CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 135 **March 30, 2022**

PCA as a defense against some adversaries

Aparna Gupte, Andrzej Banburski, Tomaso Poggio
Center for Brains, Minds, and Machines, MIT

Abstract

Neural network classifiers are known to be highly vulnerable to adversarial perturbations in their inputs. Under the hypothesis that adversarial examples lie outside of the sub-manifold of natural images, previous work has investigated the impact of principal components in data on adversarial robustness. In this paper we show that there exists a very simple defense mechanism in the case where adversarial images are separable in a previously defined (k, p) metric. This defense is very successful against the popular Carlini-Wagner attack, but less so against some other common attacks like FGSM. It is interesting to note that the defense is still successful for relatively large perturbations.

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

PCA as a defense against some adversaries

Aparna Gupte, Andrzej Banburski, Tomaso Poggio

Center for Brains, Minds and Machines, MIT

March 30, 2022

Abstract

Neural network classifiers are known to be highly vulnerable to adversarial perturbations in their inputs. Under the hypothesis that adversarial examples lie outside of the sub-manifold of natural images, previous work has investigated the impact of principal components in data on adversarial robustness. In this paper we show that there exists a very simple defense mechanism in the case where adversarial images are separable in a previously defined (k, p) metric. This defense is very successful against the popular Carlini-Wagner attack, but less so against some other common attacks like FGSM. It is interesting to note that the defense is still successful for relatively large perturbations.

1 Introduction

While deep neural networks have been hugely successful in recent years in various domains, often achieving super-human performance, they are known to be very sensitive to small changes in inputs [1]. Human-imperceptible perturbations to images or audio [2] can force these systems to misbehave arbitrarily badly, ranging from changing classification predictions [3], to no longer seeing pedestrians on the road from application of a sticker to a street sign [4]. Many defenses have been proposed, with the most popular one being adversarial training, but even that fails when it is exposed to perturbations not seen in training.

Given that the existence of such adversarial perturbations seems to be guaranteed [5] in the standard framework we use for training, it is natural to ask what their origin is, and why the human brain seems to avoid it[1]. One popular hypothesis claims that adversarial images lie orthogonal to the manifold of "natural images", and that the decision boundaries learned by ReLU networks are highly sensitive in the directions off that manifold [7]. A natural question is how such a manifold can be defined. In [7] the authors attempt to build a local linear approximation of it by using auto-encoders. In this paper, we explore whether variants of the Principal Component Analysis (PCA) can be used to define such a manifold. Previous work has shown that the distributions of principal components of natural and adversarial images are different [8]. Is it then possible to use this information to define a sort of projection onto the subspace of principal components that are more likely associated with clean images?

In [9] a variation of PCA was considered, in which principal components can be extracted per image, rather than for the whole dataset. There, a (k, p) metric was defined that measures when the top logit p switches as principal components k are reduced. Preliminary experiments in [9] suggested that the (k, p) points of some attacks are separable from those of natural images, and allow for a highly reliable detection. In this paper we further investigate those claims and test a wider range of attacks. While the original paper was very sparse on the details of the experimental setting, we find that indeed the very popular Carlini-Wagner attack [10] can be successfully separated from clean images on ImageNet according to the (k, p) metric. By investigating the pattern of prediction changes as the number of principal components is reduced, we find that for an attack that can be detected, the correct label can be extracted with high success. We propose thus

[1]There is the caveat that time-limited humans are slightly susceptible to such perturbations [6].

a simple architecture that avoids the CW attack. We find however, that some other commonly used attacks, such as PGD or FGSM are not so easily separable and the defense is less successful. We discuss the possible roots for this simple defense and find that these results suggest a deeper look into the logits of adversarial images. Very interestingly, we show that after detecting the adversarial image with the (k, p) metric, taking the first flipped prediction is a significantly stronger defense than just taking the second highest logit.

2 Related Work

Principal Component Analysis has been previously used to study adversarial examples. Besides the work of [9] that we mainly base our paper on, [8] has observed that the principal component distributions of natural and adversarial images are different – adversarial images seem to have higher weight in the large principal components. [11] has tried using PCA as a preprocessing step, forcing a deep network to only use a small subset of principal components (around $k = 20$ for MNIST), giving an increase in robustness against FGSM attacks. In [12] PCA was used as a filter in the hidden convolutional layers, but the results were not very promising.

Detection of adversarial examples has been demonstrated by [13], where they trained a neural network to distinguish adversarial samples from natural ones. [14] has shown that the Bayesian uncertainty estimates allow for detection too. However, such detection mechanisms can often be bypassed [15].

The most popular approach towards building adversarially robust models relies on training on adversarial examples [16, 17]. In [18] a connection to differential privacy was shown. For a good overview of defenses against adversarial attacks, see [19]. A recent defense relevant to this paper was explored in [20], where the defender relies purely on analyzing logits.

3 Principal components and robustness

Row-PCA With large datasets, it becomes computationally very difficult to compute the traditional PCA, where we consider each image as a data point. Instead, [9] proposed to consider each $h \times w \times d$ image (an image with h rows, w columns and d channels) as a matrix $X \in^{h \times (wd)}$, and perform PCA with truncation level k on this matrix instead.

$$X = U\Sigma W^\top$$
$$X_k = XW_kW_k^\top,$$

where $W_k \in^{(wd) \times k}$. Here, we say that $X_k = (X)$.

The (k, p) point metric [9] proposed a new metric, called the (k, p) point, that they claimed allows to detect adversarial samples from clean ones. The (k, p) point can be computed given an image $X \in^{h \times wd}$ and a classifier $N :^{h \times wd} \to^C$, that maps images to a C-dimensional vector of logits for each class, where C is the number of classes. It is defined by removing principal components from the image, until the top predicted class changes.

In their paper, [9] claimed that the Carlini-Wagner, Deepfool and JSMA attacks were highly separable from clean images in this (k, p) metric. They claim that a linear classifier could distinguish clean images from adversarial ones with 94.81% success rate. However, the experimental details of the paper lack any description on the hyperparameters used for the attacks. We repeated this experiment and found that the CW attack is indeed highly separable in this metric from natural images, this is less so for other common attacks, see Table 1. We trained both a linear classifier (logistic regression) and a 3-layer MLP and plot the decision boundaries for logistic regression in Figure 1.

