Efficient Scopeformer: Scalability and Feature Extraction Richness in Intracranial Hemorrhage Detection Challenge

Yassine Barhoumi barhou29@students.rowan.edu
Ghulam Rasool rasool@rowan.edu
Rowan University, New Jersey, USA

Editors: Under Review for MIDL 2022

Abstract

The feature map quality generated from the computed tomography (CT) scans constitutes a distinctive aspect for robust model performance in the medical computer vision field. Furthermore, well-engineered features are deterministic to ambiguous cases detection. A novel multi-CNN-based vision transformer model called Scopeformer was developed in our earlier work and used for the Intracranial Hemorrhage Detection challenge to classify various hemorrhage types within the RSNA2019 Brain CT Hemorrhage dataset. The model showed high scalability of the model size with an improved feature generating backbone. However, the model suffered from a large trainable parameter space resulting in a long training time. We adopt in this paper the Scopeformer model and aim to reduce the parameter size and enhance the global feature map richness. Effective feature projection methods were used to reduce the redundancy of the feature space. Furthermore, we used small vision transformer (ViT) versions with four different types of pretrained CNN architectures and introduced three ViT configurations to reduce the self-attention complexity within the transformer layers. Our best model achieved an accuracy of 96.03 % and a weighted logarithmic loss of 0.1088 with an eight times reduction of trainable parameter space. A second model with comparable performance further reduced the parameter space to 17 times our best-performing model.

Keywords: Computed Tomography (CT) scans and slices, Intracranial hemorrhage detection, CNN, ViT.

1. Introduction

The early detection and classification of intracranial hemorrhages within computed tomography slices is critical within the first 24 hours of a head injury for fast clinical decisions (Justine and Smith, 2012; Wardlaw, 2001). This can often require highly qualified doctors for detecting subtle details showing the existence of a lesion within the brain tissues. Emerging computer vision techniques offer faster and more robust models compared to highly trained radiologist predictions (Gong et al., 2007; Chilamkurthy et al., 2018; Ker et al., 2019; Patel et al., 2019). The success of these models resides in the vast annotated datasets offered by the community, such as the dataset provided with the RSNA intracranial hemorrhage challenge. The RSNA intracranial hemorrhage CT dataset was collected by Adam E. et al. (Flanders et al., 2020) from multiple scanner types used in different institutions worldwide and aimed to capture complex real-world details of the hemorrhage sub-types. This dataset is considered the current largest dataset available online.

Convolutional neural networks are the de-facto architectures in the medical computer vision domain for extracting high-resolution features (Mihail et al., 2020; Marissa et al., 2020). The successful implementation of the transformer model (Vaswani et al., 2017)

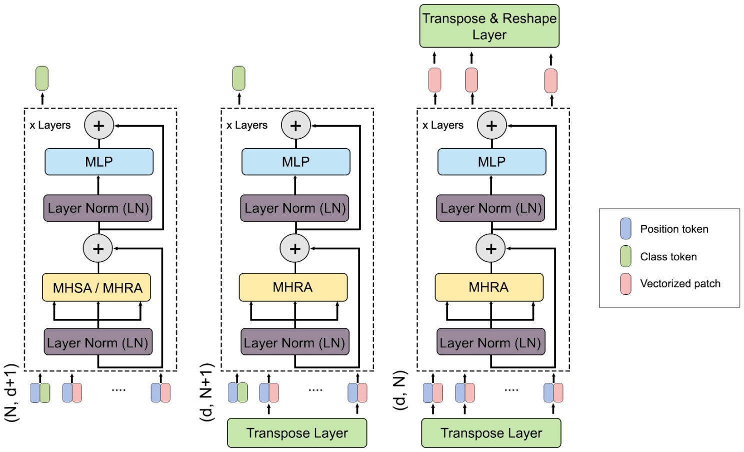


Figure 1: ViT Scopeformer configurations. The first configuration is a ViT block with an input of vectorized patches extracted from the CNNs features. The second configuration introduces a transpose layer to transform the channel-wise patches into feature-wise patches. The third configuration dismisses the token class and uses all the feature tokens as input. The output of the third block will be transposed to retrieve back the dimension of the CNN features, which we feed to the classification module.

applied to images known as vision transformer (ViT) was a milestone in the computer vision domain (Dosovitskiy et al., 2021).

Various successful implementations of the ViT in the medical field were proposed and proved to defeat standard pure convolution-based models by a wide margin (Dai and Gao, 2021). Motivated by the performance of these two models, we proposed in our earlier work (Barhoumi and Rasool, 2021), a hybrid architecture consisting of multiple Xception CNN models (Chollet, 2017) for feature extraction and several vision transformer encoders for differentially extracting significance weights of the feature map relevant to classification. Results showed that the classification accuracy is proportional to the number of Xception models and the variety of the pretraining methods used to train the CNN architectures. In this paper, we enhance our proposed n-CNN-ViT Scopeformer model by employing a more efficient version of the ViT and improved feature extraction method.

2. Methodology

Our objective is to make the Scopeformer model, a newly created convolution-based Transformer (Barhoumi and Rasool, 2021), more scalable and efficient.

2.1. Architecture

We modified the original Scopeformer architecture by introducing several changes in the feature extractor CNNs and the ViT. There are four modules as shown in Fig. 2.

