Ensembling Low Precision Models for Binary Biomedical Image Segmentation

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

Abstract

Segmentation of anatomical regions of interest such as vessels or small lesions in medical images is still a difficult problem that is often tackled with manual input by an expert. One of the major challenges for this task is that the appearance of foreground (positive) regions can be similar to background (negative) regions. As a result, many automatic segmentation algorithms tend to exhibit asymmetric errors, typically producing more false positives than false negatives. In this paper, we aim to leverage this asymmetry and train a diverse ensemble of models with very high recall, while sacrificing their precision. Our core idea is straightforward: A diverse ensemble of low precision and high recall models are likely to make different false positive errors (classifying background as foreground in different parts of the image), but the true positives will tend to be consistent. Thus, in aggregate the false positive errors will cancel out, yielding high performance for the ensemble. Our strategy is general and can be applied with any segmentation model. In three different applications (carotid artery segmentation in a neck CT angiography, myocardium segmentation in a cardiovascular MRI and multiple sclerosis lesion segmentation in a brain MRI), we show how the proposed approach can significantly boost the performance of a baseline segmentation method.

1. Introduction

Deep learning techniques, such as the U-Net [20], produce most state-of-the-art biomedical image segmentation tools. However, delineating a relatively small region of interest in a biomedical image (such as a small vessel or le-
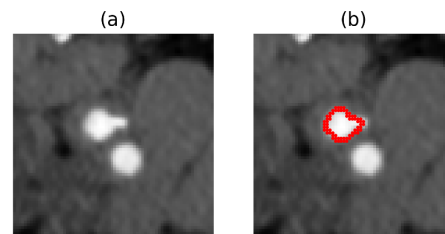

Figure 1: (a) An example CTA image with both internal and external carotid artery. (b) The image and overlaid manual ground-truth of the internal carotid artery

sion) is a challenging problem that current segmentation algorithms still struggle with. In these applications, one of the main problems is that there are different anatomic structures present in the images with similar shapes and intensity values to the foreground structure(s), making it difficult to distinguish them from each other [23]. Those intrinsically hard segmentation tasks are considered challenging even for human experts. One such task is the localization of the *internal* carotid artery in computed tomography angiography (CTA) scans of the neck. As shown in Figure 1, there are no features that separate the internal and external carotid arteries in the CTA appearance other than their relative position[18]. Learning these features, particularly with limited data, can be challenging for convolutional neural networks.

In this paper, we propose an easy-to-use novel ensemble learning strategy to deal with the challenges we mention above. Ensemble learning is an effective general-purpose

machine learning technique that combines predictions of individual models to achieve better performance. In deep learning, it often involves ensembling of several neural networks trained separately with random initialization [19, 16]. It has also been used to calibrate the predictive confidence in classification models, particularly when the individual model is over-confident as in the case of modern deep learning models [12]. In problem settings where it is more likely to make low precision predictions such as nodule detection, ensemble learning has been used to reduce the false positive rate [24, 13, 4].

Ensemble learning has previously been used for binary segmentation, where multiple probabilistic models are combined, for example by weighted averaging, and the final binary output is computed by thresholding the weighted average at 0.5 [9, 26]. The performance of a regular ensemble model depends on both the accuracy of the individual segmenters and the diversity among them [11, 31]. Previous works using ensembling for image segmentation mainly focus on having diverse base segmenters while maintaining individual segmenter accuracy [8, 14]. In this conventional ensemble segmentation framework, the diverse errors are often distributed between foreground and background pixels.

In this paper, we present a different take on ensembling for binary image segmentation. Our approach is to create an ensemble of models that each exhibit low precision (specificity) but very high recall (sensitivity). These individual models are also likely to have relatively low accuracy due to over-segmentation. We achieve this by altering widely used loss functions in a manner that tolerates false positives (FP) while keeping the false negative (FN) rate very low. During training of models in an ensemble, the relative weights of FP vs. FN predictions in each model are also selected randomly to promote diversity among them. In this way, each model will produce segmentation results that largely cover the foreground pixels, while possibly making different mistakes in background regions. These weak segmenters have high agreement on foreground pixels and low agreement on the predictions for background pixels. Compared to other ensembling strategies, our method thus does not focus on getting accurate base segmenters, but rather diverse models with high recall. We use two popular and easy-to-implement strategies to create the ensemble: bagging and random initialization. Each model is trained using a random subset of data and all parameters are randomly initialized to maximize the model diversity.

The proposed approach is general and can be used in a wide range of segmentation applications with a variety of models. We present three different experiments. In the first one, we consider the challenging problem of segmenting the internal carotid artery in a neck CTA scan. The second experiment deals with myocardium segmentation in a cardio-vascular MRI. In the third experiment, we test our method on multiple sclerosis lesion segmentation in a brain MRI. Our results demonstrate that the proposed ensembling technique can substantially boost the performance of a baseline segmentation method in all of our experimental settings.

