

Leveraging Human Selective Attention for Medical Image Analysis with Limited Training Data: Supplementary Material

Yifei Huang¹
hyf@iis.u-tokyo.ac.jp

Xiaoxiao Li²
xiaoxiao.li@aya.yale.edu

Lijin Yang¹
yang-lj@iis.u-tokyo.ac.jp

Lin Gu³
lin.gu@riken.jp

Yingying Zhu⁴
yingying.zhu@uta.edu

Hirofumi Seo¹
hseotky@gmail.com

Qiuming Meng¹
mengqiuming1993@gmail.com

Tatsuya Harada³
harada@mi.t.u-tokyo.ac.jp

Yoichi Sato¹
ysato@iis.u-tokyo.ac.jp

¹ The University of Tokyo
Japan

² The University of British Columbia
Canada

³ RIKEN, AIP
Japan

⁴ University of Texas Arlington
USA

1 Additional experiments on the segmentation task

For the segmentation task, Table 1 shows the experiment results on the validation set evaluated by the online evaluation server. The result on the validation set is very similar to that on the training set.

Mean (std)	Enhancing Tumor				Whole Tumor				Tumor Core			
Data ratio	0.2	0.3	0.5	0.7	0.2	0.3	0.5	0.7	0.2	0.3	0.5	0.7
Backbone	71.22	75.85	70.15	76.48	85.52	88.07	88.42	88.71	76.23	78.56	82.63	81.29
+gaze*	74.37	78.10	77.85	77.06	84.14	87.09	87.71	86.22	79.23	82.19	81.59	82.30
ours	75.48	76.59	78.27	79.73	85.68	88.82	89.13	89.43	77.68	83.07	81.71	83.32
ours+surv	73.20	76.91	79.72	80.07	86.12	88.52	89.36	89.64	80.07	83.39	82.92	83.87

Table 1: Results on validation set using part of training data (Dice score %)

Table 2 shows the model performances on the validation set using full training set of

BraTS 2020 dataset. In this experiment, we first initialize the training using only the training data with gaze for 20 epochs, and then use the full training set for training. We do not apply loss on SAN when it comes to the data sample without collected gaze. We can see similar performance of backbone and **+gaze***, and also performance improvement is gained by using the expert's gaze information during training.

Dice score (%)	Enhancing Tumor	Whole Tumor	Tumor Core
Backbone	79.20	88.50	84.24
+gaze*	79.88	89.71	84.11
ours	79.67	90.16	85.03
ours+surv	80.60	89.04	84.29

Table 2: Results on validation set using full training data.

2 Model parameters and computational complexity

In Table 3, we show the number of parameters and computational complexity of our framework under the 3D tumor segmentation task. Note here the difference between the model parameters and computational complexity of **ours** and **ours+surv** frameworks are negligible, so we combine them and show in Table 3 as "ours". From the table we can find that when compared with the UNet alone, our proposed SAN and Auxiliary Attention Block only add very limited amount of model parameters and computational complexity.

	Number of parameters	Computational cost
Backbone	19.17 M	777.24 GMac
Backbone + Ours	21.45 M	779.36 GMac

Table 3: Number of parameters and computational cost of each model

3 Gaze data collection process

We hire two medical experts (one with 4 years and the other with 2 years professional experience) and two non-experts for gaze data collection. The device we use for measuring gaze position is the EyeTribe eye-tracker. We first calibrate the eye-tracker and ensure the calibration quality is above "moderate" level (angular error $\leq 1^\circ$). We use the log generated from the SDK provided by the EyeTribe for getting the gaze tracking data. The user can pause the screening at any time, however after resuming the calibration is done again to preserve the gaze measurement accuracy.

The helper GUI for gaze collection is placed in a fixed position of the screen, and the timestamp when the human begins to see one frame and ends with this frame is automatically recorded by the GUI. The GUI also records which image the person is looking at, thus by processing the gaze tracking log and the GUI log, we can locate the gaze positions on every image. We take only the gaze positions inside the images and ignore all other gaze points. As for preprocessing, for all the gaze data of one sample recorded by the same person in the same session, we first classify each data point into fixation and saccade using the code from [14]. All the data labeled as "saccade" is discarded.

Name	Description	Example value
“Person”	the patient ID	100
“Mode”	the modality	1 (t1ce)
“File”	the slice	10
“Time”	timestamp when the human begins to see this image	151815.169

Table 4: Field names, description and example value in the GUI generated log file.

Mean (std)	Atelectasis		Cardiomegaly		Pleural Effusion	
Data ratio	0.7	0.5	0.7	0.5	0.7	0.5
EfficientNet-b0	61.06	52.30	57.72	54.47	61.55	50.94
MobileNet-v2	60.37	55.41	54.32	45.39	60.84	59.99
ResNet-18	54.41	44.41	52.07	47.49	58.92	59.06

Table 5: Experiment with different backbones of the classification task.

4 Experiments with different backbones of classification task

Table 5 shows the quantitative results of the performance of different backbones (EfficientNet [10], MobileNet-v2 [11], ResNet-18 [12]) in our framework. Among the three backbones, EfficientNet and MobileNet-v2 perform better than ResNet-18, which is reasonable since the amount of training data is limited, and ResNet is more complicated and requires more data for training when compared with the other two backbones.

5 Details for experiment of chest X-ray classification

We train and validate the performance of our proposed framework on the MIMIC-CXR-gaze dataset [13]. Among all the 1083 images with gaze, we use 200 images for validation and other images form the full training set. For the experiments using part of training data, the unused part of training data is added into the validation set. For training the model, we use the Adam optimizer with learning rate 0.001 and train for 100 epochs. We take the average AUROC score of the best 30 models in terms of validation loss.

References

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [2] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [3] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *European Conference on Computer Vision*, pages 754–769, 2018.

- [4] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A Krupinski, et al. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific data*, 8(1):1–18, 2021.
- [5] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019.