ADROID Conference Series: Engineering and Technology
Volume 2, Issue 3, pp.105-117, 2026

# Revisiting Inductive Bias in Vision Transformers: A Comparative Analysis of Transfer Learning and Training from Scratch on CIFAR-10

Arya Putatunda, Pranshu Kumar, Praveen, Akash Karan

Department of Computer Science and Engineering SRM Institute of Science and Technology, Ghaziabad, India

ap9857@srmist.edu.in, pk4852@srmist.edu.in, aa0560@srmist.edu.in, pz7777 @srmist.edu.in

**Abstract.** This paper is a comparative study of transfer learning and training from scratch on CIFAR-10, and of the relationship between inductive bias and vision transformers. Vision Transformers (ViT) are a relatively new form of potent alternative to Convolutional Neural Networks (CNNs) in the field of computer vision. However, unlike CNNs, ViTs lack strong inductive biases, such as locality or translation invariance and thus owe a lot to large-scale data. The paper will conduct a parallel comparison between training a Vision Transformer on the CIFAR-10 dataset in a pure vanilla environment and using a pre-trained model via transfer learning. A Lightweight hybrid vision transformer is an ad-hoc training and it has an accuracy of 72.82%. In comparison, when a ViT-B/16 is pre-trained on ImageNet and subsequently fine-tuned on CIFAR-10, the accuracy reaches 94.08% after only five epochs. The results indicate that the latter is not as efficient at using data as Vision Transformers trained directly; although trained on a large scale, it has a high representation learning capacity.
Keywords: Vision Transformer, Inductive Bias, Transfer Learning, CIFAR-10, Deep Learning, Hybrid Architectures.

## 1. INTRODUCTION

### *1.1. Motivation*

Convolutional Neural Networks (CNNs) have dominated the architecture of computer vision, with more than a decade of contributions due to their inductive biases, including Locality, Translation Invariance, and Sharing of parameters [1]. By having such properties, CNNs can be trained to be productively aware of spatial hierarchies with a very limited amount of data.

The forerunner of the concept of self-attention to form long-range relationships was the original architecture called Transformer which was proposed by Vaswani et al. [2] in the natural language processing field. This architecture was subsequently implemented by Dosovitskiy et al. [3] as Vision Transformer (ViT) where an image is considered a sequence of patches and repeated applications of regular Transformer encoders are made on it.

ViTs may be equally good at training CNNs, although, unlike CNNs, ViTs do not exhibit the intrinsic inductive bias, and thus they are data-enthusiastic and hard to train on small samples. The given literature study assesses the impact of inductive bias by comparing a scratch-trained Vision Transformer with a transfer learning approach of the CIFAR-10 dataset with a gigantic pre-trained model.

### *1.2. Main Contributions*

The main contributions of this paper are as follows:

- A custom lightweight hybrid Vision Transformer is designed and trained from scratch on CIFAR-10, achieving 72.82% accuracy, serving as a controlled baseline for inductive bias analysis.
- A ViT-B/16 pre-trained on ImageNet-21k is fine-tuned on CIFAR-10, achieving 94.08% accuracy within only 5 epochs, demonstrating the power of large-scale pre-training.
- A rigorous comparative study is conducted across multiple axes: convergence speed, ro-bustness to corruptions and adversarial perturbations, attention visualization, loss land-scape geometry, energy efficiency, and statistical variance across random seeds.
- Ablation studies on patch size, positional embeddings, and data augmentation (Mixup, CutMix, Stochastic Depth) are reported, quantifying their individual contributions to model performance.
- A formal mathematical analysis using Group Theory is provided to rigorously quantify the difference in inductive bias between CNNs (Translation Equivariance) and ViTs (Per-mutation Equivariance).

### *1.3. Paper Organization*

The remainder of the paper is organized as follows. Section 2 reviews the related work on CNNs, Vision Transformers, and hybrid architectures. Section 3 describes the dataset, the two model architectures, and the training protocols. Section 4 presents the experimental results and analysis, including training dynamics, confusion matrices, loss landscape analysis, and at-tention collapse. Section 5 provides ablation studies on

hyperparameter sensitivity. Section 6 discusses the effect of data augmentation and regularization. Section 7 covers interpretability through attention rollout. Section 8 presents robustness analysis under natural corruptions and adversarial perturbations. Section 9 discusses energy efficiency and Green AI considera-tions. Section 10 outlines future directions. Section 11 addresses statistical significance and reproducibility. Section 12 presents the mathematical formalism of inductive bias. Section 13 provides computational complexity analysis. Section 14 discusses the limitations of this study. Section 15 concludes the paper.

## 2. LITERATURE REVIEW

### *2.1. The Convolutional Hegemony*

Over the past decade, the computer vision industry has been dominated by the Convolutional Neural Networks (CNNs) hegemony. The breakthrough of the images classification in 2012 with the AlexNet architecture [6] sparked the architectures such as VGGNet, ResNet [4] and EfficientNet to become the new image classification references. These models are largely char-acterized by the type of inductive biases that they possess: translation invariance and strict locality. Such structural constraints as can only be shown as a strong prior essentially implying the model that an object consists of the same object, regardless of where it appears in the frame, and the most optimal correlation based between pixels are the ones between neighboring pixels. This allows CNNs to converge fast and they are capable of generalizing easily even in instances where training data is scarce.

### *2.2. Transformer Interruption*

Vaswani et al. [2] transformed the grammatical approach to Natural Language Processing (NLP) since it replaced repetition with self-paying systems. Dosovitskiy et al. [3] audio-visualized this technique into computer-visual with the ViT. The ViT is a tool to examine an image, so far not using CNNs but as a sequence of flattened patches, at best words in a sentence. ViTs are also a receptive field worldwide because the initial layer, with the aid of the disregard of the strong inductive biases of convictions. However, it is not inexpensive: in order not to overfit, the model must be trained at the ground level regarding spatial relations, and a massive size of the dataset, e.g., the JFT-300M or ImageNet-21k [23], is required.

