**WET PART**

**CS236781 : DEEP LEARNING**

**REPORT**

1. In this project, we focused on simple Multi-Layer Perceptron (MLP) architectures with 2-3 hidden layers, considering the addition of Dropout and Batch/Layer Normalization for regularization and stability. Our primary goal was to achieve high accuracy while maintaining simplicity and efficiency in the architecture, given the nature of the dataset.

**Reasons for Focusing on MLPs:**

1. **Initial Baseline Results:**

We started by using the default classifier provided, which included only one linear layer. After training with the Adam optimizer and using the ReduceLROnPlateau scheduler, we achieved 82% validation accuracy. This indicated that the data was largely linearly separable.

Given this result, we concluded that more complex architectures might not be necessary for this task, as a single layer was already capturing a significant portion of the decision boundaries.

1. **Supporting Literature:**

We reviewed two papers [1, 2] that applied Implicit Neural Representations (INRs) as input to neural networks for classification tasks. Both papers used simple MLP architectures and reported competitive results compared to state-of-the-art (SOTA) models.

This further motivated our decision to stick with MLPs, as the results in the literature showed that MLPs were effective for classification tasks on datasets like ours.

1. **Suitability of Other Architectures:**

Convolutional Neural Networks (CNNs): We considered whether CNNs might improve performance but concluded that they would not be appropriate for our data. Each sample in our dataset is a dense 512-dimensional vector that has already been trained on the original image and contains embedded information. Convolutions are typically used to extract spatial features from raw image data, and since our input already contains highly abstracted information, CNNs would likely perform similarly to linear layers.

Transformers and Attention: We also ruled out using transformer architectures or attention mechanisms because the data does not lend itself to the standard tokenization process used in transformers. Since each sample is a dense vector and cannot be split into patches or tokens, we concluded that attention mechanisms would not provide any significant advantages for this task.

**Our final architecture**

* Input Layer: 512 → 512 (Fully Connected) + BatchNorm + LeakyReLU + Dropout.
* Hidden Layer 1: 512 → 256 (Fully Connected) + BatchNorm + LeakyReLU + Dropout.
* Hidden Layer 2: 256 → 128 (Fully Connected) + BatchNorm + LeakyReLU + Dropout.
* Output Layer: 128 → 10 (Fully Connected)

Optimizer- AdamW with learning rate=0.001 and weight decay=0.01.

Scheduler – ReduceLROnPlateau with factor=0.1 and patience =3. The scheduler uses validation accuracy as metric.

Dropout probability- 0.1

Batch size- 64

Early stopping mechanism that stops after 5 iterations with no improvement.

The batch normalization layer uses momentum=0.05 which is more suitable for smaller batch.

Checkpoint mechanism to save the set of weights that resulted the highest accuracy on validation set.

Our final network architecture was a good choice because it provides a good balance between simplicity and performance. We used a simple MLP with 4 fully connected layers, gradually reducing the dimensions from 512 → 256 → 128 → 10.

The gradual reduction in dimensionality allows the network to progressively distill the most **Batch Normalization** stabilized training by normalizing activations, and **LeakyReLU** prevented the "dying ReLU" problem, ensuring the neurons remained active throughout training.

**Batch Normalization** stabilized training by normalizing activations, and **LeakyReLU** prevented the "dying ReLU" problem, ensuring the neurons remained active throughout training. The **ReduceLROnPlateau** scheduler helped the model fine-tune after the initial stable learning phase, reducing the learning rate to avoid oscillations or overfitting. Combined with **Dropout** and **weight decay** in the **AdamW** optimizer, overfitting was well controlled. The **early stopping** mechanism ensured that we didn't overtrain the model, stopping when there was no improvement after 5 epochs.

As we’ll explain in the next section, hyperparameter tuning was a critical part of arriving at this final design. The chosen architecture achieved the highest validation accuracy during training at **89.66%**.

In addition, we implemented a checkpoint mechanism that selects the model with the highest validation accuracy during training. After training is completed, the model is reloaded with the best weights.

**Hyper parameter tuning:**

We performed hyperparameter tuning on the following parameters:

* Learning rate: [0.001, 0.0001]
* Hidden dimensions: [ [512, 256, 128], [256, 128], [128, 128], [512,512,512], [128,64] ]
* Dropout probability: [0,0.1,0.3]
* Batch size: [32,64,128,256]
* Normalization after linear layer: [None, layer normalization, batch normalization]
* Activation: [Relu, LeakyRelu]
* Early Stopping:[5,10]

There were 1,440 possible combinations, which was too large for comprehensive training. To handle this, we used Random Search, where we randomly selected 150 combinations to train and evaluate. Each combination was trained on the training set and evaluated on the validation set. We ultimately selected the combination that achieved the highest accuracy on the validation set (as described in the final architecture section).

The maximum number of epochs was set to 500, and this parameter was not tuned. Prior to the random search, we manually experimented with different optimizers and schedulers, but none performed as well as the chosen ones, so we did not tune the optimizer or scheduler further.

We decided to use AdamW, inspired by [2] and [3], as it provided more stable results compared to the standard Adam optimizer, especially in terms of generalization and convergence.

