Attention-based Deep Multiple Instance Learning

Maximilian Ilse, Jakub Tomczak, Max Welling

AMLAB, University of Amsterdam

ICML 2018

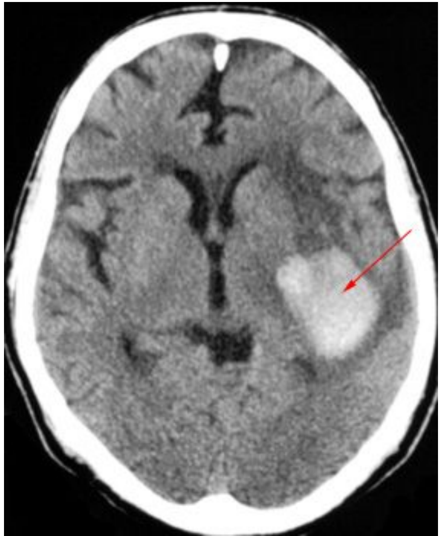
Motivation

Typical size of benchmark natural images: **up to 256x256**

Motivation

Typical size of benchmark natural images: **up to 256x256**

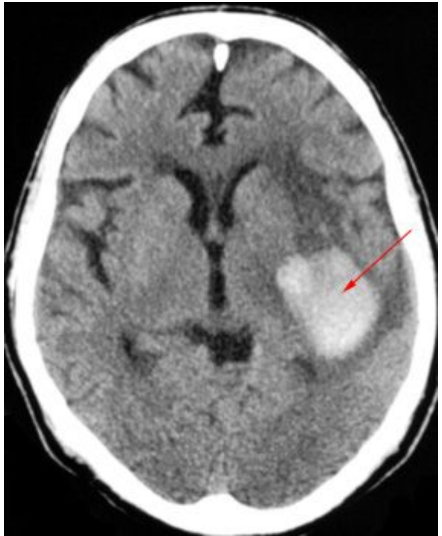
Typical size of medical images: **~10,000x10,000**



Motivation

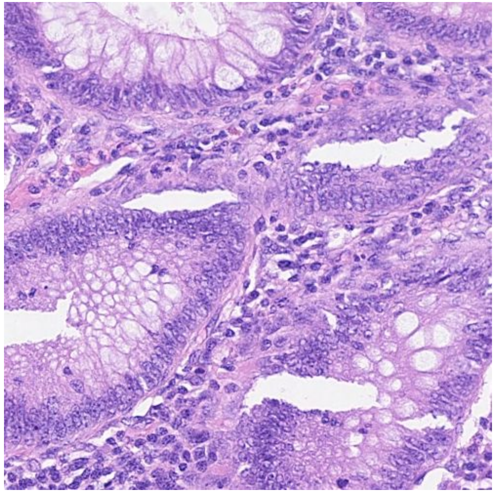
Typical size of benchmark natural images: **up to 256x256**

Typical size of medical images: **~10,000x10,000**



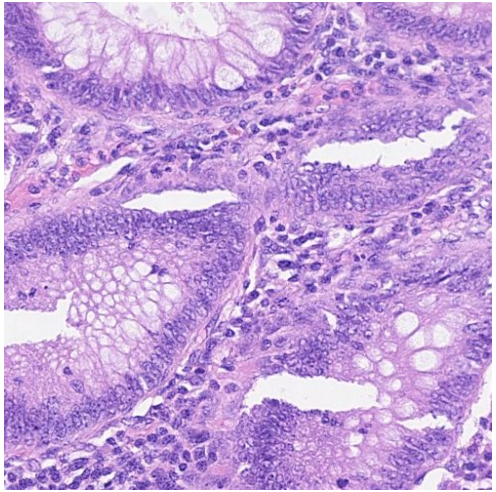
How to process it?

Motivation



Goal: Find (local) objects (abnormal changes in tissue) in an image.

Motivation



Goal: Find (local) objects (abnormal changes in tissue) in an image.

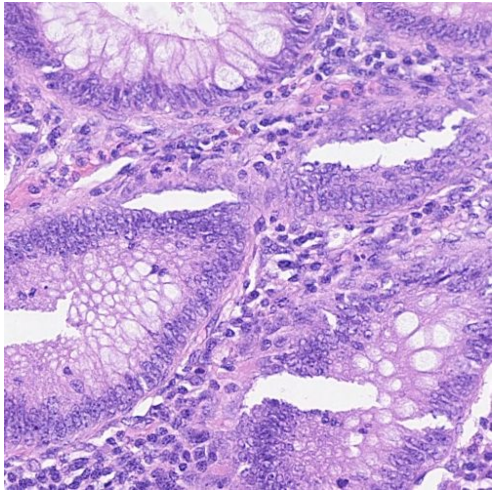
Data: billions of pixels, 10^1-10^2 scans, weak labels (for regions or a scan).

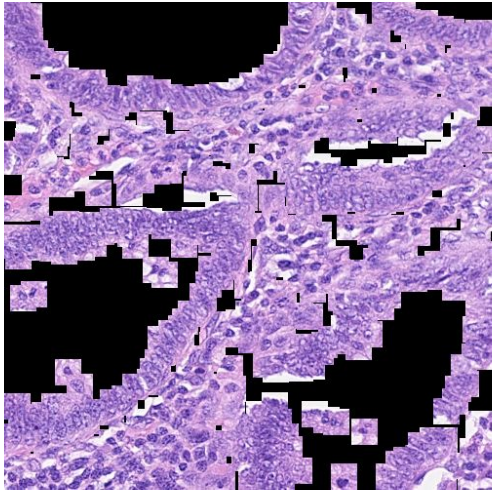
Motivation

Goal: Find (local) objects (abnormal changes in tissue) in an image.

