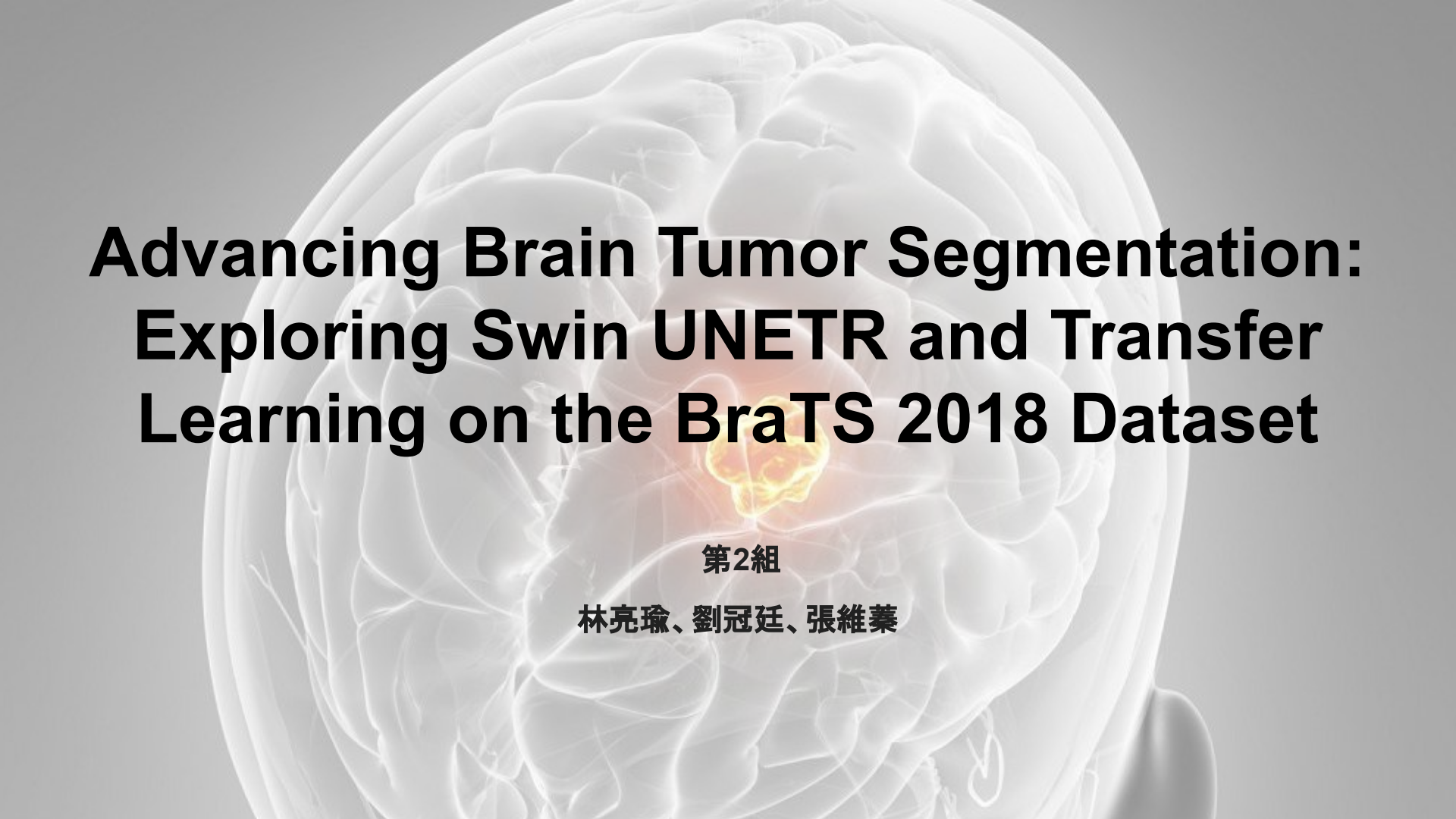


Advancing Brain Tumor Segmentation: Exploring Swin UNETR and Transfer Learning on the BraTS 2018 Dataset

