Adversarial Vision Transformer for Medical Image Semantic Segmentation with Limited Annotations

BMVC 2022

Paper ID:1002

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.

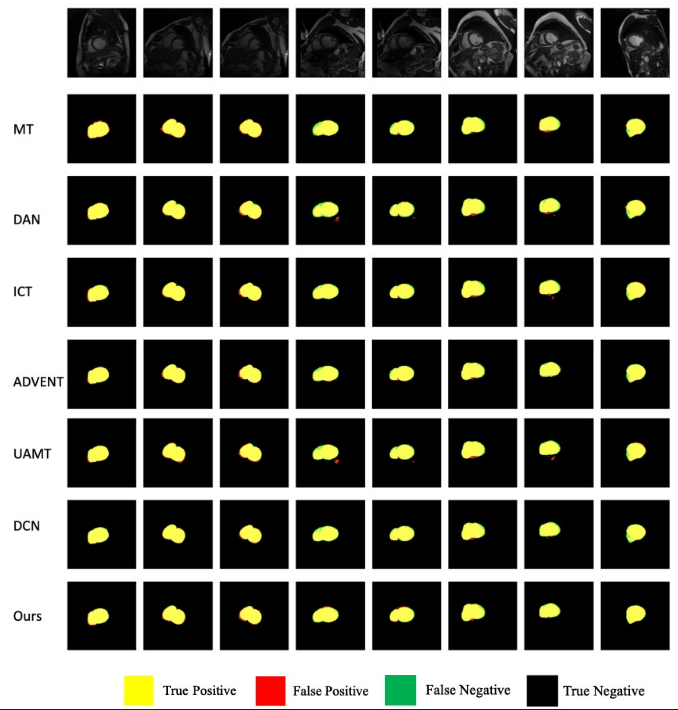
Key Contributions

Combine two SSL styles: ICT and adversarial, for medical image segmentation

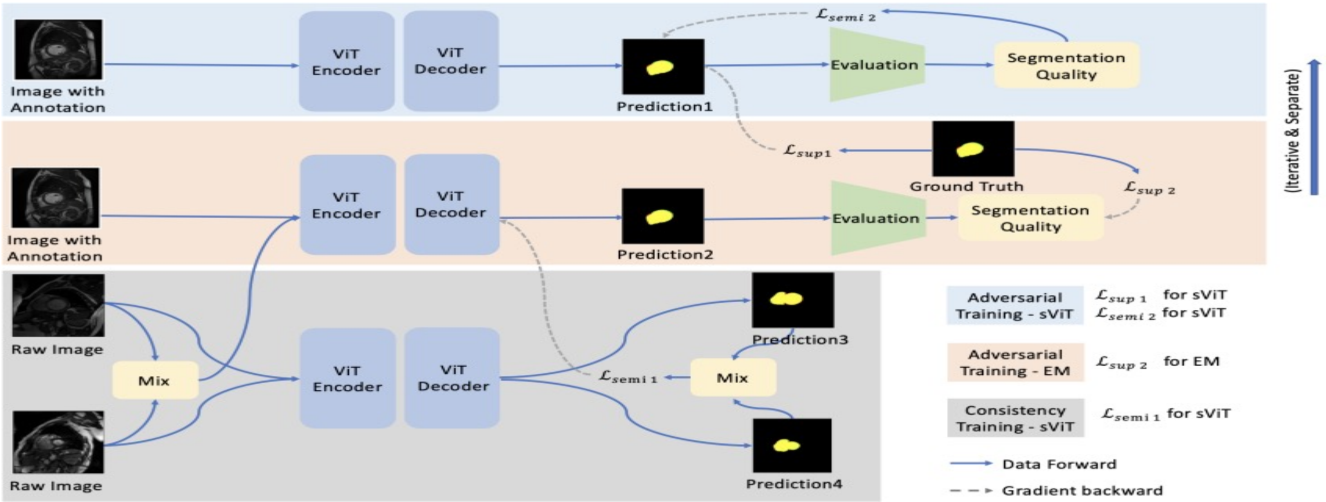
Propose a dual-view co-training semi-supervised learning framework for ViT

Demonstrate serviceable results with a proportion of labelled data as small as 2% of the total data

Qualitative Results



Framework



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.

