Non-I.I.D. Multi-Instance Learning for Predicting Instance and Bag Labels using Variational Auto-Encoder

Weijia Zhang
School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
weijia.zhang.xh@gmail.com

Abstract

Multi-instance learning is a type of weakly supervised learning. It deals with tasks where the data is a set of bags and each bag is a set of instances. Only the bag labels are observed whereas the labels for the instances are unknown. An important advantage of multi-instance learning is that by representing objects as a bag of instances, it is able to preserve the inherent dependencies among parts of the objects. Unfortunately, most existing algorithms assume all instances to be *identically and independently distributed*, which violates real-world scenarios since the instances within a bag are rarely independent. In this work, we propose the Multi-Instance Variational Autoencoder (MIVAE) algorithm which explicitly models the dependencies among the instances for predicting both bag labels and instance labels. Experimental results on several multi-instance benchmarks and end-to-end medical imaging datasets demonstrate that MIVAE performs better than state-of-the-art algorithms for both instance label and bag label prediction tasks.

1 Introduction

Multi-instance learning (MIL) [Dietterich *et al.*, 1997] was originally proposed for drug activity prediction where the task is to predict whether a molecule is suitable for binding to a receptor. Since each molecule may take many distinct low-energy conformations and scientists only know its suitability for binding at the molecule level, MIL is proposed to model the molecules as bags and the conformations of the molecules as instances, where only the bag labels are provided but the instance labels are unknown to the learner.

Since its inception, MIL has been studied extensively in various applications where the tasks have inherently structured representations or the fine-grained instance labels are expensive to obtain. For example, MIL has been studied in text categorization [Andrews *et al.*, 2002; Zhou *et al.*, 2009; Ji *et al.*, 2020] where articles are represented by bags of sentences with only article-level labels, and in image classifications [Chen *et al.*, 2006; Ilse *et al.*, 2018; Skrede *et al.*, 2020; Yao *et al.*, 2020] where the images are divided into bags of patches with only image-level labels.

As many real-world objects are inherently structured, an important advantage of multi-instance learning is that by representing objects as bags of instances it can convey more information than using a flat single-instance representation. Since instances within a bag correspond to parts of an object, they share structural and contextual information inherited from the bag and are unlikely to be independent. To see this, let us consider the example from drug activity prediction where instances within a bag represent low-energy conformations of the same molecule. The instances are evidently not independently and identically distributed (i.i.d.) since all conformations share the same structure of single bonds and only differ in how the single bonds rotate.

Unfortunately, despite the fact that neither the instances within a bag are independently and identically distributed, nor should they be treated as such [Zhou *et al.*, 2009; Zhang and Zhou, 2014], most of the existing MIL algorithms approach the problem by either explicitly assume all instances to be i.i.d. and directly predict the instance labels, or focus solely on predicting the bag labels by transforming the bags into a single-vector embedded space which buries instance-level information and prohibits instance label prediction.

In this paper, we propose the Multi-Instance Variational Autoencoder (MIVAE) algorithm which explicitly models the instances within a bag as non-i.i.d. using a generative model consisted of a shared bag-level latent factor and instance-level latent factors specific to each of the instances (Figure 1). By using the bag-level factor to capture the structural and contextual dependencies among the instances and using the instance-level factors to capture the instance-specific variations, MIVAE excels at the tasks of both instance label and bag label prediction. On the one hand, integrating the shared bag-level factor and the individualized instance-level factors encapsulate sufficient information for predicting the bag labels. On the other hand, since the shared structural and contextual dependencies are captured by the bag-level factor and the instance-level factors only capture the variations of the individual instances, the instance-level factors promotes better prediction of the instance labels .

The contributions of this work are three-fold: (i) We propose the Multi-instance Variational Autoencoder (MIVAE) model, to the best of our knowledge, the first neural network-based MIL algorithm that simultaneously models the instances as non-i.i.d. (ii) We extend the powerful variational autoencoder framework to MIL and instantiate an implementation of MIVAE for simultaneously inferencing the bag-level and instance-level latent factors and predicting both instance labels and bag labels. (iii) We empirically demonstrate the effectiveness of MIVAE over classic and neural network-based MIL algorithms using a suite of MIL benchmark datasets and end-to-end medical image diagnosis datasets for both instance label and bag label prediction.

The rest of this paper is organized as follows. We discuss related work in Section 2 and present the proposed MIVAE framework in Section 3. In Section 4 we report the experimental results. Finally, we conclude the paper in Section 5.

2 Related Work

Existing MIL algorithms can be roughly divided into two categories depending whether the algorithm operates at the level of instances or the level of bags [Amores, 2013]. The first group of algorithms tackle the MIL problem at the instance-level by training a instance classifier which separates positive instances in the positive bags from the negative ones [Andrews *et al.*, 2002; Kim and la Torre, 2010; Kandemir and Hamprecht, 2014]. The second group of methods operates at the bag-level, either by embedding the bags into a single-vector and solve a single-instance learning problem [Chen *et al.*, 2006; Wei *et al.*, 2017], or designing multi-instance kernels to measure similarities between the bags [Zhou *et al.*, 2009; Xu *et al.*, 2019].