Algorithm 1: Computing the (k, p)-point, given an image $X \in^{h \times wd}$ and neural network N. (Reproduced from [9])

Input: Image $X \in^{h \times wd}$ and classifier N.
Output: The (k, p) point of X with respect to the classifier N.

Set $n \leftarrow \min(wd, h)$.
Compute dominant class $c^* \leftarrow \arg\max N(X)$.
for $k = n$ *to* 1 **do**
  Compute $X_k \leftarrow (X)$.
  Compute the classifier's prediction $c \leftarrow \arg\max N(X_k)$.
  if $c \neq c^*$ **then**
    Set $k^* \leftarrow k$.
    Set $p^* \leftarrow N(X_k)[c]$.
    return (k^*, p^*).

Attack	ϵ	Logistic regression		3-layer MLP	
		Train accuracy	Test accuracy	Train accuracy	Test accuracy
FGSM	0.03	64.1%	69.5%	66.9%	72.0%
	0.1	64.6%	64.0%	66.6%	70.5%
	0.5	79.5%	81.0%	80.0%	84.0%
PGD	0.01	68.0%	79.5%	70.0%	82.5%
	0.05	57.5%	58.5%	62.5%	77.0%
Deepfool	0.01	72.4%	83.5%	75.9%	86.0%
	0.1	72.6%	84.5%	76.6%	85.0%
Carlini-Wagner		96.5%	99.0%	96.5%	99.0%

Table 1: Detecting adversarial examples on ResNet-18 through the (k, p) point: classification accuracy with logistic regression as well as a 3-layer MLP.

4 Experimental methods

For our experiments, we use ImageNet10 [21], a subset of the ImageNet dataset [22] with 10 classes and 50 images from each class, for a total of 500 images. We crop these images to the center 224×224 pixels as a pre-processing step. Note that we still perform full 1000-way classification on ImageNet.

We consider pre-trained models of the ResNet-18 [23] and VGG-11 [24] architectures. For the classification of the (k, p) points, we consider both logistic regression and also a multi-layer perceptron classifier with three layers $(30 \times 20 \times 10)$ and optimized with the LBFGS algorithm. We trained the classifier on 400 samples and tested on 100.

We study several attacks, including Fast Gradient Sign Method (FGSM) [3], L_∞ Projected Gradient Descent (PGD) [16], Deepfool [25] and Carlini-Wagner L_2 [26] attacks. When appropriate, we consider a range of hyperparameters (listed in Tables 1, 2). We use implementations of adversarial attacks from the Adversarial Robustness Toolbox [27]. For the PGD attacks, we further set the step size to be $\epsilon/12$ and the number of steps to be 80. We run most of the attacks on the ResNet-18 model, and only run the Carlini-Wagner attack on VGG-11 due to resource constraints.

Attack	ϵ	Defense success rate	Conditional defense success		Baseline (2nd best logit)
			Logistic regression	3-layer MLP	
FGSM	0.03	28.1%	24.8%	29.0%	6.4%
	0.1	11.1%	2.7%	5.8%	2.8%
	0.5	1.6%	1.4%	1.6%	1.0%
PGD	0.01	55.5%	64.0%	60.8%	10.2%
	0.05	37.6%	48.6%	48.0%	0.6%
Deepfool	0.01	46.8%	50.1%	48.4%	8.8%
	0.1	41.9%	45.6%	44.2%	6.6%
Carlini-Wagner		100%	100%	100%	96.4%

Table 2: Defense success rates for various attacks on ResNet-18. The conditional defense success columns give the success of the defense on images that have been detected according to their (k, p) point. The baseline defense column shows the success of the simple defense in which the prediction is switched to the second highest logit for the whole image. This baseline is surprisingly effective for the CW attack, but in all cases our proposed defense outperforms it.

5 Defense from detection

In this section, we propose a simple defense against adversarial attacks. The defense works in two steps: first, we use a previously trained classifier to detect whether an image is adversarial or not according to the (k, p) metric [9]. Then, in case we detect an adversarial image, we apply our defense as follows. The main idea is to compute the row-PCA of images for values of k, starting from the largest possible value and going downwards. The defense then outputs the neural network's prediction on the row-PCA reconstructed image for the largest value of k for which the prediction does not match the prediction on the original (adversarial) image – this is the first value of principal component k for which the predicted class changes.

Algorithm 2: Proposed defense strategy against adversarial attacks, given an adversarial image $X \in^{h \times wd}$ and neural network N.

Input: Image $X \in^{h \times wd}$ and classifier N.
Output: The (k, p) point of X with respect to the classifier N.

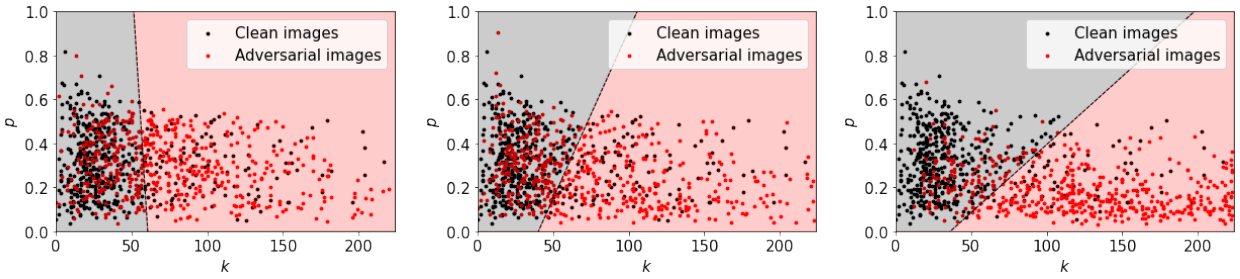
Set $n \leftarrow \min(wd, h)$.
Compute dominant class $c^* \leftarrow \arg\max N(X)$.
for $k = n$ *to* 1 **do**
  Compute $X_k \leftarrow (X)$.
  Compute the classifier's prediction $c \leftarrow \arg\max N(X_k)$.
  if $c \neq c^*$ **then**
    return c.
  return c^*.

Naively, there is no clear reason that this defense should be very successful. After all, the prediction we flip to after removing principal components on an adversarial sample does not a priori have to be correlated with the correct label, unless we have managed to project the image to the submanifold of natural images.