2. Methods

Our proposed method uses weak segmenters with low accuracy and specificity to collectively make predictions. Although individual models are prone to over segment the image, they are capable of capturing most of the true positive pixels. During the final prediction, an almost unanimous agreement is required to classify a pixel as the foreground in order to eliminate the false positive predictions made by each model. The key is for all models in the ensemble to have a high amount of overlap in the true positive parts and low overlap in the false positive predictions. We can achieve this by modifying the loss function to put more weight on false negative than false positive predictions, and use a random weight for each model in the ensemble. We will show in section 2.3 that this simple modification can encourage diverse false positive errors.

Our ensemble approach is different from existing ensemble methods, which usually combine several high accuracy models and use majority voting for the final prediction.

2.1. Supervised Learning Based Segmentation

For a binary image segmentation problem, conditioned on an observed (vectorized) image $\boldsymbol{x} \in \mathbb{R}^N$ with N pixels (or voxels), the objective is to learn the true posterior distribution $p(\boldsymbol{y}|\boldsymbol{x})$ where $\boldsymbol{y} \in \{0, 1\}^N$; and 0 and 1 stand for background or foreground, respectively.

Given some training data, $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$, supervised deep learning techniques attempt to capture $p(\boldsymbol{y}|\boldsymbol{x})$ with a neural network (e.g. a U-Net architecture) that is parameterized with θ and computes a pixel-wise probabilistic (e.g., softmax) output $\boldsymbol{f}(\boldsymbol{x}, \theta)$, which can be considered to approximate the true posterior, of say, each pixel being foreground. So $f^j(\boldsymbol{x}, \theta)$ models $p(\boldsymbol{y}^j = 1|\boldsymbol{x})$, where the superscript indicates the j'th pixel. Loss functions widely used to train these neural networks include Dice, cross-entropy, and their variants. The probabilistic Dice loss quantifies the overlap of foreground pixels:

$$L_{Dice}(\boldsymbol{y}, \boldsymbol{f}(\boldsymbol{x}, \theta)) = 1 - \frac{2\sum_{j=1}^{N} \boldsymbol{y}^j \boldsymbol{f}^j}{\sum_{j=1}^{N} \boldsymbol{y}^j + \sum_{j=1}^{N} \boldsymbol{f}^j} \quad (1)$$

Cross-entropy loss, on the other hand, is defined as:

$$L_{CE}(\boldsymbol{y}, \boldsymbol{f}(\boldsymbol{x}, \theta)) = \sum_{j=1}^{N} -\boldsymbol{y}^j \log(\boldsymbol{f}^j)$$
$$- (1 - \boldsymbol{y}^j) \log((1 - \boldsymbol{f}^j)) \quad (2)$$

Training a segmentation model, therefore, involves finding the model parameters θ that minimize the adopted loss function on the training data.

2.2. Ensemble of Segmentation Models

One can execute the aforementioned training K times using the loss functions defined above to obtain K different parameter solutions $\{\theta_1, \ldots, \theta_K\}$. Each of these training sessions can rely on slightly different training data as in the case of bagging [2] or different random initializations [12]. Given an ensemble of models, the classical approach is to average the individual predictive probabilities, which would be considered as a better approximation of the true posterior:

$$p_{ensemble}(\boldsymbol{y}^j = 1 | \boldsymbol{x}) = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{f}^j(\boldsymbol{x}, \theta_k). \qquad (3)$$

The ensemble probabilistic prediction is usually thresholded with 0.5 to obtain a binary segmentation. In the remainder of the paper, we refer to this approach as the *baseline ensembling method*.

2.3. Ensembling Low Precision Models

In this paper, we propose an ensemble learning strategy that combines diverse models with low precision but high recall. Since each model will have a relatively high recall, each model will label the ground truth foreground pixels largely correctly. On the other hand, there will be many false positives due to the low precision of each model. If the models are diverse enough, these false positives will largely be very different and thus can cancel out when averaged.

To enforce all models within the ensemble to have high recall, we experimented with two different loss functions: Tversky [22] and balanced cross-entropy loss (BCE) [27], which are generalizations of the classical Dice and cross-entropy loss functions mentioned above. These loss functions, defined below, have a hyper-parameter that gives the user the control to adopt a different operating point on the precision/recall trade-off.

$$L_{Tv}(\boldsymbol{y}, \boldsymbol{f}) = 1 - \frac{\sum_{j=1}^{N} \boldsymbol{y}^j \boldsymbol{f}^j}{\sum_{j=1}^{N} \left[\boldsymbol{y}^j \boldsymbol{f}^j + \beta \boldsymbol{y}^j (1 - \boldsymbol{f}^j) + (1-\beta)(1-\boldsymbol{y}^j) \boldsymbol{f}^j \right]}$$
$$(4)$$

$$L_{BCE}(\boldsymbol{y}, \boldsymbol{f}) = \sum_{j=1}^{N} -\beta \boldsymbol{y}^j \log(\boldsymbol{f}^j)$$
$$- (1-\beta)(1-\boldsymbol{y}^j) \log((1 - \boldsymbol{f}^j)) \quad (5)$$

Note $\beta \in [0, 1)$ is a hyper-parameter and plays a similar role for both loss functions. When $\beta = 0.5$, Tversky loss

becomes equivalent to Dice, and BCE is the same as regular cross-entropy. For higher values of β, these loss functions will penalize false negatives (pixels with $\boldsymbol{y}^j = 1$ and $\boldsymbol{f}^j < 0.5$) more than false positives (pixels with $\boldsymbol{y}^j = 0$ and $\boldsymbol{f}^j > 0.5$). E.g., for $\beta > 0.9$, the false negative rates will be kept low (such that the recall rate is greater than 90%, for instance) while producing many false positives, yielding low precision. One can achieve the opposite effect of high precision but low recall with low values of β.