Data: billions of pixels, 10^1-10^2 scans, weak labels (for regions or a scan).

Solution: Use local information in the image and look for Regions of Interest.



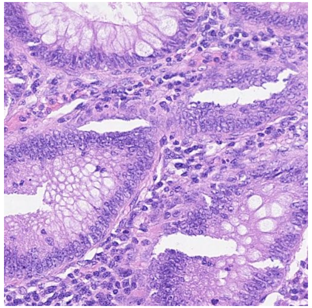


Ricci-Vitiani, L., et al. "Identification and expansion of human colon-cancer-initiating cells." *Nature* 445.7123 (2007): 111.

Supervised Learning vs. Multiple Instance Learning

One image - **one label**

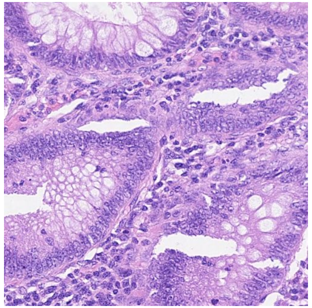
$$\mathbf{x} \in \mathbb{R}^D, \quad y \in \{0, 1\}$$



Supervised Learning vs. Multiple Instance Learning

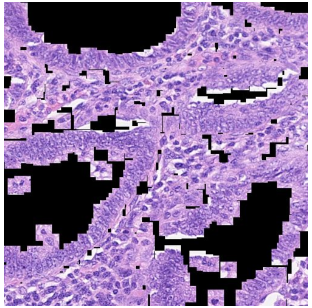
One image - **one label**

$$\mathbf{x} \in \mathbb{R}^D, \quad y \in \{0, 1\}$$



Many images - **one label**

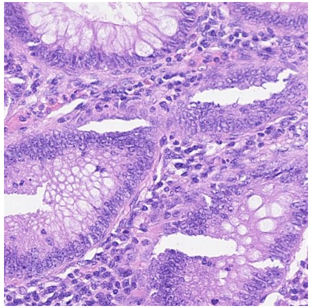
$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\},$$
$$Y \in \{0, 1\}$$



Supervised Learning vs. Multiple Instance Learning

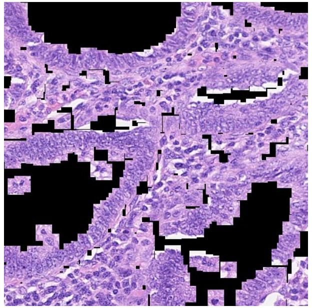
One image - one label

$$\mathbf{x} \in \mathbb{R}^D, \quad y \in \{0, 1\}$$



Many images - one label

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\},$$
$$Y \in \{0, 1\}$$



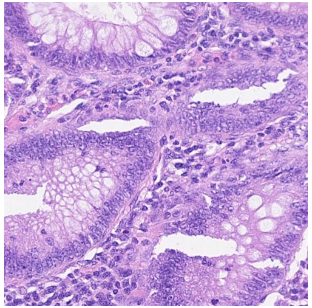
Individual labels:

$\{y_1, \ldots, y_K\}$ are **unknown**.

Supervised Learning vs. Multiple Instance Learning

One image - **one label**

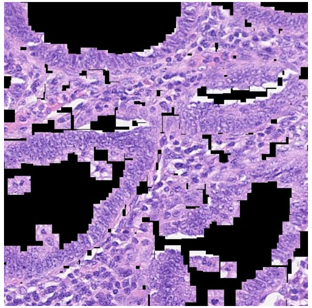
$$\mathbf{x} \in \mathbb{R}^D, \quad y \in \{0, 1\}$$



Many images - **one label**

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\},$$

$$Y \in \{0, 1\}$$



Individual labels:

$\{y_1, \ldots, y_K\}$ are **unknown**.

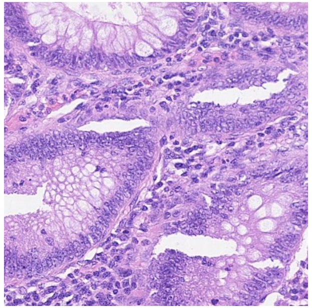
Assumptions about the label Y:

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

Supervised Learning vs. Multiple Instance Learning

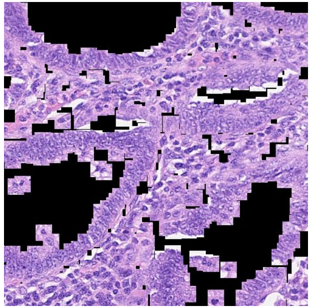
One image - **one label**

$$\mathbf{x} \in \mathbb{R}^D, \quad y \in \{0, 1\}$$



Many images - **one label**

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\},$$
$$Y \in \{0, 1\}$$



Individual labels:

$\{y_1, \ldots, y_K\}$ are **unknown**.

