Vision Transformer with EfficientNet Backbone for Severity Levels of Dementia Classification

Palawat Busaranuvong [1] **Amorn Chokchaisiripakdee** [1] **Aishwarya Ramakrishnan** [1] **Yu Zhang** [1]

Abstract

We propose a new combination of Efficient-Net backbone and Vision Transformer (ViT) for Severity levels of Dementia classification. For our contribution, we pass dementia images to a SOTA CNN backbone (EfficientNet-B0) to obtain the feature maps and pass them to the ViT. We explored several data augmentation techniques to boost the performance of the EfficientNet-VIT Classifier. Experimental results conducted on an imbalanced image dataset on Alzheimer's disease demonstrates the promising capability of our model to be compared with CNN's. The hybrid Vision Transformer (EfficientNet-B0-VIT-Ti) yields 98.85% accuracy respectively by beating the 97.10% and 94.75% balanced accuracies of Efficient-B3 and DieT-Ti respectively.

1. Introduction

Medical Image Analysis is a popular and yet challenging topic that brings the use of Deep Learning into medical society. In our project, we applied deep neural networks to classify the severity levels of dementia. Our training dataset is imbalanced and not very easy to be diagnosed by brain experts. For future usage, we would like to have a DL model that helps medical companies for early diagnosis of the dementia level.

Even though convolutional architectures remain dominant in solving computer vision tasks, the success of Transformers in NLP became an inspiration to apply them to images for classification with minimal modifications. Since the vision transformer takes the image globally and can freely look everywhere on the image, the ViT is data hungry to learn how to focus and what to focus on with the right attention. Also, leveraging pure transformers for Image Classification tasks requires massive data for training.

[1]Worcester Polytechnic Institute. Correspondence to: Every Authors <WPI>.

CS/DS 541 Deep Learning Final Project, Worcester Polytechnic Institute.

For our research contribution, we combined convolution to vision transformers by adding a state-of-the-art CNN architecture, i.e., EfficientNet-B0, at the bottom of the embedding layer of the vision transformer. That means the image will be passed to the convolutional backbone before applying its feature maps to the embedding layer of Vision Transformer. Since the size of input sequences (feature maps) is smaller than the size of the raw input image, we expect the Transformer encoders to learn to classify image patterns faster and better.

2. Related Work

As the state-of-the-art in image classification has been dominated during the past decades, the convolutional neural networks are famous approaches to solve the image classification problem. The image classification tasks have been using the advantage of convoluting images to aggregate the local information in each layer, then pass it to other layers. This CNN is too local when the image classification field becomes more global. To beat the-state-of-the-arts of CNN such as ResNet, DenseNet, VGG, the vision transformer has been invented.

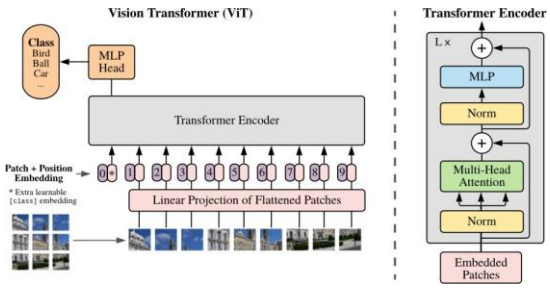


Figure 1. Vision Transformer (ViT) Architecture. Dosovitskiy et al. (2020)

The transformer is based on the natural language processing applications which are developed by Google [1]. The transformer is solely based on the attention mechanisms and is the basic of BERT tokenization and pretraining model. By replacing the recurrent layers which are commonly used in the encoder-decoder architecture with the multi-head attentions, the transformer outperforms the RNN and CNN architecture based.