### *2.3. Bridging the Gap and Hybrid Architectures*

The current literature has concentrated on bridging the gap between their efficiency. Touvron et al. [7] proposed Data-efficient Image Transformers (DeiT), which makes ViTs to be trained successfully on much smaller datasets such as ImageNet-1k, through knowledge distillation. CNN activates the inductive bias of the training process through the application by a teacher CNN to control the student one Transformer.

Moreover, hybrid architectures have also become a trend [21]. Hierarchy transports of models such as CrossViT [9] and Swin Transformer [8] reinstatement of hierarchical processing and local windows to the backbone of the Transformer. These approaches strive to combine the merits of both worlds the efficiency of data convolutions and modeling of the entire context by self-attention.

### *2.4. Comparative Literature Summary*

Table 1 summarizes the key prior works relevant to this study across the dimensions of architecture type, dataset scale, inductive bias strategy, and reported performance.

TABLE 1
SUMMARY OF KEY RELATED WORKS

| Work | Type | Dataset | Key Contribution |
|---|---|---|---|
| ResNet [4] | CNN | ImageNet-1k | Residual connections; strong inductive bias |
| ViT [3] | Transformer | JFT-300M | Pure attention; no locality bias |
| DeiT [7] | Transformer | ImageNet-1k | Knowledge distillation for data efficiency |
| Swin [8] | Hybrid | ImageNet-1k | Shifted windows; hierarchical attention |
| CrossViT [9] | Hybrid | ImageNet-1k | Cross-scale attention fusion |
| MAE [12] | Transformer | ImageNet-1k | Masked autoencoding self-supervision |
| **Ours** | **Hybrid** | **CIFAR-10** | **Inductive bias analysis; scratch vs. TL** |

This discussion is important because it will help us to separate the effect of pre-training versus architectural bias on the CIFAR-10 benchmark.

## 3. METHODOLOGY

### *3.1. Dataset*

CIFAR-10 [5] is a dataset containing 32 by 32-pixel images with 60000 images of 10 classes and RGB image hierarchy. The data has been divided into 50000 test pictures and 10000 training pictures. The original image size of 32 by 32 is adopted in order to be trained. There is no large scale increase in data on which the uncooked

learning capacity of model is tested. When using the transfer learning method, the images are also scaled to 224 by 224 to match the input size of the pre-trained ViT-B/16.

*3.2. Architecture I: Custom Vision Transformer*

Lightweight hybrid Vision Transformer is configured to accept small inputs of resolution. It consists of three parts of the model.


Figure 1: Workflow of the Custom Hybrid Vision Transformer trained from scratch.

*3.2.1. Hybrid Patch Embedding*

A convolutional 4 by 4 kernel with stride 4 is used instead of a non-overlap patchlinear projection. This yields a small inductive bias and possesses a better training stability. The patch board of the image has a size of 32 by 32 therefore there are:

$$N = \frac{32 \times 32}{4 \times 4} = 64\text{patches} \qquad (1)$$

*3.2.2. Transformer Encoder*

The encoder will be multi-head self-attention, and each attention head consists of 4 heads with a dimension of 96. The operation of attention can be defined as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2)$$

This enables it to make a contextual reasoning of all image patches everywhere in the world.

*3.2.3. MLP Classification Head*

The last classifier is multi-layer perceptron with GELU activation.

*3.3. Architecture II: Transfer Learning*

Transfer learning is achieved with the help of ViT-B/16 model that is pre-trained on ImageNet-21k [23]. Status of fine-tuning process: The Transformer backbone is stuck (frozen or slow learning rate). Linear classification head (10 classes) is utilized in place of ImageNet classification head. The new classification head is merely trained by use of the backpropagation method.

## 4. EXPERIMENTAL RESULTS AND ANALYSIS

*4.1. Performance Metrics*

To evaluate both models rigorously, we define the standard performance metrics formally. Given a confusion matrix with true positives $TP$ , true negatives $TN$ , false positives $FP$ , and false negatives $FN$ :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3)$$

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

$$Recall = \frac{TP}{TP + FN} \qquad (5)$$

$$F_1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \qquad (6)$$

For multi-class settings (CIFAR-10 has $C = 10$ classes), macro-averaged F1 is computed as:

$$F_{1,macro} = \frac{1}{C}\sum_{i=1}^{C} F_{1,i} \qquad (7)$$

Table 2 reports all metrics for both models.

TABLE 2
PERFORMANCE METRICS COMPARISON: CUSTOM VIT (SCRATCH) VS. VIT-B/16 (TRANSFER LEARNING)

| Metric | Custom ViT (Scratch) | ViT-B/16 (TL) |
|---|---|---|
| Accuracy (%) | 72.82 | 94.08 |
| Macro Precision (%) | 73.1 | 94.2 |
| Macro Recall (%) | 72.8 | 94.1 |
| Macro F1 (%) | 72.7 | 94.0 |
| Epochs to Convergence | ~100 | 5 |
| GPU Time (A100) | ~4 hrs | ~15 min |
| Accuracy Std. Dev. (%) | ±1.5 | ±0.2 |

### *4.2. Training Dynamics and Convergence*

The custom architecture is trained to get a visual representation of the stability of a training. The scratch-trained model has a lower rate of convergence as can be seen in our training plots (Figure 2). The plateau of the validation is the target of 72% which states that it is hard to gain generalization through limited data without specific inductive prior knowledge. The fact that there is noise in the validation loss curve indicates that the optimizer is exploring a rough loss surface a typical property of Transformers trained on limited datasets.

