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).

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

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

We avoid computing $f(x)_i \forall i$ by sampling a set of feature indices from the attention distribution. $Q = \{ q_i \sim a(x) \mid i \in \{1, 2, \ldots, N\} \}$ 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:

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

### 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.

#### Quantitative evaluation of attention sampling

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale</th>
<th>Test Error</th>
<th>Time/sample</th>
<th>Memory/sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>0.5</td>
<td>0.049</td>
<td>4.8 ms</td>
<td>65 MB</td>
</tr>
<tr>
<td>CNN</td>
<td>1</td>
<td>0.059</td>
<td>19.7 ms</td>
<td>250 MB</td>
</tr>
<tr>
<td>Ilse et al.</td>
<td>1</td>
<td>0.093</td>
<td>18.8 ms</td>
<td>644 MB</td>
</tr>
<tr>
<td>ATS (ours)</td>
<td>0.2/1</td>
<td>0.093</td>
<td>1.8 ms</td>
<td>21 MB</td>
</tr>
</tbody>
</table>

#### Speed Limits

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale</th>
<th>Test Error</th>
<th>Time/sample</th>
<th>Memory/sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>0.3</td>
<td>0.311</td>
<td>6.6 ms</td>
<td>86 MB</td>
</tr>
<tr>
<td>CNN</td>
<td>0.5</td>
<td>0.295</td>
<td>15.6 ms</td>
<td>239 MB</td>
</tr>
<tr>
<td>CNN</td>
<td>1</td>
<td>0.247</td>
<td>64.2 ms</td>
<td>958 MB</td>
</tr>
<tr>
<td>Ilse et al.</td>
<td>1</td>
<td>0.083</td>
<td>97.2 ms</td>
<td>1,497 MB</td>
</tr>
<tr>
<td>ATS (ours)</td>
<td>0.3/1</td>
<td>0.089</td>
<td>8.5 ms</td>
<td>86 MB</td>
</tr>
</tbody>
</table>