Vision Transformer Survey on Small Dataset

Chengjun Guo Craig Duncan

Abstract

With the growing success of Vision Transformers(ViT) on the task of Image Classification we present a survey of some of the state-of-the-art ViT architectures. We review the papers behind CrossVit, CaiT, MobileVit, and SimpleVit along with their proposed changes to improve ViT's in a variety of scenarios. There is a short discussion of the impact of the different model architectures on their theoretical performance. Finally we evaluate each model on the MNIST imageset of handwritten numbers and compare a number of metrics including training time and final accuracy.

1. Introduction

Classification tasks are an integral part of the machine learning landscape as a well defined supervised task. As a supervised task this means it has a discrete output expected for every input allowing for concrete evaluation of a models results. The past decade has seen an increased focus on the image classification subset of supervised learning. Images and video are one of the most information rich methods of sampling the world. Developing strong image classification models serves as a stepping stone for models that can see and understand the world we live in.

Neural Networks comprised primarily of Dense layers while effective in some early image classification work were quickly replaced by ones using convolution and pooling layers. Convolutional Neural Networks(CNN's) which utilize convolution computations as one of their layers have excelled in the Image classification space for a number of years with many CNN's posting State of the art results. One benefit of CNN's is that the convolutional layer repetition of the same filter across the image provides invariance to a number of transforms in the image. This means that a tilted image of a cat or one with the cat off center are still seen as a cat.

Recent advancements in Neural Network architectures have resulted in new a cutting edge model referred to as a Transformer. This new model architecture has produced cutting edge results in the realm of natural language processing. These top results have shown the potential of this new architecture and prompted attempts to leverage the architecture for other machine learning tasks including Image classification. The main feature of the Transformer architecture is its self-attention mechanism which allows the model to learn based on the relationship between different parts of the input. This relation based learning has been critical for understanding the relation between parts of a sentence such as subject and verb while focusing less on unimportant filler words.

Transformers relation based learning can be leveraged in image classification. Because the transformer learns using its self-attention mechanism it can be trained to be transform invariant in much the same way to a CNN[3]. Transformers large flexibility in learning has resulted in competitive performance in Image Classification tasks even when compared to top CNN's. This ability to perform as well in classification tasks compared to CNN's is important because training large highend CNN's can be very resource intensive. Vision Transformers (ViT's) with equivalent performance to a large CNN have been trained with a fraction of the overall compute power. This reduction in training computational power is critical to continuing to push the cutting edge as current top models are trained using 10's of thousands of compute core days.

Due to the huge potential for ViT's this paper will be discussing the fundamentals of transformers and ViT's along with some current and future improvement's for the models.

2. Background

2.1. Transformers

Transformers, introduced in the groundbreaking 2017 paper "Attention Is All You Need" by Vaswani et al[8], represented a significant shift from conventional Recurrent Neural Network(RNN)s and CNNs, emphasizing an attention-based mechanism. This shift towards self-attention mechanisms enabled parallel processing of input data, significantly enhanced performance and efficiency of Natural Language Processing(NLP) models.
The essence of the Transformer architecture is the self-attention mechanism. It allows dynamic focus on various parts of the input to understand context and relationships. With the combination of encoders and decoders, Transform-

ers led a breakthrough in scalability and parallel processing in large-scale models like GPT and BERT [4], which have redefined benchmarks in NLP tasks.

Transformers have revolutionized machine learning, especially in NLP, by enabling more accurate understanding and generation of human language. Their adaptability extends beyond NLP, as seen in image processing with ViTs, which apply self-attention to image pixels, rivaling advanced CNNs in precision.

The emergence of Transformer models signifies a transformative era in artificial intelligence, reshaping our understanding and expectations of what's achievable. Their innovative architecture and broad applicability have set a new standard in the field, underscoring their potential to drive future advancements in technology.

2.2. Vision Transformers

With an adaptation of Transformer technology to the realm of image analysis, ViT is developed. This adaptation is not just a superficial application of existing models to a new domain. It involves a fundamental rethinking of how images are processed and understood by artificial intelligence systems.