**Explanation of the choice of parameters**-

Depth+number of neourons:

We considered several options for the depth and hidden dimensions of the neural network. Inspired by [2], we aimed to test MLP architectures with reduced dimensions in each consecutive layer. Additionally, we were interested in architectures with consistent dimensions in each hidden layer, inspired by [1].

Initially, we manually tested a few MLP architectures and observed that networks with 3-4 total layers provided the best performance. Adding more than 4 layers did not lead to any noticeable performance improvement. Based on this, we concluded that the most effective approach would be to focus on MLPs with 3-4 layers during hyperparameter tuning.

Architectures Tested During Tuning:

Reduced dimensions in each consecutive layer architectures:

* (512->512) -(512->256)-(256,128)-(128,10)
* (512->256) -(256,128)-(128,10)
* (512->128)-(128,64) -(128,10)

We chose these architectures to explore how a gradual reduction in neuron count across layers impacts performance. Our goal was to maintain enough neurons in the early layers to capture important features, while progressively reducing dimensions to distill the key information. When manually testing architectures with **more than 512 neurons**, we saw no improvement, and networks with fewer than **64 neurons** didn't perform well enough.

Same size Layers architectures:

* (512->512)-(512->512)-(512->512)- (512->512)
* (512->128)-(128,128) -(128,10)

These architectures were chosen to see if consistency across layers could improve performance by maintaining a more uniform learning process throughout the network. These architectures were inspired by [1]. The specific sizes of 128 and 512 were chosen because we wanted to try the 128 size, same as in [1], and we also wanted to mimic the 4-layer architecture used in [1] but with our input size (512).

Normalization:

We aimed to explore whether normalization applied after the linear layer could improve network performance and accelerate training. Since there wasn't much information available on whether normalization would be beneficial for the INRs dataset, we decided to test whether normalization was necessary at all, and if so, whether Layer Normalization or Batch Normalization would be more effective.

Activations:

Before conducting hyperparameter tuning, we manually tested various activation functions, including Sigmoid, SiLU, ELU, GELU, ReLU, and Leaky ReLU. However, during these experiments, we observed that only ReLU and Leaky ReLUconsistently provided good performance in terms of both accuracy and training stability.

Based on these observations, we decided to focus our hyperparameter tuning specifically on ReLU and Leaky ReLU, as they proved to be the most effective for this task.

Batch size:

We decided to consider [32, 64, 128, 256] as possible batch sizes during hyperparameter tuning. Before this, we manually tested larger batch sizes like 512 and smaller batch sizes like 16, but observed that the performance was suboptimal, and the training was unstable. Larger batch sizes led to slower convergence and reduced generalization, while smaller batch sizes introduced too much noise in the gradients, leading to instability. Based on these preliminary tests, we focused on the more balanced range of 32 to 256 for tuning.

Learning Rate:

We chose to explore two learning rates, 0.001 and 0.0001, during hyperparameter tuning. Based on initial experiments, we found that larger learning rates (greater than 0.001) caused the model to oscillate and fail to converge, while smaller learning rates (below 0.0001) resulted in very slow progress.

Dropout:

We tested [0, 0.1, 0.3] for dropout during hyperparameter tuning. 0 (no dropout) was included to see if dropout was necessary. The values were chosen based on preliminary manual tests.

Early Stopping:

We considered [5, 10] for the early stopping patience. Early stopping is used to halt training if the validation performance doesn’t improve after a certain number of epochs, preventing overfitting and saving computational resources. A patience of 5 epochs was chosen for faster stopping if no improvement was seen, while 10 epochs allowed for more training before halting, in case the model needed extra time to stabilize. This gave us flexibility in balancing between early termination and giving the model enough time to improve.

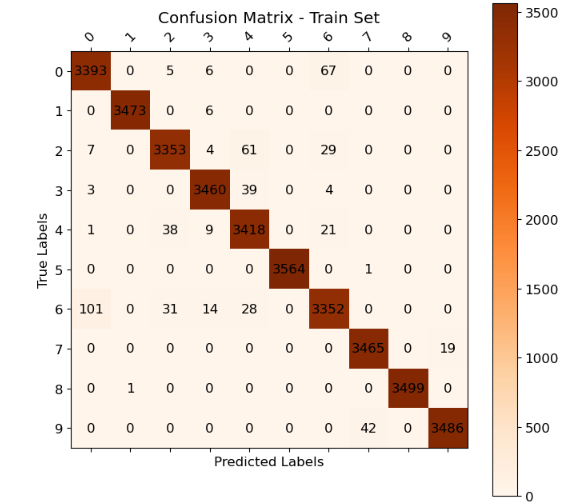
We didn’t choose a value less than 5 because our scheduler patience is set to 3, meaning the learning rate is reduced after 3 epochs without improvement. By allowing at least 5 epochs for early stopping, we ensure that the effect of the reduced learning rate can be explored before halting the training. This prevents premature stopping and allows the model a chance to improve after the learning rate adjustment.