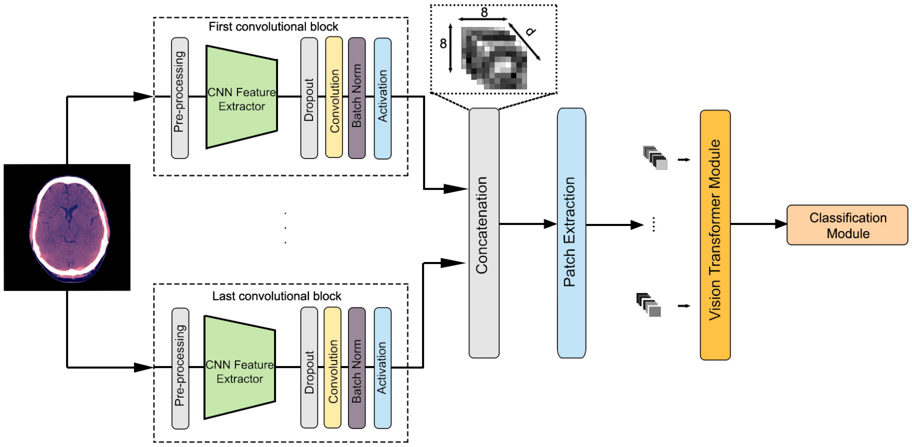


Figure 2: Scopeformer architecture. The proposed model is composed of four main modules: feature extraction, patch extraction, vision transformer encoding, and classification. A single input image will be fed to several CNN models to extract a variety of features and construct feature maps. These feature maps are processed by the patch extraction module and vectorized. These vectors form the input to the transformer encoder and the output is taken from the classification module.

2.1.1. Module 1: Scopeformer Backbone

The proposed Efficient Scopeformer uses a variety of CNNs to build the feature extraction block. The backbone CNNs include ImageNet-pretrained ResNet 152 V2, EfficientNet B5, DensNet 201, and Xception. The features generated by each CNN are concatenated along the channel axis to form a *global feature map*. However, constructing such a feature map requires that the individual feature maps generated by each CNN to have identical height and width. We propose augmenting each CNN with a single trainable 1×1 convolutional layer that projects the features to an appropriate space.

The input to the Efficient Scopeformer consists of a tensor with a dimension of $H \times W \times 3$, where H represents the height, W represents the width, and 3 is the number of channels. The image is concurrently fed to four CNNs to generate high-level feature maps. The channel dimension of all four feature maps will be reduced using 1×1 convolution layer to $8 \times 8 \times \frac{d}{4}$, where d is the size of the *global feature map*.

2.1.2. Module 2: Patch extraction

The *global feature map*, $\mathcal{X} \in \mathbb{R}^{8 \times 8 \times d}$ is passed through a patch extraction module which reshapes § into a sequence of flattened patches $X_p \in \mathbb{R}^{N \times d}$, where N is the number of the extracted patches and is assumed as 64 for the current experiments.

2.1.3. Module 3: Scopeformer ViT

We evaluated three different ViT configurations for the proposed architecture as depicted in Figure 1. These configurations include (1) Deep Scopeformer, (2) deep Scopeformer TR (Transpose), and (3) Efficient Scopeformer.

Baseline Scopeformer Configuration. In this configuration, we feed a set of vectors generated by patch extraction layer to ViT encoders. We used trainable position encoding vectors coupled with vectorized patches and a trainable class (CLS) token. The dimension of the input to ViT encoder block is $Y \in \mathbb{R}^{N\times d+1}$.

We used two self-attention variants. The first one is referred to as multi-head self-attention (MHSA) (Dosovitskiy et al., 2021) and the second variant as the multi-head re-attention (MHRA) (Zhou et al., 2021). The key difference resides in the introduction of a trainable transformation matrix. These variants are given by:

$$\mathrm{MHSA}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

$$\mathrm{MHRA}(Q,K,V) = \mathrm{Norm}\left(M^T\left(\mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\right)\right)V, \tag{2}$$

where $M \in \mathbb{R}^{h\times h}$ is a learnable transformation matrix, and h is the number of self-attention heads.

Deep Scopeformer TR Configuration. The second Scopeformer ViT configuration applies a transpose operation to the set of vectors produced by the patch extraction layer. The output of the transpose layer is summed up with the position encoded vectors and concatenated with the CLS token. The dimension of the resultant set of vectors is $Y_T \in \mathbb{R}^{d\times N+1}$. We used only MHRA self-attention variant (Eq. 2) in our experiments.

Efficient Scopeformer Configuration. The third Scopeformer module discards the CLS notion used in previous configurations. In this settings, we use all the features generated by ViT encoders for classification. As such, the dimension of the input and output of ViT encoders remain identical and is equals to $Y_T \in \mathbb{R}^{d\times N}$. We use a Transpose and Reshape layer at the ViT output to get the appropriate dimension for the feature map. We use MHRA self-attention variant to compute self-attention.

2.1.4. Module 4: Classification module

The classification module in baseline and the deep Scopeformer TR configurations receives a single CLS token. The output of this token is turned into a prediction using a multi-layer perceptron (MLP) with a sigmoid activation function and a single hidden layer. In the efficient Scopeformer configuration, the classification module receives a set of reshaped features $x_t \in \mathbb{R}^{8\times 8\times d}$. The classification module applies a 2D average pooling layer, followed by a flatten layer. Finally, the inference of the class is done via a dense layer with a sigmoid activation function.

