# Machine Intelligence Project 2
# Part 1: Comparing DCNN and Vision Transformer Performance

Kshitij Joshi, Yash Singla, Adi Asija, Carson James

Johns Hopkins University
May 8, 2024

## Abstract :

This research report is presented in two parts, both focusing on the CIFAR-10 image classification task. In Part 1, we compare the performance of a traditional Deep Convolutional Neural Network (DCNN), WideResNet, with a pre-trained Vision Transformer (ViT). Our results demonstrate that the ViT model achieves significantly higher classification accuracy, showcasing the potential of transformer-based architectures for image recognition.
Part 2 investigates the vulnerability of the WideResNet model to various adversarial attacks, including noise, FGSM, PGD, and C-W. We examine how these attacks affect the ability of our WideResNet model to perform image classification tasks in different ways. We also explore the need for robust defense mechanisms against such adversarial attacks. Our findings highlight the inherent vulnerabilities of DCNNs to adversarial perturbations and the potential benefits of incorporating adversarial training into model development pipelines.

## 1 Introduction

Deep learning has revolutionized computer vision, enabling machines to perform tasks like image classification with remarkable accuracy. Convolutional Neural Networks (DCNNs) have traditionally dominated this domain, achieving state-of-the-art performance on a variety of image recognition challenges. However, recent advancements have introduced alternative approaches like Vision Transformers (ViTs) that offer promising results.

This research aims to compare and contrast the performance of a well-established DCNN architecture, the Wide Residual Network (WideResNet), and a pre-trained Vision Transformer (ViT) model on the CIFAR-10 dataset, a benchmark commonly used for image classification tasks. We will evaluate both models based on their classification accuracy and computational cost to gain insights into their strengths and weaknesses.

By comparing these two distinct architectures, we hope to contribute to the ongoing exploration of advancements in image classification techniques. This will involve not only analyzing their effectiveness on the CIFAR-10 dataset but also understanding their underlying principles and potential trade-offs. Ultimately, this investigation aims to shed light on the future directions for deep learning in image recognition tasks.

## 2 Methods

In this section, we provide a detailed description of the methods employed in our DCNN and Vision Transformer, including the WideResNet architecture and fine-tuning a pre-trained transformer.

### *2.1 DCNN Architecture - WideResNet*

We adopted the Wide Residual Network (WideResNet) architecture proposed by **zagoruyko2016wide** for our image classification task on the CIFAR-10 dataset. The WideResNet is a variant of the ResNet architecture (**he2016deep**), characterized by its wider convolutional layers and reduced depth compared to traditional ResNets.

The WideResNet architecture consists of several ‘BasicBlock‘ modules, which are the fundamental building blocks of the network. Each ’BasicBlock’ comprises of two 3x3 convolutional layers, batch normalization, and ReLU activation. A key feature of residual blocks is the shortcut connection that allows the input to bypass the convolutional layers, aiding in the flow of gradients during training and enabling the training of deeper networks.

The network consists of three groups of residual blocks, with the number of blocks in each group determined by the depth parameter. The width parameter controls the scaling of the number of feature maps in each block.

The initial layer of the network is a 3x3 convolutional layer with 16 output channels. This is followed by the three groups of residual blocks, where the number of output channels increases with each subsequent group (16, 32, and 64, respectively), and the spatial dimensions of the feature maps are halved.

The final layers of the network include a batch normalization layer, global average pooling, and a fully connected linear layer with the number of output units corresponding to the number of classes in the dataset (10 for CIFAR-10).

To prevent overfitting, we incorporated dropout regularization after the ReLU activation in each residual block, with a dropout probability of 0.3. We used stochastic gradient descent (SGD) with momentum and a cosine annealing learning rate scheduler for optimization. The model was trained on the CIFAR-10 dataset with random cropping, horizontal flipping, and normalization as data augmentation techniques.

### *2.2 DCNN Training and Testing Procedure*

Our DCNN model was trained on the CIFAR-10 dataset, which includes 50,000 training images and 10,000 test images distributed across 10 classes. We trained our WideResNet model using the PyTorch deep learning framework. The model was initialized with the specified depth and width hyperparameters, and the loss function was set to cross-entropy loss. We employed the Stochastic Gradient Descent (SGD) optimizer with momentum and weight decay for optimizing the model parameters.

