An Effective Transformer-based Solution for RSNA Intracranial Hemorrhage Detection Competition

Fangxin Shang[1], Siqi Wang[2], Xiaorong Wang[1], Yehui Yang[2]

[1]Baidu Inc., China
[2]ByteDance Inc., China
shangfangxin@baidu.com, yangyehuisw@126.com

Abstract

We propose an effective method for the detection of intracranial hemorrhages (IHD) that exceeds the performance of the winning solution in the RSNA-IHD competition (2019)(Anouk Stein et al. 2019). Meanwhile, our model only takes quarter parameters and ten percent FLOPs compared to the winner's solution. The IHD task must predict the hemorrhage category of each slice for the input brain CT. We review the top five solutions for the IHD competition held by the Radiological Society of North America(RSNA) in 2019. Almost all the top solutions rely on 2D convolutional networks and sequential models (Bidirectional GRU or LSTM) to extract intraslice and interslice features, respectively. All the top solutions improve performance by using the ensemble of models, and the number of models varies from 7 to 31. In the past years, since much progress has been made in the computer vision regime especially Transformer-based models, we introduce the Transformer-based techniques to extract the features in both intra-slice and inter-slice views for IHD tasks. Additionally, a semi-supervised method is embedded into our workflow to further improve the performance. The code is already available online.

Code — https://github.com/PaddlePaddle/Research/tree/master/CV

Datasets — https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection

Background

Intracranial hemorrhage, bleeding that occurs inside the cranium, is a serious health problem that requires rapid and often intensive medical treatment.(Anouk Stein et al. 2019) Intracranial hemorrhages account for approximately 10% of strokes in the United States, where stroke is the fifth leading cause of death. Identifying the location and type of any hemorrhage present is a critical step in the patient's treatment.

In 2019, a competition was held by *Radiological Society of North America*(RSNA), which encourages to develop automatic algorithm for intracranial hemorrhage detection (IHD). The automatic multi-label classification algorithms were expected to determine if there is intracranial hemorrhage in each *2D slice* of the input CT scan and output a probability vector with six elements. Specifically, according to bleeding location, there are 5-categories of the hemorrhage need to detect, epidural(EDH), intraparenchymal(IPH), intraventricular(IVH), subarachnoid(SAH), subdural(SDH) with an overall indicator to indicate there is *any* hemorrhage in the slice. The performance of the solution is evaluated by the weighted multi-label logarithmic loss (log-loss).

Most solutions proposed by the top-ranked teams were constructed with the multi-stage paradigm (Wang et al. 2021; Darragh Hanley 2019; TAKUOKO 2019; MIRANDA 2019; AUTOMATA 2019): Firstly, the intra-slice features within each slice are extracted by a 2D convolutional model(2D CNN). Secondly, a sequential model (Bidirectional GRU or LSTM) is applied to extract the inter-slice features based on the intra-slice features. Finally, the features are converted into probabilities via linear models. Since the model ensemble achieve lower empirical risk than each individual model, ensemble techniques are widely used in the the competition. As listed in Table 1, the top-5 solutions contain multiple 2D CNNs and sequential models with different settings, which expand the diversity of models in the ensemble. There is no sequential model in the inter-slice feature extraction of the third and fifth solutions. They introduce the strategy named "user stacking" or model the inter-slice features by ensemble learning methods like LightGBM (Ke et al. 2017).

However, there are mainly two drawbacks of the above solutions:

1. The intra-slice model and inter-slice model are decoupled because of the extremely large memory requirements for running all models simultaneously. The gradient from the second stage is truncated and can not backward to the first stage models. Therefore, the errors of the first stage models are accumulated in the second stage, which limits the final performance.
2. The solutions based on the ensemble of a large number of models require an amount of time and computational resources consuming. For example, the first place solution of the RSNA-IHD competition integrates 15 intra-slice extractors with SE-ResNext-101(Hu, Shen, and Sun 2018), DenseNet-121(Huang et al. 2017), and DenseNet-169(Huang et al. 2017) in a 5-fold cross-validation manner, which consists of roughly 330M parameters and

Rank	Solutions Log Loss on Private LB	Amount of the models Intra-slice model	Inter-slice model	Total
ours	0.04355	3		3
	0.04368	1		1
1	0.04383	15	5	20
2	0.04484	5	2	7
3	0.04510	18	1^*	19
4	0.04529	15	15	30
5	0.04561	30	1^*	31

Table 1: The solutions of the top-5 teams.

180G FLOPs in total for each slice. As a contrast, our intra-slice featrues extractor only requires 86M parameters and 15G FLOPs.