Instances with (y_k = 1) = **key instances**

Assumptions about the label Y:

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

Multiple Instance Learning

A MIL classifier as a probabilistic model:

$$p(Y|X) = \theta(X)^{Y} \ \big(1 - \theta(X)\big)^{1-Y}$$

Multiple Instance Learning

A MIL classifier as a probabilistic model:

$$p(Y|X) = \theta(X)^{Y} \ \left(1 - \theta(X)\right)^{1-Y}$$

Must be **permutation-invariant**!

Multiple Instance Learning

A MIL classifier as a probabilistic model:

$$p(Y|X) = \theta(X)^{Y}\ \left(1 - \theta(X)\right)^{1-Y}$$

Must be **permutation-invariant**!

How?

Multiple Instance Learning

A MIL classifier as a probabilistic model:

$$p(Y|X) = \theta(X)^Y \left(1 - \theta(X)\right)^{1-Y}$$

Theorem (Zaheer et al., 2017)
A scoring function for a set of instances X, $S(X) \in \mathbb{R}$, is a symmetric function (i.e., permutation invariant to the elements in X), if and only if it can be decomposed in the following form:

$$S(X) = g\left(\sum_{\mathbf{x} \in X} f(\mathbf{x})\right)$$

where f and g are suitable transformations.

Multiple Instance Learning

A MIL classifier as a probabilistic model:

$$p(Y|X) = \theta(X)^Y \left(1 - \theta(X)\right)^{1-Y}$$

Theorem (Qi et al., 2017)
For any $\varepsilon > 0$, *a Hausdorff continuous symmetric function* $S(X) \in \mathbb{R}$ *can be arbitrarily approximated by a function in the form* $g(\max_{\boldsymbol{x} \in X} f(\boldsymbol{x}))$, *where* $\max$ *is the element-wise vector maximum operator and* f *and* g *are continuous functions, that is:*

$$|S(X) - g(\max_{\boldsymbol{x} \in X} f(\boldsymbol{x}))| < \varepsilon.$$

Multiple Instance Learning

A MIL classifier as a probabilistic model:

$$p(Y|X) = \theta(X)^Y \left(1 - \theta(X)\right)^{1-Y}$$

The theorems say that we can model a **permutation-invariant** $\theta(X)$ by composing:

- a transformation f of individual instances,
- a permutation-invariant function σ, *e.g.*, sum, mean or max (**MIL pooling**),
- a transformation of combined instances using a function g:

$$\theta(X) = g(\sigma(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_K)))$$

Multiple Instance Learning: Components

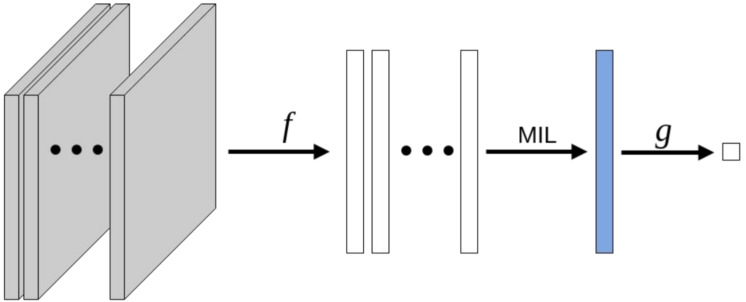
We model both transformations f and g using **neural networks**.

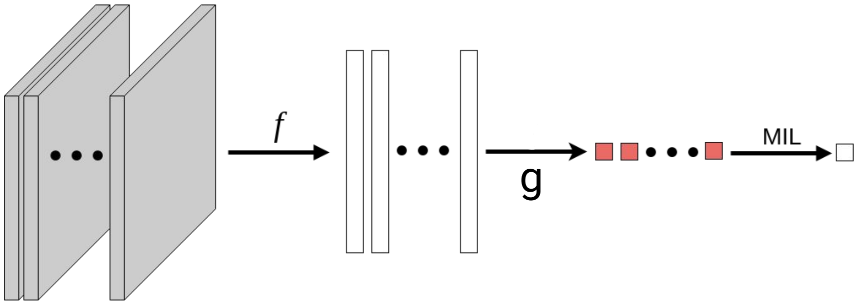
Multiple Instance Learning: Components

We model both transformations f and g using **neural networks**.

Two approaches:

- **embedded-based**
- **instance-based**



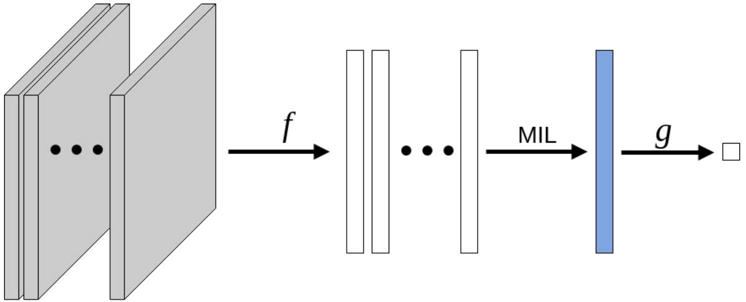


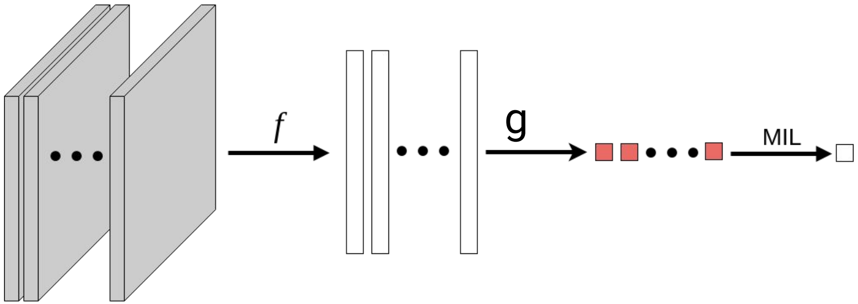
Multiple Instance Learning: Components

We model both transformations f and g using **neural networks**.