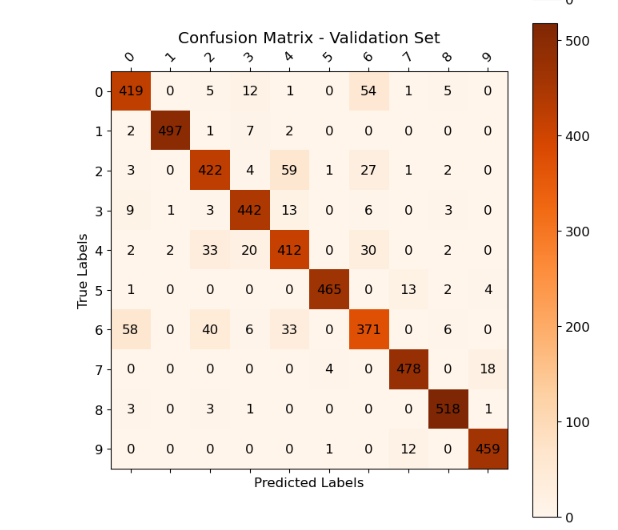
2)

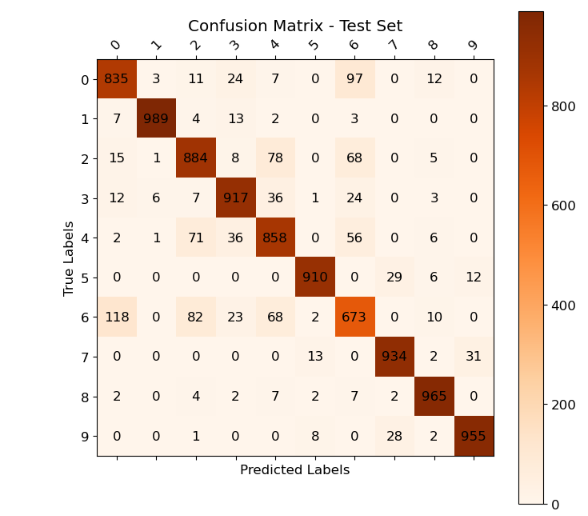
Train Accuracy: 98.47%

Validation Accuracy: 89.66%

Test Accuracy: 89.10%

3)





4) We can draw several key insights regarding the performance of the classifier across these datasets:

* The dominant diagonal across all three confusion matrices reflects that the majority of samples were correctly classified in their respective categories. This indicates that the model generally performs well across the datasets.

However, the diagonal becomes less prominent, or lighter, in the validation and test sets compared to the training set. This suggests potential overfitting.

* Major Confusions:

**Class 0 (T-shirt/top) vs. Class 6 (Shirt):** There is a confusion between these two classes, with the model often predicting class 0 as class 6, and vice versa. This bidirectional confusion indicates that the features distinguishing these two classes are not well captured by the model.

**Class 6 (Shirt), Class 4 (Coat), and Class 2 (Pullover):** There is also a confusion among these classes(in all directions). This may suggest that the model struggles with identifying visual differences between these types of clothing, possibly due to similarities in the type.

**Class 5 (Sandal) vs. Class 7 (Sneaker) and Class 7 (Sneaker) vs. Class 9 (Ankle Boot):** These confusions appear primarily in the validation and test sets, highlighting that the model may struggle with distinguishing between different types of footwear.

* The training set confusion matrix is more strictly "diagonal" compared to the validation and test sets, with fewer misclassifications. The presence of more off-diagonal elements in the validation and test sets indicates that the model has slightly overfitted to the training data.

For example, some confusions, such as misclassifying class 5 (Sandal) as class 9 (Ankle Boot), are not present in the training set but appear in both the validation and test sets.

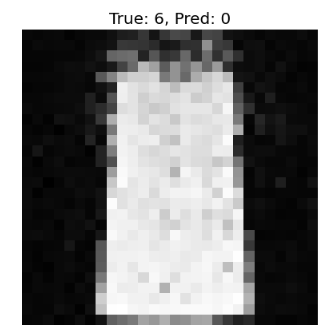
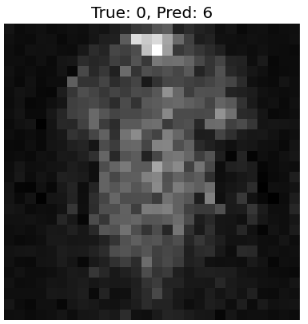
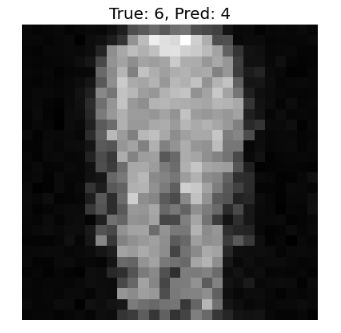
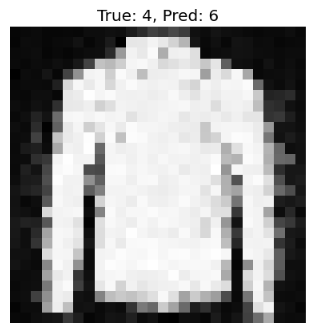
* The confusion matrix for the validation set serves as a good indicator of the model's performance on the test set, as most of the misclassifications observed in the test set are also present in the validation set.

However, there are additional confusions in the test set that are not observed in the validation set. For example, class 8 (Bag) is confused with multiple other classes (classes 2-7) in the test set, whereas in the validation set, confusion involving class 8 is more limited (mainly between classes 2 and 3).

