty based instance sampling is to compute the f-score based entropy of the instances, referred to F-Entropy:

$$\mathbf{x}_* = \arg\max_{i \in \mathcal{B}_{pos}} H[f(\mathbf{x}_i^+)], H[f(\mathbf{x}_i^+)] = -[f(\mathbf{x}_i^+) \log f(\mathbf{x}_i^+) + (1 - f(\mathbf{x}_i^+)) \log(1 - f(\mathbf{x}_i^+))] \tag{6}$$

Since the sampled instance has the largest prediction uncertainty (according to F-Entropy), labeling such an instance can effectively improve the model's instance-level performance. Active sampling using (6) is straightforward, which involves evaluating $H[f(\mathbf{x}^+)]$ for all the instances from positive training bags (note that all the instances in a negative bag are negative). Since we consider a deep learning model to better accommodate high-dimensional data, sampling one instance at a time requires frequent model training, which is computationally expensive. Instead, we sample a small batch of instances in each step based on their predicted F-Entropy. It is worth to note that, due to the highly imbalanced instance distribution, the majority of the prediction scores, including many positive instances, may be very low. The goal is to assign a relatively higher score to the potentially positive instances so that their entropy is not too low, indicating a confident negative prediction, which will be missed by the sampling function.

As discussed earlier, using the robust bag likelihood as the MIL loss can directly benefit instance sampling by increasing the chance to assign a higher prediction score to a positive instance so that it is more likely to be sampled. However, F-Entropy sampling still suffers from two major limitations. First, for some very difficult bags, such as the sample image shown in Figure 1 (a), identifying the positive instances (*e.g.,* the patch in the image containing the shadow of a bird) can be highly challenging. As a result, they may be assigned a very low f score. In fact, as shown in Figure 1 (c), all the instances in this bag receive a very low score with the highest less than 0.01, leading to a very low entropy. Some additional examples of challenging bags from the 20NewsGroup dataset are shown in Figure 6 of Appendix B, where all the instances are predicted with a very low score. Hence, all these instances are predicted as negative with low uncertainty, making them less likely to be chosen by entropy based sampling. Second, since batch-mode sampling is adopted to reduce the training cost of a deep network, it is essential to diversify the selected instances in the same batch to minimize the annotation cost. However, choosing data instances solely based on their predicted entropy may lead to the annotation of similar instances, which is not cost-effective.

The proposed P-F active sampling overcomes the above two limitations simultaneously through effective bag exploration by combining the probability vector $\mathbf{p}$ and the prediction score $\mathbf{f}$ through a $\min\max$ function according to their distinct roles in a bag. The key design rationale of P-F sampling is rooted in the standard MIL assumption that ensures at least one positive instance in each positive bag to guide effective bag exploration. Both $\mathbf{p}$'s and $\mathbf{f}$'s along with the bag structure are dynamically updated during bag exploration to increase the chance of sampling the positive instances in an under-explored bag. A hybrid loss function further utilizes labels of sampled negative instances in the same bag to boost the prediction scores of the positive instances. More specifically, let B be the total number of positive training bags, P-F sampling will choose the following data instance:

$$\mathbf{x}_*^{PF} = \arg \min_{b \in \{1,\dots,B\}} f(\mathbf{x}_{b_*}^+), \quad \text{and } b_* = \arg\max \mathbf{p}_b \tag{7}$$

where $\mathbf{p}_b$ is the probability vector of bag b. For each bag, the sampling function first identifies the instance $\mathbf{x}_{p_*}^+$ with the largest p value in each bag. Such an instance can be regarded as the most representative instance in the bag as it makes the largest contribution (according to $\mathbf{p}_b$) to the bag likelihood. According to the prediction score of $\mathbf{x}_{p_*}^+$, we can categorize bags into three groups: (1) easy bags, where $f(\mathbf{x}_{b_*}^+)$ takes a large value, indicating that the model makes confidently correct predictions, (2) confusing bags, where $f(\mathbf{x}_{b_*}^+)$ is reasonably large but uncertain, indicating the model is still confusing about its prediction, and (3) difficult bags, where $f(\mathbf{x}_{b_*}^+)$ is very low, indicating the model makes confidently wrong predictions. It is desirable to sample from both confusing and difficult bags as the model already makes accurate instance predictions for easy bags. Sampling instances from the confusing bags can be achieved through the proposed F-Entropy as the model makes uncertain predictions, which leads to a high entropy. Finally, sampling from the difficult bags are fundamentally more challenging due to low prediction scores for the entire bag. However, the MIL assumption provides a general direction for bag-level exploration of positive instances as there must be at least one positive instance in each positive bag. The P-F sampling function in (7) chooses the representative instance from the bag with the lowest prediction score. Such an instance is guaranteed to be sampled from an under-explored (*i.e.,* difficult) bag as it has the lowest prediction score despite being predicted as the most positive instance in the bag.

Extension to the batch-mode sampling is conducted in two directions, within bag and across bags, for more effective exploration while ensuring diversity of the sampled instances. First, instead of only sampling the most positive instance from the identified under-explored bag, we propose to sample

$k > 1$ instances as the positive instances may be ranked lower than multiple negative instances in the bag according to the current prediction scores (see Figure 1 (c) for an example). This helps to more effectively explore very difficult bags. To ensure diversity among the sampled instances, we keep k small but sample across multiple bags simultaneously. Only bags with a max prediction score $f(\mathbf{x}^+_{b_*})$ less than a threshold (0.3 is used in our experiments) will be explored as these represent the difficult bags as discussed above. For bags with a larger $f(\mathbf{x}^+_{b_*})$, they are either easy bags or confusing bags that can be effectively sampled using F-Entropy. Our overall P-F sampling function integrates bag exploration and F-Entropy and gives priority to the former to perform diversity-aware bag exploration first. As more bags are successfully explored along with MI-AL, less instances will be sampled by exploration and the focus will be naturally shifted to F-Entropy to perform model fine-tuning. The detailed sampling process is summarized by Algorithm 1 in Appendix D.