ViTs start by dividing an image into smaller, fixed-size patches. Like the tokens or words in NLP, each patch is flattened and linearly embedded, similar to how words are vectorized in text-based Transformer models. This process transforms the image into a sequence of vectors, enabling the self-attention mechanism of Transformers to be applied effectively. Once the image is segmented into patches, the Transformer architecture takes over. It employs layers of multi-headed self-attention, allowing the model to weigh and compare parts of the image in the context of the whole. This approach is a departure from the localized view of traditional CNNs, where filters move across the image capturing local patterns. ViTs, on the other hand, can grasp global relationships within the image, which can be crucial for understanding complex scenes.

Additionally, ViTs introduce a 'class token' - a special token added to the sequence of image patches. This token evolves during the training process and ultimately holds the information necessary for classification tasks. It's an elegant solution that integrates the classification process into the architecture.

The impact of ViTs is evident in their performance. They have shown remarkable results in image classification tasks, rivaling and sometimes surpassing traditional CNN-based models. What makes this even more impressive is the relative efficiency with which ViTs operate. They require significantly less computational power, especially when pretrained on large datasets, making them a more accessible option for various applications.

In summary, Vision Transformers represent a significant advancement in how models interprets and analyzes images. By adapting the principles of Transformers from NLP, they have opened new avenues in image processing, offering both improved performance and greater efficiency.

3. Problem Statement

As Transformer popularity has grown so have the number of models and variations being created. Even in the realm of just Vision Transformers there are hundreds of proposed variations and improvements. With such a rapidly developing field and constantly evolving Transformer models it is difficult to know which model is best. Each model claims to be the new best but many papers only compare their model to the original or one or two other models when establishing their claims. Additionally many models results are very dependent on the training method used and the dataset being evaluated.

Due to the large number of challenges with comparing ViT models it is important to have detailed comparisons of groups of State of the Art models. This comparison needs to cover not only the results of the models but also the theoretical limits and benefits of each models variations. This analysis is important for understanding how the models will behave on a specific task.

With so many new models available this paper can only cover a subset of Vision Transformer models. We have chosen to work with models which have existing Pytorch implementations publicly available for use. This was chosen so that readers looking for a model can see a comparison of the models they can implement. By using standard implementations this allows the paper to not just evaluate the architecture choices but to give an understanding of the specific implementations performance.

Finally as many papers carefully select their training and testing methods to optimize their architecture it is important to evaluate multiple models with a standard training methodology. For this reason this paper will use the same basic training method for training all models being compared. This will provide an equal footing for all models being compared without using any gimmicks such as hyperparameter tuning or transfer-learning. It is important to provide evaluation without these additional tricks as the average implementation may not benefit from these techniques.

4. Literature Review

4.1. VIT

"An Image is Worth 16x16 Words"[5] is a foundational paper for ViT's as it was the first to achieve better than CNN results on common image benchmarks using a pure

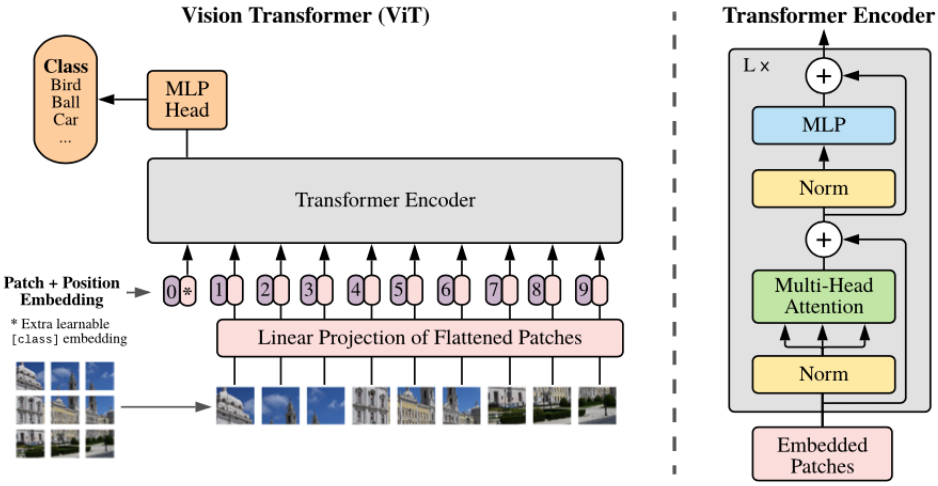

Figure 1. ViT

Transformer architecture without modification. To achieve these results a basic transformer design was used with the input image being separated into a string of 2D patches with a fixed pixel size. This string of patches is treated similarly to how a sentence is the input to a transformer when used in language processing. For larger input images the length of the sequence of patches can be increased. Overall this papers ViT implementation was fairly simple in design and can be seen in Figure1.