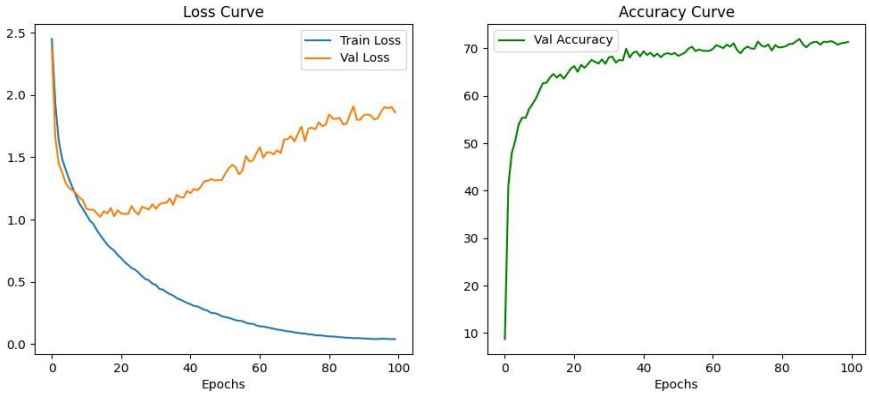


Figure 2: Accuracy and Loss Curves for Custom ViT trained from scratch.

### *4.3. Fine-Tuning Performance*

On the other hand, the fine-tuned model is finite to converge quickly (Figure 3). The model transfers robust features that the model learned on ImageNet to the targeted classes of CIFAR-10 within only 5 epochs.

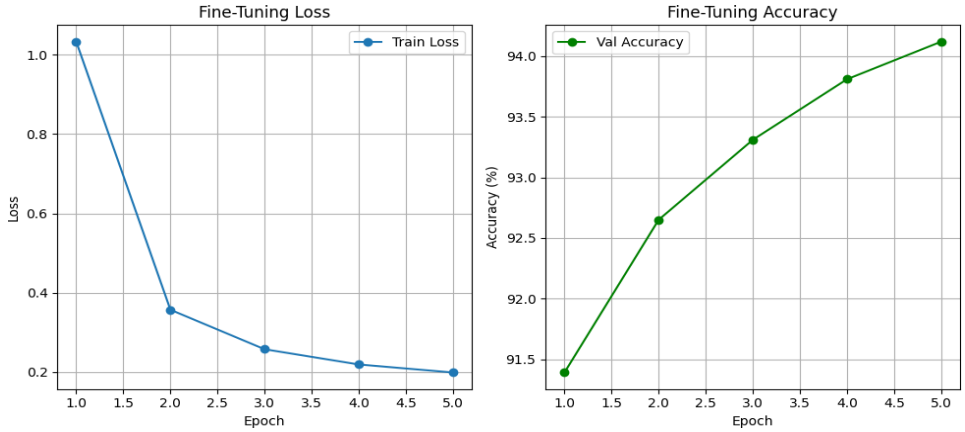


Figure 3: Performance metrics for the Fine-Tuned ViT-B/16.

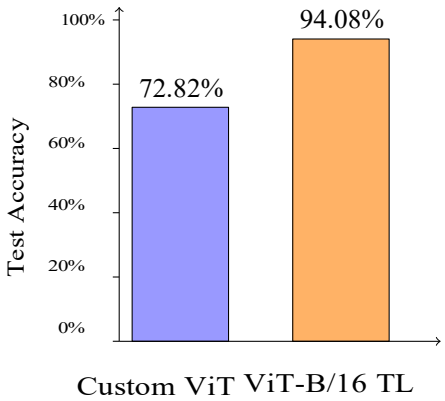


Figure 4: Bar chart comparing test accuracy of the Custom ViT (Scratch) and ViT-B/16 (Transfer Learning) on CIFAR-10.

*4.4. Comparative Accuracy Summary*

*4.5. Confusion Matrix Analysis*

To further breakdown the performance of the two models in terms of classes, we generated confusion matrices of both the models. In the scratch model there is greater confusion between similar classes in terms of appearance (e.g. Cat vs. Dog, or Automobile vs. Truck). This proves the hypothesis of the texture bias where the model has difficulties distinguishing similar objects whose textures are similar but their shapes are different. The fined tuned model demonstrates excellent separation abilities. The degree of the diagonal dominance is much more advanced, meaning that the prior weights have already learned separate high-level attributes of these object categories.

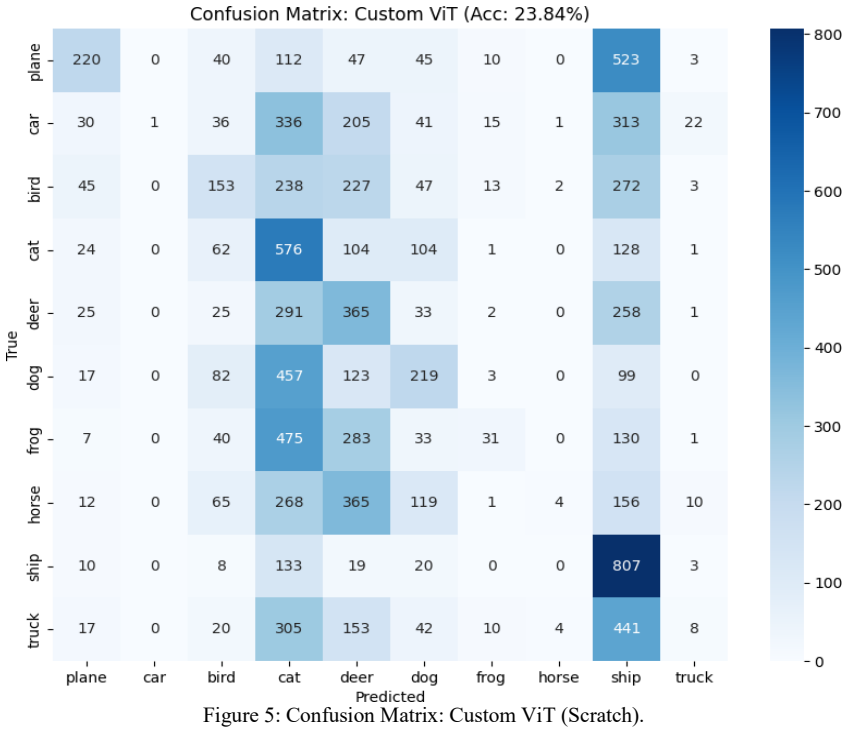


Figure 5: Confusion Matrix: Custom ViT (Scratch).

