

Smooth Attention for Deep Multiple Instance Learning: Application to CT Intracranial Hemorrhage Detection

Yunan Wu¹, Francisco M. Castro-Macías^{2,4}, Pablo Morales-Álvarez^{3,4}, Rafael Molina², and Aggelos K. Katsaggelos¹

¹ Image and Video Processing Laboratory, Department of Electrical and Computer Engineering, Northwestern University, USA

² Department of Computer Science and Artificial Intelligence, University of Granada, Spain

³ Department of Statistics and Operations Research, University of Granada, Spain

⁴ Research Centre for Information and Communication Technologies (CITIC-UGR), University of Granada, Spain

Abstract. Multiple Instance Learning (MIL) has been widely applied to medical imaging diagnosis, where bag labels are known and instance labels inside bags are unknown. Traditional MIL assumes that instances in each bag are independent samples from a given distribution. However, instances are often spatially or sequentially ordered, and one would expect similar diagnostic importance for neighboring instances. To address this, in this study, we propose a smooth attention deep MIL (SA-DMIL) model. Smoothness is achieved by the introduction of first and second order constraints on the latent function encoding the attention paid to each instance in a bag. The method is applied to the detection of intracranial hemorrhage (ICH) on head CT scans. The results show that this novel SA-DMIL: (a) achieves better performance than the non-smooth attention MIL at both scan (bag) and slice (instance) levels; (b) learns spatial dependencies between slices; and (c) outperforms current state-of-the-art MIL methods on the same ICH test set.

Keywords: Smooth attention, Multiple instance learning, CT hemorrhage diagnosis

1 Introduction

Multiple Instance Learning (MIL) [21,6] is a type of weakly supervised learning that has become very popular in biomedical imaging diagnostics due to the reduced annotation effort it requires [8,13]. In the case of MIL binary classification, the training set is partitioned into bags of instances. Both bags and instances have labels, but only bag labels are observed while instance labels remain unknown. It is assumed that a bag label is positive if and only if the bag contains at least one positive instance [10]. The goal is to produce a method that, trained on bag labels only, is capable of predicting both bag and instance labels.

Among the proposed approaches for learning in the MIL scenario [21], deep learning (DL) methods stand out when dealing with highly structured data (such as medical images and videos) [17]. The most successful deep MIL approaches combine an instance-level processing mechanism (i.e., a feature extractor) with a pooling mechanism to aggregate information from instances in a bag [8,13]. Among the pooling operators, the attention-based weight pooling proposed in [15] is frequently used as a way to discover *key instances*, i.e., those responsible for the label of a bag. However, this pooling operator was formulated under strong assumptions of independence between the instances in a bag. This is a drawback in biomedical imaging problems, where instances in a bag are often spatially or sequentially ordered and their diagnostic importance is expected to be similar for neighboring instances [24,18].

In this work, we are particularly interested in the detection of intracranial hemorrhage (ICH), a serious life-threatening emergency caused by blood leakage inside the brain [5,22]. Radiologists confirm the presence of ICH by using computed tomography (CT) scans [9], which consist of a significant number of slices, each representing a section of the head at a given height. Unfortunately, the shortage of specialized radiologists and their increasing workload sometimes lead to delayed and erroneous diagnoses [3,12,25,20], which may result in potentially preventable cerebral injury or morbidity [11,9]. For this reason, there is a growing interest in the development of automated systems to assist radiologists in making rapid and reliable diagnoses.

State-of-the-art ICH detection methods rely on DL models, specifically convolutional neural networks (CNNs), to extract meaningful ICH features [31]. However, 2D CNNs need to be coupled with other mechanisms such as recurrent neural networks (RNNs) [30,14] or 3D CNNs [7,16,27,2] to account for inter-slice dependencies. Although these approaches are quite successful in terms of performance, their use is limited by the large amount of labeled data they require [31]. To address this issue, the ICH detection task has been formulated as an MIL problem, achieving comparable performance to fully supervised models while reducing the workload of radiologists [29,26]. Note that the MIL framework is naturally suited for the ICH detection problem since a CT scan (i.e., a bag) is considered positive if it contains at least one slice (i.e., an instance) with evidence of hemorrhage (i.e., positive instance).

In this work, we improve upon the state-of-the-art deep MIL methods by introducing dependencies between instances in a sound probabilistic manner. These dependencies are formulated over a neighborhood graph to impose smoothness on the latent function that encodes the attention given to each instance. Smoothness is achieved by introducing specific first- and second-order constraints on the latent function. Our model, called SA-DMIL, is applied to the ICH detection problem, obtaining (a) significant improvements upon the performance of non-smooth models at both scan and slice levels, (b) smoother attention weights across slices by benefiting from the inter-slice dependencies, and (c) a superior performance against other popular MIL methods on the same test set.