What allowed this papers ViT model to excel was the training method used. This paper found that pre-training the ViT on very large image sets allowed the models to perform better on the relatively smaller image sets used for evaluation. This pre-training was followed by some fine-tuning on the target image set to ensure final performance was optimal for the images being evaluated in the test data. They also performed the fine-tuning using higher resolutions than the pre-training. For the higher resolution training the same 2D patch size was used but the image was broken into more patches. The positional encoding of the patches was also tweaked to reflect the different resolution. Overall this paper showed the importance of proper training for ViT's along with there overall potential to rival CNN's.

4.2. CrossViT

One of the relatively undefined components of the ViT model is the size of the image patches used. In general for ViT's the smaller the patch size the larger the model needs to be and the more accurate it is. The paper "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification"[2] proposes using two different patch sizes one large and one small patch size to create a model that is accurate without being super large and slow to train. The overall goal is to have the benefits of using both a large patch size for faster training time and a small patch size for greater accuracy of the final model. To make this work their model is comprised of 2 Transformers lines in parallel, one for each patch size as shown in Figure2. These two transformer models are linked by layers referred to as cross-attention.

The cross-attention layer works very similar to the self-attention component in a transformer. For the cross-attention layer before beginning the normal self-attention layer of a transformer the Class tokens for the two transformers are swapped. This means that the CLS token of one branch and patch tokens of the other branch are multiplied for the cross attention step. After the step they are exchanged back allowing the next normal transformer step to update each based on the cross-attention output of the other transformer branch. This repeated exchange of classification tokens between the two transformers allows for a consensus of sorts on what the final classification of the image should be.

Overall the cross-attention layer designed in this paper is very interesting and has a lot of potential for future applications where the transformers may have different inputs related to the same task. I could see one transformer having a filtered version of the image or even being fed a caption or other relevant text. Overall while the current application of the layer for linking a wide shallow network to a deep and narrow transformer network is useful I believe the future will see many other variations on this technique.

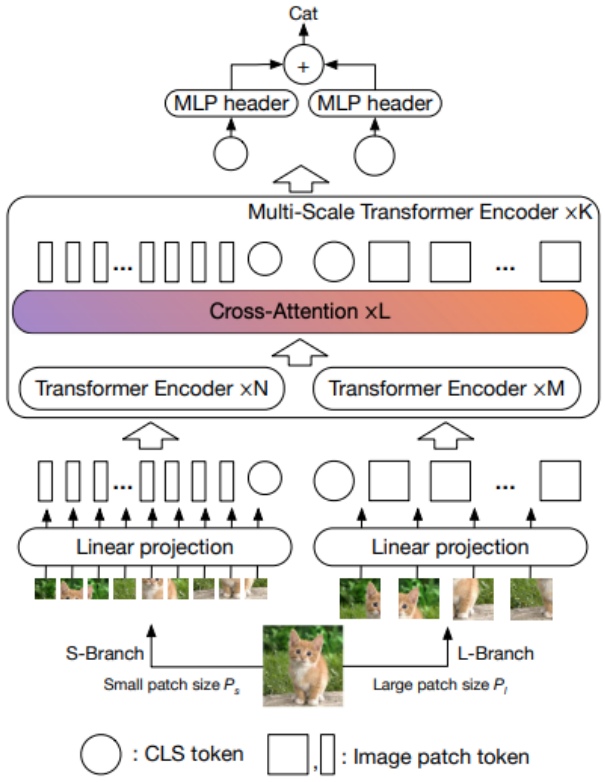

Figure 2. CrossViT uses 2 transformer stacks with a small and a large patch size tied together intermittently with cross-attention layers allowing them both to agree on a class.

