Alzheimer's diagnosis using Attention Based models

Harshit Khandelwal
Indiana University - Bloomington
hkhandel@iu.edu

Arpita Ajit Welling
Indiana University - Bloomington
aawellin@iu.edu

Sai Giridhar Rao Allada
Indiana University - Bloomington
sallada@iu.edu

Purna Chandra Sanapala
Indiana University - Bloomington
psanapal@iu.edu

Abstract

Alzheimer's disease is the most common cause for dementia (loss of memory and other cognitive abilities). Alzheimer's disease is caused when some part of the brain cells is not working well. This is mainly by plaques and tangles. Plaques are deposits of a protein called "beta-amyloid" in spaces between nerve cells and tangles are twisted fibers of protein called "tau" that build inside neurons. These deposits can be seen in MRI scans but in the early stages it is very difficult to see them with naked eye. We propose a comparative study between attention models like Vision Transformer and Compact Convolutional transformer and state-of-the-art models like VGGNet, ResNet, and AlexNet to classify MRI images into 4 categories. Compact Convolutional Transformer gives better testing accuracy than vision transformers which show that convolutions work better than image patching. Out of all the models, VGGNet gives the highest accuracy of 74.59%. Attention models require a lot of data. We had limitations of data for this project and hence, attention models are not performing to their optimum.

1. Introduction

Alzheimer's disease is the most common cause for dementia (loss of memory and other cognitive abilities). This is a progressive disease which worsens over years. In early stages, there will be only mild memory loss but as it worsens, individuals can lose their ability to carry on a conversation and their ability to respond to their environment. In our brain, there are billions of neurons that connect with each other and form groups which helps us in thinking, learning, remembering, hearing etc. Alzheimer's disease is caused when some part of this group is not working well. This is mainly by plaques and tangles. Plaques are deposits of a protein called "beta-amyloid" in spaces between nerve cells and tangles are twisted fibers of protein called "tau" that builds inside neurons. These deposits can be seen in MRI scan but in the early stages it is very difficult to see them with naked eye. Convolutional neural networks, also called ConvNets, were first introduced in the 1980s by Yann LeCun, a postdoctoral computer science researcher. This model was called LeNet-5 [5] but did not gain that much popularity due to the lack of computational power and hardware resources which could accommodate the model and show its true power. It took roughly 3 decades for convolutional neural networks to gain the recognition that it deserved which was brought by Alexnet [4] which won the most prestigious competition called ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Alexnet outperformed all its existing counterparts by a huge margin getting first place and since it's win all the winning models are some variants of CNN's, which showed the benefits of convolution operation on an image when compared to a fully connected network.

Medical field remains one of the most important fields where there is an abundant amount of data that can be processed but remains limited due to the high requirements of the outcomes of the models. General data augmentation techniques that can be used in computer vision cannot be blindly used here due to the fact that some changes in the image might result in something that is not being considered here. For example, the heart is shown to the right of the chest x-ray but when we perform a horizontal flip on it makes the heart appear on the left which represents a serious condition called situs inversus. So we have to keep in mind the subtleties of

the medical domain while performing any kind of operations on a medical image. One of the biggest considerations to be taken while training models is the final output and the importance of the results. As in medical imaging, precision of the results matter a lot and false positives and false negatives cannot be tolerated at all as when the model is deployed people's lives might matter depending on the results that are generated from the model. So the goal of the model for medical imaging should be to reduce the false positives and false negatives, while for general purpose the goal is to increase the accuracy.

Since the introduction of Transformers by researchers at Google they have been the talk of the town and and their performance on NLP related tasks have been so outstanding that they completely overshadowed any other architecture that existed at that time and now. Transformers were based on the mechanism called Attention which made it possible for models to pay dynamic attention to different parts of the sequence. The paper [12] was responsible for this revolutionary change. Attention mimics the Cognitive attention which help in fading out parts which are not important while enhancing the parts which will contribute more towards the final output. For the past few years there has been a lot of debate that has been going about the benefits of CNN and the performance of transformer based models, which has divided people on which stance to take on this matter. Since attention helps the model learn which part to focus on for the particular output, while a CNN model only focuses on the local maps. So that is what we intend to do, answer the question for medical images which is more important and performs better.