第2組

林亮瑜、劉冠廷、張維蓁

Outline

- Introduction
- Motivation and Aims
- Materials and methods
  - Datasets (BraTS 2018 vs. BraTS 2021)
  - Benchmark
  - Finetuning
- Results and Discussion
- Conclusion

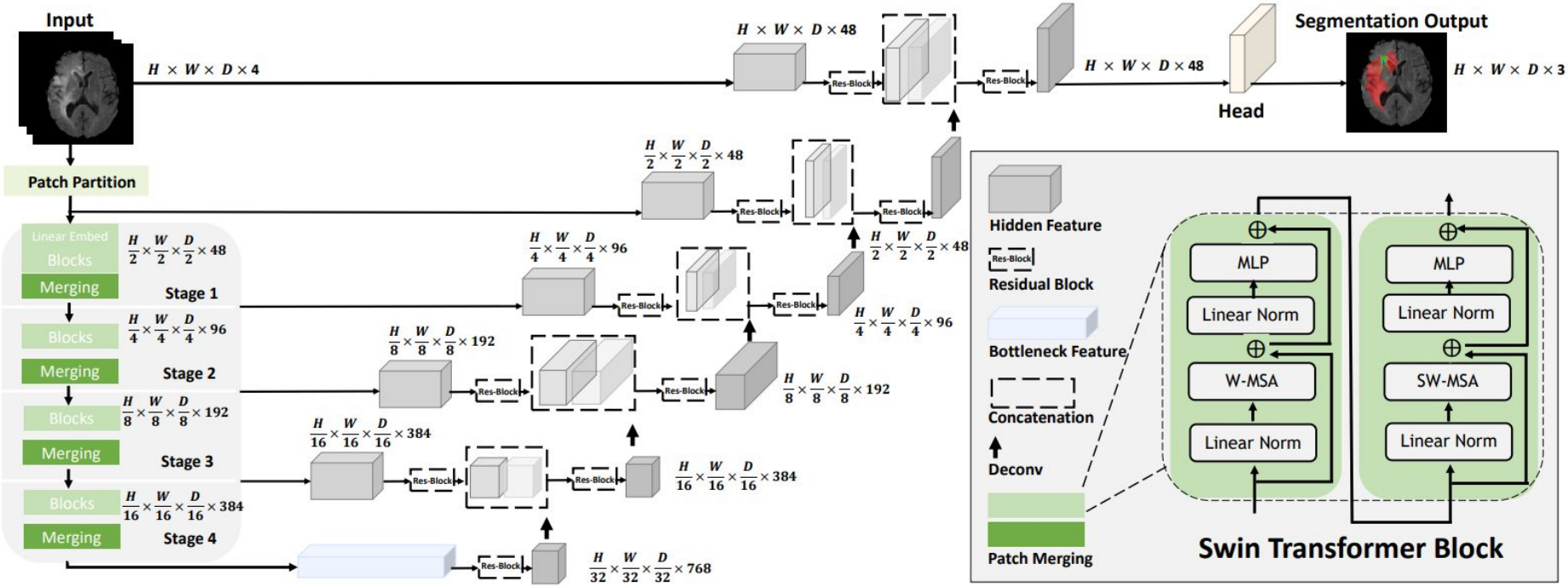
Introduction of BraTS

Brain tumor segmentation (BraTS):

Differentiate between normal and tumor tissues in brain images, increase the accuracy of tumor diagnosis.

BraTS challenge provided 3D MRI dataset and the ground truth tumor segmentation labels annotated by doctors to the challengers for model training to identify the position of the brain tumor.