Two approaches:

- **embedded-based**
- **instance-based**





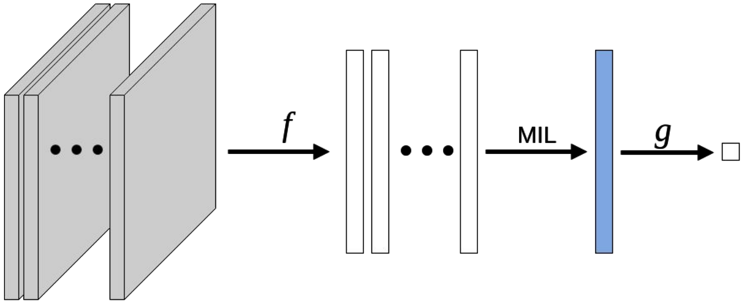
MIL pooling:

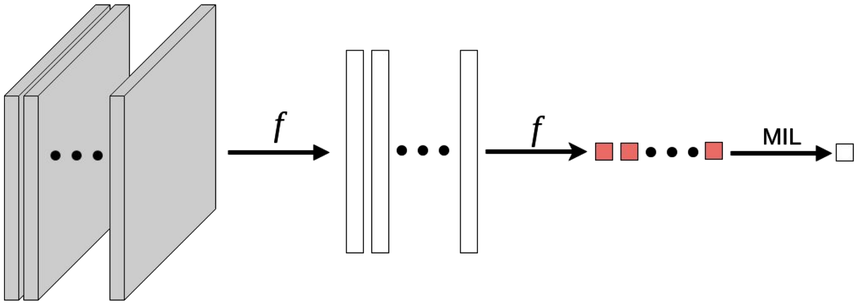
- mean,
- max,
- other (*e.g.*, Noisy-Or).

Multiple Instance Learning: Components

Issues:

- Embedded-based approach **lacks interpretability**.
- Instance-based approach **propagates error**.
- $\max$ and mean are **non-learnable**.





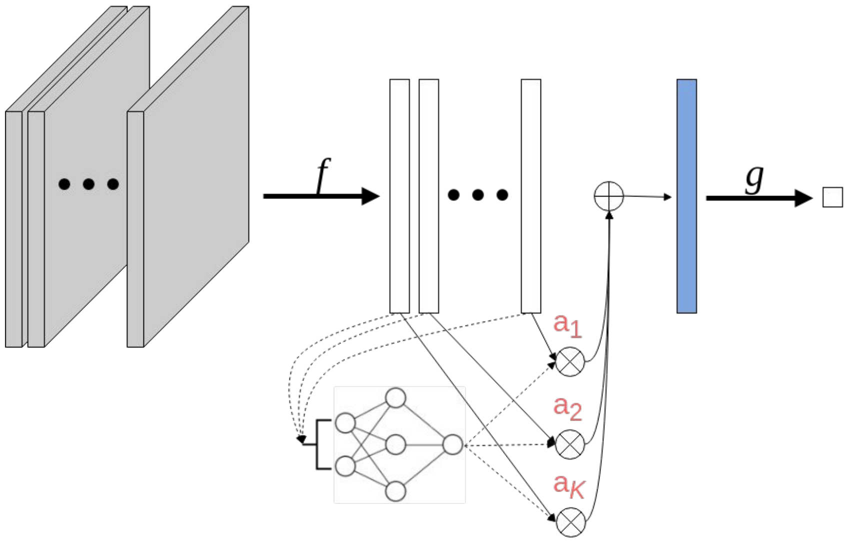
Multiple Instance Learning: Attention-based approach

We propose to use the attention mechanism as **MIL pooling**:

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

where:

$$a_k = \frac{\exp\{\mathbf{w}_k^\top \tanh\left(\mathbf{V}\mathbf{h}_k^\top\right)\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}_j^\top \tanh\left(\mathbf{V}\mathbf{h}_j^\top\right)\}},$$



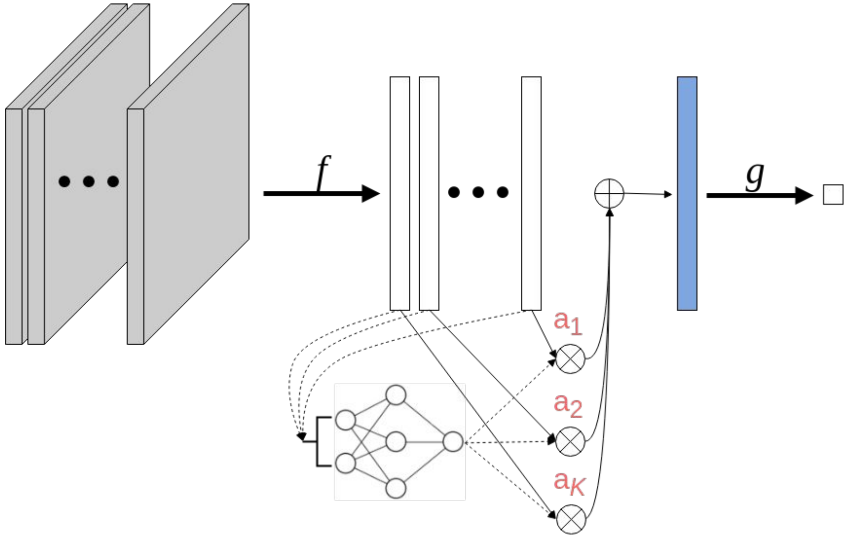
Multiple Instance Learning: Attention-based approach

