Balancing Bias and Variance for Active Weakly Supervised Learning

Hitesh Sapkota
hxs1943@rit.edu
Rochester Institute of Technology
Rochester, NY, USA

Qi Yu
qi.yu@rit.edu
Rochester Institute of Technology
Rochester, NY, USA

ABSTRACT

As a widely used weakly supervised learning scheme, modern 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 to conduct novel active deep multiple instance learning 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 model.

CCS CONCEPTS

• **Computing methodologies** → **Active learning settings**; **Learning settings**; **Machine learning**.

KEYWORDS

multiple instance learning; active learning; weak supervision

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539264

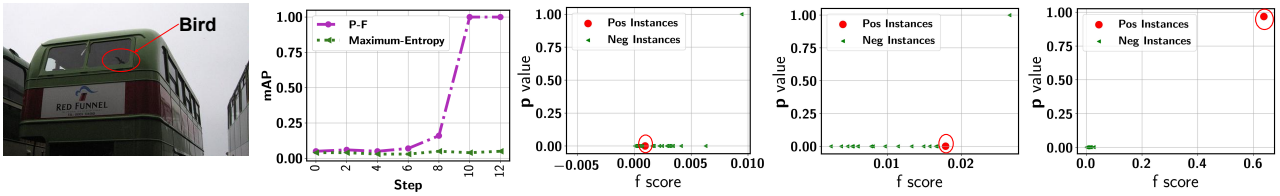
ACM Reference Format:
Hitesh Sapkota and Qi Yu . 2022. Balancing Bias and Variance for Active Weakly Supervised Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539264

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 annotation cost [6, 19, 25, 27]. 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 [27] and medical image segmentation [12]) remains largely unsatisfactory.

In MIL, a bag is considered to be positive if at least one of the instances is positive otherwise negative [6, 10]. To achieve a high bag level prediction, most existing MIL models primarily focus on the most positive instance from a positive bag that is mainly responsible for determining the bag label [1, 10, 14, 27]. 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 [3]. 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 a discrete value.

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.,* bag label) cannot be propagated to the instance level without sufficient statistical evidence. One promising direction to tackle this 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



(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 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 identify 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 [27]. 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 DRBL. 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) show 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) an 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 [1], boosting [29], graph-based models [31], attention based [11, 12], conditional random field [5] and Gaussian Processes [10, 14]. Other approaches try to relax the MIL assumption, which allows positive instances in a negative bag to handle noisy bags [19]. As MIL is commonly applied to video anomaly detection and image segmentation that involve high dimensional data, deep neural networks (DNNs) have become a popular choice for training MIL models [11, 12, 27]. 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 [10, 12, 27]. 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 [23]. Distributionally robust optimization has also been adopted in multi-class AL to address sampling bias and imbalanced data distribution [32]. 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 [9, 13, 18, 28]. 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 [2, 15, 22]. 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 [24]. 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. A reinforcement learning based AL technique is developed in [4], where segments are chosen to be labeled in each AL step . However, segmentation level annotations are required to compute the reward during the training process, which violates the assumption of MIL. Another AL framework is developed for MIL tasks in [30]. 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 METHODOLOGY

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 feature vector. Let $t_{\mathcal{B}} \in \{+1, -1\}$ indicates the bag type. 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

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 [27]. We define as

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

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 will omit $\mathbf{w}$ from $f(\mathbf{x}; \mathbf{w})$ 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 its top-k extensions suffer 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 [7], 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:

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

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 parameter C balances 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 DNN) 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 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 [7], 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 [7, 20], 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 DRBL, 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 $n \geq \max(2, \frac{\lambda}{\sigma^2}\max(8\sigma, 44))$,*

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

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. Theorem 2 below further generalizes this result to the KL-divergence.