2.2. Dataset

The Radiological Society of North America (RSNA) dataset (Flanders et al., 2020) was released in the 2019 Intracranial Hemorrhage (ICH) detection challenge hosted by the Kaggle platform. The dataset contains 755,273 annotated 16-bit grayscale computer tomography (CT) scans saved in the DICOM format. Individual images consist of pixels that have a range of 0 to 2^{16} with a resolution of 256^2, referred to as Hounsfield Units (HU). HU represents the density of the scanned matter. Trained physicians categorized each CT slice with one or more types of the brain hemorrhage, including epidural (EPH), intraparenchymal (IPH), intraventricular (IVH), subarachnoid (SAH), and subdural (SDH). An additional hemorrhage target class (ICH) was appended to indicate the existence of any hemorrhage. Attenuation HU values are indicative for the content of the scan (Broder and Preston, 2011). For instance, bones have an attenuation value ranging between 250 and 1000, and fat and muscle have attenuation values (AV) ranging between 50 and 100. Applying HU windows on a CT slice yields an 8-bit grayscale image. We use three windows of HU as channels in the input of the Scopeformer model. Our settings for HU windows were: brain AV$\in [40, 80]$ HU, subdural window AV $\in [80, 200]$ HU, and soft tissue window AV $\in [80, 200]$ HU.

Table 1: Various design configurations of Scopeformer - hyperparameters and learnable parameters.

Model	CNN Blocks	Layers	Feature Size	MLP	Heads	Parameters
Scopeformer (S)	4	8	516	3072	12	34 M
Scopeformer (B)	4	8	512	4096	16	42 M
Scopeformer (M)	4	8	512	5120	16	43 M
Scopeformer (L)/4	4	4	1024	4096	16	51 M
Scopeformer (L)/8	4	8	1024	4096	16	102 M
Scopeformer (L)/16	4	16	1024	4096	16	203 M
Deep Scopeformer (L)/8	4	8	1024	4096	16	102 M
Deep Scopeformer TR (L)/8	3	8	384	4096	16	**6 M**
Efficient Scopeformer	3	8	384	4096	16	**6 M**
Scopeformer (Barhoumi and Rasool, 2021)	3	12	3072	3072	8	755 M
Scaled Scopeformer (Barhoumi and Rasool, 2021)	4	8	4096	4096	16	870 M

2.3. Experiments

We used a host of configurations of the proposed Scopeformer architecture in our experiments. Details bout the various Scopeformer architecture in terms of various hyperparameters are presented in Table 1. In general, our experiments comprise of four main parts. In the first part, we evaluate the effect of the size of various variants of Scopeformer on the classification accuracy. Four variants are evaluated, small, base, medium, and large denoted by S, B, M, and L. In the second part, we investigate the number of ViT encoder blocks required to optimize the size and performance of the large Scopeformer model. We considered 4, 8, and 16 ViT encoders. The third part included a transition from ViT model to a different version known as DeepViT (Zhou et al., 2021), where we change MHSA module by the MHRA module. In the forth and final part of our experiments, we compared proposed Scopeformer configurations.

In all the experiments, we start with pre-training the Scopeformer model using ImageNet-1k dataset (Russakovsky et al., 2014). Later, we train all models using the RSNA dataset (Flanders AE, 2019). In the module 1 (backbone CNNs), we freeze ≈ 70% of layers in each CNN and keep top ≈ 30% layers trainable along with the newly introduced 1 × 1 convolution layer. Following guidelines of RSNA Intracranial Haemorrhage Challenge (ICH), we use *multi-label log-loss* for model training and *weighted accuracy* for model evaluation. The loss is defined as:

$$L_{\text{multi-BCE}}\left(y, \tilde{y}\right) = -\sum_{n=1}^{6} y_t \log \tilde{y_t} + \left(1 - y_t\right) \log\left(1 - \tilde{y_t}\right) \tag{3}$$

3. Results and Discussion

We evaluate the overall performance of the models based on three metrics, (1) the classification accuracy on the RSNA dataset, (2) the global feature richness generated by backbone CNNs, and (3) the model size (total number of trainable parameters).

3.1. The Effect of Size of Scopeformer

Tables 2 and 3 show the results of experiments performed with different sized Scopeformers. Our currently proposed Scopeformer models have four sizes (that is, S, B, M, and L) and use reduced number of trainable parameters compared to the original Scopeformer (Barhoumi and Rasool, 2021). The parameters reduction is linked to the 1 × 1 convolutional layer added before the ViT module. We gradually increase the model complexity of S, B, and M variants by varying the MLP dimension and the number of self-attention heads within the ViT module as depicted in table 1.

In Table 2, we note that the base model outperforms the small and medium counterparts. The *Scopeformer (L)/8* model adopts the configuration of the base variant with a global feature dimension $d = 1024$. The feature size increment resulted in a proportional increment of the model trainable parameters. The large model (L)/8 performed the best among the proposed variants.

Table 2: Performance of the different Scopeformer variants.