Based on the latest achievements in the field of machine learning, we proposed a novel IHD algorithm to extract intra-slice and inter-slice features in an end-to-end manner. The semi-supervised method is also introduced to reduce the number of ensemble models.

1. The recently proposed Swin-Transformer (Liu et al. 2021b) is applied to extract intra-slice features and a two-layer sequential transformer (Vaswani et al. 2017) extracts inter-slice features. Benefiting from the high memory efficiency of the former-based architectures, the feature extraction of the intra-slice and inter-slice is fully end-to-end in our solution, which could be held on a single modern GPU. This not only allows the gradients backward to the input layer directly but also requires only single forward propagation during model inference.
2. Our solution also benefits from model ensemble techniques. However, our solution does not rely on a large number of model ensembles in order to meet the needs of clinical applications. The semi-supervised learning method FixMatch (Sohn et al. 2020) is applied to enhance the model performance and reduce the number of ensemble models. Finally, our solution achieves 0.04368 with a single model and 0.04345 with the three-model ensemble, which is better performance (lower log-loss) compared to the first place solution (log-loss 0.04383) in the competition.

Method

End-to-end sequential modeling

In this section, we present the model architecture for extracting intra-slice and inter-slice features in an end-to-end manner. Consider a CT scan with N slices shape $H \times W$, each slice can generate a 3-channel image by setting three different window widths and levels of Hounsfield Unit (HU) values.

The convolutional neural networks are good at extracting the local features in the images while cannot effectively establish connections between long-distance image features. However, the categories of intracranial hemorrhage correspond to the position, which demands the global relation or patch-wise relation extraction of the model. Hence, we choose the recently proposed Swin-Transformer model (Liu et al. 2021b) as the intra-slice extractor, which can establish the relationship between patches by attention weights.

Our workflow is illustrated in Fig.1, each CT scan is converted into a batch of 2D images and fed into the swin-transformer model to extract intra-slice features. The sequence transformer extracts the inter-slice features with *self-attention* schema, i.e, it automatically fuses intra-slice features with attention weights among different slices.

The features of intra-slice and inter-slice are converted to the logits respectively corresponding to six classification labels by different classification heads. Inspired by InceptionNet(Szegedy et al. 2015), we introduce the auxiliary classifier to speed up the convergence and avoid gradient vanishing problem in the swin-transformer model, and the *classification head 1* is the auxiliary classifiers. The logits are supervised with ground-truth by calculating the binary cross-entropy (BCE) loss L_1, L_2. During inference, we only take the prediction generated by *classification head 2* as the final output.

A performance enhanced version named swin-transformer v2 (Liu et al. 2021a) is proposed recently. However, running the complete swin-transformer v2 requires numerous memory with an entire CT scan. To improve the performance of the intra-slice feature extractor, we optimize the original swin-transformer architectures from pre-norm manner to post-norm inspired by swin-transfomer v2 (Liu et al. 2021a), which speeds up the convergence and keeps the training procedure more stable. The pre-norm and post-norm manner indicate the position of the layer-normalization in the residual block.

Enhanced modeling by semi-supervised learning

To achieve higher performance without increasing the extra inference complexity, we introduce the semi-supervised learning (SSL) method FixMatch (Sohn et al. 2020) which can fully utilize unlabeled samples to improve the model performance.

Consistency regularization is an important component of recent state-of-the-art SSL algorithms. It assumes the model should output similar predictions for the same image, even if the is perturbed by data augmentation operations. Consider an unlabeled sample x and two types of augmentations, strong and weak, denoted by $\mathcal{A}(\cdot)$ and $a(\cdot)$ respectively. FixMatch computes an *artificial label* $\widehat{y_i'}$ for each unlabeled weakly augmented samples $a(x_i)$ based on the model pre-

