ResNet vs. ViT: Stability Prediction

Zhuoyang Hao
Student ID: 1255309

Jiani Xie
Student ID: 1025409

I. Introduction

Predicting physical characteristics from visual inputs is a pivotal task in computer vision, offering significant contributions to fields ranging from industrial automation to advanced robotics. The ability to accurately determine stable configurations from images, such as predicting the stable height of structures, not only enhances safety and operational efficiency but also drives innovations in design and testing environments. This report focuses on a comparative evaluation of two advanced neural network architectures, Residual Networks (ResNet) [1] and Vision Transformers (ViT) [2], in predicting the stable height from images of block structures.

The Shapestacks dataset [3], which forms the basis of our study, includes a variety of images with annotated stable heights. By leveraging this dataset, our research aims to uncover the strengths and limitations of ResNet [1] and ViT [2] models in handling complex spatial relationships and feature hierarchies present in the data.

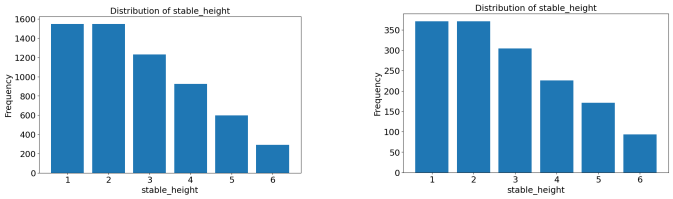
ResNet [1], known for its deep layered architecture and ability to mitigate the vanishing gradient problem, has been a staple in tasks requiring substantial depth and complexity in feature extraction. On the other hand, Vision Transformers [2], a newer paradigm adapted from natural language processing, offer a unique approach by treating image patches as sequences, potentially capturing broader contextual information across the image. The contrasting methodologies of these models provide a rich ground for analysis in terms of accuracy, computational efficiency, and model interpretability.

This report will describe the implementation and testing of ResNet [1] and ViT [2] for predicting stable height from block images. It details our experimental setup, including data preprocessing and augmentation techniques, and presents a comparative analysis of each model's performance.

II. Approach

A. Train-Validation Split

The dataset was divided into training and validation sets with an 80-20 split, allocating 80% for training and 20% for validation. This division ensures effective training and evaluation of the models, with both models utilizing the same partitioning scheme to maintain consistency in performance assessment. As illustrated in Figure 1, the training and validation sets exhibit similar distributions, with a notable bias toward labels 1 and 2.



(a) Distribution of Stable Height in Training Set

(b) Distribution of Stable Height in Validation Set

Fig. 1: Distribution of Stable Height

B. Labels and Loss Function

We define our task as a six-label classification problem and employ Cross Entropy Loss as the loss function. Since Cross Entropy Loss expects labels to start from zero, but the labels in the ShapeStacks dataset [3] start from 1, we adjusted this by subtracting 1 from each label during training.

C. Image Preprocessing

Each image is resized to (224, 224, 3). As shown in Table I, we evaluated various augmentation policies from [4]. However, the best performance on the validation set was achieved by normalizing the pixel values across the RGB channels using a mean of (0.5, 0.5, 0.5) and a standard deviation of (0.5, 0.5, 0.5), following the preprocessing method used by ViT [2]. Unlike the original approach in [2], we excluded data augmentation for the ShapeStacks dataset to preserve the physical properties depicted in the images. This normalization strategy was applied uniformly across all models and datasets.

TABLE I: Validation Accuracy and Loss for Different Policies

Policies	Val Accuracy, Val Loss
Reduced Cifar 10	Acc: 0.45, Loss: 1.31
Reduced ImageNet	Acc: 0.42, Loss: 1.33
SVHN	Acc: 0.45, Loss: 1.36
V0	Acc: 0.45, Loss: 1.46
Normalize	Acc: 0.55, Loss: 1.16

Note: All results are yielded from the same ResNet model. The first four policies are from [4].

