Semi-supervised Multiple Instance Learning using Variational Auto-Encoders

Anonymous

Editors: Under Review for MIDL 2022

Abstract

We consider the multiple-instance learning (MIL) paradigm, which is a special case of supervised learning where training instances are grouped into bags. In MIL, the hidden instance labels do not have to be the same as the label of the comprising bag. On the other hand, the hybrid modelling approach is known to possess advantages basically due to the smooth consolidation of both discriminative and generative components. In this paper, we investigate whether we can get the best of both worlds (MIL and hybrid modelling), especially in a semi-supervised learning (SSL) setting. We first integrate a variational autoencoder (VAE), which is a powerful deep generative model, with an attention-based MIL classifier, then evaluate the performance of the resulting model in SSL. We assess the proposed approach on an established benchmark as well as a real-world medical dataset.

Keywords: Multiple-Instance Learning, Variational Autoencoders, Deep Generative Models

1. Introduction

In the standard form of supervised learning, it is assumed that the learner encounters training data in a flat form where each instance, e.g., an image, belongs to a class (category). However, another setting which can be more practical in representing many real-world applications is multiple-instance learning (MIL), where training instances are grouped together into bags. In MIL, both bags and instances have labels, but an instance within a bag may have a different label from that of the bag. Only the bag label is available for learning since instance labels are not observed. Several applications can be cast as MIL problems, e.g., in medical imaging (Melendez et al., 2014; Quellec et al., 2017; Tellez et al., 2019; Weitz et al., 2021), neuromuscular disorders diagnosis (Adel et al., 2013), and computational biology (Dietterich et al., 1997).

The principal goal of MIL is to learn a model which can predict the bag label. This corresponds to the molecule binding property in the above example or to the all-important medical diagnosis in medical imaging applications. Nonetheless, inferring which instances are the most influential in predicting the bag label is of major importance due to several reasons including interpretability of the obtained prediction (especially in medical diagnosis) and related issues like GDPR (General Data Protection Regulation) which forces the right to understand in sensitive applications like self-driving cars and medical applications.

In this work, we investigate how the MIL framework fares in the semi-supervised learning paradigm (SSL, Zhu et al., 2003; Chapelle et al., 2006; Kingma et al., 2014; Siddharth et al., 2017). In SSL, the data presented to the learner typically consists of a few labeled examples as well as numerous unlabeled examples. The main goal of a semi-supervised learner is to utilize the unlabeled data in order to improve the model's performance on the supervised subset of the data. In case of the SSL MIL setting, the supervision is at the bag level. This means that the learner encounters both

© 2022 A Sutera et al.

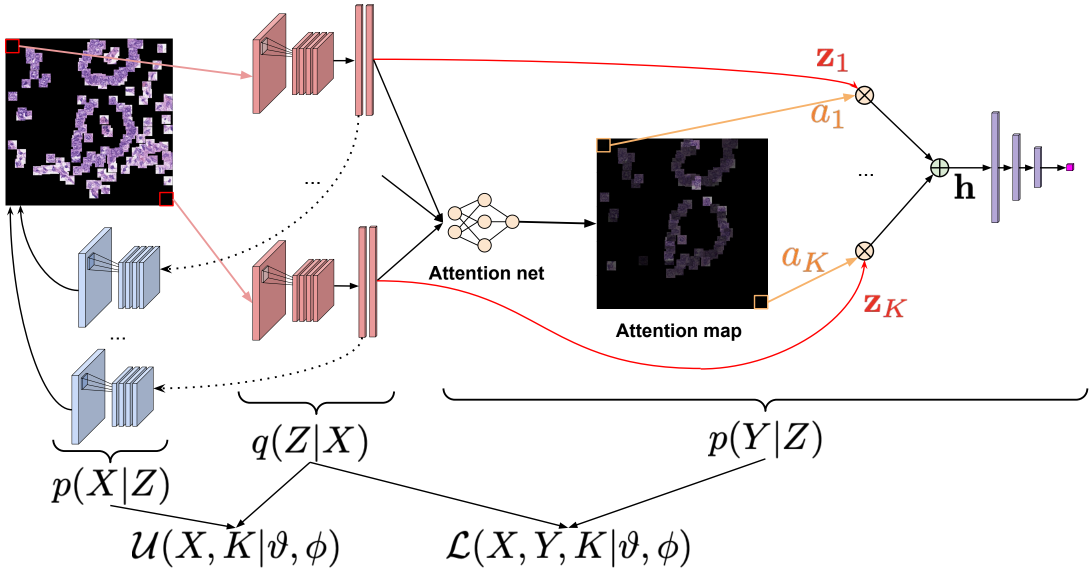


Figure 1: A schematic representation of the proposed approach. The variational posterior (in red) produces an embedding for each instance that is further processed by the attention network (in yellow). The weighted average of embeddings is fed to the classifier. Additionally, each embedding is decoded back to the pixel space (in blue). We highlight which components (in orange) are used in the unsupervised loss $\mathcal{U}(X, K|\vartheta, \phi)$ and the supervised loss $\mathcal{L}(X, Y, K|\vartheta, \phi)$. *Best viewed in color.*