A few MIL algorithms model instances within the same bag as not identically and independently distributed. [Zhou *et al.*, 2009] models the relationships among the instances with an ε-graph capturing the instance similarities. [Zhang and Zhou, 2014] addresses distribution change by considering covariate shift at both instance-level and bag-level.

Recently, several neural network-based approaches have been investigated for MIL. [Wang *et al.*, 2018] extends neural networks to MIL by using permutation-invariant pooling operations on the learned instance embeddings. [Ilse *et al.*, 2018; Shi *et al.*, 2020] propose to use attention in the pooling layer such that the attention weights can be interpreted as how much the instances contribute to the bag label. To the best of our knowledge, all of the existing neural network-based MIL algorithms assume the instances to be i.i.d. and the proposed MIVAE is the first deep generative model for MIL.

Some algorithms has approached MIL with generative models which directly predicting the instance labels by either assuming the Gaussian Process [Kim and la Torre, 2010; Haussmann *et al.*, 2017] or the Dirichlet Process [Kandemir and Hamprecht, 2014]. However, all of them explicitly assume the instances within a bag to be i.i.d. In this work we show that such assumption not only does not agree with the real-world scenarios, but also hurts empirical performance for both bag and instance level label prediction.

Perhaps the work most related to ours is the proposal [Doran and Ray, 2016], where they model the multi-instance bags as distributions over instances. However, their contribution is only applicable to classical MIL tasks with pre-computed features, while MIVAE also excels at end-to-end learning from weakly-labeled data with neural networks.

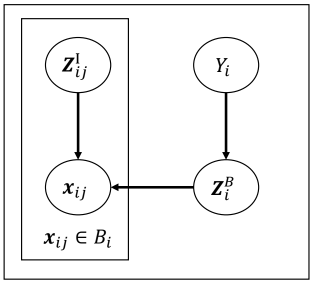


Figure 1: The generative model for the proposed Multi-Instance Variational Autoencoder (MIVAE). Y_i is the bag label for bag B_i, $x_{ij} \in B$ are the instances of B_i. $\mathbf{z}^B$ is the bag-level factor shared by all of the instances within B_i, and $\mathbf{z}^I_{ij}$ are the instance-level factors specific to each instance.

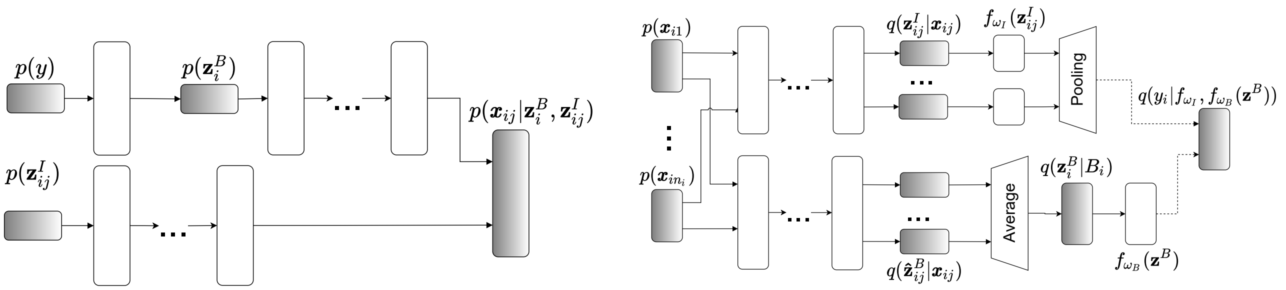
3 Method

Let $\mathcal{X} = \mathbb{R}^d$ denote the space of the instances and $\mathcal{Y} = \{0, 1\}$ denote the space of the label. The learner is given a dataset of m bags multi-instance bags $\mathcal{B} = \{\mathbf{B}_1, \cdots, \mathbf{B}_i, \cdots, \mathbf{B}_m\}$, where each bag $\mathbf{B}_i = \{\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{ij}, \cdots, \boldsymbol{x}_{x_{in_i}}\}$ is a set containing an arbitrary number of n_i instances with $\boldsymbol{x}_{ij} \in \mathcal{X}$. During training, each bag $\mathbf{B}_i$ is provided with a bag label $y_i \in \mathcal{Y}$; however, the label of the instances are unknown. The goal of multi-instance learning as two-fold, to predict both the bag label and the instance labels of unseen bags. When the context is clear, we drop the subscript i from $\boldsymbol{x}_{ij}$, $\mathbf{B}_i$ and y_i for the conciseness of the notations.

At the heart of this work lies the claim that the instances within the same bag are not independently and identically distributed and the i.i.d. assumption is only applicable at the bag-level. On the contrary, instances within the same bag share structural and contextual dependencies inherited from the bag and are only independent to each other when the dependencies have been considered. Apart from drug activity prediction, the dependencies of instances exists widely in MIL applications. For example, in text classification instances are paragraphs/sentences of an article and obviously share the styles of the author and contexts of the topic [Ji *et al.*, 2020]. In medical image diagnosis instances are patches of an organ from a patient which naturally share the pathology features and prognoses of the patient [Skrede *et al.*, 2020].