We propose to use the attention mechanism as **MIL pooling**:

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

where:

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

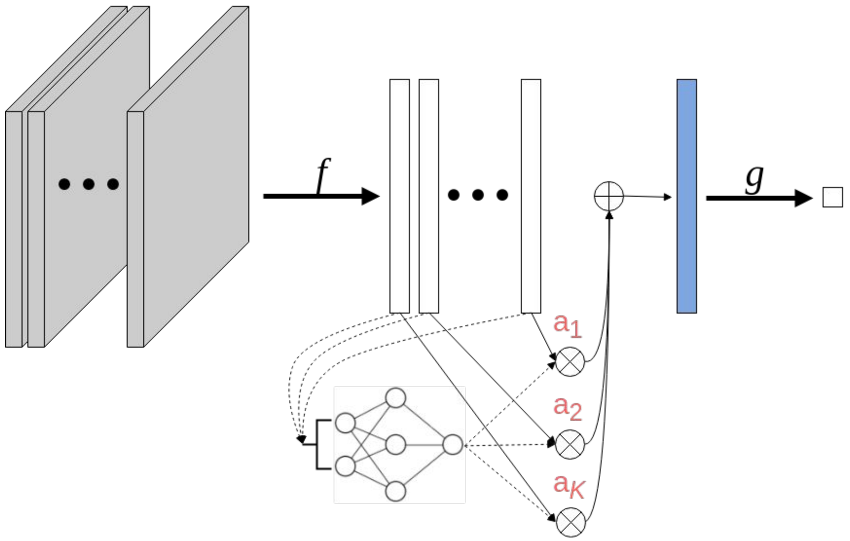


*attention with **gating mechanism***

Multiple Instance Learning: Attention-based approach

The attention mechanism as **MIL pooling**:

- MIL operator is **trainable**;
- attention weights could be **interpreted (key instances)**.



Embedded-based approach

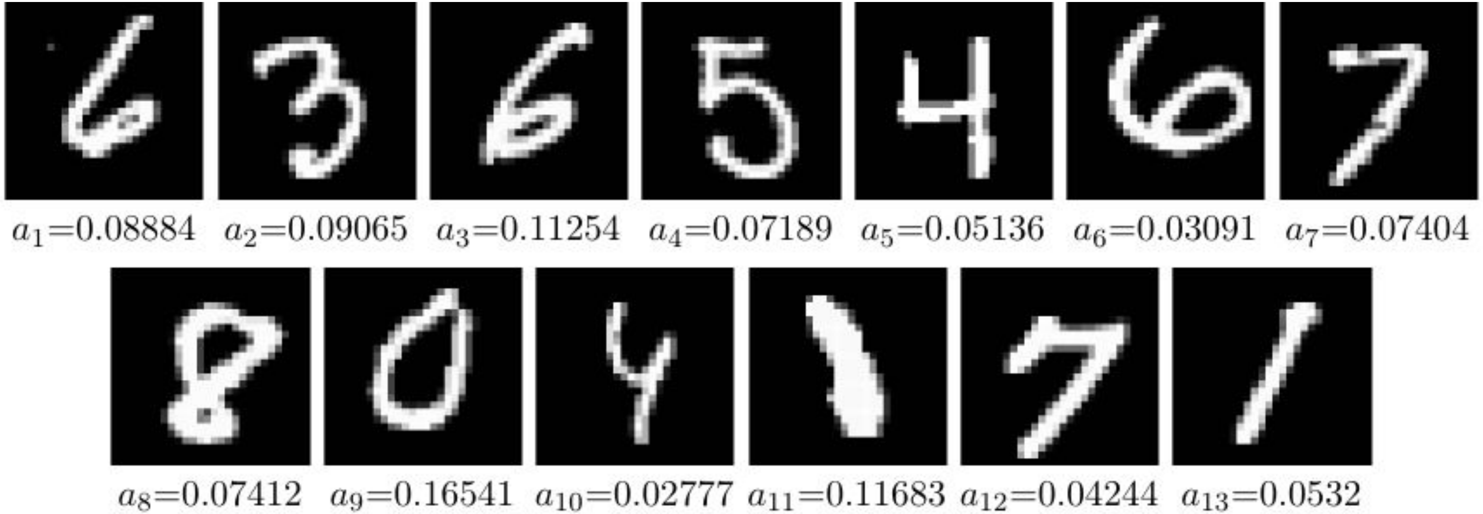
is **interpretable** and fully **trainable**.

