Multi-instance Learning with Deep AUC Maximizaiton on Medical Classification Tasks

Xinwen Zhang

Abstract

This paper explores a novel application of Deep AUC maximization (DAM) within the context of multi-instance learning (MIL), focusing on scenarios where a singular class label is assigned to a collection of instances. An inherent computational challenge arises in MIL when applying DAM, specifically when the bag size exceeds the capacity of GPU memory during backpropagation—a necessity for standard pooling methods in MIL. To overcome this computational challenge, this study proposes the utilization of stochastic pooling methods inspired by stochastic optimization. This approach involves the strategic sampling of a limited number of instances from each bag for computing prediction scores and subsequent model parameter updates. To validate the proposed methodology, a comparative analysis is conducted between AUC maximization and the conventional cross-entropy score across various medical classification tasks. The experimentation reveals the superior efficacy of the AUC score, particularly when confronted with imbalanced datasets. This research not only sheds light on the neglected yet crucial computational challenge in MIL within the DAM framework but also underscores the advantages of leveraging AUC as a robust evaluation metric in medical classification tasks.

1 Introduction

Within the expansive landscape of machine learning, weakly supervised learning stands as a distinctive paradigm, steering away from the conventional supervised learning framework by incorporating a nuanced layer of uncertainty and incompleteness into the labeling process. A significant facet of weakly supervised learning is Multi-Instance Learning (MIL)(Dietterich et al., 1997), specifically tailored for tasks where training data is organized into bags, each comprising multiple instances. Remarkably, only the label of the bag is known in MIL, leaving the individual instances within the bag unlabeled.

Initially conceptualized for drug activity prediction within medical classification tasks, MIL innovatively modeled molecules as bags(Dietterich et al., 1997), and the conformations of the molecules as instances, with only bag-level labels known. Since its inception, MIL has traversed diverse applications. For instance, in text categorization (Andrews et al., 2002), (Ji et al., 2020), articles are treated as bags of sentences with solely article-level labels, and in image classification(Oquab et al., 2015), Yao et al. (2020), images are segmented into bags of patches with only image-level labels.

The application of Multi-Instance Learning has been particularly impactful in the field of medical task classification. For example, a **bag** can be conceptualized as representing a medical image, such as an X-ray or MRI scan, where the **instances** within the bag correspond to distinct regions of interest within the image. These regions of interest may hold critical diagnostic information, but their precise location or extent of abnormality might be uncertain or variable across different

patients. The bag, in this case, is labeled to indicate the presence or absence of a specific medical condition.

However, when confronted with large bags containing multiple instances, the computational challenge from the multi instance learning arises from the practical limitations of available resources, such as the constrained memory size of GPUs. This constraint may hinder the simultaneous loading of all instances within a bag during training, leading to a significant computational bottleneck. This scenario is exacerbated when dealing with medical tasks.

A potential approach to circumvent this computational hurdle involves the adoption of a mini-batch stochastic pooling strategy. This entails the selective sampling of a subset of instances from a given bag for computing the pooled prediction and subsequently conducting the necessary parameter updates. This approach provides a pragmatic solution to the memory constraints imposed by large bags.

Recently, Deep AUC Maximization (DAM) (Liu et al., 2019) has demonstrated remarkable success across various AI applications by adeptly handling imbalanced data in traditional supervised learning settings. The integration of AUC into multi-instance learning for medical task classification introduces an intriguing dimension, elevating the complexity of the multi-instance learning paradigm.

Hence, for this project, an exploration into whether AUC scores can also excel in the context of multi-instance learning is warranted. The main tasks for this endeavor include:

- Clarify Stochastic Pooling Methods: Rigorously elucidate stochastic pooling methods to strategically reduce the computation budget, addressing the challenges posed by large bags in the context of MIL.
- Optimize DAM and Minimax Framework: Enhance the DAM optimization process to facilitate parameter updates, seamlessly integrate it into the minimax framework, and provide clarity on the chosen optimizer.
- Conduct Experiments Across Diverse Scenarios: Implement experiments across various datasets and scenarios, presenting results that substantiate the superiority of Deep AUC Maximization in multi-instance learning. This involves showcasing its efficacy in handling imbalanced data and highlighting performance improvements compared to conventional methods.