In order to capture both the shared dependencies and the individual variations of the instances, we propose the MIVAE model as depicted in Figure 1. Specifically, it assumes that for each bag the observed instances are generated from two type of latent factors: a bag-level latent factor $\mathbf{z}^B$ shared by all instances, and n_i instance-level latent factors $\mathbf{z}^I_j, j \in \{1, \cdots, n_i\}$ specific to each of the instances. While the shared bag-level factor $\mathbf{z}^B$ controls the structural and contextual dependencies among the instances, the instance-level factors $\mathbf{z}^I_j$ are responsible for the variations of each instances. When conditioning on the shared bag-level factor $\mathbf{z}^B$, the instances becomes independent to each other.

At the level of bags, MIL is supervised and we assume the bag-level factor $\mathbf{z}^B$ to be dependent on y. However, since both the instance labels and how they relate to the bag labels


$p(y)$
$p(\mathbf{z}_i^B)$
$p(\mathbf{z}_{ij}^I)$
$p(\boldsymbol{x}_{ij}|\mathbf{z}_i^B, \mathbf{z}_{ij}^I)$
$p(\boldsymbol{x}_{i1})$
$p(\boldsymbol{x}_{in_i})$
$q(\mathbf{z}_{ij}^I|\boldsymbol{x}_{ij})$
$f_{\omega_I}(\mathbf{z}_{ij}^I)$
Pooling
$q(y_i|f_{\omega_I}, f_{\omega_B}(\mathbf{z}^B))$
$q(\mathbf{z}_i^B|B_i)$
Average
$q(\hat{\mathbf{z}}_{ij}^B|\boldsymbol{x}_{ij})$
$f_{\omega_B}(\mathbf{z}^B)$

(a) Generative Model.

(b) Inference Model.

Figure 2: Architecture of the model network and the inference network for the Multi-Instance Variational AutoEncoder (MI-VAE). White nodes correspond to parametrized deterministic neural network transformations, gray nodes correspond to drawing samples from the respective distribution. Dashed arrows in the inference model represent the auxiliary classifiers $q_{\omega_y}(y|\mathbf{z}^B, \mathbf{z}^I)$.

are unknown, e.g., instances may not have explicit labels and the bag labels may not be determined by a single instance, assuming fixed generative process may restrict the expressiveness of the model. We address the relationship between the instance-level factors and the bag label during inference.

3.1 Multi-Instance Variational Autoencoder

In practice we only observe the bags and bag labels, our goal is thus to infer the unobservable posterior distributions for both the bag-level and the instance-level latent factors $p_\theta(\mathbf{z}^B|\mathbf{B})$, $p_\theta(\mathbf{z}_j^I|\boldsymbol{x}_j)$, and the posterior distribution of the observable $p_\theta(\mathbf{B}|\mathbf{z}^B, \mathbf{z}_1^I, \cdots, \mathbf{z}_j^I)$ where θ denotes the generative parameters. Since exact inference of the posterior is intractable due to the fact that both the marginal likelihood and the posterior lack analytical solution, we employ variational autoencoder [Kingma and Welling, 2014] parameterized by neural networks for efficient approximate inference.

As depicted in Figure 2(a), we utilize two separately-organized encoding networks $q_{\phi_B}(\mathbf{z}^B|\mathbf{B})$ and $q_{\phi_I}(\mathbf{z}_j^I|\boldsymbol{x}_j)$, where ϕ_B and ϕ_I denote the parameters, to serve as variational posteriors of the bag-level and instance-level latent factors. The encoded latent factors are then feed into a single decoder $p_\theta(\boldsymbol{x}_j|\mathbf{z}^B, \mathbf{z}_j^I)$ for reconstructing the bag of instances. Following standard VAE design, the prior distributions of the latent factors $p(\mathbf{z}^B)$ and $p(\mathbf{z}_j^I)$ are chosen as multivariate Gaussians. Specifically, the prior distributions of the factors and the generative model are factorized as:

$$p(\mathbf{z}^B|y) = \prod_{k=1}^{D_{z_B}} \mathcal{N}(z_k^B|f_y(y), 1);$$

$$p(\mathbf{z}_j^I) = \prod_{k=1}^{D_{z_I}} \mathcal{N}(z_{jk}^I|0, 1);$$

$$p(\boldsymbol{x}_j|\mathbf{z}^B, \mathbf{z}_j^I) = \prod_{k=1}^{d} p(x_{jk}|\mathbf{z}^B, \mathbf{z}_j^I),$$

where $p(x_{jk}|\mathbf{z}^B, \mathbf{z}_j^I)$ is the suitable distribution for the k-th feature of the instance, i.e., Gaussians for continuous features or Bernoulli for binary features; f_y is a function parameterized by neural network using the bag label as input; D_{z_B} and D_{z_I} are the parameters that determine the dimensionality of the bag-level and instance-level latent factors.