Model	Accuracy	Loss	Recall	Trainable Parameters
Small (S)	93.00%	0.1703	84.95%	34M
Base (B)	93.92%	0.1461	89.29%	42M
Medium (M)	93.88%	0.2285	88.44%	43M
Large (L)/4	93.12%	0.1378	87.81%	51M
Large (L)/8	**94.69%**	**0.1197**	**89.33%**	102M
Large (L)/16	92.57%	0.1395	87.34%	203M

3.2. The Effect of Number of ViT Encoders

We evaluate the effect of the number of ViT encoders on Scopeformer (L)/8 model using 4, 8, and 16 encoders. As presented in Table 1 and also in Table 2, the number of parameters scales linearly with the number of encoders. However, this is not the case with the model performance. Table 2 shows that Scopeformer (L)/8 performs better than all others. On the other hand, the largest model with 16 encoders, Scopeformer (L)/16 has the lowest accuracy.

In Figure 3, we plot the cosine similarity between the features generated by each ViT encoder and the last layer of the model. We can observe that the increased similarity among features of the Scopeformer (L)/16 model may have contributed to the performance decline. Also, using fewer features, as in the case of Scopeformer (L)/4 model, the model may end up performing sub-optionally. We observe (Figure 3) that shallow models lead to sub-optimal performances, and deeper models may require more data to reduce similarity among ViT features to perform optimally. In summary, there is an optimum number of ViT encoders based on the complexity of the dataset and the effectiveness of the backbone CNNs in extracting rich features.

3.3. The Effect of Two Different Self-attention Variants

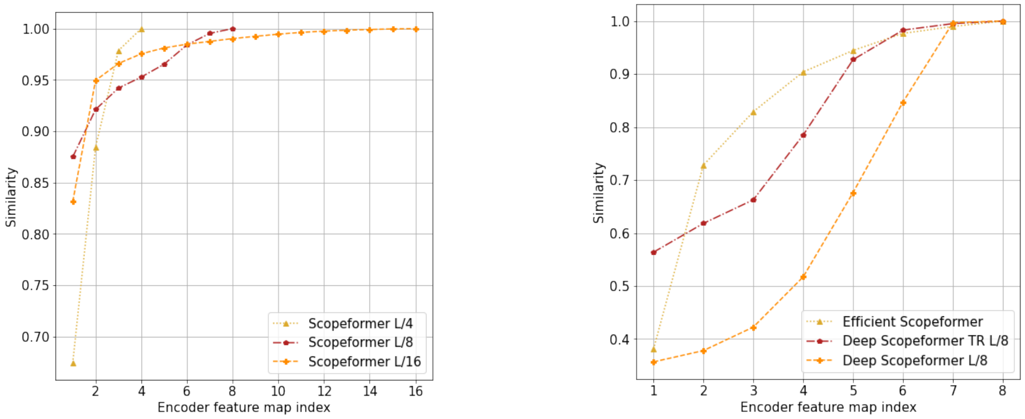
The Deep Scopeformer (L)/8 builds on the Scopeformer (L)/8 model by replacing the MHSA layer with an MHRA layer. The additional trainable matrices M adds only a small number of parameters to the Scopeformer (L)/8 model. We observe a significant dissimilarity among ViT encoders features of the *Deep Scopeformer (L)/8* in figure 3 (b), implying feature richness acquired by the model due to the MHRA heads correlations. This configuration resulted in an accuracy improvement by +1.11% as shown in table 4.

3.4. ViT Scopeformer Configurations

We approach the Self-attention complexity problem by introducing a transpose layer prior to the vision transformer module. The attention weights matrix in *Deep Scopeformer (L)/8* has a dimension of 1024^2. In the second and third ViT configurations, the attention weights matrices have dimensions of 65^2 and 64^2 respectively. The use of the transpose layer has substantially contributed to the reduction of the number of trainable parameters as indicated in 1 due to the MHRA quadratic computation complexity. Furthermore, transposing the input sequence effectively conserve the feature content retrieved by the classifier module. Table 4 shows the performance of the three proposed configurations. The *Efficient Scopeformer* variant performed similarly to the *Deep Scopeformer (L)/8*. We speculate that the role of the ViT module in this configuration is to improve the global features map. More details in Appendix A.

Table 3: Model performance on individual target classes

	Accuracy			
	Large	**Medium**	**Base**	**Small**
All	**71.34%**	60.26%	70.5%	70.83%
Epidural	96.98%	90.18%	95.73%	**98.08%**
IPH	85.94%	71.10%	**87.28%**	85.95%
IVH	90.5%	70.73%	**91.72%**	90.13%
SAH	**78.69%**	65.49%	78.57%	77.04%
SDH	**77.08%**	60.78%	74.35%	74.54%



(a) Scopeformer variants

(b) Deep Scopeformer variants

Figure 3: Cosine similarity of the vision transformer encoder feature maps with respect to the last encoder feature map.