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*

$$\max_{\mathbf{p}\in\mathcal{P}_n} \sum_{i=1}^n p_i f(\mathbf{x}_i^+) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i^+) + \sqrt{\frac{2\lambda Var_n[f(X^+)]}{n}} + \epsilon\left(\frac{\lambda}{n}\right) \tag{5}$$

where $\epsilon\left(\frac{\lambda}{n}\right) = \frac{\lambda}{3n}\frac{\kappa_3(f(X^+))}{Var_n[f(X^+)]} + O\left(\left(\frac{\lambda}{n}\right)^{3/2}\right)$ 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 $\mathrm{Var}_n[f(X^+)]$ is sufficiently large (which is true for MIL), the r.h.s. of (5) is dominated by the first two terms, which implies

$$\max_{\mathbf{p}\in\mathcal{P}_n} \sum_{i=1}^n p_i f(\mathbf{x}_i^+) \approx \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i^+) + \sqrt{\frac{2\lambda \mathrm{Var}_n[f(X^+)]}{n}} \tag{6}$$

Detailed proofs are given in Appendix A. Leveraging the above theoretical results, we formulate a DRBL-based MIL loss as

$$\mathcal{L}^{\mathrm{DRBL}} = \left\{ 1 - \max_{\mathbf{p}\in\mathcal{P}_n} \left[\sum_{i=1}^n p_i f(\mathbf{x}_i^+)\right] + \max_{j\in\mathcal{B}_{neg}} \left[f(\mathbf{x}_j^-)\right] \right\}_+ \tag{7}$$

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. A straightforward way to perform uncertainty based instance sampling is to compute the f-score based entropy of the instances, referred to F-Entropy:

$$\mathbf{x}_* = \arg\max_{i\in\mathcal{B}_{pos}} H[f(\mathbf{x}_i^+)], \tag{8}$$

where $H[f] = -[f \log f + (1 - f)\log(1 - f)]$. 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 (8) 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 8 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:

$$\mathbf{x}_*^{PF} = \arg \min_{b \in \{1,\dots,B\}} f(\mathbf{x}_{b_*}^+), \quad \text{and } b_* = \arg\max \mathbf{p}_b \tag{9}$$

where $\mathbf{p}_b$ is the probability vector of bag b. For each bag, the sampling function first identifies the instance $\mathbf{x}_{b_*}^+$ 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 $\mathbf{x}_{b_*}^+$, 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 is 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 (9) 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 a 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.

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 $\mathbf{X}^l = \{\mathbf{x}_1^l, \mathbf{x}_2^l, ..., \mathbf{x}_m^l\}$ be the m labeled instances queried by the proposed active learning function and $\mathbf{t}^l = \{t_1^l, t_2^l, ..., t_m^l\}$ with $t_l^i \in \{0, 1\}$ be the corresponding instance labels. We formulate a supervised binary cross-entropy (BCE) loss as

$$L^{\text{BCE}} = -\frac{1}{m} \sum_{i=1}^{m} \left[t_i^l \log(f(\mathbf{x}_i^l)) + (1 - t_i^l) \log(1 - f(\mathbf{x}_i^l)) \right] \tag{10}$$

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:

$$\mathcal{L}^{\text{Hybrid}} = \mathcal{L}^{DRBL}(\mathcal{B}_{pos}, \mathcal{B}_{neg}) + \beta \mathcal{L}^{BCE}(\mathbf{X}^l, \mathbf{t}^l) \tag{11}$$

where β is used to trade-off bag- 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 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.

4.1 Experimental Setup

Datasets. Our experiments involve four datasets covering both textual and image data: 20NewGroup [31], Cifar10 [16], Cifar100 [16], and Pascal VOC [8]. The detailed description of each dataset is given below and bag level statistics is summarized in the Table 1

- **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 where there are very few (≈ 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 'automobile', 'bird', and 'dog' 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:** The dataset consists of 50,000 training and 10,000 testing images with 20 different superclasses indicating different species. Bag construction is similar to Cifar10, where images in superclass flowers are treated as positive and the rest as negative.
- **Pascal VOC:** This dataset consists of 2,913 images, where images are used for segmentation. 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 the image size, the number of instances may vary. 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 considerthe bird as the object of interest. All the images consisting of bird are regarded as positive bags and others as negative.