Comparing to a standard VAE, the inference model of MI-VAE needs to overcome two challenges specific to MIL as depicted in Figure 2(b). Firstly, apart from inferring the instance-level factors $\mathbf{z}^I$ specific to each instance, MIVAE needs to infer a common bag-level factor $\mathbf{z}^B$ shared by n_i individual instances $\boldsymbol{x}_j, j = 1, \cdots, n_i$; however, in standard VAE the latent variables are specific to each of the inputs. Secondly, without any instance labels, MIVAE needs to infer the instance labels from the instance-level factors $\mathbf{z}_j^I$.

Given instances $\boldsymbol{x}_j$, the instance-level factors $\mathbf{z}_j^I$ can be inferred directly with a standard encoder $q_{\phi_I}(\mathbf{z}_j^I|\boldsymbol{x}_j)$ parameterized by neural networks. For the shared bag-level factor $\mathbf{z}^B$, a straightforward approach would be directly encoding from the concatenation of all instances within the bag; however, this approach unfortunately violates the permutation-invariant property of the instances, i.e., the encoded bag-level factor may dependent on the ordering of the instances.

To avoid this problem, we design MIVAE to first encode the intermediate bag-level factors $\hat{\mathbf{z}}_j^B$ using $q_{\phi_B}(\hat{\mathbf{z}}_j^B|\boldsymbol{x}_j)$ for each instance, and then utilize a permutation-invariant function for encouraging the intermediate instance-specific factors into a shared bag-level factor. Specifically, the variational approximations for the instance-level factors and the intermediate bag-level factors are defined as

$$q_{\phi_I}(\mathbf{z}_j^I|\boldsymbol{x}_j) = \prod_{k=1}^{D_{z_I}} \mathcal{N}(\mu = f_{\phi_I}^{\psi}(\boldsymbol{x}_j), \sigma^2 = f_{\phi_I}^{\pi}(\boldsymbol{x}_j)), \quad (1)$$

$$q_{\phi_B}(\hat{\mathbf{z}}_j^B|\boldsymbol{x}_j) = \prod_{k=1}^{D_{z_B}} \mathcal{N}(\mu = f_{\phi_B}^{\psi}(\boldsymbol{x}_j), \sigma^2 = f_{\phi_B}^{\pi}(\boldsymbol{x}_j)), \quad (2)$$

for $\boldsymbol{x}_j$ with $j = 1, \cdots, n_i$, where $f_{\phi_I}^{\psi}, f_{\phi_B}^{\psi}$ and $f_{\phi_I}^{\pi}, f_{\phi_B}^{\pi}$ are the means and variances for the Gaussian distributions parameterized by neural networks. Then, the shared bag-level factor is calculated as the mean of the intermediate factors as

$$q_{\phi_B}(\mathbf{z}_B|\mathbf{B}) = \prod_{k=1}^{D_{z_B}} \mathcal{N}(\tfrac{1}{n_i} \sum_{j=1}^{n_i} f_{\phi_B}^{\psi}(\boldsymbol{x}_j), \tfrac{1}{n_i} \sum_{j=1}^{n_i} f_{\phi_B}^{\pi}(\boldsymbol{x}_j)). \quad (3)$$

In other words, the encoding neural networks $f^{\psi}_{\phi_B}$ and $f^{\pi}_{\phi_B}$ first infer the means and variances of the posterior distributions for the intermediate factors from the instances $\boldsymbol{x}_j$, then the shared bag-level factor is calculated as the average of all the intermediate bag-level factors. It is worth noting that the variance of $\mathbf{z}^B$ then become the average of all the intermediate factor variances. In this way, as the decoder reconstructs each of the instances using the shared bag-level factor $\mathbf{z}^B$ and individual instance-level factor $\mathbf{z}^I_j$, all of the intermediate bag-level factors are encouraged to converge to the same value during optimization. Therefore, $\mathbf{z}^B$ will become a shared bag-level latent factor common to all $\boldsymbol{x}_j$, while each $\mathbf{z}^I_j$ captures as the instance-specific variations.

The generative and inference models of MIVAE can be learned by maximizing the marginal likelihood over the bags

$$\log p_\theta(\mathbf{B}, y) = \mathbb{E}_{q_{\phi_B} q_{\phi_I}}[\log \frac{p_\theta(\mathbf{B}, y, \mathbf{z}^B, \mathbf{z}^I)}{q_\phi(\mathbf{z}^B, \mathbf{z}^I|\mathbf{B}, y)}] + \mathbb{E}_{q_{\phi_B} q_{\phi_I}}[\log \frac{q_\phi(\mathbf{z}^B, \mathbf{z}^I|\mathbf{B}, y)}{p_\theta(\mathbf{z}^B, \mathbf{z}^I|\mathbf{B}, y)}], \tag{4}$$

where the second term is the Kullback-Leibler (KL) divergence from the variational approximation to the true posterior. Because the KL-divergence is non-negative, the first term serves as a lower bound of the marginal likelihood which is also called the evidence lower bound (ELBO). Maximizing the marginal likelihood is equivalent to maximizing the ELBO, which can be re-written as