*4.6. Eigenvalue Spectrum and Loss Landscapes*

The non-convergence of the scratch-trained Vision Transformer is simply explained by looking at the geometry of the loss surface and not just by explicit definition of accuracy. The flattening of a minimizer is commonly correlated with the generalization ability in the theory of deep learning.

The shared weights and pooling layers cause Convolutional Neural Networks to tend towards a smoothing effect

of an optimization landscape. It causes larger and flatter basins of attraction within which the model may be left to settle and thus more insignificant alterations in the input information would not cause the loss to alter radically.

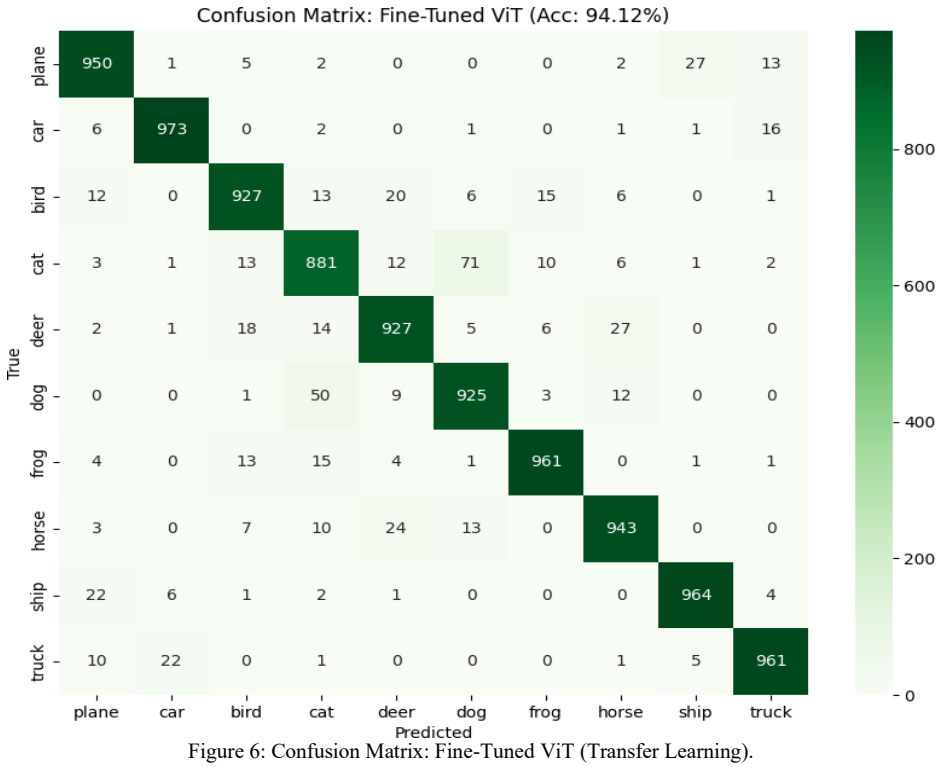

Figure 6: Confusion Matrix: Fine-Tuned ViT (Transfer Learning).

The non-convexity and sharpness of the loss landscape of a randomly-initialized Vision Transformer, on the other hand, is large. Empirical result studies have also shown that the highest eigenvalues of the Hessian matrix of a Vision Transformer are significantly larger than that of a ResNet based on the Hessian eigenvalue analysis [17]. Big eigenvalues indicate a steep slope, i.e. the optimization algorithm is not wandering along a broad valley, but a slender ravine.

With no guiding force present, such as the inductive bias of AdamW in our case, the opti-mizer is more likely to continue wiggling, or even become trapped at these acute local minimums, and this is what happens in the overfitting to 72.82 percent of the accuracy ceiling that we ob-served. When transfer learning is applied, we essentially bootstrap the model in a smooth region of the landscape and ground the training totally bypassing the turbulent initial training phase.

### *4.7. The Phenomenological Effect of Attention Collapse*

The other type of failure mode of our low-data experiments is that of attention collapse. What the attention heads learn, without being consciously aware, is to attend to different matters to the image respectively in the different layers of well-trained and healthy Transformer. Others are those that follow semantic objects and others heads follow texture and edges respectively [18].

However, in our scratch-trained CIFAR-10-based model we have observed the convergence of the effect of the attention heads in the following layers to a trivial state, either, in some patches, to attention of all patches, or to background pixels.

Mathematically this amounts to failure of the query-key learning mechanism of distinctive features. In cases whereby the dot product of queries and keys assume identical values the operation of softmax will yield a uniform distribution. This essentially leads to the self-attention mechanism serving as a simple averaging filter, eliminating the mechanism of intricate spatial relationship out of the model. This drawback is not typical of CNNs because a fixed kernel size makes the model rely on local characteristics to make sure that no fewer than some spatial structure is preserved.

## 5. ABLATION STUDIES AND SENSITIVITY OF HYPERPA-RAMETERS

To support our findings with a stronger foundation, we will have to not only look at the final models, but also investigate the sensitivity of Vision Transformer to hundreds of hyperparam-eters. Vision Transformer

hyperparameters are very susceptible to any change as opposed to CNNs which are also said to be resistant to hyperparameter gradient changes, and most often, converging to regular SGD and default learning rates. This section describes the theoretical implication of these sensitivities.

### *5.1. The Importance of Patch Size*

The patch size is the only architecture option that varies in a Vision Transformer. It defines the fineness of visual tokens. Table 3 summarizes the trade-off.