**Improvements if we would retrain/fine tune:**

* **Filtering Misrepresented Examples:**

During the analysis, we visualized several misclassified examples using the reconstructed images from the INRs:

****

**b**

As we can observe, Some of the reconstructed images are so poor that even a human eye would struggle to identify the original clothing item. This suggests that the INR representation might not be capturing the key features needed to properly reflect the original image. Since we can’t recreate or modify the INR, one idea to improve performance would be to filter out examples where the INR’s reconstruction is clearly flawed. We could look at the reconstructions for each class, identify the ones that lead to frequent misclassifications, and consider removing those examples from the training set.

That said, not all misclassified images can be blamed on poor reconstructions. In some cases, like the example on the right, the reconstruction actually looks quite good, which suggests that the misclassification might be due to other factors.

* **Re-weighting Classes in Cross-Entropy Loss:**

Another improvement would involve re-weighting specific classes in the CrossEntropyLoss function. As observed in the confusion matrices, the classifier tends to confuse certain classes more frequently than others—particularly class 0 (T-shirt), class 6 (Shirt), class 2 (Pullover), and class 4 (Coat). Given the higher frequency of confusion in these classes, increasing the weight for these classes in the loss function would give the model a stronger signal to classify them correctly.

The PyTorch CrossEntropyLoss function includes a weight parameter that allows re-weighting the importance of different classes, and we would experiment with different weight configurations to see if this helps reduce confusions for the problematic classes.

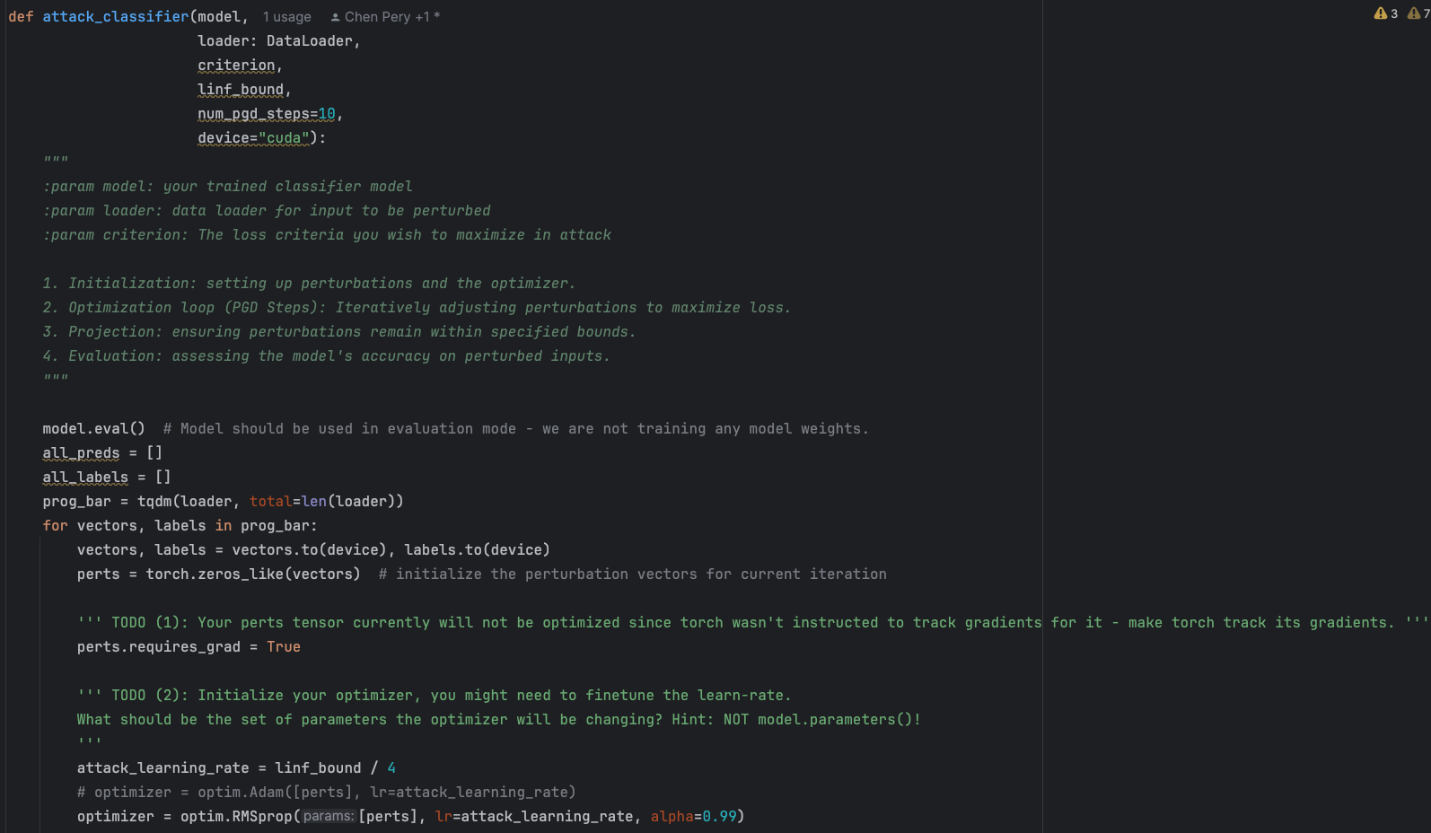
* **Reducing Overfitting:**

The confusion matrices suggest some degree of overfitting, with the model performing slightly better on the training set compared to the validation and test sets. To combat this, we could increase the dropout rate to introduce more regularization, helping the model generalize better to unseen data. Additionally, we could increase the weight decay (L2 regularization) parameter in the AdamW optimizer. This would penalize large weights, encouraging the model to avoid overfitting. Hyperparameter tuning could also be expanded to include the weight decay parameter, which we haven't fully explored before.