4.3. MobileVit

ViT's are generally more accurate and smaller than CNN's when trained and evaluated on large Datasets. This dominance of transformers over CNN models does not hold in the mobile space of small efficient models. In small models below 5 million parameters ViT's struggle to learn enough to compete with CNN's. The paper "MobileViT: Light-Weight, General-Purpose, and Mobile Friendly Vision Transformer" [6] proposes a new hybrid model to fill this gap. The MobileViT model attempts to include convolutions with the Transformer architecture with the goal of adding convolutions spatial induction biases to the ViT to improve their small model performance and their ease of training. By expanding ViT the MobileViT model hopes to maintain the global processing that has allowed ViT's to outperform CNN's on the high end. The papers authors also believe that the inclusion of Convolutions will make their model easier to optimize than vanilla Transformers. By requiring less hyper-parameter tuning this should make the model less costly to train.

MobileViT is comprised of multiple convolutional layers including MobileNetV2 blocks which perform down-scaling on the image. Along with these convolutional layers are MobileViT blocks of varying depth. MobileViT blocks consist of convolutions, unfolding and refolding the image, multiple Transformer blocks in the middle, and a skip connection linking the input to 1 convolution before the output. The overall MobileViT architecture along with a detailed view of the MobileViT block can be seen in Figure3 The multiple downscalling layers that happen before and between the transformer layers help to reduce the image size and therefore the required size of each transformer layer. For its size MobileViT was successful in producing cutting edge accuracy results alongside a general reduction in total parameters compared to other mobile CNN models. The authors did mention MobileViT having slower mobile latency than CNN models but cited this being an issue of hardware optimizations for CNN's being more prevalent than optimizations for Transformers. With upcoming hardware improvements for transformer architectures on mobile devices the performance latency should decrease.

4.4. SimpleVit

The paper "Better plain ViT Baseline for ImageNet-1k"[1] is a short addition written by some of the authors of the original ViT paper. This paper covers a number of changes to the model and the accuracy and training improvements that they provide. Some of these changes include a change in batch size, some additional augmentations, and replacing the MLP output with a single Linear layer. They also changed the position encoding for the patches from learned encoding to the use of Sine Cosine based encoding for the patch position which is more robust. The last change was replacing class(CLS) tokens with Global Average Pooling. Overall these changes allow the model to achieve slightly higher accuracy on the ImageNet-1k dataset while simplifying training.

While the original ViT needed High resolution Fine-Tuning and worked best when trained on very large datasets this new version can be trained without these tricks. The paper claims Simple ViT can be trained for 90 epochs in just 6.5 hours and reach 80% Accuracy in less than a day on a single TPUv3-8 node. This is without leveraging any new training tricks such as distillation or transfer learning. Overall this paper strives to simplify the original ViT model and training methodology while still improving on its accuracy.

4.5. CaiT

The paper "Going deeper with Image Transformers" [7] looks at how to effectively increase the depth of ViT models to improve the testing accuracy. The original ViT model peaks at a depth of 18 producing 80.% accuracy with additional layers resulting in a small loss in accuracy. The

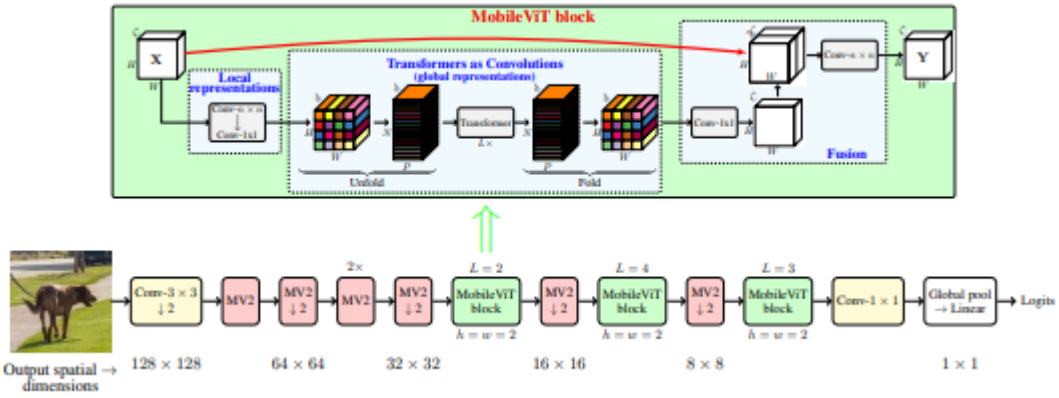

Figure 3. MobileVit contains many different layers to process and down-sample the input image along with the new MobileViT Block which contains the models Transformers.