So, we want to use modern computational power in the form of deep learning techniques to find these deposits in early stage so that, precautions can be taken before the disease worsens. For some time, attention mechanisms have been a used for transformers, but recent advancements show that it can also be applied successfully for computer vision tasks. We used attention seeking Convolutional Neural Networks for diagnosing Alzheimer's as the attention mechanism is good at isolating a part of the image. We want to compare the results of these with normal CNNs. So, we implemented standard CNNs and compared the results with attention based techniques.

We have used attention based models like Vision Transformers (ViT) and Compact Convolutional Transformers (CCT) on MRI Dataset. We have further used state-of-the-art techniques like ResNet, AlexNet and VGGNet and compared the results and performance between them.

2. Background and Related Work

Since 2017 transformer became the to-go-to architecture for any NLP task but its applications and implementation in vision was close to nothing. All of this changed with the introduction of the vision transformer paper [1]. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. It is shown in the paper that the reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train [12].

Although the transformer based models were good at incorporating temporal information they are still not good at getting the spatial information that has been incorporated into an image. This spatial information can be extracted very efficiently with the use of convolution operations. And that is what was the goal of the CCT paper [2]. The goal of the paper was to show that with the right size and tokenization, transformers can perform head-to-head with state-of-the-art CNNs on small datasets, often with better accuracy and fewer parameters. The CCT model eliminates the requirement for class token and positional embeddings through a novel sequence pooling strategy and the use of convolution/s. It is flexible in terms of model size, and can have as little as 0.28M parameters while achieving good results. The CCT model outperformed many modern CNN based approaches, such as ResNet, and even some NAS-based approaches, such as Proxyless-NAS [2].

Alzheimer's disease is a neurological disorder that is getting more traction these days. This led many people to research on detecting it. As such, there was a lot of work done related to this topic. Especially in the past decade, with an increasing use of deep learning techniques for medical applications, many techniques have been applied to detect Alzheimer's from MRI images [7] [3]. In [10], [9], the authors used a modified LeNet-5 to divide the data into 2 classes. One with Alzheimer's and the other without. They used fMRI dataset for their experiments. They used two convolution layers, each followed by pooling layers and then connected to 2 fully

connected layers and finished it with a final output layer. Their experiments gave them an average of 96.86% accuracy.In [6], attention-based model has been used to classify Glaucoma. They achieved accuracy of 95.3% in their experiments.

The optimized implementation of the vision transformer for alzheimer's disease prediction called OVi-TAD [8] [13]. [8] introduced an optimized vision transformer architecture to predict the aging effect in healthy adults (¿75 years), mild cognitive impairment, and Alzheimer's' brains within the same age group using resting-state functional and anatomical magnetic resonance imaging data. In our project, we want to do the classification using Vit and CCT. We want to compare the results of these attention-based models to normal models like AlexNet, ResNet & VGGNet and want to compare how these methods fare against each other.

3. Methodology

We used latest field of transformers on the MRI data and implemented them in vision field and learn about it's working. We want to study the effect of attention on models learning capabilities to detect the areas in the brain MRI scan which can identify the disease. We used Vision Transformer and Compact Convolutional Transformer on MRI data for detecting the plaques responsible for Alzheimer's and compare the results, performance and attention areas between these two attention-based techniques. We then applied state-of-the-art techniques like ResNet, AlexNet and VGGNet and compared the results and performance between them and also between Attention based techniques and state-of-the-art techniques and derived our conclusions on them.

3.1. Dataset

Getting the proper data source for the problem at hand was one of the biggest hurdle that we had to overcome due to fact that medical images are not available due to variuos issues like privacy issues etc. So after a lot of exploring we started of by understanding the data which compiled by Alzheimer's Disease Neuroimaging Initiative (ADNI) and maintained by University of Southern California.

We went through a lot of the data sources to get a proper source to proceed with our study and research goals. In the end we settled with an open-source dataset on Kaggle which was extracted from ADNI (Alzheimer's Disease Neuroimaging Initiative). This dataset contained 6400 images. The dataset has 4 classes: mildly demented 1, non-demented 2, very mildly demented 3, moderately demented 4.