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(\mathbf{B}, y) =& \mathbb{E}_{q_{\phi_B} q_{\phi_I}}[\log p_\theta(\mathbf{B}, y|\mathbf{z}^I, \mathbf{z}^B)] \\ &- D_{KL}[q_{\phi_B}(\mathbf{z}^B|\mathbf{B})||p_{\theta_B}(\mathbf{z}^B|y)] \\ &- \sum_{j=1}^{n_i} D_{KL}[q_{\phi_I}(\mathbf{z}^I_j|\boldsymbol{x}_j)||p_{\theta_I}(\mathbf{z}^I_j)]. \end{aligned} \tag{5}$$

A missing piece of the puzzle is how to predict the instance labels from the inferred instance-level factors $\mathbf{z}^I$ and predict the bag labels from both bag-level and instance-level factors. Instead of assuming fixed generative distributions which may restrict the applicability of the model, we employ an auxiliary classifier $q_{\omega_I}(y|\mathbf{z}^B, \mathbf{z}^I_1, \cdots, \mathbf{z}^I_{n_i})$ attached to the evidence lower bound for end-to-end prediction of both instance labels and bag labels. Using the auxiliary classifier allows MIVAE to have the flexibility of utilizing any permutation-invariant multi-instance pooling techniques, i.e., max-pooling, LSE-pooling, and attention pooling [Ilse *et al.*, 2018]. In order to demonstrate that disentangling the dependencies among instances into $\mathbf{z}^B$ and using only the instance-level factors $\mathbf{z}^I_j$ promotes better instance label prediction performance, we use the straightforward max-pooling operation which is defined as

$$f_I = \max_{i=1,\cdots,n_i}\{f_{\omega_I}(\mathbf{z}^I_1), \cdots, f_I(\mathbf{z}^I_{n_i})\}, \tag{6}$$

where f_{ω_I} is a neural network with a sigmoid activation to map the instance-level factors to probabilities. Then, the auxiliary for predicting the bag label can be expressed as

$$q_\omega(y|\mathbf{z}^B, \mathbf{z}^I_1, \cdots, \mathbf{z}^I_{n_i}) = q_\omega(y|f_{\omega_B}(\mathbf{z}^B), f_I), \tag{7}$$

where f_{ω_B} is also a parameterized neural network that outputs probabilities. Finally, the MIVAE objective can be expressed as the sum of the ELBO and the auxiliary classifier,

$$\mathcal{L}_{\text{MIVAE}} = \mathcal{L}_{\text{ELBO}}(\mathbf{B}, y) + \alpha\mathbb{E}_{q_{\phi_B} q_{\phi_I}}[\log q_\omega(y|\mathbf{z}^B, \mathbf{z}^I_1, \cdots, \mathbf{z}^I_{n_i})], \tag{8}$$

where α is a weighting parameter of the auxiliary objective. Using the re-parameterization trick [Kingma and Welling, 2014], the objective of MIVAE can be efficiently optimized using stochastic gradient descent.

Method	Musk1	Musk2	Fox	Tiger	Elephant
mi-SVM	.874±.12	.836±.09	.582±.10	.789±.09	.820±.07
MILES	.842±.08	.838±.10	**.760**±**.05**	.840±.08	**.891**±**.05**
mi-Graph	.889±.07	**.903**±**.09**	.616±.08	.801±.08	.869±.08
miFV	.878±.01	.868 ±.09	.621±.11	.813±.08	.852±.08
mi-Net	.892±.04	.858±.05	.615±.04	.839±.06	.868±.05
AttnMIL	.900±.05	.863±.04	.603±.06	.845±.05	.857±.06
MIVAE	**.904**±**.05**	.890±0.6	.626±.06	**.850**±**.05**	.870±.06

Table 1: Bag label prediction results on five benchmark MIL benchmarks. The highest average accuracy is marked in bold.

For predicting the instance labels of unseen bags, we use the trained instance-level encoder $q_{\phi_I}(\mathbf{z}^I_j|\boldsymbol{x}_j)$ to infer the mean of the posteriors for each $\boldsymbol{x}_j \in B$, and use the intermediate output $f_{\omega_I}(\mathbf{z}^I_{nj})$ from the auxiliary classifier. For predicting the bag labels, we use the bag-level encoder $q_{\phi_B}(\mathbf{z}^B|\mathbf{B})$ to infer the mean of the bag-level latent factor $\mathbf{z}^B$, and directly use the output of the auxiliary classifier $q_\omega(y|\mathbf{z}^B, \mathbf{z}^I_1, \cdots, \mathbf{z}^I_{n_i})$ for prediction.

4 Experiments

We empirically evaluate MIVAE against state-of-the-art MIL algorithms for both bag label prediction and instance label prediction tasks. Firstly, we compare against SVM-based algorithms including mi-SVM [Andrews *et al.*, 2002], miFV [Wei *et al.*, 2017], mi-Graph [Zhou *et al.*, 2009], and KI-SVM [Li *et al.*, 2009]. Secondly, we compare against neural network-based algorithms including mi-Net and MI-Net [Wang *et al.*, 2018], and AttentionMIL [Ilse *et al.*, 2018]. Thirdly, we compare against generative MIL algorithms for predicting instance labels, including Gaussian Process MIL (GPMIL) [Kim and la Torre, 2010], Dirichlet Process MIL (DPMIL) [Kandemir and Hamprecht, 2014], and Variational Gaussian Process MIL (VGPMIL) [Haussmann *et al.*, 2017].

