$\mathtt{Sm}$: enhanced localization in Multiple Instance Learning for medical imaging classification

F.M. Castro-Macías[1] P. Morales-Álvarez[1] Y. Wu[2] R. Molina[1] A. K. Katsaggelos[2]

[1]University of Granada, Spain [2]Northwestern University, USA

UNIVERSIDAD DE GRANADA

NORTHWESTERN UNIVERSITY

Multiple Instance Learning (MIL)

Multiple Instance Learning (MIL) is a type of weakly supervised learning that is particularly useful when obtaining fine-grain annotations is expensive, which is the case of medical imaging and drug discovery.

The **training data** consists of pairs of the form $(\mathbf{X}, Y)$ where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times P}$ is a bag, and $\mathbf{x}_n \in \mathbb{R}^P$ are the instances. The instances have labels $\{y_1, \ldots, y_N\} \subset \{0, 1\}$, which are not observed. Only the bag label Y is observed, and it holds $Y = \max\{y_1, \ldots, y_N\} \in \{0, 1\}$.

At **test time**, given a new bag, we want to predict the bag label (classification task), and the instance labels (localization task).

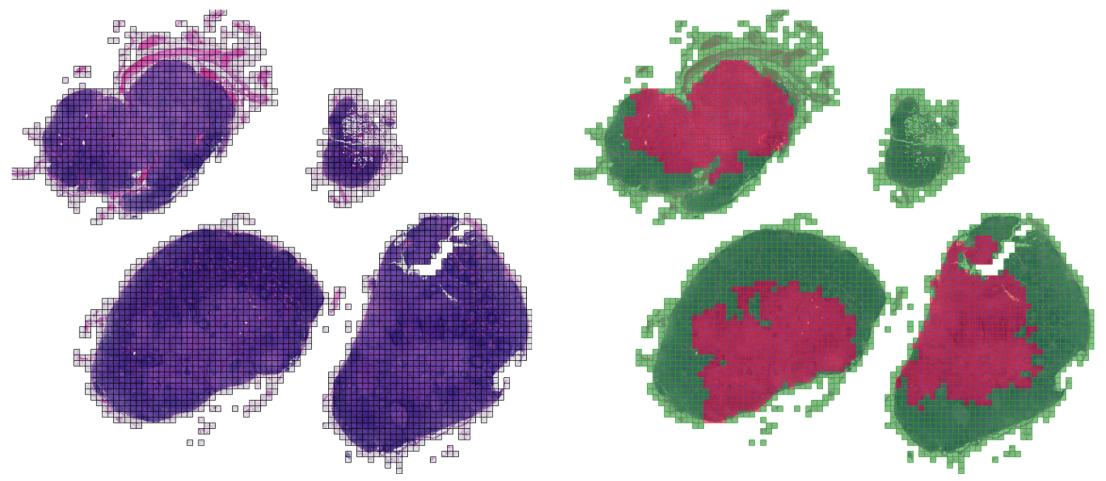


Figure 1. Whole Slide Image (WSI, bag) and labeled patches (instances).

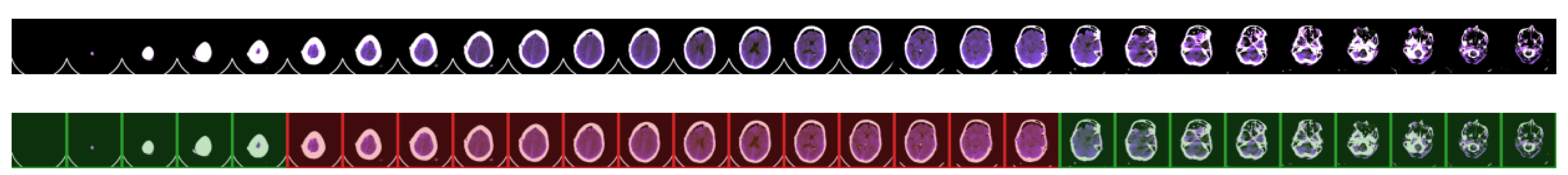


Figure 2. Computerized Tomography (CT) scan (bag) and labeled slices (instances).

Background: Deep MIL

How do the most succesful deep MIL approaches work? Two important choices:

1. They assign an **attention value** $f_n \in \mathbb{R}$ to each instance. These are used to generate the bag label prediction and as a proxy to estimate the instance labels.
2. They incorporate both global and local interactions using different mechanisms: transformers, graph neural networks...

