Transformers transforming Vision: A comprehensive study on transformers for image recognition

Umaima Rahman umaima.rahman@mbzuai.ac.ae
Muhammad Uzair Khattak uzair.khattak@mbzuai.ac.ae
Maryam Nadeem maryam.nadeem@mbzuai.ac.ae

Abstract

Transformers are now considered state-of-the-art in sequence modeling tasks. This is primarily due to the use of attention based mechanisms that help them eschew convolution and recurrence entirely, making the training parallelizable and retaining the long-range dependencies efficiently. Recent works on transformers employing self-attention and the global attention paradigm have shown remarkable performance in multiple computer vision tasks. We revisit the baseline vision transformer models and highlight their performance on image classification problem. We explain their architectural designs and compare vision transformers with their state-of-the-art ResNet-BiT for image classification. Specifically, we evaluate the performance and scalability of vision Transformers and ResNets on standard benchmark datasets: CIFAR-10, CIFAR-100 and CUB-200. Our experiments show that the accuracy curve of ResNets-BiT plateaus whereas vision transformers perform differently when the size of the pretraining dataset is increased. We also highlight the significance of DeiT with augmentation over the vanilla ViT models. Our code files and additional materials are available on Github.

1 Background

One of the most important problems in computer vision is image classification. It is a fundamental problem in computer vision, often used as a benchmark to evaluate the performance and effectiveness of a model. They are also used as a building block in many sophisticated computer vision algorithms (such as in object detection algorithms) as shown in Fig. 1(a). It has multiple applications ranging from medical imaging [Li et al. (2014); Yadav et al. (2019)] to remote sensing [Al-Doski et al. (2013); Tuia et al. (2009)], and self-driving vehicles [Sheri et al. (2018)] to robotics [Chavez-Garcia et al. (2017)]. Image classification in general, helps to assign a class from a list of predefined classes to an image after processing it. This task is straightforward for humans to accomplish, but when it comes to simulating the same task using computers, it becomes difficult.

In 2012, when AlexNet, a convolutional neural network (CNN) presented by Krizhevsky et al. (2012) achieved state-of-the-art performance on the ImageNet (ILSVRC-2012) challenge, CNNs garnered huge attention from the research and industry community alike. It opened interesting avenues for using CNNs in various applications such as image classification, segmentation, face recognition, medical image analysis, etc.

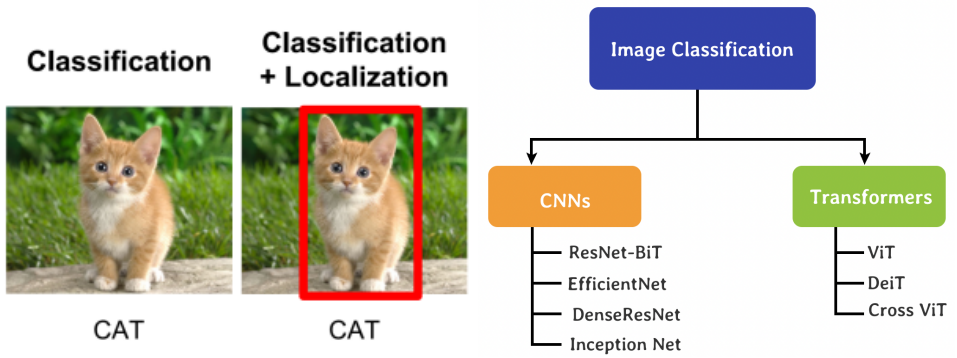
Until now, convolutional neural networks [Jmour et al. (2018)] were dominant in this area due to their intrinsic property of high inductive bias [Cohen et al. (2016)]. However, CNNs are diagnosed with the limitation of stagnated performance even with the availability of large-sized datasets, which restricts their capabilities on different computer vision problems.

Recently, with the success of transformers in the domain of NLP, the concept of attention has also been introduced to vision related applications [Parmar et al. (2018)]. Vision Transformers (ViTs) proposed by Dosovitskiy et al. (2020) for tasks related to computer vision, outperformed state-of-the-art CNN models on standard benchmark datasets. Fig. 1(b) shows different image classification techniques.

The standard vision transformers require large pretraining dataset to compensate for their weak inductive bias to give excellent results. This constraint can impose restriction on its usage as the availability of large sized datasets is often limited. As a result, data efficient vision transformers (DeiT) proposed by Touvron et al. (2021) perform better than the standard baseline vision transformers without huge data requirement.