2 Methods

2.1 Problem formulation

We start by formulating ICH detection as a Multiple Instance Learning (MIL) problem. To do so, we map slices to instances and CT scans to bags. The slices (instances) will be denoted by $\mathbf{x}_i^b \in \mathbb{R}^{3HW}$, where H and W are the height and width of the image, 3 is the number of color channels, b is the index of the scan to which the slice belongs to and i is the index of the slice inside the bag. We will denote the label of a slice by $y_i^b \in \{0, 1\}$. If the slice contains hemorrhage, then $y_i^b = 1$, otherwise $y_i^b = 0$. Note that the slice labels remain unknown since only scan labels are given. As we know, slices are grouped to form the CT scans. Each scan (bag) will be denoted by $\mathbf{X}^b = [\mathbf{x}_1^b, \dots, \mathbf{x}_{N_b}^b]^\top \in \mathbb{R}^{N_b \times 3HW}$. Here, N_b is the number of slices in bag b . We will assume that B CT scans are given, so $b \in \{1, \dots, B\}$. Given a CT scan b , we will denote its label by $T^b \in \{0, 1\}$. Notice that $T^b = 1$ if and only if some of $y_i^b = 1$, i.e., the following relationship between scan and slice labels holds,

$$T^b = \max \{y_1^b, \dots, y_{N_b}^b\}. \quad (1)$$

2.2 Attention-based Multiple Instance Learning pooling

The attention-based MIL pooling was proposed in [15] as a way to discover *key instances*, i.e., those responsible for the diagnosis of a scan. It consists of a weighted average of instances (low-dimensional embeddings) where the weights are parameterized by a neural network. Formally, given a bag of N_b embeddings $\mathbf{Z}^b = [\mathbf{z}_1^b, \dots, \mathbf{z}_{N_b}^b]^\top$, where $\mathbf{z}_i^b \in \mathbb{R}^D$, the attention-based MIL pooling computes

$$\Phi_{\text{Att}}(\mathbf{Z}^b) = \sum_{i=1}^{N_b} s(\mathbf{z}_i^b) \mathbf{z}_i^b, \quad (2)$$

where

$$s(\mathbf{z}_i^b) = \frac{\exp(f(\mathbf{z}_i^b))}{\sum_j^{N_b} \exp(f(\mathbf{z}_j^b))}, \quad f(\mathbf{z}_i^b) = \mathbf{w}^\top \tanh(\mathbf{V} \mathbf{z}_i^b). \quad (3)$$

Notice that $\mathbf{w} \in \mathbb{R}^L$ and $\mathbf{V} \in \mathbb{R}^{L \times D}$ are trainable parameters, where D denotes the size of feature vectors. We refer to $s(\mathbf{z}_i^b)$ as *attention weights* and to $f(\mathbf{z}_i^b)$ as *attention values*.

This operator was proposed under the assumption that the instances in a bag show neither dependency nor order among each other. Although this may be the case in simple problems, it does not occur in problems such as ICH detection. Note that the attention weights of slices in a bag are correlated: given a slice containing ICH, we expect that the adjacent slices will also contain ICH with high probabilities. This is essential in finding slices with ICH. In the next subsection, we show how to introduce this correlation between attention weights.

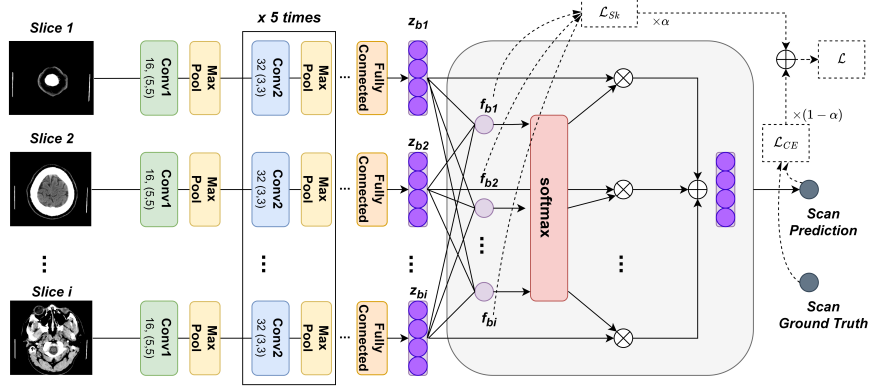


Fig. 1: SA-DMIL architecture. It consists of CNNs that extract slice level features and an attention block to aggregate slice features. The loss function is a weighted average of the binary cross entropy and a novel smooth attention loss.

2.3 Modeling correlation through the attention mechanism

Ideally, in the case of a positive scan ($T^b = 1$), high attention weights should be assigned to slices that are likely to have a positive label ($y_i^b = 1$). Given the dependency between slices, contiguous slices should have similar attention values. In other words, the differences between the attention values of contiguous slices should be *small*. Thus, for each bag b , these quantities should be small

$$\mathcal{L}_{S1}^b = 2^{-1} \sum_{i,j \in \text{Bag}(b)} A_{ij}^b (f(\mathbf{z}_i^b) - f(\mathbf{z}_j^b))^2, \quad (4)$$

$$\mathcal{L}_{S2}^b = 4^{-1} \sum_{i \in \text{Bag}(b)} \left(\sum_{j \in \text{Bag}(b)} A_{ij}^b (f(\mathbf{z}_i^b) - f(\mathbf{z}_j^b)) \right)^2, \quad (5)$$

where $A_{ij}^b = 1$ if the slices i, j are related in bag b , and 0 otherwise. We smooth $f(\mathbf{z}_i^b)$ instead of $s(\mathbf{z}_i^b)$ because a non-constrained parameter f ensures consistent smoothing while s requires a normalization across instances in a bag.

