ViT Inductive Bias Survey

Nicholas Sansoterra
sansoterra.2@osu.edu

Michael Pratt
pratt.336@osu.edu

Devin Lust
lust.65@osu.edu

Abstract

Vision Transformers (ViTs) have achieved remarkable success on large datasets, but often struggle to generalize on smaller datasets without extensive pretraining. In this work, we investigate three methods of introducing inductive biases into ViTs to improve their performance on limited data settings, using the Tiny-ImageNet classification task as a benchmark. Specifically, we evaluate (1) adding convolutional layers before transformer blocks, (2) enforcing locality through hierarchical attention masking, and (3) applying channel-wise splits between attention and convolutional operations. Our results demonstrate that each method can help in reducing overfitting on small datasets. These findings highlight the potential of structural modifications to enhance ViT performance in small-data regimes without reliance on pretraining.

1 Introduction

We are going to test and compare three methods of introducing inductive biases into ViTs. The goal is to try and improve performance of ViTs on small datasets, without prior pretraining. We will do our research on the Tiny-ImageNet classification dataset.

- **Conv Layers before the transformer block:** To reduce overfitting, [5] used the first 3 residual blocks of ResNet to extract patches for the ViT.
- **Locality enforced attention windows:** The success of the Swin Transformer [4] has showed that enforced locality can be helpful.
- **Channel-wise Splits:** Inside of each transformer block, we can split the input so that some channels go to attention operations, and others go to conv operations. This was an effective method of introducing inductive biases in [1].

2 Project Motivation

While Vision Transformers have become competitive on large-scale datasets, Convolutional Neural Networks still tend to outperform ViTs in several scenarios, especially when data or computation is limited. We want to find ways to apply the benefits of ViTs in areas where they would traditionally have more trouble than CNNs.

3 Project Plan

3.1 Baseline approach

Our experiments used two baseline benchmarks for comparison. The first benchmark was on an untrained ResNet-18[3] architecture. The second benchmark was an untrained ViT, as described in [2]. Due to the limited data, we expected this ResNet to perform better than the ViT at inference on unseen data. This expectation held true as we reported a consistent 1% increase in top-1 accuracy when using ResNet over a ViT.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Table 1: Baseline Approach Results

Name	Top-1 Accuracy
ResNet-18	41.20%
ViT	40.18%

3.2 Advanced Approaches

Our advanced approach consists of three modifications to the ViT model.

3.3 Advanced Approach 1: Conv Layers before the transformer block

The baseline ViT model lacks inductive biases, which prevents it from generalizing well for small datasets. However, convolution is translation invariant and localized. ResNet-18's architecture first uses convolutional layers with a fixed channel-width of 64 [3]. By adding these convolutional layers before the Transformer block, the model is able to learn these localized features that are then passed to the ViT's encoder. This helps to prevent the overfitting issues of the baseline ViT.

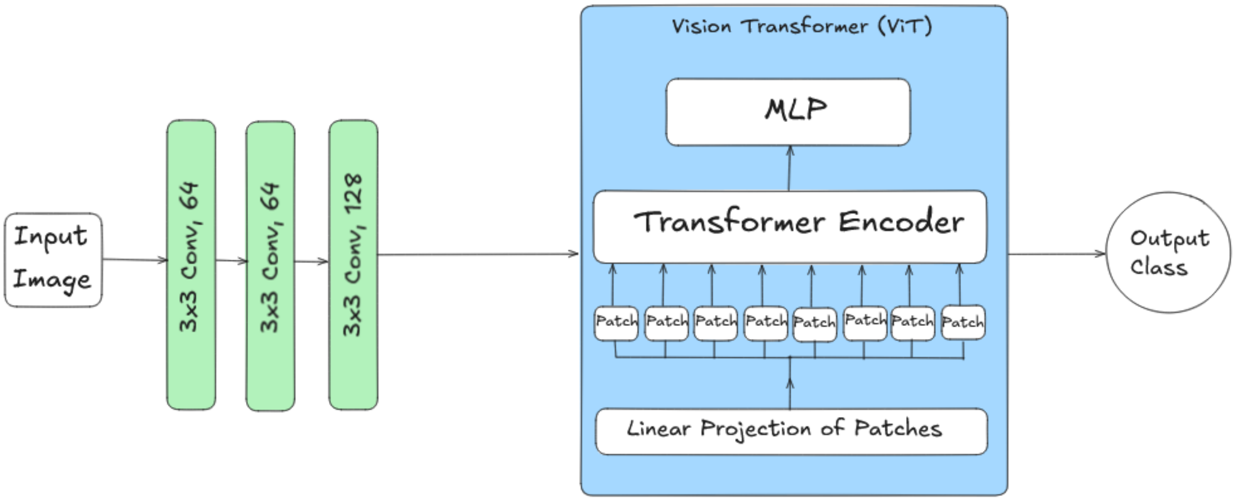

Figure 1: Early Convolution Before Transformer Block

3.4 Advanced Approach 1: Results

Early Convolution was performed using two different strategies, with varying numbers of channels in each convolution and fixed-channel width in the first layers. When the model's first convolution layers use a fixed-width of 64 channels, it achieves a marginally better accuracy to increasing channel-width. This difference reflects the intuition that more channels means better generalization.