Table 4: Model performance for different Scopeformer modalities

	Accuracy	**Loss**	**Trainable Parameters**
Scopeformer (L)/8	94.69%	0.1197	102M
Deep Scopeformer (L)/8	**96.03%**	**0.1088**	102M
Deep Scopeformer TR (L)/8	95.40%	0.1176	**6M**
Efficient Scopeformer	95.77%	0.1160	**6M**

4. Conclusion

We have explored potential model improvements on our previous implementation of the convolutional based vision transformer model called Scopeformer. The scope of the study covered model trainable parameters with respect to the model performance and the training efficiency of the Scopeformer architecture on the RSNA hemorrhage detection dataset. We explored the effect of using multiple off-the-shelf CNN models on the global feature richness of the architecture and investigated a feature projection method to reduce the large redundant feature space into a lower and efficient one. Furthermore, We conducted a parametric optimization study to evaluate the size effects on model performance and efficiency. We implemented three vision transformer configurations to evaluate the Re-attention module within the Scopeformer model and the channel-wise versus feature-wise patch extraction of the global feature map. Results show increased richness of the resultant features due to the different CNN architectures. The Re-attention module increased dissimilarities of the vision transformer features resulting in improved performances and allowing deeper models. With our proposed feature-wise patch extraction method, the model size was reduced 17 times with comparable performance. Furthermore, our Efficient Transformer module improved the global features map correlations and contributed to better performance.

References

Yassine Barhoumi and Ghulam Rasool. Scopeformer: n-cnn-vit hybrid model for intracranial hemorrhage classification, 2021.

Joshua Broder and Robert Preston. Chapter 1 - imaging the head and brain. In Joshua Broder, editor, *Diagnostic Imaging for the Emergency Physician*, pages 1–45. W.B. Saunders, Saint Louis, 2011. ISBN 978-1-4160-6113-7. doi: https://doi.org/10.1016/B978-1-4160-6113-7.10001-8. URL https://www.sciencedirect.com/science/article/pii/B9781416061137100018.

Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamal, Mustafa, Biviji Norbert G, Campeau Vasantha, Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study, 2018.

François Chollet. Xception: Deep learning with depth-wise separable convolutions, 2017.

Yin Dai and Yifan Gao. Transmed: Transformers advance multi-modal medical image classification, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Adam E. Flanders, Luciano M. Prevedello, George Shih, Safwan S. Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T. Mongan, Anouk Stein, Felipe C. Kitamura, Matthew P. Lungren, Gagandeep Choudhary, Lesley Cala, Luiz Coelho, Monique Mogensen, Fanny Morón, Elka Miller, Ichiro Ikuta, Vahe Zohrabian, Olivia McDonnell, Christie Lincoln, Lubdha Shah, David Joyner, Amit Agarwal, Ryan K. Lee, and Jaya Nath. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge, 2020.

Shih G et al (2020) Flanders AE, Prevedello LM. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. radiology, 2019.

Tianxia Gong, Ruizhe Liu andChew Lim Tan, Neda Farzad, Cheng Kiang Lee, Boon Chuan Pang, Qi Tian, Suisheng Tang, and Zhuo Zhang. Classification of ct brain images of head trauma, 2007.

Elliott Justine and Martin Smith. The acute management of intracerebral hemorrhage: a clinical review, 2012.

Justin Ker, Satya P. Singh, Yeqi Bai, Jai Rao, Tchoyoson Lim, and Lipo Wang. Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study, 2019.

DelRocini Marissa, Chris Angelini, and Ghulam Rasool. Identification of abnormalities in head computerized tomography scans, 2020.

Burduja Mihail, Radu Tudor Ionescu, and Nicolae Verga. Accurate and efficient intracranial hemorrhage detection and subtype classification in 3d ct scans with convolutional and long short-term memory neural networks, 2020.

Ajay Patel, Sil. C. van de Leemput, Mathias Prokop, Bram Van Ginneken, and Rashindra Manniesing. Image level training and prediction: intracranial hemorrhage identification in 3d non-contrast ct, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. URL http://arxiv.org/abs/1409.0575.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Joanna M Wardlaw. Overview of cochrane thrombolysis meta-analysis, 2001.

Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer, 2021.

Appendix A. Global feature map

Figure 4 , 5 and 6 plot CNNs features generated by three different architecture for an epidural example. We observe high variability of individual CNN features and no apparent similarity among the features generated by different CNNs.