BraTS 2021 Winner - Swin UNETR



Data Preprocessing

a : intensity original range
b : intensity target range

- Normalization : **zero mean** and **unit standard deviation**

```
parser.add_argument("--a_min", default=-175.0, type=float, help="a_min in ScaleIntensityRanged")
parser.add_argument("--a_max", default=250.0, type=float, help="a_max in ScaleIntensityRanged")
parser.add_argument("--b_min", default=0.0, type=float, help="b_min in ScaleIntensityRanged")
parser.add_argument("--b_max", default=1.0, type=float, help="b_max in ScaleIntensityRanged")
```

- Random axis mirror flip with a probability of 0.5 for all 3 axes

```
transforms.RandFlipd(keys=["image", "label"], prob=0.5, spatial_axis=0),
transforms.RandFlipd(keys=["image", "label"], prob=0.5, spatial_axis=1),
transforms.RandFlipd(keys=["image", "label"], prob=0.5, spatial_axis=2),
transforms.NormalizeIntensityd(keys="image", nonzero=True, channel_wise=True),
transforms.RandScaleIntensityd(keys="image", factors=0.1, prob=1.0),
transforms.RandShiftIntensityd(keys="image", offsets=0.1, prob=1.0),
transforms.ToTensord(keys=["image", "label"]),
```

Data Augmentation

```
v = v * (1 + factor)
```

- Random scale of intensity in the range (0.9, 1.1)

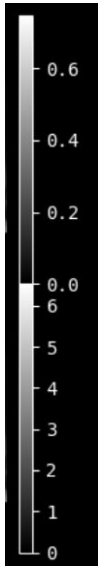
  * factors range to randomly scale.

- Random per channel intensity shift in the range (-0.1, 0.1)

  * offsets range to randomly shift.

```
transforms.RandFlipd(keys=["image", "label"], prob=0.5, spatial_axis=0),
transforms.RandFlipd(keys=["image", "label"], prob=0.5, spatial_axis=1),
transforms.RandFlipd(keys=["image", "label"], prob=0.5, spatial_axis=2),
transforms.NormalizeIntensityd(keys="image", nonzero=True, channel_wise=True),
transforms.RandScaleIntensityd(keys="image", factors=0.1, prob=1.0),
transforms.RandShiftIntensityd(keys="image", offsets=0.1, prob=1.0),
transforms.ToTensord(keys=["image", "label"]),
```

Scale



Shift

Motivation and Aims

本次報告希望能探討BraTS 2021 competition中獲勝的方法 — Swin UNETR, 是否也能適用在其他腦瘤資料集上, 並進一步優化其在其他腦瘤資料上的預測效果。