Table 2: Early Convolution Results

Channel Hierarchy	Top-1 Accuracy
64-64-128	43.17%
32-64-128	43.10%

3.5 Advanced Approach 2: Locality enforced attention windows

Hierarchical locality-enforced attention masking can help reduce overfitting by restricting the model's focus to progressively larger but still structured neighborhoods of input tokens, rather than allowing unrestricted global attention. Early layers are encouraged to capture fine-grained, local patterns without overfitting to global noise, while higher layers can gradually integrate broader contextual information. This structured inductive bias prevents the model from memorizing spurious long-range dependencies during training, promoting better generalization to unseen data.

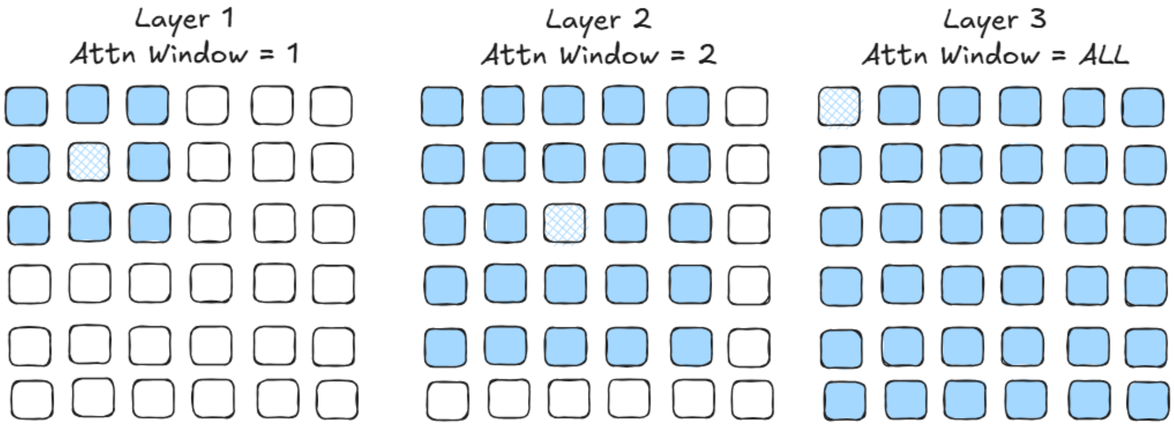

Figure 2: Hierarchical Attention-Masking Strategy

3.6 Advanced Approach 2: Results

This masking strategy seemed to have much better results when the number of transformer layers were scaled up. When starting with 3 layers, the early local layers seemed to have little help. However, when using a larger number of layers, the effects of early local attention were amplified. We believe this is due to the model needing ample time to make use of the hierarchal information.

Table 3: Hierarchical Attention-Masking Strategy Results

Name	Num Layers	Top-1 Accuracy
No Mask	3	40.18%
No Mask	6	39.92%
No Mask	8	40.12%
Local Mask	3	39.98%
Local Mask	6	42.29%
Local Mask	8	42.14%

3.7 Advanced Approach 3: Channel-wise Splits

Channel-wise splitting can help to reduce the computation required to train a transformer by a significant amount, given a significant split in favor of the convolutional layer. This simultaneously introduces inductive bias when the results are concatenated. The transformer block allows some proportion of the input to learn globally while the rest learns local information from the convolutional layer. The concatenation results in what can be thought of as an embedding with some positional encoding. This result is then put through a feedforward network that gives the final output.

A few variations of the architecture shown in figure 4 were tried, which included adding linear layers before the split and after concatenation. These architectures tended to train faster but would overfit sooner.