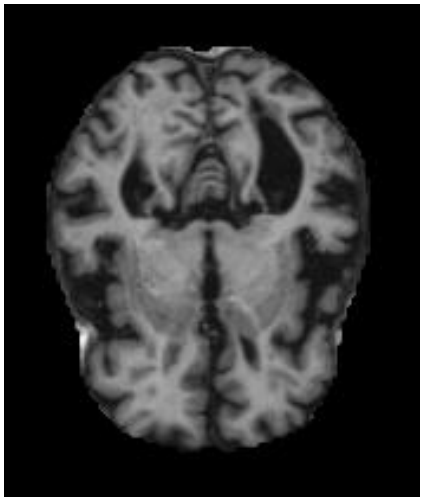

Figure 1. Mildly Demented

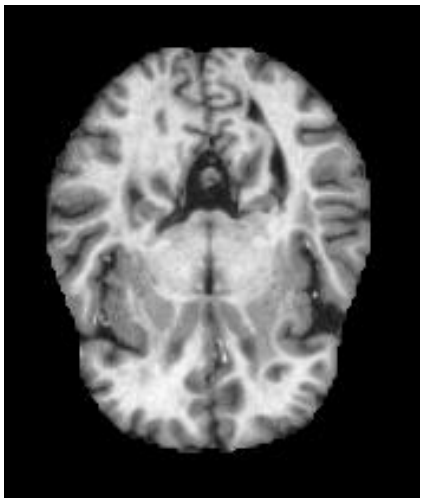

Figure 2. Non-Demented

3.2. Data Preparation

We split that into training, test and validation set. The training and test set were split into 5121 and 1279 images in order to create a disjoint set. Further, for the validation set we used the training set and created a 512 sized dataset which resembles the pattern observed in training set using various data augmentation set. The augmentation techniques used here are 10-pixel translation, random flipping, 30% saturation change and center cropping. The final size of the validation data is 10% of training dataset. Further, due to a highly imbalanced dataset one of the labels "Moderate Demented" was getting overshadowed by other labels. So, we further manufactured this label data using data augmentation tech-

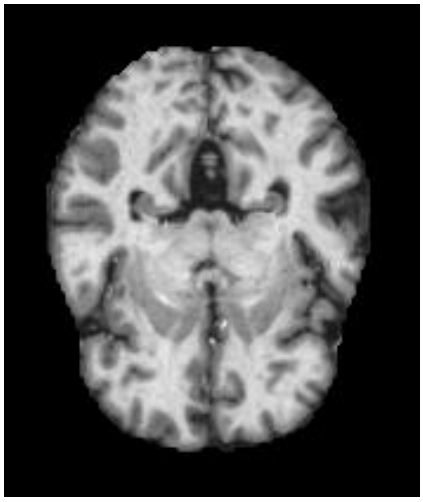

Figure 3. Very Mildly Demented

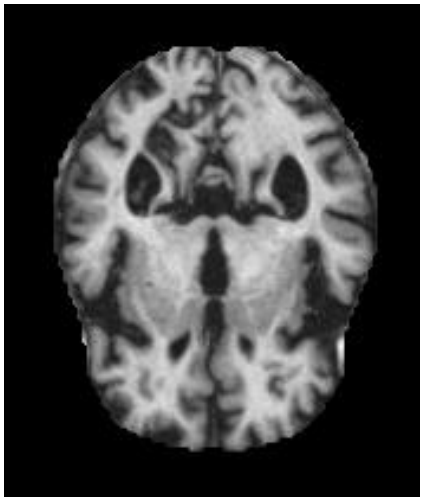

Figure 4. Moderately Demented

niques. The augmentation techniques used here are 20% change in hue, flip left and right and 30-degree rotation about the x axis. With the help of this technique, we were able to get the size of this label from 53 to 266 images.