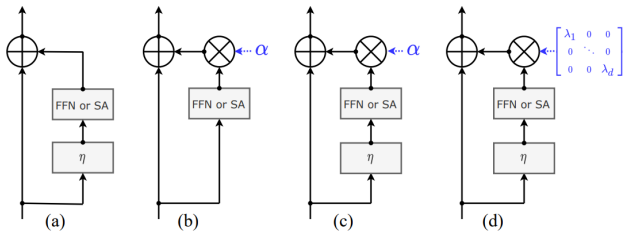

Figure 4. Variations on residual blocks in different architectures. (a) Residual block for the original ViT with prenorm(η). (b) variation removing prenorm but adding scalar α. (c) prenorm is returned to stabilize training. (d) CaiT LayerScale variation with α replaced by diagonal matrix of scalars.

paper discusses some previous attempts to improve depth training for ViT but notes that the changes made in these papers which altered the normalization of the Transformer blocks resulted in unstable training. Along with the altered normalization these models add a learn-able scalar which is multiplied by the output of the FFN and SA layers before the addition component of the residual blocks shown in Figure 4. CaiT expands on this by introducing a full diagonal matrix of scalars to multiply by the residual blocks output. These scalars are initialized to small initial values to encourage the deep networks output to start close to its input allowing for more gradual learning of the residual changes needed to fit the model.

Limiting the residual blocks allows CaiT to implement effective deeper Transformer networks. In addition to this change for deeper networks the paper also proposes introducing the CLS tokens later in the network. This allows the earlier Self-Attention layers to focus exclusively on relating the image Patches to each other without also trying to associate class tokens. The paper shows that connecting the class token later in the model does result in better training.

5. Qualitative Comparison

We are comparing 5 state-of-the-art version of Image Transformer's including the original ViT. We are evaluating the models against the MNIST image set of hand written numbers. This is a relatively small image set of low resolution. This realm of low resolution and small datasets is typically a challenging area for Transformers which usually benefit from large amounts of training data. Given this test setup the MobileViT is likely to show a strong performance as it is designed for lightweight tasks. This also means MobileViT should be the smallest model. While MobileViT is the smallest model is also performs a number of reformatting steps with the data to fold and unfold the data. These steps do not increase size but may increase the runtime and training time for the model due to the added computation. Because MobileViT performs heavy down-sampling It will likely run faster than other models on larger images but this may result in decreased accuracy due to lost information. Overall this model focuses on being lightweight and relatively performant on images around 256x256 in resolution.

Overall SimpleViT should produce similar results to the baseline ViT model. The simplifications made in SimpleViT should result in faster convergence when training as many of the harder to learn components have been reworked such as the patch position encoding. SimpleViT results in better results than ViT only when training time is constrained. When given sufficient training time and optimal tuning ViT will outperform SimpleViT. For the small Dataset we are evaluating ViT will eventually outperform SimpleViT.

CrossViT uses its dual Transformer architecture to produce

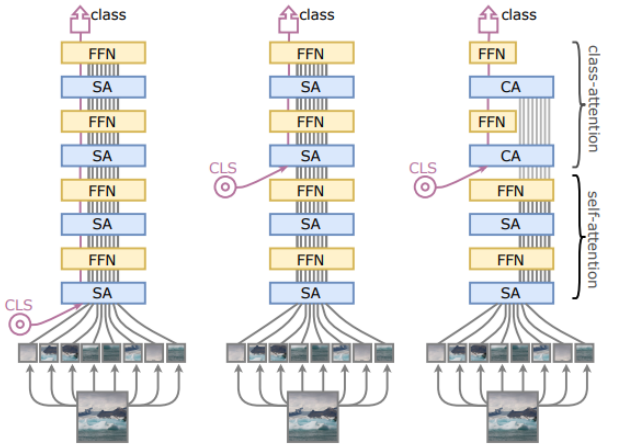

Figure 5. Variations on CLS token introduction to models. (a) standard ViT introduces CLS tokens to the first Transformer. (b) introducing the CLS tokens later results in faster training times and improved accuracy. (c) CaiT introduces the CLS tokens 2 Transformer Blocks from the end and freezes the patch embedding.