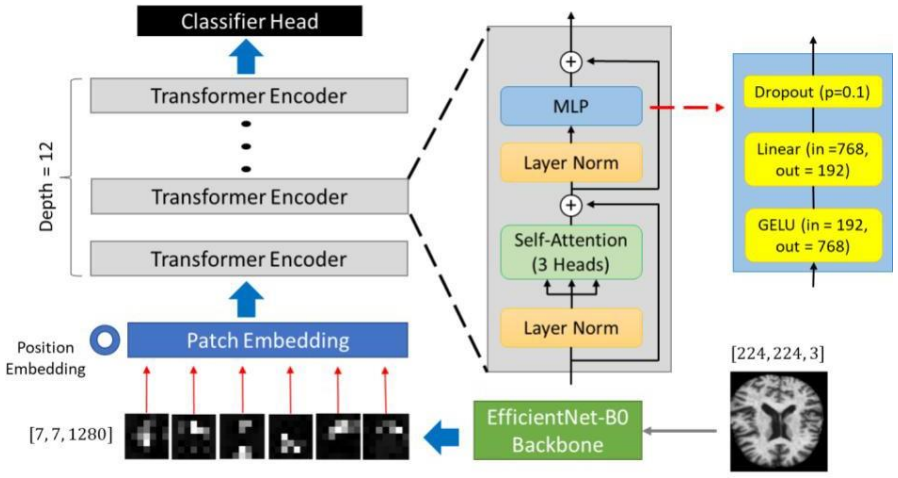


Figure 2. EfficientNet-ViT-Tiny Architecture (Ours). The demonstration of transformer encoder came from Dosovitskiy et al. (2020)

However, the transformer is originally used for the NLP purposes. To apply in the image, classification applications, the Vision Transformer (ViT) is also developed by Google.

The Vision Transformers (ViT) is based on the same architecture of the original transformer. Instead of word tokenization in the sentence, the ViT is splitting the images into patches of 16-by-16 pixels images as in figure 1. Then, the linear projection process adds the labels of position for each patch and flatten the patches. The transformer encoder takes the embedded patches to normalize and pass to the multi-head attention encoder. And the encoder normalizes and passes through the multi-layer perceptron. Finally, the classification head predicts the classes which can use the softmax function for example.

Since the vision transformer takes the image globally and can freely look everywhere on the image, the ViT is data hungry to learn how to focus and what to focus on with the right attention [2].

As the transformer emerges to the computer vision field, the ViT is not the only approach to tackle the computer vision problem. The Data-efficient image Trans- former (DeiT) [3] is developed using the ViT architecture with teacher-student strategy. The main objective is to improve the efficiency of the ViT architecture by efficiently scaling its architecture, and adding the last token called "distillation token" for accuracy improvement. The distillation token works similarly to the class token. For example,if the image has a "red giant apple" class token, the distillation token is predicted as "apple". We can see that there are many approaches to apply the transformer in the image recognition field.

3. Proposed Method

The data-efficient image transformer (DeiT) shows a significant improvement compared to the basic vision transformer (ViT) in the ImageNet competition. The DeiT with knowledge distillation beats the state-of-the-art EfficientNet architecture, which is the most efficient CNN in 2020 without using 300+ million unpublished images for training the model unlike the ViT. However, the disadvantage of this DeiT with distillation is that we need to have a good teacher network, which is "trained" beforehand, to teach a student learner of the transformer. Thus, in this paper, we want to build a stand-alone hybrid transformer architecture, which is trained independently without depending on any other networks (not include distillation).

3.1. Model Creation

We begin our creation by applying the DeiT-Ti hyperparameters without distillation to construct a tiny vision transformer encoder. The hyperparameters in the table be-low are optimized by the Facebook AI team.

Model	ViT	Shape	Multi head	No. Layers	No. Params.
DeiT-Ti	N/A	192	3	12	5M
DeiT-B.	ViT-B	768	12	12	86M

Table 1. Vision Transformer Hyperparameters Detail

We modify the DeiT-Ti by adding an EfficientNet-B0 backbone before the embedding layer. Unlike the ViT and DeiT, we cannot split an image into 16×16 fixed-size patches and linearly project them to the embedding layer. Instead, the input sequence can be formed from the output feature maps [7, 7, 1280] of the EfficientNet-B0 body by

simply flattening the spatial dimensions of the feature map and projecting to the transformer dimension. The position embedding is also included as usual to retain positional information.

The transformer encoder in Figure 2 consists of 2 main blocks that are multi-head attention and multi-layer perceptron. On the bottom of each block, there is a layer normalization, which is applied on the neuron for a single instance across all features. The advantage of this normalization is that it is independent of batches. In addition, the skip connections are also included between multi-head attention and multi-layer perceptron for preventing vanishing gradients when training deep neural networks. Inside of the MLP, the Gaussian Error Linear Units (GELU) is accounted as the activation of the first hidden layer followed by the Linear activation of the second hidden layer. GELU is addressed for faster and better loss convergence than ReLU.

3.2. Model Weights Initialization

Taking advantage of transfer learning, we assigned pre-trained weights of EfficientNet-B0 and DeiT-Ti from ImageNet Challenge to our convolution backbone and transformer encoder. However, we initialized the position and patch embedding weights randomly by kaiming uniform for connecting these 2 networks together. With the pretrained weights, we expected our architecture to con- verge faster and yield a global minimum.