labeled and unlabeled bags. SSL has a long history in MIL. However, typically speaking, models other than deep neural networks were utilized (Zhang and Goldman, 2001; Andrews et al., 2002; Rahmani and Goldman, 2006; Li et al., 2009). To deal with both the labeled and unlabeled data, we propose to learn a joint distribution over instances and a bag label within the hybrid modeling framework using deep neural networks. Hybrid models are known to combine the advantages of (standard supervised) discriminative models with those of generative models (Jaakkola and Haussler, 1999; Tulyakov et al., 2017; Nalisnick et al., 2019). Hybrid models have also been exploited in other frameworks including semi-supervised learning (Ilse et al., 2020a; Nalisnick et al., 2019) and anomaly detection (Maaloe et al., 2019; Liu and Abbeel, 2020). In this work, we propose an MIL framework (see Figure 1) which leverages the prowess of hybrid models so that they can excel in problems and applications possessing the bag-instance nature modelled by MIL. The core idea lies in using unsupervised components for learning low-dimensional embeddings (in red and blue in Figure 1) together with a classifier (in red, yellow and purple in Figure 1). As a result, we can learn an MIL model in the semi-supervised setting. We build our modelling on top of the seminal attention-based deep MIL classifier (Ilse et al., 2018, 2020b; Ing et al., 2018; Tomczak et al., 2018), mainly due to its permutation-invariant characteristics and its ability to give instance weights which can be interpreted as the contributions of each instance to the bag label. As a result, we formulate a latent variable model that could be seen as a Variational Auto-Encoder (VAE, Kingma and Welling, 2014; Rezende et al., 2014) for instances and a classifier that is fed with the outputs of the VAE's encoder.

Our main contributions are threefold: (1) Integrating an attention-based Deep MIL classifier with a deep generative model in the form of a VAE. (2) Developing an SSL framework based on the proposed hybrid MIL approach. (3) Evaluating the proposed hybrid approach on the semi-supervised

MIL scenario and comparing it with baselines on two datasets: MNIST-BAGS (Ilse et al., 2018), and COLON-CANCER (Sirinukunwattana et al., 2016).

2. Methodology

2.1. Multiple-Instance Learning

In standard binary classification, the main goal is to establish a model which predicts the target variable $y \in \{0, 1\}$ for a data instance $\mathbf{x} \in \mathbb{R}^D$. On the other hand, each data sample in an MIL paradigm comes in the form of a bag of unordered and independent[1] instances $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\}$, where the number of instances, referred to as K can differ for different bags. An MIL model must learn to predict the bag label Y, which is observed for the training data instances. In addition, there are also instance labels $y_1, y_2, \ldots, y_K$ which are all hidden even for the training data. The standard MIL rule on how to infer the bag label Y given its instance labels $y_1, y_2, \ldots, y_K$ can be expressed as follows:

$$Y = \begin{cases} 0, & \text{iff } \sum_k y_k = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

The MIL model we develop is trained by optimizing the log-likelihood (LL) function where the bag label is distributed according to a Bernoulli distribution $\theta(X) \in [0, 1]$, which depicts the probability $Y = 1$ given a bag X of instances. Also note that, since we assume bags of unordered and independent instances, the bag probability $\theta(X)$ must be permutation-invariant.

We pursue a three-step approach to predict bag labels, in which: (1) instances $\mathbf{x}_k$ are first transformed into a low-dimensional representation $\mathbf{z}_k = f_\psi(\mathbf{x}_k)$, (2) a combination of the transformed instances is formed via a permutation-invariant function (referred to as the MIL pooling), and (3) in order to form a bag representation, another transformation is applied over the combined instances, after which a classifier $\theta(X)$ is used for the resulting bag representation. We adopt a deep neural network to parameterize all the transformations. Thus, the whole model can be optimized in an end-to-end fashion via backpropagation.

2.2. Hybrid MIL

Joint distribution As mentioned earlier, we assume that instances within a bag X are identically and independently distributed. This assumption is crucial in our methodology. Further, we are interested in calculating the joint distribution over X and Y given the number of points in the bag X, $p(X, Y|K)$. Moreover, we consider the following generative model with shared latent variables:

$$\begin{aligned} p(X, Y|K) &= \int p(Y, Z, X|K)\, \mathrm{d}Z \qquad (2) \\ &= \int p(Y|Z) p(X|Z, K) p(Z|K)\, \mathrm{d}Z \qquad (3) \\ &\overset{iid}{=} \int p(Y|Z) \left(\prod_{k=1}^{K} p(\mathbf{x}_k|\mathbf{z}_k) p(\mathbf{z}_k) \right) \mathrm{d}Z, \qquad (4) \end{aligned}$$

where $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K\}$.

[1] We refer to the standard MIL case which assumes independence among instances within a bag. Nonetheless, there are a few works which study MIL settings where instances within a bag do not follow the IID assumption, e.g. (Zhou et al., 2009; Zhang, 2021)

Variational inference We parameterize these distributions using neural networks, thus, calculating the integral becomes analytically intractable. In order to overcome this issue, we propose to use variational inference which allows calculating the lower bound to the logarithm of the joint distribution (the ELBO). Considering the following family of variational posteriors $q_{\phi}(Z|X,K) = \prod_{k=1}^{K} q_{\phi}(\mathbf{z}_k|\mathbf{x}_k)$ yields:

$$\begin{aligned}\log p_{\vartheta}(X,Y|K) &= \log \int p_{\vartheta}(X,Y,Z|K) \frac{q_{\phi}(Z|X,K)}{q_{\phi}(Z|X,K)} \,\mathrm{d}Z && (5)\\ &\geq \mathbb{E}_{q_{\phi}(Z|X)} \left[\log p_{\vartheta}(Y|Z) + \sum_{k=1}^{K} \Big(\log p_{\vartheta}(\mathbf{x}_k|\mathbf{z}_k) + \log \frac{p_{\vartheta}(\mathbf{z}_k)}{q_{\phi}(\mathbf{z}_k|\mathbf{x}_k)} \Big) \right] && (6)\\ &\overset{df}{=} -\mathcal{L}(X,Y,K|\vartheta,\phi) && (7)\end{aligned}$$

Notice that in the ELBO we have a component for the classification of a bag, $\log p(Y|Z)$, and a sum of objectives for each object in the bag X that coincide with the formulation of Variational Auto-Encoders (Kingma and Welling, 2014; Rezende et al., 2014).

