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

Andreas Panteli[1,2], Jonas Teuwen[1,2,3], Hugo Horlings[1] and Efstratios Gavves[2,4]
[1]Netherlands Cancer Institute, [2]University of Amsterdam,
[3]Radboud University Medical Center, [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

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\mathcal{F}} + \mathcal{L}_{\overline{\mathcal{F}}} \\ &= \sum_{\mathcal{F}} \log p + \sum_{\overline{\mathcal{F}}} \log(1-p)\end{aligned} \tag{1}$$

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 $\mathcal{L}_{\mathcal{F}}$ corresponds to the loss on the foreground area $\mathcal{F}$ for which we have manual annotations, and $\mathcal{L}_{\overline{\mathcal{F}}}$ corresponds to the loss on the rest area $\overline{\mathcal{F}}$. As we do not have the true annotations for $\overline{\mathcal{F}}$, we cannot really compute the loss $\mathcal{L}_{\overline{\mathcal{F}}}$, at least not accurately. As usual, we rely on stochastic gradient descent for optimising the model parameters, that is

$$\begin{aligned} w_{t+1} &= w_t - \varepsilon \frac{d\,\mathcal{L}}{dw}\bigg|_t \Rightarrow \\ \frac{dw}{dt} &= -\varepsilon \frac{d\,\mathcal{L}}{dw}\bigg|_t, \end{aligned} \tag{2}$$

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

$$\begin{aligned}\frac{d\,\mathcal{L}}{dw} &= \sum_{\mathcal{F}} \frac{1}{p}\frac{dp}{dw} + \sum_{\overline{\mathcal{F}}} \frac{1}{1-p}\frac{d(1-p)}{dw} \\ &= \sum_{\mathcal{F}} \frac{1}{p}\frac{dp}{dw} - \sum_{\overline{\mathcal{F}}} \frac{1}{1-p}\frac{dp}{dw} \\ &= \frac{d\,\mathcal{L}_{\mathcal{F}}}{dw} + \frac{d\,\mathcal{L}_{\overline{\mathcal{F}}}}{dw}\end{aligned} \tag{3}$$

For binary classification, we use sigmoidal neurons in the outputs, that is

$$p = p(y=1|x) = \sigma(\sum_k w_k x_k), \tag{4}$$

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 (3) and focusing on the derivative with respect to the j-th weight, we have

$$\begin{aligned}\frac{d\,\mathcal{L}}{dw_j} &= \frac{d\,\mathcal{L}_{\mathcal{F}}}{dw_j} + \frac{d\,\mathcal{L}_{\overline{\mathcal{F}}}}{dw_j} \\ &= \sum_{\mathcal{F}} \frac{1}{\sigma(\sum_k w_k x_k)} \frac{d(\sigma(\sum_k w_k x_k))}{dw_j} \\ &\quad - \sum_{\overline{\mathcal{F}}} \frac{1}{1-\sigma(\sum_k w_k x_k)} \frac{d(\sigma(\sum_k w_k x_k))}{dw_j} \\ &= \sum_{\mathcal{F}} \frac{1}{\cancel{\sigma(\sum_k w_k x_k)}} x_j \cancel{\sigma(\sum_k w_k x_k)} \\ &\quad \cdot (1-\sigma(\sum_k w_k x_k))) \\ &\quad - \sum_{\overline{\mathcal{F}}} \frac{1}{\cancel{1-\sigma(\sum_k w_k x_k)}} x_j \sigma(\sum_k w_k x_k) \\ &\quad \cdot \cancel{(1-\sigma(\sum_k w_k x_k))} \\ &= \sum_{\mathcal{F}} x_j(1-p) - \sum_{\overline{\mathcal{F}}} x_j p\end{aligned} \tag{5}$$

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 $\overline{\mathcal{F}}$, 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

$$\begin{aligned}\frac{d\,\mathcal{L}_{\overline{\mathcal{F}}}}{dt} &= \sum_k \frac{d\,\mathcal{L}_{\overline{\mathcal{F}}}}{dw_k}\frac{dw_k}{dt} \\ &= \sum_k \frac{d\,\mathcal{L}_{\overline{\mathcal{F}}}}{dw_k}(-\varepsilon\frac{d\,\mathcal{L}_{\overline{\mathcal{F}}}}{dw_k}) \\ &= -\varepsilon\sum_k \Big(\frac{d\,\mathcal{L}_{\overline{\mathcal{F}}}}{dw_k}\Big)^2\end{aligned} \tag{6}$$