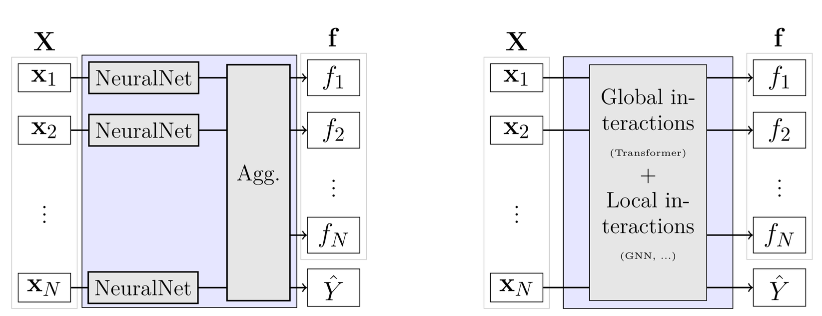


Figure 3. Architecture of deep MIL methods without interactions (left) and with interactions (right).

Problem. These methods have been designed to target the classification task... what about localization? The implications of their design choices in this task have not been studied!

How to solve this? We propose a method to be competitive in both tasks!

Our idea: attention maps should be *smooth*

Observation. Instance labels show spatial dependencies: an instance is likely to be surrounded by instances with the same label (see Figures 1 and 2).

If we want to use the attention values $\mathbf{f}$ to predict the instance labels, they should inherit this smoothing property!

Method: modelling the smoothness

We represent each bag as a graph, where the nodes are the instances and the edges represent the spatial connectivity between instances. We interpret the attention values $\mathbf{f} \in \mathbb{R}^N$ as a function defined on the bag graph.

Dirichlet energy $\mathcal{E}_D$. Measure of the variability of a function defined on a graph [2].

Goal. We want to produce smooth $\mathbf{f}$, i.e., to output $\mathbf{f}$ with low Dirichlet energy $\mathcal{E}_D(\mathbf{f})$.

Bounding $\mathcal{E}_D(\mathbf{f})$. We can bound the Dirichlet energy of the attention values using the previous layers. For example, modelling $\mathbf{f}$ as in ABMIL [1], we have

$$\mathcal{E}_D(\mathbf{f}) \leq \|\mathbf{w}\|_2^2 \mathcal{E}_D(\mathbf{F}) \leq \|\mathbf{w}\|_2^2 \|\mathbf{W}\|_2^2 \mathcal{E}_D(\mathbf{X})$$

where $\mathbf{f} = \mathbf{F}\mathbf{w}$, $\mathbf{F} = \tanh\left(\mathbf{X}\mathbf{W}^\top\right)$, and $\mathbf{w}, \mathbf{W}$ are trainable weights. This results generalizes for arbitrary depth (see the paper!).

Approach. We can act on $\mathbf{f}$ itself and on the output of previous layers. We develop the smooth operator to decrease the Dirichlet energy of any kind of layer.

Method: the smooth operator $\mathtt{Sm}$

Given $\mathbf{U} \in \mathbb{R}^{N \times D}$ and $\gamma \in \mathbb{R}^+$, the Smooth operator ($\mathtt{Sm}$) is defined as

$$\mathtt{Sm}(\mathbf{U}) = (\mathbf{I} + \gamma \mathbf{L})^{-1} \mathbf{U},$$

where $\mathbf{L}$ is the Laplacian of the bag graph.

$\mathtt{Sm}$ is principled. Trade-off between fidelity to the input signal and smoothness,

$$\mathtt{Sm}(\mathbf{U}) = \arg\min_{\mathbf{G}} \left\{ \alpha \mathcal{E}_D(\mathbf{G}) + (1 - \alpha) \|\mathbf{U} - \mathbf{G}\|_F^2 \right\}, \quad \alpha \in [0, 1).$$

$\mathtt{Sm}$ decreases the Dirichlet energy. If $\mathbf{L}$ is the normalized Laplacian matrix, then

$$\mathcal{E}_D(\mathtt{Sm}(\mathbf{U})) < \mathcal{E}_D(\mathbf{U}).$$

$\mathtt{Sm}$ is cheap to compute. It can be computed iteratively,

$$\mathtt{Sm}(\mathbf{U}) = \lim_{t \to \infty} \mathbf{G}(t),$$

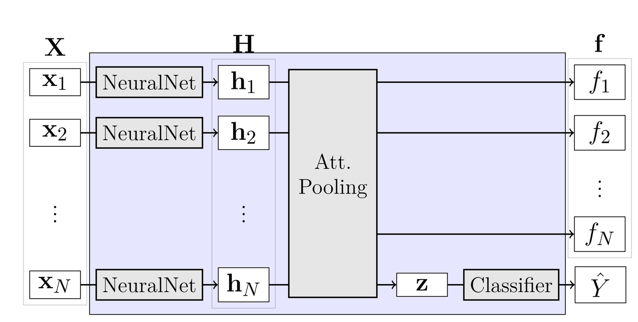
$$\mathbf{G}(0) = \mathbf{U}; \quad \mathbf{G}(t) = \alpha(\mathbf{I} - \mathbf{L})\mathbf{G}(t-1) + (1 - \alpha)\mathbf{U}.$$

