Deep Active Learning for Joint Classification & Segmentation with Weak Annotator

Soufiane Belharbi[1], Ismail Ben Ayed[1], Luke McCaffrey[2], and Eric Granger[1]
[1] LIVIA, Dept. of Systems Engineering, École de technologie supérieure, Montreal, Canada
[2] Goodman Cancer Research Centre, Dept. of Oncology, McGill University, Montreal, Canada
soufiane.belharbi.1@ens.etsmtl.ca, luke.mccaffrey@mcgill.ca,
{ismail.benayed, eric.granger}@etsmtl.ca

Due to space limitation, we provide in this supplementary material detailed hyper-parameters used in the experiments, results of the ablation study, visual results to the similarity measure, and examples of predicted masks.

1. Supplementary material for the experiments

1.1. Training hyper-parameters

Tab.1 shows the used hyper-parameters in all our experiments.

Table 1: Training hyper-parameters.

Hyper-parameter	GlaS	CUB
Model backbone	ResNet-18 [3]	
WILDCAT [2]:		
α	0.6	
$kmin$	0.1	
$kmax$	0.1	
modalities	5	
Optimizer	SGD	
Nesterov acceleration	True	
Momentum	0.9	
Weight decay	0.0001	
Learning rate (LR)	0.1 (WSL: 10^{-4})	0.1 (WSL: 10^{-2})
LR decay	0.9	0.95 (WSL: 0.9)
LR frequency decay	100 epochs	10 epochs
Mini-batch size	20	8
Learning epochs	1000	30 (WSL: 90)
Horizontal random flip	True	
Vertical random flip	True	False
Crop size	416×416	
k	40	
λ	0.1	0.001

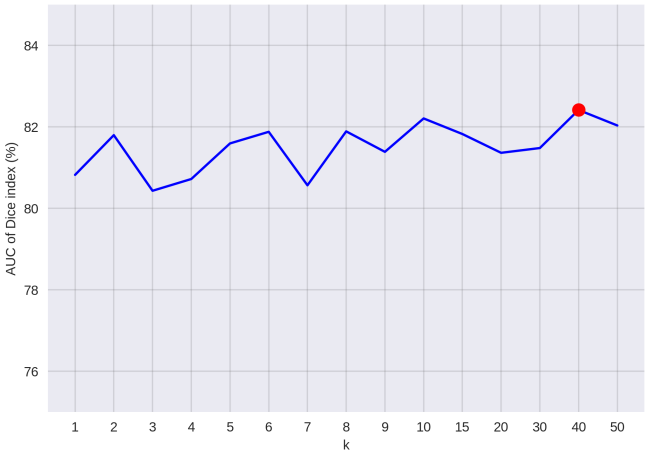

Figure 1: Ablation study over GlaS dataset (test set) over the hyper-parameter k (x-axis). y-axis: AUC of Dice index (%) of **25 queries for one trial**. AUC average $\pm$ standard deviation: 81.49 ± 0.59. Best performance in red dot: $k = 40$, $AUC = 82.41\%$.

1.2. Ablation study

We study the impact of k and λ on our method. Results are presented in Fig.1, 2 for GlaS over k, λ; and in Fig.3 for CUB over λ. Due to the expensive computation time required to perform AL experiments, we limited the experiments (k, λ, number of trials, and `maxr`). The obtained results of this study show that our method is less sensitive to k (standard deviation of 0.59 in Fig.1). In other hand, the method shows sensitivity to λ as expected from penalty-based methods [1]. However, the method seems more sensitive to λ in the case of CUB than GlaS. CUb dataset is more challenging leading to more potential erroneous pseudo-annotation. Using Large λ will systematically push the model to learn on the wrong annotation (Fig.3) which leads to poor results. In the other hand, GlaS