In this work, we attempt to understand the working of vision transformers for image classification and compare the performance of ViTs and DeiTs with state-of-the-art CNNs i.e., ResNet on benchmark datasets: CIFAR-10, CIFAR-100 and CUB-200. To study how CNN acts differently than vision transformers, we test the performance of both families of models on downstream datasets after they are pretrained on large scale datasets. These experiments helps us to highlight the performance capabilities of both CNNs and vision transformers, when the size of pretraining dataset varies.

Further, we show the importance of data augmentation for vision transformers, specially for DeiTs in order to show improvement in their performance. Also, we observe the impact of increasing the image resolution size on the model's performance during finetuning on downstream datasets.



(a) (b)

Figure 1: (a) Image classification is a problem in computer vision that forms a building block for other applications. (b) Some deep learning based image classification techniques.

2 Related Work

Image classification is one of the fundamental computer vision applications that aims to classify images into their respective categories. Conventionally CNNs and recently transformers have been performing well on this task.

2.1 Convolutional Neural Networks

Over the past several years, deep convolutional neural networks (ConvNets) have remained the state-of-the-art on ImageNet (Russakovsky et al. (2015)). Convolutional neural networks like ResNets and EfficientNet have achieved breakthrough performances on image classification tasks in recent years.

2.1.1 ResNets

In general, deep neural networks have been difficult to train due to the following factors: (i) the vanishing gradient problem which hampers convergence and (ii) by increasing the depth of the network, the accuracy gets saturated and then starts degrading rapidly. To address these problems, He et al. (2016), proposed the famous ResNet architecture which introduces "identity shortcut connection". These connections allow the gradients to flow unhindered through the shortcut connections to any previous layer. Nevertheless, scaling ConvNets helps to improve the performance accuracy. For example, ResNet152 has more learnable parameters compared to ResNet52 and hence gives better accuracy. But at the same time due to a large number of parameters, the network is not very efficient.

2.2 Vision Transformers

With the success of transformers in NLP, the emergence of vision transformers (ViT) by Dosovitskiy et al. (2020) showed a promising future of transformers in vision applications. Here, we explain the baseline models made solely of transformers for image classification. These include Vision Transformers and Data Efficient Image Transformers known as ViT and DeiT respectively.

2.2.1 ViT

ViT is a convolution free transformer based architecture which uses the encoder block of the original transformer proposed by Vaswani et al. (2017) and contains 12 such encoder blocks. Each of of these encoders contains a multi-head self-attention block (MHSA) and a multi-layer perceptron (MLP). The main architecture of ViT is shown in Figure. 2

Vision transformers start by dividing the input image into n patch tokens. These patch tokens are flattened and linearly projected to obtain the patch embeddings. To include structural information, patch embeddings are combined with learnable positional embeddings and a class `CLS` token. These combined embeddings are supplied as input to the MHSA block of the transformer encoder. The MHSA block is responsible for computing self attention i.e., to encode the interaction among these n patches in terms of global contextual information (Khan et al. (2021)). It uses three learnable weight matrices $W^Q \in \mathbb{R}^{d \times d_q}$, $W^K \in \mathbb{R}^{d \times d_k}$ and $W^V \in \mathbb{R}^{d \times d_v}$ which are multiplied with the combined embedding $X \in \mathbb{R}^{n \times d_k}$ to output the matrices: Query (Q), Key (K) and Value (V), where, $Q = XW^Q$, $K = XW^K$ and $V = XW^V$. The output $Z \in \mathbb{R}^{d \times d_v}$ is computed as follows:

$$Z = softmax\left(\frac{QK^T}{\sqrt{d_q}}\right)V \tag{1}$$

Here d_q is the dimension of the query matrix Q and $d_q = d_k$. In the ViT-Base model, the MHSA block comprises of 12 heads, each one having their own learnable weight matrices (W^Q, W^V and W^K) to produce the output Z. For a single MHSA block, these heads are computed all together in parallel. Outputs from all these heads are concatenated together and passed to a Layer Norm which is then fed to the MLP. Further, the output from one transformer encoder is passed to the next. Eventually, output of the last transformer encoder is supplied as input to the last MLP Head which is a feed forward netowrk. The MLP head provides the final outputs as class predictions.

