Insights from Attacking Interpretable Models: Style Transfer and Input Thresholding

Haizhong Zheng*, Junghwan Kim*, Do Jun Min*, Jihun Lim**
Department of Computer Science* and Electrical Engineering**
University of Michigan
Ann Arbor, MI 48109
{hzzheng,kimjhj,dojmin,jihunlim}@umich.edu

Abstract

We study the adversarial robustness with the emphasis on interpretability. We visualize the hidden space of deep models under adversarial attacks and defense. From the visualization, we formulate two hypotheses: adversarial vulnerability comes from texture biased utilization of visual cues and from mismatch in data distributions between natural and adversarial images. We propose (1) to train on style transferred images to bias model towards object shape and (2) to project images into natural image manifold before feeding into models. In our experiments, we observe performance improvement by training on style transferred images and get certified robustness using projection onto natural image manifold.

1 Introduction

Although deep learning has been successful in modeling complex mappings as evidenced by lots of breakthroughs in computer vision [18, 13], speech recognition [14] and natural language processing [17], it was demonstrated [27, 11, 19, 4, 21] that deep models are vulnerable to adversarial attack which is small imperceptible perturbation intentionally designed to deceive the model.

Understanding and defending against adversarial attack is practically significant to forecast and prevent security and safety problems of real-world machine learning applications. On the other hand, it provides evidence that our current deep models do not achieve the real perception. Since adversarial attacks do not break human perception system, the deep models seem to detect different visual cues than what human perception depends on. Better understanding of adversarial attack vulnerability provides insight on the generalization capability, the failure modes and the implicit modeling assumptions of the deep neural networks which lead to the potential model design with more desirable and more human-like generalization capability.

Most recent approaches towards the adversarial robustness are based on the mini-max optimization framework [21]. In the framework, the model is either explicitly trained on adversarially perturbed data [11, 4, 19, 21] or trained to minimize upper bound of loss under every possible attack [23, 12, 30, 31]. While providing mathematically principled approach to adversarial robustness and showing better performance empirically, these methods require prior knowledge on possible attacks and are difficult to generalize to untrained attacks. Moreover, there is no clear interpretable connection between the real perception and the models learned under the framework.

In this project, we study the interpretable approaches to achieve adversarial robustness of deep neural networks. Especially, we reconstruct image from intermediate hidden space and observe the effect of adversarial training and adversarial attacks on reconstructed images. Based on our observation, we come up with two methods for adversarially robust model training: train models on style transferred images so that the model is biased towards global object shapes rather than local texture; project

Preprint. Work in progress.

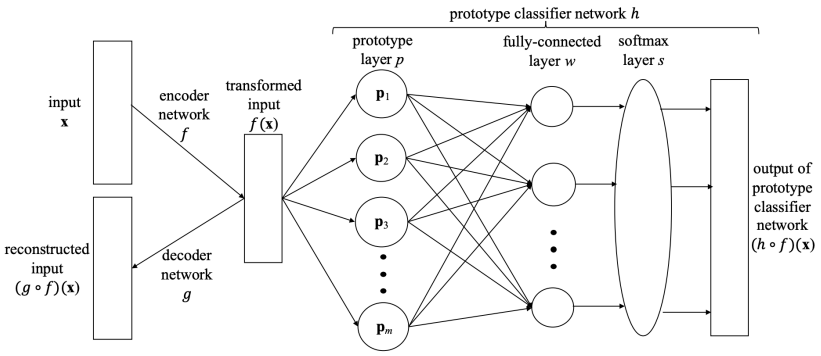


Figure 1: Network Architecture of ProtoNet

input images into natural image manifold where deep models already works well. In this report, we describe two observations that leads to the above ideas and empirically verify our methods.