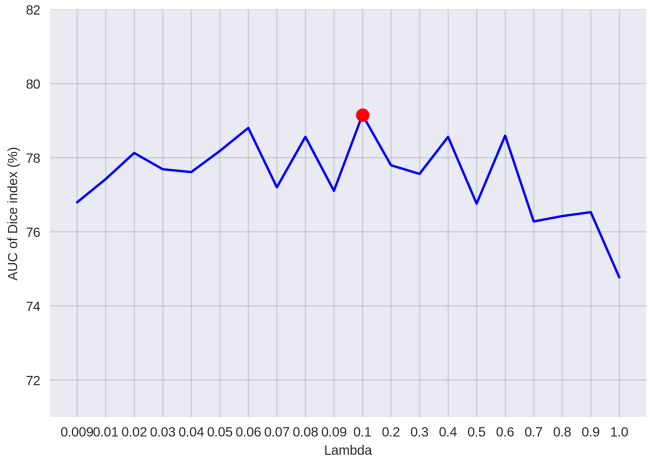

Figure 2: Ablation study over GlaS dataset (test set) over the hyper-parameter λ (x-axis). y-axis: AUC of Dice index (%) of **15 queries for one trial**. Best performance in red dot: $\lambda = 0.1, AUC = 79.15\%$.

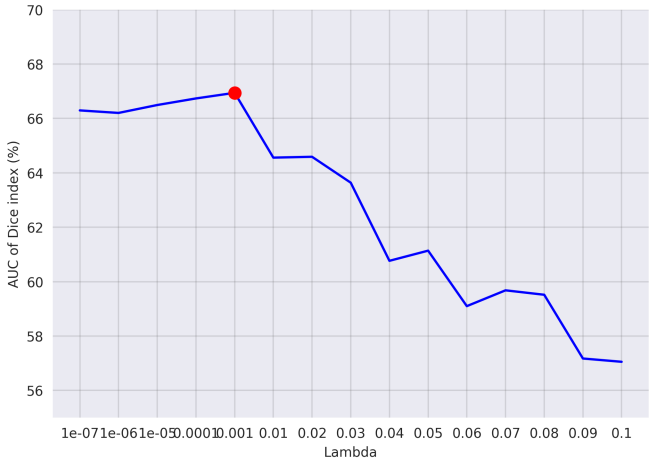

Figure 3: Ablation study over CUB dataset (test set) over the hyper-parameter λ (x-axis). y-axis: AUC of Dice index (%) of **5 queries for one trial**. Best performance in red dot: $\lambda = 0.001, AUC = 66.94\%$.

seems to allow obtaining good segmentation where using large values of λ did not hinder the performance quickly (2). The obtained results recommend using small values that lead to better and stable performance. Using high values, combined with the pseudo-annotation errors, push the network to learn erroneous annotation leading to overall poor performance.

1.3. Similarity measure

In this section, we present some samples with their nearest neighbors. Although, it is difficult to quantitatively evaluate the quality of such measure. Fig.4 shows the case of GlaS. Overall, the similarity shows good behavior of capturing the general stain of the image which is what was intended for since the structure of such histology images is subject to high variation. Since the stain variation is one of the challenging aspects in histology images [4], labeling a sample with a common stain can help the model in segmenting other samples with similar stain. The case of CUB, presented in Fig.5, is more difficult to judge the quality since the images contain always the same species within their natural habitat. Often, the similarity succeeds to capture the overall color, background which can help segmenting the object in the neighbors and also the background. In some cases, the similarity captures samples with large zoom-in where the bird color dominate the image.

1.4. Predicted mask visualization

Fig.6 shows several test examples of predicted masks of different methods over CUB test set at the first AL round ($r = 1$) where only one sample per class has been labeled by the oracle. This interesting functioning point shows that by labeling only one sample per class, the performance of the average Dice index can go from 39.08 ± 08 for WSL method up to 62.58 ± 2.15 for Label_prop and other AL methods. The figure shows that WSL tend to spot small part of the object in addition to the background leading high false positive. Using few supervision in combination with the proposed architecture, better segmentation is achieved by spotting large part of the object with less confusion with the background.