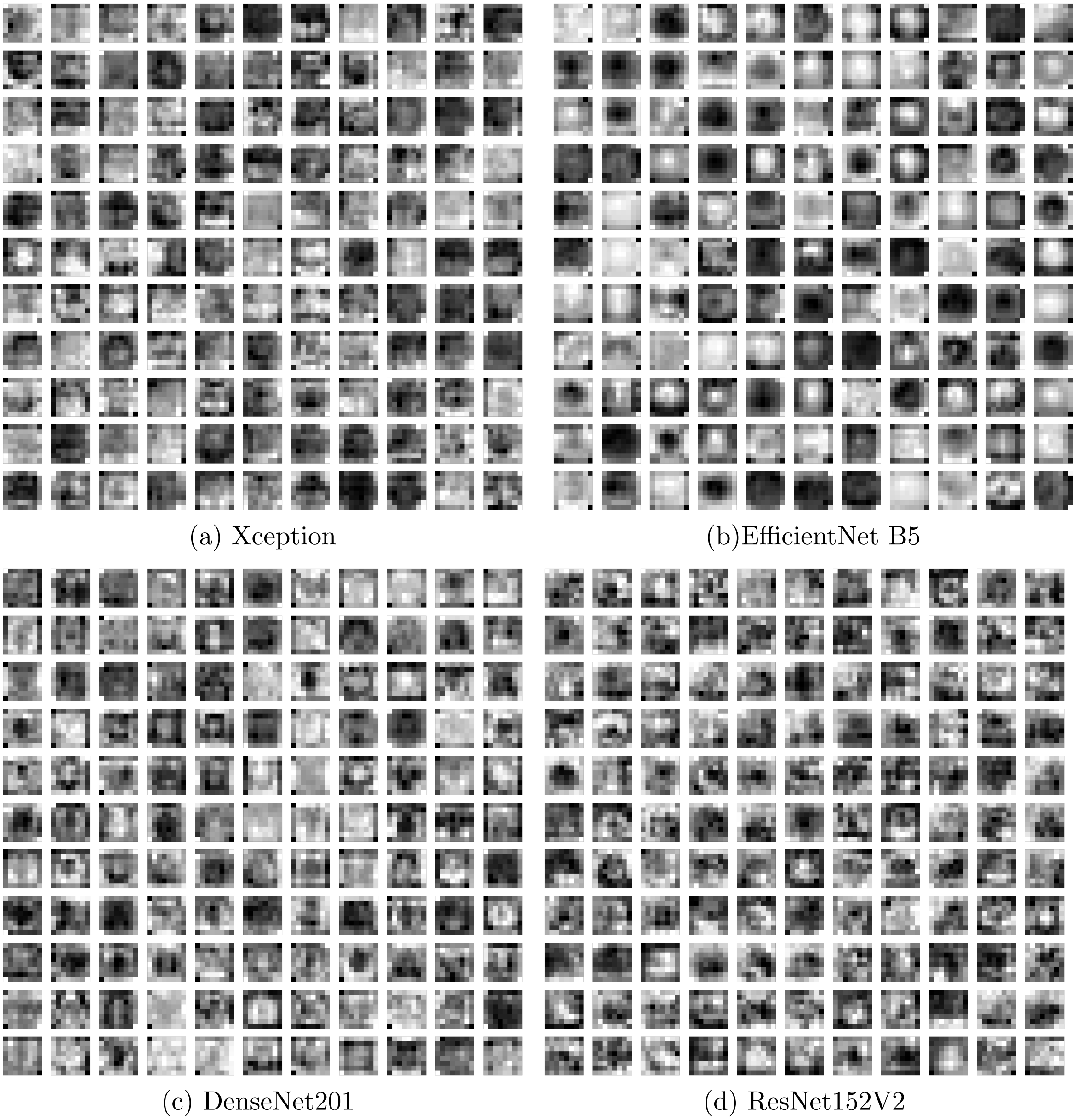
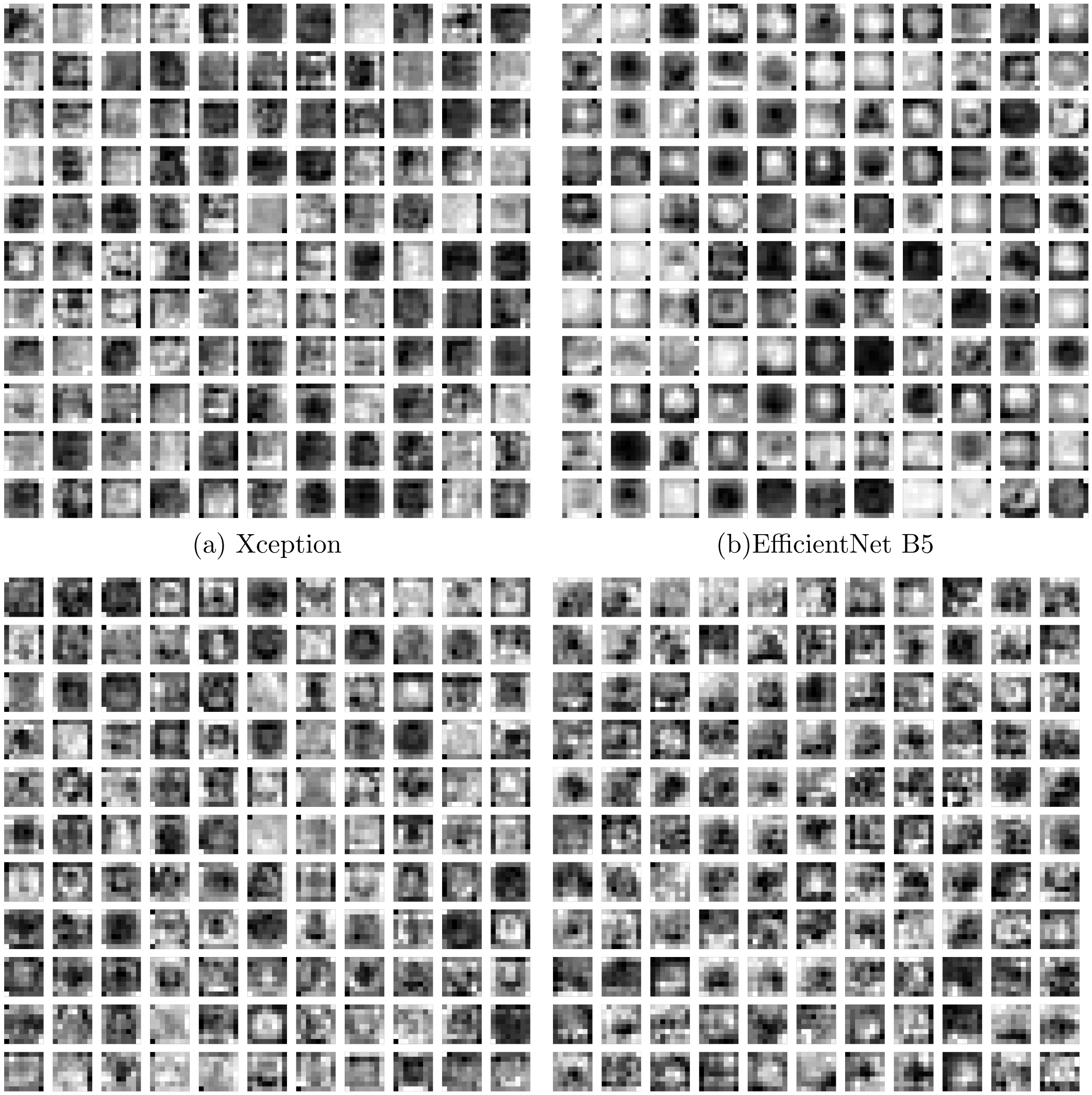


