Adversarial Vision Transformer for Medical Image Semantic Segmentation with Limited Annotations

Ziyang Wang, Chengkuan Zhao, Zixuan Ni

Introduction

The goal of image semantic segmentation is to classify each pixel of an input image as to whether or not it is part of a Region Of Interest (ROI) or background. We present practical avenues for training a Computationally-Efficient Semi-Supervised Vision Transformer (ViT) for medical image segmentation task.

Key Contributions

Combine two SSL styles: ICT and adversarial, for medical image segmentation
Propose a dual-view co-training semi-supervised learning framework for ViT
Demonstrate serviceable results with a proportion of labelled data as small as 2% of the total data

Qualitative Results

Algorithm

1. Two segmentation ViTs with the same architecture are initialized separately to train on unlabeled samples. The third is initialized as an adversarial evaluation model.

2. The mixup of two unlabeled samples are used as labels for training of two individual unlabeled samples, and then the result is sent to evaluation model, to optimize the loss between unlabeled data and labeled data.

3. The training objective is to minimize the sum of supervision loss and semi-supervision loss of the two ViTs

   \[ \mathcal{L}_{sup} = \mathcal{L}(x_{sup}, y_{sup}; \theta) = \mathcal{D}(x_{sup}; \theta) + \mathcal{C}(x_{sup}; \theta) \]

   \[ \mathcal{L}_{ss} = \mathcal{D}(x_{ss}; \theta) + \mathcal{C}(x_{ss}; \theta) \]

   \[ \mathcal{D}(x; \theta) = -\frac{1}{|X|} \sum_{x \in X} \frac{1}{2} \sum_{c \in C} p(c|x; \theta) \log \hat{p}(c|x; \theta) \]

4. The weight for semi-supervision loss is updated every 150 iterations, also using a ramp-up function

5. The ViT with the best performance on the validation set is used for the final evaluation

Quantitative Results

Table 1: Direct Comparison of Semi-supervised Frameworks on MRI Cardiac Test Set

<table>
<thead>
<tr>
<th>Framework</th>
<th>Acc′</th>
<th>Acc′′</th>
<th>Acc′′′</th>
<th>Spec</th>
<th>Spec′′</th>
<th>Spec′′′</th>
<th>Sens</th>
<th>Sens′′</th>
<th>Sens′′′</th>
<th>Dice</th>
<th>Dice′′</th>
<th>Dice′′′</th>
<th>Sensitivity</th>
<th>Sensitivity′′</th>
<th>Sensitivity′′′</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTB-ViTv</td>
<td>0.684</td>
<td>0.579</td>
<td>0.494</td>
<td>0.834</td>
<td>0.675</td>
<td>0.613</td>
<td>0.591</td>
<td>0.565</td>
<td>0.532</td>
<td>0.758</td>
<td>0.705</td>
<td>0.68</td>
<td>0.758</td>
<td>0.714</td>
<td>0.696</td>
</tr>
<tr>
<td>DAN-ViT</td>
<td>0.622</td>
<td>0.575</td>
<td>0.492</td>
<td>0.825</td>
<td>0.622</td>
<td>0.582</td>
<td>0.556</td>
<td>0.532</td>
<td>0.748</td>
<td>0.703</td>
<td>0.678</td>
<td>0.663</td>
<td>0.748</td>
<td>0.714</td>
<td>0.696</td>
</tr>
<tr>
<td>ICT-ViT</td>
<td>0.664</td>
<td>0.713</td>
<td>0.594</td>
<td>0.849</td>
<td>0.666</td>
<td>0.625</td>
<td>0.603</td>
<td>0.572</td>
<td>0.773</td>
<td>0.731</td>
<td>0.704</td>
<td>0.693</td>
<td>0.773</td>
<td>0.731</td>
<td>0.704</td>
</tr>
<tr>
<td>ADVENT-ViT</td>
<td>0.642</td>
<td>0.739</td>
<td>0.623</td>
<td>0.832</td>
<td>0.673</td>
<td>0.627</td>
<td>0.605</td>
<td>0.572</td>
<td>0.764</td>
<td>0.728</td>
<td>0.699</td>
<td>0.689</td>
<td>0.764</td>
<td>0.728</td>
<td>0.699</td>
</tr>
<tr>
<td>DCN-ViT</td>
<td>0.678</td>
<td>0.764</td>
<td>0.660</td>
<td>0.851</td>
<td>0.709</td>
<td>0.670</td>
<td>0.651</td>
<td>0.628</td>
<td>0.781</td>
<td>0.748</td>
<td>0.724</td>
<td>0.714</td>
<td>0.781</td>
<td>0.748</td>
<td>0.724</td>
</tr>
</tbody>
</table>

References


Email: cz015@bucknell.edu