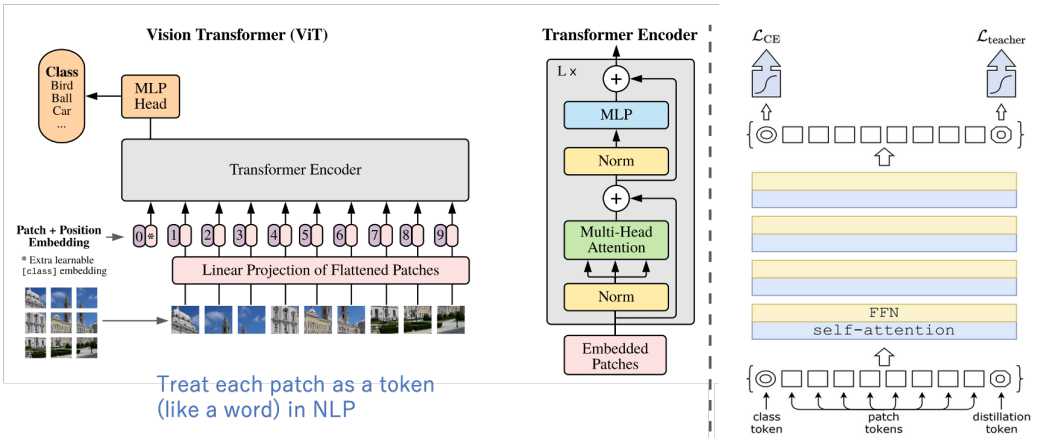


Figure 2: Architecture of two vision transformers discussed in our study: ViT along with the transformer encoder block (left) and the data efficient image transformer DeiT (right)

2.2.2 DeiT

Subsequently, many researchers have experimented with transformers and tried to improve the baseline ViTs for image classification tasks. The work done by Touvron et al. (2021) proposes data efficient image transormers (DeiTs) which are identical to ViTs in terms of their architecure except for an additional distillation token, which interacts with the class token during training. The high level architecture of DeiT is shown in Figure. 2. DeiTs perform better than ViTs without the need of large pretraining datasets. It uses a very sophisticated training pipeline, employing strong data augmentations such as RandAugment, AutoAug, CutMix, Mixup, RandomErasing and RepeatedAug. In the experiments, DeiT with such settings lead to better validation performance. In addition to that, DeiT distilled model uses the concept of knowledge distillation in which a student transformer model learns from a teacher model. On experimenting with CNNs and transformers as teacher models, it was inferred that a CNN teacher model performed better. It was shown that DeiT with such an architecture was successful in giving competitive results for the same problem.

3 Description of the methods

To carry out the experiments, we implemented our scalable and generalizable code scripts in Python. PyTorch is chosen as the framework for using deep learning models. Official computer vision library timm (PyTorch Image Models) by Wightman et al. (2019) is used for implementing ViT, DeiT and ResNet BiT models. Separate code scripts are implemented for each individual module (for train, test, and datasets, etc) to improve the readability of those modules. The advantage of using timm library is the availability of all required models and their pre-trained weights. This enables the use of general train and evaluation scripts which are suitable for all aforementioned models. Our code is available on Github.

3.1 Approach

The same approach as used in Dosovitskiy et al. (2020) was followed. We use pre-trained vision transformers and CNN models and apply transfer learning. The models are then fine-tuned on down-stream datasets by modifying their last layer and re-learning it from scratch. All the other layers are kept frozen and finally validation and test results are observed.

In the next iteration, both families of models are again pre-trained, but on a dataset larger than the one used in first iteration. Models are fine-tuned on the same downstream datasets and their test and validation accuracies are observed. Eventually, the performance of models from both the iterations are compared. We review each model's performance scalability when the pre-training dataset proportion varies. Fig. 3 summarizes our methodology.

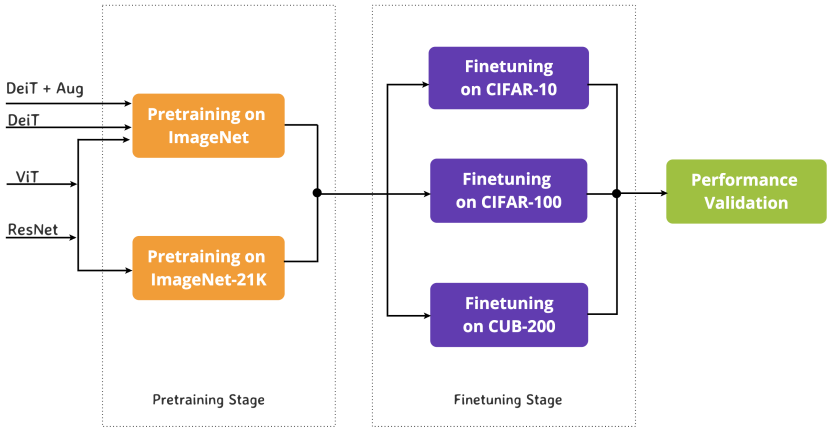


Figure 3: Our approach uses pretrained models of CNNs and transformers which are then finetuned on downstream datasets to finally validate their performance.

