Multi-scale Hierarchical Vision Transformer with Cascaded Attention Decoding for Medical Image Segmentation

Md Mostafijur Rahman and Radu Marculescu (University of Texas at Austin)

Introduction

- Transformers have shown great success in medical image segmentation. However, transformers [1-3] may exhibit a *limited* generalization ability due to the underlying *single-scale* self-attention calculation mechanism.
- We address this issue by introducing a Multi-scale hierarchical vIsion Transformer (MERIT) network, which *improves* model generalizability by performing self-attention at *multiple scales*.

Contributions

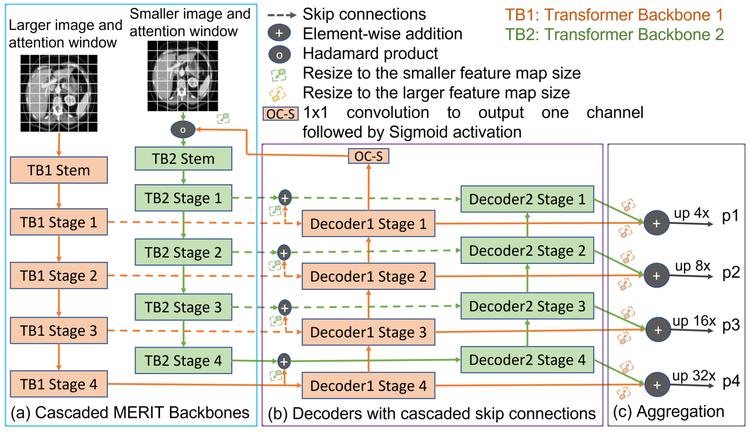
New Network Architecture

- A novel multi-scale hierarchical vision transformer (MERIT) which captures both multiscale and multi-resolution features.
- Two designs: 1) Cascaded MERIT and 2) Parallel MERIT

Multi-stage Feature-mixing Loss Aggregation

- A simple, yet effective way (i.e., MUTATION), for implicit ensembling by mixing features during loss calculation.

Cascaded MERIT Architecture



p1, p2, p3, and p4 are the aggregated multi-stage prediction maps. We can get the Parallel MERIT architecture by removing early feedback (through OC-S) from backbone 1 to backbone 2 and cascaded skip connections.

Multi-stage Feature-mixing Loss Aggregation

Algorithm 1: Multi-stage Feature-Mixing Loss Aggregation

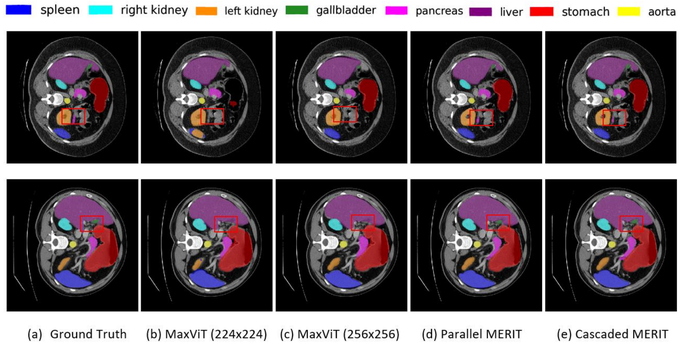
Input: y; the ground truth mask
A list $[P_i]$; $i = 0, 1, \cdots, n-1$, where each element is a prediction map
Output: *loss*; the aggregated loss

$loss \leftarrow 0.0$;
$\mathcal{S} \leftarrow$ find all non-empty subsets of prediction map indices, $\{0, \ldots, n-1\}$; // $\mathcal{S}$ is the set of non-empty subsets of $\{0, \ldots, n-1\}$
foreach $s \in \mathcal{S}$ **do**
  $\hat{y} \leftarrow 0.0$; // $\hat{y}$ is a new prediction map
  foreach $i \in s$ **do**
    $\hat{y} \leftarrow \hat{y} + P_i$;
  end
  $loss \leftarrow loss\_function(y, \hat{y})$; // $loss\_function(.)$ is any loss function (e.g., CrossEntropy, DICE)
end