The idea we are promoting in this paper is to use the Tversky or BCE loss with a relatively high β in training individual models that make up an ensemble. We believe that other loss functions that can be tuned to control the precision/recall trade-off should also work for our purpose. The comparison between using different losses need to be further explored. In our experiments, when training each model, we randomly choose a value between $[0.9, 1)$ and use that to define the loss function for that training session in order to promote diversity among individual models. The exact range of β can be adjusted based on validation performance and desired level of recall. We then average these predictions in an ensemble, as in Equation 3. We threshold the ensemble prediction at the relatively high value of 0.9, which we found is effective in removing residual false positives. We present results for alternative thresholds in our Supplemental Material. We note that the threshold is an important hyper-parameter that can be tuned for best ensembling performance during validation. In all of our three experiments presented below, however, we simply used a threshold of 0.9. In an ensemble of ten or fewer models, this strategy is similar to aggregating the individual model segmentation (e.g. obtained after thresholding with 0.5) with a logical AND operation.

The key of our method is to combine diverse high recall models to reduce the false positives in the final prediction. We empirically observe that a β value between $[0.9, 1)$ is sufficient to make the model output to have a recall rate greater than 0.9. On the other hand, changing the β used in the loss function from 0.9 to 0.99 effectively changes the relative weights of false positive and false negative from $1 : 9$ to $1 : 99$. Models trained with different weights of FP and FN have different sets of accessible hypotheses and optimization trajectories in the corresponding hypothesis space [3]. For example, a set of parameters that reduces the false negative rate by 1% at the cost of increasing the false positive rate by 20% might be rejected by models trained with $\beta = 0.9$, but considered as an acceptable hypothesis for models trained with $\beta = 0.99$. Thus, by randomizing over β during training, we can promote more diversity in our ensembles, while achieving a minimum amount of false negative error that can be determined empirically by setting the lower bound in the range of β.

2.4. Metric for Model Diversity

To quantify the diversity in an ensemble, we measured the agreement between pairs of models. More specifically, we measured the similarity between two models f_1 and f_2 for both true positive and false positive parts of the predictions, and computed the average across all model pairs in an ensemble:

$$sim(\boldsymbol{p}_1, \boldsymbol{p}_2) = \frac{2\sum_{j=1}^{N} \boldsymbol{p}_1^j \boldsymbol{p}_2^j}{\sum_{j=1}^{N} \boldsymbol{p}_1^j + \sum_{j=1}^{N} \boldsymbol{p}_2^j} \quad (6)$$

where $\boldsymbol{p}^j = \boldsymbol{f}^j \times \boldsymbol{y}^j$ or $\boldsymbol{f}^i \times (1 - \boldsymbol{y}^j)$ depending on whether we wanted the true positive or false positive similarity. Our insight is that in an ideal ensemble, true positive similarity should be high, whereas false positive similarity should be low.

3. Experiments

3.1. Internal Carotid Artery Lumen Segmentation

We first implemented our ensemble learning strategy on the challenging task of internal carotid artery (ICA) lumen segmentation. Segmentation of the ICA is clinically important because different types of plaque tissue confer different degrees of risk of stroke in patients with carotid atherosclerosis. Our IRB approved anonymized data-set consists of 76 Computed Tomography Angiogram images collected at Weill Cornell Medicine/New York Presbyterian Hospital. All CTA images were obtained from patients with unilateral $> 50\%$ extracranial carotid artery stenosis. All ground truth segmentation labels were created by a human expert (H. Ong) with clinical experience. The ICA lumen pixels were annotated within the volume of 5 slices above and below the narrowest part of the internal artery, in the vicinity of the carotid artery bifurcation point. We first used a neural network-based multi-scale landmark detection algorithm to locate the bifurcation [15], and crop a volumetric patch of $72 \times 72 \times 12$ from the original image with $0.35 - 0.7$ mm voxel spacing in x and y directions and $0.6 - 2.5$ mm voxel spacing in the z-direction, preserving the original voxel spacings. For each case, we confirmed that the annotated lumen is included in the patch. The data were then randomly split into 48 training, 12 validation, and 16 test patients. The created 3D image patches were used for all subsequent training and testing. We employed a 3D U-Net as our architecture for all of the models we trained in the ensemble, using the same design choices [6]. In this first experiment, we used the Tversky loss with high β values (we also experimented with balanced cross-entropy, and present those results in Supplemental Material). We used a mini-batch size of 4, and trained our models using the Adam optimizer [10] with a learning rate of 0.0001 for 1000 epochs.