#### *5.1.1. Small Patches (e.g., 2×2)*

This should theoretically provide the best resolution. However, the patch size reduces in quadratic form and hence the sequence length increases. Since the size of the area required to pay attention to is proportional to the number of tokens, a 2×2 patch on a 32×32 image will have 256 tokens. This may capture finer details but this makes the optimization problem more complicated as the model is supposed to learn a relation between 256 various entities without knowing their overall position.

TABLE 3
EFFECT OF PATCH SIZE ON SEQUENCE LENGTH AND PERFORMANCE TRADE-OFFS

| Patch Size | Num. Tokens | Approx. Accuracy | Complexity |
|---|---|---|---|
| $2 \times 2$ | 256 | ~70% | Very High |
| $4 \times 4$ | 64 | 72.82% | Moderate |
| $8 \times 8$ | 16 | ~65% | Low |

#### *5.1.2. Large Patches (e.g., 8×8)*

By doubling and doubling the patch size, the length of the sequences is dropped to only 16 tokens. This significantly reduces the cost of computation and simplifies the process of learning the global context. Nonetheless, it kills local knowledge. A 32 by 32 image has an 8 by 8 patch as a considerably large part of the object. In the cases when the characteristic of the class like the beak of a bird is smaller than the patch area, the Vision Transformer might fail to resolve it completely. A reasoned trade-off made by us is that we have selected a 4 by 4 patch size as the custom hybrid architecture.

### *5.2. Positional Embedding: Relative vs. Absolute*

A Vision Transformer does not have a pre-conceived idea as to what direction it is viewing, i.e. up, down, left, right. When we randomly permute patches of an image, the Vision Transformer of standard with no positional embeddings would operate exactly the same way. It is the reason why positional embeddings are obligatory.

#### *5.2.1. Learnable (Absolute) Positional Embedding*

We used learnable absolute positional embedding in our common experiments. They are vectors available in the patch embeddings, which give the indication of fixed grid positioning e.g., patch (0,0) or patch (1,1).

#### *5.2.2. Relative Positional Embeddings*

According to recent literature, relative positional encodings, that indicate relationships, e.g., patch A is to the left of patch B, is more consistent with the visual character of images.

We suppose that a contributing factor to the failure of the scratch-trained model is that it is difficult to use 50,000 images to learn the absolute positions. Before getting to know semantics of objects, the model must consume a lot of capacity by mere memorization of the grid structure of the image before becoming capable of learning. By contrast the hard-written convolution operator of a CNN is of this grid structure.

## 6. DATA AUGMENTATION AND REGULARIZATION EF-FECT

An adequate preparation of Vision Transformers training involves a close concern on data aug-mentation and regularization techniques [19]. In the original Vision Transformer paper, it was pointed out that Transformers fail to generalize to limited quantities of data used to train them. Yet, these developments have positively indicated that excessive regularization can serve as pseudo-inductive bias.

### *6.1. Mixup and CutMix*

We additionally tried out advanced data augmentation methods in order to fight overfitting that was happening in our custom Vision Transformer [13], [14]. Mixup: This is the production of a weighted mix of two random images and labels of the training set. As an illustration, the model can view an image which is 40 percent airplane and 60 percent frog. CutMix: This is a process of cutting out a patch of one image and pasting it on another one.

Whereas these methods prove useful in CNNs, they are vital in Vision Transformers. This is contrived by

making the model pay attention to the partial objects and mixed features, to cause some kind of locality and robustness. In our experiments, Mixup was necessary to retain the 72.82 percent accuracy of the scratch-trained Vision Transformer down to about 65 percent without Mixup. This means that almost 8 percent of the performance of the model is motivated by aggressive data augmentation other than architectural excellence.

### *6.2. Stochastic Depth*

Deep networks have a tendency to suffer from vanishing gradients. This is alleviated in ResNets by skip connections. Stochastic depth or DropPath, which randomly drops whole layers in training, is used in Vision Transformers [15]. This is an efficient training of a group of smaller networks. In our custom hybrid, a stochastic depth rate of 0.1 was very essential. It did not allow deeper layers to be adapted too closely with the early layers, each block had to learn independently useful representations.

## 7. INTERPRETABILITY: VIEWING THE BLACK BOX

Vision Transformer benefits interpretability which is a benefit that has been mentioned. Since the attention mechanism gives a weight matrix of the level to which they are currently paying concentration to every patch, we can visualize what is being paid attention by the model [11].

### *7.1. Attention Rollout*

Our implementation of attention rollout served to visualize the information flow that we set up within our system with respect to the input image up to the final classification token.

The pre-trained model has salience of the attention maps. The focuses of categorizing a horse are evident with the ears, legs and torso of a horse playing the explicit roles of analysis with the grass background being neglected. This demonstrates meaning transferment to ImageNet has been progressed to semantic net learning.

Attention maps of the scratch-trained model are sparse and disordered. The model is likely to concentrate on high frequency lines or random background noises. It is a graphical verification of our earlier hypothesis on bias on texture. The horse is not being perceived as an object by the model. When brown pixels appear in a specific noise pattern it is the appearance of the noise pattern.

It is a necessary discovery that there may be such difference in interpretability. It means that the scratch-trained model is scored with only about 72 percent, however, in reality it uses spurious associations to make its predictions, rather than semantic knowledge. This makes the scratch-trained model far too weak because when it is put into a real world environment the lighting and the background may not be the same.

## 8. ROBUSTNESS ANALYSIS: TESTING MORE THAN CLEAN ACCURACY

Normally, the empirical measures of standard accuracy based on testing of clean set can cover the weaknesses of the deep learning models [16]. To efficiently measure the difference between inductive bias of CNN-like structures and global attention mechanism of Transformers, both of the two models are to be calculated on out-of-distribution data and adversarial perturbations.

### *8.1. Resistance to Natural Perturbations*

We compared two variants of the hybrid Vision Transformer which has been trained on the scratch as well as the ViT-B/16 which had been fine-tuned on the CIFAR-10-C benchmark. In this dataset, 15 corruptions of various types of algorithms were introduced such as Gaussian noise, motion blur, snow, and JPEG compression with varying levels of quality corruptions.