* **Reweightning samples**

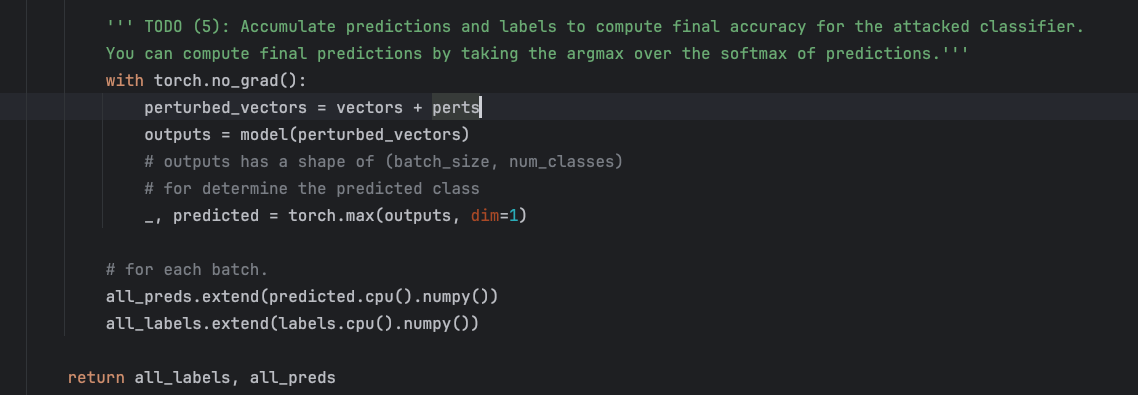
We could try reweighting some of the samples where the model consistently makes mistakes. One way to do this, for example, is by using a weighted sampler in the DataLoader, which would sample these problematic examples more frequently during training. By giving these harder samples more focus, we hope the model would learn to handle them better, potentially reducing the overall misclassification rate.

5)

Screenshot of attack\_classifier function in the attack.py:

A screenshot of a computer program

Description automatically generated



6)

A graph of a model

Description automatically generated



X-axis represents the perturbation bound (epsilon).

Y-axis shows the classification accuracy of the model after the adversarial attack.

The trend line indicated how accuracy changes as epsilon varies, from this line there is a clear inverse relationship between epsilon and the classification accuracy,

As epsilon increases the model’s accuracy decreases because larger perturbation bounds allow more substantial changes to the input features, making it easier for adversarial examples to cross the model’s decision boundaries and induce misclassifications.

As epsilon decreases the model’s accuracy increases because smaller perturbation bounds restrict the magnitude of the changes, making it harder for adversarial examples to cause misclassifications while keeping the perturbations imperceptible.

**Explanation for Choosing Learning Rate**:

In [Goodfellow et al., 2015](https://arxiv.org/pdf/1412.6572)([9]), a single-step perturbation equal to was introduced. For multi-step methods like PGD, it’s common to use a fixed step size as a fraction of to allow perturbations to accumulate without exceeding the limit. Later work, such as in [Madry et al., 2017](https://arxiv.org/pdf/2003.01690)([10]), evaluates different step sizes (e.g., ) for optimal adversarial performance across tasks. Many adversarial attack implementations use such fractional step sizes to balance effectiveness and control within multi-step attacks. For our case, we selected based on these practices and empirical performance.

A screenshot of a graph

Description automatically generated7)

A computer screen shot of text

Description automatically generated

8)

This adversarial confusion matrix highlights that the attack has successfully introduced substantial misclassifications, as seen by the increase in off-diagonal values (incorrect predictions) compared to what we would expect in a clean test set.

Certain classes, such as **true label 0**, show a strong shift in predictions toward other labels, with a significant number of samples being misclassified as class 6 (575 instances). This pattern suggests that the adversarial perturbation has altered the model’s perception of class 0, making it similar in feature space to class 6.

Other notable examples include true labels 2, 4, and 6, which show significant misclassification patterns. For instance:

* **True label 2** has 380 instances misclassified as class 4 and 436 as class 6, indicating that adversarial noise causes substantial confusion, pushing samples from class 2 toward these classes.
* **True label 4** has 402 instances misclassified as class 2 and 273 as class 6.
* **True label 6** is heavily misclassified, with 294 instances as class 0 and 270 as class 2. This consistent misclassification points to weak boundaries for class 6, making it particularly susceptible to being mistaken for these neighboring classes under attack.

This pattern suggests that adversarial perturbations lead to strong overlap between classes 4, 2, and 6.

**true label 3** shows widespread misclassification, with 111 instances incorrectly classified as class 0, 88 as class 1, 133 as class 4, and 132 as class 6. This dispersion reflects a pronounced vulnerability, where class 3 is frequently misidentified as several other classes.