To further improve the training process, we utilized a cosine annealing learning rate scheduler, which gradually reduces the learning rate during training. This technique has been shown to improve convergence and generalization performance.

During training, we monitored the loss and accuracy on both the training and validation sets using the ‘train‘ and ‘test‘ functions. The model with the highest validation accuracy was saved as the best model for subsequent evaluation and testing.

After training, we implemented a function called **visualizepredictions** to visualize our testing and analyze the model's predictions on the test dataset. This function displays a grid of test images with their true and predicted labels, allowing for qualitative evaluation of the model's performance.

By following these methods, we developed a high-performing WideResNet architecture for image classification on the CIFAR-10 dataset.

### *2.3 ViT Architecture*

In addition to the WideResNet architecture, we also explored the use of a pre-trained Vision Transformer (ViT) model (**dosovitskiy2020image**) from the Hugging Face Transformers library (**wolf2019huggingface**), which was originally trained on the ImageNet dataset. We fine-tuned this model on the CIFAR-10 dataset. Vision Transformers are a relatively new class of models that apply the transformer architecture, originally developed for natural language processing tasks, to computer vision problems.

This model divides input images into 16x16 patches, encodes their positional information, and processes them through a series of self-attention layers to capture complex image features. To adapt this model to the CIFAR-10 dataset, we modified the final classification layer to output 10 classes and fine-tuned the pre-trained weights. This fine-tuning process involved updating the model's parameters using the Adam optimizer and a categorical cross-entropy loss function over multiple epochs. Validation accuracy was monitored throughout the fine-tuning process to ensure optimal performance on the CIFAR-10 test set.

### *2.4 ViT Training and Testing Procedure*

We fine-tuned this pre-trained ViT model on the CIFAR-10 dataset by initializing the model with the pre-trained weights and updating the final classification layer to match the number of classes in CIFAR-10.

During fine-tuning, we optimized the model parameters using the same training procedure as described for the WideResNet architecture, with appropriate adjustments to the learning rate and other hyperparameters. After our training, we were able to test our ViT on our test data set and compare accuracy with our DCNN model.

## 3 Results

In this section, we compare the results of the performance of our WideResNet DCNN and our Visual Transformer on the CIFAR-10 dataset.

### *3.1 DCNN Results*

**Deep Convolutional Neural Network (DCNN):** Our WideResNet architecture achieved an accuracy of **99.99%** on the training set and a test accuracy of **90.07%** on the CIFAR-10 dataset after 100 epochs of training. The series of figures below (1-4) show the training and validation loss and accuracy.

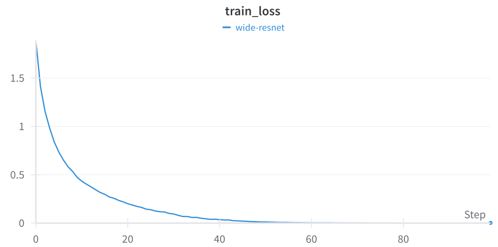

**Figure 1.** Training Loss for the WideResNet model on CIFAR-10.

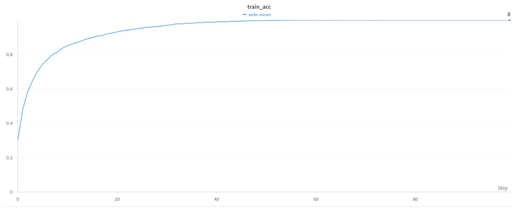

**Figure 2.** Train Accuracy for the WideResNet model on CIFAR-10.

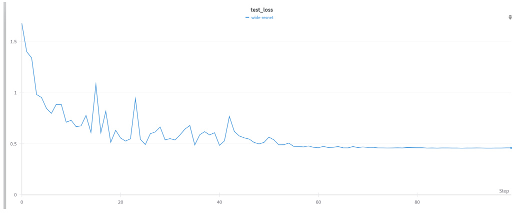

**Figure 3.** Test Loss for the WideResNet model on CIFAR-10.

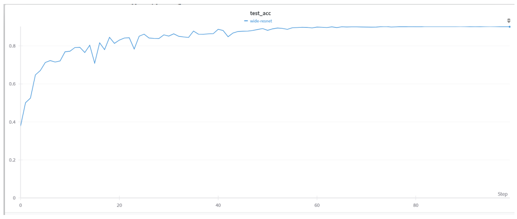

**Figure 4.** Test accuracy for the WideResNet model on CIFAR-10.

### 3.2 *ViT Results*

**Vision Transformer (ViT):** Despite the computational constraints, a smaller version of the ViT was trained, which managed an accuracy of **97.91%** on the validation set. The ViT's performance is indicative of its potential in handling image classification tasks without relying on the inductive biases of CNNs.