2 Related Work

2.1 Multi-instance learning

Multiple-instance learning (MIL) has been instrumental in shaping the landscape of machine learning, showcasing its adaptability across diverse paradigms. n the arena of conventional learning applied to tabular data, various strategies (Babenko, 2008), (Carbonneau et al., 2018) have been advanced. Concurrently, the rise of deep learning has ushered in inventive methodologies specifically tailored for unstructured data (Oquab et al., 2015), (Qi et al., 2017). Groundbreaking contributions underscore the significant headway in effectively leveraging deep learning techniques to tackle the intricacies inherent in unstructured information.

Another foundational concept in MIL is the prediction function(Zaheer et al., 2017). The process of classifying a bag of instances involves splitting it into individual instances, pooling these transformed instances using a symmetric function, and further aggregating the pooled representation. The key lies in selecting the symmetric function, commonly known as the pooling operation, which takes the transformations of all instances as input and produces a cohesive output. Various pooling strategies have been explored for these steps, encompassing traditional techniques like max

pooling, average pooling, and smoothed-max pooling of predictions (Ramon et al., 2000). Moreover, recent advancements incorporate attention-based pooling of feature representations(Ilse et al., 2018).

2.2 Deep AUC Maximization

Recently, there has been a surge in research on DAM, particularly in the context of imbalanced datasets. (Ying et al., 2016) proposed a significant advancement in optimizing AUC. Their work revolves around optimizing pairwise square loss and introduces an equivalent min-max formulation. This decomposable reformulation allows for the development of efficient stochastic methods based on mini-batches of data without the need for explicit pair construction. The min-max formulation laid the groundwork for subsequent advancements in DAM, notably in works by ((Liu et al., 2019);(Yuan et al., 2020)). (Liu et al., 2019) is the first work that explicitly considers DAM optimization and pioneers practical stochastic algorithms for DAM based on the minimax formulation but limits experiments to basic datasets. Later, (Yuan et al., 2020) further introduced a novable robust loss in the minimax form for DAM, showcasing DAM's success in various medical image classification tasks.

3 Methodology

The proposed algorithm extends the principles of Multi-Instance Learning with stochastic pooling operations. Incorporating Deep AUC Maximization, emphasis is placed on accurate classification in the context of imbalanced datasets. The design utilizes bag-level labels and instance-level information to refine the learning process, enhancing the model's capability to handle uncertainties associated with classification tasks.

3.1 Stochastic Pooling Methods

Large bags with multiple instances pose a computational challenge in multi-instance learning, especially given resource constraints like limited GPU memory. Loading all instances at once during training can be hindered, creating a significant bottleneck. To address this, a potential solution involves adopting a mini-batch stochastic pooling strategy. This approach selectively samples a subset of instances from a bag to compute pooled predictions, facilitating necessary parameter updates.

Consider $\mathcal{X}_i = \{x_i^1, \cdots, x_i^{n_i}\}$ as a bag of data instances, $\mathcal{D} = \{(\mathcal{X}_i, y_i), i = 1, \cdots, n\}$ represent the set of labeled data. Difine $h(w; \mathcal{X}_i) \in [0, 1]$ as the pooled prediction score for bag i over all its instances:

$$h(\mathcal{X}) = g(\sum_{x \in \mathcal{X}} \phi(x)) \tag{1}$$

For stochastic pooling operation, we have following definition

Assumption 3.1. *The mini-batch mean pooling can be computed as:*

$$h(w; \mathcal{B}_i) = \max_{x \in \mathcal{B}_i} \phi(w; x) \tag{2}$$

where $\mathcal{B}_i \in \mathcal{X}_i$ only holds part of instances randomly sampled from the bag of all instances.

Assumption 3.2. *The mini-batch max pooling can be computed as:*

$$h(w;\mathcal{B}_i) = \frac{1}{|\mathcal{B}_i|}\sum_{x\in\mathcal{B}_i}\phi(w;x)) \tag{3}$$

where $\mathcal{B}_i \in \mathcal{X}_i$ only holds part of instances randomly sampled from the bag of all instances.