3.3. Loss Function:

We used two different types of loss functions for our project. One is a custom loss function which helps the model to focus on all the labels and not get overshadowed by only 1 label which is in excess. So, we used the loss function called as focal loss which creates dynamic weights for the classes. This loss function is controlled by 2 parameters called alpha and gamma. These are converted into hyper-parameters which are tuned to fit a particular dataset and model. We used the values as 0.2 for alpha and 4 for gamma. The parameter gamma is responsible for controlling the curve of the loss so more gamma will result in model focusing more on the hard labels. The other loss function that we used is categorical cross entropy which is mainly used with SoftMax activation.

3.4. AlexNet:

The very first CNN which won the Imagenet competition is one of our smallest neural networks that we used here, although at the time it was one of the heaviest and best.

Here, AlexNet has a total of 8 layers. Out of these, the first five are convolutional layers. Some of these layers are followed by max pooling layers. The last three layers are fully connected. For activation, AlexNet uses ReLU activation function.

3.5. VGGNet:

VGGNet [11] is based off AlexNet but with a very small 3x3 receptive field. It also has 1x1 filters which act as a linear transformation of the input. It is followed by a ReLU unit. VGGNet has three fully connected layers. First two layers have 496 channels and th third has 1000 layers. All the hidden layers in VGGNet use ReLU.

3.6. ResNet:

After deep neural networks became popular due to AlexNet VGGNet, it gave rise to a new problem which is that, as the neural network gets deeper, the error rates started increasing. This problem was solved by ResNet by skipping training from a few layers and connecting directly to the output. ResNet50 uses a 50-layer architecture in which shortcut connections are added making the architecture into a residual network.

3.7. Vision Transformer:

The use of transformers for natural language understanding is quite prevalent. Using vision transformers for image data has been going on for a few years since it's introduction in [1]. In a vision transformer, an image is divided into patches. Patches here work like sequence of words in natural language processing. Each patch is then flattened and given to standard transformer block of the encoder-decoder architecture. A transformer block can be seen in figure 5 below.

Each embedded patch is given as input to the transformer. The transformer block has an multi-headed attention module and a multilayer perceptron module. Un-

like a transformer used in NLP, this transformer only has the encoder part. The decoder part is replaced by a MLP block.

The multi headed attention block helps the model to focus on important parts of the image. Using each patch, multihead attention layers calculate attention weight for the given token. Here, token is an image patch. n a multi-headed attention layer, each head learns different attention weights. At the end, all weights are concatenated before giving as input to MLP block. A layer normalization is used before all attention layers.

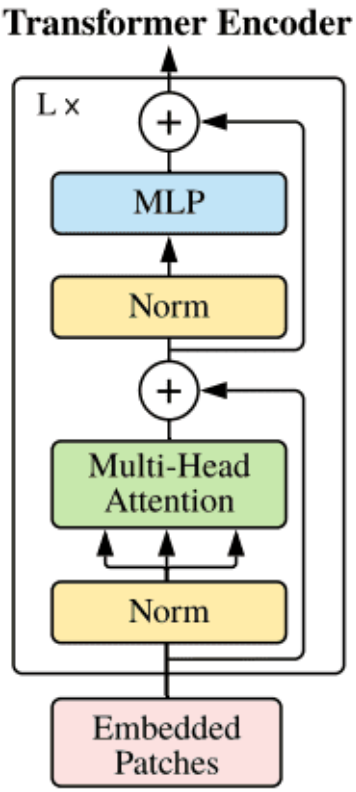

Figure 5. Transformer Block

3.8. Compact Convolutional Transformer:

Compact Convolutional Transformer was introduced in [2]. It is similar to the vision transformer. It also uses a transformer encoder block. The differences between CCT and ViT are that CCT uses convolutions. Instead of image patches, the image goes through a convolution layer. These convolutional patches are reshaped and given as input to the transformer-encoder block. Using convolutions can help inject inductive bias in the model. Although convolutions increase the sequence length because convolution operations are overlapping, it increases the performance by helping to add inductive bias to the model. Convolutions also help to maintain the local spatial information of the input. Apart from convolutions, CCT also uses sequence pooling layer. Sequence pooling essentially means using pooling over the entire sequence of data. This helps to make the model compact. The output of the sequence pooling layer is given to the MLP block. The layout of CCT can be seen in figure 6