Equations (4) and (5) correspond, respectively, to the energies of the, so called, conditional and simultaneous autoregressive models in the statistics literature [23,4]. For our problem, they model the value of f at a given location (instance) given the values at neighboring instances. From the regularization viewpoint, these terms constrain the first and second derivatives of the function f , respectively, which favors smoother functions (examine the zero of the derivative of f). That is, a *priori* all attention weights are expected to be the same because f is expected to be constant. As observations arrive, they change to reflect the importance of each instance. Note that (4) and (5) impose smoothness but they can be modified to model, for example, competition between the attention weights by simply replacing the minus sign with a plus sign.

To compute \mathcal{L}_{S1}^b and \mathcal{L}_{S2}^b efficiently we consider the simple graph defined by the dependency between slices. For a bag b , its adjacency matrix is $\mathbf{A}^b = [A_{ij}^b]$. The degree matrix $\mathbf{D}^b = [D_{ij}^b]$ is a diagonal matrix that contains the degree of

each slice (the degree of the slice i is the number of slices j such that $A_{ij}^b = 1$). This is, $D_{ii}^b = \text{degree}(i)$ and $D_{ij}^b = 0$ if $i \neq j$. Using these, one can compute the graph Laplacian matrix of a bag as $\mathbf{L}^b = \mathbf{D}^b - \mathbf{A}^b$. It is easy to show that

$$\mathcal{L}_{S1}^b = \mathbf{f}^b \top \mathbf{L}^b \mathbf{f}^b, \quad \mathcal{L}_{S2}^b = \mathbf{f}^b \top \mathbf{L}^b \mathbf{L}^b \mathbf{f}^b, \quad (6)$$

where $\mathbf{f}^b = [f(\mathbf{z}_1^b), \dots, f(\mathbf{z}_{N_b}^b)]^\top$. The sum of $\mathcal{L}_{S_k}^b$ over bags, where $k \in \{1, 2\}$, can be added to the loss function of a network to be minimized along the task-specific loss. Note that these two terms provide two different approaches to exploiting the correlations between instances through the loss function. We will refer to this approach as smooth attention (SA) loss. In the following subsection, we propose a model that can use either \mathcal{L}_{S1} or \mathcal{L}_{S2} . The effect of each term will be discussed in section 4.

2.4 SA-DMIL model description

We propose to couple the attention-based MIL pooling with the SA loss terms introduced in subsection 2.3. The proposed model, named Smooth Attention Deep Multiple Instance Learning (SA-DMIL), is depicted in Fig. 1. We use a Convolutional Neural Network (CNN), denoted by Φ_{CNN} , as a feature extractor to obtain a vector of low dimensional embeddings for each instance. That is, given a bag $\mathbf{X}^b = [\mathbf{x}_1^b, \dots, \mathbf{x}_{N_b}^b]$, where $\mathbf{x}_n^b \in \mathbb{R}^{3 \times HW}$, we compute

$$\mathbf{z}_n^b = \Phi_{\text{CNN}}(\mathbf{x}_n^b) \in \mathbb{R}^D, \quad \mathbf{Z}^b = [\mathbf{z}_1^b, \dots, \mathbf{z}_{N_b}^b]. \quad (7)$$

The CNN module in Fig. 1 is implemented with six convolutional blocks, followed by a flatten layer. \mathbf{Z}^b is then fed into the attention layer Φ_{Att} described in subsection 2.3 to obtain a scan representation. After that, the scan representation passes through a classifier Φ_c (i.e., one fully connected layer with a sigmoid activation) to predict the scan labels,

$$p(T^b | \mathbf{X}^b) \approx \Phi(\mathbf{X}^b) = \Phi_c(\Phi_{\text{Att}}(\Phi_{\text{CNN}}(\mathbf{X}^b))), \quad (8)$$

where we have written $\Phi_{\text{CNN}}(\mathbf{X}^b) = [\Phi_{\text{CNN}}(\mathbf{x}_1^b), \dots, \Phi_{\text{CNN}}(\mathbf{x}_{N_b}^b)]$. Our model, that corresponds to the composition $\Phi = \Phi_c \circ \Phi_{\text{Att}} \circ \Phi_{\text{CNN}}$, is trained using the following loss function until convergence,

$$\mathcal{L} = (1 - \alpha) \mathcal{L}_{\text{CE}} + \alpha \mathcal{L}_{S_k}, \quad (9)$$

where $\alpha \in [0, 1]$ is an hyperparameter and \mathcal{L}_{CE} the common cross-entropy loss,

$$\mathcal{L}_{\text{CE}} = \sum_b [T^b \log(\Phi(\mathbf{X}^b)) + (1 - T^b) \log(1 - \Phi(\mathbf{X}^b))], \quad (10)$$

where $k \in \{1, 2\}$, and $\mathcal{L}_{S_k} = \sum_b \mathcal{L}_{S_k}^b$ (see equations (4) and (5)). Depending on the value of k , we obtain two variations of SA-DMIL, which will be referred to as SA-DMIL-S1 and SA-DMIL-S2. The baseline model, Att-MIL (non-smooth attention), is recovered when $\alpha = 0.0$ [15]. Following the approach of previous studies [29,19], attention weights will be used to obtain predictions at the slice level (although they are not specifically designed for it). If a scan is predicted to be negative, all slices are also predicted to be negative, while if a scan is predicted to correspond to an ICH, slices whose attention weight is above a threshold (i.e., $1/N_b$, with N_b being the number of slices in that scan) are predicted as ICH.