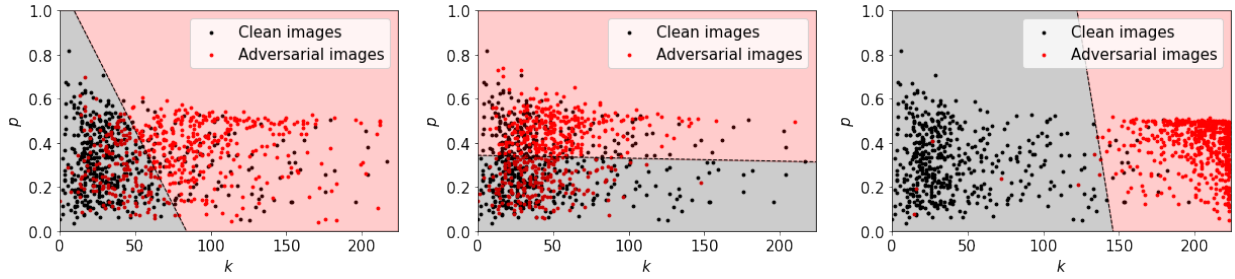
Carlini-Wagner Results We observe that this proposed defense strategy is very effective against the Carlini-Wagner attack, widely considered to be a very strong attack with minimal perturbations. We summarize our results in Table 2.

Why is this defense so surprisingly successful here? To understand this, we plotted the ten highest predicted class logits as we remove principal components from the image, see Fig. 2. By inspecting the

(a) Fast Gradient Sign Method $\epsilon = 0.03$ (b) Fast Gradient Sign Method $\epsilon = 0.1$ (c) Fast Gradient Sign Method $\epsilon = 0.5$



(d) Projected Gradient Descent $\epsilon = 0.01$ (e) Projected Gradient Descent $\epsilon = 0.05$ (f) Carlini-Wagner attack



(g) DeepFool attack $\epsilon = 0.01$ (h) DeepFool attack $\epsilon = 0.1$

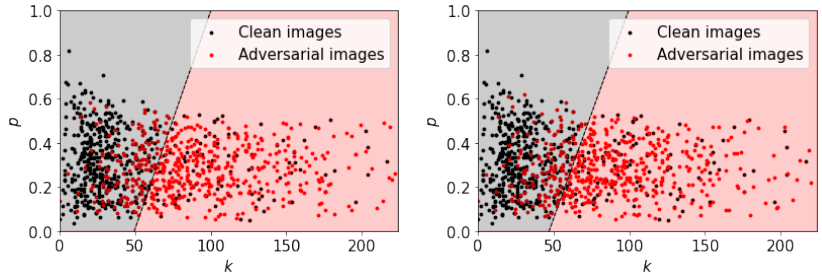


Figure 1: Detection of adversarial images from different attacks by using a linear classifier (logistic regression) on the (k, p) points of images. A pre-trained ResNet-18 model was used for these plots.

corresponding logit curves for natural and adversarial pairs, we find that for the CW attack, the label we flip to at the (k, p) point is the correct label for every single image in our dataset. Note that this is not necessarily the same as the second highest logit in the full image. As a baseline, we consider the defense of flipping to the second logit of the full image. We find that this baseline defense is actually successful on 96.4% samples, see Table 2. While all the results in Tables 1 and 2 are for the ResNet-18 model, we also tested this defense strategy against the Carlini-Wagner attack on a pre-trained VGG-11 architecture and summarize our results here. For detection based on classification in the (k, p)-space, both logistic regression achieved and the 3-layer MLP achieved 94.2% train accuracy and 98.5% test accuracy. The success rate of the proposed defense was 83.4% overall. The success rate did not change significantly conditioned on detection with logistic regression (84.0%) or the MLP (83.9%). Finally, our proposed defense does better than the naive baseline of choosing the class with second highest logit, which achieves 71.6% success rate against the attack. See the Supplementary Material for more plots.

We suspect that the reason why this defense works so well for the Carlini-Wagner attacks can be understood from the design of the attack. The CW attack tries to find the minimum perturbation to change the prediction of the neural network, and in the most common implementation does so by optimizing on the change of prediction from the true class to the target label. This would be compatible with the observation that most of