Semi-supervised learning Since the ELBO consists of a sum of two objectives, namely, one for the classifier and one for the marginal over objects, the proposed approach is well-suited for semi-supervised learning. Let us denote the part with X as follows:

$$\mathcal{U}(X,K|\vartheta,\phi) \overset{df}{=} -\mathbb{E}_{q_{\phi}(Z|X)} \left[\sum_{k=1}^{K} \Big(\log p_{\vartheta}(\mathbf{x}_k|\mathbf{z}_k) + \log p_{\vartheta}(\mathbf{z}_k) - \log q_{\phi}(\mathbf{z}_k|\mathbf{x}_k) \Big) \right]. \qquad (8)$$

For two given sources of data, namely, laballed data $(X,Y) \sim p_l(X,Y)$, and unlabelled data $X \sim p_u(X)$, we can formulate a joint learning objective by minimizing the combination of $\mathcal{L}(X,Y,K|\vartheta,\phi)$ and $\mathcal{U}(X,K|\vartheta,\phi)$. However, typically we have more unlabelled data, therefore we consider a weighted objective:

$$\mathcal{J}(\vartheta,\phi) = \alpha \cdot \sum_{(X,Y)\sim p_l} \mathcal{L}(X,Y,K|\vartheta,\phi) + \sum_{X\sim p_u} \mathcal{U}(X,K|\vartheta,\phi), \qquad (9)$$

where $\alpha > 0$. This approach is known as *hybrid modeling* (Lasserre et al., 2006).

Modeling $p(Y|Z)$ In this paper, we pursue an attention-based MIL pooling approach for modeling $p(Y|Z)$ due to several reasons: Attention-based MIL pooling is more flexible, adaptive, and more trainable than the max and mean pooling operators. It is also more interpretable due to the data-driven adjustment of instance weights according to the task and data at hand, which can potentially provide instance scores signifying the most relevant instances w.r.t. the bag label prediction. Attention-based pooling is depicted in the form of a weighted averaging with learnable parameters. To ensure invariance to the size of a bag, the weights are constrained to sum up to 1. Assuming a bag of K instance representation embeddings $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$, the MIL pooling is expressed as:

$$\mathbf{h} = \sum_{k=1}^{K} a_k \mathbf{z}_k, \qquad (10)$$

where:

$$a_k = \frac{\exp\{\mathbf{w}^\top \big(\tanh\left(\mathbf{V}\mathbf{z}_k^\top \right) \odot \mathrm{sigm}\big(\mathbf{U}\mathbf{z}_k^\top\big)\big)\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top \big(\tanh\left(\mathbf{V}\mathbf{z}_j^\top \right) \odot \mathrm{sigm}\big(\mathbf{U}\mathbf{z}_j^\top\big)\big)\}}, \tag{11}$$

where $\mathbf{w} \in \mathbb{R}^{L\times 1}$, $\mathbf{V} \in \mathbb{R}^{L\times M}$ and $\mathbf{U} \in \mathbb{R}^{L\times M}$ are parameters, and $\tanh(\cdot)$ is an element-wise hyperbolic tangent nonlinearity. Element-wise multiplication is depicted by $\odot$, and $\mathrm{sigm}(\cdot)$ refers to the sigmoid nonlinearity which grants the adoption of a gating mechanism, potentially avoiding some troublesome linearity issues associated with $\tanh(\cdot)$ (Ilse et al., 2018).

Eventually, the classifier works as follows:

1. X is transformed to Z through a shared stochastic encoder $q_\phi(Z|X, K)$, i.e., we calculate a sample $Z \sim q_\phi(Z|X, K)$.

2. An embedding $\mathbf{h}$ is calculated trough the attention-based MIL pooling operator (see Eq. 10) for given Z.

3. A neural network is used to calculate probabilities of class labels, $\theta(\mathbf{h})$.

3. Experiments

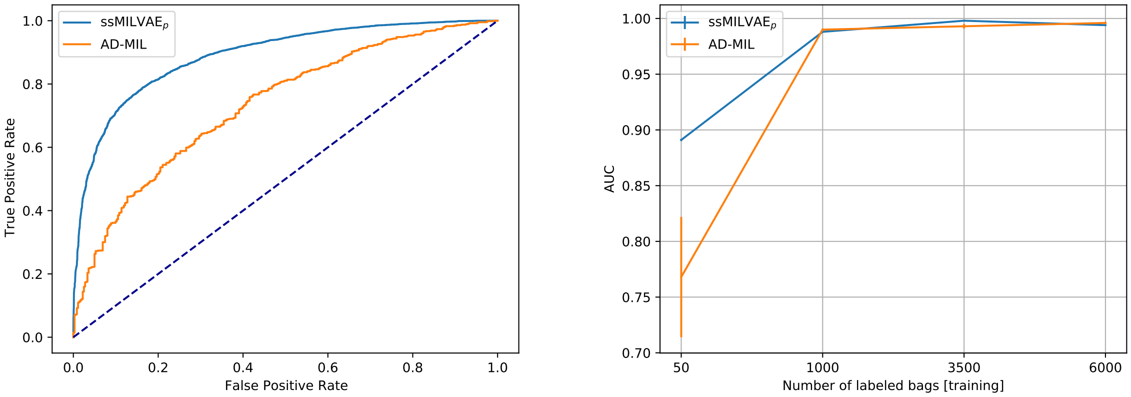
We quantitatively and qualitatively evaluate the proposed framework, which we refer to as semi-supervised multiple-instance learning variational autoencoder (ssMILVAE). The conducted experiments mainly address the following issues: (i) To assess the (accuracy) performance of the proposed ssMILVAE in the SSL paradigm, and (ii) to gauge the degree of interpretability granted by ssMILVAE and whether the learned instance weights can provide information on the contributions of each instance to the bag label prediction.