Similar to AL in standard supervised learning, the sampled annotated instances should be used to improve the model prediction performance. However, the MIL loss primarily focuses on the bag-level labels due to the lack of instance labels. To this end, we propose a hybrid loss function that integrates the bag and instance labels. Let $\mathbf{X}^l = \{\mathbf{x}^l_1, \mathbf{x}^l_2, ..., \mathbf{x}^l_m\}$ be the m labeled instances queried by the proposed active learning function and $\mathbf{t}^l = \{t^l_1, t^l_2, ..., t^l_m\}$ with $t^i_l \in \{0, 1\}$ be the corresponding instance labels. We formulate a supervised binary cross-entropy (BCE) loss as

$$L^{BCE}(\mathbf{X}^l, \mathbf{t}^l) = -\frac{1}{m}\sum_{i=1}^{m}\left[t^l_i \log(f(\mathbf{x}^l_i)) + (1 - t^l_i)\log(1 - f(\mathbf{x}^l_i))\right] \tag{8}$$

It is clear that the sampled positive instances provide important supervised signals so that the model will predict a high score for similar positive instances, which will directly benefit instance-level prediction. In contrast, the sampled negative instances, especially those chosen from the under-explored bags, contribute less to improve the prediction performance as their original prediction scores are already low. However, they play a subtle but essential role to achieve more effective bag-level exploration. First, if a sampled instance is labeled as negative, it will be removed from the bag, which does not violate the MIL assumption. Meanwhile, since we have $\sum_i p_i = 1$, the p values will be redistributed and the chance for each remaining instance to be sampled is therefore increased. Furthermore, the BCE loss will further bring down the prediction scores of negative instances that are similar to the sampled one. This may help to improve the score of the positive instance so that it can have a higher chance to be sampled in the future. Finally, the hybrid loss that combines the MIL loss and the supervised loss is used to retrain the model after a new batch of instances are queried:

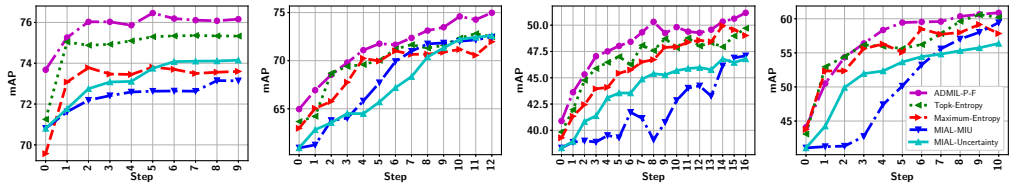
$$\mathcal{L}^{Hybrid}(\mathcal{B}_{pos}, \mathcal{B}_{neg}) = \mathcal{L}^{DRBL}(\mathcal{B}_{pos}, \mathcal{B}_{neg}) + \beta\mathcal{L}^{BCE}(\mathbf{X}^l, \mathbf{t}^l) \tag{9}$$

where β is used to trade-off bag-level and instance-level losses.

4 EXPERIMENTS

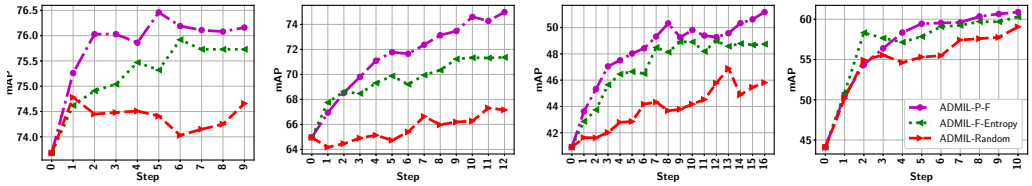
We conduct extensive experimentation over multiple real-world MIL datasets to justify the effectiveness of the proposed ADMIL model. The purpose of our experiments is to demonstrate: (i) the state-of-the-art instance prediction performance of ADMIL by comparing with existing competitive baselines, (ii) effectiveness of the proposed P-F active sampling function through comparison with other sampling mechanisms, (iii) impact of key model parameters through a detailed ablation study, and (iv) qualitative evaluation through concrete examples to provide deeper and intuitive insights on the working rationale of the proposed model.

Datasets. Our experiments involve four datasets covering both textual and image data: 20NewGroup (Zhou et al., 2009), Cifar10 (Krizhevsky, 2009), Cifar100 (Krizhevsky, 2009), and Pascal VOC (Everingham et al., 2015). For 20NewsGroup, the dataset is already available in the MIL setting, which consists of 20 topics where each topic contains 50 positive and 50 negative bags. For Cifar10 and Cifar100 datasets, bags are constructed by treating each image as an instance. For Cifar10, images corresponding to digits '1', '2', and '5' are regarded as a positive instance otherwise negative. In case of Cifar100, images in superclass 2 are treated as positive and the rest as negative. In Pascal VOC, we perform image segmentation so each image is regarded as a bag and corresponding patches cropped from the image are treated as instances. In our experiments, images containing birds as a positive bags and others as negative. Table 2 that summarizes the bag statistics and additional details for all datasets are provided in Appendix E.



(a) 20NewsGroup (b) Cifar10 (c) Cifar100 (d) Pascal VOC

Figure 2: MI-AL performance comparison



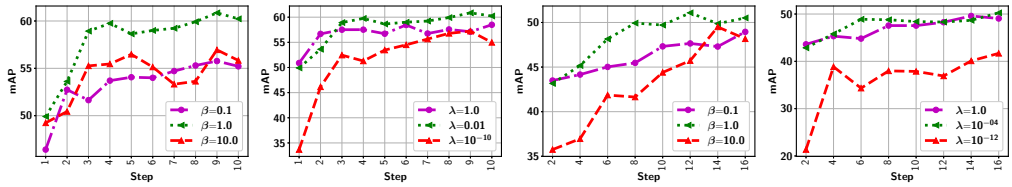
(a) 20NewsGroup (b) Cifar10 (c) Cifar100 (d) Pascal VOC

Figure 3: Effectiveness of P-F active sampling

Evaluation metric and model training. To assess the model performance, we report the instance-level mean average precision (mAP) score, which summarizes a precision-recall curve as a weighted mean of precision achieved at each threshold, with the increase in recall from the previous threshold as the weight. mAP explicitly places much stronger emphasis on the correctness of the few top ranked instances than other metrics (*e.g.,* AUC) (Su et al., 2015). This makes it particularly suitable for instance prediction evaluation as a small subset of instances with the highest prediction scores will eventually be identified as positive for further inspection (by human experts) with the rest being ignored. For Cifar10, Cifar100, and Pascal VOC datasets, we extract the visual features from the second-to-the last layer of a VGG16 network pre-trained using the imagenet dataset, yielding a 4,096 dimensional feature vector for each instance. For 20NewsGroup, we use the available 200-dimensional feature vector. In terms of network architecture, we use a 3-layer FC neural network. The first layer has 32 units followed by 16 units and 1 unit FC layers. We adopt 60% dropout between FC layers. ReLU and sigmoid activations are used for the first and last FC layers. More details regarding the learning rate and hyperparameter setting are presented in the Appendix.