Additionally, we also investigated the performance of our models by:

- Finetuning them on different image resolutions: Initially we keep the image resolution as 224×224. However, to conduct similar experiments as done by Dosovitskiy et al. (2020); Touvron et al. (2021) we repeat our experiments with an image resolution of 384×384. This was done by making changes to `pytorch`'s `Resize transform`.
- Substituting the SGD optimizer with Adam optimizer: We observed that the latter gave slightly better performance owing to the fact that it combines momentum with adaptive learning rate for better convergence.
- Training DeiT with and without strong data augmentations: The results we obtained proved that strong data augmentation definitely helps to improve the performance of the models even when pretrained on smaller datasets (as compared to JFT-300M).

4 Experimentation

In this section, we explain the choice of datasets and image classification models along with hyperparameter settings. Finally we present and reason about the results obtained from the experiments.

Dataset	Train Size	Test Size	# classes
ImageNet-21K	14M	-	21K
ImageNet	1.3M	100K	1K
CIFAR-100	50K	10K	100
CIFAR-10	50K	10K	10
CUB-200	5,994	5,794	200

Table 1: Number of train and test images along with classes for different datasets. For ImageNet-21k, no official test split is provided.

Dataset	Model Name
ImageNet	ResNet-BiT 101x3 ViT-Base DeiT-Base DeiT+Aug-Base
ImageNet-21K	ResNet-BiT 101x3 ViT-Base

Table 2: Summary of the image classification models and datasets used for pretraining in the experiments

4.1 Datasets

For pretraining of models, we use two separate datasets. First, we use ImageNet dataset from ImageNet Large Scale Visual Recognition Challenge 2012 which has about 1.3 million images covering 1000 labels. For the second iteration, its superset, ImageNet-21k dataset is used. This dataset is approximately 10x larger than the ImageNet and it has 21000 classes. For all models, we use their trained weights on ImageNet and ImageNet-21k using the timm library.

Datasets for downstream tasks include image classification on CIFAR-10, CIFAR-100 and CUB-200. CIFAR-10 contains 60,000 images in total for 10 classes with 6,000 images per class. CIFAR-100 includes 60,000 images for 100 classes, so total of 600 image per class is available. Images in both CIFAR-10 and CIFAR-100 are of dimension 32x32x3. CUB-200 is a dataset with 11,788 photos of 200 North American bird species. The images are of a higher resolution, most of them being close to a dimension of 500x300.

Some sample images from CIFAR-10, CIFAR-100 and CUB-200 are shown in Fig. 4. Once trained on large scale datasets (ImageNet and ImageNet21k separately), Vision transformers and CNNs are fine-tuned on these downstream datasets. Table. 1 shows quantified information about pretraining and finetuning these datasets.

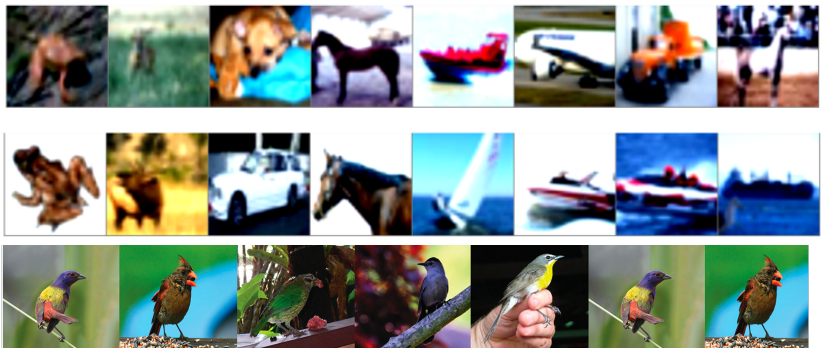


Figure 4: The first row shows a batch of 8 images for CIFAR-10 which contains 10 classes, the second row shows a batch of 8 images for CIFAR-100 which contains 100 classes. Both CIFAR-10 and CIFAR-100 have image resolution of 32x32. The last row shows a batch of 7 high resolution images belonging to the CUB-200 dataset.

4.2 Choice of Models

To experiment, we use ResNet 101x3 BiT which is a variant of ResNet Big Transfer proposed by Kolesnikov et al. (2020). ResNet-BiT is a ResNet architecture which is modified to use group normalization and standardized convolution instead of batch normalization and normal convolution respectively. ResNet 101x3 repeats the number of layers in the original architecture three times as compared to the standard ResNet-101. This is done in order to incorporate similar number of learnable parameters as compared to the vision transformer models.

