DoubleViT: Pushing Transformers Towards the End Because of Convolutions

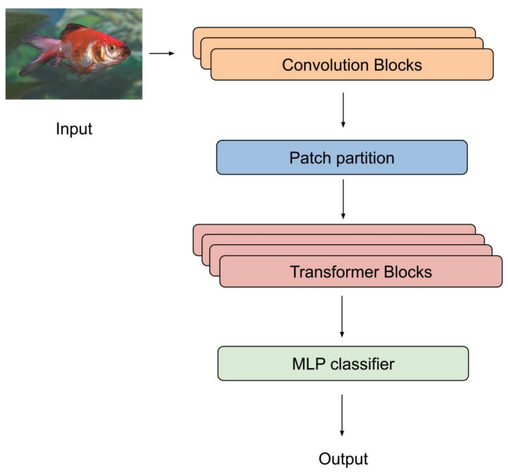
Mahendran Narayanan
mahendranNNM@gmail.com

https://github.com/mahendran-narayanan/Double-ViT

Introduction

DoubleViT

- Models transform by integrating methods, modules or layers.
- Model size is increasing constantly
- But AI models are actively used in mobile and edge devices, prompting small network development.
- In proposed method, Attention mechanism is trailing convolutional layers.



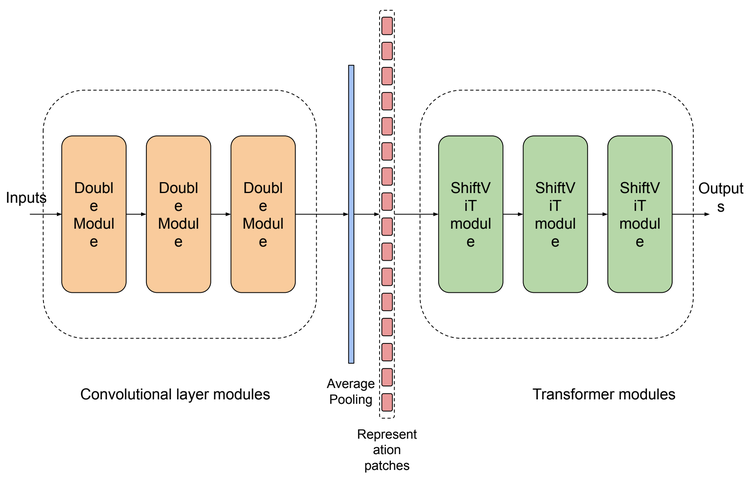
Contributions

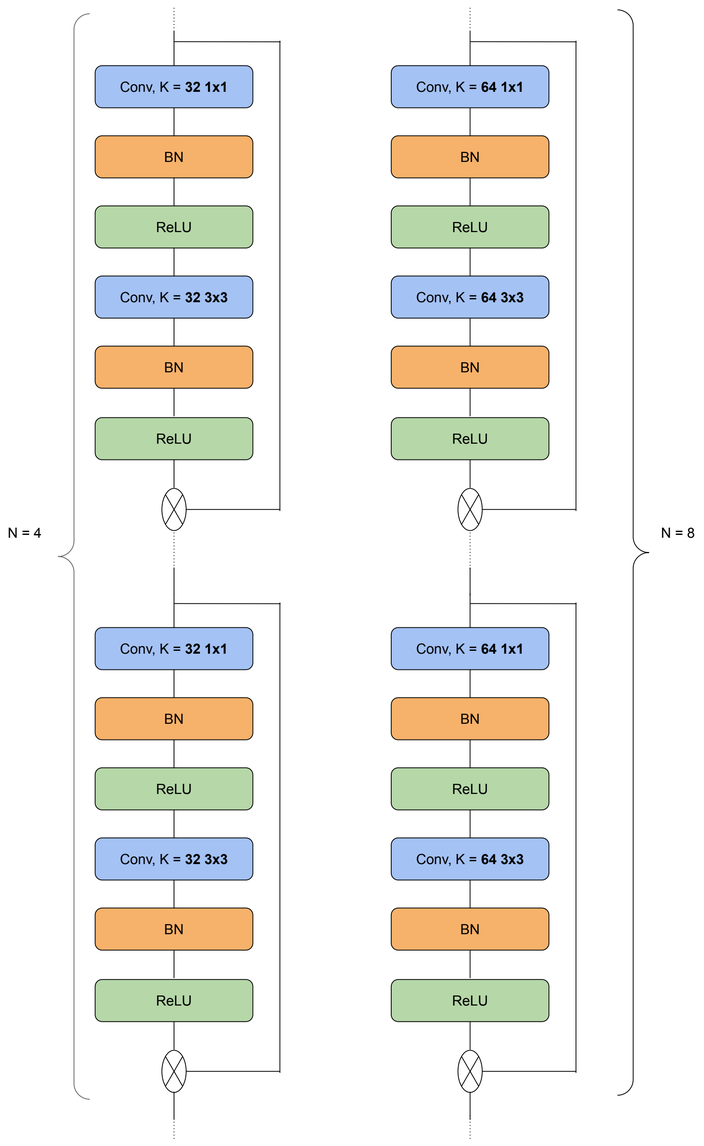
- Smallest acceptable dimensions by model is 8x8. In ViT, it is 72x72. In ShiftViT, it is 48x48
- Model consists of two modules: Double module, Shift module.
- Double module contains convolutional layers depending on input dimensions.
- Shift module consists of Transformer layers trained from the learned representations of CNNs.
- Effective substitute for the dense layers typically located at the bottom of the network.
- Transformer part learns global representations at a reduced cost, directly influencing the overall model size.

Method

- Double module, comprises multiple stages, each characterized by doubling the kernels while simultaneously reducing the size.
- At the end of each module, the size is reduced by half.
- Additionally, the number of instances of each module is doubled at each depth.

Module depth	Kernel Size	Kernels	Module repetition Count
1	3 x 3	32	4
2	3 x 3	64	8





Experiments

- DoubleViT module has a model size of ~4.85 MB for ImageNet with 4 transformer modules.
- Proposed model is ideally suited for input sizes ranging from 8x8 to 128x128 for those with limited computational resources.
