Final Report

Project 19: Glioma Classification and Biopsy Guidance with Multimodal MRI and Deep Learning

Project member: Nhat Le

1. Introduction

Glioma is a common type of brain tumor with low survival rate with less than 10 percent of patients with malignant glioma survive two years. The clinical standard for glioma diagnosis relies on magnetic resonance (MR) images and subsequent biopsies, but conventional MR are not sufficiently tissue-specific enough to guide treatment decisions. Amide proton transfer-weighted (APTw), a special MR sequence developed in JHU Radiology Dept., has shown clinical values in glioma-related tasks such as glioma grading and assessment of tumor recurrence. However, interpreting APTw images for such tasks requires radiologists to have additional expertise. In recent years, deep learning has shown excellent performance in image-based diagnosis, and therefore we want to build a fully automated pipeline that could provide accurate and explainable decisions for these tasks with MR images including APTw.

2. Project goals

Based on previous works, we want to extend the use of deep learning algorithms in classification tasks: high grade vs. Low-grade for newly diagnosed patients and tumor recurrence vs. Treatment effect for post-treatment patients with malignant gliomas. We evaluate a few baseline algorithms which are commonly used in dealing with sequence of images and aim to come up with changes that could make them work better with our dataset. Furthermore, we analyze the use of multiple APT images in the automated pipeline in order to provide helpful information that could be used to optimize the image acquisition process.

3. Summary of Dataset

a. Overview

The dataset used in this project is a private dataset curated by Dr. Jiang's APT group. It consists of 216 MRI scans in total for both newly-diagnosed and post-treatment patients. The scans are stored in NIfTI file format with each scan represented by 15 image slices of 256x256 pixels. Further technical descriptions of the scans and details for the preprocessing pipeline can be found in [1], a recent study that used a subset of this dataset. Several cases deemed not suitable for this study by Dr. Jiang for clinical reasons are removed which left us with 202 cases.

Type	# scans
Post-treatment tumor recurrence	103
Post-treatment treatment effects	61
Newly-diagnosed low-grade	18
Newly-diagnosed high-grade	20
Total	202

b. Annotations

There are 3 level of annotations for this dataset: scan-level, slice-level, and pixel-level. Scan-level annotation is demonstrated by the table above. Slice-level annotation indicates whether each slice has abnormal regions, tumors, and recurrent tumors. Pixel-level annotation indicates regions of clinical significance for glioma patients, including 4 level of labels corresponding to slice-level annotation. We obtain scan-level annotation for the whole dataset, slice-level annotation for 147 post-treatment scans, and pixel-level annotation for around 140 post-treatment scans at the time of this project.

With the annotations, we perform a quick analysis to understand the distribution of our data and later come up with suitable training and evaluation techniques for our deep learning models. A summary of slice-level annotations is presented in the table below:

Type	Train	%	Val	%	Total	%		Notes
Scan-level	118	100	29	100	147	100		
Scan-TE	47	39.83	11	37.93	58	39.45		Treatment effect
Scan-TR	71	60.17	18	62.07	89	60.55		Tumor recurrence
Slice-level	1770	100	435	100	2205	100		
Label 0	333	18.81	83	19.08	416	18.87		
Label 1	622	35.14	143	32.87	765	34.69		
Label 2	494	27.91	106	24.37	600	27.21		
Label 3	321	18.14	103	23.68	424	19.23		
Normal	333	18.81	83	19.08	416	18.87		Label 0 presence
Abnormal	1437	81.19	352	80.92	1789	81.13		Label 1 or 2 or 3 presence
Tumor	943	53.26	246	56.55	1189	53.92		Label 1 or 3 presence
Recurrent	622	35.14	143	32.87	765	34.69		Label 1 presence

c. Averaged APT images

A main difference between this dataset and the dataset described in [1] is the inclusion of multiple APT images for each scan. For each scan, APT signals are obtained multiple times under identical conditions, but the images acquired from them can be different due to the underlying mechanism of APT sequence, which is beyond the scope of this project. While earlier work [1] shows incremental improvement in deep learning pipeline with APT images from single acquisition, we propose that using the average of multiple APT images can further improve the classification performance as it reduces noise and other irregularities. Therefore, we create 6 subsets of data from 147 cases with slice-level annotations for the purpose.