We implement MIVAE using PyTorch. Unless otherwise stated, the reported results are obtained from averaging the test results of 10 times repeated 10-fold cross validation as in previous MIL literature. Training for MIVAE are conducted for 100 epochs and parameters are tuned using validation loss evaluate with 10% of the training bags. Experiments are conducted using a PC with a single Nvidia GTX 1080Ti GPU.

4.1 Bag Classification: MIL Benchmarks

We first study the performance of MIVAE for the bag label predictions on five MIL benchmark datasets including the drug activity prediction datasets Musk1 and Musk2, and three content-based image retrieval tasks Fox, Tiger, and Elephant. Following previous works, we report the average accuracy

	mi-SVM	KI-SVM	GPMIL	DPMIL	VGPMIL	MIVAE
alt.atheism	0.53	0.68	0.44	0.67	0.70	**0.745±.030**
comp.graphics	0.65	0.47	0.49	0.79	0.79	**0.800±.042**
comp.os.ms-windows.misc	0.42	0.38	0.36	0.51	0.52	**0.548±.038**
comp.sys.ibm.pc.hardware	0.57	0.31	0.35	0.67	0.70	**0.711±.034**
comp.sys.mac.hardware	0.56	0.39	0.54	0.76	**0.79**	0.783±.035
comp.windows.x	0.56	0.37	0.36	0.73	0.69	**0.754±.032**
misc.forsale	0.31	0.29	0.33	0.45	0.54	**0.553±.334**
rec.autos	0.51	0.45	0.38	0.76	0.71	**0.720±.024**
rec.motorcycles	0.09	0.52	0.46	0.69	0.76	**0.766±.029**
rec.sport.baseball	0.18	0.52	0.38	0.74	0.76	**0.764±.036**
rec.sport.hockey	0.27	0.66	0.43	0.91	**0.94**	0.925±.020
sci.crypt	0.57	0.47	0.31	0.68	**0.82**	0.773±.036
sci.electronics	0.83	0.42	0.71	0.90	0.92	**0.928±.020**
sci.med	0.37	0.55	0.32	0.73	0.73	**0.745±.025**
sci.space	0.46	0.51	0.32	0.70	0.74	**0.748±.027**
soc.religion.christian	0.05	0.53	0.45	0.72	0.73	**0.753±.035**
talk.politics.guns	0.57	0.43	0.38	0.64	**0.72**	0.714±.038
talk.politics.mideast	0.77	0.60	0.46	0.80	**0.87**	0.840±.020
talk.politics.misc	0.61	0.50	0.29	0.60	0.64	**0.650±.044**
talk.religion.misc	0.08	0.32	0.32	0.51	0.49	**0.525±.035**
W/T/L	20/0/0	20/0/0	20/0/0	20/0/0	15/0/5	

Table 2: Average test Area Under the Precision-Recall Curve (AUC-PR) of the 20 Newsgroup datasets. The results for the compared methods are obtained from the literature where standard deviations are not reported. The highest average AUC-PR is marked in bold. The last row summarizes the average of the 20 datasets.

and the standard deviation obtained from repeating 10-fold cross validation for 10 times.

Since these benchmark datasets only contains a small number of bags (approximately 100 bags for the Musk datasets and 200 bags for the image retrieval datasets) and the precomputed feature vectors, we do not expect MIVAE to significantly outperform existing methods since its deep generative model may not be well suited for these type of tasks. However, from Table 1 we can see that the performance of MIVAE is competitive to the compared baselines. Specifically, MIVAE achieves the highest average accuracy on Musk1 and Tiger datasets, and performs competitively on the others. The results demonstrate that the model of MIVAE is sucessful for handling the uncertainties of multi-instance learning.

Method	Accuracy	AUC-PR
mi-Net	0.824±0.021	0.491±0.028
MI-Net	0.845±0.015	0.466±0.031
AttentionMIL	0.900±0.021	0.500±0.034
Gated-AttentionMIL	0.895±0.026	0.513±0.023
MIVAE	**0.925±0.014**	**0.747±0.032**

Table 3: Bag label and instance label prediction results on the test data of the Colon Cancer dataset. Experiments were repeated for 5 times and the average accuracy ± standard deviation of 10-fold cross validations are reported.

4.2 Instance Classification: 20 NewsGroups

To study the performance for predicting instance labels, we evaluate MIVAE against baselines using the 20 Newsgroup dataset [Zhou *et al.*, 2009]. It contains a collection of 20 datasets where each of them is consisted of 100 bags. Each bag contains approximately 40 instances of articles from 20 different topics where each instance represents one article described by the top 200 TF-IDF features. A bag is considered to be positive if at least one of its instances belongs to a specific topic. The labels of the bags are distributed evenly; however, in each positive bag only approximately 3% of the instances are positive. Because of the proportions of positive instances are small, the dataset has been widely used as a standard benchmark for instance label prediction performances.