1. 觀察測試資料集(BraTS 2018 dataset)與BraTS 2021資料集的差異性
   ➜ Data comparison

2. 測試Swin UNETR pretrained model在BraTS 2018 dataset的表現
   ➜ Benchmark

3. 進行遷移學習(Transfer learning)更進一步優化pretrained model
   ➜ Finetuning

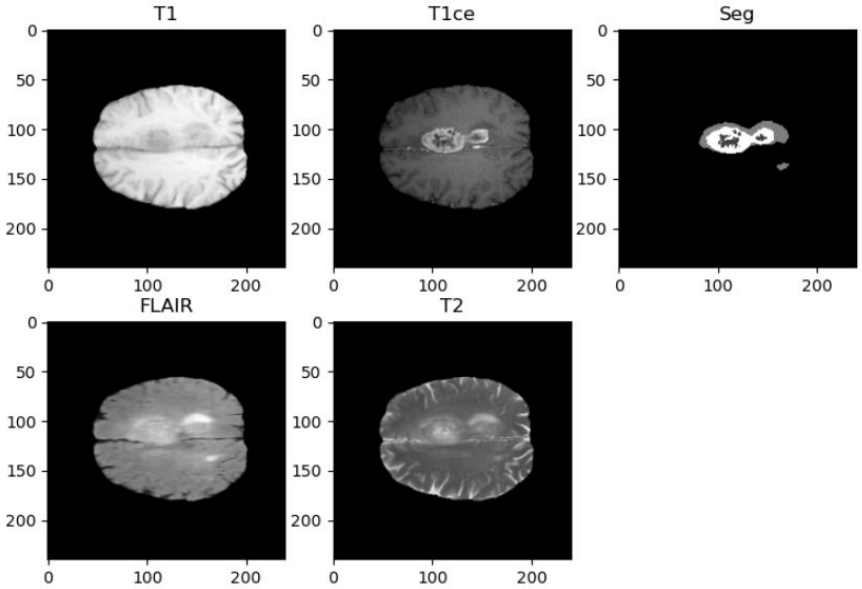
Datasets

- The input image size is 240 × 240 × 155
- Each with four 3D MRI modalities (T1, T1ce, T2 and FLAIR)
- Ground truth annotation (whole tumor (WT), tumor core (TC) and enhancing tumor (ET))

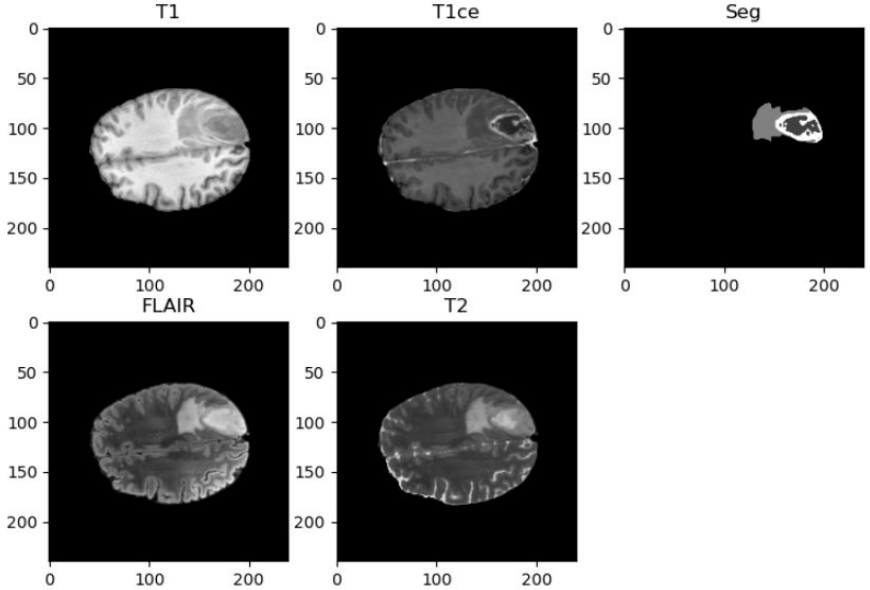
	BraTS 2018	BraTS 2021
是否有區分HGG/LGG？	O	X
Training Data 數	Total : 285 HGG(210)、LGG(75)	Total : 1251
Validation Data 數	66	219
Survival Data	O	X