Our contributions are as follows:

- We propose a method to visualize the effect of adversarial training and adversarial attacks on hidden spaces. Our method utilize reconstruction of input using decoder network.
- We observe two interesting phenomena using the visualization methods that we propose. Adversarial attack changes the shape of the reconstructed input so that the perturbation is perceptible and does not satisfy the norm bound. On the other hand, the hidden space after adversarial training projects adversarial inputs back to natural input space.
- Based on the observation, we propose to bias models towards global shapes rather than local textures by training on style transferred images. In addition, we propose a simple projection method that maps adversarial image back to the manifold of natural images.

2 Adversarial Attack on Interpretable Network and Style Transfer Augmentation

Background. We investigate the effect of adversarial attacks on the classification process of interpretable models. Specifically, we train and visualize the interpretable model proposed in [20] (which we denote ProtoNet) where each prototype neuron in prototype layer corresponds to one representative example of a high-level concept. The ProtoNet architecture is shown in Figure 1.

The ProtoNet is composed of autoencoder and neural network linked through clustering in common encoding space. The core idea of the ProtoNet is to cluster encoding around prototypes while encoding can reconstruct input and prototypes can classify the input. The loss function is composed of four different parts in order to train a network with accuracy, reconstructability and interpretability:

$$\begin{aligned} L((f,g,h),D) &= E(h \circ f, D) + \lambda R(g \circ f, D) \\ &\quad + \lambda_1 R_1(p_1, p_2, ..., p_m, D) + \lambda_2 R_2(p_1, p_2, ..., p_m, D) \end{aligned} \tag{1}$$

where f, g, h are corresponding components shown in Figure 1, D is the dataset, p_i is the prototype and $\lambda, \lambda_1, \lambda_2$ are hyperparameters.

The first part of the loss function is the simple cross-entropy loss of classification error:

$$E(h \circ f, D) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} -\mathbb{1}[y_i = k] log((h \circ f)_k(x_i)) \tag{2}$$

To train the decoder, the second loss treats the encoder and the decoder part as an autoencoder and calculate the loss between the original image and the reconstructed image:

$$R(g \circ f, D) = \frac{1}{n} \sum_{i=1}^{n} ||(g \circ f)(x_i) - x_i||_2^2 \tag{3}$$

Table 1: The architecture of the autoencoder

Encoder	Decoder
Input(28*28 grey image)	deconv3-10
conv3-32	deconv3-32
conv3-32	deconv3-32
conv3-32	deconv3-32
conv3-10	Output(28*28 grey image)

Figure 2: Prototypes of ProtoNet

The third and fourth losses couple encoding of autoencoders to prototypes:

$$R_1(p_1, p_2, ..., p_m, D) = \frac{1}{m} \sum_{j=1}^{m} \min_{i \in [1,n]} ||p_j - f(x_i)||_2^2 \tag{4}$$

$$R_1(p_1, p_2, ..., p_m, D) = \frac{1}{n} \sum_{i=1}^{n} \min_{j \in [1,m]} ||f(x_i) - p_j||_2^2 \tag{5}$$

Encodings are clustered around prototypes and each prototype corresponds to at least one encoding.

Observations. We reconstruct the adversarial latent encoding under PGD attack [21] using our decoder in ProtoNet to find how the adversarial perturbation influences the latent space. Since PGD attack applies multiple iterations of gradients, with more iterations, the adversarial example cause higher confidence for the wrong label and lower confidence for the orignal label. So we perform the attack with different attack iterations and see how the latent feature changes with the attack iterations.

We used the MNIST dataset that contains 60,000 training and 10,000 test images. The PGD attack is limited with L_∞ of 0.3. For training and testing, we used ProtoNet architecture with autoencoder having 40 encoding dimension with 15 prototypes. The autoencoder architecture is shown in Table 1 and we use one full-connected layer on top of encoder for classification. Figure 2 shows the reconstructed images of 15 prototypes of ProtoNet on MNIST dataset. Each prototype corresponds to a specific subclass of a label.