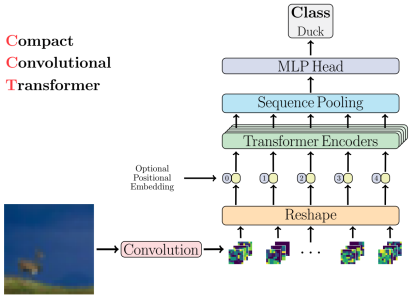

Figure 6. Compact Convolutional Transformer

4. Results

Table 1. Performance of our Deep Learning Models

Model	Training Accuracy	Validation Accuracy
CCT	66.11%	59.19%
ViT	71.12%	55.9%
VGGNet	98.32%	74.59%
VGGNet(pre-trained)	99.79%	69.27%
ResNet	100%	66.37%
AlexNet	84.03%	63.17%

Table 2. Hyper parameters of our Deep Learning Models

Model	No of epochs	Learning Rate	Total Parameters
CCT	130	0.0001	407,000
ViT	250	0.005	~1 Million
VGGNet	50	0.0001	~169 Million
VGGNet(pre-trained)	200	0.0000001	~174 Million
ResNet	200	0.001	~23 Million
AlexNet	200	0.000001	~35 Million

We applied AlexNet on the data and got a validation accuracy of 63.17%. The training accuracy was 84.03% and this converged at 200 epochs with 0.000001 learning rate. The training loss for AlexNet was 0.35 and the validation loss was 1.3373.