Images in datasets

2018 (HGG) : Brats18_2013_2_1



2021 : BraTS2021_00495



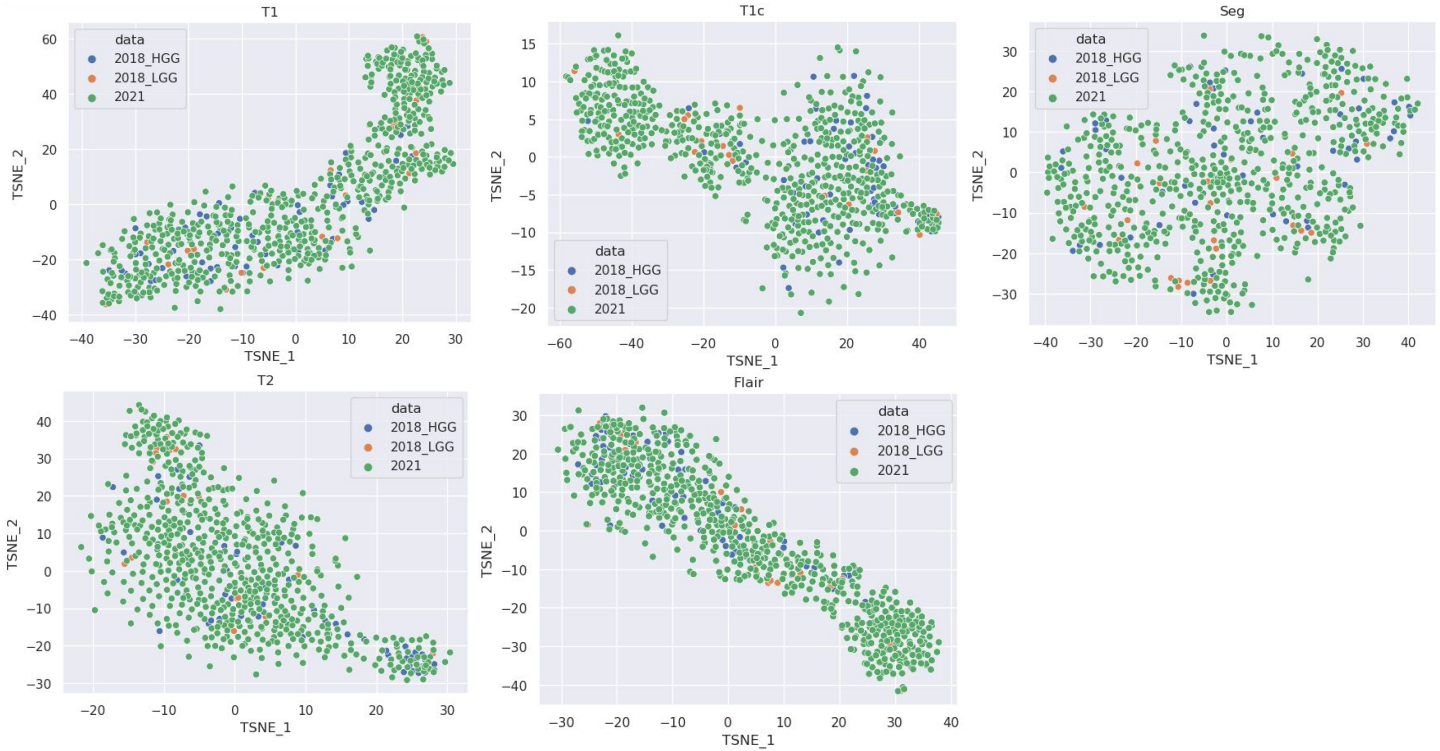
Methods of data visualization

- Randomly select 1/2 data (memory limitation) in 2018 (HGG), 2018 (LGG) and 2021 datasets from different modalities (T1, T1ce, T2 and FLAIR) and segmtation data.
- Use t-distributed stochastic neighbor embedding (t-SNE)

  reducing the dimension from 3D image to 2 dimension
- Plot!

Results of T-SNE plot of two datasets

Features in two datasets are overlapping:



Records of previous SOTA models

2018: 3D MRI

Table 2. BraTS 2018 validation dataset results. Mean Dice and Hausdorff measurements of the proposed segmentation method. EN - enhancing tumor core, WT - whole tumor, TC - tumor core.

	Dice			Hausdorff (mm)		
Validation dataset	ET	WT	TC	ET	WT	TC
Single Model	0.8145	0.9042	0.8596	3.8048	4.4834	8.2777
Single Model + TTA	0.8173	0.9068	0.8602	3.8241	4.4117	6.8413
Ensemble of 10 models	0.8233	0.9100	0.8668	3.9257	4.5160	6.8545

Table 3. BraTS 2018 testing dataset results. Mean Dice and Hausdorff measurements of the proposed segmentation method. EN - enhancing tumor core, WT - whole tumor, TC - tumor core.