Performance comparison. To demonstrate the instance prediction performance achieved by the proposed ADMIL model, we compare it with competitive baselines. First, the two MI-AL sampling strategies: MIAL-Uncertainty and MIAL-MIU (Settles et al., 2007), from the MI logistic model are included. Since our datasets involve high-dimensional data, we replace the original linear model by the exact DNN model used in our ADMIL so we can focus on comparing MI active sampling. The EGL sampling technique in (Settles et al., 2007) was not included due to the prohibitive computational cost to evaluate the gradient of each instance output with respect to the large number of DNN parameters. We also implement an MS-MIL model and its top-k variant with uncertainty sampling using entropy. Given the different sizes of the datasets, we query maximum 15 instances per step in 20NewsGroup, 30 instances in Pascal VOC, and 150 instances in Cifar10 and Cifar100. Figure 2 shows the MI-AL curves for all four datasets. ADMIL achieves the best performance in all cases. For most datasets, it shows a much better initial performance, which results from the proposed DRBL-based MIL loss that significantly benefits MIL performance in passive learning. Overall the entire MI-AL process, ADMIL consistently stays the best and converges to a higher point in the end for all datasets. For the Pascal VOC, the top-k MIL model with entropy sampling achieves closer performance towards the end, which is mainly due to the limited positive instances in this dataset. Hence, no testing bags contain similar positive instances in the challenging bags that are explored by P-F sampling. While ADMIL achieves much better instance predictions in those bags, the advantage does not transfer to the testing bags. Our qualitative study will provide a more detailed analysis on this.

Effectiveness of active sampling. To demonstrate the effectiveness of the proposed P-F active sampling function, we compare it with two other sampling methods, F-Entropy and random sampling, while keeping all other parts of the model the same. As shown in Figure 3, P-F sampling clearly outperforms others with a large margin in the first three datasets. It advantage over F-Entropy is smaller on Pascal VOC due the same reason as explained above. The performance gain is mainly attributed to the effective exploration of P-F sampling over the most challenging bags.



(a) $\lambda = 0.01$ (b) $\beta = 1$ (c) $\lambda = 0.01$ (d) $\beta = 1$

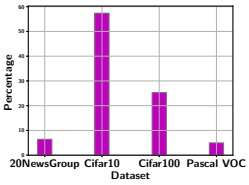
Figure 4: Impact of key model parameters: (a-b) Pascal VOC; (c-d) Cifar100

(a) SAMPLE BAG B_2

(b) SAMPLE BAG B_3

Bag	P-F	F-Entropy
B_1	1.00	0.04
B_2	0.53	0.07
B_3	0.64	0.55
B_1	Shadow of a bird	
B_2	Side view of a bird	
B_3	Part of a bird	

(c) MAP SCORES



(d) Percentage of TP bags

Figure 5: (a-b) Poorly explored bags in Pascal VOC; (b) Description of these bags and their mAP scores; (d) Additional true positive bags successfully explored by P-F sampling

Ablation Study. Figure 4 demonstrates the impact of λ and β. In particular, λ can be set according to the imbalanced instance distribution within bags, where a larger λ corresponds to a higher imbalanced distribution. We vary λ in $[10^{-10}, 1]$ and since most bags in the MIL setting are highly imbalanced, $\lambda \in [10^{-06}, 10^{-02}]$ gives very good performance in general. Figures 4 (b) and (d) show that $\lambda = 0.01$ clearly outperforms too large (or small) λ values. As for β, placing less emphasis on an instance level loss (small β), we may not fully leverage labels of queried instances. Meanwhile, with too much emphasis on the instance level loss (large β), the model overly focuses on the limited queried instances with less attention to the bag labels. Therefore, a good balance (with $\beta = 1$) results in an optimal performance, shown in (a) and (c). More ablation study results are provided in Appendix E.

Qualitative analysis. To further justify why the proposed ADMIL model and its P-F sampling function work better than other competitive baselines, we provide a few illustrative examples to offer some deeper insights on its good performance. First, we show two additional challenging bags in addition to the one shown in Figure 1 (a). In particular, as shown in Figure 5 (a-b), B_2 presents a side view of a bird while only a small portion of the bird is visible in B_3. For those difficult cases, the model originally predicts all instances as a negative with high confidence. However, by coupling the P-F sampling and the hybrid loss in (9), the positive instances from those bags are successfully queried. Figure 5 (c) shows a clear advantage in the mAP scores between P-F sampling and F-Entropy. As a further evidence, we investigate the number of true positive (TP) bags being explored by both P-F sampling and F-Entropy. TP bags refer to those that the model is being able to query at least one true positive instance. Instead of reporting the actual number of bags, which is affected by the size of the dataset, we show the additional percentage TP bags being explored by P-F sampling in Figure 5 (d). It is worth to note that neither method tries to query the easy bags as their positive instances are correctly predicted with high confidence. The major difference is from the challenging bags and the percentage of these bags varies among different datasets. Nevertheless, P-F sampling consistently explores more effectively than F-Entropy across all datasets.

5 CONCLUSIONS

To tackle the low instance-level prediction performance of existing MIL models that is essential for many critical applications, we develop a novel MI-AL model to sample a small number of most informative instances, especially those from confusing and challenging bags, to enhance the instance-level prediction while keeping a low annotation cost. We propose to optimize a robust bag likelihood as a convex surrogate of a variance regularized MIL loss to identify a subset of potentially positive instances. Active sampling is conducted by properly balancing between exploring the challenging bags (through P-F sampling) and refining the model by sampling the most confusing instances (through F-Entropy). The design of the loss function naturally supports mini-batch training, which coupled with the batch-mode sampling, makes the MI-AL model work seamlessly with a deep neural network to support broader MIL applications that involve high-dimensional data. Our extensive experiments conducted on multiple MIL datasets show clear advantage over existing baselines.

REFERENCES

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

Marc-André Carbonneau, V. Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit.*, 77:329–353, 2018.

Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pp. 746–751, 2005.

Thomas Deselaers and Vittorio Ferrari. A conditional random field for multiple-instance learning. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 287–294, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Thomas G. Dietterich, R. Lathrop, and Tomas Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89:31–71, 1997.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019. URL `http://jmlr.org/papers/v20/17-750.html`.

M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Manuel Haußmann, Fred A. Hamprecht, and Melih Kandemir. Variational bayesian multiple instance learning with gaussian processes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 810–819, 2017. doi: 10.1109/CVPR.2017.93.

Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–8. IEEE, 2008.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, 2018.

Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, pp. 2372–2379. IEEE, 2009.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

Minyoung Kim and F. Torre. Gaussian processes multiple instance learning. In *ICML*, 2010.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *arXiv preprint arXiv:1906.08158*, 2019.

A. Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 2009.

Henry Lam. Robust sensitivity analysis for stochastic systems. *Math. Oper. Res.*, 41:1248–1275, 2016.

Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):1–14, 2017.

Weixin Li and N. Vasconcelos. Multiple instance learning for soft bags via top instances. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4277–4285, 2015a.

Weixin Li and Nuno Vasconcelos. Multiple instance learning for soft bags via top instances. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4277–4285, 2015b. doi: 10.1109/CVPR.2015.7299056.

Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/5a142a55461d5fef016acfb927fee0bd-Paper.pdf.

I.R. Petersen, M.R. James, and P. Dupuis. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control*, 45(3):398–412, 2000. doi: 10.1109/9.847720.

Suju Rajan, Joydeep Ghosh, and Melba M Crawford. An active learning approach to hyperspectral data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(4):1231–1242, 2008.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pp. 441–448, 2001.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *Advances in neural information processing systems*, 20:1289–1296, 2007.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2007/file/a1519de5b5d44b31a01de013b9b51a80-Paper.pdf.

Wanhua Su, Yan Yuan, and Mu Zhu. A relationship between the average precision and the area under the roc curve. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pp. 349–352, 2015.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27 (12):2591–2600, 2016.

Xin Xu and Eibe Frank. Logistic regression and boosting for labeled bags of instances. In *Advances in Knowledge Discovery and Data Mining*, volume 3056, 08 2004. doi: 10.1007/978-3-540-24775-3_35.

Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. *arXiv preprint arXiv:2104.02324*, 2021.

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-i.i.d. samples. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1249–1256, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161.

Dixian Zhu, Zhe Li, Xiaoyu Wang, Boqing Gong, and Tianbao Yang. A robust zero-sum game framework for pool-based active learning. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 517–526. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/zhu19a.html.

Appendix

Organization. In this appendix, we first summarize the major notations in Table 1. We then provide more detailed explanation on some challenging bags using the 20NewsGroup dataset as an example in Appendix B. We extend the main theoretical result given in Theorem 1 to another important type of f-divergence (*i.e.,* KL divergence) and summarize this additional key theoretical result in Theorem 2 in Appendix C. We then provide the detailed proofs for both Theorems. We present the detailed P-F sampling algorithm in Appendix D. Additional experimental results are shown in Appendix E. Finally, we provide the link to the source code in Appendix F.

A Summary of Notations

Table 1: Notations with Descriptions

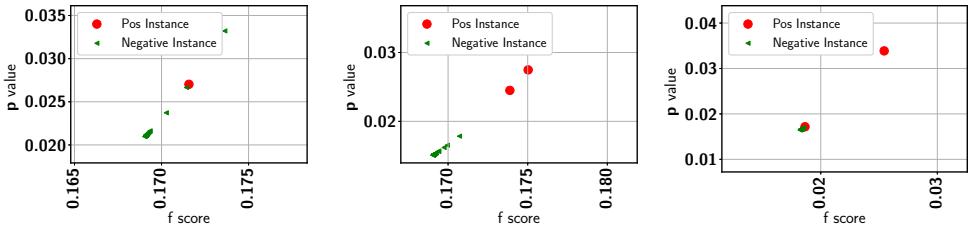
Notation	Description
$\mathbf{X}$	Set of training bag instances
Y	Binary random variable denoting bag labels
$\mathbf{x}_i$	i^{th} instance present in a bag $\mathcal{B}$
$\mathbf{x}_i^+$	i^{th} instance present in a positive bag $\mathcal{B}_{pos}$
$\mathbf{x}_j^-$	j^{th} instance present in a negative bag $\mathcal{B}_{neg}$
H	Entropy
t_i^l	Binary label of the l^{th} labeled instance
$t_{\mathcal{B}}$	Binary value indicating bag type
n	Total number of instances in a bag $\mathcal{B}$
$\mathcal{B}_{pos}$	A positive bags in the training set
$\mathcal{B}_{neg}$	A negative bags in the training set
$\mathbf{w}$	Network Parameters
Var_n	Empirical Variance
σ^2	Population Variance
$\mathcal{P}_n$	Uncertainty set
$\mathbf{p}$	n-dimensional vector weights associated with each bag instances
D_f	f-divergence indicating the distance between two distributions
f	Functional mapping of the Network parameterized by $\mathbf{w}$
N	Total number of Instances in a training set
$\mathbb{1}$	n-dimensional unit vector
λ	Radius of a ball in DRO framework
$\mathbf{z}$	n-dimensional binary indicator variable
X^+	Random variable corresponding to an instance from a positive bag

B More Examples of Challenging Bags

Figure 6 shows the p-f plots for three example challenging bags from three different topics in the 20NewsGroup dataset. As shown, the highest f-score from those bags is very low. This implies that the passive learning model predicts all the instances as negative with a high confidence. Using F-Entropy, we may not be able to query any instance from those bags because of low uncertainty. In contrast, by leveraging the standard MIL assumption, the proposed P-F sampling will effectively explore those bags. Once the positive instances from these bags are queried, they help to accurately identify similar positive instances in the same and different bags to boost the instance prediction performance, as evidenced by our experimental results.

C Extension to KL Divergence and Proofs of Theorems

In Theorem 1, we show that the optimal robust bag likelihood is equivalent to the variance regularized mean prediction score with a high probability by using the χ^2-divergence to quantify the deviation of probability vector $\mathbf{p}$ from the empirical (uniform) distribution. In this section, we extend this



(a) COMP.GRAPHICS B-8 (b) REC.MOTORCYCLES B-2 (c) TALK.POLITICS.MISC B-13

Figure 6: Example of challenging bags from different topics in 20NewsGroup

theoretical equivalence to KL-divergence and present the result in Theorem 2. We then provide the detailed proofs for both theorems.

Theorem 2. *Let X^+ be a random variable representing an instance from a positive bag, $f(X^+) \in [0,1]$ is the score assigned to an instance, $\sigma^2 = Var[f(X^+)]$ and $Var_n[f(X^+)]$ denote the population and sample variance of $f(X^+)$, respectively, and D_f takes the form of KL-divergence. We have*

$$\begin{aligned} \max_{\mathbf{p}\in\mathcal{P}_n} \sum_{i=1}^{n} p_i f(\mathbf{x}_i^+) &= \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i^+) + \sqrt{\frac{2\lambda Var_n[f(X^+)]}{n}} + \epsilon\left(\frac{\lambda}{n}\right) \\ s.t.\ \mathcal{P}_n &:= \left\{\mathbf{p}\in\mathbb{R}^n, \mathbf{p}^\top \mathbb{1} = 1, 0 \le \mathbf{p}, D_{KL}\left(\mathbf{p}||\mathbf{p}_0\right) \le \frac{\lambda}{n}\right\} \end{aligned} \tag{10}$$

where $\mathbf{p}_0 = \frac{\mathbb{1}}{n}$ is the uniform distribution indicating the center of the ball, $\epsilon\left(\frac{\lambda}{n}\right) = \frac{\lambda}{3n}\frac{\kappa_3(f(X^+))}{Var_n[f(X^+)]} + \mathcal{O}\left(\left(\frac{\lambda}{n}\right)^{3/2}\right)$ with $\kappa_3 = \mathbb{E}_0[(f(X^+) - \mathbb{E}_0[f(X^+)])^3]$ and $\mathbb{E}_0$ denotes the expectation taken over $\mathbf{p}_0$.