We trained 12 different models with uniform $\beta = 0.95$, and another 12 models with uniformly distributed random β values ranging from 0.9 to 1, resulting in low precision but high recall models. All models were randomly initialized, and a distinct set of training and validation images were used to increase model diversity. We called them *low prec ensemble ($\beta = 0.95$)* and *low prec ensemble (random β)* respectively. We also trained another 12 models with $\beta = 0.5$, creating a *baseline ensemble* of models trained with Dice loss, random initialization, and random train/validation split. Bagging were used for both baseline and low precision ensemble.

Figure 2 plots the quantitative results of the two ensemble strategies (low prec ensemble with random β and baseline ensemble) for different numbers of models in the ensemble. We randomly picked K models from all 12 models to generate the ensemble (where K is between 1 and 10). For each K, we created 10 random K-sized ensembles and computed the mean and standard deviation of the results across the ensembles.

As the number of models included in the ensemble increases, models trained with regular Dice loss show some improvement in Dice score from 0.576 to 0.614, as well as a small increase in terms of precision. On the other hand, as can be seen for $K = 1$ a single low precision model has a relatively low Dice score, but high recall (around 90%). The Dice score improves dramatically from 0.435 to 0.708 as we include more low precision models in the ensemble. The precision also improves substantially, indicating that many of the false positives are canceling each other. The ensemble recall decreases slightly as some of the true positives are removed in the ensemble. We observe that in this dataset, the low precision ensemble's performance (Dice) plateaus around 4 models.

Figure 3 visualizes example results from different low precision models trained with random β. (h) is the ground truth label, and the image has several regions including the internal and external carotid arteries with similar gray values. Note that it is hard even for a human expert to distinguish the internal and external carotid arteries during annotation (see below for inter-human agreement). Fig. 3 (b), (c), (d), (e) and (f) are predictions made by 5 models trained with random β. We can see that they all make different false positive predictions but capture the structure of interest. (g) is the result after applying our ensemble strategy, and it manages to eliminate most of the false positives.

Table 1 lists the average Dice score, recall, and precision for single baseline and low precision models (random βs), two ensemble strategies (with $K = 12$, both fixed β and random β), an additional top performing ensemble baseline M-Heads [21], as well as a secondary manual annotation by another expert who was blinded to the first annotations. By using the baseline ensemble strategy with averaging and
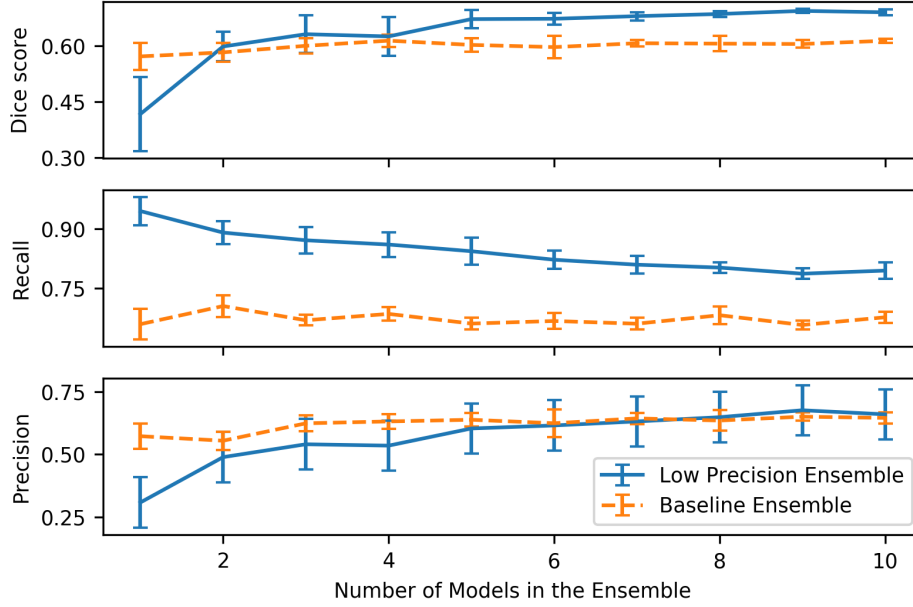
Figure 2: Dice score, recall, and precision for two ensemble strategies vs. the number of models in the ensemble for segmentation of internal carotid artery in neck CTA

Method	Dice	Recall	Precision
Single Baseline Model	0.576 ± 0.302	0.672 ± 0.393	0.563 ± 0.346
Single Low Precision Model	0.435 ± 0.192	0.944 ± 0.220	0.304 ± 0.255
Baseline Ensemble	0.614 ± 0.294	0.665 ± 0.388	0.649 ± 0.324
Low prec Ensemble (β=0.95)	0.643 ± 0.152	0.736 ± 0.191	0.628 ± 0.182
Low Prec Ensemble (random β)	$\mathbf{0.708} \pm 0.170$	0.815 ± 0.212	0.670 ± 0.202
M-Heads [21]	0.655 ± 0.134	0.757 ± 0.152	0.634 ± 0.158
Second Human Expert	0.791 ± 0.140	0.906 ± 0.161	0.714 ± 0.146

Table 1: Performance of different methods for Internal Carotid Artery Segmentation in Neck CTA. Best non-manual dice score is **boldfaced**.