D. Residual Networks

Residual Networks (ResNet) were introduced to address the challenge of training very deep neural networks by mitigating the vanishing and exploding gradient problem [1]. As networks get deeper, optimization becomes difficult, often leading to poor performance. ResNet solves this issue by introducing

residual connections that allow the network to learn residual mappings [1]. This ensures that gradients can flow through the network more effectively, stabilizing training and enabling better performance even in very deep architectures.

ResNet has been particularly successful in large-scale image classification tasks such as ImageNet [1], due to its ability to train deep networks efficiently. The architecture allows models like ResNet-50 and ResNet-152 to learn complex representations while maintaining high accuracy.

We selected ResNet for its ability to prevent overfitting, avoid vanishing/exploding gradients [1], and its demonstrated high performance on ImageNet. Given our task of image classification, ResNet's architecture was well-suited to the complexity of the data, allowing us to avoid common challenges associated with deep neural networks.

We experimented with RestNet-50 and ResNet-152:

- **ResNet-50:** In this experiment, we used the pretrained weight from ImageNet and modified the final layer to output 6 classes, instead of the original 1000 classes from the ImageNet dataset. This allowed us to fine-tune the model, utilizing the pre-trained feature representations from ImageNet while adapting the network to our specific task.
- **ResNet-152:** We also tested ResNet-152, a deeper variant of ResNet-50. As noted by [1], deeper models should not result in higher training error compared to their shallower counterparts, as additional layers can learn identity mappings. This experiment allowed us to assess whether the increased depth would enhance performance.

E. Vision Transformer

The Vision Transformer (ViT) brings the powerful self-attention mechanism from Natural Language Processing (NLP) into the domain of computer vision, marking a significant shift from traditional Convolutional Neural Networks (CNNs). ViT has become a state-of-the-art (SOTA) model in image classification tasks due to its ability to model global relationships between image patches efficiently. One challenge with the attention mechanism is its quadratic complexity, $O(n^2)$, with respect to the number of input tokens. To address this, ViT divides an image into smaller patches, analogous to the tokenization process in NLP, and processes these patches through an embedding layer, significantly reducing the computational load.

A distinctive feature of ViT is the use of a special classification token (CLS token), which is prepended to the sequence of embedded patches. This token accumulates information from the entire image during the self-attention process and is used as the final representation for classification tasks. The attention mechanism in ViT also enhances its robustness to occlusions, as demonstrated in [5]. This makes it particularly suitable for datasets like ShapeStacks [3], where some shapes may be occluded due to the high camera angle, posing challenges for traditional CNNs.

Given the computational constraints and the fact that ViT tends to perform better with large pretraining datasets [2], we opted to experiment with the ViT-Base-Patch16-224 model, using pretrained weights from ImageNet. This allowed us to leverage the power of transformer-based architectures while adapting the model to our specific image classification task.

F. Training Scheme

We incorporated early stopping and learning rate reduction to enhance model optimization. Early stopping monitors the validation loss, with a patience of 7 epochs, and restores the best model weights when training halts. Additionally, we applied the `ReduceLROnPlateau` scheduler, which decreases the learning rate by a factor of 0.5 if the validation loss plateaus for 3 consecutive epochs, with a minimum learning rate of 1×10^{-6}. This mechanism is applied to both models to ensure better convergence and prevent overfitting.

III. Results

A. Model Performance

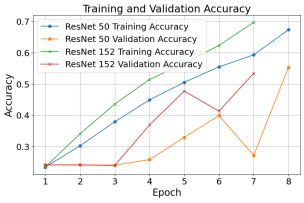
Table II presents the performance of all three models on the validation set, along with their accuracy on the Kaggle test set.