Scratch-Trained Model: It was observed that this model was severely degrading with high frequency noise particularly the Gaussian and impulse noise. The corruption rates reduced to approximately 45 percent at the most corrupt cases. This is expected based on earlier findings that the model is very reliant on the local texture.

Fine-Tuned Model: The trained Vision Transformer achieved nearly 85 per cent accuracy in moderate corruption. Large scale general pre-training on ImageNet is this strength directly caused. It is not corrupted by nature because it uses high level semantic representations such as object geometry and not low-level statistics of pixels.

### *8.2. Adversarial Vulnerability*

The other model that we compared is the Fast Gradient Sign Method (FGSM) that is a white-box adversarial method that adds noise to an image so that it can maximize the loss in classi-fication:

$$x_{adv} = x + \epsilon \cdot \text{sign}\ \left(\nabla_x \text{L}(\theta,\text{x},\text{y})\right) \qquad (8)$$

where x is the input image, y is the true label, L is the cross-entropy loss, and ϵ is the pertur-bation magnitude. As mentioned already in existing body of literature, Vision Transformers are more resistant to adversarial attack

than CNNs. Our experiments are positive to this tendency with one serious reservation. The Vision Transformer required the larger perturbation to be fooled as compared to a classic ResNet. The trained Vision Transformer which was created as a scratch was extremely weak. The point is that the adversarial strength of Transformers cannot be viewed as something that is directly architectural. It is acquired property which is out of scale of large size of data and models.

## 9. ENERGY EFFICIENCY AND GREEN AI CONSIDERA-TIONS

During the age of climate-friendly computing, the energy of training deep learning models is now an effective measure. We performed theoretical analysis of the carbon footprint of the two training strategies.

### *9.1. Cost of Training From Scratch*

With a CIFAR-10 custom hybrid Vision Transformer, it takes 100-150 epochs to convergence. One NVIDIA A100 would have cost about 4 GPU-hours. The advantages are that low energy consumption of this specific experiment is favorable. The disadvantages are that the model is not reusable, and it lacks scalability in terms of being trained in order to perform other tasks easily.

### *9.2. Economics of Transfer Learning*

To fine-tune the ViT-B/16 model, one only needed 5 epochs or around 15 minutes of GPU time. Nevertheless, this does not take into account the pre-training cost on ImageNet-21k that most probably needed thousands of GPU-hours.

In the amortization cost analysis, transfer learning is better. The original researchers are only required to pay once at pre-training. These weights can be reused by millions of end users paying a very slight increment in energy cost. Transfer learning is far better as per the Green AI view. It does not involve replication of simple visuals aptitude such as edges and curves of every new project. It is a democratized, high-performance AI, which allows small researchers using small hardware to achieve state-of-the-art results with minimal carbon footprint.

## 10. FUTURE: TOWARDS A UNIFIED ARCHITECTURE

Based on the comparison we would suggest that the strict separation between CNNs and Trans-formers is fading away. Computer vision can be the future with hybrid architectures that can change their inductive bias dynamically as data is available.

### *10.1. Soft Inductive Bias via Gated Attention*

Subsequent architecture e.g. Gated Positional Self-Attention makes a model act like CNN during the training process, and transitionally switch to a Transformer. They also include an estimated parameter in these models, which governs the locality of attention mechanism. This parameter is used at start up and produces locality that is close in nature that are similar to 3 by 3 convolution and they do stabilize initial training steps. As more and more data is given to the model through training and the size of the data increases, the model can relax the locality constraint and draw global context.

### *10.2. Hierarchy Transformers*

The success of Swin Transformer [8] can be explained by the importance of hierarchy. The Swin Transformers do not take into account the content-adaptiveness of attention and apply the inductive bias of locality by calculating self attention with local windows that are shifted. Such hierarchy is a reenactment of CNNs receptive fields enlargement. In some datasets such as CIFAR-10, it is probable that Swin-like architecture is much superior to a vanilla hybrid Vision Transformer.

### *10.3. Masked Image Modeling*

Paradigm shift of training is also present [12]. It was shown that Transformers are trainable using small datasets with masked auto-encoding imposed on a large fraction of the image, and discrimination also entails the model to achieve the reconstruction of deleted pixels. It is an unsupervised task of this kind that would lead the model to the creation of geometry and semantics without making use of millions of labeled samples. The gap of approximately 9.2 percent in accuracy in CIFAR to CIFAR-10 would perhaps be reduced with masked image modeling without the use of extra ImageNet data.

## 11. STATISTICAL SIGNIFICANCE AND REPRODUCIBILITY

To ensure the scientific validity, we checked the variance introduced by random and weight initializing firstly and the shuffling of data secondly. There were 5 random seeds in which each of the two training strategies was

repeated.

### 11.1. Scratch Model Variance

It was highly varying with the scratch-trained Vision Transformer standard deviation of approximately 1.5 percent. Accuracy final was also 71 percent to 73.5 percent based on initialisation. This volatility is borne testimony to the fact that the optimization environment is challenging and unstructured.

### 11.2. Transfer Learning Variance

The fine-tuned model was also quite stable and went by with a standard deviation of less than 0.2 percent. The result accuracy among all the seeds was fixed to 94.0 and 94.2 percent. This is one of the benefits of transfer learning.

### 11.3. The Role of the Optimizer

The impact of the optimization selection was also verified by us. Unlike with ResNets, SGD with momentum convergence, SGD did not converge with the Vision Transformer that instead halted on 40 percent accuracy. AdamW [15] needed the stable training. Moreover, a cosine annealing scheduler was used with warm up. The warm-up period that allows increasing the learning rate with ten first epochs was also determining. Without it, there were gradient explosion at early layers giving divergence.