Following previous work, we report results obtained from repeating 10-fold cross validation for 10 times using the train/test splits provided in [Zhou *et al.*, 2009]. Because the ground truth instance labels are highly imbalanced, we use Area Under the Precision-Recall Curve (AUC-PR) instead of the area under the receiver operating characteristic curve. Furthermore, in order to demonstrate that the advantage of MIVAE lies in its model instead of parameter tuning, the parameters of MIVAE are selected using only the first of the 20 Newsgroup datasets (*alt.atheism*), and the same parameters are applied across all the rest of the datasets.

As shown by the Table 2, without tuning parameters for each specific dataset, MIVAE outperforms the compared baselines on 15 out of 20 datasets for predicting instance labels despite the small number of training bags. Since all compared algorithms assume the instances to be i.i.d., these results confirm the advantage of MIVAE for modeling the dependencies among the instances.

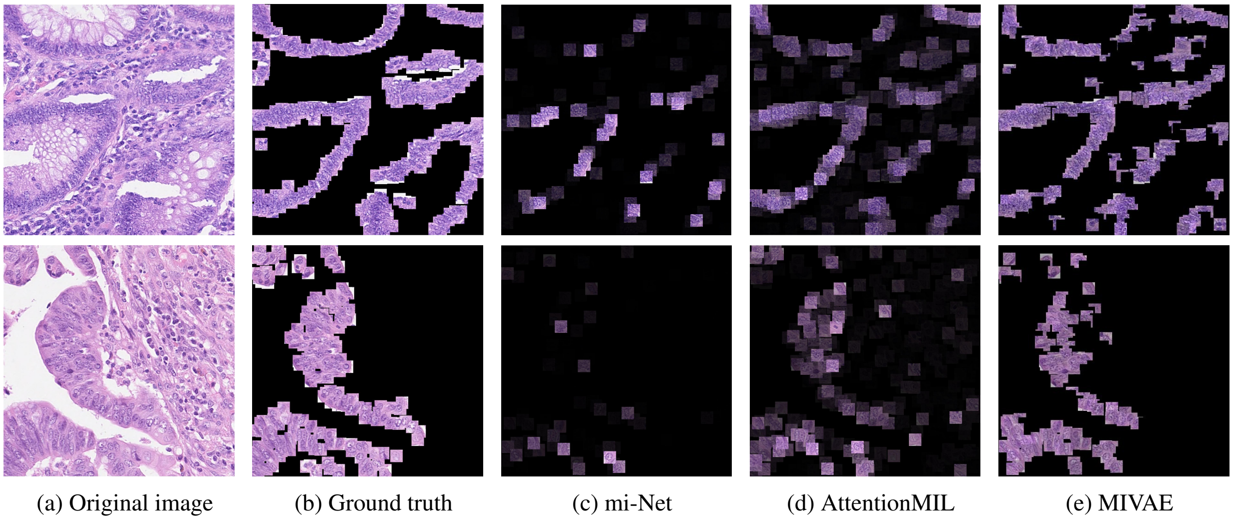


Figure 3: (a) Original H&E stained images of the Colon Cancer dataset. (b) Ground truths. (c) Prediction heatmaps of mi-Net. (d) Prediction heatmaps of AttentionMIL. (e) Prediction heatmaps of MIVAE. This figure is best viewed in color.

4.3 Histopathology Images: Colon Cancer

Computer-aided histopathology using hematoxylin and eosin (H&E) stained whole-slide images is an important task for MIL since traditional supervised approaches require pixel-level annotations, which is extremely labor-intensive for pathologists [Skrede *et al.*, 2020]. Therefore, a label efficient multi-instance approach using only image-level labels is essential. For this purpose, we evaluate MIVAE against state-of-the-art end-to-end neural network-based MIL algorithms on both bag and instance label prediction using H&E histopathology images from the colon cancer dataset provided in [Sirinukunwattana *et al.*, 2016].

The dataset contains 100 H&E images where each image is from a patient's tissue of either normal or malignant regions. For each image bag, the instances are generated as patches 27×27 pixels using markings of major nuclei for each cell. A total amount of 22,444 nuclei instances are provided with ground truth instance class labels, i.e. epithelial, inflammatory, fibroblast, and miscellaneous. The bag is considered to be positive if it contains at least one epithelial patch.

For fairness of comparison, the training images are color normalized and augmented with the same random horizontal flipping/vertical flipping, and random rotation procedures. Furthermore, they also use the same convolutional structures for feature extraction (please refer to the supplementary for details). When evaluating bag label prediction, we use accuracy since the bag labels are evenly distributed. Because the instances label are imbalanced, we use AUC-PR for quantitative evaluating instance label prediction performance as in the last section. The reported results are the average performances from 5 times repeated 10-fold cross-validation.

As shown in the results of Table 3, MIVAE achieves both the highest classification accuracy for predicting bag labels and the best AUC-PR scores for predicting instance labels, indicating that MIVAE outperforms state-of-the-art neural network-based MIL algorithms at both tasks.