**true label 1 and 8** remain relatively robust under attack; however, there are still notable misclassifications. The diagonal entries for these classes indicate a high number of correct classifications, showing that the model’s perception of these classes remains stable under adversarial perturbations to some extent.

**true labels 5, 7, and 9** demonstrate moderate resilience under attack, but they are still impacted by notable misclassifications.

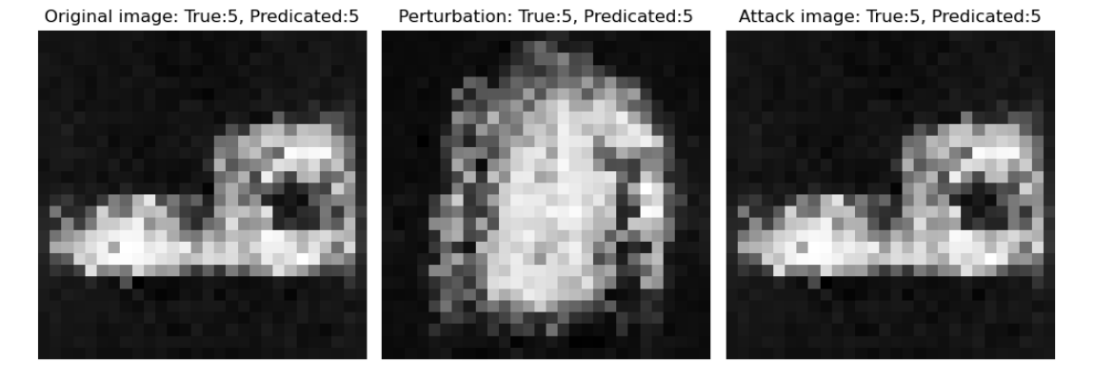
These patterns indicate that the adversarial attack is exploiting specific weaknesses in the model's boundaries, leading to consistent misclassifications in certain classes. Some classes are more vulnerable, with frequent misclassifications into neighboring classes, while others remain relatively robust, showing fewer errors

**Similarities and Differences Compared to the Clean Matrix**

In both the clean and adversarial matrices, some inherent confusion patterns are observed, particularly between classes that share similar features. For example, in the clean matrix, there are occasional misclassifications between classes like true labels 0 and 6 or 4 and 2. These mild overlaps likely arise from similarities in the visual or feature-based cues the model uses, resulting in slight misclassifications even without adversarial influence.

However, the adversarial matrix amplifies these misclassifications considerably, revealing the model’s vulnerabilities under attack. While the clean matrix shows mostly accurate classifications with isolated errors, the adversarial matrix demonstrates a drastic increase in off-diagonal misclassifications. For instance, true label 0 is heavily misclassified as class 6, and true label 3 is frequently misclassified into classes 0, 1, and 4 under attack conditions—patterns that were minimal in the clean case.

This comparison shows that the adversarial attack exploits weaknesses in the model’s decision boundaries, leading to more consistent and severe misclassifications in classes that already had slight confusion in the clean matrix. The clean matrix reflects mostly well-defined boundaries with minor overlaps, while the adversarial matrix exposes and magnifies these weak spots, making certain class boundaries more fragile and prone to errors.

****9)

A close-up of a person's face

Description automatically generated

A close-up of a ultrasound

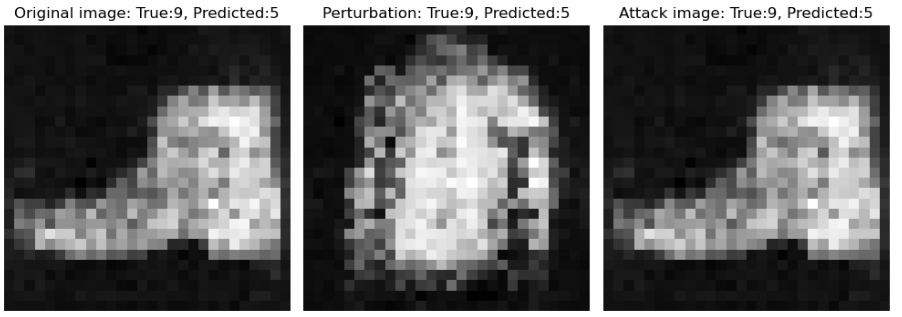
Description automatically generated

A close-up of a scan

Description automatically generated

**A close-up of a ultrasound

Description automatically generated**

****

**10. Bonus**

The approach we chose for training the classifier is adversarial training using the PGD (Projected Gradient Descent) attack to generate adversarial examples.

This approach was introduced in [4].

The inspiration for this approach comes from robust optimization, where a typical optimization problem is formulated as: .

The goal is to find a solution x that minimizes the objective function under the worst realization of u.

In the context of adversarial robustness, this framework can be applied to model training by considering perturbation attacks.

The optimization objective suggested in the paper was:

The inner maximization problem aims to find an adversarial version of a given data point x that achieves a high loss. This is precisely the problem of attacking a given neural network. On the other hand, the goal of the outer minimization problem is to find model parameters so that the “adversarial loss” given by the inner attack problem is minimized. This is precisely the problem of training a robust classifier using adversarial training techniques.