Experiments: MNIST-based problem

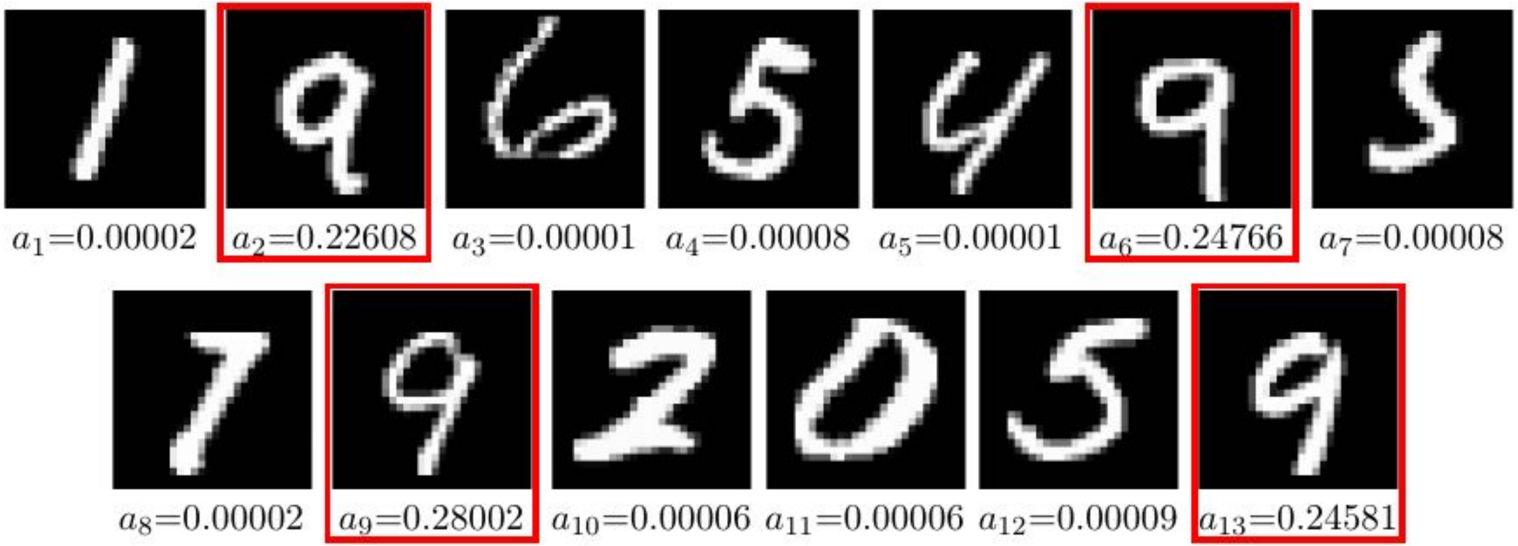
$Y = 0$

$Y = 1$

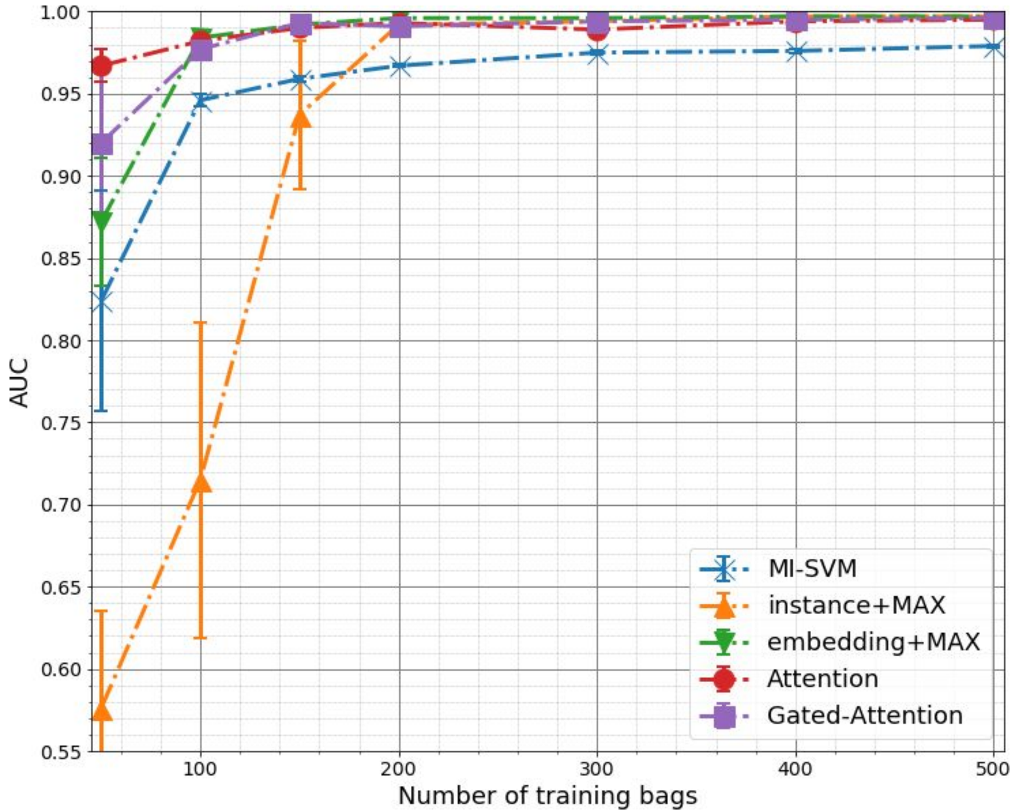
Experiments: MNIST-based problem



Experiments: MNIST-based problem



Experiments: MNIST-based problem



Experiments: Breast Cancer

METHOD	ACCURACY	PRECISION	RECALL	F-SCORE	AUC
Instance+max	0.614±0.020	0.585±0.03	0.477±0.087	0.506±0.054	0.612±0.026
Instance+mean	0.672±0.026	0.672±0.034	0.515±0.056	0.577±0.049	0.719±0.019
Embedding+max	0.607±0.015	0.558±0.013	0.546±0.070	0.543±0.042	0.650±0.013
Embedding+mean	**0.741**±0.023	**0.741**±0.023	0.654±0.054	0.689±0.034	**0.796**±0.012
Attention	**0.745**±0.018	0.718±0.021	**0.715**±0.046	**0.712**±0.025	0.775±0.016
Gated-Attention	**0.755**±0.016	**0.728**±0.016	**0.731**±0.042	**0.725**±0.023	**0.799**±0.020