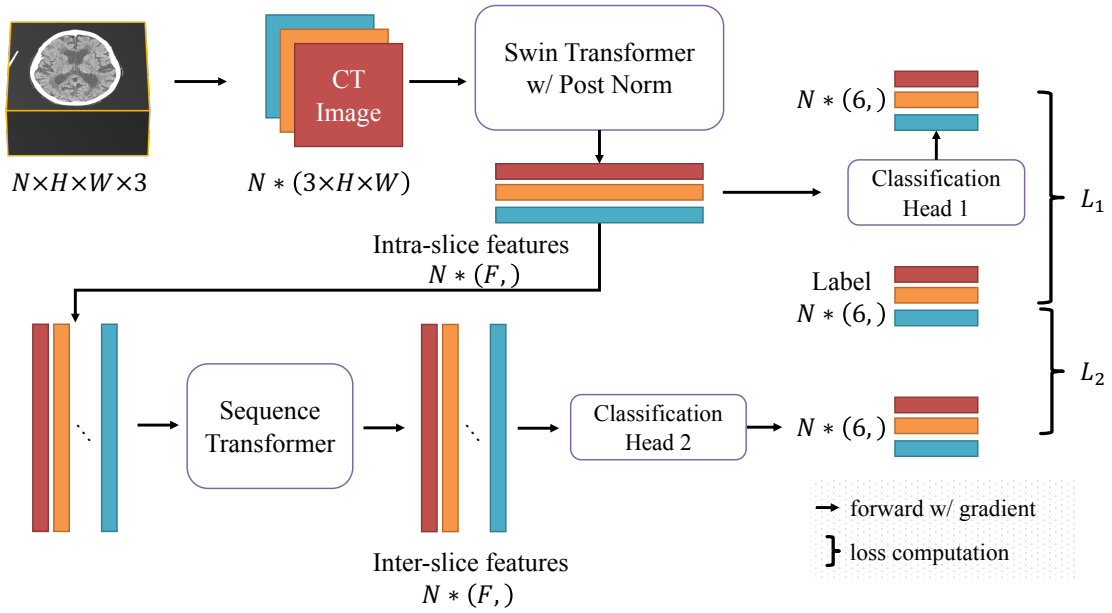

Figure 1: Workflow of our end-to-end model. One CT scan is transformed into a batch of 2D images with 3 channels. The intra-slice features are extracted by the swin-transformer model and transformed into inter-slice features by the sequential transformer. The auxiliary classifier is introduced to speed up the convergence speed.

diction $y_i' = M(a(x_i))$. The model is fine-tuned with the strong augmented samples $\mathcal{A}(x_i)$ and pseudo label $\widehat{y_i'}$ over N samples:

$$\mathcal{L}_u = \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}(\widehat{y_i'}, M(\mathcal{A}(x)), \tag{1}$$

with the $\widehat{y_i'} = argmax(y_i)$. The cost function $\mathcal{H}$ is binary cross-entropy function in this paper.

Inspired by FixMatch, the test set of the RSNA-IHD competition can be the unlabeled sample set x and the ensemble results can be seen as the predictions of weakly augmented samples. The augmentations are applied to the images to acquire $\mathcal{A}(x)$ including random crop, flip and rotation, gaussian blur, distortion, etc. As proposed in FixMatch, only the confident enough samples are chosen to fine-tune the model. The confident score s is defined as below:

$$s_i = max(1 - y_i', y_i'). \tag{2}$$

Our workflow of introducing the unlabeled slices into the model training is illustrated in Fig.2. Firstly, we choose the CT series from the ensemble results which the confident scores s of the model predictions are larger than the threshold τ_s for all slices in the series. Secondly, the predictions are binarized to acquire the artificial label, i.e, the pseudo label by the threshold τ_p. We simply choose $\tau_p = 0.5$ in our solution. Thirdly, the unlabeled slices and pseudo labels are mixed into a novel training set along with the annotated slices.

The FixMatch method improves the performance of the model. The model with the best performance is added to the model zoo, replacing the model previously involved in the ensemble. Therefore, the ensemble results could be improved continuously and the virtuous circle is established. The performance improvement would converge as the consistent prediction of the model and ensemble results increases during the cycle.

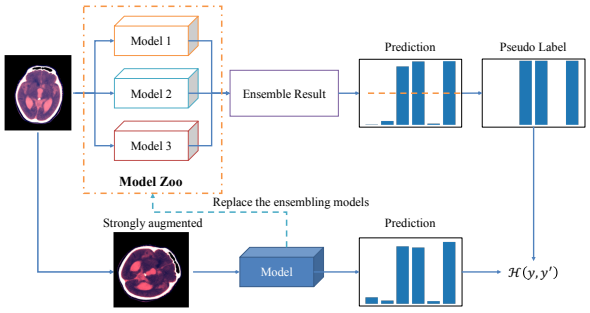

Figure 2: Semi-supervised learning procedure to reduce the number of the ensemble models. Inspired by FixMatch(Sohn et al. 2020), the ensemble result can be seen as the pseudo-label of the unlabeled samples(test set), the model is optimized by L_u with strong-aug unlabeled samples and L_1+L_2 with labeled samples simultaneously.