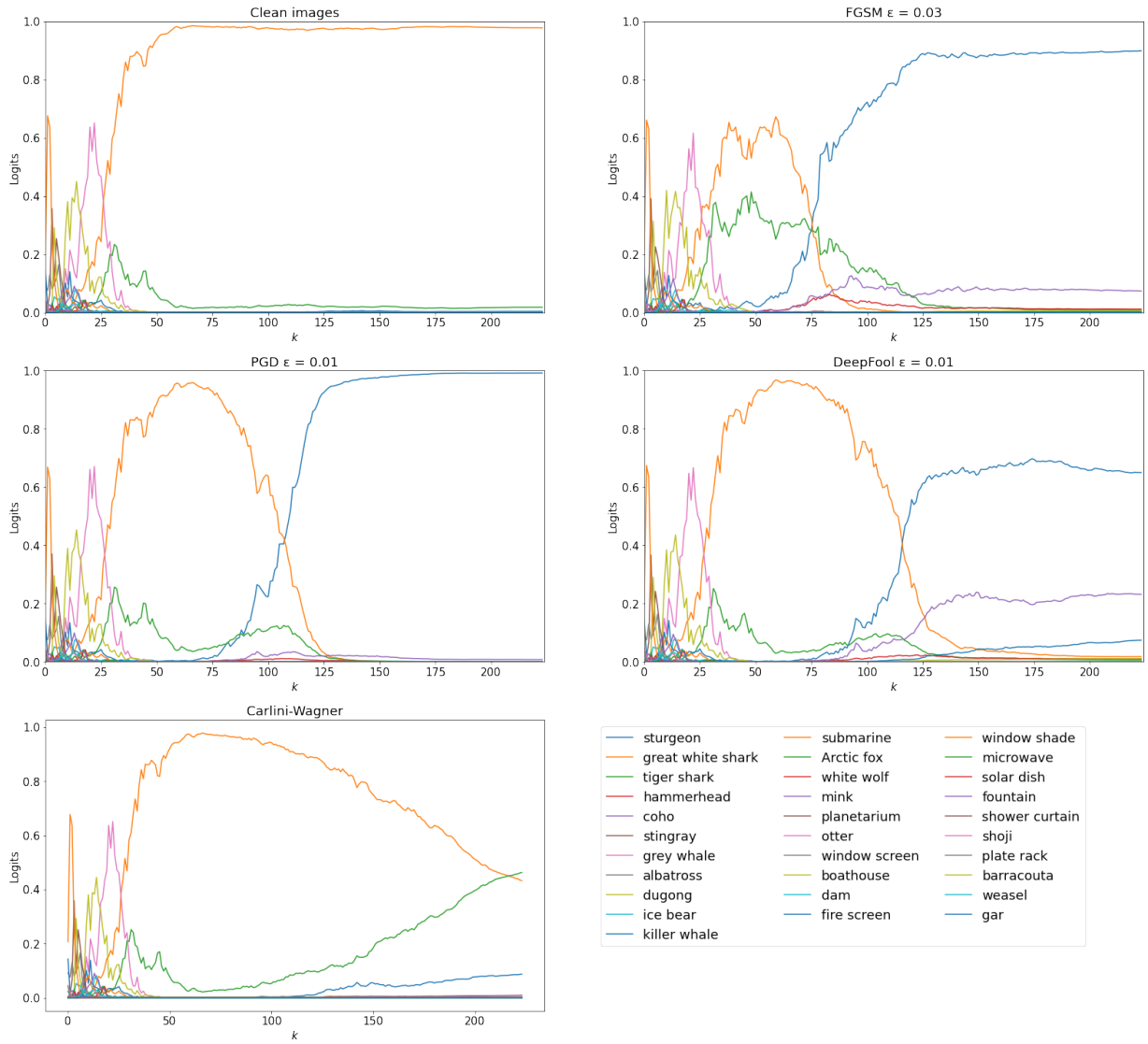


Figure 2: Logits of the top predictions of the network for a clean and adversarial images, as the number of principal components k gets reduced. Notice that the Carlini-Wagner attack mostly reduces the highest logit and promotes the second one to be just above the true label one. This is very different from the behavior for other attacks. In particular, FGSM leads to changes in many of the logits. This plot shows result for a single image of a great white shark, but the behavior is representative of the general behavior for all images, see Supplementary Material for more examples.

the time the correct class has the second-highest logit value for the adversarial image (but not all of the time).

Other attacks We saw previously that the attacks other than the Carlini-Wagner are less separable. Similarly, we find that our defense is less successful here too. Again, we consider the baseline of a defense in which we simply flip the highest and second highest logits of an adversarial image (assuming that we have previously detected adversarial noise somehow, perhaps via the (k, p) metric). As we can see in Table 2, our proposed defense vastly outperforms this simple baseline. We report both the success of the defense on all adversarial images, as well as the success conditioned on previous detection via the (k, p) metric. For small perturbations,

both the PGD and Deepfool attacks can be successfully repelled more than half of the time. Surprisingly, the very simple FGSM attack bypasses this proposed defense the most. Inspecting the pattern in which the top prediction changes as we reduce the number of principal components k in Fig. 2, we can see that there is a lot more impact on all the logits in FGSM compared to the other attacks.

6 Discussion & Conclusions

We found that the results in [9] were highly optimistic, but important details on the experimental paradigm were missing. Repeating their experiment, we discovered that indeed some attacks, like the CW attack, are highly separable in the (k, p) metric, while other common attacks are more difficult to separate. Nonetheless, results in Table 1 suggest that we can have a high success rate in detection even for PGD attacks.

This motivated us to define the defense algorithm based on the (k, p) point of an image. In our experiments we have noticed that often the prediction we flip to at the (k, p) point is the correct label (for CW attacks this is universally true). From inspecting the CW attack logits, a simpler defense scheme seemingly presents itself – one could simply take the second highest logit given a detection. We show in Table 2 that while this is minimally weaker for CW attacks (success rate of 96.4% as opposed to 100%), our proposed defense is significantly stronger for other attacks. For Deepfool, we get jumps from 8.8% up to 50.1% and for PGD a jump from 10.2% up to 64% for small perturbations. It is thus clear that our defense is significantly stronger than would be expected from just flipping logits. It is of note that for this defense, we did not need to restrict ourselves to a subset of classes in ImageNet - unlike adversarial training, this defense seems to be agnostic to the total number of classes.

While we suspect that this mechanism can be bypassed by an adversary with the knowledge of this defense, it is nonetheless interesting and surprising that such a simple algorithm using a variation of PCA can be so successful. We motivated the approach taken in this paper by an attempt to define the manifold of natural images. It would be very interesting to understand what the manifold defined by our projection exactly is – hopefully an extension of this work could demonstrate that such a manifold is more closely aligned with human robustness. The rates of detection and successful defense we found here motivate a closer look at the distribution of principal components in natural and adversarial images and how those interact with the predicted logits in classification problems. Our proposed defense is very simplistic in choosing the classification based on the second decision we flip to at the (k, p) point, but the results here are suggestive that it might be possible to learn the correct class label from the pattern of changes as k is reduced.

Acknowledgments This material is based upon work supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216. This research was also sponsored by a grant from the Lockheed Martin Corporation.

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[2] Nicholas Carlini and David A. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. *CoRR*, abs/1801.01944, 2018.

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[4] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2774–2783, 2017.

[5] Peter Bartlett, Sebastien Bubeck, and Yeshwanth Cherapanamjeri. Adversarial examples in multi-layer random reLU networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[6] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[7] Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The dimpled manifold model of adversarial examples in machine learning. *CoRR*, abs/2106.10151, 2021.

[8] Dan Hendrycks and Kevin Gimpel. Visible progress on adversarial images and a new saliency map. *CoRR*, abs/1608.00530, 2016.

[9] Malhar Jere, Sandro Herbig, Christine Lind, and Farinaz Koushanfar. Principal component properties of adversarial samples. *CoRR*, abs/1912.03406, 2019.

[10] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[11] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Dimensionality reduction as a defense against evasion attacks on machine learning classifiers. *ArXiv*, abs/1704.02654, 2017.

[12] Xin Li and Fuxin Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5775–5783, 2017.

[13] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *ArXiv*, abs/1702.04267, 2017.

[14] Reuben Feinman, Ryan R. Curtin, Saurabh Shintre, and Andrew B. Gardner. Detecting adversarial samples from artifacts. *ArXiv*, abs/1703.00410, 2017.