3 Experimental design

3.1 Data and data preprocessing

The dataset used in this work was obtained from the 2019 Radiological Society of North America (RSNA) challenge [1], which included 39650 CT slices from 1150 subjects. The data were split among subjects, with 1000 scans (ICH: Normal scans = 411: 589; ICH: Normal slices = 4976: 29520) used for training and validation, and the remaining 150 scans (ICH: Normal scans = 72: 78; ICH: Normal slices = 806: 4448) used for held-out testing. The number of slices in the scans varied from 24 to 57. All CT slices underwent the same preprocessing procedure as described in [29]. Each CT slice had three windows applied to its original Hounsfield Units by changing the window Width (W) and Center (C) to manipulate the display of specific tissues, as radiologists typically do when diagnosing brain CTs. Here, we selected the brain (W: 80, C:40), subdural (W:200, C:80) and soft tissue (W:380, C: 40) windows. All images were then resized to the same size of 512×512 and normalized to the range $[0, 1]$. CTs were annotated at both the scan and slice levels, but slice labels were used for evaluation only, while scan labels were used for training and evaluation.

3.2 Experimental settings

We fix $D = 128$ and $L = 50$ in equation (3). We use the Adam optimizer with the learning rate starting at 10^{-4} . The batch size is set to 4, the maximum number of epochs is set to 200 and the patience for early stopping is set to 8. We test different values of the α hyperparameter, between 0 and 1 with a jump of 0.1. All experiments were run 5 independent times and the mean and standard deviation were reported in the held-old testing set at both scan and slice levels. The average training time is 10.3 hours for SA-DMIL-S1 and 10.5 hours for SA-DMIL-S2. The prediction time is approximately 15.8 seconds for each scan. All experiments were conducted using Tensorflow 2.11 in Python 3.8 on a single GPU (NVIDIA Quadro RTX 8000). The code will be available via GitHub.

4 Results and discussion

4.1 Hyperparameters tuning

In this subsection, we study the effect of SA loss in terms of performance. Table 1 compares the performance of models for different values of α . The standard deviation and other values of α can be found in the appendix, Tables S1 and S2. The results show that at both scan and slice levels, adding a smoothness term to the loss function ($\alpha > 0.0$) achieves better performance than Att-MIL ($\alpha = 0.0$). These improvements are significant, with increases in accuracy, F1 and AUC scores of approximately 7%, 9% and 5% respectively, at scan level, and increases in accuracy and F1 score of 8% and 11% respectively, at slice level. The recall is the only metric in which our model does not excel, where the baseline Att-MIL obtains the best value. However, this is associated with very low precision values. Note that, as α increases, the performance of the model first

improves and then drops, which is consistent with the role played by the SA loss as a regularization term. The difference between \mathcal{L}_{S1} and \mathcal{L}_{S2} is not significant although \mathcal{L}_{S1} performs slightly better. In fact, when using \mathcal{L}_{S1} , $\alpha = 0.5$ gives the best diagnostic performance with an AUC of 0.879 (± 0.003) at scan level and an accuracy of 0.834 (± 0.010) at slice level.

Table 1: Performance of SA-DMIL and other MIL methods at slice and scan levels on the RSNA dataset. The average of 5 independent runs is reported. For space constraints, the standard deviation is reported in the appendix.

Model	Scan level					Slice level				
	Acc	Pre	Rec	F1	AUC	Acc	Pre	Rec	F1	
SA-DMIL-S1	$\alpha = 0.9$	0.753	0.803	0.681	0.735	0.839	0.789	0.670	0.541	0.598
	$\alpha = 0.7$	0.806	0.763	0.784	0.775	0.860	0.828	0.679	0.576	0.639
	$\alpha = 0.5$	0.813	0.805	0.806	0.806	0.879	0.834	0.732	0.608	0.686
	$\alpha = 0.3$	0.767	0.734	0.806	0.768	0.859	0.775	0.702	0.551	0.624
	$\alpha = 0.1$	0.747	0.783	0.652	0.712	0.841	0.766	0.649	0.540	0.584
SA-DMIL-S2	$\alpha = 0.9$	0.753	0.817	0.613	0.714	0.816	0.768	0.733	0.551	0.598
	$\alpha = 0.7$	0.767	0.776	0.722	0.748	0.843	0.807	0.734	0.591	0.638
	$\alpha = 0.5$	0.800	0.828	0.736	0.780	0.867	0.823	0.748	0.596	0.659
	$\alpha = 0.3$	0.763	0.797	0.686	0.721	0.853	0.790	0.738	0.561	0.622
	$\alpha = 0.1$	0.747	0.736	0.740	0.736	0.833	0.767	0.683	0.547	0.593
Att-MIL ($\alpha = 0.0$) [15]	0.740	0.674	0.832	0.719	0.829	0.751	0.623	0.543	0.579	
MIL + Max agg. [28]	0.617	0.856	0.447	0.575	0.743	0.732	0.441	0.373	0.406	
MIL + Mean agg. [28]	0.677	0.670	0.734	0.693	0.801	0.741	0.502	0.386	0.447	
Att-CNN + VGPMIL [29]	0.765	0.724	0.851	0.773	0.868	0.807	0.714	0.538	0.597	