Experimental Results

Implementation details

Dataset and evaluation metric In our solution experiments on the RSNA-IHD dataset (Anouk Stein et al. 2019), the annotated slices are split into train and validation sets without

losing the consistency of the CT series, i.e, the slices in the same CT scan are split to the same partition. The distribution of the datasets is described in Table 2. With the obvious label imbalance, the metric of the online evaluator for the submission files is *weighted log-loss* for each slice and averaged across all samples, i.e, the lower loss leads to the higher rank.

Pre-processing As the categories of intracranial hemorrhages are corresponded to the location of the blood, the solutions need to consider both brain and skull information in the slice. Following the commonly practice (APPIAN 2019) of the top solutions, we choose three configuration of HU windows by W_c for the level and W_r for the width of the HU window. The HU window strategy converts a CT slice into a 3-channel image by $W_c = 40, W_r = 80$ for brain tissues, $W_c = 80, W_r = 200$ for blood features and $W_c = 40, W_r = 380$ for soft tissues. The conversion is illustrated in Fig.3.

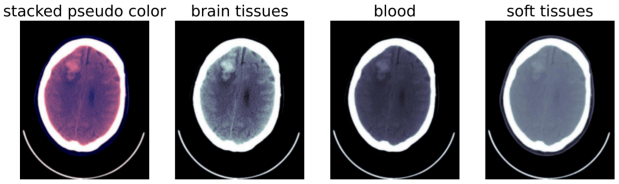

Figure 3: The conversion from CT slice to the 3-channel image. The slice is intraparenchymal hemorrhage positive and one can easily see the blood in the brain tissue and blood windows.

Every CT scan is consists of up to 60 slices with shape 512px $\times$ 512px, which brings a lot of memory requirements. After all the slices are converted to a batch of 3-channel images, the following steps are applied to complete the preprocess steps: Firstly, we remove the black edges outside of the brain region in every image by the opening operation. Secondly, the images are resized to 224px $\times$ 224px to fit the expected resolution of the swin-transformer. Finally, the images belonging to the same CT series are stacked into a 4D tensor in the slice stacking order of the *dicom* files.

Model Configuration The configuration named *Swin-B* is applied for the intra-slice feature extractor with the input resolution 224px $\times$ 224px, embedding-dim 128 and $(2, 2, 18, 2)$ swin-blocks for 4-stages respectively. The intra-slice features are converted by positional embedding before being fed into the inter-slice extractor. The standard two-layer sequential transformer is applied as the inter-slice feature extractor with two heads for each layer. The single-layer perceptron (linear layer) is applied as the classification head to transfer the features into the classification logits.

Training details All of our experiments are implemented with PaddlePaddle 2.2.0 and NVIDIA-P40. The intra-slice and inter-slice feature extractors are trained in an end-to-end manner with the weighted log-loss. The weights of the loss are $2, 1, 1, 1, 1, 1$ respectively for *any* and the others hemorrhage categories. The training procedure lasted 80,000 iterations (every series iterated 4 times) with the SGD optimizer. The warm-up and cosine-decay learning rate (LR) strategy is applied with the 300 warm-up iterations and the warmed LR is 0.001. The slices in train-set with human annotations are augmented by *RandomResizedCrop* with [0.8, 1.2] scale range, horizontal and vertical flip. The stronger augmented strategies are applied to the slices with pseudo-label, which is described in Section 2.2.

Model ensemble Our ensemble solution consists of the following steps: Firstly, the model of the best performance trained without pseudo-label is reversed with the best two models trained with pseudo-label. Secondly, these three models are used to acquire ensemble results with a weighted average strategy. Thirdly, we set the upper and lower thresholds τ_h, τ_l to binarize the predictions for the most confident probabilities. The ensemble weights of the models are related to their ranks on the *private leaderboard* as follows:

$$w_i = 1 - \frac{rank_i}{\sum rank_i}, i \in 1, 2. \tag{3}$$

For example, our best ensemble results are consists of three submissions, which are rank 1, rank 2 (w/ pseudo-label) and rank 33 (w/o pseudo-label) respectively. Firstly, we calculate the weights of the two groups by $w_1 = 1 - \frac{1+2}{1+2+33})$ and $w_2 = 1 - \frac{33}{1+2+33})$. Secondly, the weight for the pseudo-label results are further splitted into $w_{11} = (1 - \frac{1}{1+2}) * w_1$ and $w_{12} = (1 - \frac{2}{1+2}) * w_1$.