Figure 4: Feature maps visualization of an epidural type hemorrhage example. Scopeformer (L)/8

Subsequently, the resultant global feature map has low redundancy and higher feature richness comprised in the reduced feature space. However, DenseNet model showed the highest redundancy of the resultant features. To this end, we conducted an ablation study for the *Deep Scopeformer TR* and *Efficient Scopeformer* resulting in removing the model from the backbone for further parameter reduction.



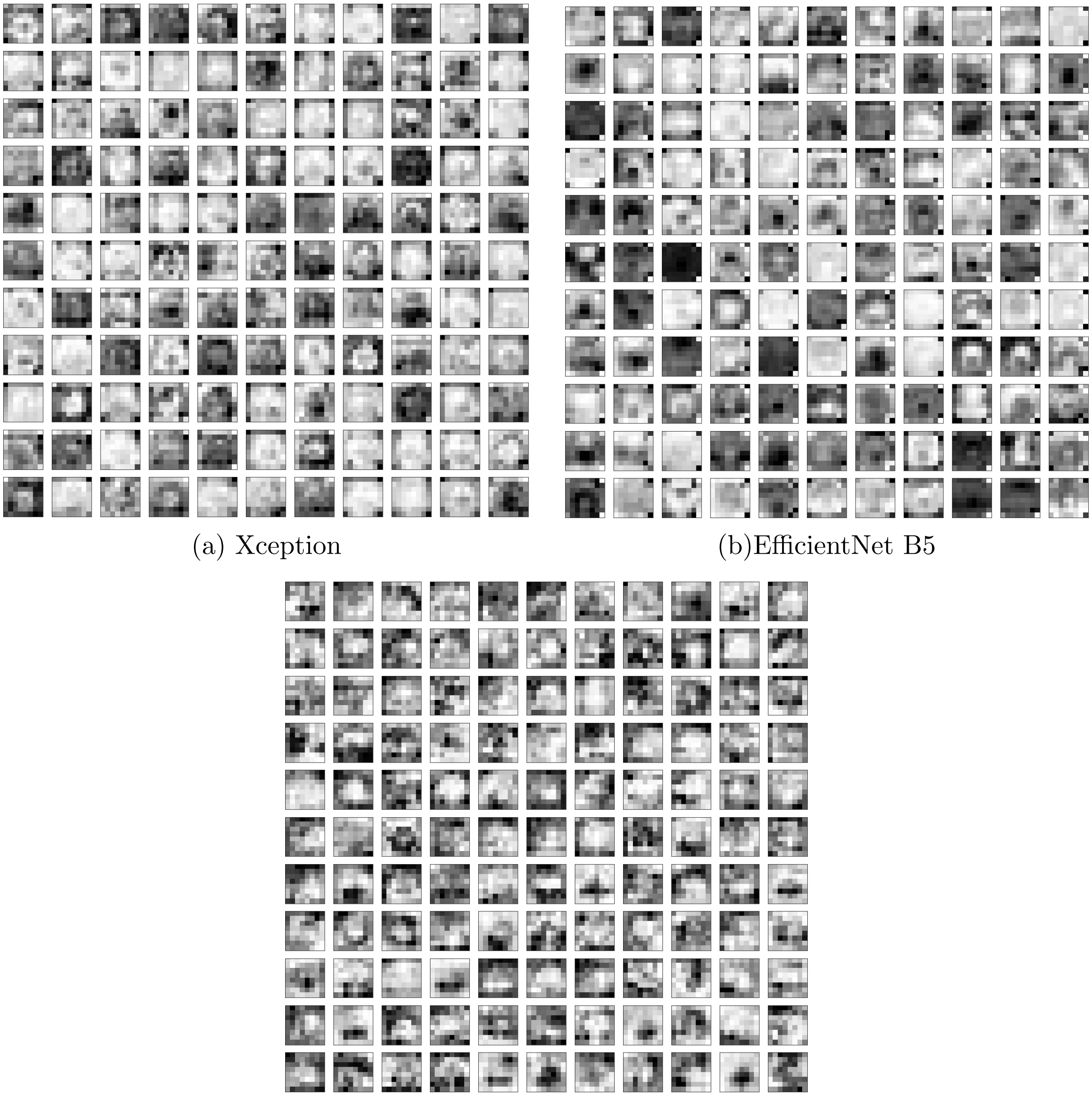
(a) Xception

(b)EfficientNet B5

(c) DenseNet201

(d) ResNet152V2

Figure 5: Feature maps visualization of an epidural type hemorrhage example. Deep Scopeformer (L)/8