Firstly, we create 3 versions of the 147 cases dataset above with mixture of averaged APT images, namely 147-APT-1, 147-APT-2, 147-APT-3. Each scan in each dataset contains 4 common structural MR images T1w, T2w, T1-Gdw, FLAIR, and an APTw image. The difference is the number of APT images used to calculate the APTw image for each scan:

- 147-APT-1: the APTw image comes from a single APTw image (apt-1)
- 147-APT-2: the APTw image comes from the average of 2 APTw images (apt-2) if available, else use a single APTw image
- 147-APT-4: the APTw image comes from the average of 4 APTw images (apt-4) whenever available, if not then from 2 images, if not then from a single image

The following tables show the distribution of APT images in these datasets:

147-APT-1	Train	%	Val	%
apt-1	118	100	29	100
Total	118	100	29	100

147-APT-2	Train	%	Val	%
apt-1	2	1.69	4	13.79
apt-2	116	98.31	25	86.21
Total	118	100	29	100

147-APT-4	Train	%	Val	%
apt-1	2	1.69	4	13.79
apt-2	18	15.25	4	13.79
apt-4	98	83.06	21	72.42
Total	118	100	29	100

(Figure: distribution of APT images in 147-case datasets)

Secondly, we create another 3 datasets named 119-APT-1, 119-APT-2, and 119-APT-4. These datasets come from a subset of 119 cases from 147 cases above, but these 119 cases all have at least 4 APTw images such that there is no mixture of average APT images in each dataset.

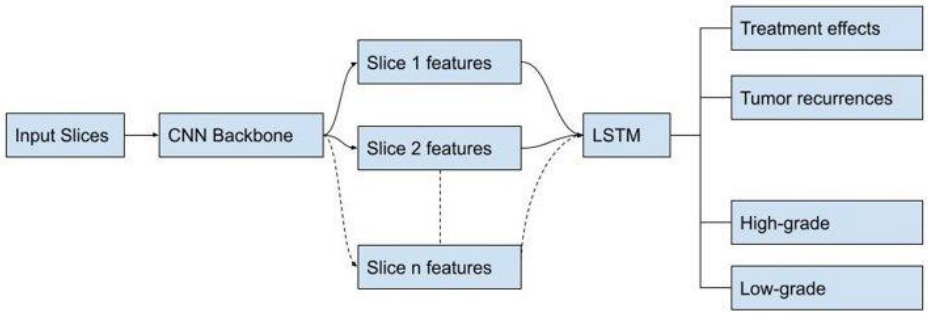
The APT images for each case are co-registered to each other using built-in registration functionality of ITK-Snap before the average is taken. Then, structural MR images will be co-registered and resampled to the space of the average APT images.

4. Technical approach

We treat each scan as a sequence of slices and benchmark several deep learning algorithms for video classification. The main task used for evaluation of baseline algorithms and development of our improved method is detecting tumor recurrence at slice-level and scan-level with the datasets of patients with post-treatment malignant glioma. The dataset used for slice-level and scan-level classification is 147-APT-1.