Assumption 3.3. *The mini-batch smoothed-max pooling can be computed as:*

$$h(w;\mathcal{B}_i) = \tau log\Big(\frac{1}{|\mathcal{B}_i|}exp(\phi(w;x)/\tau)\Big) \tag{4}$$

where $\mathcal{B}_i \in \mathcal{X}_i$ only holds part of instances randomly sampled from the bag of all instances.

Assumption 3.4. *The mini-batch attention pooling can be computed as:*

$$h(w;\mathcal{B}_i) = \sigma\Big(\sum_{x\in\mathcal{B}_i}\frac{exp(g(w;x))\delta(w;x)}{\sum_{x'\in\mathcal{B}_i}exp(g(w;x'))}\Big) \tag{5}$$

where $\mathcal{B}_i \in \mathcal{X}_i$ only holds part of instances randomly sampled from the bag of all instances.

3.2 Deep AUC Maximization

Training classifiers through AUC optimization ((Hanley and McNeil, 1982)) proves effective for managing highly imbalanced datasets. Yet, conventional AUC maximization models often rely on pairwise sample input, restricting their applicability to large-scale data. Recently, (Ying et al., 2016) innovatively formulated the AUC maximization model as a minimax optimization problem, defined as follows:

$$\begin{aligned}\min_{w,a,b}\max_{\alpha}\mathcal{L}_{AUC}(w,a,b,\alpha;\mathbf{z}) \triangleq (1-p)(h_w(x)-a)^2\mathbb{I}_{[y=1]} - p(1-p)\alpha^2 \\ + p(h_w(x)-b)^2\mathbb{I}_{[y=-1]} + 2(1+\alpha)(ph_w(x)\mathbb{I}_{[y=-1]} - (1-p)h_w(x)\mathbb{I}_{[y=1]}) \ ,\end{aligned} \tag{6}$$

where h represents the classifier parameterized by $\mathbf{w} \in \mathbb{R}^d$, while $a \in \mathbb{R}, b \in \mathbb{R}, \alpha \in \mathbb{R}$ serve as parameters for measuring the AUC score, $z = (x, y)$ denotes the sample's feature and label, p substitutes the prior probability of the positive class, and $\mathbb{I}$ is an indicator function that evaluates to 1 if the argument is true and 0 otherwise. This minimax objective function effectively decouples the dependence of pairwise samples, enabling its application to large-scale datasets.

Algorithm 1 PESG

Require: v_0, α_0, $\eta_v \in (0,1)$, $\eta_\alpha \in (0,1)$, $\gamma > 0$, $\lambda > 0$.
1: **for** $t = 0, \cdots, T-1$ **do**
2: Compute $\nabla_v F(v_t, \alpha_t; z_t)$ and $\nabla_\alpha F(v_t, \alpha_t; z_t)$
3: Update primal variable: $v_{t+1} = v_t - \eta_v(\nabla_v F(v_t, \alpha_t; z_t) + \gamma(v_t - v_{ref})) - \lambda\eta_v v_t$
4: Update dual variable: $\alpha_{t+1} = [\alpha_t + \eta_\alpha \nabla_\alpha F(v_t, \alpha_t; z_t)]_+$
5: **end for**

This objective function with the minimax formulation decomposes over individual examples, enabling the development of efficient primal-dual stochastic algorithms to update the model parameter w without the necessity of explicitly constructing positive-negative pairs.

Therefore, the proximal epoch stochastic method, denoted as PESG is proposed. Here, $v = (w, a, b)$ is used to represent all primal variables. Stochastic gradient descent methods are applied