Method	True Positive	False Positive	All Positive
Baseline Ensemble	0.768	0.637	0.696
Low prec Ensemble (β=0.95)	0.951	0.572	0.655
Low Prec Ensemble (random β)	0.954	0.515	0.643

Table 2: Average pairwise model similarity scores of true positive, false positive, and all positive (foreground) predictions for the different ensemble methods in the neck CTA experiment. Lower values indicate more diversity. A good ensemble should have high diversity in its (e.g. false positive) errors, but less diversity in correct predictions.

thresholding at 0.5, we boost the single model dice from 0.576 to 0.614, demonstrating the effectiveness of a regular ensemble strategy in our application. Our low precision ensemble method, on the other hand, is capable of greatly enhancing the performance from a relatively low single model dice of 0.435 to 0.708, utilizing weak segmenters to make a more accurate prediction. We can see that the biggest dice score improvement (from 0.643 to 0.708) comes from using random β instead of a single fixed value. The proposed method also has a better dice score compared to the M-Heads method. Compared to the baseline ensemble method, the low-precision ensemble (random β) has a higher Dice score, a lower false negative rate, and a comparable false positive rate. However, there is still room for improvement,

particularly in recall rates, as can be observed from the second human expert performance.

For our diversity analysis, Table 2 lists the pairwise similarity scores of the true positive, false positive, and all positive (foreground) predictions for different ensemble methods. A higher score means less diverse predictions. Compared to a baseline ensemble, models in the low precision ensemble have a higher (lower) score for their true positive (false positive) predictions. Thus, our low precision ensemble is capable of making more diverse false positive errors but consistent true positive predictions. We observe that the false positive diversity is higher in the random β ensemble, relative to the fixed $\beta = 0.95$ ensemble. With an average
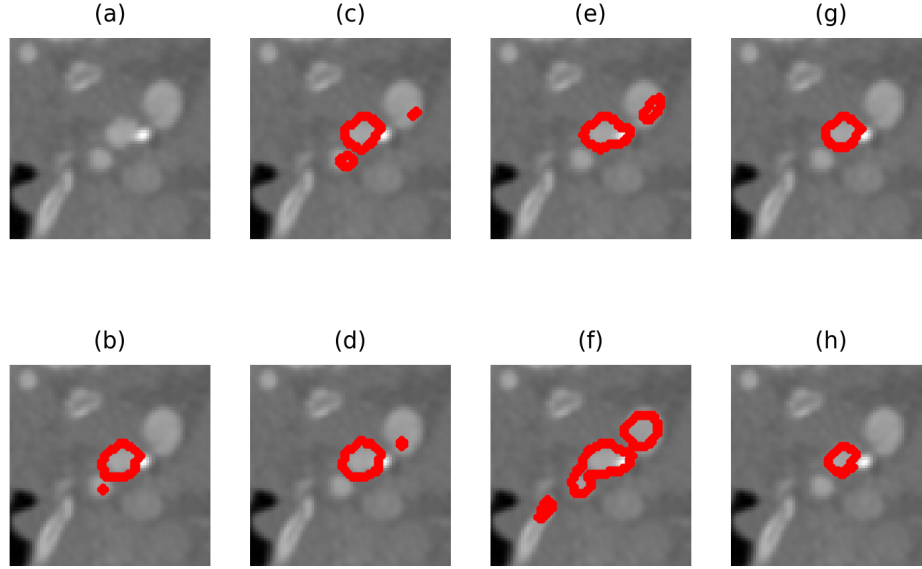
Figure 3: An example CTA image and overlaid predictions of the internal carotid artery. (a) is the raw image. (b),(c),(d),(e),(f) are predictions made by different low precision models. (g) is the ensemble result. (h) manual ground-truth

true positive similarity of ∼ 0.95, the correctly identified internal carotid artery can be mostly preserved in low precision ensembles.

3.2. Myocardium Segmentation

In our second experiment, we employed the dataset from the HVSMR 2016: MICCAI Workshop on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI [17]. This dataset consists of 10 3D cardiovascular magnetic resonance (CMR) images. The image dimension and voxel spacing varied across subjects, and averaged at $390 \times 390 \times 165$ and $0.9 \times 0.9 \times 0.85$ mm. The ground truth manual segmentations of both the blood pool and ventricular myocardium are provided. In our experiment, we focused on the myocardium segmentation because it is a more challenging task with a low state-of-the-art average dice score. Before training, certain pre-processing steps were carried out. We first normalized all images to have zero mean and unit variance intensity values. Data augmentation was also performed via image rotation and flipping in the axial plane. We implemented a 5 fold cross-validation, holding 2 images for testing and 8 images for training.