a. Baseline evaluation

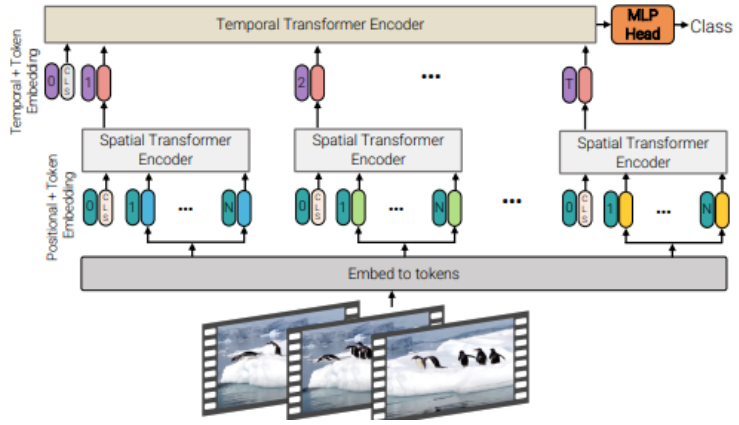
Algorithm 1: CNN-LSTM



This algorithm involves a combined architecture of a convolutional neural network (CNN – ResNet 18) and long-short term memory (LSTM). The CNN backbone acts as feature extractor on each multimodal slice with each sequence being used as an input channel. Then, the LSTM module aggregates all features of 15 slices for each scan and perform the classification at scan-level.

Implementation & Training Note: Rather than training the combined architecture end-to-end, we train a ResNet-18 to perform slice-level classification and then freeze the weights and use it as feature extractor to train the LSTM module.

Algorithm 2: Video Vision Transformer (Vivit) [3]



(Figure: Factorised encoder – Model 2 ViViT from [3])

This is a pure-transformer based architecture for video classification. We use Model 2 (Factorised encoder) from the original paper [3] as it shares similarities with algorithm 1, except that both CNN and LSTM modules are replaced with transformer modules. As illustrated, the first one, spatial transformer encoder, models interactions between tokens extracted from the same slice, and the second one, temporal transformer encoder, model interactions between representations all slices in a scan. The output of the temporal transformer encoder is then used in a multilayer perceptron to classify the instance.

b. Analysis of using averaged APT images

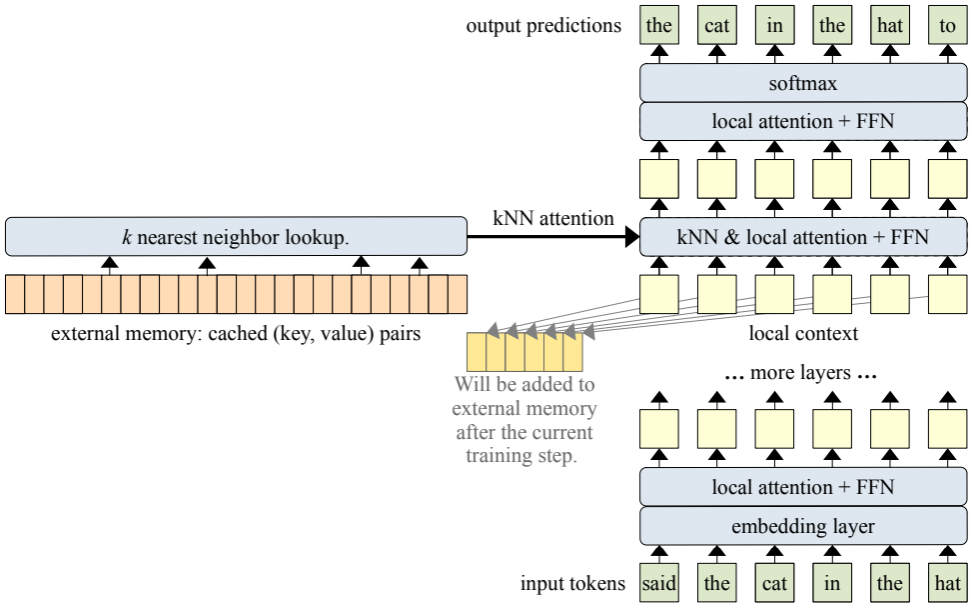
We train and evaluate the performance of ResNet-18 on the slice-level classification task with 6 different datasets described in section 3.c. We choose this over the scan-level classification because it provides more instances for training and evaluation as well as the practicality of generating class activation maps for each slice.

c. Our proposed method

We also focus on the slice-level classification while developing our proposed method for the same reasons as above. In recent years, transformers and attention mechanism, a deep learning architecture originally used in natural language processing (NLP), have gained popularity in other deep learning domains. Transformer-based models in computer vision [1] start taking over CNN and providing state-of-the-art performance in different vision tasks. In general, the architecture differences of vision transformers allow them to model long-range dependence and give attention to pixel-level details better than CNN, which we believe is important when dealing with image markers of APTw. However,

many of these models only achieve good results if they are trained on large dataset which is not what we have for our problem.