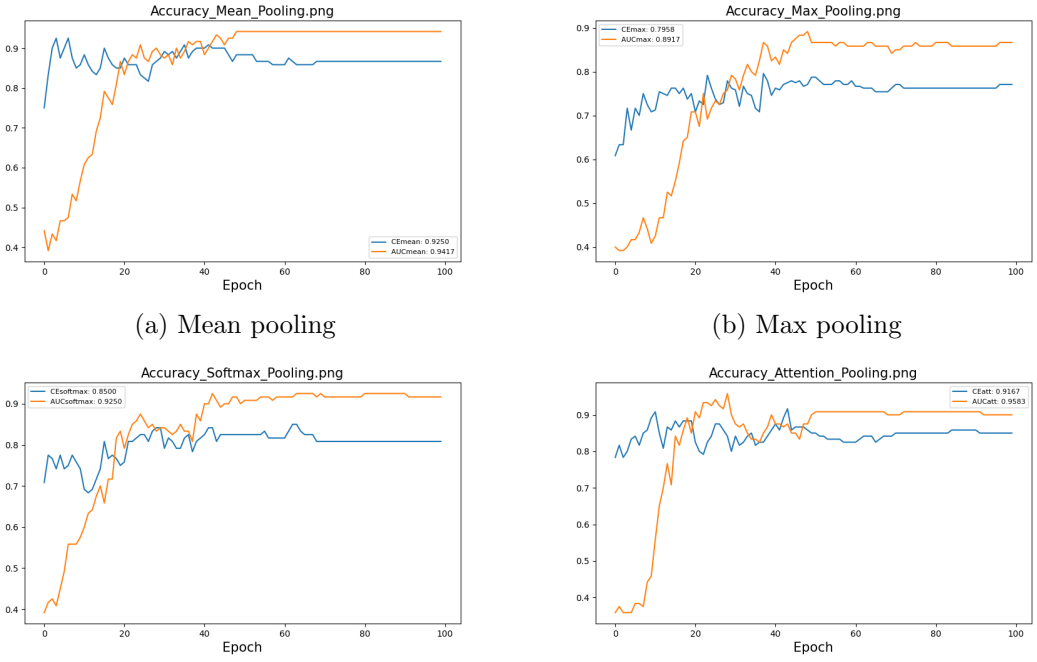
to update primal variables, with λ as the standard regularization parameter, γ as an algorithmic regularization parameter enhancing generalization, and v_{ref}as a periodically updated reference solution, computed using the accumulated average of v_t from the preceding stage before decaying the learning rate. Additionally, for the update on the dual variable α, a projection step is enforced after the gradient ascent.

4 Experiment

This section concentrates on medical experimental Multiple-Instance Learning (MIL) tasks. Experiments are conducted on three medical datasets tailored for MIL tasks, and the dataset details are outlined in the Table 1.

Dataset	Pos	Neg	Avg Bag Size	Features
MUSK1	47	45	5.17	166
MUSK2	39	63	64.69	166
Breast Cancer	26	32	672	32×32× 3

Table 1: Dataset Detail



(a) Mean pooling

(b) Max pooling

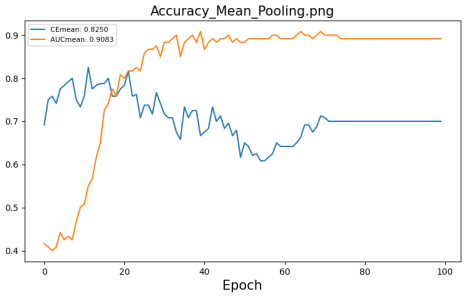
(c) Softmax pooling

(d) Attention pooling

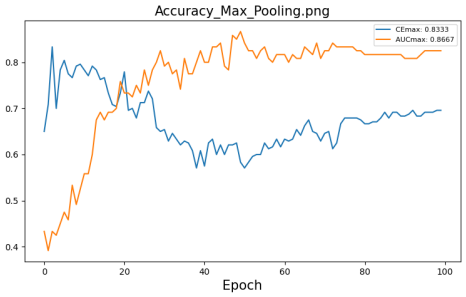
Figure 1: The test auc score versus the number of training epoch on MUSK1 Dataset.

Concerning the MUSK1 and MUSK2 datasets, they encompass drug molecules either binding strongly or not binding to a target protein. Each molecule, functioning as a bag, can assume various shapes or conformations, representing instances. A positive molecule possesses at least one shape conducive to strong binding though the specific shape is unknown, while a negative molecule lacks any shapes supporting effective binding.(Dietterich et al., 1997)