Remark: Given a bag with a decent size $n \gg 1$ and since λ is usually set to $\lambda \ll 1$ (0.01 is used in our experiments), we have $\epsilon\left(\frac{\lambda}{n}\right) \to 0$. When the empirical variance $\text{Var}_n[f(X^+)]$ is sufficiently large (which is true for MIL), the r.h.s. of (10) is dominated by the first two terms, which implies

$$\max_{\mathbf{p}\in\mathcal{P}_n} \sum_{i=1}^{n} p_i f(\mathbf{x}_i^+) \approx \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i^+) + \sqrt{\frac{2\lambda \text{Var}_n[f(X^+)]}{n}} \tag{11}$$

Next, we present the detailed proofs for both Theorems 1 and 2.

Proof of Theorem 1. Our proof of Theorem 1 is adapted from (Duchi & Namkoong, 2019) by making extensions that fit the unique design of the distributionally robust bag likelihood (DRBL). We start by introducing the following lemma, which will later be used in our proof.

Lemma 1 (Maurer and Pontil Theorem 10)**.** *Let Y be a random variable taking values in [0, L]. Let $\sigma^2 = Var[Y]$ and $Var_n[Y] = \frac{1}{n}\sum_{i=1}^n Y_i^2 - (\frac{1}{n}\sum_{i=1}^n Y_i)^2$ be the population and sample variance of Y, respectively. Then for for $n \ge 2$,*

$$\left(\sigma - t \le \sqrt{Var_n[Y]} \le \sigma + t\right) \ge 1 - \exp\left(-\frac{nt^2}{2L^2}\right) \tag{12}$$

The distributionally robust bag likelihood (DRBL), *i.e.,* the l.h.s. of (4), can be formulated as the following constrained optimization problem:

$$\begin{aligned} &\max_{\mathbf{p}\in\mathcal{P}_n} \sum_{i=1}^{n} p_i f(\mathbf{x}_i^+) \\ &\text{s.t. } \mathcal{P}_n := \left\{\mathbf{p}\in\mathbb{R}^n, \mathbf{p}^\top\mathbb{1} = 1, 0 \le \mathbf{p}, D_f\left(\mathbf{p}||\frac{\mathbb{1}}{n}\right) \le \frac{\lambda}{n}\right\} \end{aligned} \tag{13}$$

Since the $D_f(\mathbf{p}||\mathbf{q})$ is assumed to be the χ^2-divergence and $\mathbf{q}$ follows the uniform distribution, $D_f(\mathbf{p}||\mathbf{q})$ is reduced to the squared Euclidean distance. We first introduce the mean of

$f(\mathbf{x}_i^+)$'s, which is denoted as $\bar{f} = \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i^+)$. Also, recall we denote the score vector by $\mathbf{f} = (f(\mathbf{x}_1^+), ..., f(\mathbf{x}_n^+))^\top$ in Section 3.2. Thus, the empirical variance of $f(X^+)$ is given by $\text{Var}_n[f(X^+)] = \frac{1}{n}||\mathbf{f}||_2^2 - \bar{f}^2 = \frac{1}{n}||\mathbf{f} - \bar{f}\mathbb{1}||_2^2$. We further introduce $\mathbf{u} = \mathbf{p} - \frac{\mathbb{1}}{n}$, so the objective in (13) can be transformed as

$$\mathbf{p}^\top \mathbf{f} = (\mathbf{u} + \frac{\mathbb{1}}{n})^\top \mathbf{f} = \bar{f} + \mathbf{u}^\top \mathbf{f} = \bar{f} + \mathbf{u}^\top (\mathbf{f} - \bar{f}\mathbb{1}) \tag{14}$$

where the last equality holds because $\mathbf{u}^\top \mathbb{1} = 0$. Thus, the optimization problem in (13) can be further transformed into

$$\max_{\mathbf{u}\in\mathbb{R}^n} \bar{f} + \mathbf{u}^\top (\mathbf{f} - \bar{f}\mathbb{1}) \quad \text{s.t.} \quad ||\mathbf{u}||_2^2 \le \frac{\lambda}{n^2}, \mathbf{u}^\top \mathbb{1} = 0, \mathbf{u} \ge -\frac{1}{n} \tag{15}$$

where the first constraint is derived by replacing D_f with the χ^2-divergence. Now, using the Cauchy-Schwarz inequality, which states that $\mathbf{u}^\top \mathbf{v} \le ||\mathbf{u}||_2 ||\mathbf{v}||_2$, gives the following condition

$$\mathbf{u}^\top (\mathbf{f} - \bar{f}\mathbb{1}) \le \frac{\sqrt{\lambda}}{n} ||\mathbf{f} - \bar{f}\mathbb{1}||_2 = \sqrt{\frac{\lambda \text{Var}_n[f(X^+)]}{n}} \tag{16}$$

where the equality holds if and only if

$$u_i = \frac{\sqrt{\lambda}(f(\mathbf{x}_i^+) - \bar{f})}{n||\mathbf{f} - \bar{f}\mathbb{1}||_2} = \frac{\sqrt{\lambda}(f(\mathbf{x}_i^+) - \bar{f})}{n\sqrt{n\text{Var}_n[f(X^+)]}} \tag{17}$$

Since we also have a constraint $\mathbf{u} \ge -\frac{1}{n}$, which satisfies if and only if

$$\min_{i\in[n]} \frac{\sqrt{\lambda}(f(\mathbf{x}_i^+) - \bar{f})}{\sqrt{n\text{Var}_n[f(X^+)]}} \ge -1 \tag{18}$$

Thus, if inequality (18) holds for vector $\mathbf{f}$, we have

$$\max_{\mathbf{p}\in\mathcal{P}_n} \mathbf{p}^\top \mathbf{f} = \bar{f} + \sqrt{\frac{\lambda \text{Var}_n[f(X^+)]}{n}} \tag{19}$$

which will prove the Theorem given in (4).

What remains is to prove inequality (18) holds with a high probability. To show this, we leverage the concentration inequality given by Lemma 1. Since $f(\mathbf{x}_i^+) \in [0, 1]$, we have $|f(\mathbf{x}_i^+) - \bar{f}| \le 1$. To satisfy inequality (18), it is sufficient to have

$$\frac{\lambda}{n\text{Var}_n[f(X^+)]} \le 1 \quad \text{or} \quad n \ge \frac{\lambda}{\text{Var}_n[f(X^+)]} \quad \text{or} \quad \text{Var}_n[f(X^+)] \ge \frac{\lambda}{n} \tag{20}$$

Let us define the following event

$$\epsilon_n := \{\text{Var}_n[f(X^+)] \ge \frac{1}{43}\sigma^2\} \tag{21}$$

In Theorem 1, we suppose $n \ge \frac{4\lambda}{\sigma^2}\max\{2\sigma, 11\})$. Then, on event ϵ_n, we have $n \ge \frac{44\lambda}{\sigma^2} \ge \frac{44\lambda}{43\text{Var}_n[f(X^+)]} \ge \frac{\lambda}{\text{Var}_n[f(X^+)]}$, so that the sufficient condition (20) holds and the (19) becomes true.