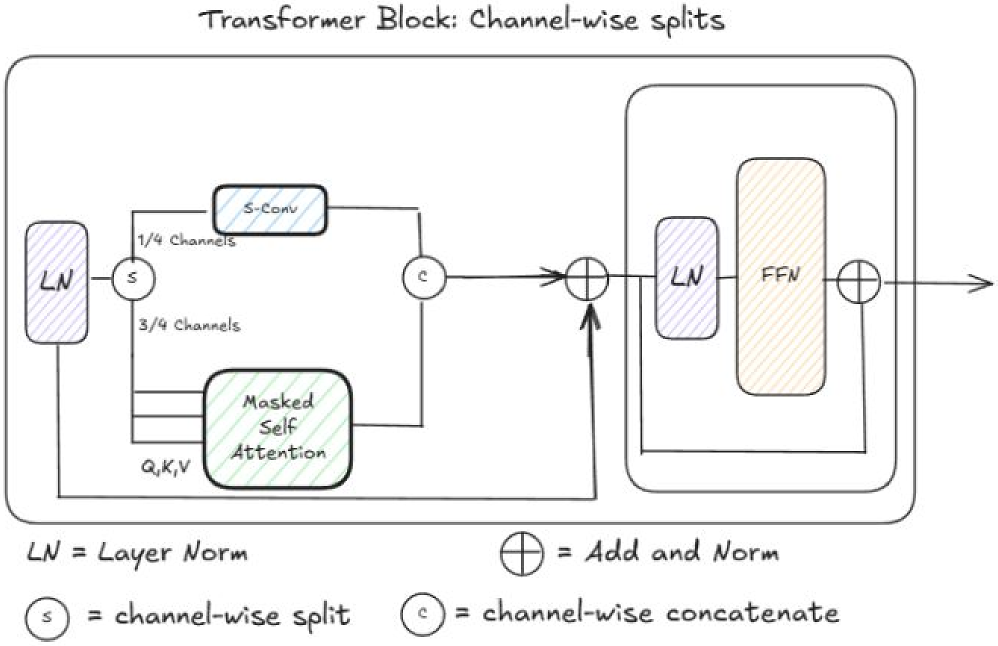

Figure 3: Channel-wise Split Architecture

3.8 Advanced Approach 3: Results

The main hyperparameter tuned in this strategy was the proportion of input to split for the transformer and the rest to go to the convolutional layer. This approach seems to perform best when a higher percentage of the input goes to the transformer block. This could be because transformers generally require a lot of data, and anything less than 50% of the input might be too little to get any gains from the split.

Table 4: Channel-wise split Strategy Results

Transformer split	Top-1 Accuracy
0.75	42.16%
0.50	41.49%
0.25	39.94%

4 Workload

The first portion of the workload was to research and get familiar with different approaches to training ViTs on small datasets. This involved reading many research papers on the approaches as well as how to set them up with PyTorch. This also involved finding baseline approaches that would be appropriate to compare our approaches with.

The actual training time for each transformer could take about 90 minutes when using the OSC resource and all the data. Each model also had their own unique hyper-parameters to tune. This resulted in many training iterations for each architecture, where the one with the most potential was picked from each approach. Random initialization was also a problem that had to be addressed. Due to limited time, the models could only train for around 100 epochs, and so random initialization could have a large impact on the already close results. Out of the best hyper-parameters, each was ran multiple times to get the best out of those.

5 Insights

The baseline ResNet-18 architecture shows that traditional convolution neural networks have inductive biases inherent to them that allow them to outperform ViTs for tasks with noisy data or small datasets, as it was able to achieve a 1.02% increase over the baseline ViT model. However, through different techniques of introducing similar inductive biases to the ViT, we were able to reduce this overfitting.

The use of early convolution produced noticeable improvement over the baseline, as it was able to achieve a 2.99% increase for its top-1 accuracy. This suggests the model was generalizing localized features, which allowed it to learn more abstract patch embeddings. Similarly, restricting the context window via attention masking was able to produce a 1.94% increase over the baseline. This suggests the model was able to learn local patterns, preventing overfitting to the noise that results from global attention. Increasing the number of layers of the attention-masking enabled additional noise reduction. Splitting the input channels between the transformer and a convolutional layer showed mixed results. For an equal split, the model was able to achieve a 1.31% increase over baseline. When an even greater percentage of the input went to the transformer block, it showed an even greater increase. However, less of a split and the model underperformed. This could reflect the need of transformer blocks to consume large amounts of data to prevent overfitting and reducing the proportion of input to be passed to the transformer reduces valuable context.

Despite the increases over baseline using varying advanced approaches, each modification to the ViT model displayed clear signs of overfitting. This could reflect on our choice of dataset. Tiny-ImageNet uses images of low resolution, which makes for noisy data. Similarly, transformers need to consume large amounts of data to learn more meaningful connections. This could suggest that future experiments in which the size and diversity of the dataset were scaled up could achieve greater improvements.

6 Conclusion

In this work, we explored three methods of introducing inductive biases into Vision Transformers to improve their performance on small datasets without relying on pretraining. Our experiments showed that each approach—adding early Convolutional layers, enforcing locality through hierarchical attention masking, and splitting channels between attention and convolution operations—provided measurable improvements over a standard ViT baseline. Among these, early convolution yielded the highest top-1 accuracy gain, while attention masking proved increasingly effective as model depth increased. Channel-wise splitting also provided benefits, though the effectiveness depended heavily on how much of the input was allocated to the transformer branch. Despite these gains, overfitting remained a persistent challenge, underscoring the inherent difficulty of training data-hungry models like ViTs on limited datasets such as Tiny-ImageNet.

While each method of introducing inductive biases individually improved Vision Transformer performance on small datasets, we were limited in our ability to explore combinations of these techniques. In future work, we would like to experiment with architectures that integrate multiple inductive bias strategies simultaneously. We believe that layering these approaches could lead to complementary effects, further improving generalization without heavy pretraining. Additionally, future experiments on larger or higher-resolution datasets could help verify the scalability of these improvements.

References

[1] Jeonghyeok Do and Munchurl Kim. Skateformer: Skeletal-temporal transformer for human action recognition, 2024.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[5] Tianyu Zhang, Longhui Wei, Lingxi Xie, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Spatiotemporal transformer for video-based person re-identification, 2021.