Experiments: Colon Cancer

METHOD	ACCURACY	PRECISION	RECALL	F-SCORE	AUC
Instance+max	0.842 ± 0.021	0.866 ± 0.017	0.816 ± 0.031	0.839 ± 0.023	0.914 ± 0.010
Instance+mean	0.772 ± 0.012	0.821 ± 0.011	0.710 ± 0.031	0.759 ± 0.017	0.866 ± 0.008
Embedding+max	0.824 ± 0.015	0.884 ± 0.014	0.753 ± 0.020	0.813 ± 0.017	0.918 ± 0.010
Embedding+mean	0.860 ± 0.014	0.911 ± 0.011	0.804 ± 0.027	0.853 ± 0.016	0.940 ± 0.010
Attention	$\mathbf{0.904} \pm 0.011$	$\mathbf{0.953} \pm 0.014$	$\mathbf{0.855} \pm 0.017$	$\mathbf{0.901} \pm 0.011$	$\mathbf{0.968} \pm 0.009$
Gated-Attention	$\mathbf{0.898} \pm 0.020$	$\mathbf{0.944} \pm 0.016$	$\mathbf{0.851} \pm 0.035$	$\mathbf{0.893} \pm 0.022$	$\mathbf{0.968} \pm 0.010$

Experiments: Colon Cancer

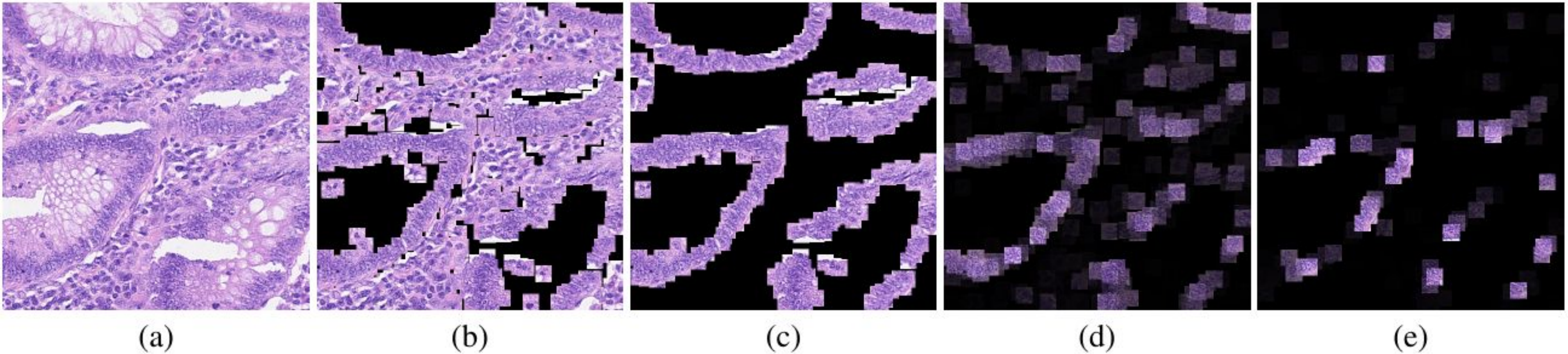
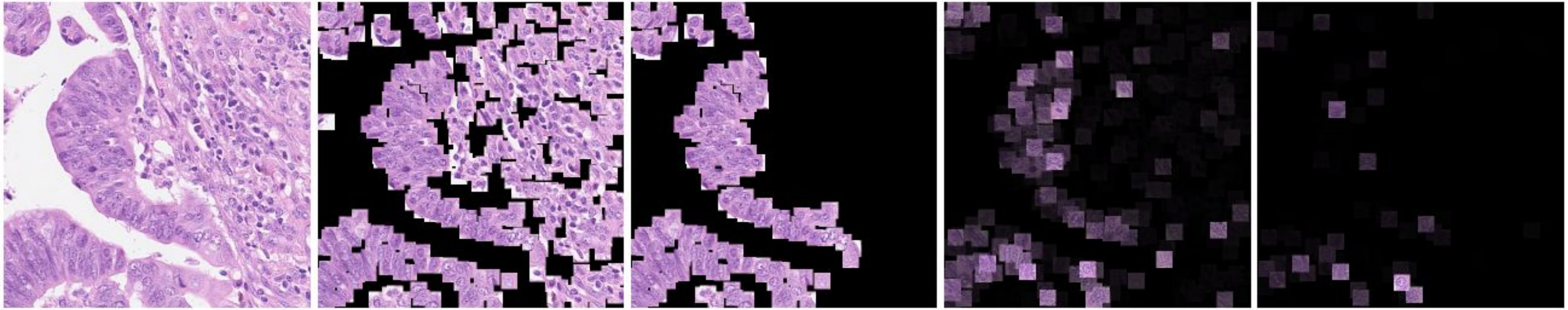


Figure 10. Colon cancer example 1: (a) H&E stained histopathology image. (b) 27×27 patches centered around all marked nuclei. (c) Ground truth: Patches that belong to the class epithelial. (d) Attention heatmap: Every patch from (b) multiplied by its attention weight. (e) Instance+max heatmap: Every patch from (b) multiplied by its score from the Instance+max model. We rescaled the attention weights and instance scores using $a'_k = a_k - \min(\mathbf{a})/(\max(\mathbf{a}) - \min(\mathbf{a}))$.

