Kent State University

MIS-64061: Advanced Machine Learning

FALL 2022

Final Project Report

Skin Cancer MNIST Image Classification using CNN & VIT - TensorFlow

By:

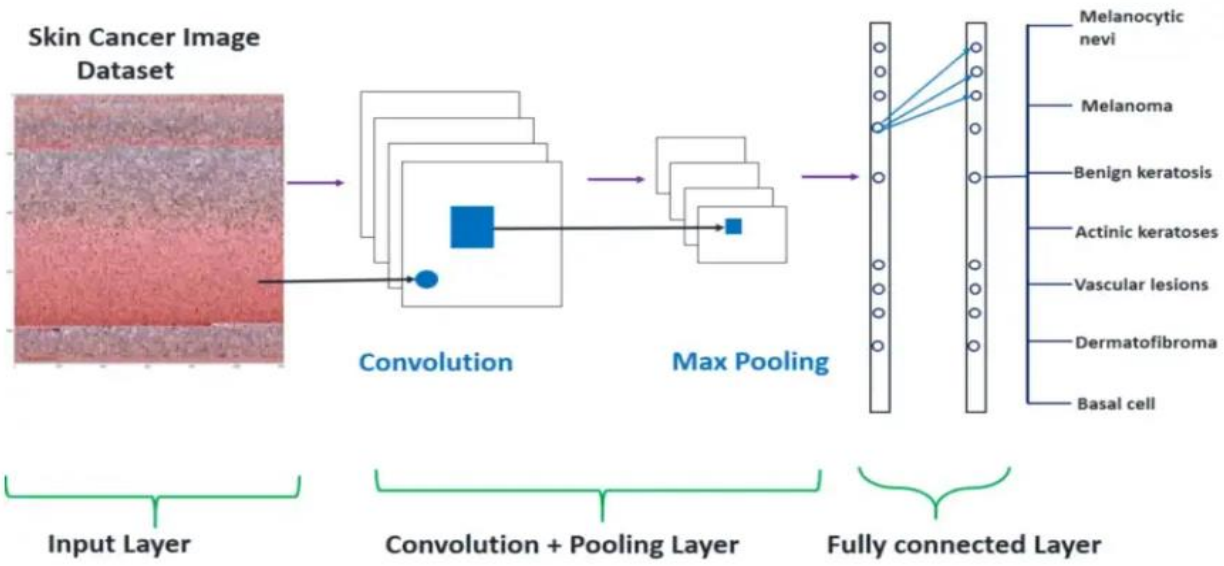
Manasa Chelukala

Table of Contents

Introduction:

A new neural network architecture called Transformer was introduced by Google AI in 2017, and it claimed to perform better than the leading approaches. In just four years, it has revolutionized the field of natural language processing. Transformers bring new capabilities to Machine Learning along with neural architectures. In 2021, 'An Image is Worth 16X16 Words' was presented in International Conference for Representation Learning (ICLR) by Alex Dosovitskiy et. al. It demonstrated, for the first time, that Transformers could be implemented for Computer Vision tasks and outperformed CNN in image classification. An explanation of the Vision Transformer architecture will be provided, as well as a detailed comparison between CNN and VIT models on the Skin Cancer MNIST (HAM10000) dataset.

Implementation of CNN Model:



Source: [Skin Cancer Detection using Convolution Neural Network(CNN)]

Input -> [[Conv2D -> relu] *2 -> MaxPool2D -> Dropout]*2 -> Flatten -> Dense -> Dropout -> Output

In Keras Sequential API, one layer is added at a time, beginning from the input.

1. The first layer is the convolutional layer (Conv2D). It is similar to a set of learnable filters. For the two first conv2D layers, I chose 32 filters, and for the two last, 64 filters. Using the kernel filter, each filter transforms a part of the image (defined by the kernel size). The kernel filter matrix is applied to the entire image. An image can be transformed by filters. In these transformed images (feature maps), CNN can isolate features that are useful everywhere.