Method: the proposed model

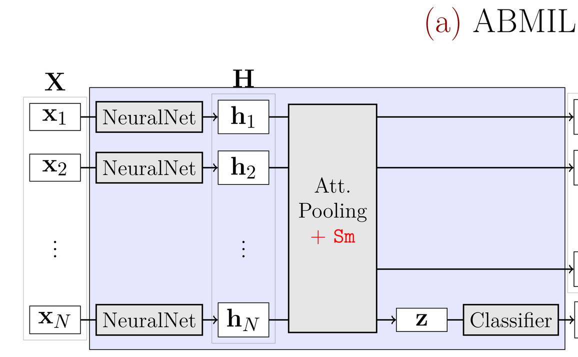
We build on top of the well-known ABMIL [1], proposing two modifications.

$\mathtt{Sm}$AP. We add the smooth operator $\mathtt{Sm}$ in the attention pooling. It accounts for local interactions.

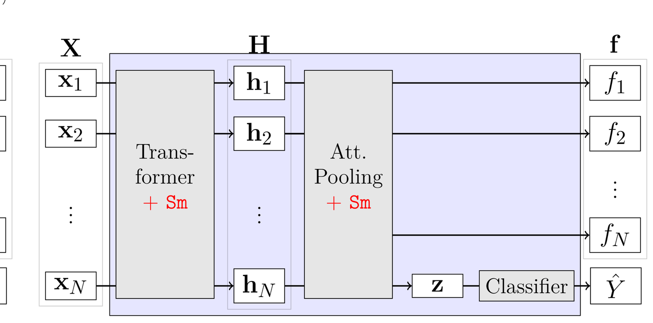
$\mathtt{Sm}$TAP. We use a transformer encoder to account for global interactions. We add the smooth operator $\mathtt{Sm}$ both in the transformer and in the attention pooling, accounting for both global and local interactions.



(a) ABMIL [1], the baseline.



(b) $\mathtt{Sm}$AP.



(c) $\mathtt{Sm}$TAP.

Figure 4. Proposed models.

Experiments: quantitative evaluation

Experimental setup. We evaluate the proposed models

- in 3 different medical imaging datasets: RSNA (CT scans), PANDA (WSIs), and CAMELYON16 (WSIs),
- using 4 different feature extractors, trained with and without self-supervised learning,
- considering up to 13 different SOTA methods for comparison,
- in both localization and classification tasks.

		Instance localization	Bag classification
Without global interactions	$\mathtt{Sm}$AP	$\mathbf{1.500}_{0.548}$	$\mathbf{1.833}_{0.753}$
	ABMIL	$\underline{2.500}_{1.225}$	$2.500_{1.049}$
	CLAM	$4.167_{1.329}$	$4.500_{0.837}$
	DSMIL	$4.333_{0.516}$	$4.167_{0.753}$
	DFTD-MIL	$2.500_{1.049}$	$\underline{2.000}_{1.265}$
With global interactions	$\mathtt{Sm}$TAP	$\mathbf{1.500}_{1.225}$	$\mathbf{1.833}_{0.983}$
	TransMIL	$3.083_{1.429}$	$4.083_{0.917}$
	SETMIL	$3.667_{0.816}$	$3.583_{2.010}$
	GTP	$3.917_{1.429}$	$2.750_{0.987}$
	CAMIL	$\underline{2.833}_{1.169}$	$\underline{2.750}_{1.173}$

Table 1. Average rank (lower is better).

Results. The proposed methods with $\mathtt{Sm}$ achieve the best performance in localization and remain very competitive in classification.

Experiments: attention histograms and attention maps

We examine the attention histograms and the attention maps produced by each model on the CAMELYON16 dataset. The proposed $\mathtt{Sm}$AP and $\mathtt{Sm}$TAP stand out at separating positive and negative instances.