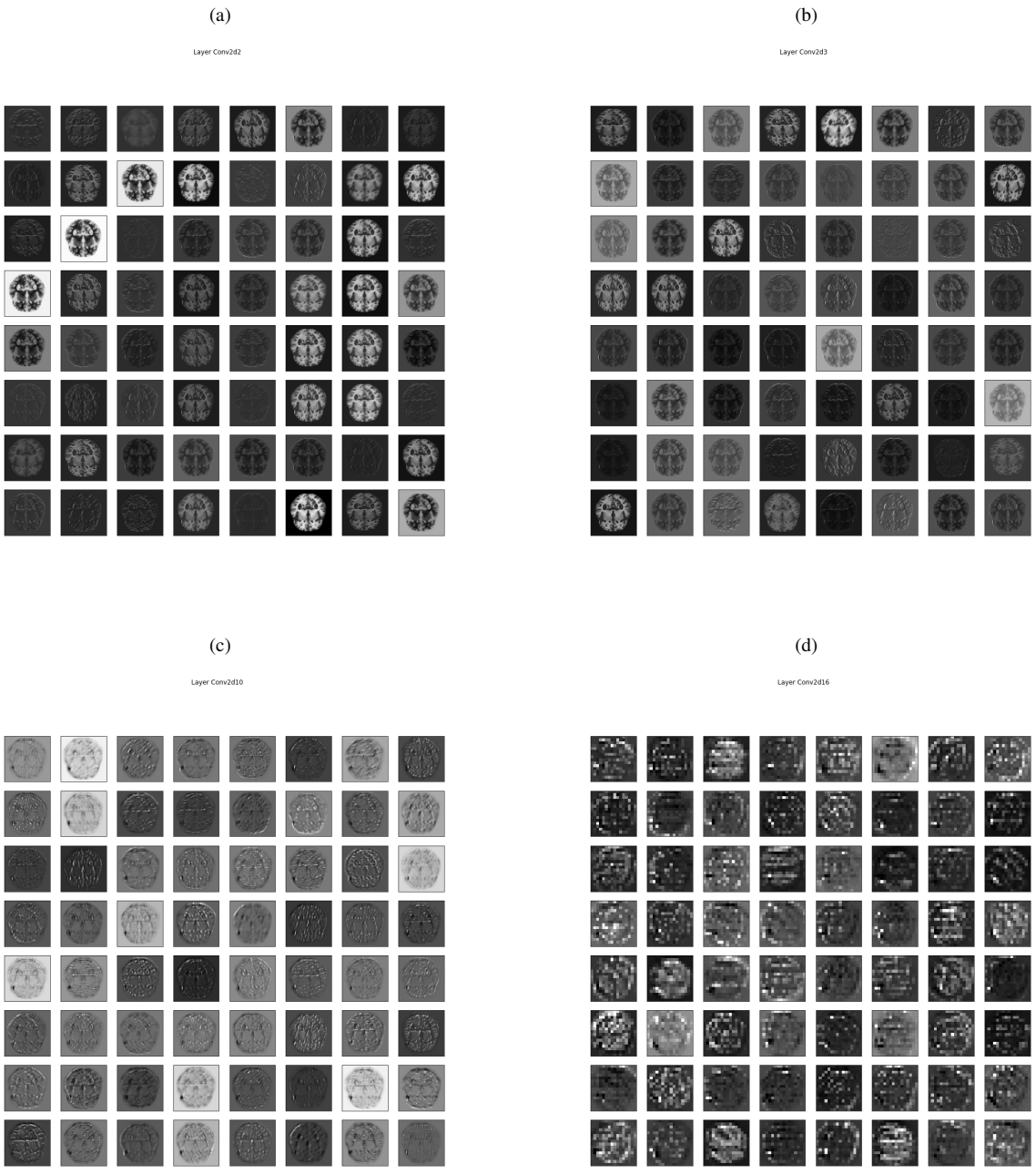

Figure 7. Layerwise knowledge extracted by VGG

For ResNet, the validation accuracy is slightly better at 66.37% with a validation loss of 1.74082. ResNet converged at 200 epochs at 0.001 learning rate. The training accuracy was 100% with a training loss of 4.2386x10e-9.

Pre-trained VGGNet gave us better validation accuracy compared to AlexNet and ResNet. Pre-trained VGGNet gave a validation accuracy of 69.27% and a validation loss of 0.9508 which converged at 200 epochs at a learning rate of 0.0000001. The training accuracy of

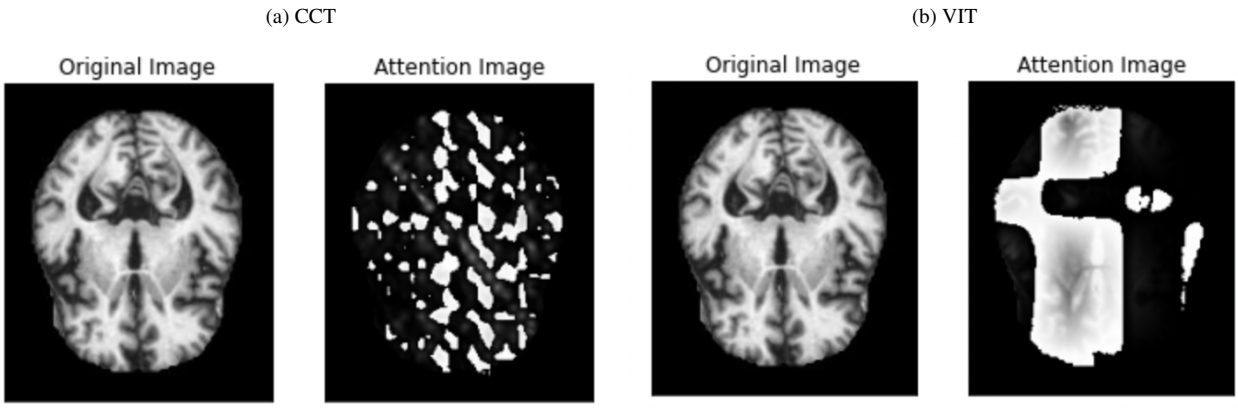

Figure 8. Areas of Interest identified by Attention Based Models

pre-trained VGGNet was 99.79% with a training loss of 0.0670.

VGGNet gave us the best results of all the experiments that we have conducted. The validation accuracy for VGGNet was at 74.59% with a validation loss of 1.5493 with a batch size of 32 which converged at 50 epochs with 0.0001 learning rate. The training accuracy of VGGNet was 98.32% with a training loss of 0.0510.

The first transformer that we have used is Vision Transformer (Vit). Transformers require a lot of data but with just 5121 images to train, we got a validation accuracy of 55.9% with validation loss of 0.9815 which converged at 250 epochs with 32 batch size at a learning rate of 0.005. The training accuracy was 71.12% with a training loss of 0.6373.

