CS 182 Project Proposal: Computer Vision Project

Chirag Sharma, Varun Jadia
shchirag@berkeley.edu, jvarun@berkeley.edu

1 Objective:

In this project, we aim to build an image classifier, trained on the ImageNet dataset, that is robust to perturbations and noise in test-time data. While achieving good classification performance on a test set generated from the same underlying distribution as the training set is a good metric to evaluate our learning quality on, we are ultimately interested in being able to make good quality predictions on real-world data, which is often noisy and sometimes adversarial. For example, a self-driving car needs to be able to correctly identify that there is a car in front of it even if the car is of a color it hasn't seen before or if it is covered in stickers or dirt. Small perturbations in inputs can significantly distort the feature embeddings, and consequently the output, of a deep network. Thus, we hope to propose a stabilized architecture that has a good performance on the original dataset, while also achieving good scores on perturbed data.

2 Related Work:

Zheng et al. [Zhe+16] stabilized the Inception CNN architecture [Sze+14] by generating Gaussian-noise augmented versions of samples during each training step and then adding the stability objective with respect to this perturbation to the loss function. Their stabilized network achieved better ranking scores on both the original and explicitly distorted test sets. The authors claim that this approach works better than explicitly augmenting the input dataset. However, this naive Gaussian noise approach is unlikely to help in the case of targeted adversarial attacks (for example, overlaying the original image with something).
Rusak et al. [Rus+20] highlight the differences between adversarial and common corruptions, and focus on building a neural network robust to Gaussian and Speckle noise as a way to combat both corruption types. Specifically, they have a two way approach: first, they train the classifier using data augmented with noise, and secondly, they train a generator neural net to produce worst-case noisy data, and jointly train the classifier model along with this (this is called ANT in the paper). They achieve state-of-the-art results on several corruption types, but for some corruptions, like Motion and Zoom blurs, the performance is subpar.

3 Technical Approach:

Our goal is to build stability features on top of architectures like Inception V3 that have a very good baseline performance on the ImageNet dataset. After initial tests, we plan on combining techniques seen in the above referenced papers to combat both 'common' image corruptions and adversarial attacks – some of these techniques include explicit dataset augmentation with various perturbed version of the original data, loss function modification as described in [Zhe+16], training adversarial networks to expose our network to maximally perturbed data as in [Rus+20], etc. Apart from these more conventional approaches, we will also increase robustness of our classifier using a heuristics-based approach – we will test our classifier on different types of occluded images (using the adversarial patches approach in [McC+20]) to defend against patch-attacks. Moreover, for the explicitly perturbed samples in our dataset, we aim to visualize the behavior of our filters on them to understand the types of activations triggered by the perturbations that we explore and use this understanding to heuristically develop regularization methods to combat it. Finally, we will develop a method by which we analyze the types of perturbations of the original images that our network performs poorly on to generate entirely new images that are designed to combine them and train our network to overcome these obstacles.

Thus, by combining several current state-of-the-art methods with this heuristic-based visualization approach, we hope to achieve optimal performance on real-world image data. In using several approaches to achieving prediction stability, we expect to obtain a novel architecture that is not only able to defend against seen perturbations and adversarial attack types but also performs well on new types of adversarial attacks. Finally, to ensure that our architecture maintains good performance on the original dataset, we plan to construct our training batches to include optimal amounts of legitimate and noisy images.

References

[McC+20] Michael McCoyd et al. *Minority Reports Defense: Defending Against Adversarial Patches*. 2020. arXiv: 2004.13799 [cs.LG].

[Rus+20] Evgenia Rusak et al. *A simple way to make neural networks robust against diverse image corruptions*. 2020. arXiv: 2001.06057 [cs.CV].

[Sze+14] Christian Szegedy et al. *Going Deeper with Convolutions*. 2014. arXiv: 1409.4842 [cs.CV].

[Zhe+16] Stephan Zheng et al. *Improving the Robustness of Deep Neural Networks via Stability Training*. 2016. arXiv: 1604.04326 [cs.CV].