Analysis for Adversarial Attack on Image on Different Models

Lupin Cai
CS Department
Emory University
Atlanta, Georgia
lupin.cai@emory.edu

Helen Jin
CS Department
Emory University
Atlanta, Georgia
helen.jin@emory.edu

Jinghan Sun
CS Department
Emory University
Atlanta, Georgia
jinghan.sun@emory.edu

***Abstract*—Adversarial attacks pose significant challenges to the robustness and reliability of machine learning (ML) and deep learning (DL) models. This paper investigates the effects of the Iterative Fast Gradient Sign Method (iFGSM) on the robustness of diverse ML and DL architectures, including Random Forests, Decision Trees, ResNet18, Vision Transformers (ViT), and CLIP models. Using the CIFAR-10 dataset as a benchmark, we evaluate model performance under varying levels of adversarial perturbation. Principal Component Analysis (PCA) is employed to visualize the disruption of feature embeddings caused by adversarial attacks. The results reveal a hierarchy of model robustness, with Vision Transformers showing the highest resilience and traditional ML models being the most vulnerable. Our findings highlight the need for improved defense mechanisms and provide insights into the architectural factors influencing adversarial robustness. This study underscores the importance of holistic evaluations across different model types to advance the field of adversarial machine learning.**

I. Introduction

Adversarial attacks have emerged as a critical challenge in machine learning (ML) and deep learning (DL), where carefully crafted perturbations can significantly degrade model performance. These attacks exploit vulnerabilities in the models' decision boundaries, often producing outputs that are incorrect but imperceptible to human observers. With the increasing deployment of ML/DL systems in sensitive domains, understanding and mitigating their susceptibility to adversarial attacks is paramount.

Among the various methods for generating adversarial examples, the Iterative Fast Gradient Sign Method (iFGSM) has gained prominence for its effectiveness in crafting perturbations that are more refined than single-step approaches like FGSM [5]. By iteratively maximizing the model's loss within a constrained perturbation budget, iFGSM often leads to adversarial examples that are harder to defend against. In this project, we focus on the impact of iFGSM on a diverse set of models, including both traditional ML algorithms and state-of-the-art DL architectures.

Using the CIFAR-10 [1] dataset as a benchmark, we evaluate the robustness of models such as Random Forest, Decision Tree, ResNet18 [4], Vision Transformer (ViT) [6], and CLIP [7] under varying levels of adversarial perturbation. To further understand the effects of these perturbations, we leverage Principal Component Analysis (PCA) to visualize how adversarial attacks disrupt the clustering structure of feature embeddings. This analysis provides insights into the vulnerability of diverse model architectures when subjected to iterative adversarial attacks like iFGSM.

II. Related Work

Adversarial robustness has been a focal point of research in ML and DL due to the significant risks posed by adversarial attacks. Early foundational work by Madry et al. [8] introduced adversarial training as a min-max optimization problem, training models to minimize worst-case loss under adversarial perturbations. This approach has proven effective but is computationally intensive, particularly for iterative attack methods like iFGSM.

To address these challenges, Wong et al. [9] proposed fast adversarial training, which reduces computational overhead while maintaining robustness against common attacks, including FGSM and its iterative variant, iFGSM. The iterative nature of iFGSM often leads to stronger adversarial examples, highlighting the need for efficient defenses that can generalize across attack methods.

As DL architectures evolve, researchers have investigated their varying susceptibility to adversarial attacks. Mahmood et al. [10] examined the robustness of Vision Transformers (ViTs) [6] and found that their self-attention mechanisms provide some inherent resistance to adversarial perturbations compared to convolutional neural networks (CNNs). This aligns with a broader interest in exploring how architectural features influence adversarial robustness, especially under iterative attacks like iFGSM.

Self-supervised learning has also been explored as a defense mechanism. Kim et al. [11] introduced adversarial

self-supervised contrastive learning, which combines adversarial training with self-supervised objectives to improve robustness against iterative attacks. Their approach demonstrates that pretraining can enhance resilience to attacks like iFGSM while maintaining strong performance on clean data.

Despite these advancements, the integration of traditional ML models, such as Random Forests and Decision Trees, with adversarial defenses remains an underexplored area. Comparative studies evaluating the adversarial robustness of diverse algorithms, particularly under iterative attacks like iFGSM, are limited. Additionally, the use of dimensionality reduction techniques such as PCA to analyze how adversarial attacks affect feature embeddings is relatively underutilized. These gaps underscore the need for research that bridges traditional and modern approaches, providing a comprehensive evaluation of adversarial robustness.

