Sm: enhanced localization in Multiple Instance Learning for medical imaging classification

F.M. Castro-Macías, P. Morales-Álvarez, Y. Wu, R. Molina, A. K. Katsaggelos

2024 Conference on Neural Information Processing Systems (NeurIPS)

Multiple Instance Learning (MIL)

Training data: pairs of the form $(\mathbf{X}, Y)$.

- Bag: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times P}$, $\mathbf{x}_n \in \mathbb{R}^P$.
- Instance labels (not observed): $\{y_1, \ldots, y_N\} \subset \{0, 1\}$.
- Bag label (observed): $Y = \max\{y_1, \ldots, y_N\} \in \{0, 1\}$.

Multiple Instance Learning (MIL)

Training data: pairs of the form $(\mathbf{X}, Y)$.

- Bag: $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times P}$, $\mathbf{x}_n \in \mathbb{R}^P$.
- Instance labels (not observed): $\{y_1, \dots, y_N\} \subset \{0, 1\}$.
- Bag label (observed): $Y = \max\{y_1, \dots, y_N\} \in \{0, 1\}$.

Test time: given a new bag, we want to predict

- the bag label (classification task),
- the instance labels (localization task).

Multiple Instance Learning (MIL)

Training data: pairs of the form $(\mathbf{X}, Y)$.

- Bag: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times P}$, $\mathbf{x}_n \in \mathbb{R}^P$.
- Instance labels (not observed): $\{y_1, \ldots, y_N\} \subset \{0, 1\}$.
- Bag label (observed): $Y = \max\{y_1, \ldots, y_N\} \in \{0, 1\}$.

Test time: given a new bag, we want to predict

- the bag label (classification task),
- the instance labels (localization task).

Why is it useful? Minimal annotation effort.

MIL in medical imaging

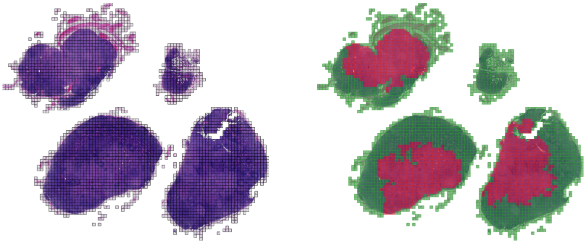


Figure: Whole Slide Image (WSI, bag) and labeled patches (instances).

MIL in medical imaging

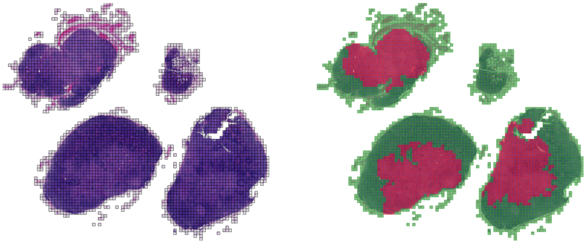


Figure: Whole Slide Image (WSI, bag) and labeled patches (instances).

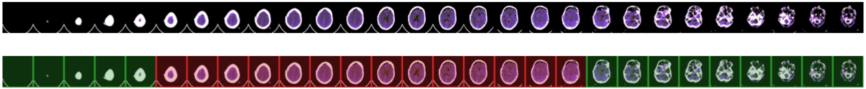
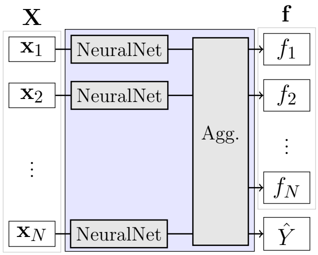
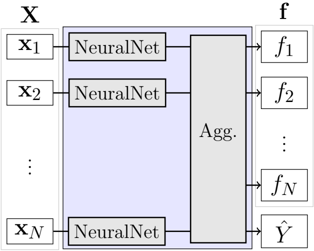


Figure: Computerized Tomography (CT) scan (bag) and labeled slices (instances).

Background: deep MIL

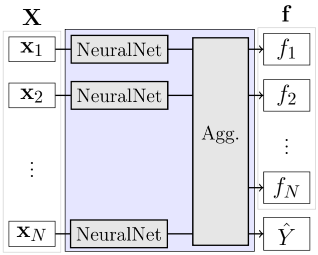


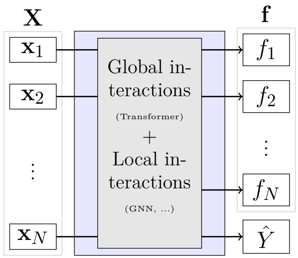
Background: deep MIL



- Attention values ($f_n \in \mathbb{R}$) are used as a proxy to estimate the instance labels.