	Dice			Hausdorff (mm)		
Testing dataset	ET	WT	TC	ET	WT	TC
Ensemble of 10 models	0.7664	0.8839	0.8154	3.7731	5.9044	4.8091

2021: Swin UNETR

Table 2. Five-fold cross-validation benchmarks in terms of mean Dice score values. ET, WT and TC denote Enhancing Tumor, Whole Tumor and Tumor Core respectively.

	Swin UNETR				nnU-Net				SegResNet				TransBTS			
Dice Score	ET	WT	TC	Avg.	ET	WT	TC	Avg.	ET	WT	TC	Avg.	ET	WT	TC	Avg.
Fold 1	**0.876**	**0.929**	**0.914**	**0.906**	0.866	0.921	0.902	0.896	0.867	0.924	0.907	0.899	0.856	0.910	0.897	0.883
Fold 2	**0.908**	**0.938**	**0.919**	**0.921**	0.899	0.933	**0.919**	0.917	0.900	0.933	0.915	0.916	0.885	0.919	0.903	0.902
Fold 3	**0.891**	**0.931**	**0.919**	**0.913**	0.886	0.929	0.914	0.910	0.884	0.927	0.917	0.909	0.866	0.903	0.898	0.889
Fold 4	**0.890**	**0.937**	**0.920**	**0.915**	0.886	0.927	0.914	0.909	0.888	0.921	0.916	0.908	0.868	0.910	0.901	0.893
Fold 5	**0.891**	**0.934**	**0.917**	**0.914**	0.880	0.929	**0.917**	0.909	0.878	0.930	0.912	0.906	0.867	0.915	0.893	0.892
Avg.	**0.891**	**0.933**	**0.917**	**0.913**	0.883	0.927	0.913	0.908	0.883	0.927	0.913	0.907	0.868	0.911	0.898	0.891

Table 3. BraTS 2021 validation dataset benchmarks in terms of mean Dice score and Hausdorff distance values. ET, WT and TC denote Enhancing Tumor, Whole Tumor and Tumor Core respectively.

	Dice			Hausdorff (mm)		
Validation dataset	ET	WT	TC	ET	WT	TC
Swin UNETR	0.858	0.926	0.885	6.016	5.831	3.770

Myronenko, Andriy. "3D MRI brain tumor segmentation using autoencoder regularization." Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4. Springer International Publishing, 2019.

Hatamizadeh, Ali, et al. "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images." *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*. Cham: Springer International Publishing, 2022.

Method of pre-trained model evaluation

Use pretrain model to evaluate 2018 BraTs Training dataset

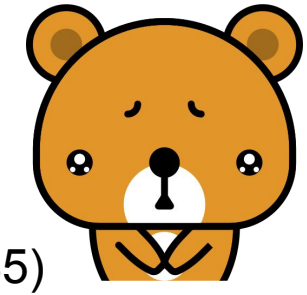
```
(swin) 109304021@c007:~/swin$ python main.py --resume_ckpt --workers 0 --val_every 1 --fold 3 --json_list=jsons/brats2018_fold.json --data_dir=2018_trai
ning/MICCAI_BraTS_2018_Data_Training/ --distributed --max_epochs 1
train: 231
validation: 54
0  gpu 0
Batch size is: 1 epochs 1
Using pretrained weights
Total parameters count 62191941
Writing Tensorboard logs to  ./runs/test
0 Mon May 29 14:24:03 2023 Epoch: 0
/home/109304021/anaconda3/envs/swin/lib/python3.7/site-packages/monai/metrics/meandice.py:78: UserWarning: y_pred should be a binarized tensor.
  warnings.warn("y_pred should be a binarized tensor.")
Val 0/1 0/54 , Dice_TC: 0.42521814 , Dice_WT: 0.9305793 , Dice_ET: 0.3455174 , time 20.14s
Val 0/1 1/54 , Dice_TC: 0.529399 , Dice_WT: 0.6273658 , Dice_ET: 0.51985264 , time 13.02s
Val 0/1 2/54 , Dice_TC: 0.35293266 , Dice_WT: 0.45141873 , Dice_ET: 0.34656844 , time 12.80s
Val 0/1 3/54 , Dice_TC: 0.2646995 , Dice_WT: 0.38517028 , Dice_ET: 0.25992632 , time 12.74s
Val 0/1 4/54 , Dice_TC: 0.3357696 , Dice_WT: 0.41443592 , Dice_ET: 0.3586741 , time 12.79s
Val 0/1 5/54 , Dice_TC: 0.31555295 , Dice_WT: 0.37798536 , Dice_ET: 0.32788244 , time 12.88s
```