amazing results while still reducing the total parameters in the model. This comes from having different depths for the 2 transformer allowing the thicker model to have a shallower depth without reducing the total depth of the network as a whole. Depending on the patch sizes chosen CrossViT could still be larger than ViT due to having nearly 2 copies of the ViT involved. For good choices of patch sizes CrossViT should achieve better accuracy with a smaller model. The added complexity that achieves these improved results will result in longer training times. The training time for CrossViT will be disproportionate to its size as the smaller ViT path in CrossViT is still tied to the longer path and will have to wait on it as the computation progresses forward.

Finally CaiT focuses on improving the training of Deep networks through reducing the initial learning rate and adding additional parameters to the layers that dampen the residual layers. These additional parameters and the added depth of the CaiT model will increase the overall size of the model. CaiT also alters the logic for determining class tokens choosing to introduce the CLS tokens later in the architecture as seen in figure5. This late introduction of CLS tokens allows the earlier self-Attention layers to focus on the sole task of patch association and by freezing the patch embedding in the layers with CLS tokens introduced this reduces the computation required for these later layers. Overall CaiT is designed to be competitive with large ViT models running on very large datasets. For small datasets CaiT is an oversized model and may struggle to converge meaningfully.

6. Numerical Comparison

6.1. Test Procedure

We provide a numerical comparison of various Vision Transformer (ViT) models based on the MNIST dataset. The models under consideration are SimpleViT, MobileViT, CaiT, CrossViT, and the original ViT. Since the dataset is small and less complex compared to the dataset such as ImageNet, we correspondingly reduced the depths of the models when setting up. Models like SimpleViT and MobileViT, which are presumably designed for efficiency and lower computational demand, are particularly well-suited for MNIST. On the other hand, CaiT and CrossViT, despite their more complex designs, also had their depths adjusted to match the simplicity of MNIST. The original ViT, serving as the baseline, were similarly adapted with fewer layers or parameters for optimal performance on MNIST. Overall most transformer models are very powerful and can produce very high accuracy's on the simple MNIST dataset.

6.2. Accuracy

The accuracy trends observed across training epochs are highly indicative of each model's effective learning capabilities on the MNIST dataset. Consistently, all models demonstrated a steady increase in accuracy, aligning with expectations for a dataset of MNIST's relatively simple nature. Notably, some models exhibited a faster convergence, likely attributable to the limited complexity and variability inherent in MNIST. Among the five models, MobileViT displayed an exceptionally rapid convergence. Remarkably, it achieved a test accuracy of 100% and reduced the loss to zero within just 7 epochs, significantly outpacing the average convergence time of approximately 20 epochs for the other models.

For the accuracy, The MNIST dataset is separated into 60000 training samples and 5000 testing samples for training and 5000 validating samples for performance. All the accuracy are based on the performance on validation dataset. MobileViT and CrossViT reach higher accuracy than the ViT baseline. CaiT and SimpleViT get worse performance than the ViT baseline. The SimpleViT's worse performance is because of the simplification on the structure such as the linear layer replacing the MLP head. It is trading off the performance with a faster training time. As it turns out, they achieves faster training time. CaiT reaches the lowest performance. One reason could be that the improvements are intended for the deep transformer based on the large dataset. The other reason could be the short training epoch.

The CrossViT improves the performance 0.81% with the two vision transformers processing the image at different scales. The MobileViT reaches the highest performance. It is a lightweight transformer for mobile device intended to specialize