4.2 Smooth Attention MIL vs. other MIL methods

The performance of other popular MIL methods is also included in Table 1. All methods share the same CNN architecture to extract slice features, but they differ in the pooling operator they use: Max [28], Mean [28], Attention [15] or Gaussian Process (GP) [29]. These results show that the performance of SA-DMIL is consistently better than other methods across different metrics and at both scan and slice levels. Only the precision of MIL+Max agg. and the recall of AttCNN+VGPMIL at scan level are higher than those obtained by SA-DMIL. However, considering the trade-off between precision and recall given by F1, our method achieves a superior performance. In tasks like ICH detection, where neighbouring instances are expected to have similar diagnostic importance. Unlike other MIL methods that assume each instance to be independently distributed, SA-DMIL stands out by considering the spatial correlation between instances, which compels it to learn more meaningful features for making accurate bag predictions. Notably, this is achieved by simply adding a smoothing term to the loss function without increasing the number of model parameters. This can potentially be applied to existing architectures to further improve performance without adding complexity.

4.3 Visualizing smooth regularizing effects at slice level

So far we have observed enhanced performance through the SA term. In this subsection, we visually illustrate how this novel term imposes smoothness be-

tween attention scores of consecutive slices, leading to more accurate predictions. Figure 2 shows plots of the attention scores assigned by SA-DMIL-*S1* and Att-MIL to the slices of three different scans (Fig. S1 in the appendix contains an analogous plot for SA-DMIL-*S2*). As expected, introducing the SA loss results in smoother attention weights. Note that the smoothness constraint of SA-DMIL effectively penalizes the appearance of isolated non-smooth attention weights that incorrectly jump over or below the threshold.

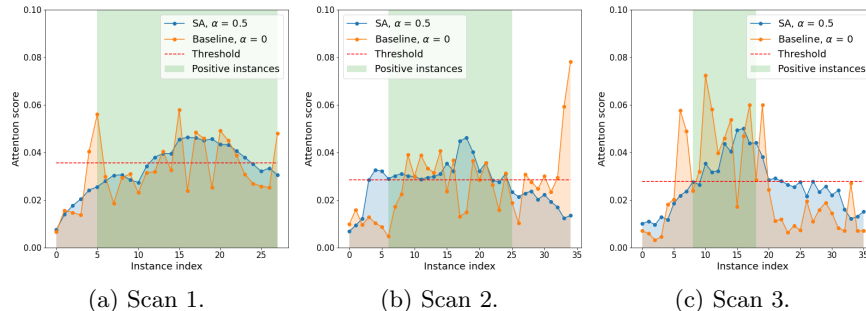


Fig. 2: Attention weights of SA-DMIL-*S1* (blue lines, $\alpha = 0.5$) and Att-MIL [15] (orange lines, $\alpha = 0.0$). Slices with values above the threshold ($1/N_b$) are predicted as ICH, while those below are predicted as Normal. The green area highlights those slices whose ground truth label is ICH.

We also include visual examples of consecutive CT slices in Fig. 3. In Scan 1, the baseline Att-MIL produces a wrong prediction at scan level. When using SA, the prediction is correct since dependencies between adjacent slices have been learned. In Scan 2, both models produce correct predictions at scan level, but SA-DMIL is more accurate at slice level. This occurs thanks to the SA loss, that turns the attention scores into smoother values and, therefore, avoids random *jumps* up and down the decision threshold.

5 Conclusion

In this study we have proposed SA-DMIL, a new model that obtains significant improvements in ICH classification compared to state-of-the-art MIL methods. This is done by adding a smoothing regularizing term to the loss function. This term imposes a smoothness constraint on the latent function that encodes the attention weights, which forces our model to learn dependencies between instances rather than training each instance independently in a bag. This flexible approach does not introduce any additional complexity, so similar ideas can be applied to other methods to model dependencies between neighboring instances.

Data use declaration

The dataset used in this study is from the 2019 RSNA Intracranial Hemorrhage Detection Challenge and is publicly available in this link.

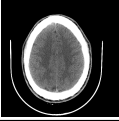



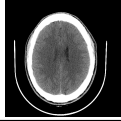
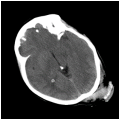
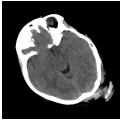
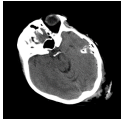
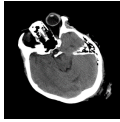
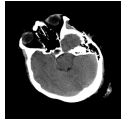
Scan 1						Scan label
Ground truth	Normal	Normal	Normal	Normal	Normal	Normal
Att-MIL ($\alpha = 0.0$) [15]	ICH ●	ICH ●	Normal ●	ICH ●	Normal ●	ICH ●
SA-DMIL-S1 ($\alpha = 0.5$)	Normal ●	Normal ●	Normal ●	Normal ●	Normal ●	Normal ●
Scan 2						Scan label
Ground truth	ICH	ICH	Normal	Normal	Normal	ICH
Att-MIL ($\alpha = 0.0$) [15]	Normal ●	ICH ●	ICH ●	ICH ●	Normal ●	ICH ●
SA-DMIL-S1 ($\alpha = 0.5$)	ICH ●	ICH ●	Normal ●	Normal ●	Normal ●	ICH ●

Fig. 3: Predictions of Att-MIL [15] and SA-DMIL-S1 at CT slice level in two different scans. SA improves predictions at both scan and slice level. Red color: incorrect prediction, green color: correct prediction.