2. Next is the pooling layer (MaxPool2D) in CNN. The purpose of this layer is to downsample data. It picks the maximal value based on the two neighboring pixels. As a result, computational costs are reduced, and overfitting is reduced to some extent as well. When the pooling dimension is high, more downsampling is important. We have to choose the size of the pooling each time. As a result of combining convolutional and pooling layers, CNN is able to learn more global features of an image while combining local features.

3. A dropout method involves ignoring a proportion of nodes in the layer (setting their weights to zero) for each training sample. Using this method, a portion of the network is dropped randomly and the network is forced to learn features in a distributed fashion. Furthermore, this technique improves generalization and reduces overfitting.

4. Relu is the activation function of the rectifier (max(0,x)). To add nonlinearity to the network, the rectifier activation function is used.

5. In the Flatten layer, the final feature maps are converted into one single 1D vector. The flattening step allows you to use fully connected layers after convolutional/maxpool layers. As a result, all the local features found in previous convolutional layers are combined in this layer.

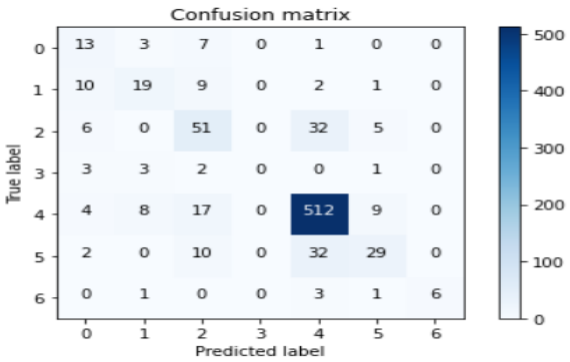
6. At the end, I used two fully-connected (Dense) layers, which are artificial neural networks (ANNs). The final layer (Dense(10,activation="softmax")) outputs a probability distribution for each class.

The full implementation of the CNN on Image classification is found here:

https://github.com/mcheluka/64061_mcheluka/blob/main/Final%20Project/Skin_Cancer_mnist_

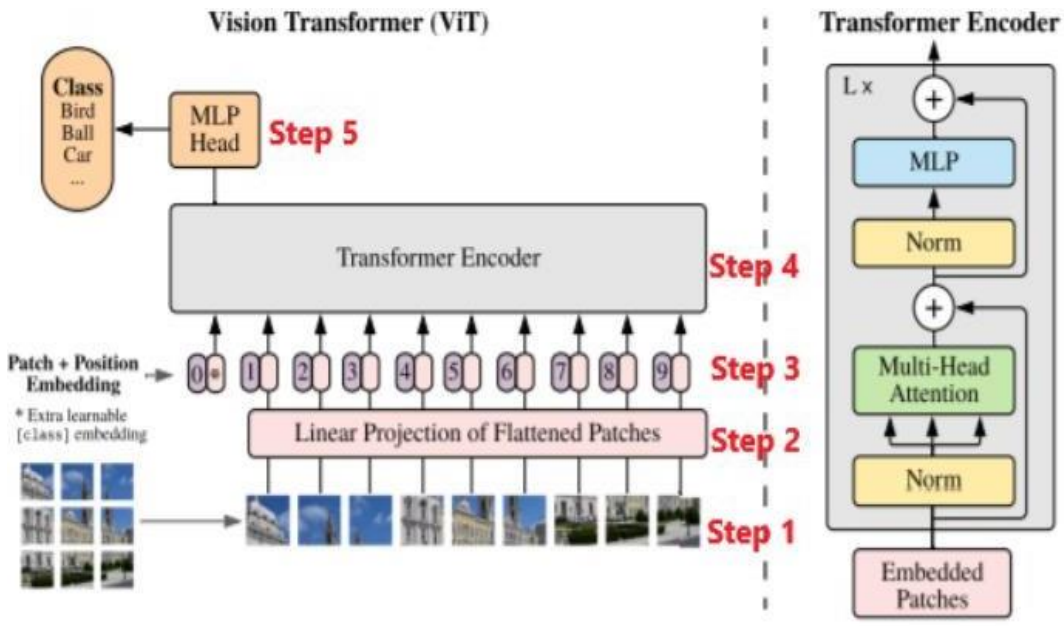
HAM_10000_CNN.ipynb

Output of the CNN Model(Confusion Matrix):



-- From the above confusion matrix, we can say that the cancer type of number - 4 which is Melanocytic nevi has the highest number of correct prediction i.e., 512.