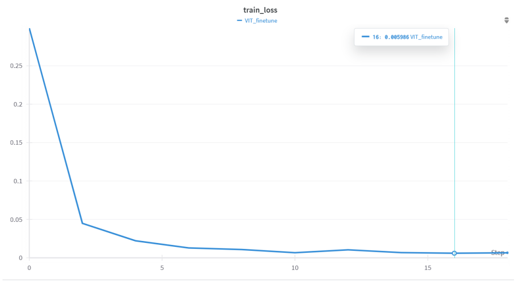

**Figure 5.** Vision Transformer Training Loss.

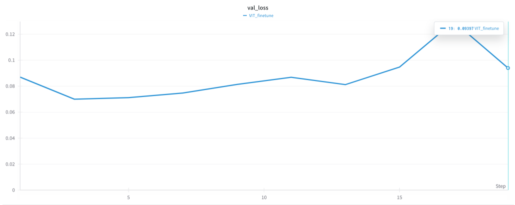

**Figure 6.** Vision Transformer Validation Loss.

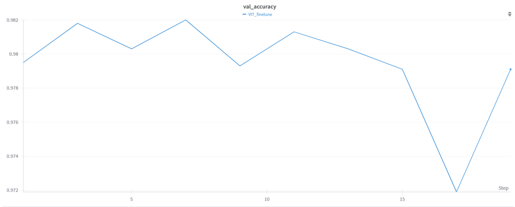

**Figure 7.** Vision Transformer Validation Accuracy.

## 4 Discussion

Our investigation into the performance of WideResNet (DCNN) and a pre-trained Vision Transformer (ViT) for image classification on the CIFAR-10 dataset yielded interesting results. The ViT model achieved a significantly higher validation accuracy (97.91%) compared to the WideResNet's test accuracy (90.007%). This observation suggests that the ViT model possesses a greater capacity for learning complex image features and relationships.

Several factors might contribute to this performance difference. ViTs, by design, excel at capturing long-range dependencies within images through their self-attention mechanisms. This allows them to learn relationships between distant image regions, potentially leading to a more comprehensive understanding of the visual content. Conversely, DCNNs primarily rely on local convolutions, which might limit their ability to capture intricate interactions between non-adjacent image patches.

Furthermore, the pre-trained nature of the ViT model could be another contributing factor. The pre-training on the massive ImageNet dataset likely equipped the ViT with a rich set of generic image features that could be readily transferred and adapted to the CIFAR-10 classification task. In contrast, the WideResNet, trained solely on CIFAR-10, might lack this initial advantage.

However, it's important to consider potential limitations of our findings. While the validation accuracy of the ViT is impressive, it's crucial to acknowledge the potential for over-fitting on the validation set. A more robust evaluation could involve employing a separate testing set not used during training or fine-tuning. Additionally, computational cost is another aspect to consider. ViTs are generally more computationally expensive to train compared to DCNNs due to their reliance on self-attention mechanisms. This could be a significant factor in resource-constrained scenarios.

# Part 2: DCNN Attack Models and Defense

Kshitij Joshi, Yash Singla, Adi Asija, Carson James

Johns Hopkins University

May 8, 2024

## 5 Introduction

Part 1 of this report established the effectiveness of a Wide Residual Network (WRN) architecture for classifying images within the CIFAR-10 benchmark dataset. However, modern deep neural networks are known to be vulnerable to adversarial attacks – crafted perturbations that can mislead models into misclassifications. Part 2 of this report investigates the robustness of the trained WRN against several adversarial attack techniques. Specifically, we will implement noise, FGSM, PGD, and C-W attacks to assess the model's susceptibility under varying noise magnitudes. We will compare how these attacks affect the classification accuracy of our WRN model and the different effects of each.