To demonstrate that our method is not restricted to a spe-

Method	Dice	Recall	Precision
Single Baseline Model [28]	0.786 ± 0.045	0.845 ± 0.047	0.747 ± 0.081
Single Low Precision Model	0.757 ± 0.065	0.974 ± 0.062	0.607 ± 0.121
Baseline Ensemble	0.790 ± 0.033	0.872 ± 0.022	0.750 ± 0.078
Low Prec Ensemble (β=0.95)	0.796 ± 0.046	0.904 ± 0.028	0.715 ± 0.092
Low Prec Ensemble (random β)	$\mathbf{0.815} \pm 0.052$	0.949 ± 0.034	0.719 ± 0.097

Table 3: Performance for Segmentation of Ventricular Myocardium in MRI. Best dice score is **boldfaced**.

cific network architecture and loss function, we adopted the network and method used by the challenge winner [28]. We implemented the 3D FractalNet, with the same parameters and experimental settings proposed by the authors, trained with regular cross-entropy loss. We trained 12 different models, and we call this the *Baseline Ensemble*. To train models with high recall and low precision, we used balanced binary cross-entropy loss with random $\beta \in [0.9, 1)$. Note that results obtained with Tversky loss are presentede in Supplemental Material. We trained 12 different models to create the low precision ensemble (random β) and another 12 models with fixed β=0.95.

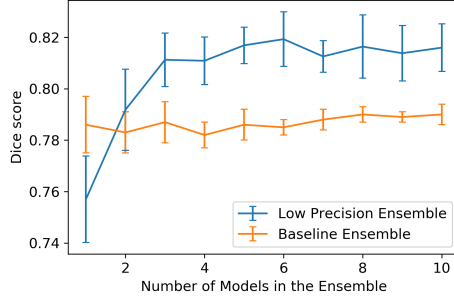Similar to the previous experiment, we randomly picked

Figure 4: Performance of Low-Precision Ensemble vs Number of Models: Segmentation of Ventricular Myocardium in MRI

Method	True Positive	False Positive	All Positive
Baseline Ensemble	0.926	0.760	0.898
Low prec Ensemble (β=0.95)	0.971	0.644	0.842
Low Prec Ensemble (random β)	0.974	0.621	0.818

Table 4: Average pairwise model similarity scores of true positive, false positive, and all positive (foreground) predictions for the different ensemble methods in the myocardium segmentation experiment. Lower values indicate more diversity. A good ensemble should have high diversity in its (e.g. false positive) errors, but less diversity in correct predictions.

K models from all 12 models to generate an ensemble, where K goes from 1 to 10. For each K, we created 10 random K-sized ensembles, and computed the mean and standard deviation of the results across the ensembles (see Figure 4). Table 3 shows the experimental results of the average of 5 fold testing. The low precision ensemble model trained with fixed $\beta = 0.95$ value does not show significant improvement over the regular ensemble. The Dice score improves from 0.790 (for the state-of-the-art baseline) to 0.815 (for the low precision ensemble with random β). The low precision ensemble with random β method has higher Dice score and recall, but lower precision. We also observe a higher improvement in terms of average dice scores from single models to the ensemble.

Additionally, we perform a diversity analysis for different ensemble models. As we observe in Table 4, models in the low precision ensemble have more consistent true positive predictions but more diverse false positive errors.

3.3. Multiple Sclerosis (MS) Lesion Segmentation

We conduct our third experiment on MS lesion segmentation. MS is a chronic, inflammatory demyelinating disease of the central nervous system in the brain. Pre-

cise segmentation can help characterize MS lesions and provide important markers for clinical diagnosis and disease progress assessment [29]. However, MS lesion segmentation is challenging and complicated as lesions vary vastly in terms of location, appearance, and shape. Concurrent hyper-intensities make MS lesion tracing more difficult even for experienced neural radiologists (as shown in Fig 5). Dice score between masks traced by two raters from a ISBI dataset is only 0.732 [5].

We employ the dataset from ISBI 2015 Longitudinal MS Lesion Segmentation Challenge [5] to verify our method. The ISBI dataset contains MRI scans from 19 subjects, and each subject has 4-6 time-point scans. For each scan, FLAIR, PD-weighted, T2-weighted, and T1-weighted images are provided. All image modalities are skull-stripped, dura-stripped, N4 inhomogeneity corrected, and rigidly co-registered to a $1mm$ isotropic MNI template. Each image contains 182 slices with $FOV = 182 \times 256$. Two experienced raters manually traced all lesions, so there are two gold-standard masks. To our knowledge, using the intersection of the two masks to train our model yields the best performance. Only 5 training subjects (21 images) have publicly available gold-standard masks. We can evaluate our model on an online website by submitting predicted results of the remaining 14 subjects (61 images). Unlike ICA lumen and myocardium segmentation, MS patients usually contain numerous lesions with different sizes. Thus, to evaluate the performance of the method, we apply metrics of lesion-wise precision (L-precision) and lesion-wise recall (L-recall), which are defined as follows.

- Lesion-wise Precision

$$\text{L-Precision} = \frac{LTP}{PL}, \quad (7)$$

- Lesion-wise Recall

$$\text{L-Recall} = \frac{LTP}{GL}, \quad (8)$$

where LTP denotes the number of lesion-wise true positives, GL is the total number of lesions in the gold-standard segmentation, and PL is the total number of lesions in the predicted segmentation. We calculate lesion-wise F_1 score as harmonic mean of lesion-wise recall and precision.