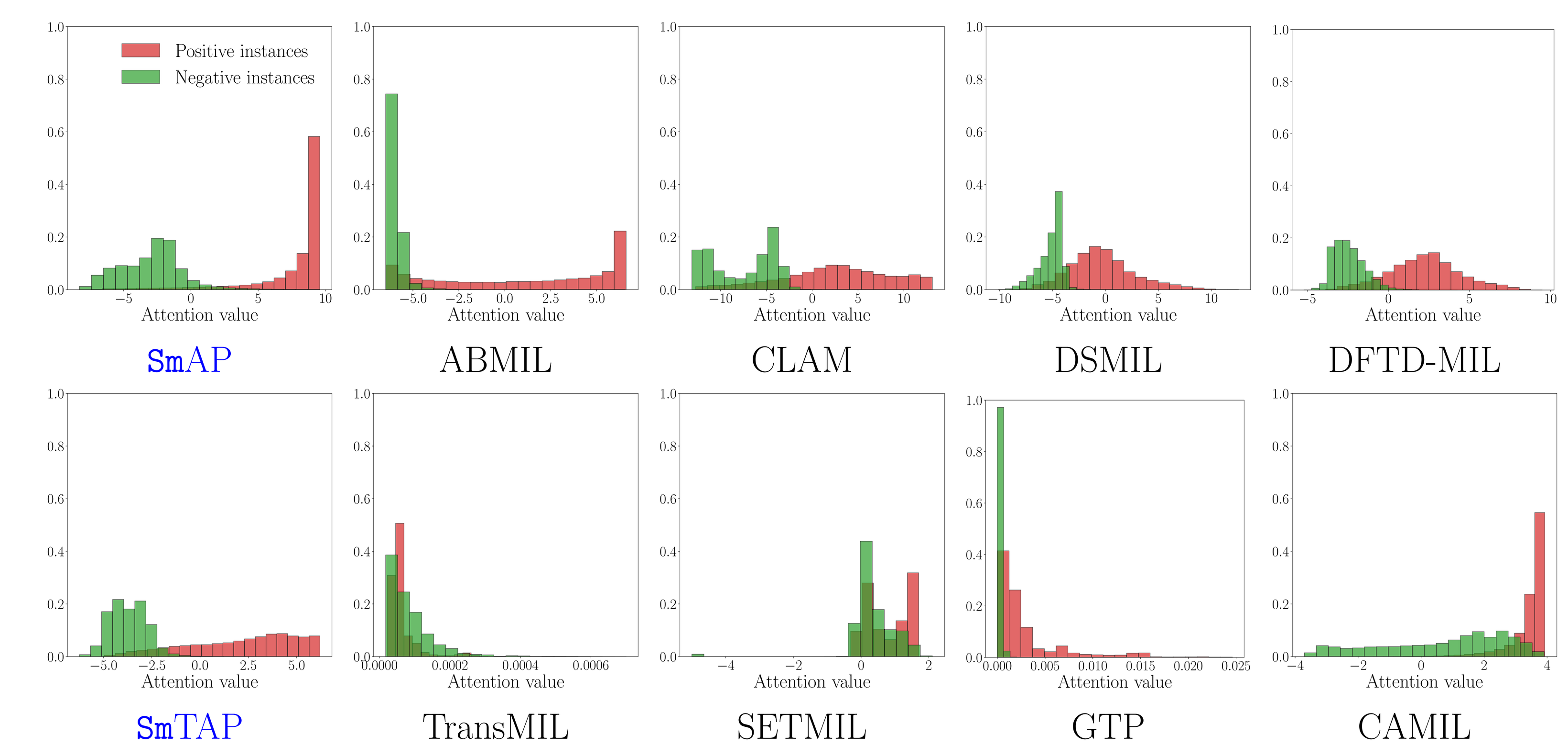


Figure 5. Attention histograms on CAMELYON16.

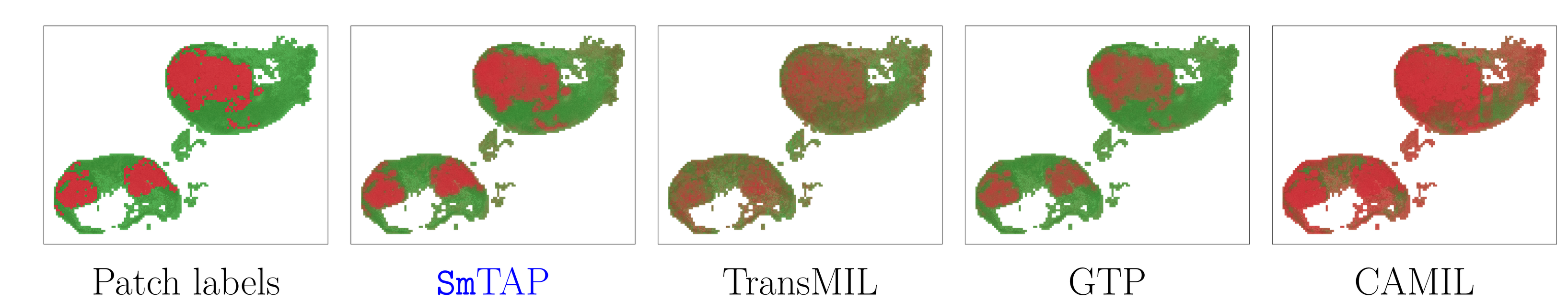


Figure 6. Attention maps on CAMELYON16.

References

[1] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[2] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16, 2003.

Check our code!

github.com/Franblueee/SmMIL

fcastro@ugr.es, pablomorales@ugr.es