Acknowledgements

This work was supported by project PID2019-105142RB-C22 funded by Ministerio de Ciencia e Innovación and by project B-TIC-324-UGR20 funded by FEDER/Junta de Andalucía and Universidad de Granada. The work by Francisco M. Castro-Macías was supported by Ministerio de Universidades under FPU contract FPU21/01874.

References

1. RSNA intracranial hemorrhage detection, <https://kaggle.com/c/rsna-intracranial-hemorrhage-detection>
2. Arbabshirani, M.R., Fornwalt, B.K., Mongelluzzo, G.J., Suever, J.D., Geise, B.D., Patel, A.A., Moore, G.J.: Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ digital medicine* **1**(1), 9 (2018)
3. Arendts, G., Manovel, A., Chai, A.: Cranial ct interpretation by senior emergency department staff. *Australasian radiology* **47**(4), 368–374 (2003)
4. Belkin, M., Niyogi, P., Sindhvani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research* **7**(11) (2006)
5. Caceres, J.A., Goldstein, J.N.: Intracranial hemorrhage. *Emergency medicine clinics of North America* **30**(3), 771–794 (2012)
6. Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition* **77**, 329–353 (2018)
7. Chang, P.D., Kuoy, E., Grinband, J., Weinberg, B.D., Thompson, M., Homo, R., Chen, J., Abcede, H., Shafie, M., Sugrue, L., et al.: Hybrid 3d/2d convolutional

- neural network for hemorrhage evaluation on head ct. *American Journal of Neuroradiology* **39**(9), 1609–1616 (2018)
8. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis* **54**, 280–296 (2019)
 9. Cordonnier, C., Demchuk, A., Ziai, W., Anderson, C.S.: Intracerebral hemorrhage: current approaches to acute management. *The Lancet* **392**(10154), 1257–1268 (2018)
 10. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* **89**(1-2), 31–71 (1997)
 11. Elliott, J., Smith, M.: The acute management of intracerebral hemorrhage: a clinical review. *Anesthesia & Analgesia* **110**(5), 1419–1427 (2010)
 12. Erly, W.K., Berger, W.G., Krupinski, E., Seeger, J.F., Guisto, J.A.: Radiology resident evaluation of head ct scan orders in the emergency department. *American journal of neuroradiology* **23**(1), 103–107 (2002)
 13. Gadermayr, M., Tschuchnig, M.: Multiple instance learning for digital pathology: A review on the state-of-the-art, limitations & future potential. arXiv preprint arXiv:2206.04425 (2022)
 14. Grewal, M., Srivastava, M.M., Kumar, P., Varadarajan, S.: Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 281–284. IEEE (2018)
 15. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
 16. Ker, J., Singh, S.P., Bai, Y., Rao, J., Lim, T., Wang, L.: Image thresholding improves 3-dimensional convolutional neural network diagnosis of different acute brain hemorrhages on computed tomography scans. *Sensors* **19**(9), 2167 (2019)
 17. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *nature* **521**(7553), 436–444 (2015)
 18. Li, H., Yang, F., Xing, X., Zhao, Y., Zhang, J., Liu, Y., Han, M., Huang, J., Wang, L., Yao, J.: Multi-modal multi-instance learning using weakly correlated histopathological images and tabular clinical information. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. pp. 529–539. Springer (2021)
 19. López-Pérez, M., Schmidt, A., Wu, Y., Molina, R., Katsaggelos, A.K.: Deep gaussian processes for multiple instance learning: Application to ct intracranial hemorrhage detection. *Computer Methods and Programs in Biomedicine* **219**, 106783 (2022)
 20. McDonald, R.J., Schwartz, K.M., Eckel, L.J., Diehn, F.E., Hunt, C.H., Bartholmai, B.J., Erickson, B.J., Kallmes, D.F.: The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic radiology* **22**(9), 1191–1198 (2015)
 21. Quellec, G., Cazuguel, G., Cochener, B., Lamard, M.: Multiple-instance learning for medical image and video analysis. *IEEE reviews in biomedical engineering* **10**, 213–234 (2017)
 22. Qureshi, A.I., Tuhim, S., Broderick, J.P., Batjer, H.H., Hondo, H., Hanley, D.F.: Spontaneous intracerebral hemorrhage. *New England Journal of Medicine* **344**(19), 1450–1460 (2001)
 23. Ripley, B.: *Spatial statistics* john wiley & sons. New York, New York (1981)

24. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems* **34**, 2136–2147 (2021)
25. Strub, W., Leach, J., Tomsick, T., Vagal, A.: Overnight preliminary head ct interpretations provided by residents: locations of misidentified intracranial hemorrhage. *American journal of neuroradiology* **28**(9), 1679–1682 (2007)
26. Teneggi, J., Yi, P.H., Sulam, J.: Weakly supervised learning significantly reduces the number of labels required for intracranial hemorrhage detection on head ct. arXiv preprint arXiv:2211.15924 (2022)
27. Titano, J.J., Badgeley, M., Schefflein, J., Pain, M., Su, A., Cai, M., Swinburne, N., Zech, J., Kim, J., Bederson, J., et al.: Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nature medicine* **24**(9), 1337–1341 (2018)
28. Wang, Y., Li, J., Metze, F.: A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 31–35. IEEE (2019)
29. Wu, Y., Schmidt, A., Hernández-Sánchez, E., Molina, R., Katsaggelos, A.K.: Combining attention-based multiple instance learning and gaussian processes for ct hemorrhage detection. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 582–591. Springer (2021)
30. Ye, H., Gao, F., Yin, Y., Guo, D., Zhao, P., Lu, Y., Wang, X., Bai, J., Cao, K., Song, Q., et al.: Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. *European radiology* **29**, 6191–6201 (2019)
31. Yeo, M., Tahayori, B., Kok, H.K., Maingard, J., Kutaiba, N., Russell, J., Thijs, V., Jhamb, A., Chandra, R.V., Brooks, M., et al.: Review of deep learning algorithms for the automatic detection of intracranial hemorrhages on computed tomography head imaging. *Journal of neurointerventional surgery* **13**(4), 369–378 (2021)

Supplementary Materials

Table S1: Performance of SA-DMIL and other MIL methods at scan level on the RSNA dataset. The average of 5 independent runs is reported.

Model	Scan level					
	Acc	Pre	Rec	F1	AUC	
SA-DMIL-S1	$\alpha = 0.9$	0.753 \pm 0.014	0.803 \pm 0.006	0.681 \pm 0.017	0.735 \pm 0.011	0.839 \pm 0.007
	$\alpha = 0.8$	0.754 \pm 0.010	0.801 \pm 0.004	0.713 \pm 0.014	0.740 \pm 0.009	0.851 \pm 0.006
	$\alpha = 0.7$	0.806 \pm 0.007	0.763 \pm 0.008	0.784 \pm 0.010	0.775 \pm 0.008	0.860 \pm 0.010
	$\alpha = 0.6$	0.801 \pm 0.009	0.791 \pm 0.004	0.806 \pm 0.009	0.799 \pm 0.006	0.869 \pm 0.007
	$\alpha = 0.5$	0.813 \pm 0.010	0.805 \pm 0.003	0.806 \pm 0.011	0.806 \pm 0.006	0.879 \pm 0.003
	$\alpha = 0.4$	0.773 \pm 0.015	0.806 \pm 0.007	0.694 \pm 0.018	0.746 \pm 0.012	0.863 \pm 0.004
	$\alpha = 0.3$	0.767 \pm 0.012	0.734 \pm 0.011	0.806 \pm 0.007	0.768 \pm 0.008	0.859 \pm 0.010
	$\alpha = 0.2$	0.747 \pm 0.021	0.681 \pm 0.010	0.889 \pm 0.004	0.771 \pm 0.006	0.857 \pm 0.011
	$\alpha = 0.1$	0.747 \pm 0.017	0.783 \pm 0.010	0.652 \pm 0.018	0.712 \pm 0.014	0.841 \pm 0.014
SA-DMIL-S2	$\alpha = 0.9$	0.753 \pm 0.012	0.817 \pm 0.010	0.613 \pm 0.019	0.714 \pm 0.013	0.816 \pm 0.010
	$\alpha = 0.8$	0.733 \pm 0.021	0.813 \pm 0.008	0.656 \pm 0.017	0.747 \pm 0.009	0.825 \pm 0.008
	$\alpha = 0.7$	0.767 \pm 0.014	0.776 \pm 0.012	0.722 \pm 0.013	0.748 \pm 0.010	0.843 \pm 0.008
	$\alpha = 0.6$	0.760 \pm 0.009	0.846 \pm 0.007	0.711 \pm 0.014	0.778 \pm 0.008	0.852 \pm 0.007
	$\alpha = 0.5$	0.800 \pm 0.008	0.828 \pm 0.010	0.736 \pm 0.017	0.780 \pm 0.011	0.867 \pm 0.008
	$\alpha = 0.4$	0.773 \pm 0.010	0.896 \pm 0.005	0.697 \pm 0.013	0.737 \pm 0.007	0.880 \pm 0.004
	$\alpha = 0.3$	0.763 \pm 0.012	0.797 \pm 0.007	0.686 \pm 0.016	0.721 \pm 0.010	0.853 \pm 0.011
	$\alpha = 0.2$	0.753 \pm 0.010	0.818 \pm 0.006	0.625 \pm 0.021	0.709 \pm 0.017	0.838 \pm 0.012
	$\alpha = 0.1$	0.747 \pm 0.019	0.736 \pm 0.011	0.740 \pm 0.012	0.736 \pm 0.013	0.833 \pm 0.016
Att-MIL ($\alpha = 0.0$) [15]	0.740 \pm 0.015	0.674 \pm 0.024	0.832 \pm 0.011	0.719 \pm 0.014	0.829 \pm 0.009	
MIL + Max agg. [28]	0.617 \pm 0.031	0.856 \pm 0.030	0.447 \pm 0.109	0.575 \pm 0.068	0.743 \pm 0.015	
MIL + Mean agg. [28]	0.677 \pm 0.028	0.670 \pm 0.032	0.734 \pm 0.041	0.693 \pm 0.040	0.801 \pm 0.016	
Att-CNN + VGPMIL [29]	0.765 \pm 0.017	0.724 \pm 0.012	0.851 \pm 0.008	0.773 \pm 0.010	0.868 \pm 0.007	

Table S2: Performance of SA-DMIL and other MIL methods at slice level on the RSNA dataset. The average of 5 independent runs is reported.