Then, we applied CCT which uses convolutions and got a better result compared to Vit. With our very limited dataset of images, we got a validation accuracy of 59.19% with validation loss of 1.0015. With a batch size of 64, the CCT converged at just 130 epochs with a learning rate of 0.0001. The training accuracy for CCT was 66.11% with a training loss of 0.9007.

5. Discussions

As discussed in the previous section, we have experimented with both traditional CNNs and attention based methods. The main purpose of our project was to explore how attention based methods would perform for our particular use case i.e medical image processing. As observed in table 1, the VGG net tends to perform the best with about 74 %. Though this traditional CNN architecture performs the best in terms of accuracy, the features learnt by the model would paint a different picture. As you can see in fig. 7, the features learnt by the model wouldn't necessarily be medically relevant in our opinion. The heatmap literally highlights the entire mri as you could observe which wouldn't be relevant.

We then explored attention based methods like CCT and VIT. As you can see in table 1, these models don't perform as well as the traditional CNN architectures with around 55 % each. This could be attributed to the amount of data we have readily available. Since we're dealing with medical data we could not compensate with too many augmentations. But as you could observe in fig. 8, we see that the VIT slightly attributes it's prediction to the slight grey matter in the upper right corner. We could assume that that's where the attention is based. The same is the case with the CCT, the model kind of highlights the grey matter regions. However, we cannot make a conclusive decision as these visualizations are not the most clear. And the model accuracy is not high enough either. But we do believe that this could be improved with the presence of more data. It could lead to better visual representation of what it's learning as well.

There is a huge limitations to attention based models that they require a lot of data which was evident form the low performing ViT model. But it seems that this problem was also transferred to the CCT model which is in contradiction to what was present/claimed in the CCT paper.

It can be explained with our preliminary results that attention models although low performing with less data pays more attention to the areas which are relevant to the problem at hand. For example, here the model learns to predict a class based on the plaque accumulation while paying attention to it using transformers.

6. Conclusion

With all experimentation we can conclude that attention based models are better at learning the subtleties of problem as compared to conventional convolutional neural networks like VGG, Alexnet etc. Although they have a limitation of the amount of data that is available to them which restricts the learning process for the model.

If the models are able to learn to recognize the areas in the brain which are responsible for various neurological diseases then diagnosis and treatment of these diseases would become much better than what already exists while it will also help in reducing the false positives and the false negatives that a model might predict for an input.

The future for the attention based models in vision tasks are still developing but if it keeps going in the direction that we were able to conclude from these experiments then the future is promising. We could also work on some pre-hoc attention based models where the model knows what to look before hand. For the future work we can collaborate with the medical professional for the purpose of development.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," *arXiv preprint arXiv:2104.05704*, 2021.

[3] J. Islam and Y. Zhang, "A novel deep learning based multi-class classification method for alzheimer's disease detection using brain mri data," in *International conference on brain informatics*, Springer, 2017, pp. 213–222.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[6] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu, "Attention based glaucoma detection: A large-scale database and cnn model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 571–10 580.

[7] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, "Early diagnosis of alzheimer's disease with deep learning," in *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, IEEE, 2014, pp. 1015–1018.

[8] S. Sarraf, A. Sarraf, D. D. DeSouza, J. A. Anderson, M. Kabia, A. D. N. I. ADNI, *et al.*, "Ovitad: Optimized vision transformer to predict various stages of alzheimer's disease using resting-state fmri and structural mri data," *bioRxiv*, 2021.

[9] S. Sarraf and G. Tofighi, "Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks," *arXiv preprint arXiv:1603.08631*, 2016.

[10] ——, "Deep learning-based pipeline to recognize alzheimer's disease using fmri data," in *2016 future technologies conference (FTC)*, IEEE, 2016, pp. 816–820.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[13] X. Xing, G. Liang, Y. Zhang, S. Khanal, A.-L. Lin, and N. Jacobs, "Advit: Vision transformer on multi-modality pet images for alzheimer disease diagnosis," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2022, pp. 1–4.