Algorithm 1: P-F Active Sampling

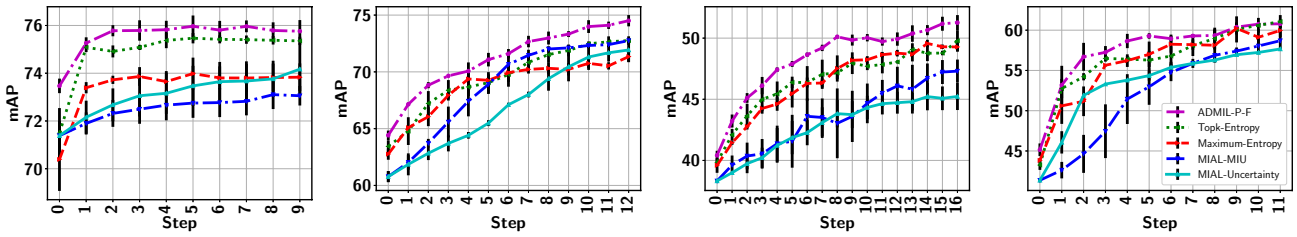
Input: $\mathbf{p}_{\mathcal{B}_{pos}}$, $\mathcal{Q}_{prev}$, Th_{PF}, Th_H, $BSize$, k
Output: $\mathcal{Q}$
Data: B positive training bags `// Feature vector for each bag`

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 $f(\mathbf{x}^+_{b_*}) \le Th_{PF}$ **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*$\ge$ *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;`
$H[f(\mathbf{x}^+_i)] =$
$-\left[f(\mathbf{x}^+_i \log f(\mathbf{x}^+_i)) + (1 - f(\mathbf{x}^+_i)) \log(1 - f(\mathbf{x}^+_i))\right]$ `*/`
$C_{idx} = \arg\text{sortDesc}_i\left(H[f(\mathbf{x}^+_i)] \ge Th_H\right)$
for $i \in C_{idx}$ **do**
  if *count*$\ge$ *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
$\mathcal{Q} = \mathcal{Q}_{prev} \cup Q_F$

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) [26].

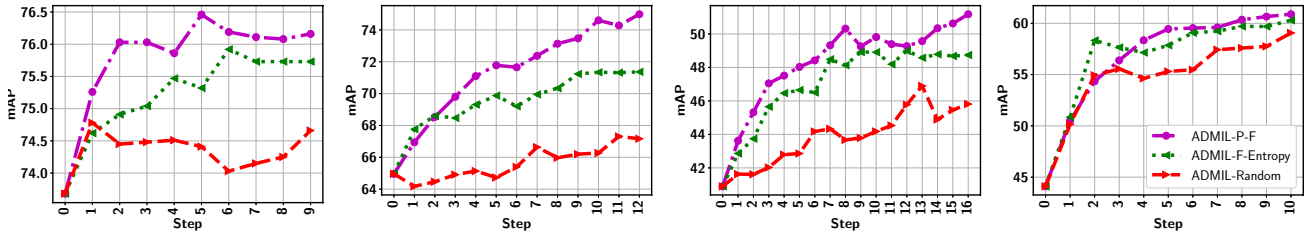
Table 1: Number of positive and negative bags on different datasets

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

Figure 2: MI-AL performance



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

Figure 3: Effectiveness of P-F active sampling

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. Learning rate 0.01 is used for all dataset except for 20NewsGroup which is 0.1.

4.2 Performance Comparison

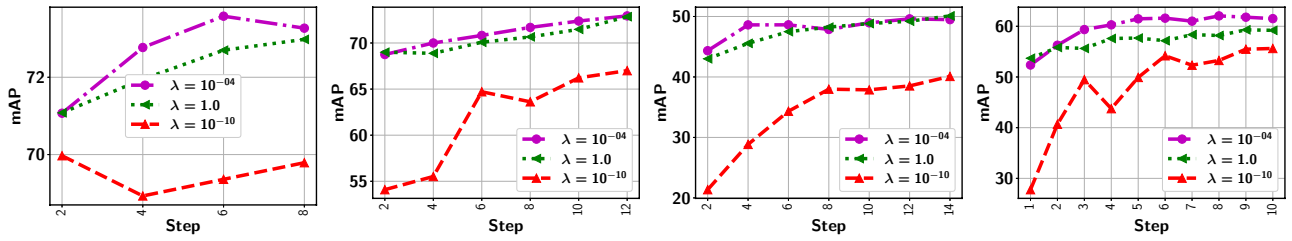
To demonstrate the instance prediction performance achieved by the proposed ADMIL model, we compare it with competitive baselines. First, the two MI-AL sampling strategies: MIAL-Uncertainty and MIAL-MIU [24], from the MI logistic model are included. Since our datasets involve high-dimensional data, we 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 [24] 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