4. Experiment

The data we are using is the Alzheimer's Dataset from Kaggle [8]. There are a total 5121 images belonging to four classes. However, the data is very imbalanced as two smallest classes only contain 52 and 717 images each, respectively. For the purpose of training and evaluating the classification models, we split the whole datasetinto the following datasets: train (65%), validation (15%),test (20%), with the same proportion for each class.

We are using Fast.ai platform to train the models. The model used here for image augmentation evaluation is our hybrid architecture EfficientNet ViT-Ti that we have created. The fast.ai platform allows us to quickly search for the best learning rate, use variable learning rate during the multiple-epoch training to reach the best results fast. It also has tools available for image data augmentation and switch to customized loss function such as focal loss function, which focus on training hard negatives. We choose focal loss (FL) over cross-entropy loss (CE), especially for the unbalanced data.

Focal loss is the same as cross entropy loss except easy-to-classify observations are down-weighted in the loss calculation. The strength of down-weighting is proportional to the size of the γ parameter. Put another way, the larger γ the less the easy-to-classify observations contribute to the loss.

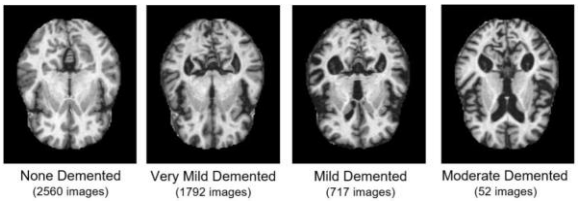


Figure 3. Alzheimer's disease sample training images

For in-batch image augmentation, we first resize the images to 224, which is standard for many existing image classification networks for comparison purposes. Then the rotation, zoom, affine, lighting and horizontal flip are used. We ensure all models are trained in about 30 epochs, which reaches reasonably good models. The data augmentation is done in batches randomly to make image more various during a training process.

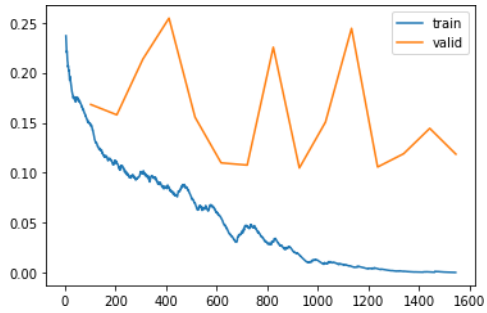
When training the model with original dataset without augmentation, with the comparison that the image augmentation data are used, (see Figure 4), we noticed that the loss for training data is very low after 15 epochs, but the loss for validation data is still high. This indicates the first model might be over fitting the training data too much and the second one on the right is a much better model.

Augment	None		In Batch
Loss func.	CE	Focal	Focal
Bal Acc.	0.813	0.856	**0.984**
Accuracy	0.813	0.835	**0.983**
F1 score	0.803	0.832	**0.983**

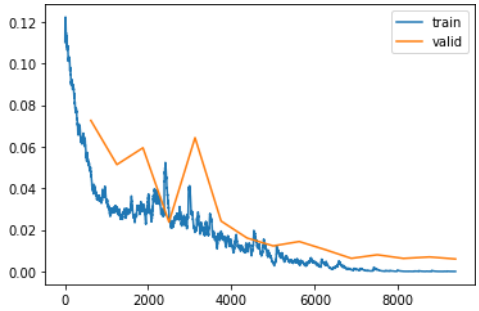
Table 2. Comparison of test evaluation scores of approaches applied for training process

We did some experiments to evaluate which way of augmentations is more helpful and gather the accuracy and other related scores for the different approaches (Table 2). From the table we can see that the data augmentation does have very positive results in training the model. The advantage of using in-batch data augmentation is that it does not require extra disk space to generate more images. Thus, it requires about the same computational power as no data augmentation training.

The other point in this experiment is the loss function. In Table 2, without data augmentation, the balanced accuracy on the test images when using cross-entropy loss (81.28%) are lower than when using focal loss (85.57%).



(a) No Data Augmentation



(b) With Data Augmentation

Figure 4. Learning loss vs training

This encourages us to use a focal loss and data augmentation for training models on imbalanced dataset.

5. Results

We also trained some other state-of-the-art CNN models using focal loss and in-batch augmentation. The balanced accuracy is shown in Figure 5. Since we are handling highly imbalanced datasets, we considered balanced accuracy as one of the evaluation scores for comparison because balanced accuracy still cares about the negative data points unlike the F1 score which most likely focuses more on detecting positives rather than negatives. We compared the various models across Balanced Accuracy and Number of Parameters. From the graph, it could be observed that the EfficientNet-ViT-Ti model (transformer + CNN) has achieved the highest balanced accuracy score (98.85%) compared to the SOTA CNN models like EfficientNet, ResNet, and DenseNet, and transformer models like DeiT-Ti. In addition, the total number of parameters for our hybrid model accounts to around 10 million which is less than the EfficientNet-B3 and all ResNet family.