Now we find the probability of holding the above event in (21) using Lemma 1. First, $L = 1$ in our case, which gives

$$P(\sigma - t \le \sqrt{\text{Var}_n[f(X^+)]} \le \sigma + t) \ge 1 - \exp\left(-\frac{nt^2}{2}\right)$$

The following also holds true:

$$P\left(\sigma - t \le \sqrt{\text{Var}_n[f(X^+)]}\right) \ge P(\sigma - t \le \sqrt{\text{Var}_n[f(X^+)]} \le \sigma + t) \ge 1 - \exp\left(-\frac{nt^2}{2}\right)$$

Let $t = \left(1 - \sqrt{\frac{1}{43}}\right)\sigma$, which gives $\sigma - t = \sqrt{\frac{1}{43}}\sigma$. Substituting this to (12) leads to

$$P\left(\sqrt{\frac{1}{43}}\sigma \le \sqrt{\text{Var}_n[f(X^+)]}\right) \ge 1 - \exp\left(-\frac{nt^2}{2}\right); P(\epsilon_n) \ge 1 - \exp\left(-\frac{nt^2}{2}\right)$$

Further substituting the value of $t = \left(1 - \sqrt{\frac{1}{43}}\right)\sigma$ gives rise to

$$P(\epsilon_n) \geq 1 - \exp\left(-0.359n\sigma^2\right) \geq 1 - \exp\left(-\frac{7n\sigma^2}{20}\right)$$

This completes the proof of Theorem 1.

Proof of Theorem 2. In order to prove this theorem, we consider two assumptions, which both hold true for our MIL setting.

Assumption 1: Random variable $f(X^+)$ has a finite exponential moment in a neighborhood of 0 under the distribution $\mathbf{p}_0$ *i.e.,*, $\mathbb{E}_0[\exp(\tau f(X^+))] < \infty$ for $\tau \in [-\tau_0, \tau_0]$ for some $\tau_0 > 0$.

Assumption 2: Random variable $f(X^+)$ is non-constant under $\mathbf{p}_0$.

Assumption 1 is true in our case as $f(X^+)$ is bounded in [0, 1]; Assumption 2 also empirically holds true as there are both positive and negative instances in a positive bag so the output scores are distinct over different instances in a bag. The second assumption ensures that the uniform distribution $\mathbf{p}_0$ is not a locally optimum, which means there exists an opportunity to upgrade the value by re-balancing the probability between positive and negative instances in a positive bag.

Consider $\mathbf{p}$ that is absolutely continuous with respect to $\mathbf{p}_0$ and therefore the likelihood ratio $g = \frac{d\mathbf{p}}{d\mathbf{p}_0}$ (a.k.a., Radon–Nikodym derivative) exists. Using a change of measure, the optimization problem in the l.h.s. of (10) can be written as

$$\begin{aligned} &\max_{g \in \mathcal{L}_1(\mathbf{p}_0)} \mathbb{E}_0[gf(X^+)] \\ &\text{s.t.} \left\{ \mathbb{E}_0[g \log g] \leq \frac{\lambda}{n}, \mathbb{E}_0[g] = 1, g \geq 0 \right\} \end{aligned} \tag{22}$$

where $\mathcal{L}_1(\mathbf{p}_0)$ is $\mathcal{L}_1$-space with respect to the measure $\mathbf{p}_0$. To solve the optimization problem above, we formulate its Lagrangian,

$$\max_{g \in \mathcal{L}_1(\mathbf{p}_0)} \mathbb{E}_0[gf(X^+)] - \alpha \left(\mathbb{E}_0[g \log g] - \frac{\lambda}{n} \right) \tag{23}$$

where α is the Lagrange's multiplier. The solution of the above objective function is given by the following proposition (Petersen et al., 2000; Lam, 2016):

Proposition 1. *Under Assumption 1, when $\alpha > 0$ is sufficiently large, there exists an unique optimizer of* (23) *given by*

$$g^*(\mathbf{x}^+) = \frac{\exp(\frac{f(\mathbf{x}^+)}{\alpha})}{\mathbb{E}_0\left[\exp\frac{f(X^+)}{\alpha}\right]} \tag{24}$$

Assume that such α^* and g^* exist and that α^* is sufficiently large then

$$\frac{\lambda}{n} = \mathbb{E}_0[g^* \log g^*] = \frac{\mathbb{E}_0[g^* f(X^+)]}{\alpha} - \log \mathbb{E}_0\left[\exp\left(\frac{f(X^+)}{\alpha^*}\right)\right]$$

$$= \frac{\beta^* \mathbb{E}_0[f(X^+)\exp(\beta^* f(X^+))]}{\mathbb{E}_0[\exp(\beta^* f(X^+))]} - \log \mathbb{E}_0[\exp \beta^* f(X^+)] = \beta^* \psi^{'}(\beta^*) - \psi(\beta^*)$$

In the above expression, we define $\beta^* = \frac{1}{\alpha^*}$ and $\psi(\beta) = \log \mathbb{E}_0[\exp(\beta f(X^+))]$ is the logarithmic moment generating function of $f(X^+)$.

We can write the optimal solution of the objective function (22) as follows

$$\mathbb{E}_0[f(X^+)g^*] = \frac{\mathbb{E}_0[f(X^+)\exp(\frac{f(X^+)}{\alpha^*})]}{\mathbb{E}_0[\exp(\frac{f(X^+)}{\alpha^*})]} = \psi^{'}(\beta^*) \tag{25}$$

Now let us perform Taylor expansion of the following

$$\beta\psi^{'}(\beta) - \psi(\beta) = \sum_{m=0}^{\infty} \frac{1}{m!}\kappa_{m+1}\beta^{m+1} - \sum_{m=0}^{\infty} \frac{1}{m!}\kappa_m\beta^m = \sum_{m=1}^{\infty} \left[\frac{1}{(m-1)!} - \frac{1}{m!}\right]\kappa_m\beta^m$$

$$= \sum_{m=2}^{\infty} \frac{1}{m(m-2)!}\kappa_m\beta^m = \frac{1}{2}\kappa_2\beta^2 + \frac{1}{3}\kappa_3\beta^3 + \frac{1}{8}\kappa_4\beta^4 + \mathcal{O}(\beta^5)$$

In the above expression, $\kappa_m = \psi^{(m)}(0)$ is the m-th derivative of ψ with evaluated at $\beta = 0$ and $\mathcal{O}(\beta^5)$ is continuous in β. By Assumption 2, we have $\kappa_2 > 0$. Therefore, for small enough $\frac{\lambda}{n}$, above equation reveals that there is a small $\beta^* > 0$ that is root to the equation $\frac{\lambda}{n} = \beta\psi^{'}(\beta) - \psi(\beta)$ and the root is unique. This is because by Assumption 2, $\psi(.)$ is strictly convex, and therefore, $(\frac{d}{d\beta})(\beta\psi^{'} - \psi(\beta)) = \beta\psi^{''}(\beta) > 0$ for $\beta > 0$, so that $\beta\psi^{'}(\beta) - \psi(\beta)$ is strictly increasing.