In this project, we address these gaps by focusing on iFGSM, an iterative attack method, to assess the adversarial robustness of a range of ML and DL models. By combining quantitative evaluations with PCA-based visualizations, we aim to contribute to a deeper understanding of how iterative attacks disrupt model performance and feature representations.

III. Proposed approaches

We first trained a CNN with 3 convolutional layers that would return an output with 128 features and will be fed into the neural network for classification using the 50k samples from the CiFar10 data training set. Then with the trained model m, we perform the Iterative Fast Gradient Sign Method (iFGSM), which is an extension of the Fast Gradient Sign Method (FGSM), designed to generate more potent adversarial examples by applying perturbations iteratively. At each iteration, iFGSM adjusts the input in the direction of the loss function gradient with respect to the input, scaled by a step size $\alpha = \epsilon/2/iterations$. The process is repeated for a predefined number of iterations, with each update constrained within an ϵ ball around the original input to ensure the perturbation remains imperceptible. This iterative approach allows iFGSM to refine adversarial perturbations, leading to examples that are more effective at deceiving machine learning models compared to single-step attacks. This method forms the foundation of our adversarial attack strategy in this study. We perform iFGSM on the CiFar10 test set with different ϵs.

IFGSM Attack Function

1) **Input:**
   - **Model:** A trained neural network.
   - **Images:** Input images (batch).
   - **Labels:** True labels for the images.
   - $\boldsymbol{\epsilon}$**:** Perturbation limit.
   - $\boldsymbol{\alpha}$**:** Step size.
   - **iters:** Number of iterations.
2) Initialize images as trainable tensors.
3) **Repeat for** $\boldsymbol{iters}$ **iterations:**
   - Compute model outputs: `outputs` $\leftarrow$ `model(images)`.
   - Compute loss: `loss` $\leftarrow$ `CrossEntropyLoss(outputs, labels)`.
   - Backpropagate gradients: `grad` $\leftarrow \nabla$`loss`.
   - Update images: `images` $\leftarrow$ `images` + $\boldsymbol{\alpha \cdot sign}$`(grad)`.
   - Clip images to valid range: `images` $\leftarrow$ `clip(images, [0, 1])`.
   - Ensure perturbation limit: `images` $\leftarrow$ `clip(images, original_images` $-\ \boldsymbol{\epsilon}$`, original_images` $+\ \boldsymbol{\epsilon}$`)`.
4) **Output:** Adversarial images.

For visualizing the perturbed difference of the original pictures and the adversarial examples. We perform PCA on both of the sets.

For evaluation of the effectiveness of ϵ, we fit the adversarial attacked images with different ϵ to DL and ML models to see how accuracy changes in accordance to different ϵ.

IV. Datasets

CIFAR-10 dataset is a widely used benchmark in computer vision, consisting of 60,000 32x32 pixcel color images (3 channels)divided into 10 distinct classes, such as airplanes, automobiles, birds, and cats. An example of CIFAR-10 dataset, class truck, is shown in

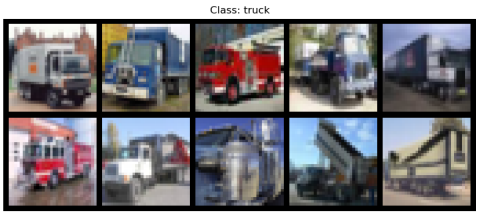

Fig. 1. An example image from the CIFAR-10 dataset, showing the class "truck."

The dataset is split into 50,000 training images and 10,000 test images, providing a balanced and diverse representation of objects commonly encountered in real-world scenarios. In this project, we utilized the CIFAR-10 dataset to evaluate the robustness of various machine learning and deep learning models against adversarial

attacks. Specifically, the training set was used to train models such as ResNet18, Vision Transformer, Random Forest, and Decision Tree, while the test set served as the basis for generating adversarial examples using the Fast Gradient Sign Method (FGSM). This allowed us to compare model performance on both clean and perturbed test sets, as well as analyze how adversarial perturbations disrupted feature representations and clustering patterns within the dataset.