Vision Transformer Architecture:



Source: AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Below are the high-level steps to implement the Vision Transformer in TensorFlow.

Step 1: The image should be split into patches of fixed size.

Step 2: Using a fully connected layer, flatten the 2D image patches to 1D patch embedding.

Step 3: In contrast to other architectures, transformers have no idea of sequence or position. In order to retain positional information, we embed positional embeddings into patches.

Step 4: A Transformer Encoder consists of alternating layers of self-attention and MLP blocks. Layer normalization (LN) is applied before the self-attention and MLP blocks. To avoid vanishing gradient problems, residual connections are applied after every block. Using this block, helps to learn about local and global dependencies in an image.

Step 5: A classification head is implemented using MLP with one hidden layer at pre-training time and a single linear layer at fine-tuning time.

Implementation of VIT:

In the above section, VIT architecture has been summarized. For the implementation, I used the Skin Cancer MNIST (HAM10000) dataset.

Dataset: HAM-10000- skin cancer MNIST

Now-a-days, Skin cancer has became the most common human disease, which is primarily diagnosed visually, starting with an initial clinical screening and followed potentially by dermoscopic analysis, a biopsy & histopathological examinations.
usually, the skin cancer occurs due to the mutations caused in the DNA of the cells and these cells grow out-of-control and form a mass of cancer cells.

HAM10000 is a dataset of images of Skin Cancer—consisting of a training set of 10,000 examples. Each example is a 450 X 600 X 3 grayscale image, associated with a label from SEVEN classes.

I have used the vision transformers with B-16 and B-32.

The full implementation of the VIT on Image classification using TensorFlow code is found

here:

https://github.com/mcheluka/64061_mcheluka/blob/main/Final%20Project/SKIN_CANCER_MNIST_HAM10000_VIT.ipynb

As the main objective of this project is to provide the comparison between CNN and Transformers for the Skin Cancer (HAM-10000) MNIST dataset.

Comparison between CNN and VIT:

CNN	VIT
Feature Map	Attention Map
Pixels	Patches
Batch Normalization	Layer Normalization
High Inductive Bias	Weaker Inductive Bias
2D neighborhood - Kernel	Spatial Relationships – training from scratch
It is not sensitive to the size of the dataset	It is sensitive to the size of the dataset

I would like to highlight a few topics that require a detailed comparison.

- **Receptive field range:**

Since CNNs use kernels of 3x3 or 5x5, each layer can only have a corresponding field of view. Likewise, the field of view expands as it propagates through the layers, but the expansion increases linearly with depth.