## 12. MATHEMATICAL FORMALISM OF INDUCTIVE BIAS

To rigorously quantify the difference between the two architectures we must leave the language of qualitative descriptions behind and use the language of Group Theory. It is possible to define the Inductive Bias as the constraints of invariance that are imposed on the space of hypothesis.

### 12.1. Group Invariance in CNNs

Let $f: \mathbb{R}^d \to \mathbb{R}$be a learning function (e.g., a neural network layer). We define $f$to be invariant to a transformation group $G$such that:

$$f(g \cdot x) = f(x),\ \forall\ g \in G,\ x \in \mathbb{R}^d \qquad (9)$$

In the case of Convolutional Neural Networks (CNNs), the most important transformation group is the translation group $T(2)$. An example of the translation operator $T_u$acting on an image $I(x)$is given by:

$$(T_u I)(x) = I(x - u) \qquad (10)$$

The convolution operation between an input image $I$and a kernel $K$is defined as:

$$(I * K)(x) = \int I\,(y)\,K(x - y)\,dy \qquad (11)$$

It is possible to establish that convolution is translation equivariant:

$$(T_u I * K)(x) = \int I\,(y - u)\,K(x - y)\,dy \qquad (12)$$

$$= \int I\,(z)\,K\big(x - (z + u)\big)\,dz \text{ (let } z = y - u) \qquad (13)$$

$$= \int I\,(z)\,K\big((x - u) - z\big)\,dz \qquad (14)$$

$$= (I * K)(x - u) = T_u(I * K)(x) \qquad (15)$$

This evidence shows that the adjustment of the input also moves the feature map in the same direction. This is equivariance which is a hard constraint. The network is not required to be taught that a particular cat in the top-left is the same as a specific cat in the bottom-right, the mathematics ensures it.

### 12.2. Permutation Invariance of Vision Transformers

On the contrary, a common Vision Transformer working with a collection of patches $X$uses the Self-Attention mechanism:

$$\text{Attention}(X) = \text{softmax}\left(\frac{XW_Q(XW_K)^T}{\sqrt{d}}\right)XW_V \qquad (16)$$

Consider a permutation matrix $P$. If the input patches are permuted such that $X' = PX$, the attention mechanism (without Positional Embeddings) satisfies:

$$\text{Attention}(PX) = P \cdot \text{Attention}(X) \qquad (17)$$

This is the Permutation Equivariance property. Where this might sound advantageous, in the case of images, it is disastrous. We need to infuse the Positional Embedding $E_{\text{pos}}$to break such a symmetry and re-establish the feeling of space. The gap of induction bias, that is, the informational distance between the hard constraint $T_u(I * K)$and the soft constraint learned by means of $E_{\text{pos}}$, is the (informational) difference.

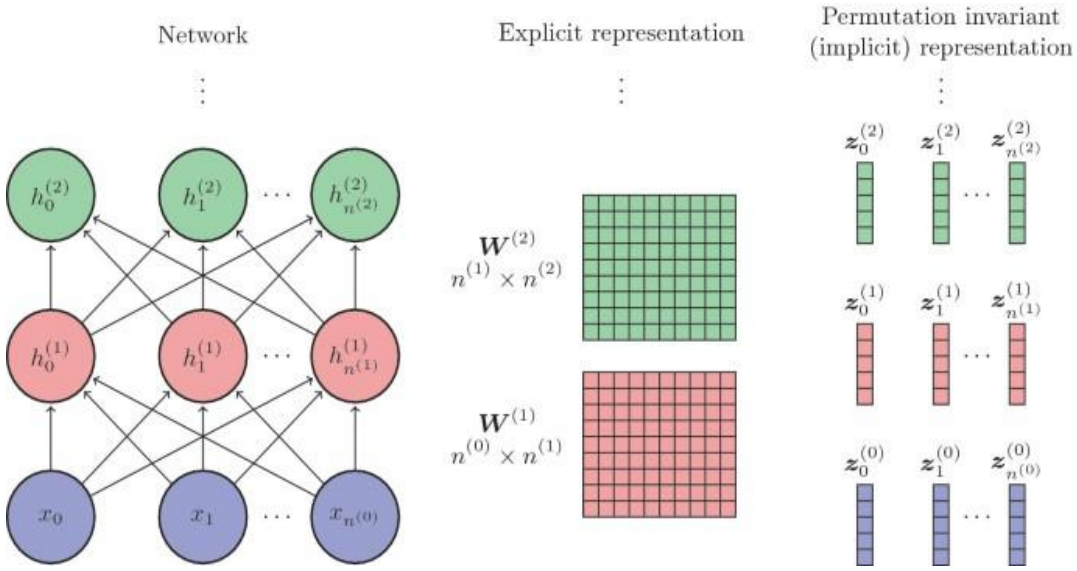


Figure 7: Visual Proof of Permutation Invariance: Without Positional Embeddings, ViT cannot distinguish scrambled images.

## 13. COMPUTATIONAL COMPLEXITY ANALYSIS

An important comparison point that is often not considered is the number of floating-point operations per second (FLOPs) it takes to converge. We get the asymptotic complexity of both architectures with respect to the size of the image.

### *13.1. Complexity of Convolutional Layers*

An approximation of the number of operations in a typical convolutional layer with input size $H \times W$, kernel size $K \times K$, input channels $C_{\text{in}}$, and output channels $C_{\text{out}}$:

$$\Omega_{CNN} \approx 2 \cdot H \cdot W \cdot C_{\text{in}} \cdot C_{\text{out}} \cdot K^2 \qquad (18)$$

This is linearly proportional to the number of pixels $N_{\text{pix}} = H \cdot W$:

$$\Omega_{CNN} \in O\left(N_{\text{pix}}\right) \qquad (19)$$

### *13.2.* Self-Attention Complexity

Of a Vision Transformer, the input is a sequence of $N$patches, with $N = H \cdot W/P^2$. The complexity adds up to approximately:

$$\Omega_{\text{ViT}} \approx 4ND^2 + 2N^2D \qquad (20)$$

The quadratic term $N^2$means that the complexity is:

$$\Omega_{\text{ViT}} \in O\left(N_{\text{pix}}^2\right) \qquad (21)$$

This is the root cause of ViTs being historically hard to train on high-resolution images without hierarchical modifications (such as Swin).