[15] Nicholas Carlini and David Wagner. *Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods*, page 3–14. Association for Computing Machinery, New York, NY, USA, 2017.

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[17] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *arXiv: Computer Vision and Pattern Recognition*, 2019.

[18] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.

[19] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[20] Yifeng Li, Lingxi Xie, Ya Zhang, Rui Zhang, Yanfeng Wang, and Qi Tian. Defending adversarial attacks by correcting logits. *ArXiv*, abs/1906.10973, 2019.

[21] Manish Reddy Vuyyuru, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically inspired mechanisms for adversarial robustness. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2135–2146. Curran Associates, Inc., 2020.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[26] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 3–14, 2017.

[27] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018.

A Detection of Carlini-Wagner attack with VGG-11 model

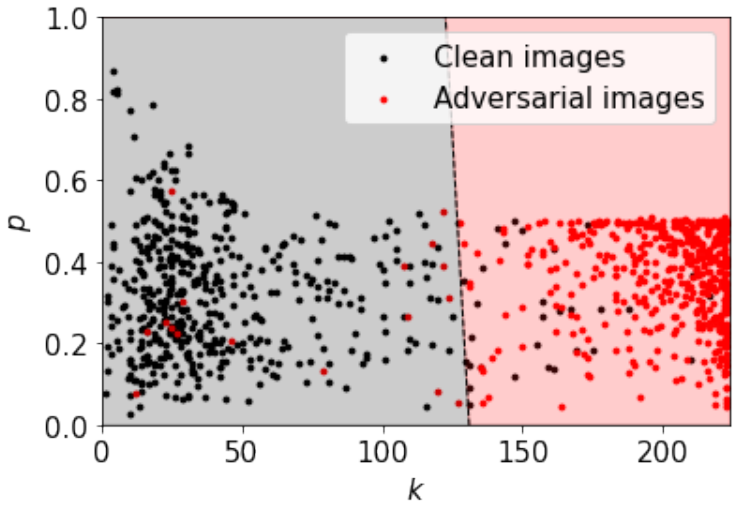
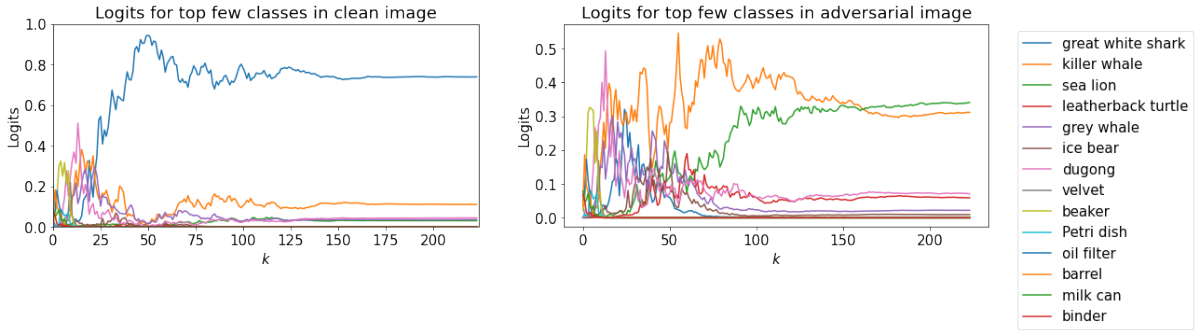


Figure 3: Linear classification of (k, p) points for Carlini-Wagner attack on pre-trained VGG-11 model.

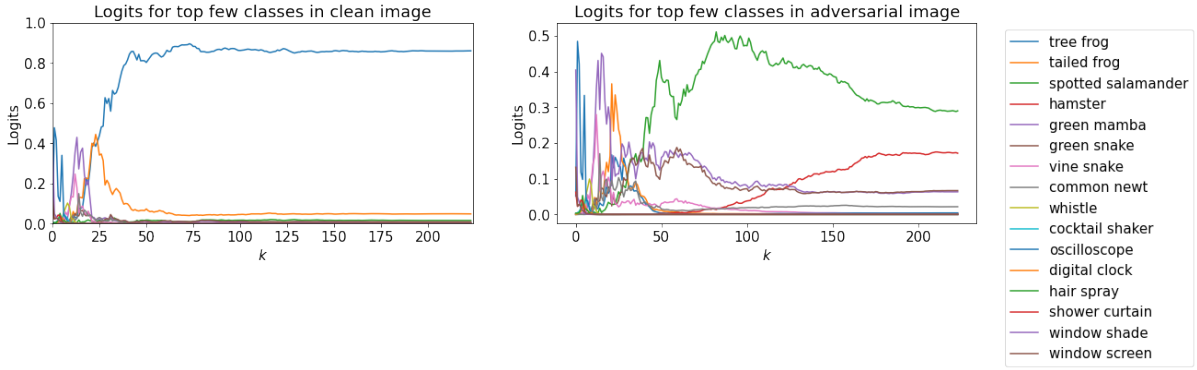
B Top few logits for different number of principal components k