For vision transformers, we use the ViT-Base model "ViT-B/12" which has about 86 million parameters and the data efficient DeiT-B model which has similar number of parameters. Table 2 shows the choice of pre-training dataset used for the models in our experiments.

4.3 Train, validation and test datasets

The official train, test and validation images provided by imageNet (1k classes) and imageNet-21k (21k classes) datasets are used with no other custom splits. We do not use these datasets directly, but rather use model pretrained on it. For CIFAR-10 and CIFAR-100, we split the training images and use 2% for validation and remaining 98% for training, whereas for CUB-200 we split the training images and use 10% for validation and remaining 90% for training. All the train images are shuffled before splitting. For testing, the available official test splits for CIFAR-10, CIFAR-100 and CUB-200 were used.

4.4 Dataset preprocessing and training

Pretrained model weights available in the timm libarary for ViTs, ResNet-BiT and DeiT are independently trained on ImageNet and imageNet21k datasets. These models are then fine-tuned on CIFAR10, CIFAR100 and CUB-200. Only training images are subjected to random horizontal flipping with a batch size of 256. The images from train, validation and test set are first resized to image resolution of 224×224 and later to 384×384 . Each model is fine-tuned for 50 epochs using Adam optimizer.

5 Results

We evaluate the baseline vision transformers against ResNet-BiT for different downstream datasets: CIFAR-10, CIFAR-100 and CUB-200, as discussed subsequently. We used Adam optimizer for all our experiments. Validation and test accuracies of all the models finetuned on different datasets are summarized in Table 3.

5.1 Validating results on CIFAR-10

When finetuned on CIFAR-10, ResNet BiT pretrained on ImageNet achieves a test accuracy of 95.33%. The performance of ResNet-BiT is improved by 0.09% when the pretraining dataset is switched from ImageNet to ImageNet-21k.

ViT-Base from the vision transformers family gives test accuracy of 84.39% and 72.68% when it is pretrained on ImageNet and ImageNet-21k dataset respectively. Surprisingly, the performance decreases sharply when ImageNet-21k dataset is chosen for pre-training as shown in Fig. 5(top-left).

DeiT+Aug-Base which uses strong augmentation techniques outperforms all the other models by achieving an accuracy of 97.70 %. The effect of augmentation is prominent because the augmentation deprived DeiT-Base model fall behind by 7%. Both the DeiT models were pretrained on ImageNet dataset.

5.2 Validating results on CIFAR-100

For CIFAR-100, we observe similar results as obtained for CIFAR-10. ResNet BiT again outperforms ViT-Base and DeiT-Base models with an accuracy of 96.42% when it is pretrained on ImageNet. Its test performance decreases by roughly 1% when the pretraining dataset is changed from ImageNet to ImageNet-21k.

The maximum test accuracy that ViT achieves is 65.48% when it is pretrained on ImageNet dataset. Similar to the behaviour observed for CIFAR-10, the accuracy of the ViT drops significantly when imageNet-21k dataset is used for pretraining as seen in Fig. 5(top-right). DeiT-Base achieves 73.55% accuracy whereas with augmentation it rises up to 87.50%.

Pretrained on	**Model** @384	**CIFAR-10**		**CIFAR-100**		**CUB-200**	
		Val. Acc. (%)	**Test Acc. (%)**	**Val. Acc. (%)**	**Test Acc. (%)**	**Val. Acc. (%)**	**Test Acc. (%)**
ImageNet	ResNet-BiT	94.10	95.33	95.60	**96.42**	79.33	81.00
	ViT-Base	84.98	84.39	64.26	65.48	60.68	79.53
	DeiT-Base	89.69	90.18	72.46	73.55	57.81	72.88
	DeiT+Aug-Base	86.10	**97.70**	71.50	87.50	69.62	84.80
ImageNet-21K	ResNet-BiT	**94.34**	95.42	**95.80**	95.67	**84.17**	**85.07**
	ViT-Base	74.70	72.68	49.51	49.40	58.20	75.39

Table 3: Summary of the results obtained with finetuning on CIFAR-10, CIFAR-100 and CUB-200 datasets.

5.3 Validating results on CUB-200

Lastly, we perform similar experiments on the CUB-200 dataset. ResNet BiT provides the highest performance among all models and provide 85.07% test accuracy when pre-trained on ImageNet-21k. Only for CUB200 dataset, we observe that the accuracy for ResNet models increases from 81% to 85.07% when the pretraining increases from ImageNet to ImageNet-21k.