We assess ssMILVAE on two datasets, MNIST-BAGS which is an MNIST-based image dataset, and COLON CANCER which is a real-world histopathology dataset. We use 10-fold cross-validation and repeat each experiment five times. To compare on common ground, we follow most of the settings and modelling choices pursued by Ilse et al. (2018). We refer to the latter method here as AD-MIL. The MIL pooling layers are located right below the top layer of the model. We compare the bag level performance based on the area under the receiver operating characteristic curve (AUC). All the experiments were run for 100 epochs. We used Adam (Kingma and Ba, 2015) optimizer with values of β_1 and β_2 set equal to 0.9 and 0.999, respectively. Weights are initialized according to (He et al., 2015). The hyperparameter α (i.e., the weighting between the labelled objective and the unlabelled objective) was determined through model selection on the validation set. In the experiments, we use various number of labeled bags, while the rest of bags are treated as unlabeled. A detailed description of the experiments could be found in the Appendix A and B.

3.1. MNIST-BAGS

In this experiment, we sample images from the MNIST training (test) set to form training (test) bags, respectively. Each bag consists of a random number of 28 × 28 greyscale handwritten MNIST images. Number of images within a bag is Gaussian distributed where the closest integer value is the chosen bag size. Since the number ‘9’ can possibly be confused with ‘7’ and ‘4’, we rate a bag as positive if it contains at least one image of the digit ‘9’.

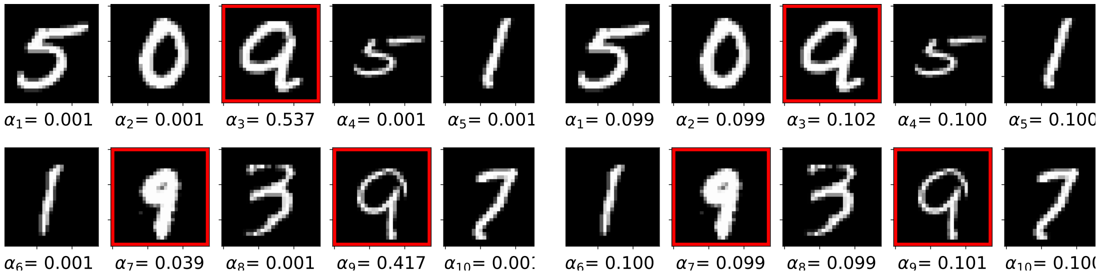
We present the results in terms of the True Positive Rate vs. False Positive Rate in Figure 2(a) for the case with only 10 labeled bags, and we compare the AUC in Figure 2(b) for various number of labeled data. The results demonstrate the supremacy of the proposed ssMILVAE when the learner encounters a small number of labeled bags. The performance of ssMILVAE is nearly equalled by AD-MIL with a larger number of labeled bags.



(a) **(b)**

Figure 2: A comparison between ssMILVAE and AD-MIL. **A:** The ROC curve results for a bag size of 10 instances on the MNIST-BAGS dataset. **B:** The bag AUC results for 10-instance bags on the MNIST-BAGS dataset.

We present the attention mechanism of the proposed ssMILVAE algorithm on the MNIST-BAGS dataset, and compare it with AD-MIL. The comparison is done based on a rather limited number of labeled bags, which is 50 bags. The bags displayed in Figure 5 are correctly classified by both algorithms and not cherry-picked. The proposed ssMILVAE is capable of assigning higher weights to the positive instances than AD-MIL. This suggests that ssMILVAE may provide more *interpretable* bag label predictions than AD-MIL, when trained on a limited number of labeled bags, since the instance weights convey the relevance of the respective instances for the bag labeling decision.



(a) ssMILVAE **(b) AD-MIL**

Figure 3: Evaluation of the attention mechanism of the proposed ssMILVAE compared to that of AD-MIL, tested on bags containing multiple positive (’9’) instances from MNIST-BAGS.

3.2. COLON CANCER

The COLON CANCER dataset consists of real-world histopathology data (Sirinukunwattana et al., 2016). The data contains cancerous regions in hematoxylin and eosin (H&E) stained whole-slide images. There are a total of 22,444 nuclei labeled as epithelial, inflammatory, fibroblast or miscellaneous. It consists of 100 H&E images originating from a variety of tissue appearances from healthy and malignant regions (Ilse et al., 2018). Each bag consists of 27 × 27 patches. A bag is labeled as positive if it contains at least one epithelial nuclei. Colon cancer clinically originates from epithelial cells, and this is why epithelial nuclei are very informative about the diagnosis here.

The detailed results can be found in the Appendix B. The AUC results are displayed in Figure 4. We experiment with the following number of labeled training bags: 22, 92 and 162. Interestingly, the proposed ssMILVAE is more accurate when trained on a small number of training bags. However, when the number of available labeled training bags increases, AD-MIL begins to outperform ssMILVAE. We hypothesize that the explanation lies in the fact that our hybrid model was not properly tuned and it focused too much on the generative part.