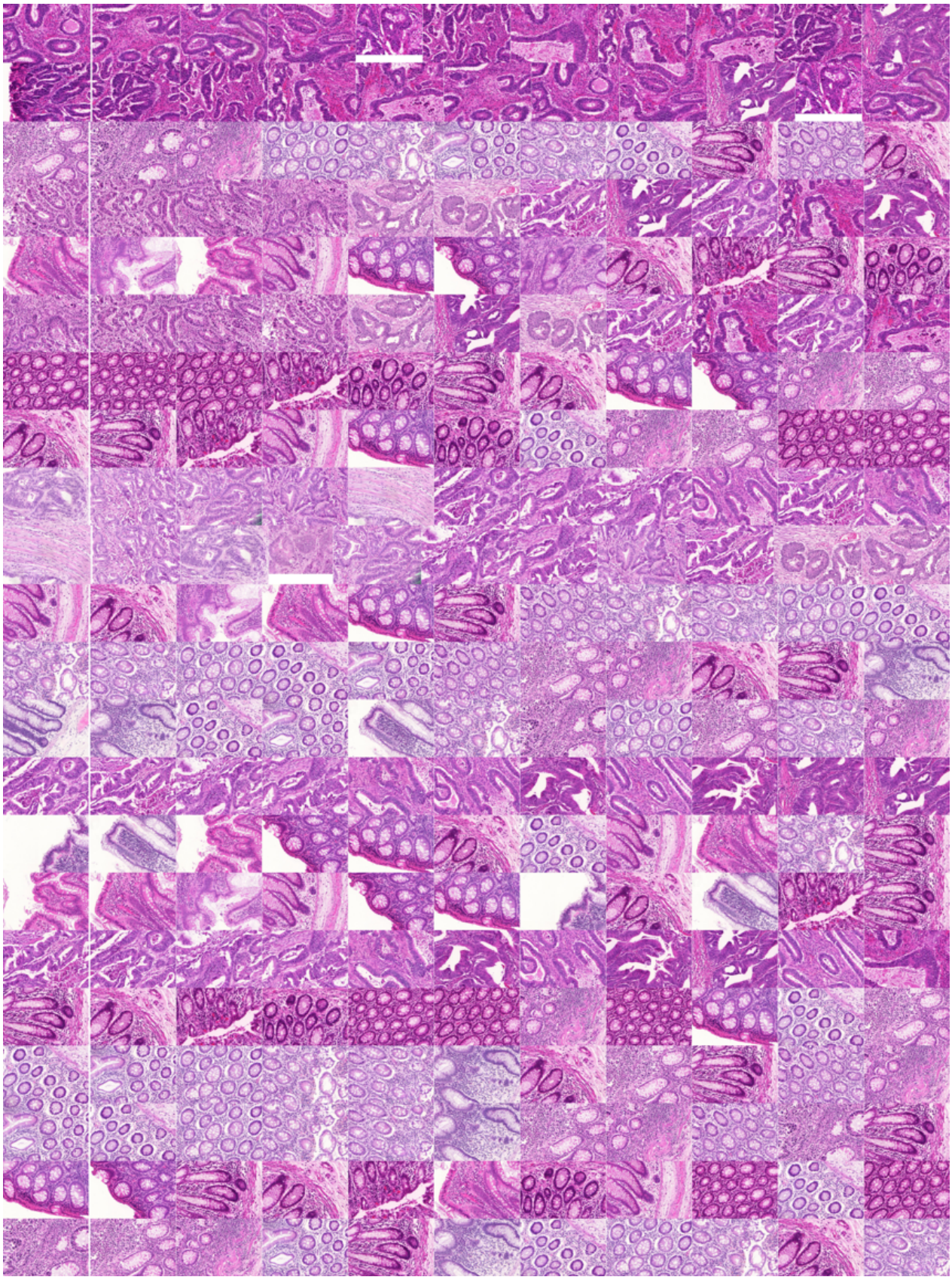

Figure 4: Examples of k-nn over GlaS dataset. The images represents the 10 nearest images to the first image in the extreme left ordered from the nearest.

Figure 5: Examples of k-nn over CUB dataset. The images represents the 10 nearest images to the first image in the extreme left ordered from the nearest.