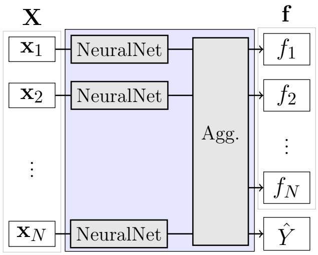
Background: deep MIL

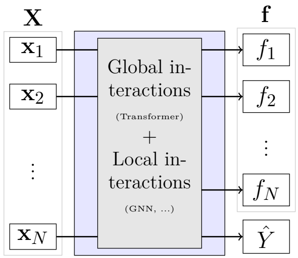




- Attention values ($f_n \in \mathbb{R}$) are used as a proxy to estimate the instance labels.

Background: deep MIL





- Attention values ($f_n \in \mathbb{R}$) are used as a proxy to estimate the instance labels.
- Interactions have shown to improve the classification performance.

Background: deep MIL

- Attention values ($f_n \in \mathbb{R}$) are used as a proxy to estimate the instance labels.
- Interactions have shown to improve the classification performance.
- **Problem:** previous works have been designed to target the classification task... what about localization?

Method: the idea

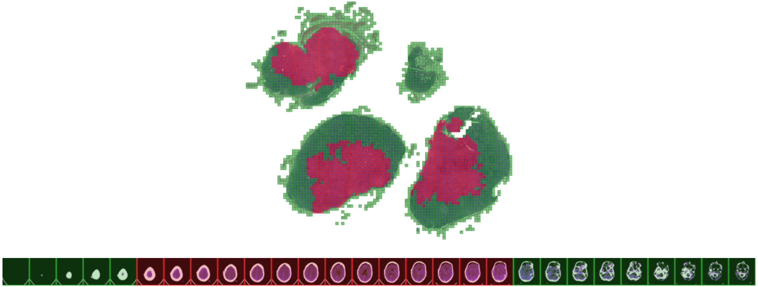


Figure: Map of labeled instances.

- Instance labels show spatial dependencies: an instance is likely to be surrounded by instances with the same label.

Method: the idea

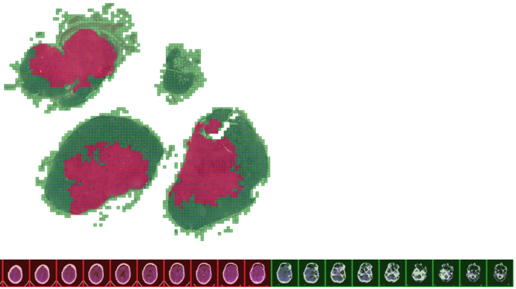


Figure: Map of labeled instances.

- Instance labels show spatial dependencies: an instance is likely to be surrounded by instances with the same label.
- Attention values f_n should inherit this smoothing property... How?

Method: modelling the smoothness

Let $\mathbf{f} \in \mathbb{R}^N$ be attention values; interpreted as a function defined on a graph.

Method: modelling the smoothness

Let $\mathbf{f} \in \mathbb{R}^N$ be attention values; interpreted as a function defined on a graph.

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

Method: modelling the smoothness

Let $\mathbf{f} \in \mathbb{R}^N$ be attention values; interpreted as a function defined on a graph.

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

Goal. Output $\mathbf{f}$ with low $\mathcal{E}_D(\mathbf{f})$.

Method: modelling the smoothness

Let $\mathbf{f} \in \mathbb{R}^N$ be attention values; interpreted as a function defined on a graph.

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

Goal. Output $\mathbf{f}$ with low $\mathcal{E}_D(\mathbf{f})$.

Bounding $\mathcal{E}_D\left(\mathbf{f}\right)$.

- $\mathcal{E}_D\left(\mathbf{f}\right)$ is bounded by the Dirichlet energy of previous layers.

Method: modelling the smoothness

Let $\mathbf{f} \in \mathbb{R}^N$ be attention values; interpreted as a function defined on a graph.

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

Goal. Output $\mathbf{f}$ with low $\mathcal{E}_D(\mathbf{f})$.

Bounding $\mathcal{E}_D(\mathbf{f})$.

- $\mathcal{E}_D(\mathbf{f})$ is bounded by the Dirichlet energy of previous layers.
- Consequence: We can act on $\mathbf{f}$ itself and/or on the output of previous layers.

Method: Smooth operator ($\mathtt{Sm}$)

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

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

Method: Smooth operator ($\mathtt{Sm}$)

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

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

Theoretical guarantees. If $\mathbf{L}$ is the normalized Laplacian matrix, then

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

Consequence: It can be used in the different layers of a neural network to decrease $\mathcal{E}_D$.

Method: Smooth operator ($\texttt{Sm}$)

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

$$\texttt{Sm}\left(\mathbf{U}\right) = \left(\mathbf{I} + \gamma \mathbf{L}\right)^{-1} \mathbf{U}.$$

Theoretical guarantees. If $\mathbf{L}$ is the normalized Laplacian matrix, then

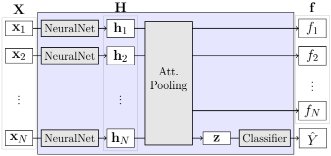
$$\mathcal{E}_D\left(\texttt{Sm}\left(\mathbf{U}\right)\right) < \mathcal{E}_D\left(\mathbf{U}\right).$$

Consequence: It can be used in the different layers of a neural network to decrease $\mathcal{E}_D$.