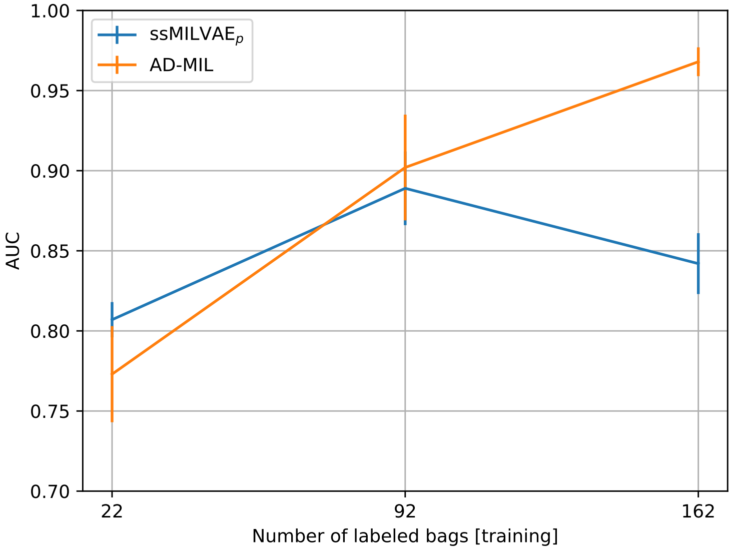
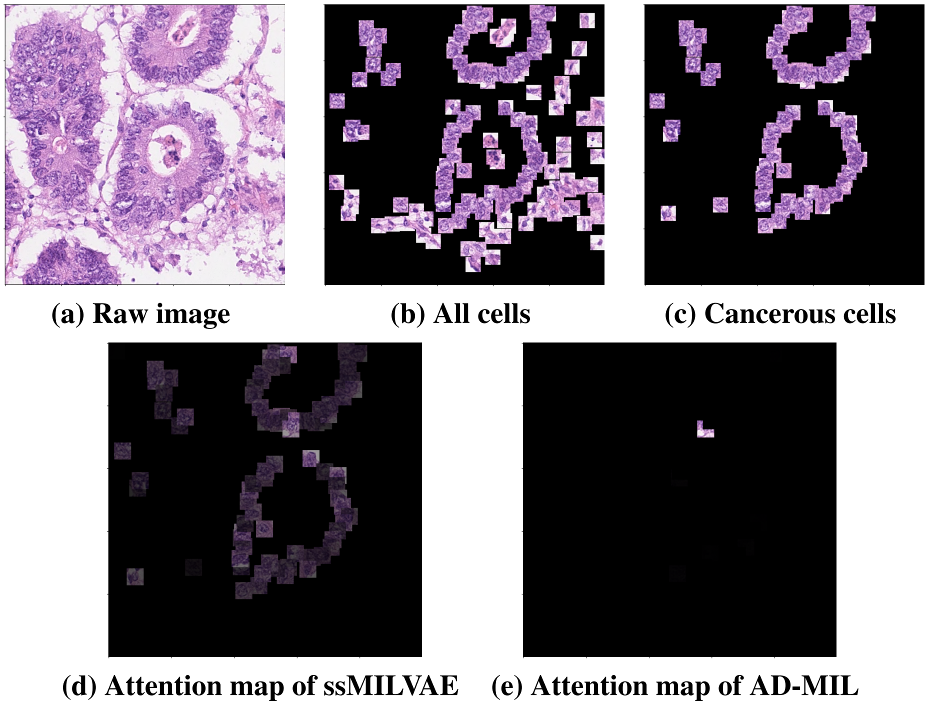


Figure 4: The bag AUC results on the COLON CANCER dataset for the proposed ssMILVAE and the baselineAD-MIL given a small number of labeled training bags.

Regarding the attention mechanism, we compare the proposed ssMILVAE with AD-MIL in terms of the resulting regions of interest (ROIs), which are of paramount importance in medical diagnosis. The raw histopathological image is displayed in Figure 5(a). The histopathological image is split into smaller tile containing single cells, Figure 5(b). An attention map is generated by multiplying cell images by their respective attention weights. The attention weights are then rescaled using $a' = \frac{a_k - min(a)}{max(a) - min(a)}$. As can be noticed in Figure 5(d), the proposed attention mechanism by ssMILVAE achieves a much better outcome in spotting the relevant cells compared to AD-MIL. As such, it seems that the attention mechanim trained within the proposed ssMILVAE framework provides more interpretable predictions by identifying the key patches responsible for the diagnosis.



(a) Raw image **(b) All cells** **(c) Cancerous cells**

(d) Attention map of ssMILVAE **(e) Attention map of AD-MIL**

Figure 5: (a) Raw image from the COLON CANCER. All cells are depicted in (b) and the cancerous cells are presented in (c). The attention maps for ssMILVAE are in (d) and for AD-MIL in (e).

4. Conclusion

In this paper, we have introduced an extension of the multiple-instance learning classification problem to learning a joint distribution in the semi-supervised setting. We have proposed a latent variable model for the multiple-instance learning generative model with a shared parameterization between the classifier and the unsupervised part. The resulting hybrid model allows joint training as well as switching between the discriminative and generative modes. In the experiments, we have shown that the proposed approach is beneficial in cases with a limited number of labeled data on both datasets, MNIST-BAGS and COLON CANCER. Moreover, we have indicated that the attention mechanism seems to benefit from being learnt within the proposed hybrid modeling framework.