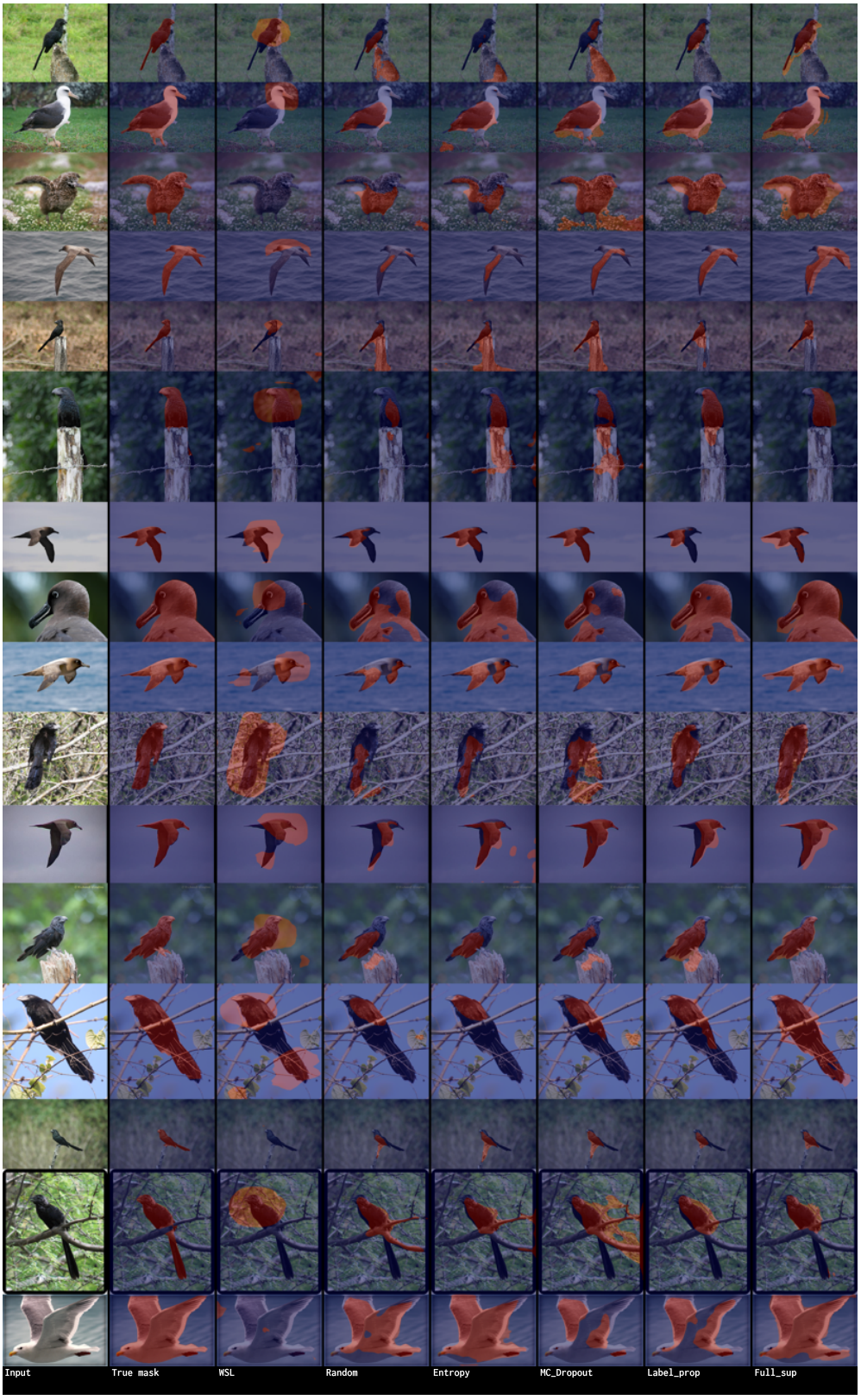

Figure 6: Qualitative results (on several CUB test images) of the predicted binary mask for each method after being trained in the first round $r = 1$ (*i.e.* after labeling 1 sample per class) using seed=0. The average Dice index over the test set of each method is: 40.16% (WSL), 55.32% (Random), 55.41% (Entropy), 55.52% (MC_Dropout), 59.00% (Label_prop), and 75.29% (Full_sup). (Best visualized in color.)

References

[1] DP Bertsekas. Nonlinear programming, 2nd ed. chapter 4. 1999.

[2] T. Durand, T. Mordan, N. Thome, and M. Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *CVPR*, volume 2, 2017.

[3] K. He, X. Zhang, S.g Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] J. Rony, S. Belharbi, J. Dolz, I. Ben Ayed, L. McCaffrey, and E. Granger. Deep weakly-supervised learning methods for classification and localization in histology images: a survey. *CoRR*, abs/1909.03354, 2019.