Supplementary material: Sparse-shot Learning with Exclusive Cross-Entropy for Extremely Many Localisations

Andreas Panteli\textsuperscript{1,2}, Jonas Teuwen\textsuperscript{1,2,3}, Hugo Horlings\textsuperscript{1} and Efstratios Gavves\textsuperscript{2,4}
\textsuperscript{1}Netherlands Cancer Institute, \textsuperscript{2}University of Amsterdam, \textsuperscript{3}Radboud University Medical Center, \textsuperscript{4}Ellogon.AI
\{a.panteli, j.teuwen, h.horlings\}@nki.nl, egavves@uva.nl

1. Second derivative

We start from our cross entropy loss function $\mathcal{L}$, which we divide in two loss terms

$$
\mathcal{L} = \mathcal{L}_F + \mathcal{L}_{\overline{F}} = \sum_F \log p + \sum_{\overline{F}} \log(1 - p)
$$

(1)

As a reminder, for convenience we set $p = p(y = 1|x)$ and $1 - p = p(y = 1|x)$ given a binary classification problem. The $\mathcal{L}_F$ corresponds to the loss on the foreground area $F$ for which we have manual annotations, and $\mathcal{L}_{\overline{F}}$ corresponds to the loss on the rest area $\overline{F}$. As we do not have the true annotations for $\overline{F}$, we cannot really compute the loss $\mathcal{L}_{\overline{F}}$, at least not accurately. As usual, we rely on stochastic gradient descent for optimising the model parameters, that is

$$
w_{t+1} = w_t - \varepsilon \frac{d\mathcal{L}}{dw}_t \
\frac{dw}{dt} = -\varepsilon \frac{d\mathcal{L}}{dw}_t,
$$

(2)

where we have approximated the discrete change in weights over two subsequent time steps $w_{t+1} - w_t$ with the continuous derivative $\frac{dw}{dt}$.

The derivative of the loss with respect to the weights is

$$
\frac{d\mathcal{L}}{dw} = \sum_F \frac{1}{p} \frac{dp}{dw} + \sum_{\overline{F}} \frac{1}{1 - p} \frac{dp}{dw}
$$

$$
= \sum_F \frac{1}{p} \frac{dp}{dw} - \sum_{\overline{F}} \frac{1}{p} \frac{dp}{dw}
$$

(3)

where $k$ is an index running over all the dimensions of the input sample $x \in I$, to the sigmoidal output neuron. As a reminder, the derivative of the sigmoid with respect to its inputs is $\frac{d\sigma(z)}{dz} = \sigma(z)(1 - \sigma(z))$. Replacing equation (4) to (5) and focusing on the derivative with respect to the $j$-th weight, we have

$$
\frac{d\mathcal{L}}{dw_j} = \frac{d\mathcal{L}_F}{dw_j} + \frac{d\mathcal{L}_{\overline{F}}}{dw_j}
$$

$$
= \sum_F \frac{1}{p} \frac{dp}{dw_j} \frac{d(\sigma(\sum_k w_k x_k))}{dw_j} - \sum_F \frac{1}{1 - p} \frac{dp}{dw_j} \frac{d(\sigma(\sum_k w_k x_k))}{dw_j}
$$

$$
= \sum_F \frac{1}{p} \frac{dp}{dw_j} \left( \sum_k x_j \sigma(\sum_k w_k x_k) \right)
$$

$$
\cdot \left( 1 - \sigma(\sum_k w_k x_k) \right)
$$

(5)

$$
- \sum_{\overline{F}} \frac{1}{1 - p} \frac{dp}{dw_j} \left( \sum_k x_j \sigma(\sum_k w_k x_k) \right)
$$

$$
\cdot \left( 1 - \sigma(\sum_k w_k x_k) \right)
$$

$$
= \sum_F x_j (1 - p) - \sum_{\overline{F}} x_j p
$$

Next, we want to examine what is the dynamics of learning, as well as the speed of learning. We associate the dynamics of learning with $\frac{d\mathcal{L}}{dt}$, in that the derivative with respect to time indicates how the loss decreases with time and the learning improves with time. Then, the speed of learning is associated with the second derivative $\frac{d^2\mathcal{L}}{dt^2}$. As a side note, the more frequently appearing derivative with respect to weights, $\frac{d\mathcal{L}}{dw}$, indicates the optimal direction for learning but not the dynamics of learning. Since we are more interested in the unannotated area $\overline{F}$, we will focus only on the respective terms. The same derivations can be made for the other terms also.
Using equation (2) the first derivative with respect to time is
\[
\frac{d L_{\mathcal{T}}}{dt} = \sum_k \frac{d L_{\mathcal{T}}}{dw_k} \frac{dw_k}{dt} \\
= \sum_k \frac{d L_{\mathcal{T}}}{dw_k} (-\varepsilon \frac{d L_{\mathcal{T}}}{dw_k}) \\
= -\varepsilon \sum_k (\frac{d L_{\mathcal{T}}}{dw_k})^2
\] (6)