TABLE II: Model Performance

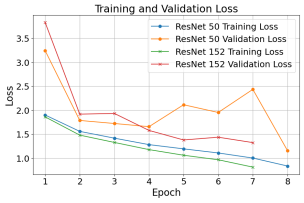
ResNet Models	Val Accuracy, Val Loss	Test Accuracy
ResNet-50	Acc: 0.55, Loss: 1.16	0.57
ResNet-152	Acc: 0.53, Loss: 1.32	0.53
ViT-Base-Patch16-224	Acc: 0.52, Loss: 1.17	0.52

B. ResNets

Figure 2 illustrates the learning curves of ResNet-50 and ResNet-152. Both variants employ the original architecture of the ResNet model as described in [1], with modifications only to the final layer after global average pooling, where the dimension is adjusted to 6 to suit our specific problem. The learning rate for ResNet-50 starts at 0.001 and is reduced to 0.0005 in the final epoch, whereas ResNet-152 maintains a consistent learning rate of 0.0005 throughout the training process. We terminated the training for ResNet-152 at Epoch 8 and for ResNet-50 at Epoch 7 due to signs of overfitting observed beyond these points.



(a) Training and Validation Accuracy



(b) Training and Validation Loss

Fig. 2: Comparison of training and validation curves for ResNet 50 and ResNet 152

C. Vision Transformers

We retained the original architecture of ViT-Base-Patch16-224, adding two additional linear layers for classification after the CLS token, as illustrated in Figure 3. The results for the ViT model, shown in Table II, were obtained after training for two epochs with a learning rate of 0.0001.

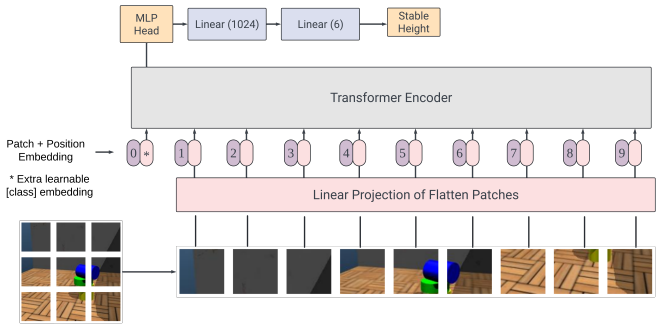
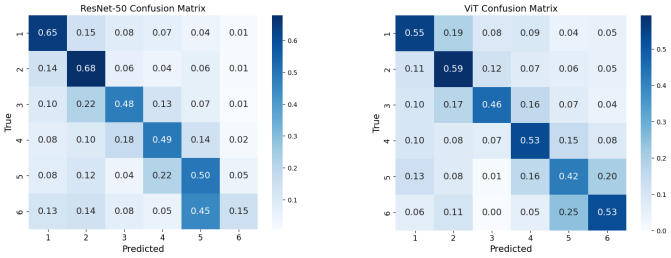


Fig. 3: Vision Transformer Architecture

Note: For demonstration purposes, the image is divided into 9 patches. The actual model architecture, as described in [2], divides a 224 x 224 image into 196 patches, each measuring 16 x 16.

D. ResNets vs. ViT

Figure 4 presents the confusion matrices for ResNet-50 and ViT. As the confusion matrix for ResNet-152 is similar to that of ResNet-50, it is not included here.



(a) ResNet-50 (b) ViT

Fig. 4: Confusion Matrix of ResNet-50 and ViT

E. Attention Maps

As shown in Figure 5 and 6, we generate attention maps by extracting the attention weights from multiple heads within each layer of the ViT. Each head captures unique patterns of interaction between different parts of the image, highlighting specific regions that the model deems important. To create a unified attention map, we average the attention weights across all heads in the same layer, providing an overall view of where the model is focusing its attention.

To gain deeper insights into how information is processed throughout the network, we recursively multiply the attention matrices across layers. This produces a joint attention map, which reflects the cumulative effect of attention from the input to the final layer. By visualizing this joint map, we can observe how the model progressively aggregates information from different parts of the image, enabling a more comprehensive understanding of the regions that influence the model's decisions. This approach allows us to trace how attention flows through the network, revealing which features and patterns are prioritized during the classification process.