Figure 3 shows the results of adversarial attacks on reconstructed images from latent space. With more attack iterations, there are more attack perturbations in the adversarial examples. Although the input adversarial examples are unrecognizable for humans, reconstructed images show clear perturbation towards the target label. For example, in the first column of Figure 3, after 100 attack iterations, the latent space of images are closer to target label 6. There is a targeted attack from 8 to 0, we can find that the adversarial perturbation removes the central part of 8, which make the latent space be more closer to 0.

Hypothesis and Proposed Method. The result in Figure 3 suggests that the model translates the adversarial perturbation into shape deformation. We hypothesize that if the model focused on capturing the global shape information in the first place, then the adversarial attack will fail to deform the latent space. We investigate the adversarial robustness of models trained on dataset augmented with style transfer in order to test this hypothesis. Style transfer alters local statistics of images but preserves the global shapes and semantics. In the process, some visual cues, that models might depend on but not related to semantics, are lost. By leveraging dataset using multiple styles, we expect the models to be biased towards invariant shapes rather than changing textures.

In addition, models trained on stylized images do not assume any adversarial attacks on training. Customized methods for specific adversarial attacks have a risk of overfitting. The previous history of arm race between adversarial attack and defense have shown that defense against existing attacks are

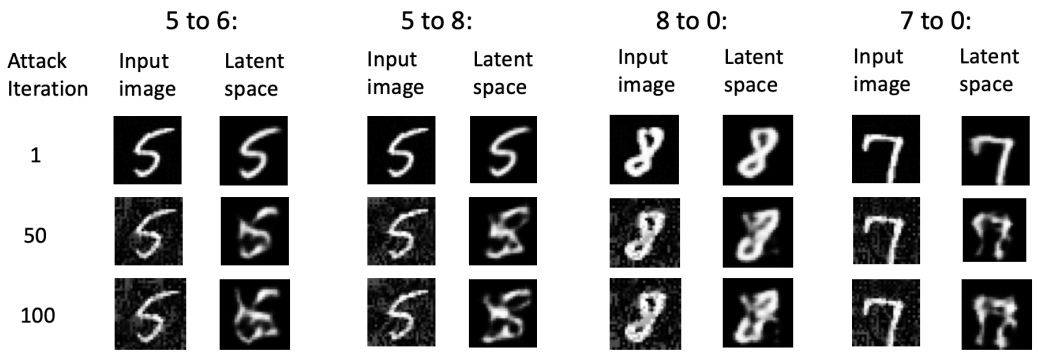


Figure 3: Reconstruction of adversarial images with different attack iteration.

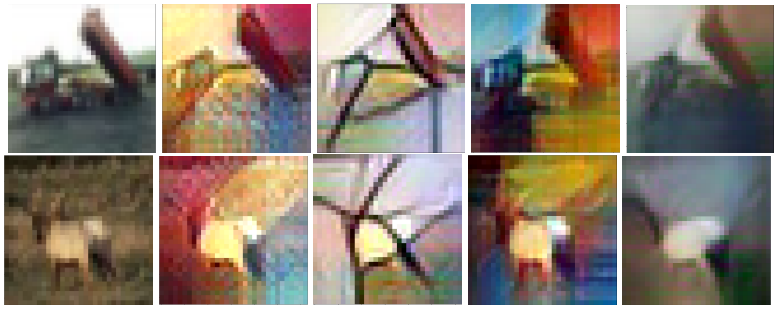


Figure 4: The first row is an image of a truck and four stylized versions of it. The second row is an image of a deer with the same four stylized versions.

vulnerable to unseen new attacks. In addition, most of the adversarial defense methods are based on mini-max framework which seems unnatural since any inperceptible attacks are the same for human eyes. In comparison, stylized image augmentation is a natural way to bias models towards object shape independently from any specific adversarial attack algorithms.

Finally, we note that there are many efficient style transfer algorithms and libraries available [8, 9, 16, 15, 29], making style transfer a relatively cheap way to augment datasets. Moreover, by utilizing various style templates as well as style transfer algorithm, variability of the resulting augmented datasets is easy to achieve.