Using equations (6) and the term in (5) that corresponds to $\mathcal{T}$, and dropping the $\sum_{\mathcal{T}}$ for notation clarity (the total result is the sum over all samples in $\mathcal{T}$), the second derivative is then
\[
\frac{d^2 L_{\mathcal{T}}}{dt^2} = \frac{d}{dt} \left[ -\varepsilon \sum_k \left( \frac{d L_{\mathcal{T}}}{dw_k} \right)^2 \right] \\
= -\varepsilon \sum_k \frac{d L_{\mathcal{T}}}{dw_k} \frac{d}{dt} \left( \frac{d L_{\mathcal{T}}}{dw_k} \right) \\
= -2\varepsilon \sum_k (-x_k p) \frac{d}{dt}(x_k p) \\
= -2\varepsilon \sum_k x_k^2 p^2 (1-p) \frac{d}{dt} \sum_r w_r x_r \\
= -2\varepsilon \sum_k x_k^2 p^2 (1-p) \frac{d}{dt} \sum_r w_r x_r
\] (7)

We know that
\[
\frac{dw_r}{dt} = -\varepsilon \frac{d L_{\mathcal{T}}}{dw_r} \\
= -\varepsilon (-x_r p) = \varepsilon x_r p
\] (8)

By combining equations (7) and (8), we obtain that
\[
\frac{d^2 L_{\mathcal{T}}}{dt^2} = -2\varepsilon \sum_k x_k^2 p^2 (1-p) \sum_r x_r^2 p \\
= -2\varepsilon p^3(1-p) \sum_{k,r} x_k^2 x_r^2
\] (9)

where $m$ and $n$, $m > n$, indicate integers powers forming a polynomial equation for output probabilities $p$ and $1-p$ as roots.

In figure 1 we show that unlike standard cross entropy for exclusive cross entropy the $\frac{d^2 L_{\mathcal{T}}}{dt^2}$ is close to zero, effectively reducing the speed of learning for the background and delaying biased gradients.

2. Dataset information

In table 1 an overview of all datasets can be observed. All datasets except the TIL localisation dataset are originally exhaustively annotated. All datasets except WBC-NuClick contain areas from H&E stained slides and from various cancer types. The WBC-NuClick dataset, contains white blood cells in blood sample images synthetically generated for cell segmentation [2]. In addition, there exists data overlap between dataset Kumar and MoNuSeg, due to the fact that Kumar was later enhanced with additional data and has functioned into a benchmark challenge dataset online [1].

For the TNBC dataset, the split decision for training, validation and testing sets was done on a slide level, avoiding mixing images of the same slide in different sets. For our TIL localisation dataset, a similar decision was made on a patient level to aver overlap between testing and training sets. For datasets CoNSep, CPM15, CPM17, Kumar, WBC-NuClick, the testing and training splits were available, but not the validation. In this case, 30% of the training set was used as an independent validation set. Datasets CRCHisto, TNBC contained no split information and a 60-20-20 split was performed for the training, validation and testing sets respectively. For the MoNuSeg dataset, all split sets were provided.

Pre-processing For datasets TNBC and WBC-NuClick because there is only foreground level information, cells are indistinguishable from each other in the mask, a watershed method is applied to segment the datasets into separate cells [1]. For every image, we calculate local maximum points for the foreground areas. We define a specific cell size for each dataset: a circle with 25 and 70 pixels in diameter for

