Processing Megapixel Images with Deep Attention-Sampling Models

Angelos Katharopoulos[1,2] François Fleuret[1,2]
[1]Idiap Research Institute [2]École Polytechnique Fédérale de Lausanne

Funded by FNSNF

http://attention-sampling.com/

Processing Large Images with DNNs

Common pitfalls:

- **Downsampling** results in **loss of useful information**
- **Processing only parts** of the image **requires per-part annotations**
- **Attention** has been shown to overcome the need for per-part annotations, however **processing the whole image** is still required (Ilse et al. 2018)

The speed limit is unrecognizable in low resolution

We propose a **fully differentiable** end-to-end trainable model that processes only **a fraction of the input** by **sampling from an attention distribution** computed in low resolution.

Attention Sampling

Given a sample x, the output of the neural network $\Psi(x;\Theta)$ that uses features $f(x;\Theta) \in \mathbb{R}^{K\times D}$ and attention $a(x;\Theta) \in \mathbb{R}_+^K$ is

$$\Psi(x;\Theta) = g\left(\sum_{i=1}^{K} a(x;\Theta)_i f(x;\Theta)_i\right) = g\left(\mathbb{E}_{I\sim a(x;\Theta)}[f(x;\Theta)_I]\right).$$

We avoid computing $f(x)_i \,\forall i$ by **sampling a set of feature indices from the attention distribution**, $Q = \{q_i \sim a(x) \,|\, i \in \{1,2,\dots,N\}\}$ and approximate the output as

$$\Psi(x;\Theta) \approx g\left(\frac{1}{N}\sum_{q\in Q} f(x;\Theta)_q\right)$$

We show that for a fixed feature norm, namely $\|f(x)_i\|_2 = \|f(x)_j\|_2 \;\forall i,j$ our estimator is the **minimum variance approximation** of $\Psi(x)$.

Deriving Gradients

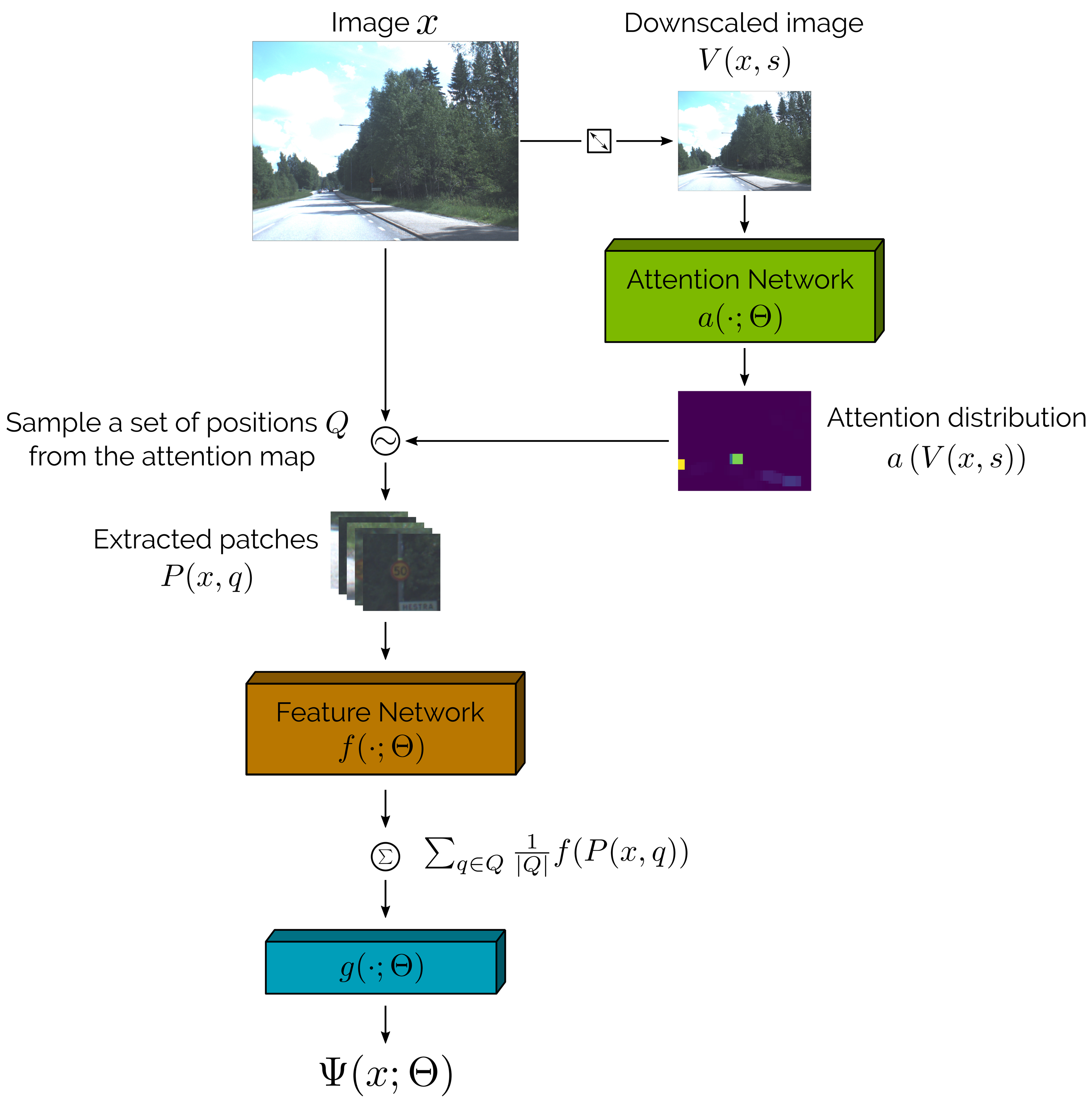
To train the network **we need to compute gradients** with respect to the parameters of the attention and the feature functions.