ViT-Base models provides test accuracies of 79.53% and 75.39% when it is pretrained on ImageNet and ImageNet-21k respectively. Here again, we observe a decrease in performance (of about 4%) when the ImageNet-21k weights are used instead of ImageNet as shown in Fig. 5(bottom-center).

We observe significant performance jump when DeiT-Base is finetuned with augmentations. As we see, the augmentation rich DeiT-Base model pretrained on ImageNet, competes with ResNet-BiT model pretrained on ImageNet-21k by achieving an accuracy of 84.80% which is very close to 85.07% provided by ResNet BiT model. However, the DeiT model finetuned without augmentations gives poor performance and its test accuracy is 72.88%.

5.4 Effect of resolution:

To understand the importance of image resolution on the performance of models used in our study, we repeat our experiments with two different image resolutions of 224×224 and 384×384.

When finetuned on CIFAR-10 and CIFAR-100 dataset, ResNet-BiT and DeiT+Aug performs better with a higher resolution of 384×384 compared to its 224×224 counterpart, whereas the transformer models give better accuracies on the smaller resolution. This can be accounted by the fact that CIFAR-10 and CIFAR-100 datasets consists of low resolution images of size 32×32. Resizing these small resolution images to an even higher resolution of 384×384 will include redundant information leading to poor performance. However, on CUB-200 dataset, which consists of high resolution images, all the models give better accuracy when finetuned on the image resolution of 384×384 because it captures more information as opposed to 224×224. The results of our experiments are recorded in Table 4.

Pretrained on	**Model**	**CIFAR-10**		**CIFAR-100**		**CUB-200**	
		Test Acc.% @224	**Test Acc.%** @384	**Test Acc.%** @224	**Test Acc.%** @384	**Test Acc.%** @224	**Test Acc.%** @384
ImageNet	ResNet-BiT	94.22	**95.33**	83.92	**96.42**	71.87	**81.00**
	ViT-Base	**92.09**	84.39	**77.90**	65.48	76.95	**79.53**
	DeiT-Base	**92.28**	90.18	**77.31**	73.55	71.33	**72.88**
	DeiT+Aug-Base	96.20	**97.70**	84.40	**87.50**	79.30	**84.80**
ImageNet-21k	ResNet-BiT	94.54	**95.42**	84.92	**95.67**	79.77	**85.07**
	ViT-Base	**87.40**	72.68	**70.69**	49.40	72.95	**75.39**

Table 4: Summary of the results obtained with finetuning on CIFAR-10, CIFAR-100 and CUB-200 datasets on different image resolutions using Adam optimizer.

Training Time: Our experiments were conducted on an Intel Xeon Silver 4215 CPU with 128 GB RAM and Nvidia Quadro RTX 6000 GPU. Finetuning the ViT models as well as the DeiT model on all the datasets: CIFAR-10, CIFAR-100and CUB-200 is faster than the ResNet 101x3 model. The training time of ResNet is almost 2x that of ViT. For example, when we trained ResNet 101x3 and ViT-B on ImageNet using Adam as the optimizer on the same configuration, we saw that the time taken to train ResNet 101x3 was 5.1 hours whereas to train ViT-B was 2.76 hours. This proves that ViT-B is faster to train compared to its CNN counterparts, which had comparatively lesser number of parameters.

6 Discussion:

From our experiments, we have observed that performance of ResNets are eventually plateaued even when the pretraining dataset was increased. These trends are evident from the plots shown in Figure. 5. On the otherhand, this is not the case for vision transformers as they show different behaviour when pretraining dataset is varied.

Overall, results for ViT-Base model, differs from the results presented in Dosovitskiy et al. (2020) in which the performance of ViTs increases with increase in pretraining dataset size. The reason behind their performance could be as follows:

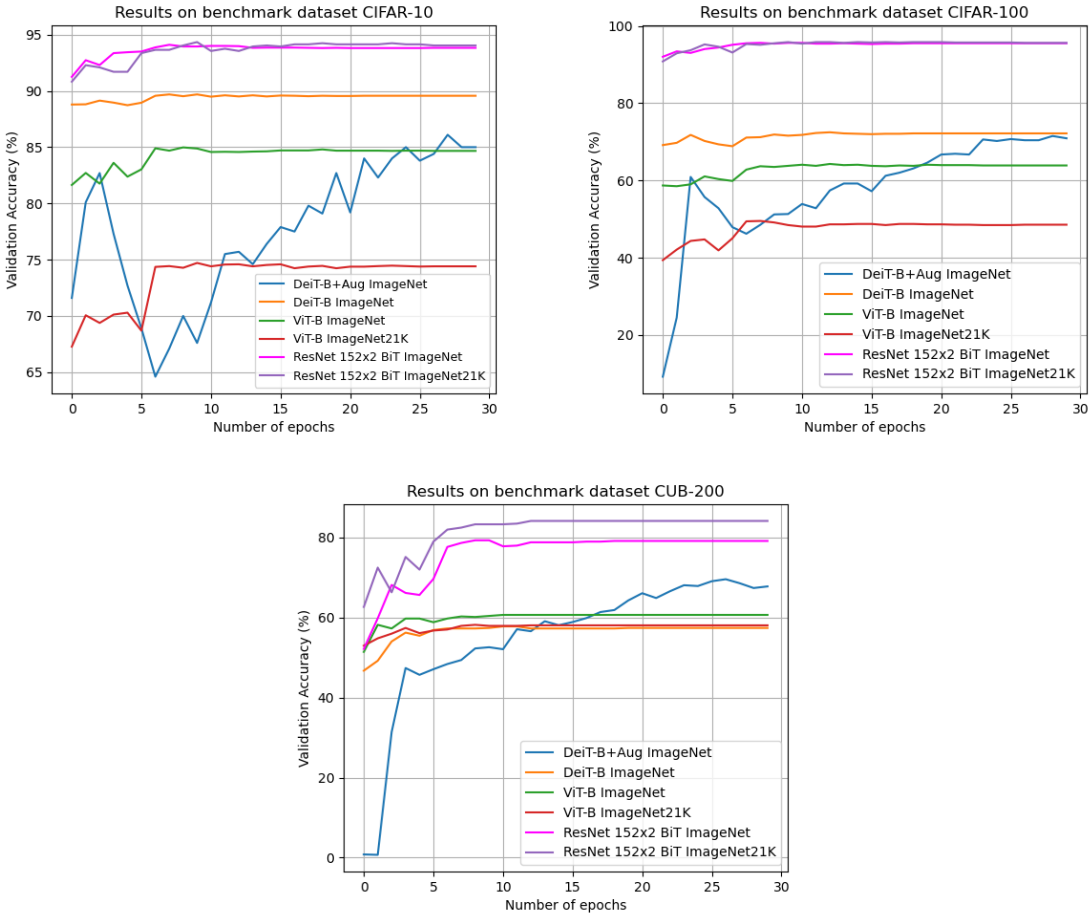


Figure 5: Performance validation on CIFAR-10(top-left), CIFAR-100(top-right) and CUB-200(bottom-center).

Firstly, Dosovitskiy et al. (2020) trains ViTs for 300 epochs, while due to compute limitations we have trained each model for only 50 epochs. Similarly, DeiT models are typically trained for 300-1000 epochs which shows that transformer based models are very compute hungry. We know that, by default vision transformers do not have any prior inductive bias to understand images, it needs a lot of rigorous training to gain a good understanding of the image.

Secondly, we have finetuned the models on the downstream datasets by training only the last MLP prediction head of ViT while keeping all other layers frozen. The official finetuning carried by the Dosovitskiy et al. (2020) used a different finetuning technique, which involved learning of the entire model (with different learning rate schedules), in addition to learning of head layer only. Conducting such experiments again were not feasible for us to finetune a complete model.

Thirdly, we resize images of CIFAR-10 from original resolution of 32x32 to 224x224 which does not improve the details present in those images. Moreover, the distribution of images in CIFAR-10 and CIFAR-100 is very different from that in ImageNet-21k. As a result, we might be facing suppressed peformance because of very low resolution of the original dataset. For CUB-200, the higher resolution images did help in improving the overall performance.

Moreover, it can be also concluded that performance of DeiT models are highly dependent on strong data augmentations. The major reason that data augmentation helps vision transformers more than CNNs is mainly because vision transformers lack the property of transnational equivariance which is naturally inherited by CNNs (Dai et al. (2021)). As there is no convolution operation in transformer blocks, different augmented images of same parent image needs to be shown to the models in order to make it artifically co-variant to translation. Strong