Pre-trained model evaluation on BraTS 2018

Model: Swin UNETR pre-trained model on BraTS21 dataset (Fold1)

	Dice_TC	Dice_WT	Dice_ET	Accuracy
Fold-0 (65)	0.28285	0.53255	0.22598	0.34713
Fold-1 (64)	0.22941	0.32770	0.17163	0.24291
Fold-2 (47)	0.29992	0.42889	0.23711	0.32198
Fold-3 (54)	0.29015	0.47305	0.24850	0.33724
Fold-4 (55)	0.29225	0.40019	0.21032	0.30092
Average	0.27892	0.43248	0.21871	0.31004
Fold-1 (2021 best)	0.90075	0.92517	0.88792	0.90461

Method of finetuning on 2018 BraTS dataset



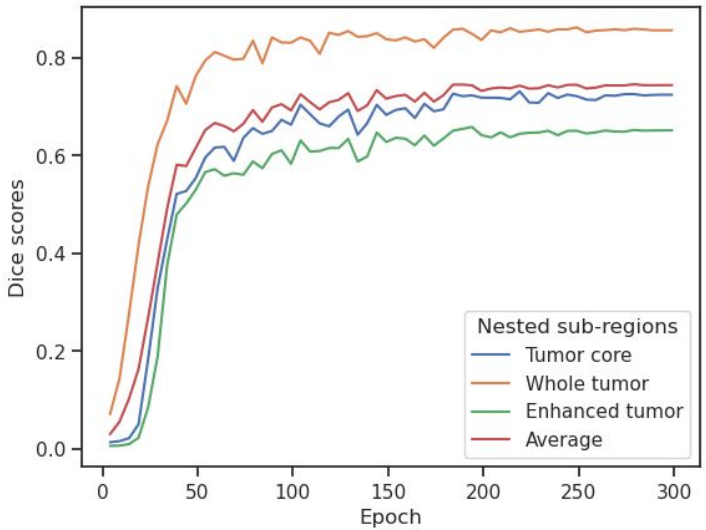
- Randomly allocate data into the range of 0 to 4 folds
- Training data (fold0, 2-4): 221, validation data(fold1): 64 (Total: 285)
- Pretrain model: fold-1 model (Mean Dice: 90.59)

```
(swin) 109304021@c007:~/swin$ python main.py --json_list=jsons/brats2018_fold.json --val_every=5 --data_dir=2018_training/MICCAI_BraTS_2018_Data_Training/ --noamp --pretrained_model_name=model.pt --pretrained_dir=pretrained_models/fold1_f48_ep300_4gpu_dice0_9059 --fold=1 --roi_x=128 --roi_y=128 --roi_z=128  --in_channels=4 --spatial_dims=3 --use_checkpoint --feature_size=48 --sw_batch_size=4 --workers=1 --distributed
Found total gpus 1
train: 221
validation: 64
0  gpu 0
Batch size is: 1 epochs 300
Total parameters count 62191941
Writing Tensorboard logs to  ./runs/test
0 Mon May 22 10:46:01 2023 Epoch: 0
Epoch 0/300 0/221 loss: 0.9077 time 27.78s
Epoch 0/300 1/221 loss: 0.9526 time 3.19s
Epoch 0/300 2/221 loss: 0.9524 time 3.17s
```