V. System design

As shown in Fig. 2, this system design evaluates the impact of adversarial attacks on machine learning and deep learning models using the CIFAR-10 dataset. The process begins by training a Convolutional Neural Network (CNN) on the CIFAR-10 training set, extracting features from its last layer for further analysis. The CIFAR-10 test set is then subjected to adversarial perturbations using the Fast Gradient Sign Method (FGSM) to create an attacked test set. A performance comparison is conducted between the original and attacked test sets using three approaches: traditional machine learning models, such as Decision Trees and Random Forests, trained on the extracted CNN features, the result of run PCA on two sets, and deep learning models, including CNNs, Vision Transformers, ResNet-18, and Contrastive Language-Image Pretraining (CLIP) models. This workflow aims to assess and compare the robustness of various models to adversarial perturbations.

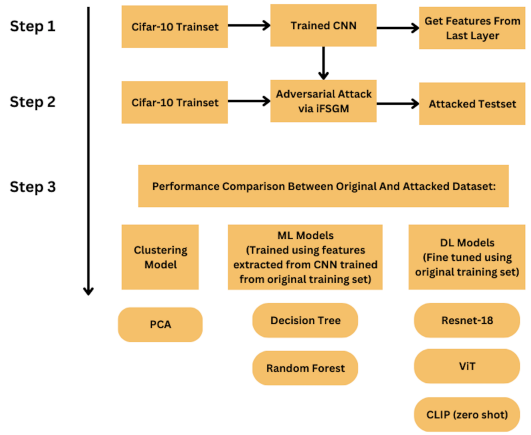

Fig. 2. Project Work Flow

TABLE I
Initial Model Accuracy on CIFAR-10 Dataset (Clean Test Set)

Model	Accuracy (%)
ResNet18	91.05
Vision Transformer	53.98
CLIP	84.71
Random Forest	79.05
Decision Tree	65.75

VI. Experiment results

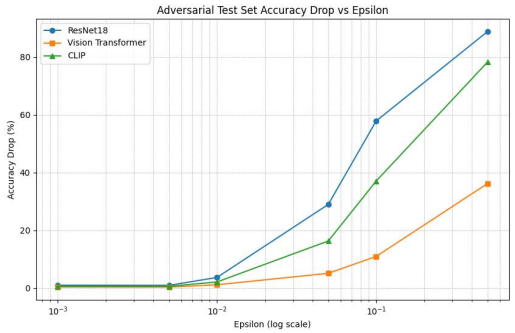

Fig. 4. Adversarial Test Set Accuracy Drop vs Epsilon (ResNet18, Vision Transformer, and CLIP)

The first graph (Figure 4) highlights the comparative robustness of deep learning models against adversarial perturbations. Quantitatively, ResNet18 experiences the steepest accuracy drop, exceeding 80% at higher epsilon values, revealing its significant sensitivity to adversarial noise. Vision Transformer (ViT), on the other hand, shows a relatively lower accuracy drop, remaining under 40%, even at the highest epsilon values, which indicates better robustness. CLIP lies in between, with moderate robustness, performing better than ResNet18 for mid-range epsilon values but less so compared to ViT.

The results emphasize a robustness hierarchy of $ViT > CLIP > ResNet18$, which can be attributed to their architecture. ViT's global attention mechanism likely provides resilience to localized adversarial perturbations, while CLIP's pretraining on diverse datasets strengthens its robustness. ResNet18, despite its strong baseline performance, struggles under adversarial attacks, indicating limitations in its convolutional architecture's capacity to counter adversarial noise.

PCA Analysis (Epsilon = 0.005)

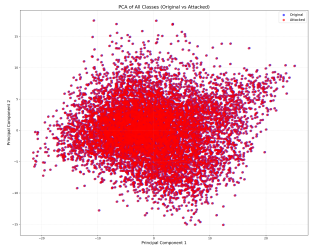

Fig. 5. PCA Visualizations of All classes with Epsilon value = 0.005

PCA Analysis (Epsilon = 0.05)

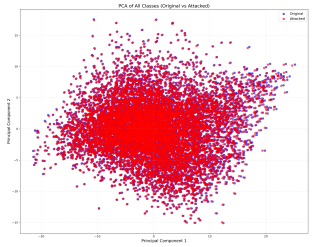

Fig. 6. PCA Visualizations of All classes with Epsilon value = 0.05

PCA Analysis (Epsilon = 0.5)