Since $\alpha^* = \frac{1}{\beta^*}$, this shows that for any sufficiently small $\frac{\lambda}{n}$, we can find a large $\alpha^* > 0$ such that the corresponding g^* in 24 satisfies $\frac{\lambda}{n} = \mathbb{E}_0[g^* \log g^*]$. This means we can write the following

$$\frac{\lambda}{n} = \frac{1}{2}\kappa_2{\beta^*}^2 + \frac{1}{3}\kappa_3{\beta^*}^3 + \frac{1}{8}\kappa_4{\beta^*}^4 + \mathcal{O}({\beta^*}^5) \tag{26}$$

We can obtain β^* as follow

$$\beta^* = \sqrt{\frac{2\lambda}{n\kappa_2}}\left(1 + \frac{2}{3}\frac{\kappa_3}{\kappa_2}\beta^* + \frac{1}{4}\frac{\kappa_4}{\kappa_2}{\beta^*}^2 + \mathcal{O}({\beta^*}^3)\right)^{-\frac{1}{2}} = \sqrt{\frac{2\lambda}{n\kappa_2}}\left(1 - \frac{1}{3}\frac{\kappa_3}{\kappa_2}\beta^* + \mathcal{O}({\beta^*}^2)\right)$$

$$= \sqrt{\frac{2}{\kappa_2}}\left(\frac{\lambda}{n}\right)^{1/2} - \frac{2}{3}\frac{\kappa_3}{\kappa_2^2}\frac{\lambda}{n} + \mathcal{O}\left(\left(\frac{\lambda}{n}\right)^{\frac{3}{2}}\right)$$

In the above expression, first we use the binomial expansion $(1+x)^{\frac{-1}{2}} = 1 - \frac{1}{2}x + \frac{3}{8}x^2$.... followed by substitution of β^* in the second term. Now, the corresponding optimal solution becomes following

$$\mathbb{E}_0[f(X^+)g^*] = \psi^{'}(\beta^*) = \kappa_1 + \kappa_2\beta^* + \kappa_3\frac{{\beta^*}^2}{2} + \mathcal{O}({\beta^*}^3) = \kappa_1 + \sqrt{2\kappa_2}\left(\frac{\lambda}{n}\right)^{\frac{1}{2}} + \frac{1}{3}\frac{\kappa_3}{\kappa_2}\frac{\lambda}{n} + \mathcal{O}\left(\left(\frac{\lambda}{n}\right)^{\frac{3}{2}}\right)$$

In the above equation $\kappa_1 = \bar{f}, \kappa_2 = \text{Var}_n[f(X^+)], \kappa_3 = \mathbb{E}_0[(f(X^+) - \mathbb{E}_0[f(X^+)])^3]$. This completes the proof of Theorem 2.

D THE P-F SAMPLING ALGORITHM

Algorithm 1 shows the detailed description of the proposed P-F active sampling technique. First, we find the unexplored bags from a pool of positive training bags. To determine the unexplored bags, we identify the instance with highest p-value from each bag (*i.e.,* b_* for bag b according to (7)). Next, we sort the f-scores of corresponding instances in an non-decreasing order. Based on the sorted scores, we pick bags whose highest f-scores are less than a given threshold Th_{PF} as defined in Algorithm 1. From each unexplored bag, we pick k-instances with highest scores. During this process, we avoid bags, where at least one true positive instance is already queried in previous steps. The maximum number of instances queried from unexplored bags is bounded by the batch size in each step. During query, we give higher priority to the least explored bag, whose highest instance score is the smallest.

If the batch size is not reached, we continue to query instances whose uncertainty are the highest based on their F-Entropy. To ensure that the queried instances are indeed uncertain, we compare the corresponding F-Entropy with a threshold Th_H to avoid querying confident instances.

Algorithm 1: P-F Active Sampling

Input: $\mathbf{p}_{\mathcal{B}_{\mathbf{pos}}}$, $\mathcal{Q}_{prev}$, Th_{PF}, Th_H, BSize, k,
Output: $\mathcal{Q}$
Data: B positive training bags `// Consists of a feature vector for each bag`

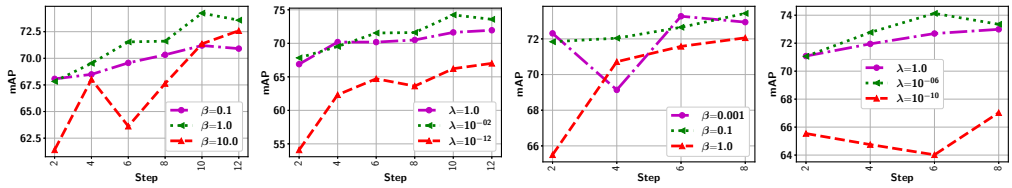
Initialization: $\mathcal{U}_B = \{\}$, count = 0, $\mathcal{Q}_{P-F} = \{\}$, $\mathcal{Q}_F = \{\}$
for $b \in [B]$ **do**
  $\mathbf{p}_b \leftarrow \mathbf{p}_{\mathcal{B}_{pos}}[b]$
  $b_* \leftarrow \arg\max \mathbf{p}_b \setminus \mathcal{Q}_{prev}[b]$
  if $f(\mathbf{x}_{b_*}^+) \leq Th_{PF}$ **then**
    $\mathcal{U}_B \leftarrow b_*$

`/* Adding instances from unexplored bags */`
$\mathcal{U}_B = \arg\text{sortAsc}_{b_* \in \mathcal{U}_B} f(\mathbf{x}_{b_*}^+)$
for $b_* \in \mathcal{U}_B$ **do**
  if $b_* \in \mathcal{Q}_{prev}$ **then**
    if *positive ins* $\in \mathcal{Q}_{prev}[b]$ **then**
      continue
  else
    $\mathcal{X}^{PF} = \arg\text{sortDesc}_{b_*} \left(f(\mathbf{x}_{b_*}^+) \setminus \mathcal{Q}_{prev}[b_*]\right)[:k]$
    for $\mathbf{x}_i \in \mathcal{X}^{PF}$ **do**
      if *count*$\geq$ *BSize* **then**
        break
      $\mathcal{Q}_{P-F}[b_*] \leftarrow \mathbf{x}_i$
      count $\leftarrow$ count+1