Results on Synapse Multi-organ Dataset

Architectures	Average DICE↑	Average HD95↓	Aorta	GB	KL	KR	Liver	PC	SP	SM
UNet (Ronneberger et al., 2015)	70.11	44.69	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
AttnUNet (Oktay et al., 2018)	71.70	34.47	82.61	61.94	76.07	70.42	87.54	46.70	80.67	67.66
R50+UNet (Chen et al., 2021)	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50+AttnUNet (Chen et al., 2021)	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
SSFormerPVT (Wang et al., 2022b)	78.01	25.72	82.78	63.74	80.72	78.11	93.53	61.53	87.07	76.61
PolypPVT (Dong et al., 2021)	78.08	25.61	82.34	66.14	81.21	73.78	94.37	59.34	88.05	79.4
TransUNet (Chen et al., 2021)	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUNet (Cao et al., 2021)	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
MT-UNet (Wang et al., 2022a)	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MISSFormer (Huang et al., 2021)	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
CASTformer (You et al., 2022)	82.55	22.73	**89.05**	67.48	86.05	82.17	**95.61**	67.49	91.00	81.55
PVT-CASCADE (Rahman et al., 2023)	81.06	20.23	83.01	70.59	82.23	80.37	94.08	64.43	90.1	83.69
TransCASCADE (Rahman et al., 2023)	82.68	17.34	86.63	68.48	87.66	84.56	94.43	65.33	90.79	83.52
Parallel MERIT (Ours)	84.22	16.51	88.38	73.48	87.21	84.31	95.06	69.97	91.21	84.15
Cascaded MERIT (Ours)	**84.90**	**13.22**	87.71	**74.40**	**87.79**	**84.85**	95.26	**71.81**	**92.01**	**85.38**

MERIT results reported for MERIT + CASCADE decoder (Additive) + MUTATION. Cascaded MERIT outperforms SOTA methods, such as TransUNet [1] and SwinUNet [2] by 7.42% and 5.57%, respectively.



MERIT architecture can segment both small (red rectangular box) and large organs better than the single scale MaxViT for both 224×224 and 256×256 input resolutions.

Effect of Multi-scale Self-Attention

Architectures	Input Resolutions	Attention Windows	Params (M)/ FLOPS (G)	Avg DICE (%)
(single) MaxViT	224×224	7×7	82.62/14.2	79.83
(single) MaxViT	256×256	8×8	82.62/19.11	80.20
Parallel Double MaxViT	224×224, 224×224	7×7, 7×7	147.86/28.4	80.81
Parallel Double MaxViT	256×256, 256×256	8×8, 8×8	147.86/38.22	82.15
Cascaded Double MaxViT	224×224, 224×224	7×7, 7×7	147.86/28.4	81.06
Cascaded Double MaxViT	256×256, 256×256	8×8, 8×8	147.86/38.22	83.02
Parallel MERIT (**our**)	256×256, 224×224	8×8, 7×7	147.86/33.31	**82.91**
Cascaded MERIT (**our**)	256×256, 224×224	8×8, 7×7	147.86/33.31	**83.35**

Cascaded MERIT achieves the **best** DICE score (83.35%) which improves the baseline 256×256 resolution single-scale MaxViT (see 2nd row entries in the table) by 3.15%.

Effect of CASCADE Decoder and MUTATION

Architectures	CASCADE decoder	MUTATION	Avg DICE (%)
Parallel MERIT	No	No	80.44
Parallel MERIT	No	Yes	81.06
Parallel MERIT	Yes	No	82.91
Parallel MERIT (**our**)	Yes	Yes	84.22
Cascaded MERIT	No	No	80.76
Cascaded MERIT	No	Yes	82.03
Cascaded MERIT	Yes	No	83.35
Cascaded MERIT (**our**)	Yes	Yes	**84.90**

CASCADE decoder with MUTATION loss aggregation for implicit ensembling/augmentation achieves the best DICE score (84.90%).

Conclusion

- Experiments on two medical image segmentation benchmarks demonstrate the superior performance of MERIT over SOTA methods.
- Our MERIT architectures and MUTATION loss aggregation can improve other downstream medical image and semantic segmentation tasks.

References

1. J. Chen et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021.
2. H. Cao et al. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537, 2021.
3. Z. Tu, et al. Maxvit: Multi-axis vision transformer. ECCV, 2022.