For every parameter $\theta \in \Theta$, even the ones affecting $a(\cdot)$, we show that the gradient is:

$$\frac{\partial}{\partial\theta}\frac{1}{N}\sum_{q\in Q} f(x;\Theta)_q = \frac{1}{N}\sum_{q\in Q}\frac{\frac{\partial}{\partial\theta}\left[a(x;\Theta)_q f(x;\Theta)_q\right]}{a(x;\Theta)_q}$$

Attention Sampling for Images

Computing the attention in low resolution and features only for some parts of the image based on the attention distribution results in **an order of magnitude lower memory use and faster computation**.



Experiments

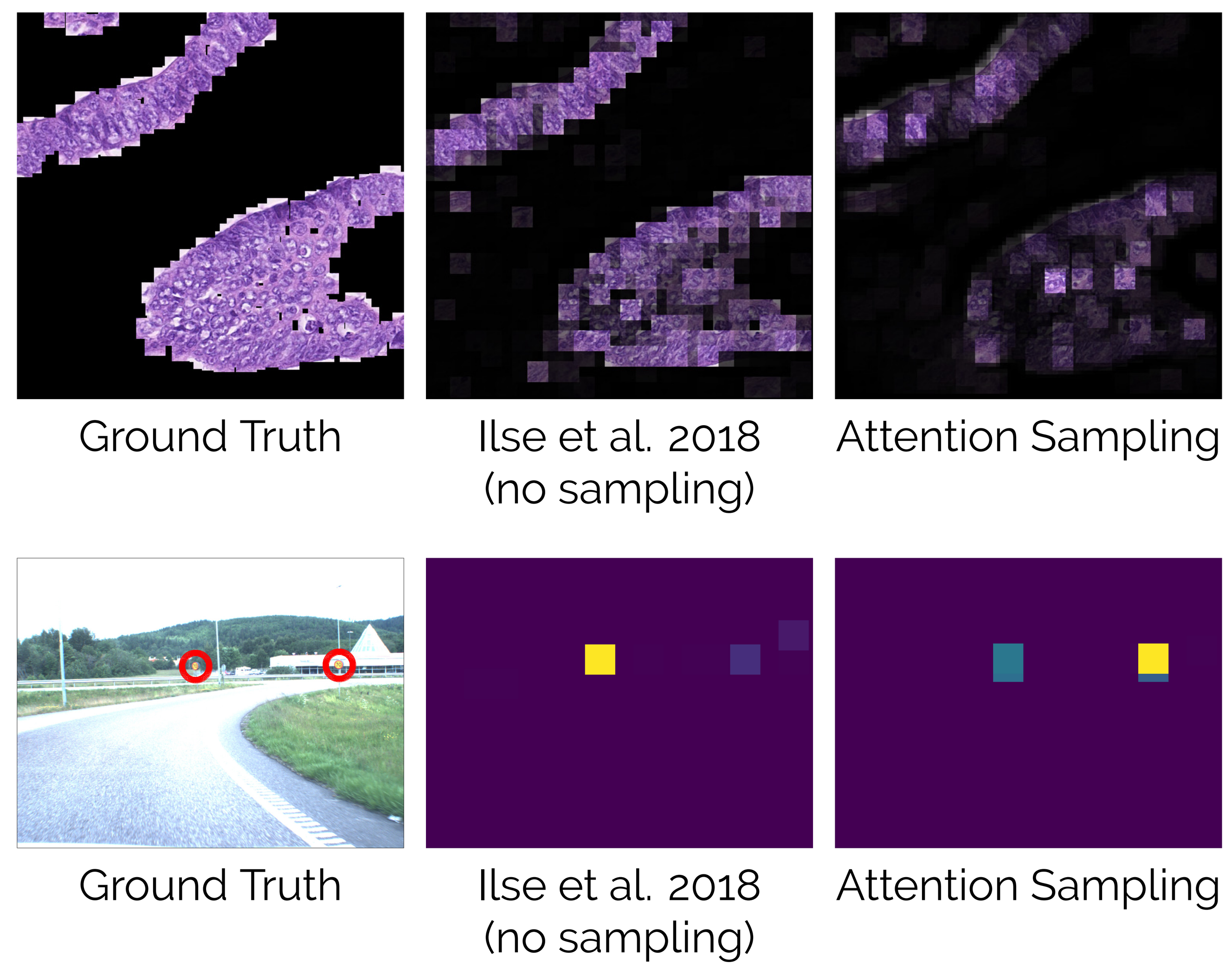
Baselines

- Attention-based Deep Multiple Instance Learning (Ilse et al. 2018) that computes the attention from the per patch features
- Shallow ResNets at various input scales, denoted below as CNN

Datasets

- Histopathology dataset for detecting images that contain epithelial cells
- Speed limit sign detection, adapted from the Swedish traffic signs dataset, for detecting and classifying the speed limit in the image

Qualitative evaluation of attention sampling



Quantitative evaluation of attention sampling

Histopathology

Method	Scale	Test Error	Time/sample	Memory/sample
CNN	0.5	0.104 ± 0.009	4.8 ms	65 MB
CNN	1	0.092 ± 0.012	18.7 ms	250 MB
Ilse et al. 2018	1	0.093 ± 0.004	48.5 ms	644 MB
ATS (ours)	**0.2/1**	**0.093** ± **0.014**	**1.8 ms**	**21 MB**

Speed Limits

Method	Scale	Test Error	Time/sample	Memory/sample
CNN	0.3	0.311 ± 0.049	6.6 ms	86 MB
CNN	0.5	0.295 ± 0.039	15.6 ms	239 MB
CNN	1	0.247 ± 0.001	64.2 ms	958 MB
Ilse et al. 2018	1	0.083 ± 0.006	97.2 ms	1,497 MB
ATS (ours)	**0.3/1**	**0.089** ± **0.002**	**8.5 ms**	**86 MB**