In many applications, (especially in the medical domain), it is difficult to obtain huge sizes of labeled observations, and in such cases our proposed ssMILVAE seems to represent a recommended choice due to its ability to learn from limited numbers of labeled bags (medical cases). Moreover, the attention mechanism allows the model to assist a human expert (e.g. a physician) in interpreting results, which is of great importance in practice.

In this work, we assumed that instances are i.i.d. This assumption may be limiting or, in other words, we may learn better latent representations by introducing dependencies among instances. We believe that considering the non-i.i.d. assumption is an interesting future direction (Zhang, 2021).

References

Tameem Adel, Ruth Urner, Benn Smith, Daniel Stashuk, and Daniel J Lizotte. Generative multiple-instance learning models for quantitative electromyography. In *29th Conference on Uncertainty in Artificial Intelligence, UAI 2013*, pages 2–11, 2013.

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15, 2002.

O. Chapelle, B. Scholkopf, and A. Zien. *Semi-supervised learning*. MIT Press, 2006.

T. Dietterich, R. Lathrop, and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71, 1997.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on Imagenet classification. *IEEE International Conference on Computer Vision (ICCV)*, 2015.

M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. *International Conference on Machine Learning (ICML)*, 2018.

M. Ilse, J. Tomczak, C. Louizos, and M. Welling. DIVA: Domain Invariant Variational Autoencoders. *Medical Imaging with Deep Learning*, 2020a.

Maximilian Ilse, Jakub M Tomczak, and Max Welling. Deep multiple instance learning for digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, pages 521–546. Elsevier, 2020b.

Nathan Ing, Jakub M Tomczak, Eric Miller, Isla P Garraway, Max Welling, Beatrice S Knudsen, and Arkadiusz Gertych. A deep multiple instance model to predict prostate cancer metastasis from nuclear morphology. *MIDL*, 2018.

T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems (NIPS)*, 1999.

D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2015.

D. Kingma and M. Welling. Auto-encoding variational Bayes. *International Conference on Learning Representations (ICLR)*, 2014.

D. Kingma, D. Rezende, S. Mohamed, and M. Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems (NIPS)*, 28:3581–3589, 2014.

Julia A Lasserre, Christopher M Bishop, and Thomas P Minka. Principled hybrids of generative and discriminative models. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 87–94. IEEE, 2006.

Wu-Jun Li et al. Mild: Multiple-instance learning via disambiguation. *Ieee transactions on knowledge and data engineering*, 22(1):76–89, 2009.

H. Liu and P. Abbeel. Hybrid discriminative-generative training via contrastive learning. *arXiv preprint arXiv:2007.09070*, 2020.

L. Maaloe, M. Fraccaro, V. Lievin, and O. Winther. BIVA: A very deep hierarchy of latent variables for generative modeling. *Advances in neural information processing systems (NeurIPS)*, 2019.

Jaime Melendez, Bram van Ginneken, Pragnya Maduskar, Rick HHM Philipsen, Klaus Reither, Marianne Breuninger, Ifedayo MO Adetifa, Rahmatulai Maane, Helen Ayles, and Clara I Sánchez. A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays. *IEEE Transactions on Medical Imaging*, 34(1):179–192, 2014.

E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Hybrid models with deep and invertible features. *International Conference on Machine Learning (ICML)*, 2019.

G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard. Multiple-instance learning for medical image and video analysis. *IEEE Reviews in Biomedical Engineering*, 2017.

Rouhollah Rahmani and Sally A Goldman. Missl: Multiple-instance semi-supervised learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 705–712, 2006.

D. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *International Conference on Machine Learning (ICML)*, 31, 2014.

N. Siddharth, B. Paige, J. van den Meent, A. Demaison, N. Goodman, P. Kohli, F. Wood, and P. Torr. Learning disentangled representations with semi-supervised deep generative models. *Advances in neural information processing systems (NIPS)*, 2017.

K. Sirinukunwattana, S. Raza, Y. Tsang, D. Snead, I. Cree, and N. 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.

David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Jakub M Tomczak, Maximilian Ilse, Max Welling, Marnix Jansen, Helen G Coleman, Marit Lucas, Kikki de Laat, Martijn de Bruin, Henk Marquering, Myrtle J van der Wel, et al. Histopathological classification of precursor lesions of esophageal adenocarcinoma: A deep multiple instance learning approach. *MIDL*, 2018.

S. Tulyakov, A. Fitzgibbon, and S. Nowozin. Hybrid VAE: Improving deep generative models using partial observations. *arXiv preprint arXiv:1711.11566*, 2017.

Philippe Weitz, Yinxi Wang, Johan Hartman, and Mattias Rantalainen. An investigation of attention mechanisms in histopathology whole-slide-image analysis for regression objectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 611–619, 2021.

Qi Zhang and Sally Goldman. Em-dd: An improved multiple-instance learning technique. *Advances in neural information processing systems*, 14, 2001.

W. Zhang. Non-I.I.D. multi-instance learning for predicting instance and bag labels using variational auto-encoder. *arXiv preprint arXiv:2105.01276*, 2021.

Z. Zhou, Y. Sun, and Y. Li. Multi-instance learning by treating instances as non-I.I.D. samples. *International Conference on Machine Learning (ICML)*, 2009.

X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. *International Conference on Machine Learning (ICML)*, 2003.

Appendix A. The MNIST-BAGS

In this appendix, we provide further experimental details related to the MNIST-BAGS experiment. We begin by listing the hyperparameter values in Table 1, followed by details about the architecture of the VAE used in the MNIST-BAGS experiment in Table 2.