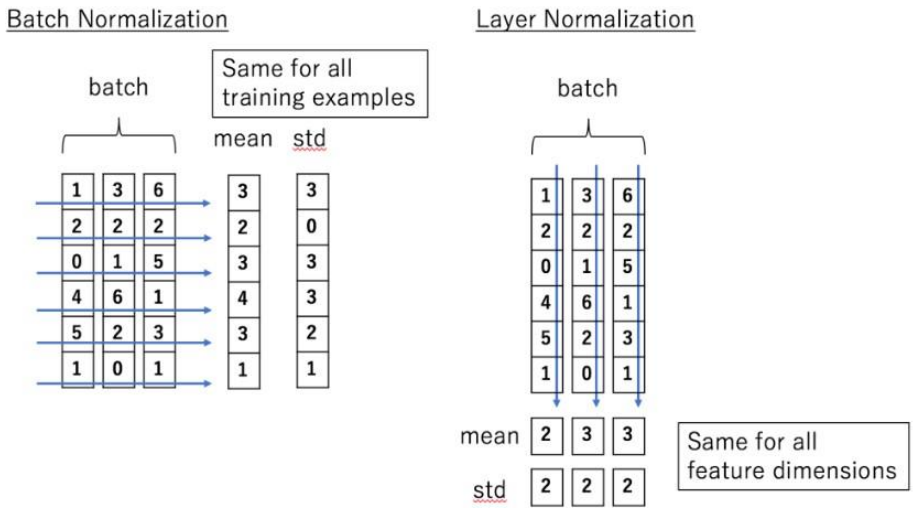
Transformer, however, uses Self-Attention, which allows the network to see the entire image from the beginning. It is possible to learn global features from the beginning since each patch is treated as a token and all of them are correlated and calculated.

- **Impact on size of the dataset and the inductive bias:**

The main difference is — In CNN, the kernels help us to learn/understand about the 2D neighborhood structure; whereas in transformers, no 2D structure is used, and the positional embeddings do not carry information about the 2D positions of the patches when initialized, and all spatial relationships have to be learned from scratch. Since it must learn the relations from scratch, it requires a lot of data to perform better than CNN. As a result, the size of the dataset is important when it comes to the vision transformer.

- **Batch Normalization and Layer Normalization:**

To improve the stability and performance of the model, we used Batch Normalization in CNN, which is done by scaling with the mean and standard deviation for each training example (Batch), while Layer normalization, which is used in transformers, computes standard deviation and mean for each feature separately.



Source: Batch Normalization and Layer Normalization

In batch normalization, a less known issue is how difficult it is to parallelize batch-normalized models. There is an additional need for synchronization across devices due to the dependent nature of elements. While most vision models do not have this problem, since they use a limited number of devices, Transformers suffer from this issue due to their quadratic complexity. Therefore, layer normalization provide some normalization while avoiding batch-wise dependence.

- **Effect of texture perturbations:**

In comparison to CNN, the Transformer model (ViT) is relatively robust to texture perturbations [1].

Results:

- **Accuracy & Loss:**

As part of the implementation process, all models were evaluated.All the models are trained for different epochs.

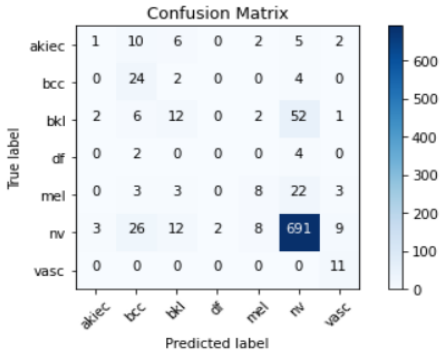
	VALIDATION		TEST	
MODEL	**LOSS**	**ACCURACY**	**LOSS**	**ACCURACY**
CNN	0.6838	0.7855	0.695	0.7638
VIT- B-32	0.1632	0.7803	0.18	0.7964
VIT - B-16	0.2202	0.8497	0.2303	0.85

Vision transformer models have been trained from scratch incorporating data augmentation and drop out techniques, and CNN models have been trained without regularization techniques.

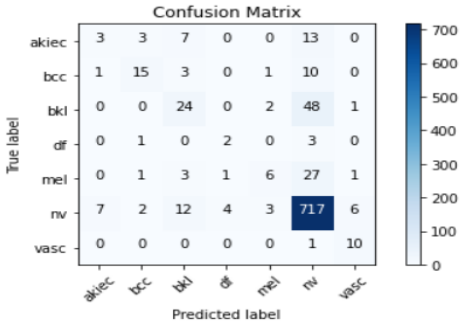
Based on the above results, we can see that the VIT models are more accurate than CNNs. Nevertheless, we can see a slight improvement in accuracy over CNN baseline model performance. As a result, the input dataset was medium in size. A large input would have resulted in exceptional performance.