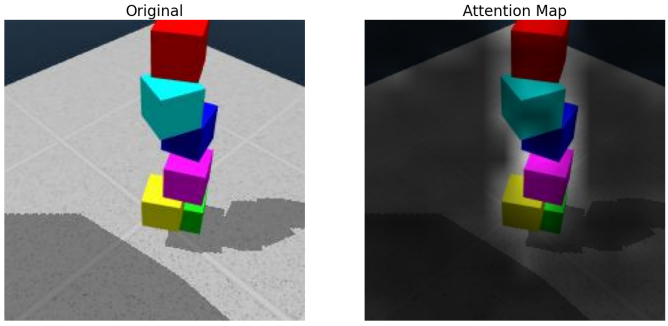


Fig. 5: Attention Example 1

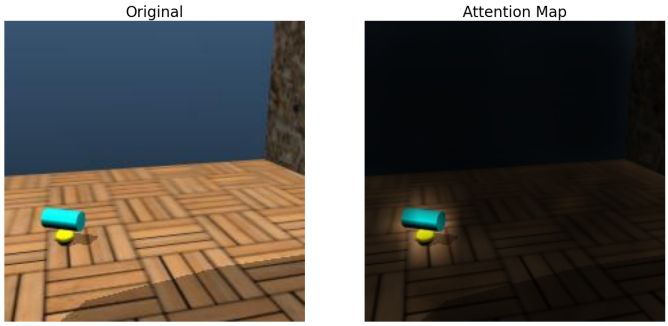


Fig. 6: Attention Example 2

IV. Analysis

A. ResNets

As illustrated in Figure 2, ResNet-152 demonstrates relatively stable convergence in both training and validation loss, employing a lower learning rate compared to ResNet-50. However, it is important to note that in our experiments with the ShapeStacks dataset, when using the same learning rate as ResNet-50 or by adding additional layers after the global average pooling, ResNet-152 tends to underperform due to underfitting. Contrary to the performance trends observed in ImageNet [1], where ResNet-152 generally outperforms ResNet-50, our findings indicate that ResNet-50 achieves higher validation and test accuracies. The final evaluation results, as documented in Table II, suggest that for this specific application, a less complex model such as ResNet-50 might be more effective. This could be attributed to its capability to adequately handle the unique characteristics and challenges of the ShapeStacks dataset without the need for excessive model complexity.

B. ResNet-50 vs. ViT

In comparing the ViTs to ResNets architectures, it's observed that ViTs require a more complex structure for processing the classification token (CLS) (Figure 3), involving additional dense layers to model complex interactions and predict probabilities. This complexity, however, enables ViTs to converge more rapidly, necessitating fewer epochs and lower learning rates. Such dynamics are attributed to the transformer's capacity to parallelly process and integrate global contextual information from all image patches, which contrasts with the sequential data processing of CNNs like ResNets. Despite their efficiency in convergence, the enhanced model capacity of ViTs necessitates careful hyperparameter tuning to mitigate risks of overfitting, particularly in less complex datasets.

Both ResNet-50 and ViT perform well overall, as observed from the confusion matrices (as shown in Figure 4), but they exhibit differences in handling certain classes. The confusion matrices reveal key differences in classification performance between ResNet-50 and ViT, further influenced by the imbalanced data distribution. ResNet-50 performs well on overrepresented classes (e.g., 65% accuracy for class 1 and 68% for class 2) but struggles significantly with underrepresented classes like class 6, misclassifying 45% of instances as class 5. This is likely due to its reliance on convolutional layers, which may be more sensitive to the data imbalance. Conversely, ViT's utilization of an attention mechanism facilitates a more uniform classification across all classes, notably maintaining a consistent 53% accuracy for class 6. The model's capability to selectively focus on pertinent features across the image allows it to effectively capture essential characteristics irrespective of class prevalence, enhancing its robustness against data imbalance. This attribute of ViT is particularly beneficial for distinguishing subtle variances between classes, providing an advantage in scenarios with underrepresented categories.