Using equations (6) and the term in (5) that corresponds to $\overline{\mathcal{F}}$, and dropping the $\sum_{\overline{\mathcal{F}}}$ for notation clarity (the total result is the sum over all samples in $\overline{\mathcal{F}}$), the second derivative is then

$$\begin{aligned}\frac{d^2\,\mathcal{L}_{\overline{\mathcal{F}}}}{dt^2} &= \frac{d}{dt}\Big[-\varepsilon\sum_k\Big(\frac{d\,\mathcal{L}}{dw_k}\Big)^2\Big] \\ &= -\varepsilon\sum_k 2\frac{d\,\mathcal{L}_{\overline{\mathcal{F}}}}{dw_k}\frac{d}{dt}\Big(\frac{d\,\mathcal{L}_{\overline{\mathcal{F}}}}{dw_k}\Big) \\ &= -2\varepsilon\sum_k(-x_k p)\frac{d}{dt}(-x_k p) \\ &= -2\varepsilon\sum_k x_k^2 p\Big(p\cdot(1-p)\Big)\frac{d}{dt}\sum_r w_r x_r \\ &= -2\varepsilon\sum_k x_k^2 p^2(1-p)\frac{d}{dt}\sum_r w_r x_r\end{aligned} \tag{7}$$

We know that

$$\begin{aligned}\frac{dw_r}{dt} &= -\varepsilon\frac{d\,\mathcal{L}_{\overline{\mathcal{F}}}}{dw_r} \\ &= -\varepsilon(-x_r p) = \varepsilon x_r p\end{aligned} \tag{8}$$

By combining equations (7) and (8), we obtain that

$$\begin{aligned}\frac{d^2\,\mathcal{L}_{\overline{\mathcal{F}}}}{dt^2} &= -2\varepsilon\sum_k x_k^2 p^2(1-p)\sum_r x_r^2 p \\ &= -2\varepsilon p^3(1-p)\sum_{k,r} x_k^2 x_r^2 \\ &\propto p^m(1-p)^n\end{aligned} \tag{9}$$

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 $\frac{d^2\,\mathcal{L}_{\overline{\mathcal{F}}}}{dt^2}$ 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].

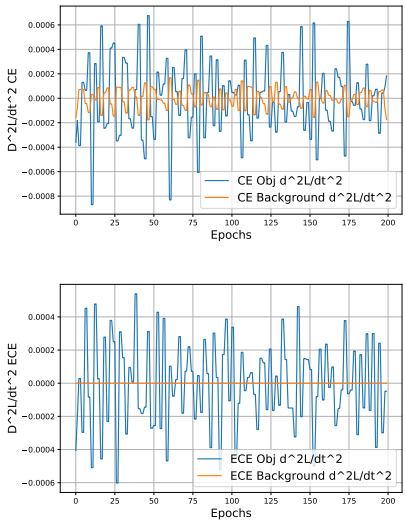

Figure 1: Second order derivative graphs for the cross-entropy and the exclusive cross-entropy losses, split per Object "Obj" and Background loss components on the 60% non-exhaustive set of the TNBC dataset

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 avert 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

[1] https://monuseg.grand-challenge.org/Data/

Table 1: Summary information of datasets.

	CoNSeP	CPM15	CPM17	CRCHisto	Kumar	MoNuSeg	WBC-N.	TNBC	TIL
Number of images	41	15	74	100	30	44	1'463	50	440'734
Size of images	10^3x10^3	Varying	Varying	500x500	10^3x10^3	10^3x10^3	512x512	512x512	256x256
Number of cells	24'332	2'905	7'570	29'748	16'954	23'610	10'821	4'053	45'127
Exhaustive	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No

TNBC and WBC-NuClick respectively. Using this information, cell centroids are selected from the distinct local maxima separated by at least the dataset-specific diameter. 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 dimenions 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 512×512.