- **Impact of Patch size on the performance:**

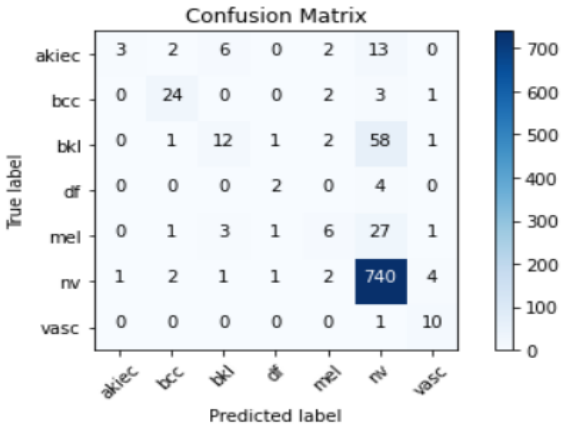
In the VIT architecture, patch size is a significant hyperparameter. I evaluated the model performance by tuning patch size and found that accuracy increases as patch size decreases, as can be seen in the chart below.



Confusion Matrix- vit-B-32



Confusion Matrix- vit-B-16



Fitted_model Confusion Matrix- vit-B-16

- Using the patch size of 32, we can see that the cancer type which is nv i.e. Melanocytic nevi has highest number of correct prediction i.e., 691.
- Using the patch size of 16,we can see that the cancer type which is nv i.e. Melanocytic nevi has highest number of correct prediction i.e., 717.
- So,by using the vit – B-16 with the patch size of 16 and using the optimizer sgd and activation function Gelu the model predicted well the type of cancer so I have fitted this model and evaluated its performance on the classification of skin cancer.
- Hence, we can see that the cancer type which is nv i.e. Melanocytic nevi has highest number of corr ect prediction i.e., 740.

Conclusion:

Although Vision Transformer has achieved good accuracy compared to CNN, it does have a limitation. As a result of weak inductive bias, the model requires large datasets to achieve good accuracy. To overcome this, various improvement methods have been proposed.

As one approach, CNNs can be used to reduce data requirements: DeiT [3] uses a knowledge distillation framework where CNNs are used as the teacher model, and knowledge is fed to the transformer model. As a result, transformers can outperform CNN regardless of the size of the dataset.

Lastly, I would like to conclude that Transformers have replaced RNNs in NLP completely. Currently, they are attempting to replace Convolutional Neural Networks (CNNs). The model might make CNNs extinct in the future, but not yet. The model still faces challenges when it comes to smaller datasets and other computer vision tasks, such as image segmentation and detection.

Contribution:

In the class, I have learnt how the CNN and vision transformers are implemented. So, here I have used the skin cancer image classification dataset and implemented using CNN and Vit-B-32,Vit-B-16 and have used different optimizers like adam,SGD and activation functions like sigmoid,GELU,RELU in determining the time of skin cancer from all the Seven type of skin cancer and have made the comparison between CNN and Vision transformers and analyzed how well there performances are in predictions.

References:

1. Shikhar Tuli, Ishita Dasgupta, Erin Grant, Thomas L. Griffiths. Are Convolutional Neural Networks or Transformers more like human vision? arXiv(2021)
2. https://towardsdatascience.com/understand-and-implement-vision-transformer-with-tensorflow-2-0-f5435769093
3. https://www.researchgate.net/publication/347797071_Training_data-efficient_image_transformers_distillation_through_attention
4. https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000
5. Recent Developments and Views on Computer Vision x Transformer | by Akihiro FUJII | Towards Data Science
6. https://www.sciencedirect.com/science/article/pii/S0010482522006746
7. https://keras.io/examples/vision/image_classification_with_vision_transformer/#build-the-vit-model
8. https://www.sciencedirect.com/science/article/pii/S0010482522006746
9. https://medium.com/@sid321axn/skin-lesion-classification-using-deep-cnns-a-step-wise-approach-347d47868196