(a) Xception

(b)EfficientNet B5

(d) ResNet152V2

Figure 6: Feature maps visualization of an epidural type hemorrhage example. Efficient Scopeformer

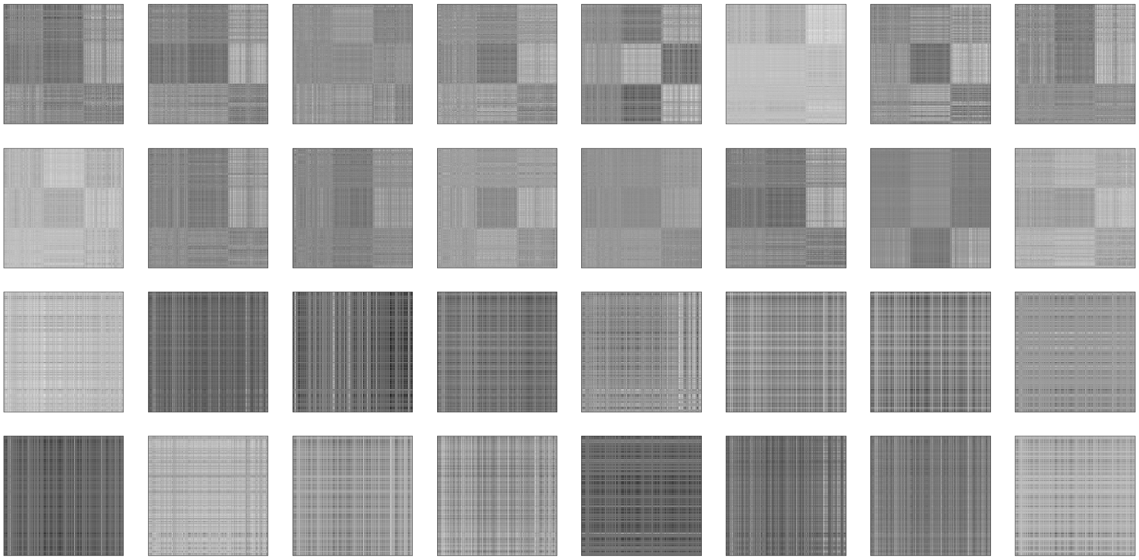
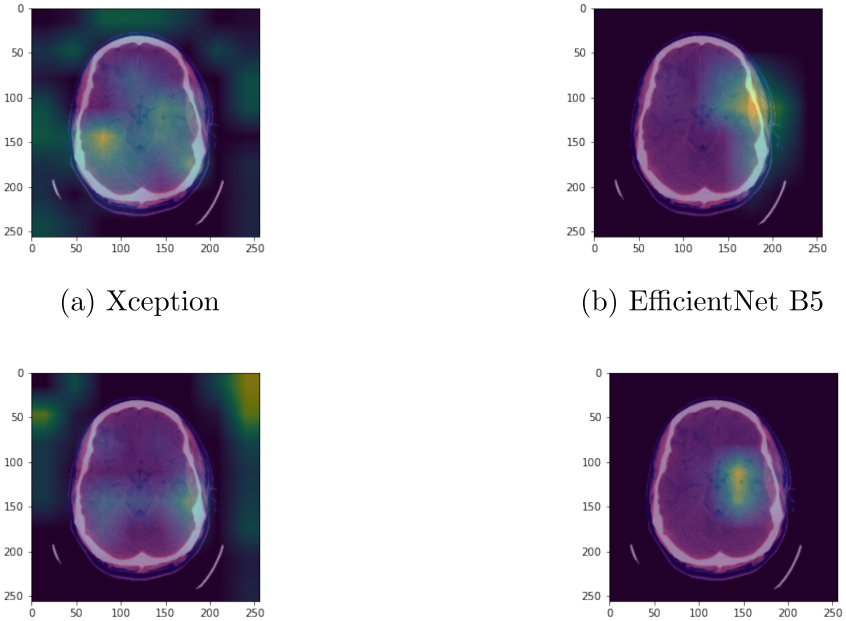


Figure 7: Attention pattern visualization of the Efficient Scopeformer model. The first and second row represent the 16 attention heads of the first encoder layer. The third and fourth row represent the 16 attention heads of the last encoder layer. Each attention map has a dimension of 384 × 384

Figure 7 shows the attention patterns visualizations of the 16 MHRA heads concerning the first and last ViT encoders. In the first layer, the model learns to correlate individual features across the features derived from different CNNs. Each head learns different correlations patterns among the set of features. The last layer shows a global correlations patterns among all features

Appendix B. Grad-CAM

Figure 8 plots a Grad-CAM visualization of an epidural type hemorrhage example for the Deep Scopeformer (L)/8 model. We observe high variability of the regions where the model considers to conduct the classification. We note that in many cases the DenseNet model contributed the least to the classification and in some cases, was shown to be mapping to the wrong regions on the image.



(a) Xception

(b) EfficientNet B5

(c) DenseNet201

(d) ResNet152V2

Figure 8: Grad-CAM visualization of an epidural type hemorrhage example.