DRAGONFIST

Joseph Boudreau, Andrew Ferrazzutti, Kia Shakiba, Zhongyu Zhao

***Abstract*—The purpose of this paper is to examine the effects of applying transformations to images in the context of image classification neural networks. We investigate how using filtered images during training and inference affects the classification accuracy and resilience to adversarial attacks. The rationale supporting this methodology is that transforming the input images should obstruct or diffuse the intended perturbations of an adversarial example, while still preserving the significant features of the image so the model can effectively generalize and classify. Our findings indicate a marginal increase in both accuracy and a significant improvement in adversarial robustness for the ensemble model against black-box attacks and white-box attacks. Despite the improvement in accuracy, adversarial images are still misclassified 50% of the time in most cases, so this model structure cannot be considered completely effective at this time.**

I. Introduction

In this paper we will describe and provide quantitative results for two model structures, each utilizing image filters in some way to create a more robust neural network. The first structure is an ensemble model composed of several individually trained CNNs, each with a different image filter applied prior to classifying. The outputs of every model are combined using a fitted logistic regression function. The second design is a single CNN trained on a data set which has been augmented sets of filtered versions of each image in the original training data set. During inference, the model will classify each filtered version of the image sequentially (having learned to classify different filtered versions of the images during training) and then similarly combine each prediction vector using logistic regression. In both designs, we experiment with different combinations of filters in order to determine which are effective. Our findings indicate a marginal increase in both accuracy (3-5%) and a significant improvement in adversarial robustness (30-40%) for the ensemble model against black-box attacks. The second model design achieved less favorable results, exhibiting only marginal increases in performance and accuracy. With regards to transformations, we investigate a multitude of existing filters, both linear and non-linear, in addition to implementing some custom filters and looking at their effectiveness. Previous research indicates that median filters are effective at removing adversarial noise[1] - we found this to be the case in our experiments as well. The most effective combination of filters to ensemble was Identity, Edge Detection, Min, Max, Rank, Median, and Gaussian. In general, most filtered models exhibited small accuracy reductions compared to the same model trained with no filter. The only filter which could achieve the same classification accuracy as an unfiltered model was a Sobel (edge detection) filter. Extremely lossy filters performed, as expected, much worse on classification. In general, we want to leverage the trade-off between distortion and accuracy in the ensemble to maximize resilience to adversarial perturbations.

For the purposes of convenience, as well as to limit the scope of variability in our research, every model in the ensemble shares the same underlying layer structure. The models are implemented using the Keras Sequential interface and are composed of several 2D convolutional layers, a 2x2 max pooling layer, a final densely connected convolutional layer followd by an output Softmax activation function. Unfortunately, due to a incompatibility between TensorFlow and the Cleverhans library, dropout regularization could not be utilized.

II. Terminology

The name of the project, which encompasses the entire research scope of this work, is ”Dimensionally Reduced Adversarial Generations on Networks Filtered by Integrated Static Transformations“ which is shortened to Project DRAGONFIST.

Each model and filter combination is referred to as a ”Classification Worker“ (CLAW). The only parameters for CLAWs are the choice of filter as well as optional feature-wise preprocessing applied to the filtered dataset.

The ensembled model, composed of a set of CLAWs, is named the “Final Intuition Stage” (FIST). The model trained on the training set augmented with filtered versions of the training images is called the ”Package Aggregate Learning Model” (PALM)

III. Image Filters

The proper selection of filters is important to the overall effectiveness of the algorithm. As previously mentioned, the intention of DRAGONFIST is to improve

TABLE I
IMAGE FILTER ACCURACIES.

Filter	Accuracy
Edge detection	0
Gabor	0
Gaussian	0
Average rows	0
Average columns	0
Average	0
Rank	0
Maximum	0
Minimum	0
Median	0

a machine learning model's resiliency against adversarial noise. Based on the work done by ——, adversaries can target specific pixels in an image to which noise will be applied. As such, filters which combine the values of pixels with their neighbours are intuitively more desirable as they increase the difficulty of targeting specific areas of the image to which noise will be applied. Therefore, the following filters were considering when designing DRAGONFIST:

- Edge detection
- Gabor
- Gaussian
- Average rows
- Average columns
- Average
- Rank
- Maximum
- Minimum
- Median

A. Descriptions

In order to understand what the filters are doing to the images, a brief explanation of each is given.

1) Edge detection: The edge detection filter uses the Python library `skimage.filters.sobel` [2].

B. Filter accuracy

When applying each filter to the system, it is important to remember the overall goal of the model – to accurately classify its input. Therefore, each filter's accuracy in classification is important to consider to ensure the overall model maintains a high accuracy as well. Each filter was tested individually in order to determine its classification accuracy. The results of these tests can be observed in table I.

IV. White Box Attacks

V. Black Box Attacks

Blackbox attacks were implemented using a modified version of the attack provided by the Cleverhans library[3]. The attack threat model assumes unlimited oracle ability on the target model, but no knowledge of the internal weights or biases of the model itself. This threat model is arguably a more realistic and important scenario, compared to a model which assumes knowledge of the interior model. This is because a model's structure should never be externally exposed and defense against these attacks can be mitigated by traditional security controls to prevent data theft and unwanted network access. On the other hand, it is entirely valid for a model to expose query access publicly, either as part of an API or a GUI. Although true "oracle" access can be prevented with query limitations imposed on the user, there is still risk of gaining enough information from rate-limited access or of the controls being circumvented somehow. Therefore, finding effective strategies for minimizing the effectiveness of black-box attacks are an important field of research in machine learning security.

The attack works by training a substitute CNN using the output labels of the target model in conjunction with the images given to the model to be classified. This substitute model is a two layer CNN with ReLu activation functions and a Softmax normalization final layer. This substitute model is then used to generate adversarial examples using the Fast Gradient Sign Method developed by GoodFellow et. al. These adversarial examples are then tested against the original target model with the intention of misclassifying the input. This attack utilizes the transferability characteristic of adversarial examples - an adversarial image generated using one model will generally have a high probability of being misclassified by other models which perform the same classification task but may have been trained differently and use different internal structures.

We tested the effectiveness of both PALM and FIST model designs in increasing the classification accuracy of these transferred adversarial images, relative to a single "standard" model, as defined above.

VI. Future Research

VII. Conclusion

References

[1] M. Osadchy, J. Hernandez-Castro, S. Gibson, O. Dunkelman, , and D. P'erez-Cabo, "No bot expects the deepcaptcha! introducing immutable adversarial examples with applications to captcha," *IACR Cryptology ePrint Archive*, 2016.

[2] *Module: filters – skimage v0.15.dev0 docs*, https://scikit-image.org, scikit-image.

[3] N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Xie, Y. Sharma, T. Brown, A. Roy, A. Matyasko, V. Behzadan,

K. Hambardzumyan, Z. Zhang, Y.-L. Juang, Z. Li, R. Sheatsley, A. Garg, J. Uesato, W. Gierke, Y. Dong, D. Berthelot, P. Hendricks, J. Rauber, and R. Long, "Technical report on the cleverhans v2.1.0 adversarial examples library," *arXiv preprint arXiv:1610.00768*, 2018.