B.1 FGSM attack on ResNet-18 model

(a) Logits for a sample image of a great white shark



(b) Logits for a sample image of a tree frog



(c) Logits for a sample image of a boa constrictor

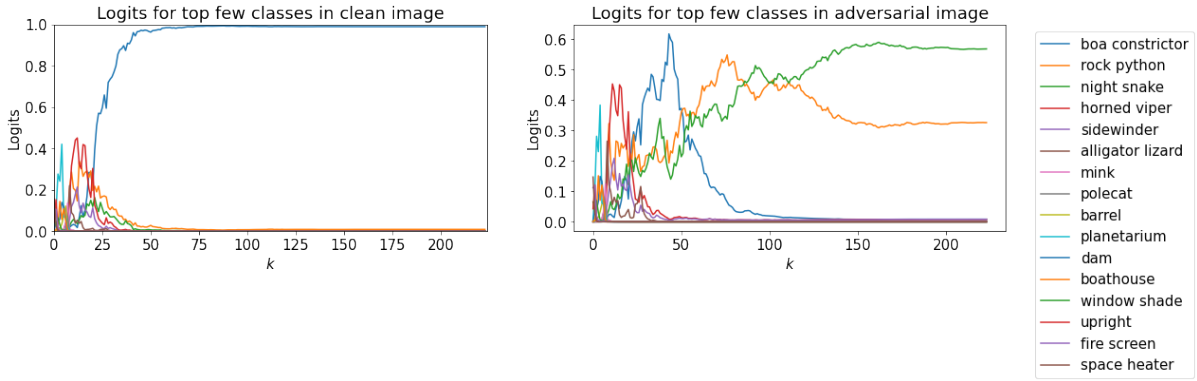
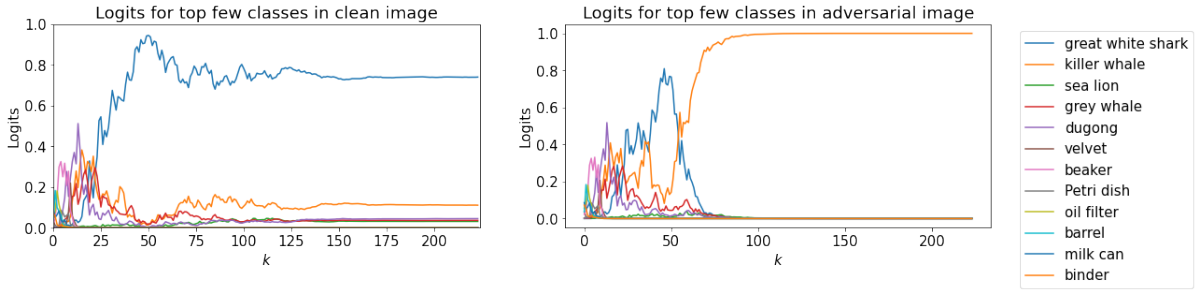


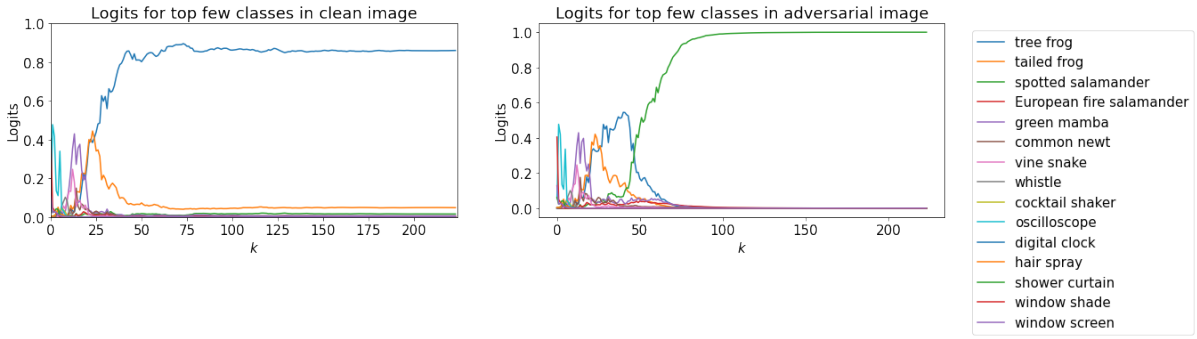
Figure 4: Logits of top few classes as the number of principal components k changes, for ResNet-18 model on clean images (left) and adversarial images using the FGSM attack(right) with $\epsilon = 0.03$.

B.2 PGD attack on ResNet-18 model

(a) Logits for a sample image of a great white shark



(b) Logits for a sample image of a tree frog



(c) Logits for a sample image of a boa constrictor

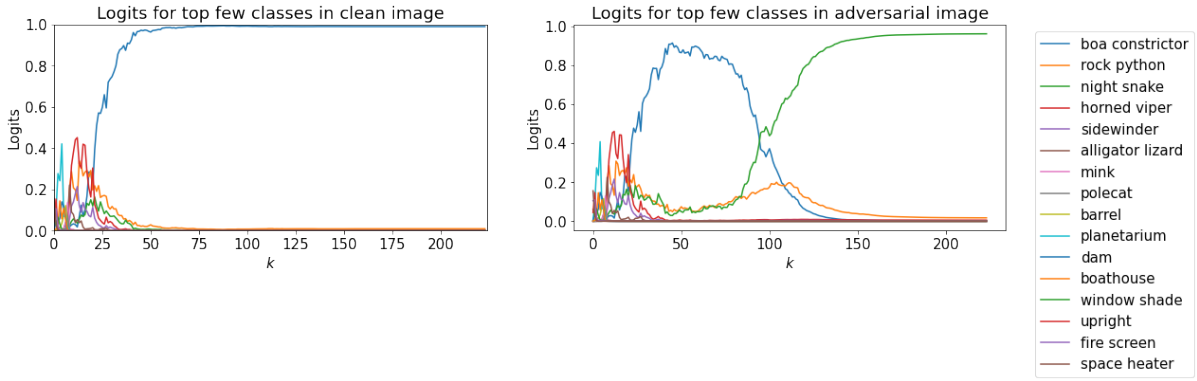
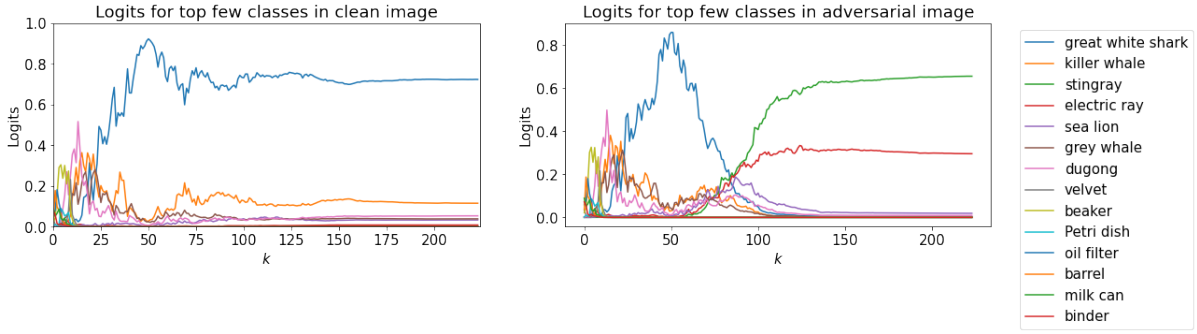