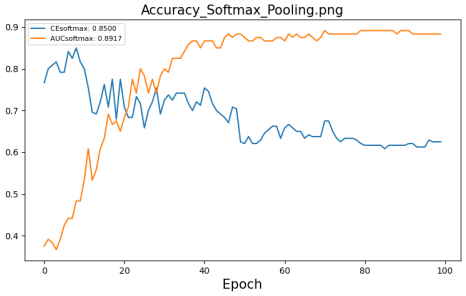
For these two datasets, a simple 2-layer feed-forward neural network (FFNN) serves as the foundational model, with the neuron count matching the data dimension, which is 166 in this case. The activation function for the middle layer is tanh, and sigmoid is employed as a normalization function for predicting scores in computing the AUC loss function. Data is uniformly and randomly split into training and testing sets with a ratio of 0.9/0.1, and 5-fold cross-validation experiments are conducted to avoid overfitting. Here, use the traditional cross-entropy loss as the baselines. The initial learning rate is 1e-1, and it decreases by 10 at the end of the 50th and 75th epochs within the 100-epoch training period. For all experiments in this study, the weight decay remains fixed at $1e^{-4}$. Each iteration involves sampling 8 positive bags and 8 negative bags. The testing AUC score is then plotted against the number of epochs. From Figures 1 and 2, it is evident that the application of Deep AUC Maxmization on Multiple-Instance Learning (MIL) tasks can outperform traditional cross-entropy loss with a large margin.



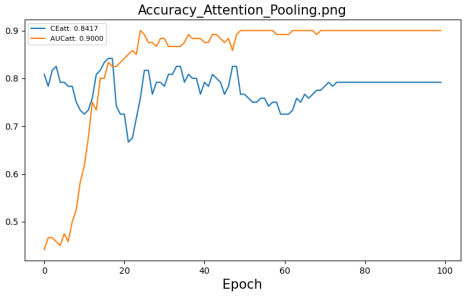
(a) Mean pooling



(b) Max pooling



(c) Softmax pooling



(d) Attention pooling

Figure 2: The test auc score versus the number of training epoch on MUSK2 Dataset.

The Breast Cancer dataset consists of microscopic images used for examining tissues in cancer diagnosis. The images are high-resolution, making the analysis of the entire picture challenging. To overcome this, the 896×768 images are partitioned into 32×32 patches, allowing for multi-instance learning. Employing ResNet20 as the backbone model, Figure 3 displays test auc score with stochasitic max pooling. The application of Deep AUC Maximization on this medical image dataset yields improved performance, addressing resolution challenges and highlighting its efficacy in

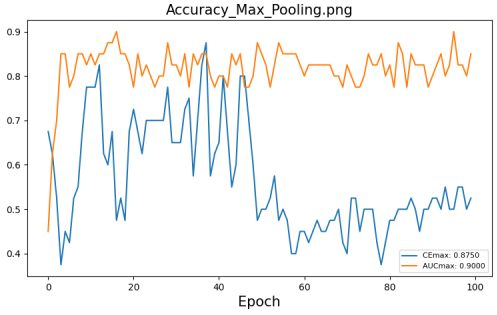


Figure 3: The test auc score on Breast Cancer

enhancing accuracy for cancer diagnosis. The utilization of multi-instance learning, coupled with deep AUC maximization, emerges as a promising strategy in advancing medical image classification.

5 Conclution

This report outlines a novel approach to medical classification tasks by combining stochastic Multi-Instance Learning with the Deep AUC Maximization. The algorithm's design and experimentation showcase promising results, highlighting its potential for improving model robustness in the challenging context of medical tasks analysis. Further research and refinement of the proposed algorithm could contribute significantly to the evolving landscape of machine learning applications in healthcare.

6 References

S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15, 2002.

B. Babenko. Multiple instance learning: algorithms and applications. *View Article PubMed/NCBI Google Scholar*, 19, 2008.

M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.

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

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

Y. Ji, H. Liu, B. He, X. Xiao, H. Wu, and Y. 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)*, pages 7012–7023, 2020.

M. Liu, Z. Yuan, Y. Ying, and T. Yang. Stochastic auc maximization with deep neural networks. *arXiv preprint arXiv:1908.10831*, 2019.

M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 685–694, 2015.

C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

J. Ramon, L. De Raedt, and S. Kramer. Multi instance neural networks. In *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*, pages 53–60, 2000.

J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

Y. Ying, L. Wen, and S. Lyu. Stochastic online auc maximization. *Advances in neural information processing systems*, 29, 2016.

Z. Yuan, Y. Yan, M. Sonka, and T. Yang. Robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. *arXiv preprint arXiv:2012.03173*, 8, 2020.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.