Table 2: MIL Performance in Passive Setting

Approach	20NewsGroup	Cifar10	Cifar100	Pascal VOC
Ilse et al. [12]	60.85	65.16	40.15	40.15
Hsu et al. [11]	42.08	63.84	41.57	34.83
ADMIL	73.47(**75.42**)	64.41(**74.50**)	40.41(**51.26**)	45.15(**60.79**)

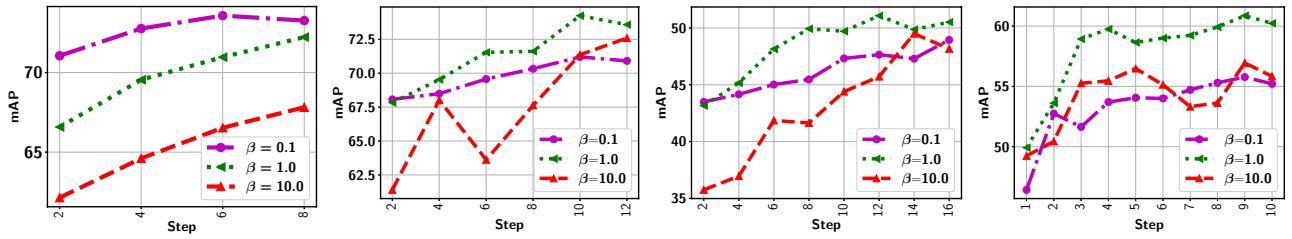
MI-AL curves with one standard deviation (computed over three runs) represented by vertical black line for all four datasets. ADMIL achieves the best performance in all cases. For most datasets, it shows a much better initial performance, which results from the proposed DRBL-based MIL loss that significantly 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. For reference, we also compare ADMIL with two recently developed MIL models, including Ilse et al. [12] and Hsu et al. [11], under the passive setting. As shown in Table 2, ADMIL achieves better or at least comparable performance as compared with these competitive baselines. This clearly justifies of using ADMIL as a base model for active sampling. After labeling a small set of actively sampled instances, the performance is significantly boosted (as shown in the parenthesis), which further justifies the benefits of combining AL with MIL. 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



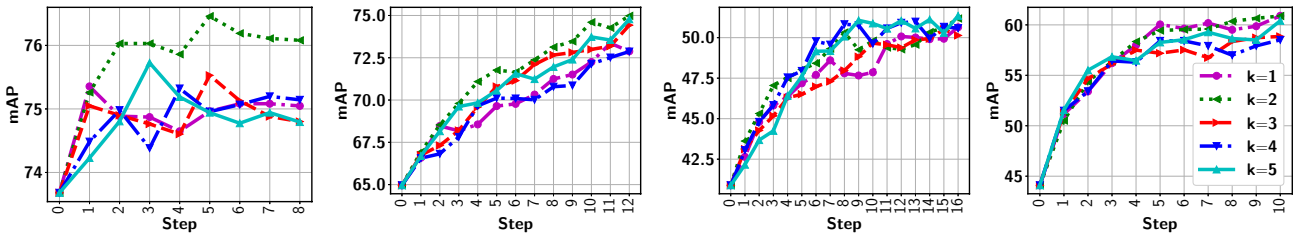
(a) 20NewsGroup (b) Cifar10 (c) Cifar100 (d) Pascal VOC

Figure 4: Impact of model parameter λ



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

Figure 5: Impact of model parameter β



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

Figure 6: Impact of hyperparameter k

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's advantage over F-Entropy is smaller on Pascal VOC due to 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.

4.3 Ablation Study

Impact of λ and β: Figures 4 and 5 demonstrate the impact of λ (with $\beta = 1$) and β (with $\lambda = 0.01$) to the model performance. 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, relatively higher λ value gives very good performance in general. Figure 4 shows that $\lambda = 0.0001$ 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 results in an optimal performance, shown in Figures 5.

Impact of k: Figure 6 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.