- For those with more robust computational capabilities, the model can accommodate larger input sizes.

CIFAR-100 dataset

Model	Accuracy (%)	Input shape	Model size
ViT	51.42	64	62.26 MB
ShiftVit	47.75	48	44.86 MB
DoubleViT	**50.18**	32	**6.45 MB**
DoubleViT	**56.36**	**48**	**6.45 MB**
DoubleViT	**58.26**	**64**	**6.45 MB**

CIFAR-10 dataset

Model	Accuracy (%)	Input shape	Model size
ViT	81.22	64	61.91 MB
ShiftViT	77.91	48	44.60 MB
DoubleViT	**79.30**	**32**	6.39 MB
DoubleViT	**88.95**	**48**	6.39 MB

ImageNet dataset

Model	Accuracy (%)	Input shape
ViT	22.33	64
ShiftViT	23.44	48
DoubleViT	**24.84**	**32**

- Model performs exceptionally well with smaller input sizes up to 128x128.
- However, for larger input sizes, the model's depth doubles after each module, leading to a significant increase in model size and computational demands.
- The model convolutional layer shape and depth have been decided in accordance with the model input. For better results, depth can be increased, supported by computational capabilities.
- The proposed architecture adopts simpler kernels, with each increase in depth being a power of 2.
- The connection between convolutional layers and transformers is achieved by directly merging them, allowing the layer to learn and send representations to shift transformer layers, unlike traditional average pooling layers in CNNs.

Takehome messages

The proposed model presents a combination of convolutional layers and transformer modules, strategically integrated to enhance performance in computer vision tasks. More experiments in this research area in future, can lead to smaller size models with on par performance.