Table 1: The grid search parameters for MNIST dataset.

Hyperparamter	Values
Number of hidden layers	[3, 4, 5]
Number of hidden units	[32, 64, 128, 256, 512, 1024]
Latent dimensions	[16, 32, 64]
α	[1, 3, 5, 32, 50, 100]
Learning Rate	[1e-6, 1e-5, 3e-5, 1e-4, 3e-4, 1e-3]
Weight Decay	[1e-5, 1e-4, 1e-3, 0]

Table 2: The VAE Architecture for MNIST dataset.

Encoder		**Decoder**	
Layer	Type	Layer	Type
1	Conv(3, 2, 1)-32 + LeakyReLU(0.2)	1	ConvTranspose(1, 1, 0)-512 + LeakyReLU(0.2)
2	Conv(3, 2, 1)-64 + LeakyReLU(0.2)	2	ConvTranspose(4, 1, 0)-256 + LeakyReLU(0.2)
3	Conv(3, 2, 1)-128 + LeakyReLU(0.2)	3	ConvTranspose(4, 2, 0)-128 + LeakyReLU(0.2)
4	Conv(3, 2, 1)-256 + LeakyReLU(0.2)	4	ConvTranspose(4, 2, 0)-64 + LeakyReLU(0.2)
5	Conv(3, 2, 1)-512 + LeakyReLU(0.2)	5	ConvTranspose(4, 1, 0)-32 + LeakyReLU(0.2)
6	Conv(1, 1, 0)-32	6	ConvTranspose(4, 1, 0)-1 + Sigmoid

Appendix B. The COLON CANCER

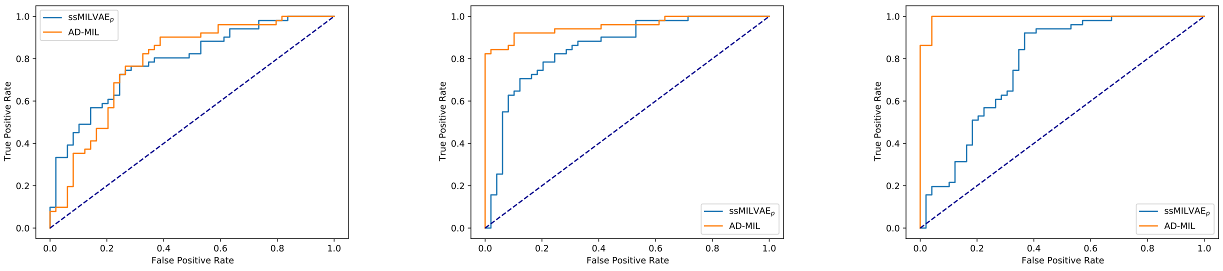
Values of the hyperparameters used in the experiment on the COLON CANCER dataset are listed in Table 3. The Structure of the VAE used with the COLON CANCER data is described in Table 4. Results (in terms of different performance metrics) on the COLON CANCER dataset are presented in Table 5, followed by the ROC curves for the following numbers of labeled bags: 22, 92 and 162 in Figure 6. Finally, the attention maps resulting from a COLON CANCER data image after applying training on 92 and 162 bags, are displayed in Figures 7 and 8, respectively. In both cases, similar to the case presented in the main text, the attention maps resulting from training the proposed framework ssMILVAE clearly outperform the corresponding attention maps obtained by the previous SOTA (AD-MIL).

Table 3: The grid search parameters for the COLON CANCER dataset.

Hyperparamter	Values
Number of hidden layers	[3, 4, 5]
Number of hidden units	[32, 64, 128, 256, 512, 1024]
Latent dimensions	[16, 32, 64]
α	[100, 1000, 10000]
Learning Rate	[1e-6, 1e-5, 3e-5, 1e-4, 3e-4, 1e-3]
Weight Decay	[1e-5, 1e-4, 1e-3, 0]

Table 4: The VAE Structure for the COLON CANCER dataset.

	Encoder		**Decoder**
Layer	Type	Layer	Type
1	Conv(5, 1, 0)-64 + LeakyReLU(0.2)	1	ConvTranspose(3, 1, 0)-128 + LeakyReLU(0.2)
2	MaxPool(2, 2)	2	Upsample(2)
3	Conv(4, 1, 0)-128 + LeakyReLU(0.2)	3	ConvTranspose(4, 1, 0)-64 + LeakyReLU(0.2)
4	MaxPool(2, 2)	4	Upsample(2)
5	Conv(3, 1, 0)-64	5	ConvTranspose(7, 1, 0)-100



(a)

(b)

(c)

Figure 6: The ROC results of (a) 22, (b) 92, and (c) 162 labeled bags from the COLON CANCER test bags under an experiment conducted for 5 times.

Table 5: The results on the COLON CANCER. Experiments were run 5 times and an average (± a standard error of the mean) is reported.