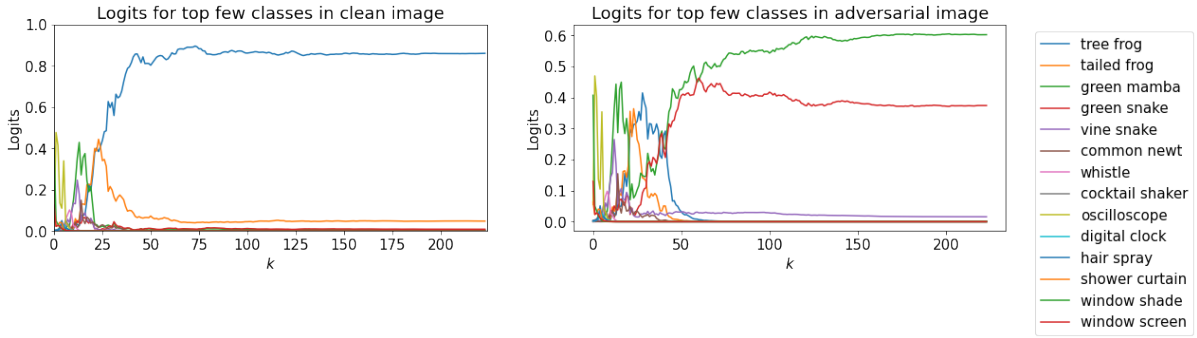
Figure 5: Logits of top few classes as the number of principal components k changes, for ResNet-18 model on clean images (left) and adversarial images using the PGD attack(right) with $\epsilon = 0.01$.

B.3 DeepFool attack on ResNet-18 model

(a) Logits for a sample image of a great white shark



(b) Logits for a sample image of a tree frog



(c) Logits for a sample image of a boa constrictor

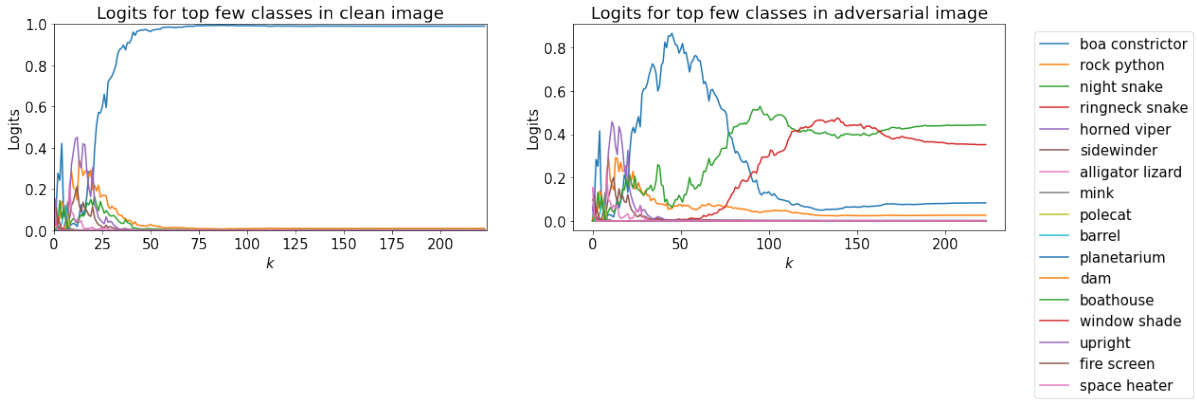
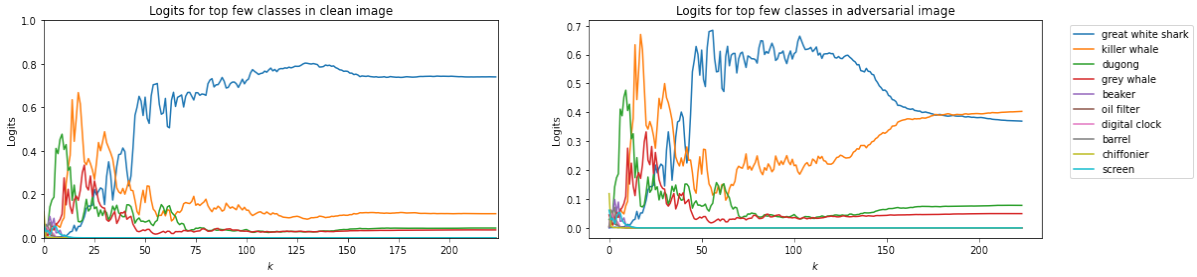


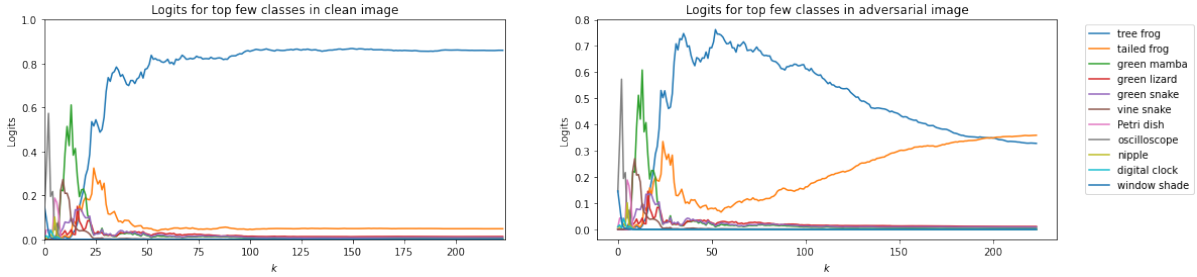
Figure 6: Logits of top few classes as the number of principal components k changes, for ResNet-18 model on clean images (left) and adversarial images using the DeepFool attack(right) with $\epsilon = 0.01$.