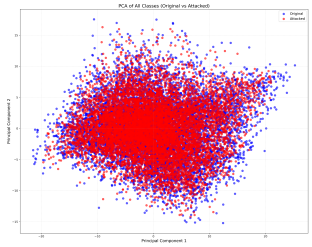

Fig. 7. PCA Visualizations of All classes with Epsilon value = 0.5

The PCA analysis across the three graphs demonstrates the distribution of features in the presence of varying levels of adversarial perturbations, represented by different Epsilon values. The attacked data points are in red and the original data points are in blue. In Fig 5, (epsilon = 0.005), the data points from the original and adversarially perturbed samples are closely clustered, indicating minimal impact from the adversarial noise. In this case, the perturbations are not significant enough to distort the feature representation.

In Fig 6, (epsilon = 0.05), the separation between the original and adversarially perturbed samples becomes a bit more distinct, with still significant overlapped regions in their distributions. This reflects a moderate degradation in the feature space, showing that the adversarial perturbations are starting to impact the principal components extracted from the images.

In the Fig 7, (epsilon = 0.5), a more distinct division attacked data points and original data points is observed, demonstrating the distortion of feature embeddings caused by adversarial noise. This means that at this level of perturbation, the difference between original and adversarially perturbed samples can be more clearly observed. Quantitatively, the attacked data points shift closer to a central point, indicating a disruption in the original clustering structure.

Overall, adversarial perturbations affect the features extracted from images, making classification more challenging for models. As epsilon increases, such differences become more distinct, and image data points become more central. PCA effectively visualizes the extent of these disruptions, illustrating how adversarial attacks compromise the integrity of the feature space, thereby reducing model confidence and accuracy.

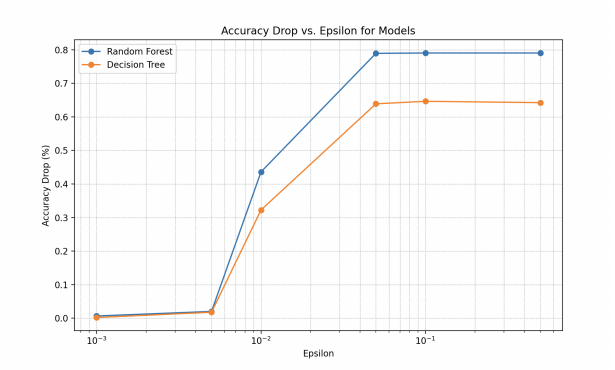

Fig. 8. Accuracy Drop vs Epsilon for Random Forest (RF) and Decision Tree (DT)

The second graph (Figure 8) demonstrates the behavior of traditional machine learning models under adversarial perturbations, each almost zero accuracy at fourth data point, with epsilon value of 0.05. Random Forest initially achieves a baseline accuracy of 79.05% but drops to nearly 0% at higher epsilon values, showcasing its extreme vulnerability. Comparatively, decision Tree performs slightly better in accuracy drop. Similarly, with a baseline accuracy of 65.75%, after attack, Decision Tree retain minimal predictive capability (1.5%) at epsilon level of 0.05.

Qualitatively, traditional ML models are significantly more vulnerable compared to deep learning models, as expected. This vulnerability stems from the lack of feature extraction mechanisms, such as convolutional layers or attention mechanisms, found in DL architectures. The higher baseline accuracy of Random Forest provides limited advantage against adversarial noise, indicating that

ensemble methods alone are insufficient for adversarial robustness.

VII. Discussions

The results of this study underscore the vulnerabilities of machine learning (ML) and deep learning (DL) models to adversarial attacks, with implications for their deployment in real-world applications. Several key points emerge from the analysis:

A. Differentiated Robustness Across Model Architectures

The observed robustness hierarchy among DL architectures—Vision Transformers (ViTs) outperforming CLIP and ResNet18—indicates that model design significantly influences resistance to adversarial perturbations. ViTs demonstrated resilience due to their global self-attention mechanisms, which appear less sensitive to localized perturbations. This aligns with prior findings suggesting that the ability to integrate context across an image can mitigate the impact of adversarial noise. On the other hand, convolutional architectures like ResNet18, while effective on clean data, showed susceptibility to adversarial examples, highlighting a potential trade-off between architectural simplicity and robustness.