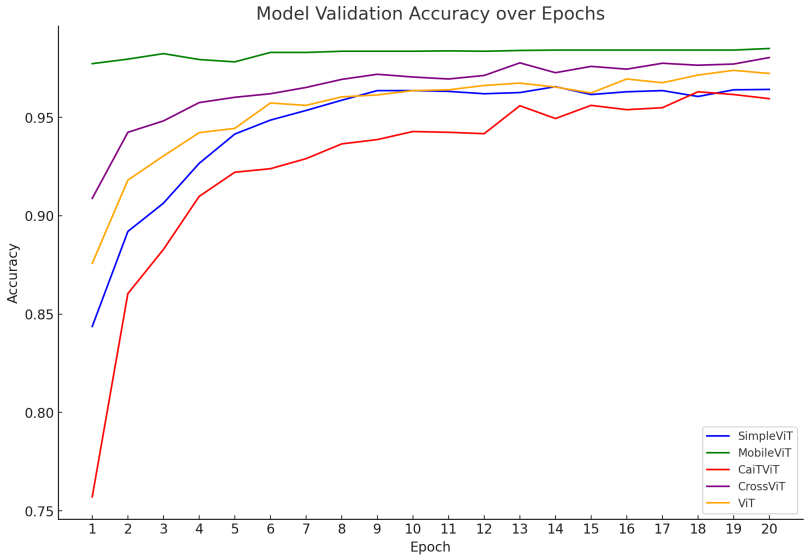

Figure 6. Run time accuracy

in smaller images. Due to the large amount of downscalling in the MobileViT, we upscaled the mnist image from (28,28) to (32,32) to keep the essence of the model. this small increase in input image size could potentially improve the performance of this model.

6.3. Training time

The models are all trained on google colab with a T4 GPU on a PyTorch platform with 60000 training sample and 20 epoch. The training time and resource utilization of these models, when applied to MNIST, are anticipated to be more efficient than in scenarios involving more complex datasets. Models with simpler architectures, such as SimpleViT and MobileViT, are expected to benefit from shorter training times and reduced resource demands. Conversely, more intricate models like CaiT and CrossViT may necessitate a modest increase in computational resources, even considering their adaptations for MNIST. CrossViT, for instance, requires roughly triple the training time of the baseline ViT model, reflecting its more complex structure. SimpleViT shows a reduced training time, whereas CaiT demands about 50% longer, aligning with expectations. SimpleViT started strong quickly surpassing ViT's accuracy at epoch 7 but plateaued early and was beaten by ViT's final accuracy.

Surprisingly, MobileViT, despite its efficient design, takes the longest training time through 20 epochs. The slower processing could be due to the slightly larger size of each input image used for MobileViT. The extended duration is primarily attributed to the reformatting of data before and after each Transformer layer. The original MobileViT paper also mentions a relatively slow processing speed for the network but attributed it to mobile hardware not supporting Transformers as well as CNN's. Given that MobileViT was still slower in comparison to larger transformer models we can see that it was not just a hardware issue causing MobileViT to have long processing times.

6.4. Model size

For model size, the original ViT and SimpleViT are the same, each comprising 5.86 million parameters. SimpleViT uses 28 layers while the ViT have 37 layers. CrossViT utilizes only 3.42 million parameters with 237 layers. MobileViT, at the lower end, has a mere 2.00 million parameters, but it has 239 layers. A large amount of them are the downsampling blocks for the architecture. The heavy use of downsampling in MobileViT allows it to use less parameters in many of the layers. Standing in contrast is CaiT, with the highest parameter count at 11.59 million – almost double that of the original ViT with 77 layers. This increase is a predictable outcome of CaiT's design philosophy, which aims to optimize deeper transformer networks by adding parameters and layers to the ViT network. This means each of CaiT's layers is larger than the layers in MobileViT resulting in nearly 6x the parameters with 1/3rd the layers.

Table 1. Model sizes in millions of parameters, training time in seconds, and final accuracy after training for 20 epochs.

MODEL	PARAMS	TRAINING TIME	ACCURACY
VIT	5.86M	457S	97.23%
CROSSVIT	3.42M	1221S	98.04%
CAIT	11.59M	654S	95.95%
MOBILEVIT	2.00M	1431S	98.50%
SIMPLEVIT	5.86M	439S	96.42%

References

[1] L. Beyer, X. Zhai, and A. Kolesnikov. Better plain vit baselines for imagenet-1k, 2022.

[2] C.-F. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification, 2021.

[3] J.-B. Cordonnier, A. Loukas, and M. Jaggi. On the relationship between self-attention and convolutional layers, 2020.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[6] S. Mehta and M. Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer, 2022.

[7] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou. Going deeper with image transformers, 2021.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.