The perturbation that maximizes to create the adversarial example is generated by PGD attack, starting from random restart points within the ball around the original x.

This approach encourages robustness because the problem's formulation ensures that if we achieve a small robust loss, no allowed attack will be able to fool the model. This is because the model becomes resistant even to the worst-case adversarial attack.

However, the paper notes that solving this saddle point problem is challenging due to the non-convex nature of the outer minimization problem and the non-concave nature of the inner maximization problem. Despite this, the authors show empirically that the problem can be effectively addressed in practice by using PGD to solve the inner maximization, yielding strong robust results.

On a more intuitive level, we believe this method promotes robustness because the model is continuously exposed to adversarial examples during training. These examples push the model to learn more complex decision boundaries that are better equipped to handle subtle changes in input.

This approach has been widely adopted in the field of adversarial training.

After this paper was published, other research suggested that the formulation of the problem leads to an inherent trade-off between accuracy and robustness [5].

This has inspired the development of hybrid formulations, such as the one proposed in [6]:

In this formulation, is the adversarial example generated by the same PGD attack as before.

We implemented adversarial training in training.py, inspired by the code from [7], and the attack is implemented in attack.py. Since our classifier architecture includes batch normalization layers, the mean and variance statistics of these layers are affected by every forward pass.

To address this, we followed the approach suggested in [8]: the model is put into evaluation mode during the attack generation to prevent batch normalization statistics from being updated by the adversarial examples. Additionally, we process the model on a concatenated batch of clean and adversarial examples, ensuring that batch normalization statistics are updated based on both types of inputs in each iteration. We found that computing the model separately on clean and adversarial examples caused instability in the batch normalization layers, as the statistics of the two groups are often different, leading to inconsistent behavior.

We used the same original classifier with the same hyperparameters as described in the first part of the project. Instead of using a specific optimizer for generating adversarial examples, we manually implemented the gradient step. For adversarial example generation, we chose to use 10 pgd steps, bound=0.01, step size = bound/4, where represents the weight in the hybrid approach.

Additionally, we did not use random restarts in generating adversarial examples because we observed that it led to less improvement in robustness.

All parameters were chosen through manual experimentation, testing different combinations to find a good balance between accuracy and adversarial robustness.Initially, we experimented with larger bounds, expecting better robustness, but found that too large of a bound actually reduced robustness. This may suggest that overly large bounds fail to create subtle adversarial examples, which are necessary for creating sharper decision boundaries.

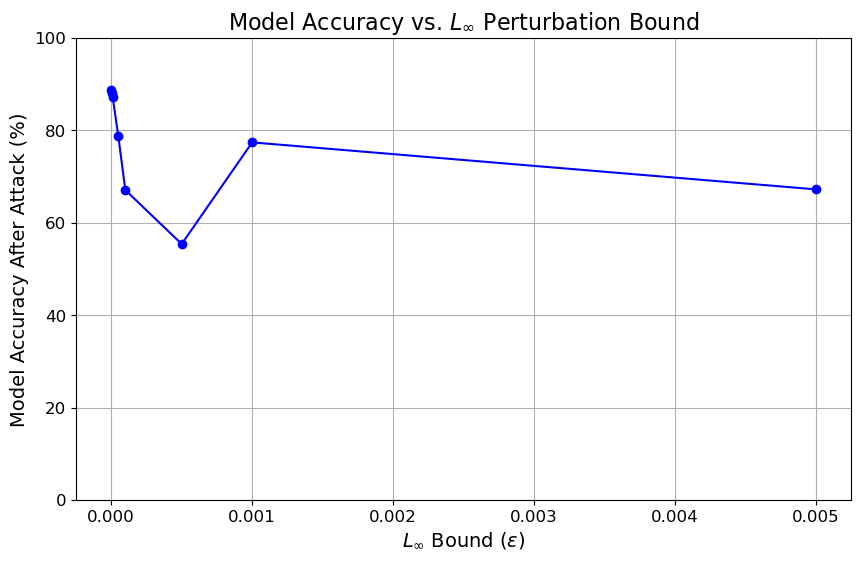
We also experimented with increasing the weight of the adversarial loss by trying different values of , but this reduced the classifier's performance to below 80% and did not result in significant improvements in robustness under attacks.