B.4 Carlini-Wagner attack on ResNet-18 model

(a) Logits for a sample image of a great white shark



(b) Logits for a sample image of a tree frog



(c) Logits for a sample image of a boa constrictor

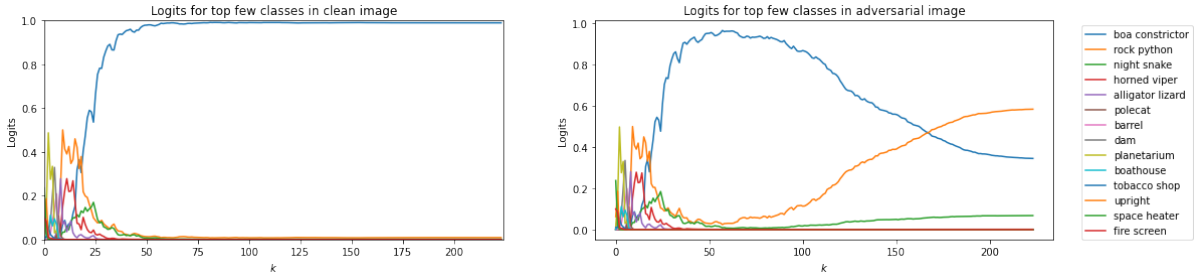
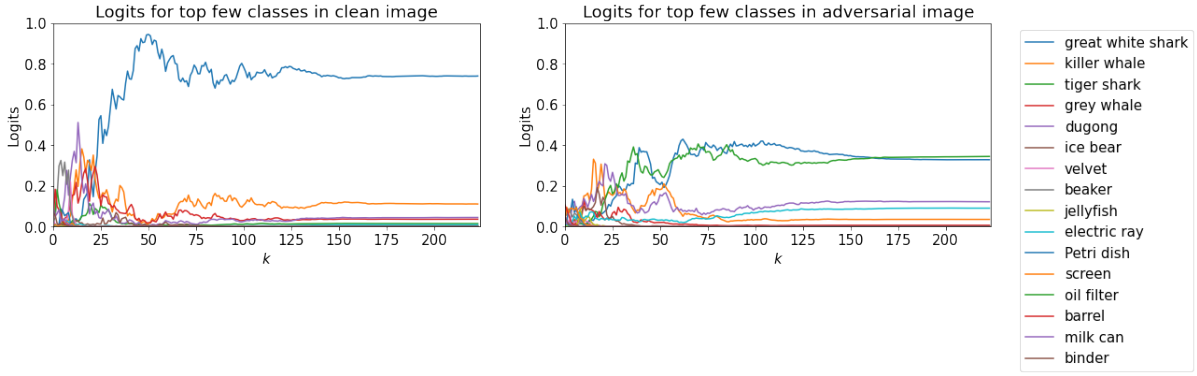


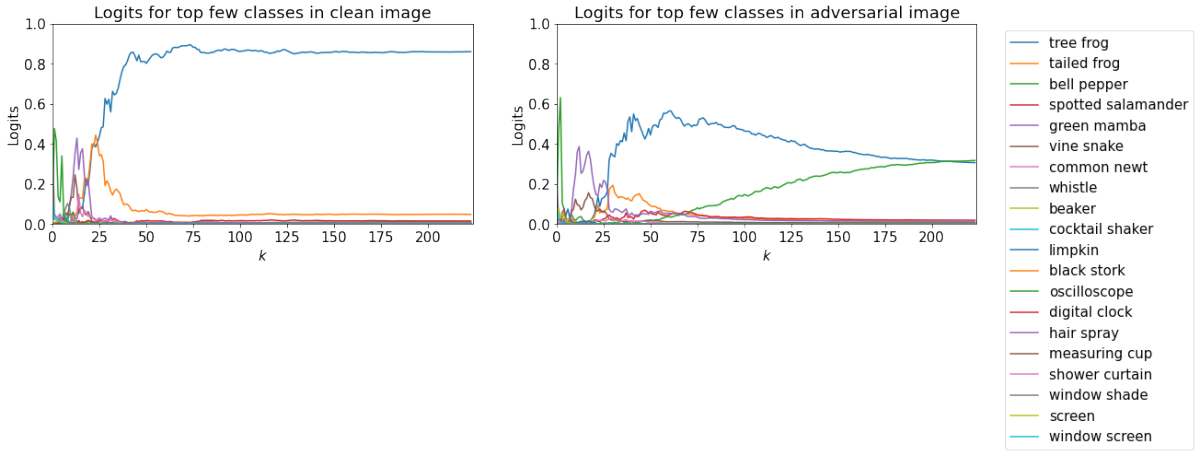
Figure 7: Logits of top few classes as the number of principal components k changes, for ResNet-18 model on clean images (left) and adversarial images using the Carlini-Wagner attack(right).

B.5 Carlini-Wagner attack on VGG-11 model

(a) Logits for a sample image of a great white shark



(b) Logits for a sample image of a tree frog



(c) Logits for a sample image of a boa constrictor

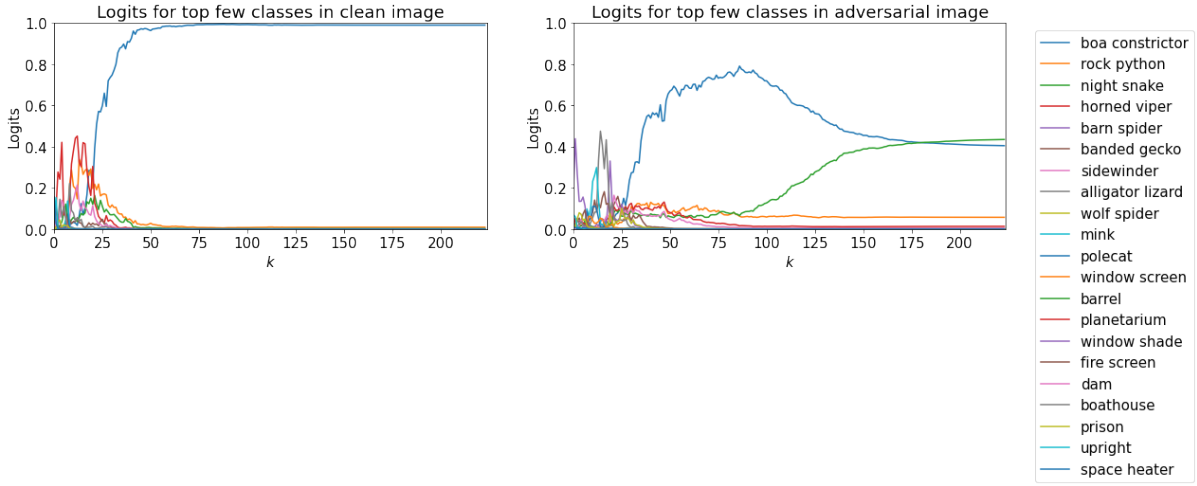


Figure 8: Logits of top few classes as the number of principal components k changes, for VGG-11 model on clean images (left) and adversarial images using the Carlini-Wagner attack(right).