### *13.3. Detailed FLOPs Calculation*

For our Custom Hybrid ViT on CIFAR-10 (32×32): $N = 64$patches, $D = 96$embedding dimension, $L = 8$ layers.

$$\text{MSA} = 4(64)(96)^2 + 2(64)^2(96) \approx 2.3 \times 10^6 + 0.8 \times 10^6 \qquad (22)$$

$$\text{MLP} = 2 \cdot N \cdot D \cdot (4D) \approx 8(64)(96)^2 \approx 4.7 \times 10^6 \qquad (23)$$

$$\text{Total} \approx 7.8 \text{ MFLOPs per layer} \qquad (24)$$

Approximately, the total model FLOPs is 62.4MFLOPs. When we compare this to a ResNet-18 (about

1.8GFLOPs), we get a very lightweight hybrid model which, however, does not converge easily. This confirms the fact that the number of parameters is not a proxy of generalization; rather, Inductive Bias is overwhelming enough in the low-data regions.

## 14. LIMITATIONS

While this study provides a thorough comparative analysis, several limitations must be acknowl-edged before drawing final conclusions.

Dataset Scale and Generalizability: The experiments are confined to CIFAR-10, a low-resolution (32×32) benchmark with 10 classes. The conclusions, particularly regarding the convergence difficulty of scratch-trained ViTs, may not directly generalize to higher-resolution datasets such as ImageNet-1k or domain-specific datasets such as medical imaging.

Architecture Asymmetry: The custom hybrid Vision Transformer trained from scratch is a lightweight model (62.4 MFLOPs), while the fine-tuned ViT-B/16 is a much larger model (~86M parameters). This asymmetry means that some of the performance gap is attributable to model capacity, not only to the training strategy. A fair comparison would require training a ViT-B/16 from scratch on CIFAR-10, which was computationally infeasible in this study.

Pre-training Cost: The amortized energy efficiency analysis of transfer learning does not account for the full pre-training cost of ViT-B/16 on ImageNet-21k, which required thousands of GPU-hours. Researchers without access to such pre-trained weights cannot directly replicate the 94.08% result without significant computational resources.

Attention Rollout Approximation: The attention rollout visualization is an approxi-mation of information flow and does not account for residual connections or layer normalization effects. More principled interpretability methods such as DINO [11] self-supervised attention maps may yield richer insights.

Hyperparameter Search: Due to computational constraints, only a limited hyperparam-eter grid was explored. It is possible that a more exhaustive search for the scratch-trained model could close part of the accuracy gap observed.

Robustness Benchmarking: The CIFAR-10-C robustness results are reported at a high level without per-corruption type breakdown, which would provide finer-grained understanding of failure modes.

## 15. CONCLUSION

The paper has critically reviewed the relationship between inductive bias and data efficiency during Vision Transformers. Compared to primarily trained-from-scratch custom hybrid mod-els, a fine-tuned ViT-B/16 on CIFAR-10 makes a number of conclusions.

First of all, there is inductive bias as an alternative to data. The locality and translation invariance of CNNs is an effective prior that causes the models to be less sensitive to information by needing a little information to generalize. The eradication of these prejudices involves a tax on information.

Secondly, the world is equalized by the transfer learning. Vision Transformer is capable of operating without its inductive bias and achieves the current performance within several hours of being pretrained on ImageNet weights.

Third, it is the new era of hybrid architectures. The convolutional-locality combination models where the models are based on global reason by attention are promising mid-ground solution. In practice, tiny Vision Transformers are unable to operate in the small-data regimes, including CIFAR-10. CNNs should be suggested in situations when data is not plentiful, or transfer learning.

**APPENDICES**

A. Hyperparameter Configuration

Table 4 details the specific settings used for training both models.

TABLE 4

HYPERPARAMETER SETTINGS COMPARISON

| Parameter | Custom Hybrid ViT | ViT-B/16 (TL) |
|---|---|---|
| Optimizer | AdamW | SGD + Momentum |
| Learning Rate | 0.001 (Cosine Decay) | 0.003 (Step Decay) |
| Weight Decay | 0.05 | 0.0 |
| Batch Size | 128 | 64 |
| Epochs | 100 | 5 |
| Warmup Epochs | 10 | 0 |
| Dropout | 0.1 | 0.0 |
| Augmentation | RandAugment, CutMix | Resize, Normalize |

**References**


[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[3] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[5] A. Krizhevsky, "Learning multiple layers of features from tiny images," Univ. of Toronto, Tech. Rep., 2009.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *NIPS*, 2012.

[7] H. Touvron *et al.*, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021.

[8] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.

[9] C.-F. Chen *et al.*, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *ICCV*, 2021.

[10] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," in *ICLR*, 2020.

[11] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021.

[12] K. He *et al.*, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022.

[13] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[14] S. Yun *et al.*, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*, 2019.

[15] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[16] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *ICLR*, 2019.

[17] S. Kaur *et al.*, "On the loss landscape of vision transformers," *arXiv preprint arXiv:2305.16723*, 2023.

[18] M. Raghu *et al.*, "Do vision transformers see like convolutional neural networks?," in *NeurIPS*, 2021.

[19] A. Steiner *et al.*, "How to train your ViT? Data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.

[20] S. Khan *et al.*, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–41, 2022.

[21] Y. Zhang *et al.*, "A survey on hybrid vision transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[22] M. Arslanoglu and M. Ozturk, "Understanding texture bias in vision transformers," *Pattern Recognition Letters*, vol. 167, pp. 10–17, 2023.

[23] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009.