Ablation Study for Various Tricks

To completely study what tricks are helpful in our solution, we show the enhancement of each trick for the single model in Table 3. With all the tricks, our single model achieve the log-loss 0.04368 which outperforms the ensemble solution of the first place solution in RSNA-IHD competition. The **IntraExt** is the abbreviated form for intra-slice feature extractor, **InterExt** for inter-slice feature extractor, **RBE** for remove black edges, **DS** for deep-supervision schema of the intra extractor, **PN** for the post-norm manner of the swin-transformer, **DW** for the dynamic weighted log-loss (according to the proportion of positive samples), **SW** for the static weighted log-loss and **FM** for the FixMatch.

Besides the effectively tricks, there are also some attempts during the algorithm development that do not improve the performance:

- The external data. We introduced the CQ500 (Chilamkurthy et al. 2018) dataset into our training sets, which contains the same annotation format as RSNA-IHD. However, there is a series-level annotation different from the slice-level of RSNA-IHD. We tried to extract the maximum predictions over the slices to acquire the series-level predictions, but the model did not benefit from this attempt. We believe that our model are effective enough so that it can not profit by the coarse-grained multi-sample learning paradigm.
- The logical *any* prediction. As the *any* prediction is equivalent to *any positive* in the other five categories, we tried to adjust the classification heads to output five logits other than six, and

Dataset Split	Series	Slices	Positive label distribution					
			Epidural	Intra-parenchymal	Intra-ventricular	Subarachnoid	Subdural	Any
train	19569	677485	2841	32814	23955	31884	42438	97510
validation	2175	75318	304	3304	2250	3791	4728	10423
test	3518	121232			-			

Table 2: Sample and label distribution of the RSNA-IHD dataset.

Log loss on Private LB	Tricks							
	Intra-Ext	Inter-Ext	RBE	DS	PN	DW	SW	FM
0.09495	•							
0.06582	•	•						
0.06490	•	•	•					
0.06117	•	•	•	•				
0.05808	•	•	•	•	•			
0.05409	•	•	•	•	•	•		
0.05117	•	•	•	•	•	•	•	
0.04368	•	•	•	•	•	•	•	•
0.04383	The winner of RSNA-IHD with 20 ensemble models							

Table 3: Submission logs for the single model. The ablation studies reveal the effectiveness of our tricks, and our final single model (shown in red) outperform the ensemble solution of the winner team in RSNA-IHD competition.

$P_{any} = max(P_{EDH}, P_{IPH}, P_{IVH}, P_{SAH}, P_{SDH})$. Although we got performance improvement, it was so slight that we guessed it was caused by random factors. Moreover, we did not observe significant *any-others* mismatch problems in the six-logits outputs. Therefore, the logical relationship between the categories may not need to be explicitly expressed.

Conclusion

In this paper, we proposed an intracranial hemorrhage detection algorithm that outperforms the first place solution by a single model. The intra-slice and inter-slice feature extractor are integrated into one model with the end-to-end manner, which avoid cumulative errors. Moreover, to reduce the amount of the ensemble models, we introduced the recently proposed semi-supervised learning method which take the test-samples into the training set and improve the performance. Some open issues still exist, including more precisely detecting hemorrhages and extending our solution to other slice-wise applications. We release the source code and hope that other researchers will achieve more breakthroughs.

References

Anouk Stein, M.; Wu, C.; Carr, C.; Shih, G.; Kalpathy-Cramer, J.; Elliott, J.; kalpathy; Prevedello, L.; Marc Kohli, M.; Lungren, M.; Culliton, P.; Ball, R.; and MD, S. H. 2019. RSNA Intracranial Hemorrhage Detection. https://kaggle.com/competitions/rsna-intracranial-hemorrhage-detection. Kaggle.

APPIAN. 2019. 11th place solution (with updated code on github).

AUTOMATA. 2019. 5th place solution.

Chilamkurthy, S.; Ghosh, R.; Tanamala, S.; Biviji, M.; Campeau, N. G.; Venugopal, V. K.; Mahajan, V.; Rao, P.; and Warier, P. 2018. Development and validation of deep learning algorithms for detection of critical findings in head CT scans. *arXiv preprint arXiv:1803.05854*.

Darragh Hanley, D. L. 2019. 2nd place solution.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2021a. Swin Transformer V2: Scaling Up Capacity and Resolution. *arXiv preprint arXiv:2111.09883*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

MIRANDA. 2019. 4th place solution.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33: 596–608.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, 1–9.

TAKUOKO. 2019. 3rd place solution.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, X.; Shen, T.; Yang, S.; Lan, J.; Xu, Y.; Wang, M.; Zhang, J.; and Han, X. 2021. A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head CT scans. *NeuroImage: Clinical*, 32: 102785.