## 6 Methods

### 6.1 *Adversarial Attacks and Evaluation*

To assess the robustness of the trained WRN, a series of adversarial attacks were conducted on the CIFAR-10 test set. Four attack models were employed: Noise, FGSM, PGD, and C-W. For each attack, three different values of epsilon (.01, .03, .1) were used to evaluate the impact on model accuracy.

### 6.2 *Attack Models*

Below are explanations of the different implementations of attack models we tested.

- Noise Attack: adding random Gaussian noise to the input images with varying standard deviations corresponding to different epsilon values.
- FGSM Attack: Implemented as a single-step perturbation of the input image in the direction of the gradient of the loss with respect to the input.
- PGD Attack: Multi-step variant of FGSM, iteratively perturbing the input image in the direction of the gradient of the loss, while constraining the perturbed image to remain within a specified maximum distance (epsilon) from the original image.
- C-W Attack: Optimizes the perturbation to minimize the distance between the original and adversarial image while still causing misclassification.

For each attack and epsilon value, the classification accuracy of the model on the perturbed test set was recorded and compared to the accuracy on the original, unperturbed test set. This provided a measure of the model's vulnerability to each type of attack and perturbation strength.

## 7 Results

In this section we will compare the effectiveness of different adversarial attacks on our DCNN and compare these to the accuracy after implementing our [defense mechanism]. We will compare these results for 3 different values of epsilon.

### 7.1 *DCNN Classification Performance Accuracy Against Attack Models*

**Epsilon = .01**

- **No Attack** = 90.23%
- **Noise** = 90.16%
- **FGSM** =48.12%
- **PGD** =35.69%
- **C-W** = .13%

**Epsilon = .03**

- **No Attack** = 90.23%
- **Noise** = 90.17%
- **FGSM** = 41.52%
- **PGD** = 13.85%
- **C-W** = .13%

**Epsilon = .1**

- **No Attack** = 90.23%
- **Noise** = 88.09%
- **FGSM** = 25.68%
- **PGD** = 1.03%
- **C-W** = .13%

### 7.2 *CIFAR-10 Images After Attack Models (Epsilon = .1)*

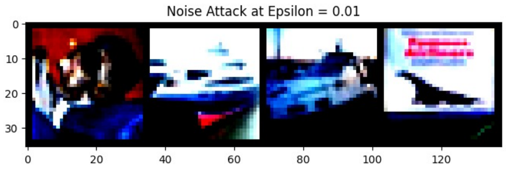

**Figure 8.** Noise Attack Image

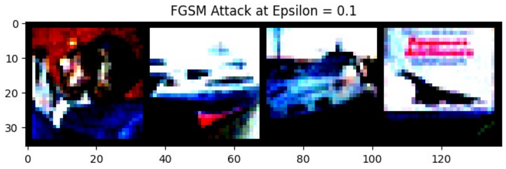

**Figure 9.** FSGM Attack Image

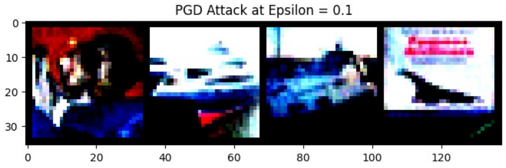

**Figure 10.** PGD Attack Image

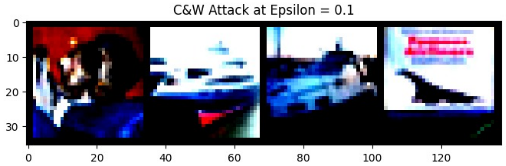

**Figure 11.** C-W Attack Image

## 8 Discussion

The results of our adversarial attack experiments highlight the inherent vulnerabilities of the WideResNet (WRN) architecture to carefully crafted perturbations. The baseline classification accuracy of 90.23% on unperturbed images was significantly degraded by each type of attack, with the degree of degradation depending on both the attack method and the perturbation magnitude (epsilon).