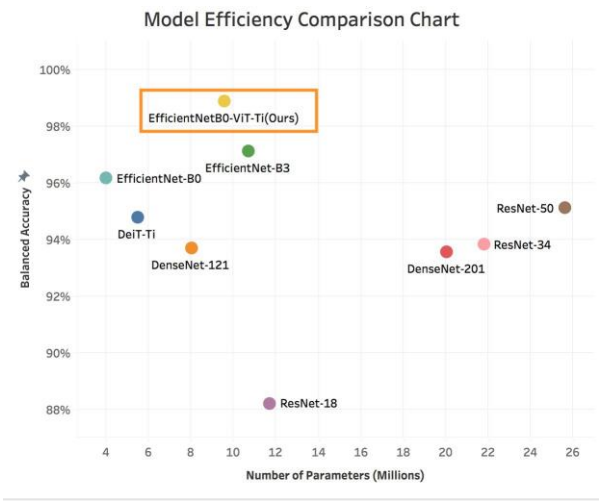


Figure 5. Model Efficiency Comparison Chart

6. Discussion

Our training experiments using in-batch training data augmentation demonstrated that it is capable of achieving comparable high accuracy results, which are much better than simply using the available data for training without augmentation. However, the concern that in-batch training data augmentation may not be able to address the issue of data imbalance relative to manual augmentation, which generate augmented images for minority classes. However, this manual augmentation also takes more time for training due to the fact that dataset is larger. Thus, we decide that the in-batch augmentation is the right choice as it has the benefit of smaller original training data footprint, avoids extra steps in data preparation, and faster training speed.

7. Conclusions and Future Work

In conclusion, Transformers are used to solve problems that are not only limited to NLP, but they could also solve computer vision problems even when getting rid of regular convolutional layers producing SOTA results. Leveraging SOTA CNN backbone with pretrained weights rather than building custom convolutional layers on top of the patch embedding, will improve the model performance of ViT.

The highest testing balanced accuracy has been achieved by our model with limited total network parameters compared to pretrained SOTA CNN models as well as ViT models.

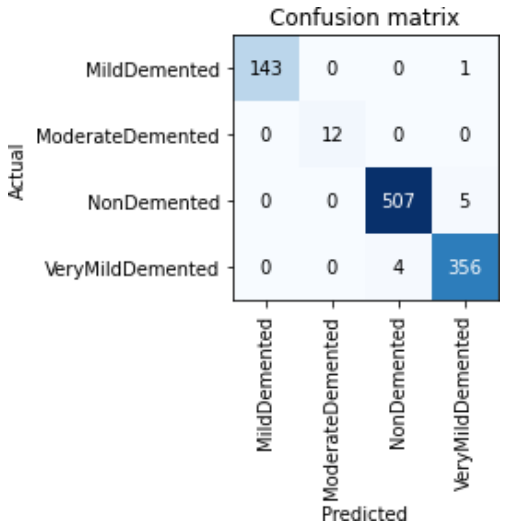


Figure 6. Confusion matrix of EfficientNet-ViT-Ti performing on the testing set with in-batch augmentation technique

One of the limitations of our contribution is that our model was not trained on larger datasets like ImageNet to compare performance with other architectures due to the lack of time and availability of resources.

References

[1] Vaswani, Ashish, et al. ”Attention is all you need.” arXiv preprint arXiv:1706.03762 (2017).

[2] Dosovitskiy, Alexey, et al. ”An image is worth 16x16 words: Transformers for image recognition at scale.” arXiv preprint arXiv:2010.11929 (2020).

[3] Touvron, Hugo, et al. “Training data-efficient image transformers & distillation through attention.” arXiv preprint arXiv:2012.12877 (2021)

[4] Tan, Mingxing, et al. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.” arXiv preprint arXiv:1905.11946 (2020)

[5] ViT Pytorch GitHub - lucidrains/vit-pytorch: Implementation of Vision Transformer, a simple way to achieve SOTA in vision classification with only a single transformer encoder, in Pytorch

[6] Vision Transformers google-research/vision_transformer

[7] FastAi focal loss https://www.kaggle.com/dromosys/fast-ai-v1-focal-loss

[8] Alzheimer’s Dataset (4 Class of Images) https://www.kaggle.com/tourist55/alzheimers-dataset-4-class-of-images