Neural Networks And Deep Learning
Mini Project

Tahmine Tavakoli

July 19, 2024

1 Introduction

In this project, I explore a novel approach to improving classification and face recognition by integrating ideas from two influential papers: *"A Discriminative Feature Learning Approach for Deep Face Recognition"* and *"The Enemy of My Enemy is My Friend: Exploring Inverse Adversaries for Improving Adversarial Training."* The primary goal is to investigate if inverse adversaries, as introduced in the second paper, can replace traditional centers in center loss to enhance the discriminative power of deep models.

2 Background

2.1 Deep Face Recognition with Center Loss

The paper *"A Discriminative Feature Learning Approach for Deep Face Recognition"* introduces center loss, which aims to enhance the discriminative power of features learned by deep neural networks. Center loss works by minimizing the distance between features of the same class and their corresponding class centers in the feature space. This reduction in intra-class variance helps improve the robustness and accuracy of the face recognition model.

2.2 Inverse Adversaries for Adversarial Training

The second paper, *"The Enemy of My Enemy is My Friend: Exploring Inverse Adversaries for Improving Adversarial Training,"* presents the concept of inverse adversaries. These adversaries are constructed using a modified Projected Gradient Descent (PGD) method and are positioned in high-likelihood regions. Unlike traditional adversaries that maximize loss to deceive the model, inverse adversaries are crafted to minimize the loss within specific constraints, potentially serving as more meaningful examples for improving model training.

3 Objective

The main objective of this project is to examine whether inverse adversaries can be used as a substitute for class centers in center loss. This involves evaluating and comparing the position of centers generated by inverse adversarial features and by center loss.

4 Methodology

4.1 Data Preparation

The experiments are conducted on the MNIST dataset. The dataset is divided into training and testing sets, ensuring a balanced representation of all classes.

4.2 Model Training

Two models are trained using the same architecture:

- **Model A**: Trained with center loss.
- **Model B**: Trained with cross-entropy loss.

Both models are trained for 25 epochs. The Stochastic Gradient Descent (SGD) optimizer is adopted with Nesterov momentum factor 0.9, cyclic learning rate schedule with a maximum learning rate of 0.01, and a weight decay factor of 5×10^{-4}.

4.3 Generating Inverse Adversaries

Using the trained models, inverse adversaries are generated through a modified PGD approach. The steps involved are:

1. Initialize a perturbation δ within a specified range.
2. Iteratively update δ to minimize the loss function while ensuring δ remains within the allowable range.
3. Apply the perturbation to the input data to generate inverse adversaries.

$$\check{\mathbf{x}}^{t+1} = \Pi_{\mathbb{B}(\mathbf{x},\epsilon')}\left(\check{\mathbf{x}}^{t} - \alpha' \cdot \operatorname{sign}\left(\nabla_{\check{\mathbf{x}}^t} \mathcal{L}_{\text{Inv}}\left(\check{\mathbf{x}}^{t}, y\right)\right)\right), \quad (3)$$

where α' is the gradient descent step size, $\check{\mathbf{x}}^t$ represents t^{th} iteration update, and $\mathcal{L}_{\text{Inv}}$ denotes the loss function for the inverse adversary generation. Generally, the cross-entropy

The inverse perturbation radius is set as $\epsilon' = 4/255$. The iteration steps for instance-wise inverse perturbation is 5 times with the step size of $\alpha' = 2/255$.

4.4 Comparison Metrics

The comparison of feature centers is conducted using two metrics:

1. **Cosine Similarity**: Measures the cosine of the angle between two vectors, indicating the directional similarity between feature vectors.
2. **Mean Squared Error (MSE)**: Measures the average of the squared differences between corresponding elements of two vectors, indicating the absolute difference in magnitude.

5 Experiments and Results

5.1 Model Training

Both models converged successfully, with Model A demonstrating lower intra-class variance due to the center loss. Model B showed typical behavior for a model trained with cross-entropy loss.

5.2 Inverse Adversary Generation

The inverse adversaries generated exhibited characteristics as expected, lying in high-likelihood regions and providing meaningful perturbations to the input data.

5.3 Feature Extraction and Analysis

The extracted features from clean images and inverse adversaries were used to compute the following centers:

1. Traditional center loss centers (from Model A).
2. Inverse adversary centers (from both Model A and Model B).
3. Clean feature centers (from both Model A and Model B).

5.4 Comparison of Centers

The cosine similarity and MSE were calculated between:

1. Inverse adversary centers and traditional center loss centers.
2. Inverse adversary centers and clean feature centers.
3. Inverse adversary centers and model weights.

5.4.1 Cosine Similarity Results

Model A:

- Inverse adversary centers vs. Center loss centers: High similarity indicating that inverse adversaries maintain the directional characteristics of traditional centers.
- Inverse adversary centers vs. Clean feature centers: Moderate similarity showing some divergence due to adversarial nature.
- Inverse adversary centers vs. Model weights: Lower similarity indicating that weights do not align directly with feature space centers.

Model B:

- Inverse adversary centers vs. Clean feature centers: Similar to Model A, showing that inverse adversaries preserve some class-specific characteristics.
- Inverse adversary centers vs. Model weights: Similar trend as Model A.

5.4.2 MSE Results

Model A:

- Inverse adversary centers vs. Center loss centers: Low MSE indicating that inverse adversary centers are close in magnitude to traditional centers.
- Inverse adversary centers vs. Clean feature centers: Higher MSE reflecting the perturbation introduced by adversaries.
- Inverse adversary centers vs. Model weights: Highest MSE showing a significant difference between feature centers and weights.

Model B:

- Inverse adversary centers vs. Clean feature centers: Moderate MSE reflecting some divergence due to adversarial nature.
- Inverse adversary centers vs. Model weights: Similar trend as Model A.

6 Discussion

The results indicate that inverse adversary centers exhibit high cosine similarity and low MSE with traditional center loss centers, especially in Model A. This suggests that inverse adversaries can potentially serve as a substitute for traditional centers in center loss, maintaining intra-class compactness and improving model robustness.

Moreover, the moderate similarity and MSE with clean feature centers in both models demonstrate that inverse adversaries preserve class-specific features while introducing beneficial perturbations. The significant differences with model weights confirm that feature space characteristics are distinct from learned weights.

7 Conclusion

This project demonstrates the feasibility of using inverse adversaries as an alternative to traditional centers in center loss for deep classification. The approach shows promise in enhancing model robustness and maintaining feature discriminativeness. Further research can explore optimizing the generation of inverse adversaries and integrating this approach into other discriminative learning frameworks.

References

[1] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In 6th International Conference on Learning Representations, ICLR 2018, 2018.

[2] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In International conference on machine learning, pages 7472–7482. PMLR, 2019.

[3] Junhao Dong, Seyed-Mohsen Moosavi-Dezfooli, Jianhuang Lai, Xiaohua Xie. The Enemy of My Enemy Is My Friend: Exploring Inverse Adversaries for Improving Adversarial Training. CVPR, 2023.

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In 6th International Conference on Learning Representations, ICLR 2018, 2018.

[5] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 499–515.

[6] Zhanglei Shi, Hao Wang , and Chi-Sing Leung , Senior Member. Constrained Center Loss for Convolutional Neural Networks. IEEE

[7] H.-M. Yang, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Robust classification with convolutional prototype learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3474–3482.