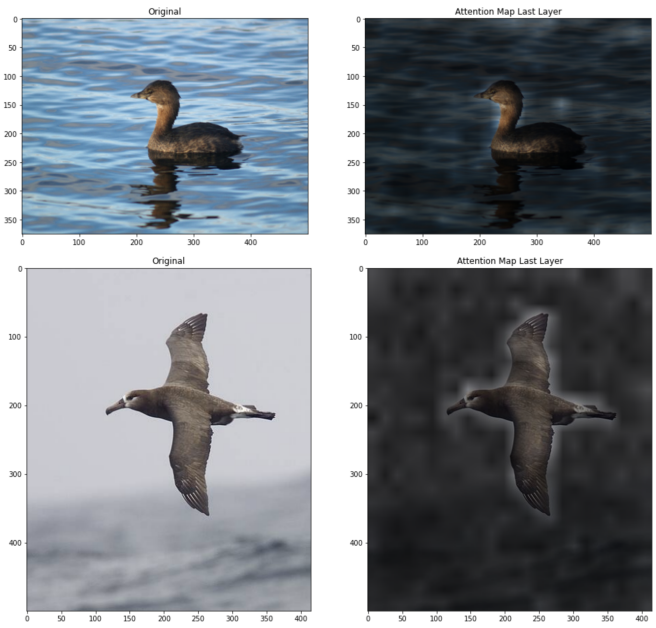


Figure 6: Attention maps (right) for images on (left) are obtained by computing the attention matrix from the last layer of the transformer encoder.

augmentation techniques like CutMix, Mixup, Auto-Augment, Rand-Augment and random erasing helps to induce more generic priors to the vision transformer working pipeline.

Additionally, to show how the self attention works to attend to image regions that are semantically relevant, we obtain attention maps from the last layer of the transformer encoder and map it onto the image as shown in Fig. 6

7 Conclusion

In this work, we attempted to understand vision transformers and validated the results of main baseline vision transformers along with ResNets on standard benchmark datasets: CIFAR-10, CIFAR-100 and CUB-200. From the experiments, we observed that CNN based models i.e., ResNet suffer from performance saturation when the pretraining dataset increases. Surprisingly, for our experiments, the vanilla ViT-Base and DeiT-Base did not outperform ResNet-BiT, however, DeiT+Aug-Base pretrained on ImageNet gave better results than ResNet-BiT pretrained on ImageNet-21K on CIFAR-10 and comparable results on CUB-200. This proves that strong augmentation can significantly help to improve the performance of transformer models. Moreover, our experiments with different image resolutions during finetuning also highlighted their effect on the accuracy of our models. We also conclude that, training ViT for very few epochs and finetuning (for downstream tasks) only their last layers cannot leverage their full potential and therefore it requires substantial amount of training as well as thoughtful finetuning which will lead to promising results.

References

Wightman, R. (2019). Pytorch image models. URL https://github. com/rwightman/pytorch-image-models.(cited on p.).

Jmour, N., Zayen, S., & Abdelkrim, A. (2018, March). Convolutional neural networks for image classification. In 2018 International Conference on Advanced Systems and Electric Technologies (IC_ASET) (pp. 397-402). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2020). Big transfer (bit): General visual representation learning. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16 (pp. 491-507). Springer International Publishing.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning (pp. 10347-10357). PMLR.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2021). Transformers in vision: A survey. arXiv preprint arXiv:2101.01169.

Mahmood, K., Mahmood, R., & Van Dijk, M. (2021). On the robustness of vision transformers to adversarial examples. arXiv preprint arXiv:2104.02610.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

Tan, M., Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (pp. 6105-6114). PMLR.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., Chen, M. (2014, December). Medical image classification with convolutional neural network. In 2014 13th international conference on control automation robotics vision (ICARCV) (pp. 844-848). IEEE.

Yadav, S. S., Jadhav, S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big Data, 6(1), 1-18.

Al-Doski, J., Mansorl, S. B., Shafri, H. Z. M. (2013). Image classification in remote sensing. Department of Civil Engineering, Faculty of Engineering, University Putra, Malaysia, 3(10).

Tuia, D., Ratle, F., Pacifici, F., Kanevski, M. F., Emery, W. J. (2009). Active learning methods for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 47(7), 2218-2232.

Sheri, R., Jadhav, N., Ravi, R., Shikhare, A., Sannakki, S. (2018). Object detection and classification for self-driving cars. International of journal of Engineering and Techniques, 4.

Chavez-Garcia, R. O., Guzzi, J., Gambardella, L. M., Giusti, A. (2017, September). Image classification for ground traversability estimation in robotics. In International Conference on Advanced Concepts for Intelligent Vision Systems (pp. 325-336). Springer, Cham.

Cohen, N., Shashua, A. (2016). Inductive bias of deep convolutional networks through pooling geometry. arXiv preprint arXiv:1605.06743.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D. (2018, July). Image transformer. In International Conference on Machine Learning (pp. 4055-4064). PMLR.

Dai, Z., Liu, H., Le, Q. V., Tan, M. (2021). CoAtNet: Marrying Convolution and Attention for All Data Sizes. arXiv preprint arXiv:2106.04803.