In Figure 3, we use heatmaps to visually compare the instance label prediction performances. For each image, the heatmaps are obtained by multiplying the instances prediction scores of each algorithm (normalized to the range of $[0, 1]$) to the pixel values of the image patches. From Figure 3 (b)-(e), it is evident that the heatmaps generated by MIVAE are the closest to the ground truths, while mi-Net exhibits lower recalls and the predictions of AttentionMIL are not as confident as the predictions of MIVAE.

It is worth noting that both mi-Net and AttentionMIL assume the instances to be independently distributed. Moreover, mi-Net also uses max-pooling when linking instance prediction scores to bag label. Therefore, contrasting the heatmaps of MIVAE and mi-Net paints a clear picture for the benefits of modelling the instance dependencies: mi-Net only selects a small portion of positive patches, while MIVAE is able to identify most of them because the shared dependencies are excluded from instance label predictions. Furthermore, comparing MIVAE with AttentionMIL shows that by using the instance-level factors, max-pooling provides more confident predictions than the attention weights. For more examples, please kindly refer to the supplementary materials.

To sum up, the experiments results demonstrates that the proposed MIVAE algorithm is effective at both bag label and instance label prediction tasks. Moreover, the instance prediction scores of MIVAE can be used for explaining the region of interests which is valuable for weakly-labeled tasks where fine-grained labels are costly to obtain.

5 Conclusion

In this paper, we propose MIVAE, a generative multi-instance learning algorithm which models the instances within a bag as non-i.i.d.. Experiments demonstrate that MIVAE is effective for both instance-level and bag-level label prediction. An interesting future direction would be to investigate how other MIL pooling techniques affect the performance of MIVAE.

References

[Amores, 2013] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

[Andrews *et al.*, 2002] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Proceedings of the 15th International Conference on Neural Information Processing Systems (NeurIPS '02)*, pages 577–584, 2002.

[Chen *et al.*, 2006] Yixin Chen, Jinbo Bi, and J.Z. Wang. MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1931–1947, 2006.

[Dietterich *et al.*, 1997] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[Doran and Ray, 2016] Gary Doran and Soumya Ray. Multiple-instance learning from distributions. *Journal of Machine Learning Research*, 17(128):1–50, 2016.

[Haussmann *et al.*, 2017] 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 '17)*. IEEE, 2017.

[Ilse *et al.*, 2018] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML '18)*, pages 2127–2136, 2018.

[Ji *et al.*, 2020] Yunjie Ji, Hao Liu, Bolei He, Xinyan Xiao, Hua Wu, and Yanhua Yu. Diversified multiple instance learning for document-level multi-aspect sentiment classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP '20)*, pages 7012–7023, 2020.

[Kandemir and Hamprecht, 2014] Melih Kandemir and Fred A. Hamprecht. Instance label prediction by dirichlet process multiple instance learning. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI '14)*, 2014.

[Kim and la Torre, 2010] Minyoung Kim and Fernando De la Torre. Gaussian processes multiple instance learning minyoung. In *Proceedings of the 27th International Conference on Machine Learning (ICML '10)*, 2010.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR '14)*, 2014.

[Li *et al.*, 2009] Yu-Feng Li, James T.Kwok, Ivor W. Tsang, and Zhi-Hua Zhou. A convex method for locating regions of interest with multi-instance learning. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD '09)*, 2009.

[Shi *et al.*, 2020] Xiaoshuang Shi, Fuyong Xing, Yuanpu Xie, Zizhao Zhang, Lei Cui, and Lin Yang. Loss-based attention for deep multiple instance learning. *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI '20)*, pages 5742–5749, 2020.

[Sirinukunwattana *et al.*, 2016] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David R. J. Snead, Ian A. Cree, and Nasir M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1196–1206, 2016.

[Skrede *et al.*, 2020] Ole-Johan Skrede, Sepp De Raedt, Andreas Kleppe andTarjei S Hveem, Knut Liestøl, John Maddison, Hanne A Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, Inger Nina Farstad, Enric Domingo, David N Church, Arild Nesbakken, Neil A Shepherd, Ian Tomlinson, Rachel Kerr, Marco Novelli, David J Kerr, and Håvard E Danielsen. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet*, 35(10221):350–360, 2020.

[Wang *et al.*, 2018] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018.

[Wei *et al.*, 2017] Xiu-Shen Wei, Jianxin Wu, and Zhi-Hua Zhou. Scalable algorithms for multi-instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 28(4):975–987, 2017.

[Xu *et al.*, 2019] Bi-Cun Xu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation set-kernel and its application to multi-instance learning. In *Proceedings of The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19)*, pages 941–949, 2019.

[Yao *et al.*, 2020] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

[Zhang and Zhou, 2014] Weijia Zhang and Zhi-Hua Zhou. Multi-instance learning with distribution change. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI '14)*, pages 2184–2190, 2014.

[Zhou *et al.*, 2009] 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 International Conference on Machine Learning (ICML '09)*, pages 1249–1256, 2009.