In this project, we explore the idea of augmenting transformer with 'memories' and the use of cross-attention mechanism to tackle this problem. This approach shares some similarities with few-shot learning, in which we 'compare' the instance we want to classify with other instances of known classes to leverage external information to make the decision, instead of using only local context of the instance. In a recent work [9], a transformer model for NLP is introduced with a non-differentiable memory slot, and we want to extend this model for our project. The memory slot contains the (key, value) pairs of tokens in previous training steps. Besides the common self-attention mechanism, it employs kNN-augmented attention, which computes similarity scores between the query token and tokens retrieved from a K nearest neighbor lookup into the memory with the query token, as shown in the figure below from the original paper [9].



(Figure: kNN-augmented transformer architecture from [9])

To make more suitable for the problem, we implement a few changes on the original model and the workflow. We remove the causal attention mask in the original model because it is irrelevant to our task. Then, we change the input and output head of the model. The output head of the model is replaced by a multilayer perceptron with binary classification head. As we want to use the same data loader across different methods, we modify the embedding layer to accept 2D images of N channels and perform patch and position embedding as usually seen in other vision transformer architecture. Lastly, for the training loop, we decide to only add to the memory the tokens of slices with recurrent tumors, which is believed to contain key features of the APTw images for this task, and we follow the rest of the add/drop scheme for the memory in the original paper. We keep the design choice of 6 layers of transformer decoders with the 5^{th} layer has kNN-attention for our main evaluation. The memory has size of 8192 tokens, which can contains information from around 40 slices, each slice represented by 196 tokens from 16x16 pixel image patches. Mathematical formulation of kNN-augmented transformer module can be found in the original paper.

We compare our performance of our method with a CNN method ResNet-18, similar to what is used in baseline evaluation and analysis of using average APT. We also plan to perform further experiments to optimize the choice of which transformer layers will use the kNN-attention other than second-to-last by default and explore potential improvements with spatial pyramid pooling layer with our architecture.

d. Evaluation methods

The two metrics used for evaluating different models and dataset is overall classification accuracy and corresponding AUC, or area under the receiver operating curve, which are commonly used in similar problems.

5. Results and Discussion

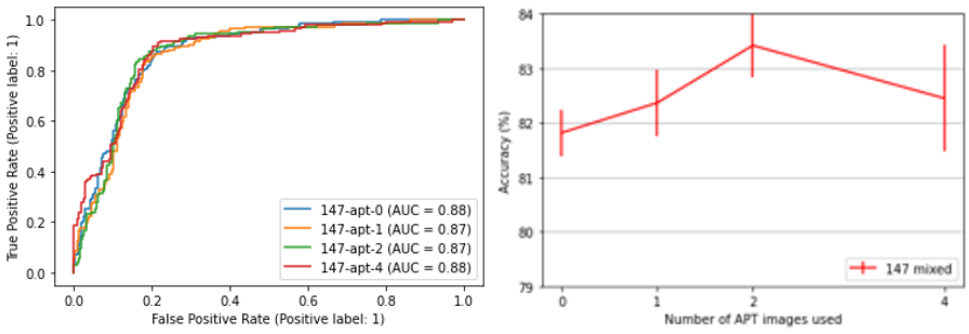
a. Baseline evaluation

Baseline evaluation	Accuracy (%)	+/-
CNN-LSTM	78.9	1.7
CNN-LSTM **	63.3	1.3
ViViT	76.2	1.6

(Figure: Scan-level classification results of baseline algorithms)

The two baseline algorithms provide comparable results for scan-level classification task with the CNN-LSTM provides better results. However, CNN-LSTM benefits from training the CNN module on slice-level classification which improves gradient flows to CNN and stabilize the training for LSTM, while ViViT is training end-to-end at scan level only. A quick experiment (CNN-LSTM **) shows that the combined architecture could only achieve around 63% classification accuracy if CNN and LSTM modules are trained from scratch concurrently. This result shows that pure-transformer based method ViViT is better at modeling interactions within a slice and between slices when training at scan-level only.