Avoiding matrix inversion. It holds that

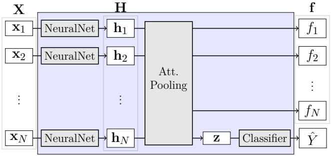
$$\texttt{Sm}\left(\mathbf{U}\right) = \lim_{t \to \infty} \mathbf{G}(t),$$
$$\mathbf{G}(0) = \mathbf{U}; \quad \mathbf{G}(t) = \alpha\left(\mathbf{I} - \mathbf{L}\right)\mathbf{G}(t-1) + (1-\alpha)\,\mathbf{U}.$$

Method: the proposed model

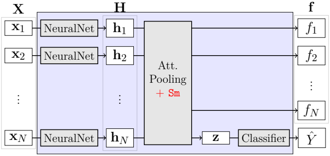


(a) ABMIL, the baseline.

Method: the proposed model

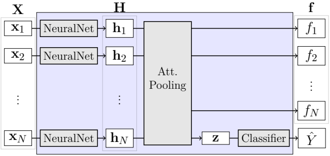


(a) ABMIL, the baseline.

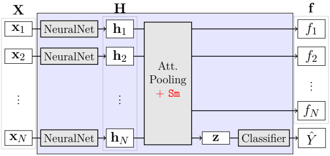


(b) SmAP.

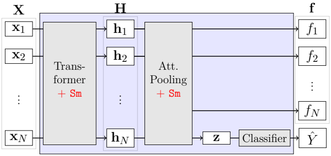
Method: the proposed model



(a) ABMIL, the baseline.



(b) SmAP.



(c) SmTAP.

Experiments

- 3 different medical imaging datasets: RSNA (CT scans), PANDA (WSIs), and CAMELYON16 (WSIs).

Experiments

- 3 different medical imaging datasets: RSNA (CT scans), PANDA (WSIs), and CAMELYON16 (WSIs).
- 4 different feature extractors, with and without self-supervised learning.

Experiments

- 3 different medical imaging datasets: RSNA (CT scans), PANDA (WSIs), and CAMELYON16 (WSIs).
- 4 different feature extractors, with and without self-supervised learning.
- Up to 13 different SOTA methods considered for comparison.

Experiments

- 3 different medical imaging datasets: RSNA (CT scans), PANDA (WSIs), and CAMELYON16 (WSIs).
- 4 different feature extractors, with and without self-supervised learning.
- Up to 13 different SOTA methods considered for comparison.
- Results: the proposed methods with `Sm` achieve the best performance in localization and remain very competitive in classification.

Table: Average rank (lower is better).

		Instance localization	Bag classification
Without global interactions	`SmAP`	$\mathbf{1.500}_{0.548}$	$\mathbf{1.833}_{0.753}$
	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	`SmTAP`	$\mathbf{1.500}_{1.225}$	$\mathbf{1.833}_{0.983}$
	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	$\underline{2.833}_{1.169}$	$\underline{2.750}_{1.173}$

Experiments: WSI visualization.

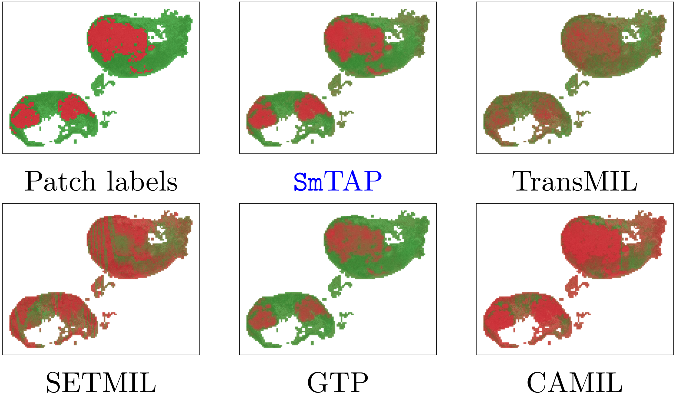


Figure: Attention maps on CAMELYON16. The novel SmTAP produces the most accurate map.

Conclusions

- We draw attention to the localization task: MIL methods need to be evaluated at the instance level.

Conclusions

- We draw attention to the localization task: MIL methods need to be evaluated at the instance level.
- The proposed `Sm` introduces local interactions in a principled way.

Conclusions

- We draw attention to the localization task: MIL methods need to be evaluated at the instance level.
- The proposed `Sm` introduces local interactions in a principled way.
- It achieves the best performance in localization while being highly competitive in classification.

Conclusions

- We draw attention to the localization task: MIL methods need to be evaluated at the instance level.
- The proposed `Sm` introduces local interactions in a principled way.
- It achieves the best performance in localization while being highly competitive in classification.
- Future work: MIL methods need to quantify uncertainty so they can be deployed in clinical settings.

Thank you!