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.

Table 2: Hyper-parameters used for training.

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

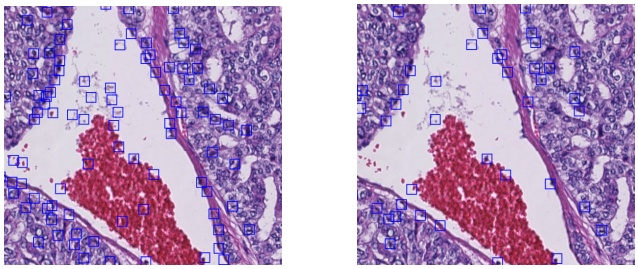

Figure 2: Qualitative results for losses CE (left), ECE (right) on the TIL dataset. Blue squares indicate predictions for lymphocytes.

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 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.

Dataset	Method	DICE			AJI		
		30	60	100	30	60	100
TNBC	ELR asymmetric	0.55	0.75	0.82	0.35	0.41	0.71
	ELR CIFAR 10	0.62	0.73	0.80	0.51	0.64	0.67
	ELR CIFAR 100	0.63	0.74	0.79	0.57	0.61	0.66
	Cross-entropy	0.43	0.61	**0.84**	0.24	0.43	0.73
	Exclusive cross-entropy	**0.80**	**0.83**	0.82	**0.58**	**0.73**	**0.74**
CPM15	ELR asymmetric	0.24	0.66	**0.81**	0.16	0.58	0.65
	ELR CIFAR 10	0.39	0.65	0.76	0.26	0.33	0.68
	ELR CIFAR 100	0.59	0.67	0.76	0.32	0.44	0.56
	Cross-entropy	0.47	0.61	0.80	0.25	0.52	**0.69**
	Exclusive cross-entropy	**0.79**	**0.82**	**0.81**	**0.55**	**0.64**	**0.69**
CoNSeP	ELR asymmetric	0.53	0.62	0.76	0.43	0.45	0.52
	ELR CIFAR 10	0.60	0.66	0.74	0.36	0.42	0.43
	ELR CIFAR 100	0.66	0.67	0.69	0.42	0.67	0.48
	Cross-entropy	0.47	0.61	**0.80**	0.33	0.45	0.67
	Exclusive cross-entropy	**0.78**	**0.80**	0.79	**0.58**	**0.70**	**0.69**

sake of generality, exclusive cross-entropy was developed using as reference the collective scores of the 30%, 60% and 80% sets of the TNBC dataset. This was done to satisfy high performance for the intended task of sparse-shot learning without having a costly trade-off. This resulted in the drop of 1-3 percentage points at the fully exhaustive sets. In table 4 we present results with the exclusive cross-entropy by adjusting the weighing factor of the no-objects group in the loss function [3]. The results show that it is possible to optimise further the exclusive cross-entropy training for the exhaustive cases specifically if required.

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

Dataset	Method	90%	100%
CoNSeP	CE	0.50	0.51
	ECE, current	0.48	0.48
	ECE, exhaustive-specific	**0.53**	**0.52**
Kumar	CE	0.64	0.63
	ECE, current	0.61	0.62
	ECE, exhaustive-specific	**0.66**	**0.68**

Weakly supervised learning and early learning regularization We compare our exclusive cross-entropy with the related work of Early Learning Regularization (ELR) [4] 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 $\text{Rec}_{\text{exc}}(y)$ 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

[1] Nezamoddin N Kachouie, Paul Fieguth, and Eric Jervis. Watershed deconvolution for cell segmentation. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 375–378. IEEE, 2008.

[2] Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir Rajpoot. Nuclick: A deep learning framework for interactive segmentation of microscopic images. *Medical Image Analysis*, 65:101771, 2020.

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[4] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *arXiv:2007.00151*, 2020.

[5] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 850–855. IEEE, 2006.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[7] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.

[8] Mart van Rijthoven, Zaneta Swiderska-Chadaj, Katja Seeliger, Jeroen van der Laak, and Francesco Ciompi. You only look on lymphocytes once. 2018.