Ensembling Low Precision Models for Binary Biomedical Image Segmentation Supplemental Materials

Tianyu Ma
Cornell University
tm478@cornell.edu

Hang Zhang
Cornell University
hz459@cornell.edu

Hanley Ong
Weill Cornell Medical College
hao2007@med.cornell.edu

Amar Vora
Weill Cornell Medical College
apv7002@med.cornell.edu

Thanh D. Nguyen
Weill Cornell Medical College
tdn2001@med.cornell.edu

Ajay Gupta
Weill Cornell Medical College
ajg9004@med.cornell.edu

Yi Wang
Cornell University
yw233@cornell.edu

Mert R. Sabuncu
Cornell University
msabuncu@cornell.edu

1. Tversky and Balanced Cross-Entropy Loss

As we mention in our paper, we can use either Tversky or balanced cross-entropy as the loss function to encourage high recall predictions. In all three experiments, we conduct analyses using both loss functions but only include the subset results that produces the best dice score in the paper. Here, we show the dice scores obtained by either Tversky or BCE loss for all three experimens. Tables 1,2, and 3 list the numeric values.

Method	Dice (BCE)	Dice (Tversky)
Single Baseline Model	0.557	0.576
Single Low Precision Model	0.432	0.435
Baseline Ensemble	0.606	0.614
Low Prec Ensemble (β=0.95)	0.633	0.643
Low Prec Ensemble (random β)	0.675	0.708

Table 1. Dice Scores for Internal Carotid Artery Segmentation in Neck CTA with both BCE and Tversky loss

Method	Dice (BCE)	Dice (Tversky)
Single Baseline Model	0.786	0.762
Single Low Precision Model	0.757	0.749
Baseline Ensemble	0.790	0.774
Low Prec Ensemble (β=0.95)	0.796	0.792
Low Prec Ensemble (random β)	0.815	0.811

Table 2. Dice Scores for Segmentation of Ventricular My-ocardium in MRI with both BCE and Tversky loss

Method	Dice (BCE)	Dice (Tversky)
Baseline Ensemble	0.608	0.624
Low Prec Ensemble (β=0.95)	0.626	0.645
Low Prec Ensemble (random β)	0.647	0.662

Table 3. Dice Scores for Segmentation of MS lesions MRI with both BCE and Tversky loss

2. Generalized Ostu's and Other Thresholding Methods

Generalized Ostu's method (GHT) [1], which based on histogram thresholding method, has shown potentials in many binary segmentation tasks including medical image segmentation. One advantage of such approach is that it does require any labels for training. We explore the idea of thresholding in all the datasets we used in our experiments.

Task	GHT	Oracle Threshold
Internal Carotid Artery	0.165	0.214
Ventricular My-ocardium	0.283	0.307
MS lesions	0.203	0.259

Table 4. Dice Scores for Segmentation tasks using different thresholding methods

Table 4 shows the results from using GHT and oracle threshold, which is the best thresholds one can obtain given labels. We can see from the table that the dice scores for all tasks using thresholding approach are substantially lower than results from training with neural networks. It also indicates that our tasks have a lot of hyper-intensities that are hard to distinguish using their intensity levels.

3. Aggregation with Different Thresholds

In all of our experiments, our baseline ensemble are obtained by average all models and thresholded at 0.5. On the other hand, the low precision ensembles use 0.9 as the threshold for the final binary segmentation results. We choose these numbers based on the validation results. In figure 1, we show the aggregated results using thresholds from 0.3 to 0.9 for all experiments. For baseline models trained with $\beta = 0.5$, a threshold of 0.5 always produces the highest dice score. For low precision models trained with random βs greater than 0.9, the dice scores monotonically increase with threshold values.

References

[1] Jonathan T Barron. A generalization of otsu's method and minimum error thresholding. *arXiv preprint arXiv:2007.07350*, 2020.

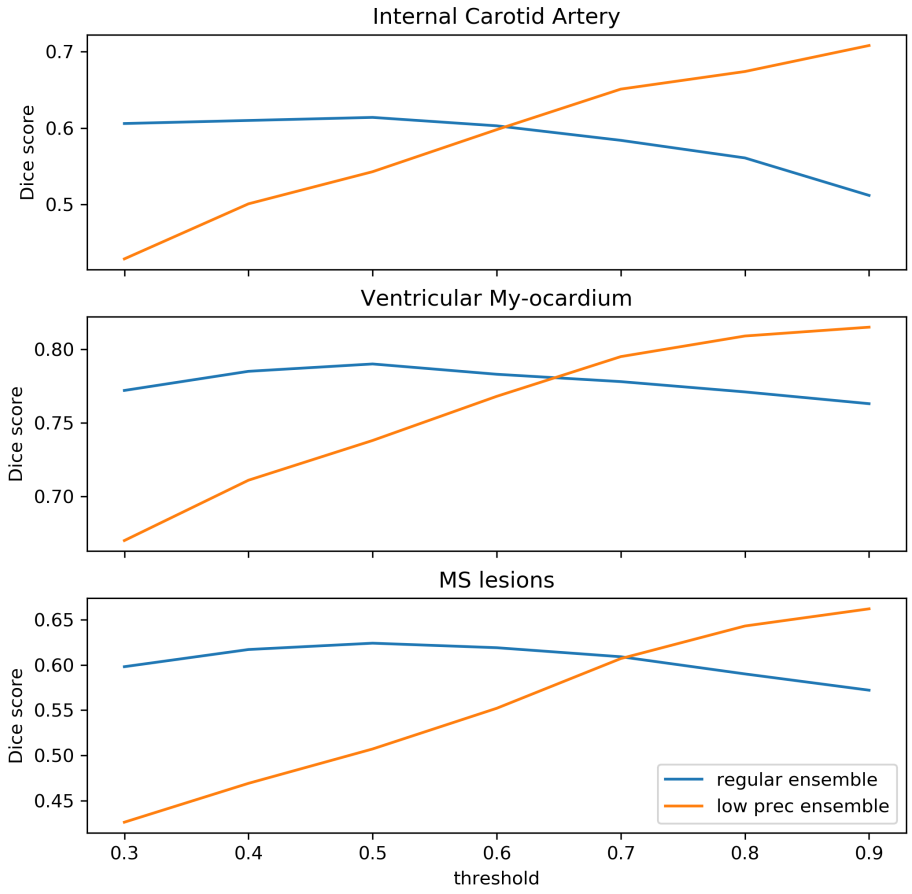

Figure 1. Dice score for different thresholds in segmentation of internal carotid artery, ventricular my-ocardium, and MS lesion