4.4 Qualitative analysis

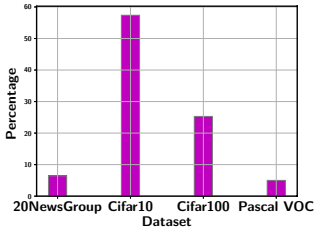
To further justify why the proposed ADMIL model and its P-F sampling function work better than other baselines, we provide a few illustrative examples to offer deeper insights on its good performance. First, we show two challenging bags in addition to the one shown in Figure 1 (a). As shown in Figure 7 (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 (11), the positive instances from those bags are successfully queried. Figure 7 (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

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

Figure 7 (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 CONCLUSION

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.

ACKNOWLEDGEMENT

This research was supported in part by an NSF IIS award IIS-1814450 and an ONR award N00014-18-1-2875. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

REFERENCES

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. 2002. Support Vector Machines for Multiple-Instance Learning. In *NIPS*.
[2] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv* (2019).
[3] Marc-André Carbonneau, V. Cheplygina, Eric Granger, and Ghyslain Gagnon. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit.* 77 (2018), 329–353.
[4] Arantxa Casanova, Pedro H. O. Pinheiro, Negar Rostamzadeh, and Christopher Joseph Pal. 2020. Reinforced active learning for image segmentation. *ArXiv* abs/2002.06583 (2020).
[5] Thomas Deselaers and Vittorio Ferrari. 2010. A Conditional Random Field for Multiple-Instance Learning. In *ICML*. 287–294.
[6] Thomas G. Dietterich, R. Lathrop, and Tomas Lozano-Perez. 1997. Solving the Multiple Instance Problem with Axis-Parallel Rectangles. *Artif. Intell.* 89 (1997), 31–71.
[7] John Duchi and Hongseok Namkoong. 2019. Variance-based Regularization with Convex Objectives. *Journal of Machine Learning Research* 20, 68 (2019), 1–55.
[8] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *IJCV* 111, 1 (2015), 98–136.
[9] Yarin Gal and Zoubin Ghahramani. 2015. Bayesian convolutional neural networks with Bernoulli approximate variational inference. *arXiv* (2015).
[10] Manuel Haußmann, Fred A. Hamprecht, and Melih Kandemir. 2017. Variational Bayesian Multiple Instance Learning with Gaussian Processes. In *CVPR*. 810–819.
[11] Yen-Chi Hsu, Cheng-Yao Hong, Ming-Sui Lee, and Tyng-Luh Liu. 2020. Query-Driven Multi-Instance Learning. In *AAAI*.
[12] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. 2018. Attention-based Deep Multiple Instance Learning. In *ICML*.
[13] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. 2015. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv* (2015).
[14] Minyoung Kim and F. Torre. 2010. Gaussian Processes Multiple Instance Learning. In *ICML*.
[15] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *arXiv* (2019).
[16] A. Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images. *University of Toronto* (2009).
[17] Henry Lam. 2016. Robust Sensitivity Analysis for Stochastic Systems. *Math. Oper. Res.* 41 (2016), 1248–1275.
[18] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. 2017. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports* 7, 1 (2017), 1–14.
[19] Weixin Li and N. Vasconcelos. 2015. Multiple instance learning for soft bags via top instances. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 4277–4285.
[20] Hongseok Namkoong and John C Duchi. 2017. Variance-based Regularization with Convex Objectives. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.
[21] I.R. Petersen, M.R. James, and P. Dupuis. 2000. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Trans. Automat. Control* 45, 3 (2000), 398–412.
[22] Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv* (2017).
[23] Burr Settles. 2009. Active learning literature survey. (2009).
[24] Burr Settles, Mark Craven, and Soumya Ray. 2007. Multiple-instance active learning. *NIPS* 20 (2007), 1289–1296.
[25] Burr Settles, Mark Craven, and Soumya Ray. 2008. Multiple-Instance Active Learning. In *NIPS*, J. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), Vol. 20.
[26] Wanhua Su, Yan Yuan, and Mu Zhu. 2015. A relationship between the average precision and the area under the ROC curve. In *ICTIR*. 349–352.
[27] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-World Anomaly Detection in Surveillance Videos. In *CVPR*.
[28] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. 2016. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 12 (2016), 2591–2600.
[29] Xin Xu and Eibe Frank. 2004. Logistic Regression and Boosting for Labeled Bags of Instances. In *PAKDD*.
[30] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. 2021. Multiple instance active learning for object detection. *CoRR* abs/2104.02324 (2021). arXiv:2104.02324
[31] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. 2009. Multi-Instance Learning by Treating Instances as Non-I.I.D. Samples. In *ICML*. 1249–1256.
[32] Dixian Zhu, Zhe Li, Xiaoyu Wang, Boqing Gong, and Tianbao Yang. 2019. A Robust Zero-Sum Game Framework for Pool-based Active Learning. In *AISTATS*, Vol. 89. 517–526.