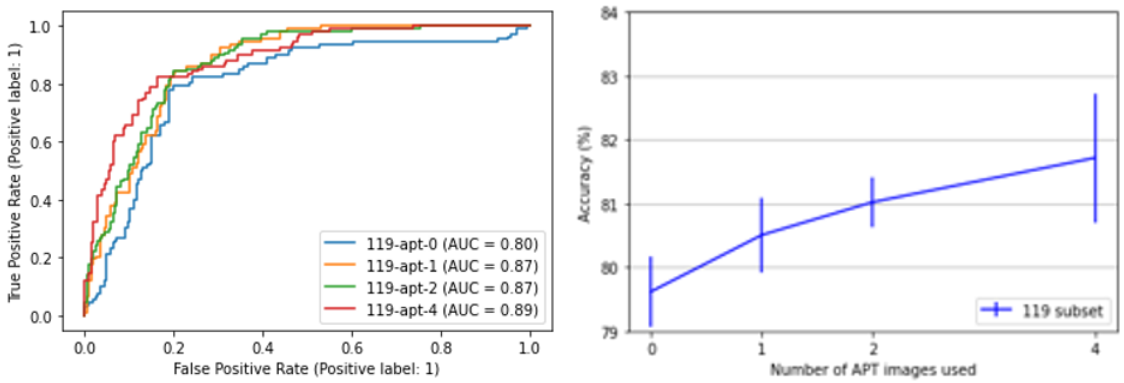
b. Analysis of using averaged APT images



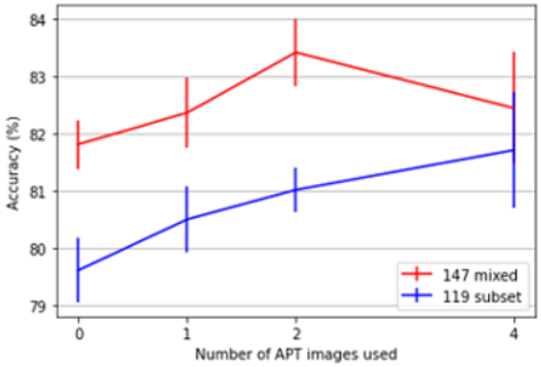
(Figure: ROC curve and slice-level classification performance with 147-APT datasets)

The results on different versions of 147-case datasets shows that using the average of 2 APTw images yields better slice-level classification performance than using no APTw images (t-test, p=0.90). However, performance degrades when a more diverse mixture of APTw images in 147-APT-4 is used as it leads to

inhomogeneous data in training and potential distribution shift between the train and validation set. The following results with homogenous 119-case datasets supports our prediction:



(Figure: ROC curve and slice-level classification performance with 119-APT datasets)



(Figure: slice-level classification performance with 147-APT and 119-APT datasets)

The slice-level classification performance on 119-case datasets is lower than that on 147-case datasets which is expected with smaller dataset size. Similarly on this dataset, using 2 APT images yields better slice-level classification performance than using no APTw images (t-test, p=0.90) and there is a clearer association between using more APTw images and higher classification accuracy. We would need more data and slice-level annotations to confirm this result.

c. Our method

Slice-level classification	Accuracy (%)	+/-
CNN (ResNet-18)	82.3	0.6
kNN-augmented transformer (ours)	79.6	0.8
kNN-augmented transformer * (ours)	81.4	0.5

(Table: Slice-level classification results for baseline and evaluation method. * denotes the training scheme in which the tokens added to memory is filtered as described in 4c)

While the performance of our model is not better than CNN method, it is promising as earlier attempts using vision transformers (ViT-Base) on slice-level classification only gave around 70% accuracy. We also show that adding only the tokens from slices with recurrent tumors to the memory improves the

classification result by nearly 2%. This could be because these slices only account for 34.7% of our dataset (see 3b) and by doing so could help the network 'compare' the query slice with the memory more efficiently as only the chosen tokens might include unique APT information for recurrent tumor regions. Further analysis and ablation study will be performed to confirm our speculation.