$\mathcal{Q}_{prev} = \mathcal{Q}_{prev} \cup \mathcal{Q}_{P-F}$
`/* Adding instances with highest F-Entropy;`
$H[f(\mathbf{x}_i^+)] = -\left[f(\mathbf{x}_i^+ \log f(\mathbf{x}_i^+)) + (1 - f(\mathbf{x}_i^+)) \log(1 - f(\mathbf{x}_i^+))\right]$ `*/`
$\mathcal{C}_{idx} = \arg\text{sortDesc}_i \left(H[f(\mathbf{x}_i^+)] \geq Th_H\right)$
for $i \in \mathcal{C}_{idx}$ **do**
  if *count*$\geq$ *BSize* **then**
    break
  if $\mathbf{x}_i^+ \in \mathcal{Q}_{prev}[b_i]$ **then**
    break
  $\mathcal{Q}_F[b_i] \leftarrow \mathbf{x}_i^+$
  count $\leftarrow$ count+1
$\mathcal{Q} = \mathcal{Q}_{prev} \cup Q_F$

Table 2: Bag level distributions on different datasets

Split	20NewsGroup		Cifar10		Cifar100		Pascal VOC	
	Positive	*Negative*	*Positive*	*Negative*	*Positive*	*Negative*	*Positive*	*Negative*
Train	30	30	500	500	500	500	124	124
Test	20	20	100	100	100	100	84	84



(a) $\lambda = 0.01$ (b) $\beta = 1$ (c) $\lambda = 1$ (d) $\beta = 0.1$

Figure 7: Impact of key model parameters: (a-b) Cifar10; (c-d) 20NewsGroup

E Additional Experimental Results

In this section, we first give a detailed description of the datasets. We then present additional ablation study results that complement the ones presented in the main paper. Finally, we demonstrate some qualitative examples where our approach is able to sample the true positive instance from the bag containing outlier but Maximum-Entropy can not.

E.1 Dataset Description

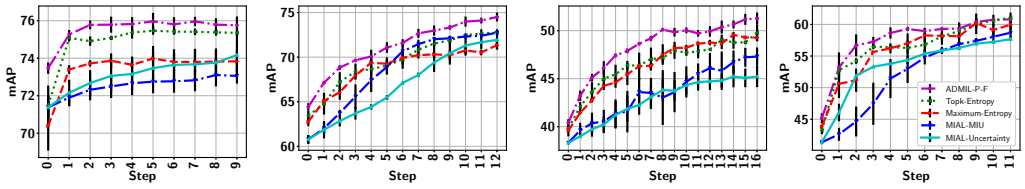
Table 2 summarizes the bag level statistics. We also present the details of each dataset below.

- **20NewsGroup:** In this dataset, an instance refers to a post from a particular topic. For each topic, a bag is considered as positive if it contains at least one instance from that topic and negative otherwise. This dataset is particularly challenging because of the severe imbalance in terms of instances where there are very few ($\approx 3\%$) positive instances in each positive bag. While number of instances per bag may vary, on average there are around 40 instances per bag.
- **Cifar10:** In the original dataset, there are 50,000 training and 10,000 testing images with 10 classes indicating different images. The bags are constructed as follows. First, we pick '1', '2', and '5' related images as positive instances and the rest as negative. To construct a positive bag, we choose a random number from 1 to 3 and pick the positive instances equal to the randomly generated number. The rest of the instances are selected from a negative instances pool. For negative bags, all instances are selected from the negative instance pool. For each bag, we consider 32 instances.
- **Cifar100:** In the original dataset, there are 50,000 training and 10,000 testing images with 20 different superclasses indicating different species. Bag construction is similar to Cifar10. In this case, images from superclass 2 are treated as positive and the rest as negative.
- **Pascal VOC:** This dataset consists of 2,913 images, where images are used for segmentation. In particular, each image is treated as a bag and instances are obtained as follows. We define a grid size of 60×75 and partition the images. Depending on an image size, the number of instances may very. We treat an instance as positive if at least 5% of the total pixels in a given instance are related to the object of interest otherwise negative. In our case, we consider bird as the object of interest. All the images consisting of bird are regarded as positive bags and other as negative.

E.2 Additional Ablation Study

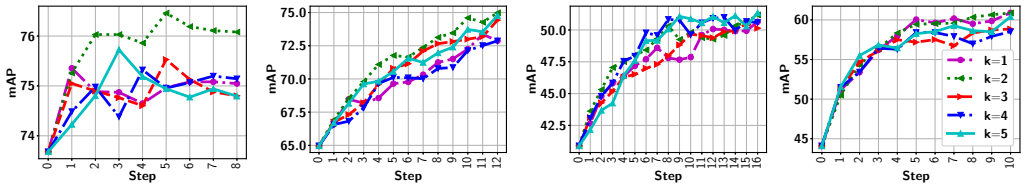
In this section, we present some additional ablation study results to demonstrate the impact of key model parameters and the stability of the model performance over multiple runs.

Impact of β and λ: Since we have already shown the results on Cifar100 and Pascal VOC, Figure 7 show the impact of λ and β on Cifar10 and 20NewsGroup datasets. Similar to the findings on the other two datasets, Figures 7 (b) and (d) demonstrate that a λ in the middle range outperforms too large (or small) λ values. In case of β, placing too less (or too much) emphasis may result in overly



(a) 20NewsGroup (b) Cifar10 (c) Cifar100 (d) Pascal VOC

Figure 8: MI-AL performance with standard deviation



(a) 20NewsGroup (b) Cifar10 (c) Cifar100 (d) Pascal VOC

Figure 9: Impact of hyperparameter k

(or poorly) leveraging annotated instances. Therefore, a good balance between the bag-level and instance-level losses achieves the best result, as shown in Figures 7 (a) and (c).

Performance stability. Figure 8 reports the performance comparison with one standard deviation (computed over three runs), which is represented by the vertical black line. As discussed in the main paper, the mean MI-AL curve of ADMIL clearly outperforms other competitive baselines. Meanwhile, the standard deviation of the proposed ADMIL model is also relatively small, which indicates its overall stable MI-AL performance across different datasets in multiple runs.

Impact of k: Figure 9 shows the impact of the hyperparameter k, which is the number of instances queried in each unexplored bag, on model performance. As can be seen, $k = 2$ achieves a generally decent performance across all the datasets. For datasets with a larger size (*e.g.,* Cifar100), a larger k leads to a slightly better performance.

F Link to Source Code

For the source code of our experiments, please click here.