We compare our ensemble method with four recent works [30, 1, 25, 7] on MS lesion segmentation. We build our baseline network architecture based on a publicly available implementation [30] designed for MS lesion segmentation. Other three methods for comparison are multi-branch network (MB-Net) [1], multi-scale network (MS-Net) [7], and cascaded-network (CD-Net) [25]. Similar to myocardium segmentation, all images are normalized to have zero mean and unit variance intensity values. We use random crop, intensity shifting, and elastic deformation to augment our data.
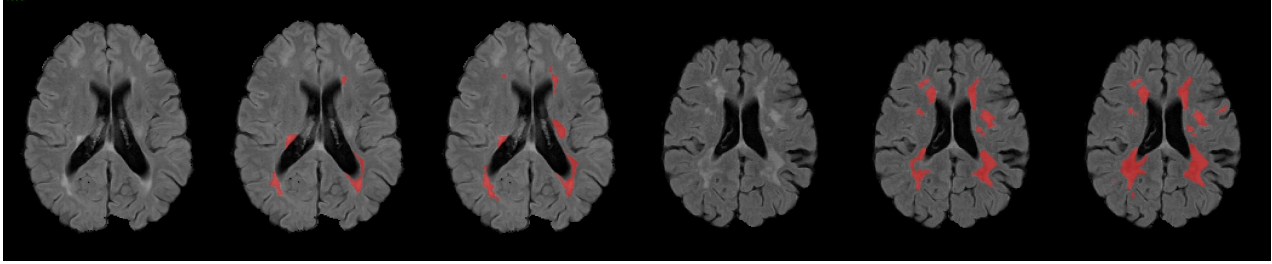
Figure 5: Example FLAIR images and corresponding masks traced by two human experts and marked in red. The left three are a flair image, 1st rater's mask, 2nd rater's mask from subject-01's first time-point scan. The right three are from subject-02's first time-point scan. We can see from the figure, besides ms lesion areas, there are many other hyperintensities that can confuse algorithms or even human experts.

Method	Dice	L-Recall	L-Precision	L-F_1
Baseline Ensemble [30]	0.624	0.458	0.889	0.605
Low Prec Ensemble ($\beta = 0.95$)	0.645	0.473	0.823	0.601
Low Prec Ensemble (random β)	**0.662**	0.491	0.849	**0.622**
MB-Net [1]	0.611	0.410	0.860	0.555
CD-Net [25]	0.630	0.367	0.847	0.512
MS-Net [7]	0.501	0.429	0.434	0.431

Table 5: Performance comparison for segmenting MS lesions in brain MRI. Best dice and L-F_1 scores are **bold-faced**.

Method	Pairwise Similarity of Foreground Pixels
Baseline Ensemble	0.814
Low prec Ensemble (β=0.95)	0.753
Low Prec Ensemble (random β)	0.737

Table 6: Pairwise model similarity scores of foreground predictions for the different ensemble methods. Lower values indicate more diversity.

We trained 10 different models with regular dice loss, 10 models using Tversky loss with random high $\beta \in [0.9, 1)$, and another 10 models with high $\beta = 0.95$ [1]. We refer these three methods as *Baseline Ensemble*, *Low Prec Ensemble (random β)* and *Low Prec Ensemble ($\beta = 0.95$)* in Table 5.

We can see from Table 5 that compared with the baseline ensemble model, the low-precision ensemble achieves a slightly higher lesion-wise F_1 score. In terms of overall dice score, the low precision ensemble is 6% higher than the baseline ensemble. Also, compared to the recently proposed MB-Net [1] , CD-Net [25], and MS-Net [7], the proposed low precision ensemble exhibits superior performance in all aspects. We also note that randomizing β improves the quality of segmentations, yielding a dice score boost of 1.7 points and an increase in lesion-wise F_1 score.

Because we do not have ground truth labels for the test images, we cannot show diversity measurements for true positive and false positive predictions. For overall positive predictions, however, the pairwise similarity scores are listed in Table 6. We observe that the low precision ensemble achieves the lowest among all three methods, indicating a more diverse set of results generated by models trained with random high β's.

4. Conclusion

In this paper, we presented a novel low-precision ensembling strategy for binary image segmentation. Similar to regular ensemble learning, predictions from multiple models are combined by averaging.

However, in contrast to regular ensemble learning, we encourage the individual models to have a high recall, typically at the expense of low precision and accuracy. Our goal is to have a diverse ensemble of models that largely capture the foreground pixels, but make different types of false positive predictions that can be canceled after averaging.

We conducted experiments on three different data-sets, with different loss functions and network architectures. The proposed method achieves better Dice score compared to using a single model or a regular ensembling strategy that does not combine high recall models. We believe that our method can be applied to a wide range of hard segmentation problems, with different loss functions and architectures. Finally, the proposed strategy can also be used in other types of challenging classification problems.

Acknowledgments

---

[1] Results obtained with balanced cross-entropy loss are included in the Supplemental Material.