The number of labeled bags: **22**

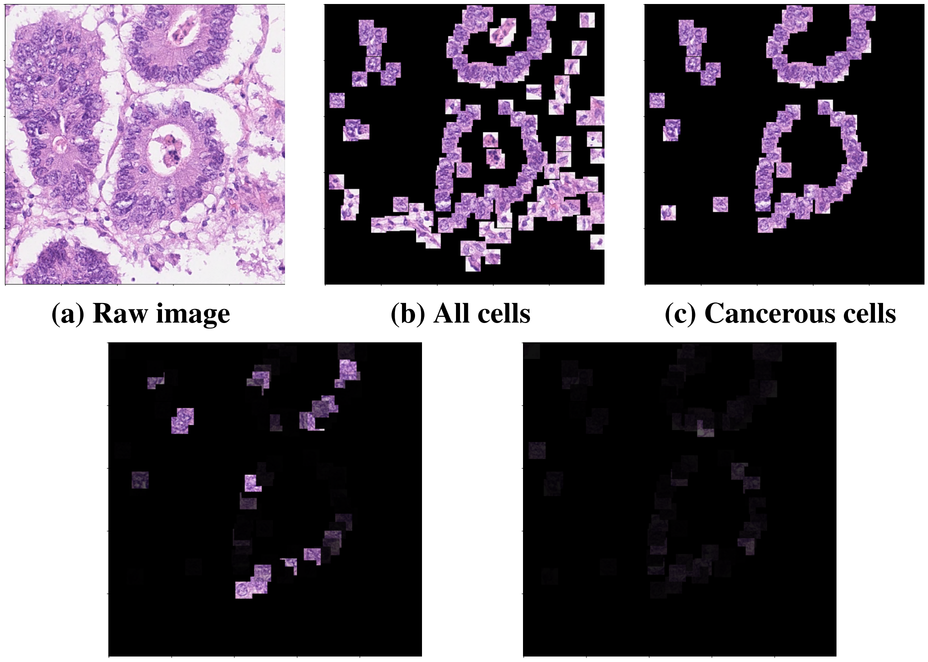
METHOD	ACCURACY	PRECISION	RECALL	F-SCORE	AUC
ssMILVAE	**0.727** ± 0.027	0.679 ± 0.027	**0.915** ± 0.013	**0.779** ± 0.021	**0.787** ± 0.027
AD-MIL	0.663 ± 0.024	**0.790** ± 0.065	0.583 ± 0.115	0.611 ± 0.061	0.773 ± 0.030

The number of labeled bags: **92**

METHOD	ACCURACY	PRECISION	RECALL	F-SCORE	AUC
ssMILVAE	**0.806** ± 0.019	0.859 ± 0.024	0.757 ± 0.013	0.804 ± 0.015	0.889 ± 0.023
AD-MIL	0.805 ± 0.032	**0.872** ± 0.059	**0.759** ± 0.024	**0.805** ± 0.028	**0.902** ± 0.033

The number of labeled bags: **162**

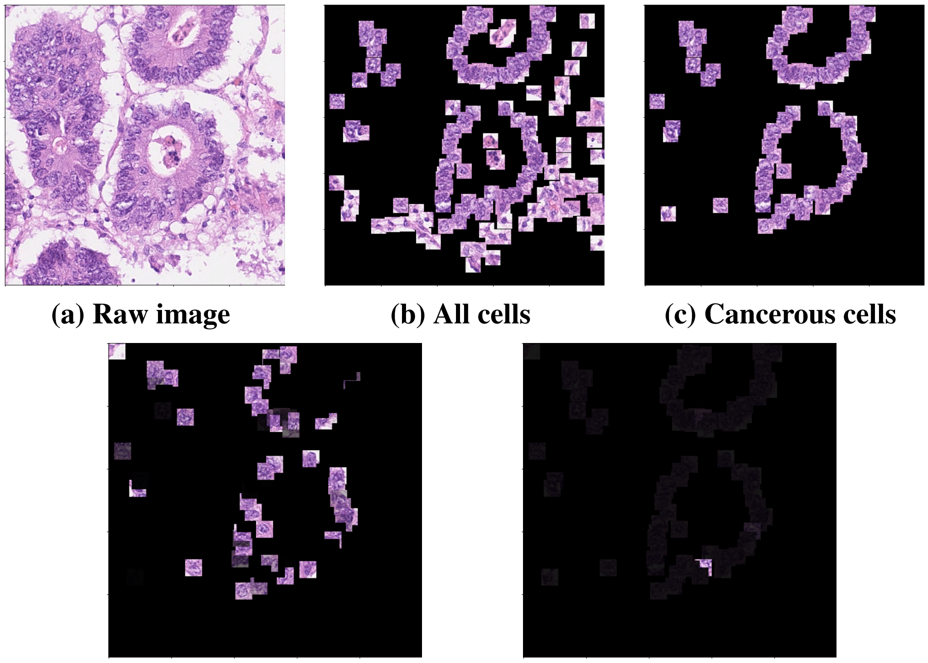
METHOD	ACCURACY	PRECISION	RECALL	F-SCORE	AUC
ssMILVAE	0.775 ± 0.010	0.743 ± 0.018	**0.870** ± 0.016	0.801 ± 0.011	0.842 ± 0.019
AD-MIL	**0.904** ± 0.011	**0.953** ± 0.014	0.855 ± 0.017	**0.901** ± 0.011	**0.968** ± 0.009



(a) Raw image **(b) All cells** **(c) Cancerous cells**

(d) Attention map of ssMILVAE **(e) Attention map of AD-MIL**

Figure 7: (a) Raw image from the COLON CANCER. All cells are depicted in (b) and the cancerous cells are presented in (c). The attention maps for ssMILVAE are in (d) and for AD-MIL in (e). This example is for a model trained with 92 labeled bags.



(a) Raw image

(b) All cells

(c) Cancerous cells

(d) Attention map of ssMILVAE

(e) Attention map of AD-MIL

Figure 8: (a) Raw image from the COLON CANCER. All cells are depicted in (b) and the cancerous cells are presented in (c). The attention maps for ssMILVAE are in (d) and for AD-MIL in (e). This example is for a model trained with 162 labeled bags.