CSC413 / CSC2516 Winter 2023 Project Proposal

Wilson Gao 1005053987
Yuwei Liu 1004735824
Liang Chen 1005735126

Abstract

Improving the adversarial robustness of deep neural networks has been an active research topic in recent years. Adversarial training and generative models have shown promise in improving the robustness of deep neural networks against adversarial attacks. However, there is still much work to be done in this area, especially when dealing with novel datasets, such as art. In this paper, we investigate the effectiveness of two methods for improving adversarial robustness, image transformation and new network architectures, on a novel dataset called ArtEmis, which ties emotions to art. We evaluate our methods using the Fast Gradient Sign Method (FGSM) and the Projected Gradient Descent (PGD) attack, which are commonly used in evaluating adversarial robustness. We aim to explore the performance of these methods on different models and the ArtEmis dataset.

Introduction

Deep neural networks have shown remarkable performance in various computer vision tasks, including image classification. However, recent research has shown that these models are vulnerable to adversarial attacks, which can lead to misclassification with severe consequences. Therefore, improving the adversarial robustness of these models has become an active research topic. Adversarial training and generative models have shown promising results in improving the robustness of deep neural networks. In adversarial training, the models are trained with adversarial examples generated during the training process, while generative models learn a representation of the input space that is robust to adversarial attacks. Despite these methods' effectiveness, their performance on art datasets is not yet clear. One such dataset is the ArtEmis dataset. Therefore, in this paper, we investigate the performance of image transformation and new network architectures for improving the adversarial robustness of deep neural networks on the ArtEmis dataset. We consider two papers that apply pre-processing for the data, namely "Countering Adversarial Image using Input Transformation" by Guo et al. and "Feature Denoising for Improving Adversarial Robustness" by Xie et al. We also evaluate the performance of these methods using commonly used attacks such as FGSM and PGD. Our findings will contribute to the understanding of the effectiveness of these methods in improving the adversarial robustness of deep neural networks on art datasets such as ArtEmis.

Related Work

In recent years, there has been significant interest in improving the robustness of image classification models against adversarial attacks. A number of methods have been proposed to mitigate the impact of such attacks on deep neural networks.

One common approach to improve adversarial robustness is adversarial training. This technique involves augmenting the training data with adversarial examples generated during the training process. Goodfellow et al. (2014) proposed the fast gradient sign method (FGSM), which is a simple and effective method for generating adversarial examples. Subsequent work has extended this method, such as the projected gradient descent (PGD) method proposed by Madry et al. (2018), which is currently considered one of the strongest attacks for evaluating the robustness of a model.

Recent work has also explored the use of generative models to improve adversarial robustness. Samangouei et al. (2018) proposed the use of generative adversarial networks (GANs) to learn a representation of the input space that is robust to adversarial attacks. Other work has focused on using variational autoencoders (VAEs) to generate "clean" reconstructions of adversarial examples (e.g., Song et al., 2018).

While these approaches have shown promise in improving the robustness of deep neural networks, there is still much work to be done in this area. In our project, we mainly consider two papers "Countering Adversarial Image using Input Transformation" by Guo et al. (2018) and "Feature Denoising for Improving Adversarial Robustness" by Xie et al. (2019), which both apply some process for the data before training the model, and applying both ImageNet latter. We want to explore the performance of two methods in different models and the dataset.

Method/Algorithm

In this project, we investigate two methods to improve adversarial robustness: image transformation and new network architectures.

The former modifies the structure of adversarial perturbations to undo their effects. There are five main image transformation methods: image cropping and rescaling, bit-depth reduction, JPEG compression, total variance minimization, and image quilting.Image cropping and rescaling involves cropping and rescaling images at training time as part of data augmentation. At test time, we average predictions over the random images. This method alters the spatial positioning of the adversarial perturbation, which is essential in making attacks successful. Bit-depth reduction performs a simple type of quantization that can remove small adversarial variations in pixels. Similarly, JPEG compression removes small perturbations.Total variance minimization randomly selects a small set of pixels and reconstructs the "simplest" image consistent with the selected pixels. The reconstructed image does not contain adversarial perturbations because these perturbations tend to be small and localized. Image quilting is a non-parametric technique that synthesizes images by piecing together small patches from a database of image patches. It places appropriate patches in the database for a predefined set of grid points and computes minimum graph cuts in all overlapping boundary regions to remove edge artifacts.

Alternatively, new network architectures can increase adversarial robustness by performing feature denoising. Adding denoising blocks at intermediate layers of the convolutional network improves the robustness of image classification models against adversarial attacks. The denoising blocks are trained jointly with all layers of the network in an end-to-end manner using adversarial training. Four different instantiations of the denoising operation in our denoising blocks are non-local means, bilateral filter, mean filter, and median filter. The basic idea of adversarial training is to train the network on adversarially perturbed images generated by a

given white-box attacker based on the current parameters of the models. The Projected Gradient Descent (PGD)2 is used as the white-box attacker for adversarial training.

References

- ArtEmis: Affective Language for Visual Art: Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., & Guibas, L. (2021). ArtEmis: Affective Language for Visual Art. arXiv. https://doi.org/10.48550/ARXIV.2101.07396
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations. https://arxiv.org/abs/1412.6572
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations. https://arxiv.org/abs/1706.06083
- Samangouei, P., Kabkab, M., & Chellappa, R. (2018). Defense-GAN: Protecting classifiers against adversarial attacks using generative models. International Conference on Learning Representations. https://arxiv.org/abs/1805.06605
- Song, J., Kim, T., Nowozin, S., Ermon, S., & Kushman, N. (2018). Adversarial examples for generative models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7224-7233). https://arxiv.org/abs/1702.06832
- Guo, C., Rana, M., Cisse, M., & van der Maaten, L. (2018, January 25). *Countering adversarial images using input transformations*. arXiv.org. Retrieved February 16, 2023, from https://arxiv.org/abs/1711.00117
- Xie, C., Wu, Y., Maaten, L. van der, Yuille, A., & He, K. (2019, March 25). *Feature denoising for improving adversarial robustness*. Retrieved February 16, 2023, from https://openaccess.thecvf.com/content_CVPR_2019/papers/Xie_Feature_Denoising_for_Improving_Adversarial_Robustness_CVPR_2019_paper.pdf