A PROOFS OF THEOREMS

In this section, we provide the detailed proofs for both theorems.

Proof of Theorem 1. Our proof of Theorem 1 is adapted from [7] 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 3 (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 $n \geq 2$,*

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

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

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

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

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

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

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

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 $\mathbf{u} \geq -\frac{1}{n}$, which satisfies if and only if

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

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 3. Since $f(\mathbf{x}_i^+) \in [0, 1]$, we have $|f(\mathbf{x}_i^+) - \bar{f}| \leq 1$. To satisfy inequality (18), it is sufficient to have

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

Let us define the following event

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

In Theorem 1, we suppose $n \geq \frac{4\lambda}{\sigma^2}\max\{2\sigma, 11\})$. Then, on event ϵ_n, we have $n \geq \frac{44\lambda}{\sigma^2} \geq \frac{\lambda}{\text{Var}_n[f(X^+)]}$, 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 3. First, $L = 1$ in our case, which gives

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

The following also holds true:

$$\begin{aligned} P\left(\sigma - t \leq \sqrt{\text{Var}_n[f(X^+)]}\right) &\geq P(\sigma - t \leq \sqrt{\text{Var}_n[f(X^+)]} \leq \sigma + t) \\ &\geq 1 - \exp\left(-\frac{nt^2}{2}\right) \end{aligned}$$

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

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

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 rebalancing 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 (5) can be written as

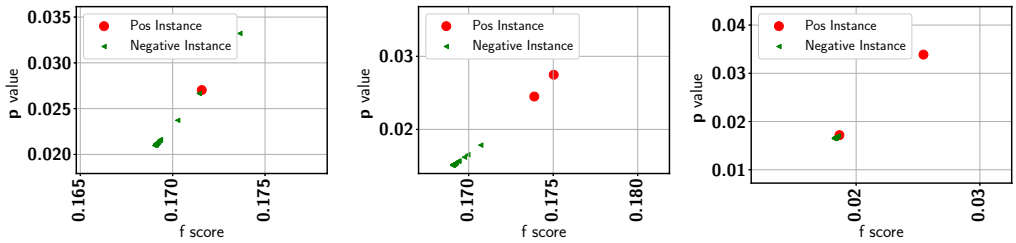


Figure 8: Example of challenging bags from different topics in 20NewsGroup

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

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 [17, 21]:

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

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

where 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

$$\beta\psi'(\beta) - \psi(\beta) = \sum_{m=0}^{\infty}\frac{1}{m!}\kappa_{m+1}\beta^{m+1} - \sum_{m=0}^{\infty}\frac{1}{m!}\kappa_m\beta^m$$
$$= \sum_{m=1}^{\infty}\left[\frac{1}{(m-1)!} - \frac{1}{m!}\right]\kappa_m\beta^m$$
$$= \sum_{m=2}^{\infty}\frac{1}{m(m-2)!}\kappa_m\beta^m = \frac{1}{2}\kappa_2\beta^2 + \frac{1}{3}\kappa_3\beta^3 + \frac{1}{8}\kappa_4\beta^4 + O(\beta^5)$$

In the above expression, $\kappa_m = \psi^{(m)}(0)$ is the m-th derivative of ψ with evaluated at $\beta = 0$ and $O(\beta^5)$ 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, $\frac{d(\beta\psi'-\psi(\beta))}{d\beta}) = \beta\psi''(\beta) > 0$ 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

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

We can obtain β^* as follow

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

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

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

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

B MORE EXAMPLES OF CHALLENGING BAGS

Figure 8 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 LINK TO SOURCE CODE

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