**Background:**

There are many image classifiers that have come to prominence in the past few years, but some of the best performing models have used Convolutional neural networks (CNNs). CNNs take image pixels as input and through convolutions layers extract and learn features from the image which can be used to classify them. There are three main layers to highlight in the architecture of a CNNs, the first is the convolution layer which multiplies to the input pixel data to produce a output feature map which is done by multiplying the filter one pixel at a time then sliding horizontally and vertically the next layer is the ReLu layer in which the Rectified Linear Unit transformation this is done to the convoluted layer to introduce non-linearity, and the last layer is the pooling layer which reduces the dimensionality of the feature map while maintaining the extracted feature information.

Vision Transformers are derived from transformer models commonly used in natural language processing applications, where a ViT takes an image broken into a series of patches or visual tokens, a traditional transformer takes in a sentence broken into a series of word tokens. It also include the position of each patch of the image as an input to the network. One of the essential elements of ViT is the concept of attention, specifically self-attention which put simply is a method to aggregate info from the entire series of input patches.

The architecture of a ViT is as follows: A image is broken into patches, the patches are then flattened and used to create lower dimension linear embeddings with corresponding positional embeddings that are then fed into the transformer encoder that is pretrained on a large dataset. The hyperoperators in this architecture are network depth, optimizer choice and dataset specific choices, all of which have been found difficult to tune to obtain the most optimal performance and tends to be much more data hungry than CNNs.

In terms of performance results of [1] when the ViTs were trained on a mid-size dataset they underperformed CNNs which the authors contribute to inductive biases in CNNs such as

translational equivariance, however when pretrained on a large dataset the model can outperform the CNN based model.

**System Project:**

For this project I will build upon my topic for my research paper and conduct a comparison of a vision transformers (ViT)and a convolutional neural network (CNN) for image classification. The aim of this comparison is to explore the performance of the two models and to determine the benefits and drawback of each . For this project I aim to apply the 6Ds framework of creating AI enabled systems which consist of Decomposition, Domain Expertise, Data, Design, Diagnosis and Deployment. I will discuss how each of the first four elements will be considered.

Starting with Decomposition this project is intended to achieve new insights into image classification by comparing ViTs and CNNs. One insight to be gained is which framework has better explainability along with lower computational resources. To compare explainability ViT are well known for their attention maps and CNNs have some level of explainability when looking at how an image progress through the various layers in the network.

The second is Domain expertise which for this dataset is being familiar with the 5 types of plants used for classification and pictured in the figure below.

The third D is Data, and for this project I used the cassava-leaf-disease dataset.

**Data:**

In this endeavor the use of a simple CNN is compared to a ViT on a Cassava leaf dataset to determine which network is better at predicting the following 5 categories of Plant Disease from a dataset of 17000 images (from the cassava-leaf-disease dataset):

Chart, bar chart

Description automatically generated

"0": Bacterial Blight

"1": Brown Streak Disease

"2": Green Mottle

"3": Mosaic Disease

"4": Healthy

The five types of disease can be visualized in the 5 photos below:

  A close-up of some leaves

Description automatically generated with medium confidence

2

1

0

 

4

3

**Design:**

Diagram

Description automatically generatedUsing tensorboard the CNN architecture can be visualized in the figure 1, where the input image is passed through convolutional layers to extract features and then passed through a max pooling layer before final being passed through a multi-layer perceptron.

Similarly, the ViT architecture used can be visualized in the figure below, where an image is broken into patches or “vision tokens” along with its positional embeddings which are then passed to the encoder to be trained:

Diagram

Description automatically generated with low confidence

**Results:**

After both models were trained, their classification performance was compared by looking at accuracy and loss. CNN results figure below shows a plateu in performance at around 17 epochs at which accuracy was ~88% while while the ViT plateu was around 15 epochs with an accuracy ~88%. The results are not conclusive in determining which network has better performance in terms of better classification accuracy however from a computations resources perspective the ViT trained 10 epochs in around 3 hours while the CNN took around 4 hours for 10 epochs.

Chart, line chart

Description automatically generated

ViT results:

Chart, line chart

Description automatically generated Chart, line chart, scatter chart

Description automatically generated

The ViT self-attention mechanism can be used to generate an attention map to highlight what connections were drawn on the data from patch to patch, an example of this can be seen in the image below:

A picture containing graphical user interface

Description automatically generated

The benefit of having an attention map is it allows the user to intuitively understand the most significant features to the network, however this relies on having some positional information in the image for the network to learn, for imagery from a long-range infrared sensor the ViT may not be able to extract any relevant positional features.

**Conclusion:**

The architecture of vision transformers clearly shows its potential as a top performing classifier however the comparison done here does not clearly demonstrate its ability to outperform a CNN, a further study should be conducted on a variety of large datasets of various industries to better answer the question which classifier has the superior classification accuracy.

References:

[0] Vaswani, Ashish, et al. “Attention Is All You Need.” *ArXiv.org*, 6 Dec. 2017, https://arxiv.org/abs/1706.03762?context=cs.

[1] Dosovitskiy, Alexey, et al. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.” *ArXiv.org*, 3 June 2021, <https://arxiv.org/abs/2010.11929>.

[2] Ma, Hailong, et al. “MoCoViT: Mobile Convolutional Vision Transformer.” *ArXiv.org*, 26 May 2022, https://arxiv.org/abs/2205.12635.

[3] Liu, Yahui, et al. “Breaking the Chain of Gradient Leakage in Vision Transformers.” *ArXiv.org*, 25 May 2022, https://arxiv.org/abs/2205.12551.

[4] Liu, Yahui, et al. “Breaking the Chain of Gradient Leakage in Vision Transformers.” *ArXiv.org*, 25 May 2022, https://arxiv.org/abs/2205.12551.

[5] Cuenat, S., & Couturier, R. (2021). Convolutional Neural Network (CNN) vs Visual Transformer (ViT) for Digital Holography. *arXiv preprint arXiv:2108.09147*.

[6] Xie, Y., Yan, J., Kang, L., Guo, Y., Zhang, J., & Luan, X. (2022). FCT: fusing CNN and transformer for scene classification. *International Journal of Multimedia Information Retrieval*, 1-8.

[7] Wang, W., Tang, C., Wang, X., & Zheng, B. (2022). A ViT-based multiscale feature fusion approach for remote sensing image segmentation. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1-5.

[8] Mehta, S., & Rastegari, M. (2021). Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*.

[9] Dataset 1: <https://www.kaggle.com/competitions/cassava-leaf-disease-classification>

* <https://www.kaggle.com/code/piantic/vision-transformer-vit-visualize-attention-map>
* https://www.kaggle.com/code/maksymshkliarevskyi/cassava-leaf-disease-best-keras-cnn/notebook

[9] Dataset 2: <https://www.kaggle.com/c/aerial-cactus-identification>

* <https://github.com/asyml/vision-transformer-pytorch/blob/main/src/model.py>
* <https://www.kaggle.com/code/masonblier/aerial-cactus-simple-cnn>