Model		Slice level			
		Acc	Pre	Rec	F1
SA-DMIL-S1	$\alpha = 0.9$	0.789 ± 0.025	0.670 ± 0.031	0.541 ± 0.051	0.598 ± 0.048
	$\alpha = 0.8$	0.794 ± 0.017	0.683 ± 0.027	0.548 ± 0.047	0.622 ± 0.040
	$\alpha = 0.7$	0.828 ± 0.011	0.679 ± 0.028	0.576 ± 0.038	0.639 ± 0.037
	$\alpha = 0.6$	0.821 ± 0.014	0.687 ± 0.025	0.579 ± 0.037	0.643 ± 0.029
	$\alpha = 0.5$	0.834 ± 0.010	0.732 ± 0.021	0.608 ± 0.027	0.686 ± 0.018
	$\alpha = 0.4$	0.788 ± 0.017	0.718 ± 0.023	0.563 ± 0.031	0.647 ± 0.024
	$\alpha = 0.3$	0.775 ± 0.018	0.702 ± 0.028	0.551 ± 0.036	0.624 ± 0.032
	$\alpha = 0.2$	0.768 ± 0.023	0.661 ± 0.034	0.551 ± 0.032	0.611 ± 0.035
	$\alpha = 0.1$	0.766 ± 0.022	0.649 ± 0.032	0.540 ± 0.041	0.584 ± 0.039
SA-DMIL-S2	$\alpha = 0.9$	0.768 ± 0.030	0.733 ± 0.024	0.551 ± 0.048	0.598 ± 0.046
	$\alpha = 0.8$	0.766 ± 0.022	0.736 ± 0.022	0.573 ± 0.046	0.625 ± 0.042
	$\alpha = 0.7$	0.807 ± 0.018	0.734 ± 0.028	0.591 ± 0.049	0.638 ± 0.039
	$\alpha = 0.6$	0.801 ± 0.020	0.746 ± 0.026	0.594 ± 0.037	0.655 ± 0.029
	$\alpha = 0.5$	0.823 ± 0.012	0.748 ± 0.022	0.596 ± 0.029	0.659 ± 0.024
	$\alpha = 0.4$	0.812 ± 0.017	0.751 ± 0.020	0.583 ± 0.027	0.637 ± 0.019
	$\alpha = 0.3$	0.790 ± 0.021	0.738 ± 0.028	0.561 ± 0.034	0.622 ± 0.031
	$\alpha = 0.2$	0.784 ± 0.024	0.742 ± 0.024	0.551 ± 0.037	0.617 ± 0.028
	$\alpha = 0.1$	0.767 ± 0.031	0.683 ± 0.029	0.547 ± 0.040	0.593 ± 0.037
Att-MIL ($\alpha = 0.0$) [15]		0.751 ± 0.024	0.623 ± 0.037	0.543 ± 0.047	0.579 ± 0.041
MIL + Max agg. [28]		0.732 ± 0.041	0.441 ± 0.108	0.373 ± 0.152	0.406 ± 0.128
MIL + Mean agg. [28]		0.741 ± 0.038	0.502 ± 0.110	0.386 ± 0.117	0.447 ± 0.107
Att-CNN + VGPMIL [29]		0.807 ± 0.022	0.714 ± 0.028	0.538 ± 0.039	0.597 ± 0.036

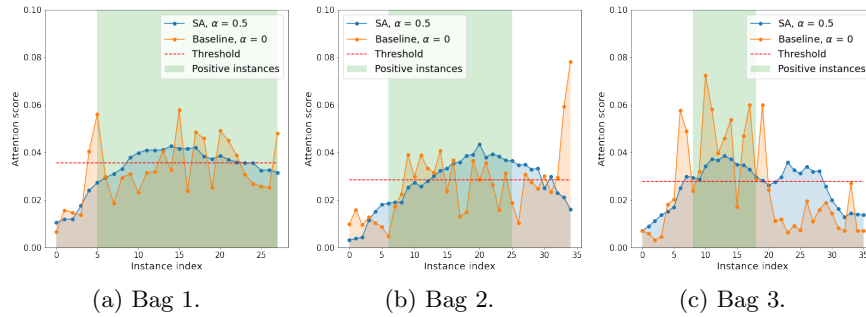


Fig. S1: Attention weights of SA-DMIL-S2 (blue lines, $\alpha = 0.5$) and Att-MIL [15] (orange lines, $\alpha = 0.0$). Slices with values above the threshold ($1/N_b$) are predicted as ICH, while those below are predicted as Normal. The green area highlights those slices whose ground truth label is ICH.