https://monuseg.grand-challenge.org/Data/
Table 1: Summary information of datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CoNSeP</th>
<th>CPM15</th>
<th>CPM17</th>
<th>CRCHisto</th>
<th>Kumar</th>
<th>MoNuSeg</th>
<th>WBC-N.</th>
<th>TNBC</th>
<th>TIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of images</td>
<td>41</td>
<td>15</td>
<td>74</td>
<td>100</td>
<td>30</td>
<td>44</td>
<td>1'463</td>
<td>50</td>
<td>440'734</td>
</tr>
<tr>
<td>Size of images</td>
<td>10^3 x 10^3</td>
<td>Varying</td>
<td>Varying</td>
<td>500 x 500</td>
<td>10^3 x 10^3</td>
<td>10^3 x 10^3</td>
<td>512 x 512</td>
<td>512 x 512</td>
<td>256 x 256</td>
</tr>
<tr>
<td>Number of cells</td>
<td>24'332</td>
<td>2'905</td>
<td>7'570</td>
<td>16'954</td>
<td>23'610</td>
<td>10'821</td>
<td>4'053</td>
<td>45'127</td>
<td></td>
</tr>
<tr>
<td>Exhaustive</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

We then apply watershed with origin points being the cell centroids to acquire cell shapes and boundaries.

Bounding box labels for all datasets were created by calculating the width, height and centre position of each cell. No other pre-processing steps were taken that affected the raw image signal. Datasets CoNSeP, CPM15, CPM17, CRCHisto, Kumar, and MoNuSeg, do not have constant image size dimensions of power of 2, e.g. $2^8 = 256$. This situation is prone to practical errors during training different sized images using one model. To mediate this effect, a random crop operation is applied to reduce the image dimensions to the nearest power of 2 number. For example, an image of size $1'000 \times 1'000$ is cropped at a random location to form an input sample of $512 \times 512$.

**Data augmentation**  Only for the training sets, for each sample we randomly apply the following augmentations: (1) image blurring (2) Gaussian additive noise (3) rotations around 90 degrees (4) left-right, up-down flips, and (5) diagonal flips. For the image blurring, we use a Gaussian kernel of randomly selected size chosen from the set {3, 5, 7}. For the additive noise, we generate noise by sampling out of a Gaussian distribution with zero mean and standard deviation equal to 5. For the rotational augmentations we randomly select an angle from the set {90, 180, 270}. All 5 augmentation variants are randomly applied with a probability of 0.5 per sample.

In addition, for the datasets with H&E stained images we apply a normalisation method, introduced in the work of [7], which perturbs stain concentrations. As recommended by the original work of [7], this step helps model staining variability in histological images; such the ones which are used in this work.

**3. Experimental details**

The Unet network was implemented as originally introduced in the work of [6]. No changes were made to the model architecture. The YOLLO network, as described in the work of [8], comprises 7 convolutional layers with batch normalisation and max pooling operations. In table 2 the hyper-parameters used for training the two models are listed. No change of parameters is made when using a different model or when changing datasets.

For the detection task the predicted bounding boxes were further processes with a non-maximum suppression algorithm, as described in the work of [5]. From the remaining bounding boxes the interest-over-union (IOU) is calculated for each available label box and if is above the IOU threshold then the object is considered correctly detected.

It was observed, that in some datasets of the detection task for the 90% and 100% sets, exclusive cross-entropy did not attend top performance over the traditional cross-entropy method by a couple of percentage points. For the

Table 2: Hyper-parameters used for training.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam optimiser betas</td>
<td>(0.9, 0.999)</td>
</tr>
<tr>
<td>Learning rate</td>
<td>$10^{-4}$</td>
</tr>
<tr>
<td>Batch size</td>
<td>16</td>
</tr>
<tr>
<td>Epochs</td>
<td>300</td>
</tr>
<tr>
<td>Detection $\lambda_{Background}$</td>
<td>10</td>
</tr>
<tr>
<td>Detection $\lambda_{Objects}$</td>
<td>1</td>
</tr>
<tr>
<td>Detection iou threshold</td>
<td>0.5</td>
</tr>
<tr>
<td>Non-maximum suppression threshold</td>
<td>0.2</td>
</tr>
<tr>
<td>ECE, background group annealing epoch</td>
<td>50</td>
</tr>
<tr>
<td>ECE, negative sampling annealing epoch</td>
<td>150</td>
</tr>
<tr>
<td>Gradient accumulation iterations</td>
<td>4</td>
</tr>
</tbody>
</table>