### *8.1 Attack effectiveness*

The C-W attack proved the most effective, reducing accuracy to near zero for all epsilon values. This demonstrates the power of optimization-based attacks to generate highly misleading perturbations. PGD, a multi-step variant of FGSM, also caused substantial accuracy drops, particularly with higher epsilon values, indicating the increased difficulty of defending against iterative attacks. PGD adjusts the attack iteratively based on the gradient loss, and the steep loss of PGD in this case represents the potent nature of higher perturbation levels. Interestingly, FGSM, while less effective than PGD, still caused significant accuracy degradation, highlighting the need for robust defenses even against simpler attacks. The Noise Attack had a relatively minor impact, suggesting that the WRN is more susceptible to structured, targeted perturbations than random corruptions.

### *8.2 Impact of Epsilon*

As expected, the effect of adversarial attacks increased with the magnitude of the perturbation. Higher epsilon values allowed for greater distortion of the input image, making it easier for the attack to manipulate the model's predictions. This underlines the importance of evaluating models across different perturbation levels to assess their robustness in diverse real-world scenarios.

### *8.3 The Need For Defense*

The substantial vulnerability of the WRN to adversarial attacks underscores the critical need for defense mechanisms to ensure the reliability of image classification models in security-sensitive applications. In our code for part b, we implemented adversarial training to our data set. Adversarial training is a defense strategy where a model is trained on both clean (original) and adversarial examples. The goal is to expose the model to a variety of potential attacks during training, enabling it to learn to recognize and resist these perturbations. Adversarial training offers a promising approach to enhance the robustness of deep learning models against adversarial attacks. However, adversarial training is not without its limitations. It can be computationally expensive, requiring additional resources and training time to generate and incorporate adversarial examples. Additionally, the effectiveness of adversarial training can vary depending on the specific attack methods used to generate the adversarial examples, and there's no guarantee of complete robustness against all possible attacks. Furthermore, adversarial training might lead to a trade-off between robustness and accuracy on clean data, as the model's focus shifts towards defending against adversarial perturbations.

## 9 Conclusion

In this comprehensive study, we explored two critical aspects of image classification on the CIFAR-10 dataset: model architecture comparison and adversarial robustness.

Our initial investigation compared a traditional Deep Convolutional Neural Network (DCNN), WideResNet, with a pre-trained Vision Transformer (ViT). The results clearly demonstrated the ViT's superior performance, achieving significantly higher accuracy due to its ability to capture long-range dependencies and leverage pre-trained knowledge from the ImageNet dataset. However, we also acknowledged the potential for over-fitting in ViTs and their increased computational demands compared to DCNNs.

We further delved into the adversarial vulnerabilities of the WideResNet model, subjecting it to various attack types – Noise, FGSM, PGD, and C-W. The significant drop in accuracy under these attacks, especially the highly effective C-W attack, underscored the inherent fragility of deep learning models to adversarial perturbations. The varying degrees of vulnerability across different attack types and perturbation magnitudes emphasized the importance of comprehensive robustness evaluation in diverse scenarios.

These findings emphasize the ongoing challenges and future directions in the field of image classification. While Vision Transformers demonstrate promising capabilities, their computational cost and potential for over-fitting warrant further research. Similarly, the vulnerability of DCNNs to adversarial attacks necessitates the development and refinement of effective defense mechanisms. Future work could involve exploring strategies like adversarial training, input denoising, or gradient masking to enhance model robustness against adversarial perturbations.

## References

1. dosovitskiy2020image: Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *arXiv preprint arXiv:2010.11929* (2020).
2. he2016deep: He, Kaiming, et al. "Deep Residual Learning for Image Recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (2016).
3. wolf2019huggingface: Wolf, Thomas, et al. "Hugging Face's Transformers: State-of-the-art Natural Language Processing." *arXiv preprint arXiv:1910.03771* (2019).
4. zagoruyko2016wide: Zagoruyko, Sergey, and Nikos Komodakis. "Wide Residual Networks." *arXiv preprint arXiv:1605.07146* (2016).
5. goodfellow2014explaining: Goodfellow, Ian J., et al. "Explaining and Harnessing Adversarial Examples." *arXiv preprint arXiv:1412.6572* (2014).
6. madry2017towards: Madry, Aleksander, et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." *arXiv preprint arXiv:1706.06083* (2017).

7. szegedy2013intriguing: Szegedy, Christian, et al. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199* (2013).