More Birds in the Hand - Medical Image Segmentation using a Multi-Model Ensemble Framework

Yung-Han Chen1, Pei-Hsuan Kuo2, Yi-Zeng Fang3 and Wei-Lin Wang4

1. Department of Electronics Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, contact email: g0410440@gmail.com
2. Institute of Electronics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
3. Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
4. Institute of Electronics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

Abstract
In this paper, we introduce a multi-model ensemble framework for medical image segmentation. We first collect a set of state-of-the-art models in this field and further improve them through a series of architecture refinement moves and a set of specific training skills. We then integrate these fine-tuned models into a more powerful ensemble framework. Preliminary experiment results show that the proposed multi-model ensemble framework performs well under the given polyp and instrument datasets.

Keywords: medical image segmentation, polyp segmentation, instrument segmentation, model ensemble

Introduction
Most state-of-the-art (SOTA) models dealing with medical image segmentation are based on either convolutional neural networks (CNN) or Transformers. Besides, no single model outperforms all others across all public datasets, suggesting each model has unique strengths. Hence, it seems a good idea to integrate a set of SOTA models into an even more powerful ensemble framework. To that end, we select a CNN-based model HarDNet-MSEG [1] and a Transformer-based model TransFuse [2] as baseline models. To fully exploit the advantages of selected models, we refine them first and then create their variants, where each variant is carefully fine-tuned for specific targets. In the end, we have multiple high-quality models in hand for further ensemble. In addition, we apply several data augmentation techniques to the given training dataset and then utilize the augmented ones to further improve models included in the final ensemble framework.

Materials and methods
Data pre-processing and augmentation
Without a rich set of training data, it is generally difficult to shape a model very well. Since the training data is limited, common augmentation techniques are adopted here, such as random image flips and rotations. It is also reported that the datasets after augmentation can help alleviate the possible over-fitting problem. In addition to the limited training data issue, we found the model accuracy is also bound due to severe deformation of resized training images. To resolve this issue, an image is properly zero-padded before resizing to preserve the original aspect ratio. This simple strategy effectively boosts the accuracy according to our experiments.

Baseline models and their variants
TransFuse: In the original paper [2], the authors use BiFuse layers to assemble features extracted from CNN and Transformer. Several attempts have been made for better performance. We replace the original backbone of TransFuse, ResNet-34 [3], with another powerful backbone model HarDNet-85 [4] to improve the model accuracy. Another try is to include four BiFuse layers instead of the original three for extracted feature merging. An extra BiFuse layer provides a meaningful improvement according to our experiments. To unify the resolution, all input images are resized to 352×352 for better outcomes. After investigating several optimizers [5, 6, 7], Ranger [5] becomes our final choice since it not only reaches a convergence faster but provides more accurate prediction outcomes. Besides, the default activation function is ReLU [8], while we replace ReLU with Mish [9] to further increase the accuracy.

© 2021 Author(s). This is an open access article licensed under the Creative Commons Attribution License 4.0. (http://creativecommons.org/licenses/by/4.0/).
HarDNet-MSEG: The primary target of the original model [1] is to achieve an extremely fast inference speed and fairly good accuracy. Hence, on the one hand, we still try to equip it with a more powerful backbone, e.g., HarDNet-85 [4] and ResNet101 [3], to see if the prediction accuracy can be improved. We observed that the decoder part of HarDNet-MSEG is not powerful enough to deal with various image resolutions. Therefore, we replace it with the context module and axial reverse attention, which have been used in CaraNet [10], to further enhance the accuracy. As for the receptive field, we adopt the SPP part of YOLOv4 [11] to replace the original RFB part in HarDNet-MSEG. It should not be a surprise that models with different backbones, decoders, and receptive fields perform differently. Therefore, a well-trained multi-model ensemble framework should be able to produce better outcomes than any single one.

Ensemble: According to our experiments, no single model can always outperform the others on various datasets, which suggests the attributes of input images do affect the model performance. As a consequence, a multi-model ensemble framework is a good strategy to exploit the full advantages of all models. The key idea of our ensemble framework is to properly weight the results of all models and then integrate them into the final one based on the features of an input image (such as color, resolution).

Post-processing
During the inference phase, to suppress erroneous predictions, especially false positive ones, we adopt the “test-time augmentation” technique. It is an inference method that integrates outputs of several augmentations as the final result. Four augmentation skills are utilized here: horizontal/vertical flipping and clockwise/counter-clockwise rotation by 90 degrees. In addition, the global threshold is carefully fine-tuned and the final result of each pixel is jointly determined by test-time augmentation through voting, which helps eliminate less confident pixels and thus produces masks with higher mean IoU.

Results
Our submission includes the following five models:
(1) Ensemble with different lateral-maps and resolutions (HarDNet-MSEG based)
(2) HarDNet-MSEG with ResNet-101 backbone
(3) Ensemble with Model(1), Model(2), and Model(2) with HarDNet-85 backbone
(4) TransFuse with HarDNet-85 backbone
(5) Ensemble with four different models (HarDNet-85, ResNet-101 with different receptive fields)

The following table shows that a multi-model ensemble produces better results in general.

<table>
<thead>
<tr>
<th>Instruments</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td>0.9729</td>
<td>0.9737</td>
<td><strong>0.9742</strong></td>
<td>0.9718</td>
<td>0.9735</td>
</tr>
<tr>
<td>jaccard</td>
<td>0.8412</td>
<td>0.8339</td>
<td>0.8393</td>
<td>0.8022</td>
<td><strong>0.8418</strong></td>
</tr>
<tr>
<td>dice</td>
<td>0.8882</td>
<td>0.8800</td>
<td>0.8856</td>
<td>0.8542</td>
<td><strong>0.8885</strong></td>
</tr>
<tr>
<td>fl</td>
<td>0.8882</td>
<td>0.8800</td>
<td>0.8856</td>
<td>0.8452</td>
<td><strong>0.8885</strong></td>
</tr>
<tr>
<td>recall</td>
<td><strong>0.8847</strong></td>
<td>0.8786</td>
<td>0.8753</td>
<td>0.8324</td>
<td>0.8794</td>
</tr>
<tr>
<td>precision</td>
<td>0.9310</td>
<td>0.9224</td>
<td>0.9297</td>
<td>0.9223</td>
<td><strong>0.9319</strong></td>
</tr>
</tbody>
</table>

Figure 1: The overall architecture of the proposed multi-model ensemble framework. The left-hand side illustrates the flow of how to determine the weight of every model in use based on the input image, while the right-hand side represents actual models included in the framework.

Figure 2: Result visualization on two tasks.

Discussion
The experiment results related to polyp segmentation clearly show that HarDNet-MSEG variants perform better than TransFuse ones, and model ensembles outperform each of the single models. The results also suggest that the proposed model ensemble framework can truly increase the prediction accuracy of medical imaging segmentation. A wise proverb once says, "A bird in the hand is worth two in the bush." Nevertheless, why bother if you have more birds in the hand?
References