References

[1] Shahab Aslani, Michael Dayan, Loredana Storelli, Massimo Filippi, Vittorio Murino, Maria A Rocca, and Diego Sona. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *NeuroImage*, 196:1–15, 2019.

[2] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[3] Gavin Brown. *Diversity in neural network ensembles*. PhD thesis, Citeseer, 2004.

[4] Peng Cao, Jinzhu Yang, Wei Li, Dazhe Zhao, and Osmar Zaiane. Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule cad. *Computerized Medical Imaging and Graphics*, 38(3):137–150, 2014.

[5] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[7] Mohsen Ghafoorian, Nico Karssemeijer, Tom Heskes, Mayra Bergkamp, Joost Wissink, Jiri Obels, Karlijn Keizer, Frank-Erik de Leeuw, Bram van Ginneken, Elena Marchiori, et al. Deep multi-scale location-aware 3d convolutional neural networks for automated detection of lacunes of presumed vascular origin. *NeuroImage: Clinical*, 14:391–399, 2017.

[8] Konstantinos Kamnitsas, Wenjia Bai, Enzo Ferrante, Steven McDonagh, Matthew Sinclair, Nick Pawlowski, Martin Rajchl, Matthew Lee, Bernhard Kainz, Daniel Rueckert, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. In *International MICCAI Brainlesion Workshop*, pages 450–462. Springer, 2017.

[9] Anuj Karpatne, Ankush Khandelwal, and Vipin Kumar. Ensemble learning methods for binary classification with multimodality within the classes. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 730–738. SIAM, 2015.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in neural information processing systems*, pages 231–238, 1995.

[12] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017.

[13] Chaofeng Li, Guoce Zhu, Xiaojun Wu, and Yuanquan Wang. False-positive reduction on lung nodules detection in chest radiographs by ensemble of convolutional neural networks. *IEEE Access*, 6:16060–16067, 2018.

[14] Hongwei Li, Gongfa Jiang, Jianguo Zhang, Ruixuan Wang, Zhaolei Wang, Wei-Shi Zheng, and Bjoern Menze. Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images. *NeuroImage*, 183:650–665, 2018.

[15] Tianyu Ma, Ajay Gupta, and Mert R Sabuncu. Volumetric landmark detection with a multi-scale shift equivariant neural network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 981–985. IEEE, 2020.

[16] Debapriya Maji, Anirban Santara, Pabitra Mitra, and Debdoot Sheet. Ensemble of deep convolutional neural networks for learning to detect retinal vessels in fundus images. *arXiv preprint arXiv:1603.04833*, 2016.

[17] Danielle F Pace, Adrian V Dalca, Tal Geva, Andrew J Powell, Mehdi H Moghari, and Polina Golland. Interactive whole-heart segmentation in congenital heart disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 80–88. Springer, 2015.

[18] Jong Won Park et al. Connectivity-based local adaptive thresholding for carotid artery segmentation using mra images. *Image and Vision Computing*, 23(14):1277–1287, 2005.

[19] Robi Polikar. Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer, 2012.

[20] 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. Springer, 2015.

[21] Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3591–3600, 2017.

[22] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 379–387. Springer, 2017.

[23] Neeraj Sharma and Lalit M Aggarwal. Automated medical image segmentation techniques. *Journal of medical physics/Association of Medical Physicists of India*, 35(1):3, 2010.

[24] Atsushi Teramoto, Hiroshi Fujita, Osamu Yamamuro, and Tsuneo Tamaki. Automated detection of pulmonary nodules in pet/ct images: Ensemble false-positive reduction using a convolutional neural network technique. *Medical physics*, 43(6Part1):2821–2827, 2016.

[25] Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. *NeuroImage*, 155:159–168, 2017.

[26] Shuangling Wang, Yilong Yin, Guibao Cao, Benzheng Wei, Yuanjie Zheng, and Gongping Yang. Hierarchical retinal

blood vessel segmentation based on feature and ensemble learning. *Neurocomputing*, 149:708–717, 2015.

[27] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[28] Lequan Yu, Xin Yang, Jing Qin, and Pheng-Ann Heng. 3d fractalnet: dense volumetric segmentation for cardiovascular mri volumes. In *Reconstruction, segmentation, and analysis of medical images*, pages 103–110. Springer, 2016.

[29] Hang Zhang, Jinwei Zhang, Rongguang Wang, Qihao Zhang, Pascal Spincemaille, Thanh D Nguyen, and Yi Wang. Efficient folded attention for 3d medical image reconstruction and segmentation. *arXiv preprint arXiv:2009.05576*, 2020.

[30] Hang Zhang, Jinwei Zhang, Qihao Zhang, Jeremy Kim, Shun Zhang, Susan A Gauthier, Pascal Spincemaille, Thanh D Nguyen, Mert Sabuncu, and Yi Wang. Rsanet: Recurrent slice-wise attention network for multiple sclerosis lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 411–419. Springer, 2019.

[31] Zhilu Zhang, Adrian V Dalca, and Mert R Sabuncu. Confidence calibration for convolutional neural networks using structured dropout. *arXiv preprint arXiv:1906.09551*, 2019.