Lastly, preliminary results on the experiment to optimize which layer to have kNN-attention show that there is no significant performance difference between the different positions of layer with kNN-attention while having multiple layers with memories and kNN-attention will increase the computational cost. More comparison will be made once we can visualize the attention from our method.

6. Management Summary

a. Personal takeaway

I obviously learned a lot from reading related papers for this project and working on the design and implementation of the proposed architecture. I think I could have done better in keeping track of subtasks of the project with a productivity tool separate from or in parallel to my general personal calendar.

b. Plan vs. Accomplishment

In the proposal, we planned to tackle the high-grade/low-grade classification task for newly-diagnosed patients but was later instructed to drop it in this project due to the size of the dataset and clinical relevance. For our proposed architecture, we have not yet been able to produce attention maps for visualization and qualitative comparison to other methods. This is important for interpretability purposes.

c. Next Steps

Additional experiments with our proposed network architecture will be performed in the summer to validate and extend the usage for scan-level classification. We will also train our method on larger public dataset to test for generalizability. With our results on analysis of using averaged APT images, we will take the work of [8] into consideration to introduce meaningful data augmentation techniques to our training schemes. For the data of newly-diagnosed patients, we will merge with the data of post-treatment patients to train on the task of detecting active tumors.

d. Management

I communicate with my mentors in their weekly lab/group meetings and by emails. We also met a few times in one-on-one meetings for resolving dependencies and after milestones. The link for code repository is on the project wiki page.

7. References

[1] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2021). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*.

[2] Zhou, J., Heo, H. Y., Knutsson, L., van Zijl, P. C., & Jiang, S. (2019). APT-weighted MRI: Techniques, current neuro applications, and challenging issues. *Journal of Magnetic Resonance Imaging*, *50*(2), 347-364.

[3] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6836-6846).

[4] Lee, J., Wang, N., Turk, S., Mohammed, S., Lobo, R., Kim, J., ... & Rao, A. (2020). Discriminating pseudoprogression and true progression in diffuse infiltrating glioma using multi-parametric MRI data through deep learning. *Scientific reports*, *10*(1), 1-10.

[5] Liu, L., Hamilton, W., Long, G., Jiang, J., & Larochelle, H. (2020). A universal representation transformer layer for few-shot image classification. *arXiv preprint arXiv:2006.11702*.

[6] Hou, R., Chang, H., Ma, B., Shan, S., & Chen, X. (2019). Cross attention network for few-shot classification. *Advances in Neural Information Processing Systems*, *32*.

[7] Guo, Pengfei and Unberath, Mathias and Heo, Hye-Young and Eberhardt, Charles and Lim, Michael and Blakeley, Jaishri and Jiang, Shanshan, Learning-Based Analysis of Amide Proton Transfer-Weighted MRI to Identify Tumor Progression in Patients with Post-Treatment Malignant Gliomas. http://dx.doi.org/10.2139/ssrn.4049653

[8] Guo, P., Wang, P., Zhou, J., Patel, V.M., Jiang, S. (2020). Lesion Mask-Based Simultaneous Synthesis of Anatomic and Molecular MR Images Using a GAN. In: , *et al.* Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science(), vol 12262. Springer, Cham. https://doi.org/10.1007/978-3-030-59713-9_11

[9] Wu, Y., Rabe, M. N., Hutchins, D., & Szegedy, C. (2022). Memorizing transformers. *arXiv preprint arXiv:2203.08913*.