B. Vulnerability of Traditional Machine Learning Models

Traditional ML models, such as Random Forests and Decision Trees, exhibited significant performance degradation under adversarial attacks. Unlike DL models, these algorithms rely on handcrafted features or embeddings extracted from other models and lack inherent mechanisms for spatial or semantic feature extraction. The near-complete loss of accuracy before reaching highest perturbation levels suggests that integrating adversarial defenses into these models remains an open challenge, particularly for scenarios where such models are used due to their interpretability and computational efficiency.

C. Impact of Adversarial Perturbations on Feature Space

The PCA visualizations provide a compelling narrative on how adversarial attacks disrupt the feature space. At low perturbation levels, the feature cluster differences between the original and the attacked data were similar with lots of overlaps, but as the attack strength increased, the deviation of attacked data points from the original ones became distinct. All the principal component value were deviated from the original ones, and the data points were moved toward the center. This degradation illustrates the mechanism by which adversarial perturbations mislead models—by rendering features extracted from attacked samples indistinguishable from those of other classes. This phenomenon highlights the need for defenses that preserve the integrity of feature representations, such as adversarial training or feature-denoising techniques.

D. Trade-Off Between Accuracy and Robustness

While higher robustness was observed in some models, it often came at the cost of lower baseline accuracy on clean data. For example, ViTs had lower clean accuracy than ResNet18, but they experienced a smaller drop in performance under attack. This trade-off underscores the importance of context-specific model selection, particularly in safety-critical applications where robustness may outweigh the need for peak accuracy on clean data.

VIII. Implications for Adversarial Defense Strategies

The findings indicate that existing defense mechanisms must evolve to address the increasing sophistication of adversarial attacks. For DL models, integrating adversarial training or pretraining on diverse datasets (as seen in CLIP) appears to improve resilience. For traditional ML models, there is a need for hybrid approaches that combine the interpretability of these models with the robustness of DL architectures, potentially through embedding-based defenses or ensemble methods.

A. Practical Considerations

The effectiveness of adversarial attacks, as demonstrated in this study, highlights the importance of considering robustness during model development, especially in applications like autonomous vehicles, healthcare, and security systems. Regular evaluations of models under adversarial scenarios and the incorporation of robust training methodologies should become standard practices for practitioners in the field.

IX. Limitations and Future Work

While this study provides valuable insights, it has several limitations. First, the experiments were conducted using a single dataset (CIFAR-10), which may not be generalized to more complex real-world scenarios. Future work could extend this analysis to larger and more diverse datasets, such as ImageNet. Second, only iFGSM was considered as an attack method, but adversarial robustness should be evaluated against a broader range of attacks, including black-box and physical attacks. Thirdly, when fine-tuning the vision transformer, the number of epochs is 4, which might not be sufficient due to the limited computing power that we have. Finally, while PCA offered a useful tool for visualizing feature space disruptions, more advanced methods such as t-SNE or UMAP could provide additional insight.

Future work should extend this study to more diverse datasets and adversarial attack methods to generalize the findings. Exploring advanced defense mechanisms, such as adversarial training, feature denoising, or hybrid model architectures, is essential to improving robustness. Furthermore, investigating the relationship between architectural

design choices and adversarial resilience can guide the development of inherently robust ML/DL models.

X. Conclusions

This study provides a comprehensive evaluation of the robustness of machine learning (ML) and deep learning (DL) models against adversarial attacks, using the Iterative Fast Gradient Sign Method (iFGSM) and the CIFAR-10 dataset as a benchmark. Several key findings emerge from our analysis.

Model Robustness Hierarchy: Among DL architectures, Vision Transformers (ViTs) exhibit the highest resilience to adversarial perturbations, followed by CLIP and ResNet18. ViTs leverage global attention mechanisms that appear to mitigate the impact of localized adversarial noise. In contrast, traditional ML models, such as Random Forests and Decision Trees, are highly susceptible, with performance collapsing under moderate perturbations.

Feature Space Disruption: Adversarial attacks distort feature embeddings, reducing the spread of image clusters. Principal Component Analysis (PCA) visualizations reveal this disruption, particularly at higher perturbation levels, where the original and adversarially perturbed samples clearly deviates.