C. Attention and Inductive Bias

Figure 5 and 6 present two examples from the test set, illustrating the ViT's ability to focus on pertinent regions of the image despite variations in object size and occlusion. In the first example, although some shapes are partially obscured due to a high camera angle, ViT effectively highlights relevant areas, demonstrating its capacity to manage significant portions of the image space [5]. The second example, which occupies a smaller portion of the image area, also shows ViT's proficiency in concentrating on the essential elements of the scene.

Convolutional Neural Networks (CNNs) inherently possess natural inductive biases, such as locality, which is absent in ViTs [2]. Locality in CNNs pertains to the architecture's ability to apply convolutional filters over small, localized regions of the input (receptive fields), which is crucial for learning spatially coherent features [6]. This is particularly significant in our dataset where the shapes are in close contact. However, despite lacking this locality, ViTs, as suggested in [2], can overcome such limitations by leveraging pretraining on extensive datasets. Our results demonstrate that ViTs can adeptly handle scenes with closely interacting objects, effectively resolving the challenge of locality by focusing on relevant features across the entire image, thereby confirming their robustness and versatility in handling complex visual data.

V. Conclusion and Future Works

In conclusion, this report has rigorously evaluated the performance of Residual Networks (ResNet) [1] and Vision Transformers (ViT) [2] on the ShapeStacks [3] dataset for predicting the stable height of block structures. Our findings highlight the strengths and limitations of both architectures, underscoring their potential in handling complex visual tasks with significant implications for industrial automation and robotics.

ResNet, with its deep layered structure, excels in processing well-represented classes, leveraging its depth to extract detailed features that are spatially localized. This ability makes it particularly effective in scenarios where precision in detail is crucial. However, its performance dips with underrepresented classes, likely due to the convolutional layers' sensitivity to data imbalance. In contrast, ViT [2] demonstrates a remarkable capacity to maintain consistent performance across diverse classes, thanks to its attention mechanism that captures global dependencies effectively. This feature enables it to treat image patches as sequences, thus not only providing a comprehensive view of the image but also enhancing robustness against variations and occlusions within the visual input.

The experimental results and comparative analysis clearly show that while ResNet is more suited for tasks where data is abundant and well-distributed, ViT offers a compelling advantage in environments where the data is imbalanced or complex interactions are present. The ability of ViT to adapt to different parts of the image dynamically makes it exceptionally suitable for future applications that require understanding of intricate spatial relationships, such as in advanced robotics and complex design tasks.

Furthermore, our study paves the way for specific advancements in model architecture, particularly through the exploration of Swin Transformers [7], which integrate the locality benefits of CNNs with the global processing capabilities of traditional transformers. This approach promises to enhance model robustness and adaptability, addressing the complexities inherent in visual processing tasks. By leveraging Swin Transformers, future research could achieve significant breakthroughs in handling intricate image-based interactions within diverse technological and scientific applications.

Ultimately, the choice between using ResNet or ViT should be guided by the specific requirements of the application, including considerations of dataset characteristics and computational resources. As the field of computer vision continues to evolve, the insights gained from this comparative study will undoubtedly contribute to the development of more sophisticated and adaptable image processing technologies.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[3] O. Groth, F. B. Fuchs, I. Posner, and A. Vedaldi, "Shapestacks: Learning vision-based physical intuition for generalised object stacking," in *Proceedings of the european conference on computer vision (eccv)*, 2018, pp. 702–717.

[4] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 113–123.

[5] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang, "Intriguing properties of vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 296–23 308, 2021.

[6] Z. Wang and L. Wu, "Theoretical analysis of the inductive biases in deep convolutional networks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.