- resize:128x128x128
- batch size: 1
- epochs: 300

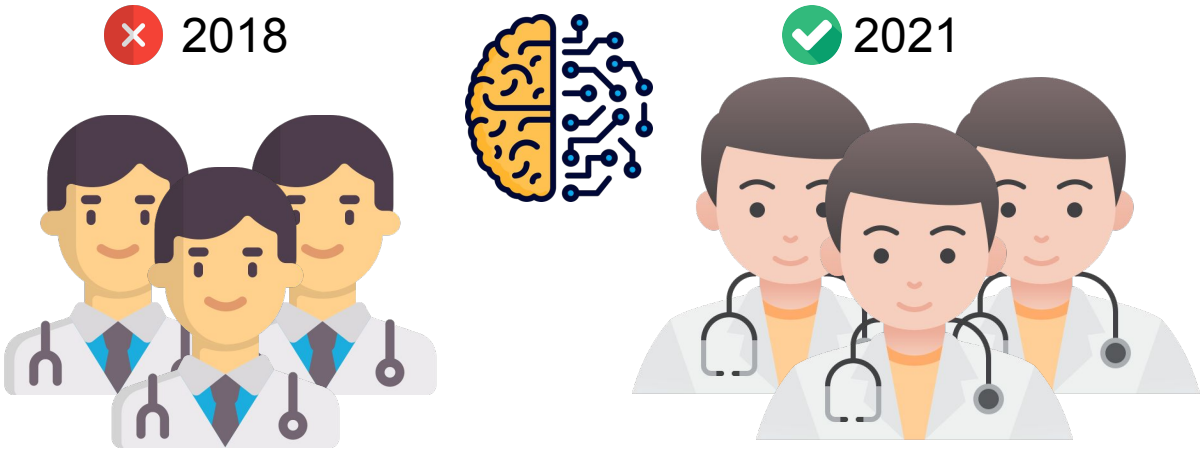
Finetuning results

- Best Accuracy: 0.746
- Best validation stats at epoch 280/300
- Dice_TC: 0.726 , Dice_WT: 0.860 , Dice_ET: 0.653



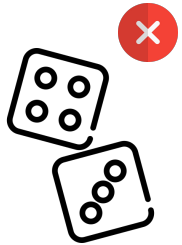
Discussion

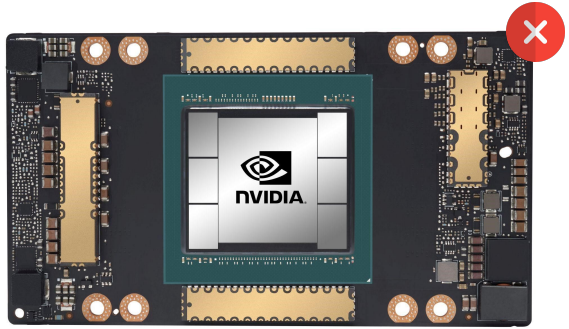
- 觀察到pre-trained model在2018資料上表現不如預期，沒有外推性?
  a. 不同醫生的判斷不盡相同，對兩個datasets的ground truth可能有不同的標註標準
     ➜ 使得模型學習到2021的醫師判斷，但無法學習到2018醫師判斷。
  b. 可能對2021 dataset overfitting?



Limitations

- Only one observation (without robustness)
- Finetuning hyperparameters didn’t optimize
- Randomly data labeling is uncommon
- GPU out of memory (School’s server)
- We don't have enough time… (1-fold: 5 days)
- Unable to apply 5-fold validation





Conclusion

1. Qualitative analysis and T-SNE dimension reduction yield similar features across 2018 and 2021 datasets.
2. Pre-trained model cannot accurately predict the 2018 dataset.
3. Finetuning of pre-trained model can enhance the accuracy from 0.310 to 0.746 (SOTA of 2018: 0.859)
4. Varying interpretations among doctors and differing annotation standards for the two datasets hinder the model's ability to learn from the 2018 physician judgments.
5. Overfitting to the 2021 dataset could also be a contributing factor.