Trade-Offs in Accuracy and Robustness: A trade-off exists between clean-data accuracy and robustness. Models like ResNet18 achieve high accuracy in clean data, but suffer steep accuracy drops under attack. In contrast, ViTs maintain higher robustness, although with lower baseline accuracy, highlighting the need for application-specific model selection. Meanwhile, low baseline accuracy of ViTs might also attribute to limited traning time we had during the project.

Practical Implications: These findings underscore the importance of integrating robustness as a core design consideration for models used in safety-critical applications, such as autonomous driving, healthcare, and security. The vulnerability of traditional ML models highlights the need for hybrid solutions that combine the interpretability of traditional models with the robustness of DL architectures.

In conclusion, this study highlights the persistent challenges posed by adversarial attacks and the need for innovative strategies to mitigate their impact. By bridging the gap between traditional and modern approaches, this work contributes to advancing adversarial robustness research and fostering safer, more reliable machine learning systems.

References

[1] Krizhevsky, A. (2009). *Learning Multiple Layers of Features from Tiny Images*. Technical Report, University of Toronto. Retrieved from https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[2] O'Shea, K., & Nash, R. (2015). *An Introduction to Convolutional Neural Networks*. arXiv:1511.08458. Retrieved from https://arxiv.org/abs/1511.08458

[3] Zügner, D., Akbarnejad, A., & Günnemann, S. (2018). *Adversarial Attacks on Neural Networks for Graph Data*. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). ACM. http://dx.doi.org/10.1145/3219819.3220078

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. arXiv:1512.03385. Retrieved from https://arxiv.org/abs/1512.03385

[5] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and Harnessing Adversarial Examples*. arXiv:1412.6572. Retrieved from https://arxiv.org/abs/1412.6572

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv:2010.11929. Retrieved from https://arxiv.org/abs/2010.11929

[7] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. arXiv:2103.00020. Retrieved from https://arxiv.org/abs/2103.00020

[8] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2019). *Towards Deep Learning Models Resistant to Adversarial Attacks*. arXiv:1706.06083. Retrieved from https://arxiv.org/abs/1706.06083

[9] Wong, E., Rice, L., & Kolter, J. Z. (2020). *Fast is Better than Free: Revisiting Adversarial Training*. arXiv:2001.03994. Retrieved from https://arxiv.org/abs/2001.03994

[10] Mahmood, K., Mahmood, R., & Van Dijk, M. (2021). *On the Robustness of Vision Transformers to Adversarial Examples*. In Proceedings of the IEEE/CVF International Conference on Computer Vision.

[11] Kim, M., Tack, J., & Hwang, S. J. (2020). *Adversarial Self-Supervised Contrastive Learning*. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 2983–2994). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/1f1baa5b8edac74eb4eaa329f14a0361-Paper.pdf

XI. Appendix

The detailed parameter of CNN model trained during iFGSM processes shown in 9 Its embeddings of the convolution layer are also extracted to be used for running machine learning models.

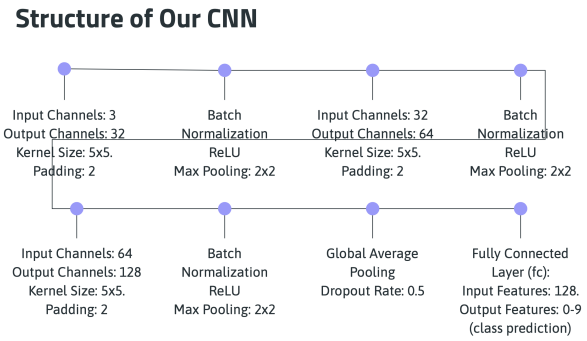

Fig. 9. Structure of the CNN We Used

The detail codes can be found with https://github.com/LupinC/Adversarial-Attack-Analysis-on-CNN-Image-Classification.

A. Task assignment for each member

Lupin Cai: Mainly responsible for training the CNN that is used for performing the adversarial attack. Creating adversarial images with different epsilons. Fine-tuning ResNet18 and Vision Transformer. Evaluating the accuracy drops of accuracy for Cifar10 test set classification of fine-tuned ResNet18, Vision Transformer, and zeroshot CLIP.

Jinghan Sun: Mainly responsible for Performing the PCA analysis to the attacked and original dataset and making relevant visualizations.

Helen Jin Mainy responsible for extracting embedding space for respective images from CNN models. Using image embeddings at various epsilons, trained Random Forest and Decision tree model. Evaluating the accuracy drops for machine learning classification tasks.