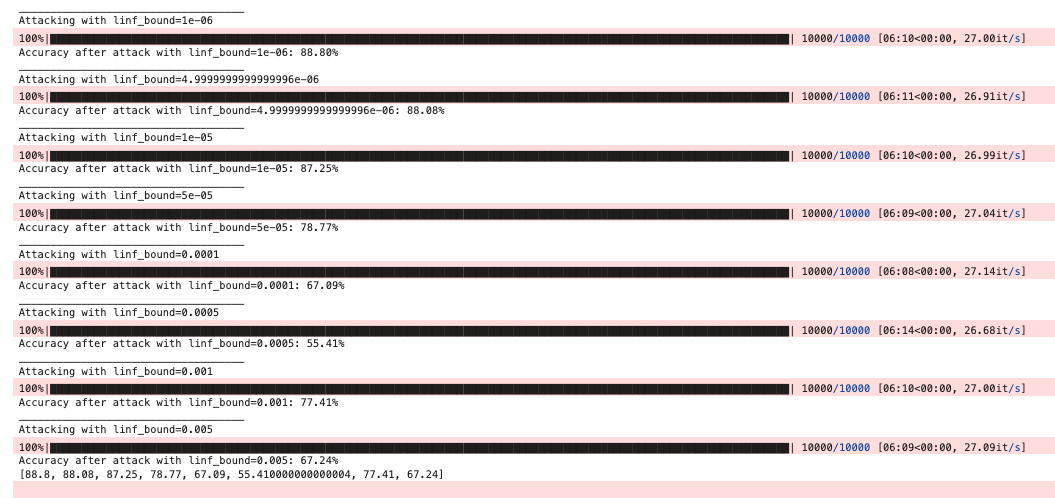
**Results:**

**Clean classification accuracy:**Train Accuracy: 94.92%

Validation Accuracy: 89.46%

Test Accuracy: 88.93%



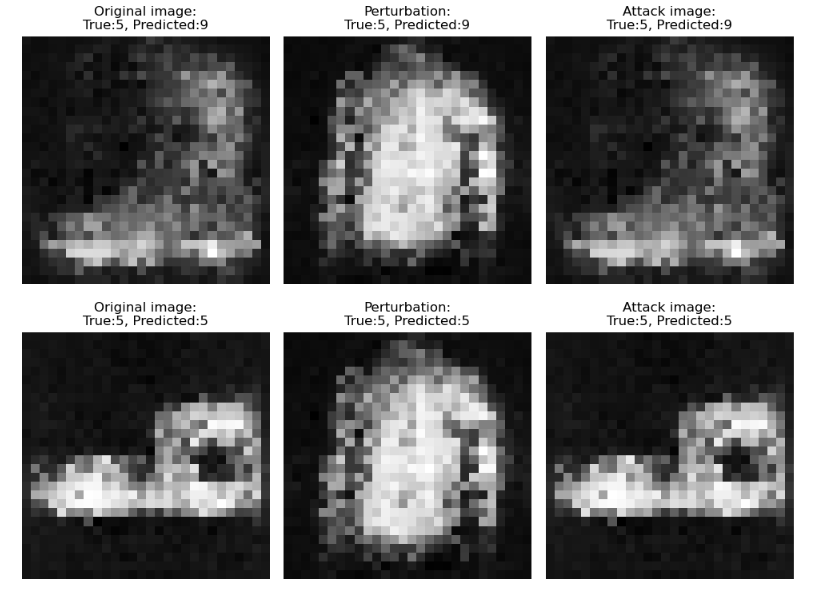
****

As observed from the results, we were able to achieve a significant improvement in the robustness metric. Compared to the classifier from the previous sections, which saw its robustness drop to zero for bounds of 0.0005 and higher, this classifier demonstrates high robustness, particularly for larger bounds.

However, we observe an interesting phenomenon: the model exhibits lower robustness for mid-sized bounds compared to larger bounds. This could be due to the fact that we trained our model using a large bound (). Training on such a large bound likely generated adversarial examples that increased the model's robustness against larger perturbations but did not sufficiently improve robustness for mid-sized perturbations.  
This might occur because the model’s decision boundaries became more focused on handling extreme cases, allowing mid-sized or smaller perturbations—which are more subtle—to still bypass the model's defenses.

To improve robustness across different perturbation sizes, it might be necessary to train on multiple perturbation sizes, ensuring the model can effectively defend against both subtle and large adversarial examples.

**Plotted images from class 5:**



**Bibliography**

[1] Emilien Dupont, Hyunjik Kim, S. M. Ali Eslami, Danilo Jimenez Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you can treat it like one. In International Conference on Machine Learning, pp. 5694–5725. PMLR, 2022.

[2] L. De Luigi, A. Cardace, R. Spezialetti, P. Zama Ramirez, S. Salti, and L. Di Stefano. Deep learning on implicit neural representations of shapes. In International Conference on Learning Representations (ICLR), 2023.

[3] Kim, T. Generalizing MLPs with dropouts, batch normalization, and skip connections. arXiv preprint arXiv:2108.08186, 2021.

[4] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations (ICLR), 2018.

[5] Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., & Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning (ICML), pp. 7472–7482, 2019.

[6] Haotao Wang, Aston Zhang, Shuai Zheng, Xingjian Shi, Mu Li, & Zhangyang Wang. Removing batch normalization boosts adversarial training. In ICML 2022, pp. 23433–23445. PMLR, 2022.

[7] Adversarial Machine Learning Tutorial. (n.d.). Adversarial Training. Retrieved October 22, 2024, from <https://adversarial-ml-tutorial.org/adversarial_training/>

[8] Stutz, D. (n.d.). 47.9% robust test error on CIFAR10 with adversarial training and PyTorch. Retrieved October 22, 2024, from <https://davidstutz.de/47-9-robust-test-error-on-cifar10-with-adversarial-training-and-pytorch/>

[9] Goodfellow, I. J., Shlens, J., & Szegedy, C. Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR), 2015. [Goodfellow et al., 2015](https://arxiv.org/pdf/1412.6572)

[10] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations (ICLR), 2017. [Madry et al., 2017](https://arxiv.org/pdf/2003.01690)