$$\mathcal{L} = \mathcal{L}_{\text{sup1}} + \mathcal{L}_{\text{sup2}} + \lambda_1 \mathcal{L}_{\text{semi1}} + \lambda_2 \mathcal{L}_{\text{semi2}}$$

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

The supervision loss is based on Dice and Cross-Entropy

$$\mathcal{L}_{\text{sup1}} = \text{CE}(Y_{\text{gt}}, f_{ViT}(\boldsymbol{X}_l;\boldsymbol{\theta})) + \text{Dice}(Y_{\text{gt}}, f_{ViT}(\boldsymbol{X}_l;\boldsymbol{\theta}))$$

The training of evaluation model is based on two BCE losses

$$\mathcal{L}_{\text{sup2}} = \text{BCE}(f_{CNN}(f_{ViT}(\boldsymbol{X}_l;\theta_{ViT}),\boldsymbol{X}_l;\theta_{CNN}),1) + \text{BCE}(f_{CNN}(f_{ViT}(\boldsymbol{X}_u;\theta_{ViT}),\boldsymbol{X}_u;\theta_{CNN}),0)$$

The training of two segmentation models is based on Cross-Entropy

$$\mathcal{L}_{\text{semi1}} = \text{CE}(f_{ViT}(Mix_\lambda(X_{u1},X_{u2});\theta), f_{ViT}(\boldsymbol{X}_l;\theta), Mix_\lambda(f_{ViT}(X_{u1};\overline{\theta}), f_{ViT}(X_{u2};\overline{\theta})))$$

The update of evaluation model is based on Cross-Entropy

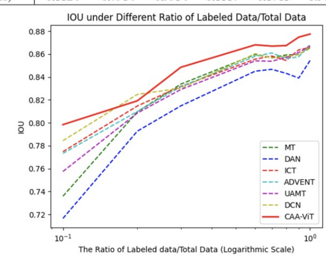
$$\mathcal{L}_{\text{semi1}} = \text{CE}(f_{ViT}(Mix_\lambda(X_{u1},X_{u2});\theta), f_{ViT}(\boldsymbol{X}_l;\theta), Mix_\lambda(f_{ViT}(X_{u1};\overline{\theta}), f_{ViT}(X_{u2};\overline{\theta})))$$

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

Framework	mDice↑	mIOU↑	Acc↑	Pre↑	Sen↑	Spe↑	HD↓	ASD↓
MT[ ]+ViT	0.8384	0.7359	0.9939	0.8362	0.8415	0.9716	23.7689	2.2801
DAN[ ]+ViT	0.8232	0.7165	0.9932	0.8243	0.8222	0.9686	23.5824	2.8411
ICT[ ] +ViT	0.8663	0.7748	0.9948	0.8607	0.8722	**0.9775**	22.5653	2.2404
ADVENT[ ]+ViT	0.8654	0.7733	0.9949	0.8689	0.8625	0.9743	20.4493	1.9022
UAMT[ ] +ViT[ ]	0.8542	0.7575	0.9945	0.8523	0.8575	0.9752	23.6615	2.2159
DCN[ ] +ViT	0.8728	0.7844	0.9950	0.8770	0.8690	0.9738	22.3759	1.9762
CAA-ViT(ours)	**0.8824**	**0.7984**	**0.9954**	**0.8887**	**0.8765**	0.9748	**18.9200**	**1.7587**



References

[MT] Tarvainen, Antti, and Harri Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results." Advances in neural information processing systems. 2017.

[DAN] Zhang, Yizhe, et al. "Deep adversarial networks for biomedical image segmentation utilizing unannotated images." International conference on medical image computing and computer-assisted intervention. 2017.

[ADVENT] Vu, Tuan-Hung, et al. "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[UAMT] Yu, Lequan, et al. "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. 2019.

[DCN] Qiao, Siyuan, et al. "Deep co-training for semi-supervised image recognition." Proceedings of the European conference on computer vision. 2018.

[ICT] Verma, Vikas, et al. "Interpolation consistency training for semi-supervised learning." Neural Networks. 2022.

[ViT] Cao, Hu, et al. "Swin-unet: Unet-like pure transformer for medical image segmentation." arXiv preprint arXiv:2105.05537. 2021.

Email: cz015@bucknell.edu