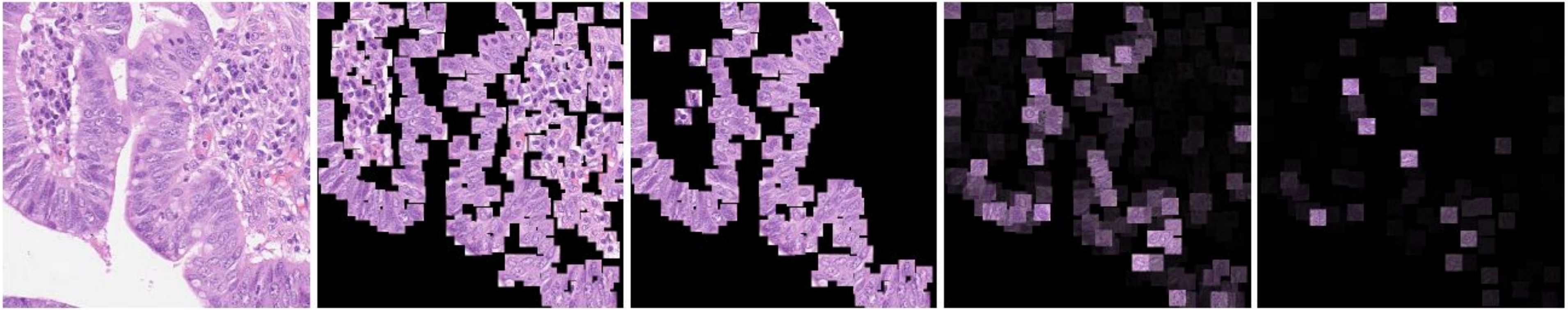
Experiments: Colon Cancer



(a) (b) (c) (d) (e)

Figure 11. Colon cancer example 2: (a) H&E stained histopathology image. (b) 27×27 patches centered around all marked nuclei. (c) Ground truth: Patches that belong to the class epithelial. (d) Attention heatmap: Every patch from (b) multiplied by its attention weight. (e) Instance+max heatmap: Every patch from (b) multiplied by its score from the Instance+max model. We rescaled the attention weights and instance scores using $a'_k = a_k - \min(\mathbf{a})/(\max(\mathbf{a}) - \min(\mathbf{a}))$.

Experiments: Colon Cancer



(a) (b) (c) (d) (e)

Figure 12. Colon cancer example 3: (a) H&E stained histopathology image. (b) 27×27 patches centered around all marked nuclei. (c) Ground truth: Patches that belong to the class epithelial. (d) Attention heatmap: Every patch from (b) multiplied by its attention weight. (e) Instance+max heatmap: Every patch from (b) multiplied by its score from the Instance+max model. We rescaled the attention weights and instance scores using $a'_k = a_k - \min(\mathbf{a})/(\max(\mathbf{a}) - \min(\mathbf{a}))$.

Conclusion

Deep MIL: a flexible approach to cope with large images.

Attention mechanism: interpretable and **learnable** MIL pooling.

Next step: Application to **whole-slide classification.**

Next step: **taking into account spatial dependencies (non i.i.d. instances).**

Conclusion

Deep MIL: a flexible approach to cope with large images.

Attention mechanism: interpretable and **learnable** MIL pooling.

Next step: Application to **whole-slide classification.**

Next step: **taking into account spatial dependencies (non i.i.d. instances).**

Conclusion

Deep MIL: a flexible approach to cope with large images.

Attention mechanism: interpretable and **learnable** MIL pooling.

Next step: Application to **whole-slide classification.**

Next step: **taking into account spatial dependencies (non i.i.d. instances).**

Conclusion

Deep MIL: a flexible approach to cope with large images.

Attention mechanism: **interpretable** and **learnable** MIL pooling.

Next step: Application to **whole-slide classification**.

Next step: **taking into account spatial dependencies (non i.i.d. instances).**

Conclusion

Deep MIL: a flexible approach to cope with large images.

Attention mechanism: interpretable and **learnable** MIL pooling.

Next step: Application to **whole-slide classification**.

Next step: **taking into account spatial dependencies (non i.i.d. instances)**.

Code on github:

https://github.com/AMLab-Amsterdam/AttentionDeepMIL

Contact:

ilse.maximilian@gmail.com

jakubmkt@gmail.com

The research conducted by Maximilian Ilse was funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Grant DLMedIa: Deep Learning for Medical Image Analysis).

Marie Skłodowska-Curie Actions

The research conducted by Jakub M. Tomczak was funded by the European Commission within the Marie Skłodowska-Curie Individual Fellowship (Grant No. 702666, ”Deep learning and Bayesian inference for medical imaging”).