Figure 2: Qualitative results for losses CE (left), ECE (right) on the TIL dataset. Blue squares indicate predictions for lymphocytes.
Table 3: Dice and Aggregated Jaccard Index (AJI) scores results for the segmentation task using the Early Learning Regularization (ELR) method. ELR asymmetric refers to the parameter configuration of the Early Learning Regularization method using the training under their asymmetric setup in the original work. Similarly, for the CIFAR 10 and CIFAR 100 configurations.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>DICE 30</th>
<th>DICE 60</th>
<th>DICE 100</th>
<th>AJI 30</th>
<th>AJI 60</th>
<th>AJI 100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.55</td>
<td>0.75</td>
<td>0.82</td>
<td>0.35</td>
<td>0.41</td>
<td>0.71</td>
</tr>
<tr>
<td>TNBC</td>
<td>ELR asymmetric</td>
<td>0.62</td>
<td>0.73</td>
<td>0.80</td>
<td>0.51</td>
<td>0.64</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>ELR CIFAR 10</td>
<td>0.63</td>
<td>0.74</td>
<td>0.79</td>
<td>0.57</td>
<td>0.61</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>Cross-entropy</td>
<td>0.43</td>
<td>0.61</td>
<td>0.84</td>
<td>0.24</td>
<td>0.43</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Exclusive cross-entropy</td>
<td>0.80</td>
<td>0.83</td>
<td>0.82</td>
<td>0.58</td>
<td>0.73</td>
<td>0.74</td>
</tr>
<tr>
<td>CPM15</td>
<td>ELR asymmetric</td>
<td>0.24</td>
<td>0.66</td>
<td>0.81</td>
<td>0.16</td>
<td>0.58</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>ELR CIFAR 10</td>
<td>0.39</td>
<td>0.65</td>
<td>0.76</td>
<td>0.26</td>
<td>0.33</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Cross-entropy</td>
<td>0.59</td>
<td>0.67</td>
<td>0.76</td>
<td>0.32</td>
<td>0.44</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>Exclusive cross-entropy</td>
<td>0.79</td>
<td>0.82</td>
<td>0.81</td>
<td>0.55</td>
<td>0.64</td>
<td>0.69</td>
</tr>
<tr>
<td>CoNSeP</td>
<td>ELR asymmetric</td>
<td>0.53</td>
<td>0.62</td>
<td>0.76</td>
<td>0.43</td>
<td>0.45</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>ELR CIFAR 10</td>
<td>0.66</td>
<td>0.67</td>
<td>0.69</td>
<td>0.42</td>
<td>0.67</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>Cross-entropy</td>
<td>0.47</td>
<td>0.61</td>
<td>0.80</td>
<td>0.33</td>
<td>0.45</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Exclusive cross-entropy</td>
<td>0.78</td>
<td>0.80</td>
<td>0.79</td>
<td>0.58</td>
<td>0.70</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Table 4: F1-score results for the detection task training the YOLLO model using the Exclusive cross-entropy loss with a larger $\lambda_{\text{noobj}}$ value manually defined for exhaustive labels.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNSeP</td>
<td>CE</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>ECE, current</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>ECE, exhaustive-specific</td>
<td>0.53</td>
<td>0.52</td>
</tr>
<tr>
<td>Kumar</td>
<td>CE</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>ECE, current</td>
<td>0.61</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>ECE, exhaustive-specific</td>
<td>0.66</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Weakly supervised learning and early learning regularization

We compare our exclusive cross-entropy with the related work of Early Learning Regularization (ELR) indicating a unique example for noisy label learning, using a weakly supervised approach.

In table 3, the Dice and Aggregated Jaccard Index (AJI) scores are shown comparing the methods. It is important to note that the same pattern observed in the traditional weakly supervised method is also shown in the results for the early learning regularisation. In sparser sets, there exists too much noise for the noisy label learning to be able to work, while the exclusive cross-entropy outperforms this weakly supervised variant. In addition, early learning regularisation adapts its parameters for three separate difficulty levels, whilst for the exclusive cross-entropy all 161 ECE experiments with YOLLO and Unet on nine datasets use the same hyperparameters.

4. Qualitative results for TIL dataset

In 2, the difference between these two losses is seen qualitatively. Because the exclusive recall $\text{Rec}_{exc}(y)$ cannot account for missed true positives, the quantitative results on the TIL localisation dataset show only a small increase in performance for the exclusive cross-entropy. However, when inspecting the detections visually, it can be observed that the exclusive cross-entropy makes more conservative predictions, reducing the number of false positives signifi-
cantly; which cannot be reflected in any performance score of non-exhaustively annotated dataset. This is consistent with previous results on all other datasets where the exclusive cross-entropy demonstrates a significant drop in false positive counts, compared to cross-entropy.

References