Experimental Results. We test the performance of the deep neural networks trained with the stylized image augmentation. We varied the ratio between the orignal images and stylized images to see the effect of the augmentation. We used the CIFAR10 dataset that contains 50,000 training images and 10,000 test images. In the implementations, we use the fast-neural-style algorithm as introduced in [16] and implemented in PyTorch [1]. One notable feature of this style transfer algorithm is that the transfer network is trained with respect to a fixed style template. That is, the algorithm only transfer the style of the image it was trained on. The lack of flexibility in terms of available styles is offset by the fast speed of the algorithm, and we chose four pretrained models as shown in Figure 4.

One crucial hyperparameter of the style transfer augmentation is the ratio of unstylized examples to stylized examples. One can imagine there is a trade-off between robustness of the learned classifier and its accuracy on test examples, since increasing the fraction of stylized image in the dataset will likely reduce its performance on non-adversarial examples. We use PGD (Projected Gradient Descent) on the cross-entropy loss, which currently holds the top place at Madry [21]'s public black-box attack leaderboard [2]. We used the ResNet-50 architecture, which is known to achieve $\sim$94% accuracy on the original CIFAR10 dataset.

The test accuracy of the model trained with stylized image augementation is shown in Table 2. The learned classifier consistently performs poorly even on the original dataset. We come up with the following reasons for the poor performance:

[1] `https://github.com/pytorch/examples/tree/master/fast_neural_style`

[2] `https://github.com/MadryLab/cifar10_challenge`

Table 2: The performance of model trained with stylized image augmentation

Original: Stylized Ratio	Accuracy on Original Test Set	Accuracy on Adversarial Test Set
1:1	0.2559	0.0911
1:2	0.2830	0.0935
1:3	0.3157	0.0971
1:4	0.3762	0.0627

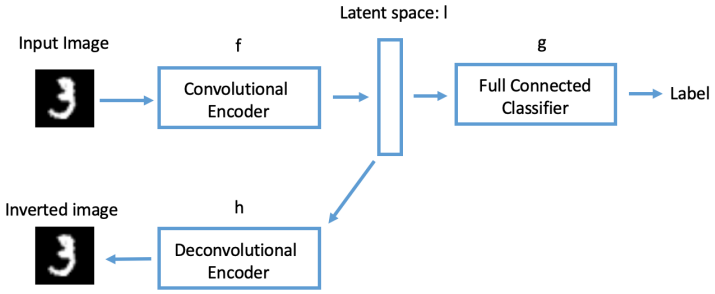


Figure 5: MNIST network

- CIFAR-10 images are too small for the style transfer technique to have the desired fine-grain effect of perturbing the image. In other words, the stylized images are too distorted for the classifier to learn anything from it. As depicted in Figure 4, the style transferred images show too much distortion so that it is difficult for even human to distinguish the object in the image. Since the performance on original dataset is very poor while state-of-the-art systems achieve human-level performance, we believe that larger image dataset should be used in this experiment.
- The performance on adversarial dataset is even lower than the already low performance on original dataset. Style transfer is different from adversarial perturbation in that the former explicitly seeks to preserve the semantic content of the input while the latter attempts to mislead the classifier as to the class/label represented by the image.

3 Understanding Adversarial Training and Input Thresholding

Background. We study the effect of adversarial training on the difference in hidden layer activations between adversarial and natural images. To visualize the hidden space, we invert hidden activations using the architecture shown in Figure 5. The input image is mapped into latent hidden space by the convolutional encoder f. Then, the fully connected classifier g on top of f predicts the labels of the input image and the deconvolutional decoder h reconstructs the image.

We first train a natural classification model ($g \circ f$) on the natural training dataset of MNIST and separately train an adversarially robust model ($g^* \circ f^*$) using adversarial augmentation as in Madry et al. [21]. Then, we freeze the parameters in f, g, f^*, g^* and train the adaptive autoencoders ($h \circ f$ and $h^* \circ f^*$) for both models on adversarial examples against the autoencoder. Our design aims to train the encoder f, f^* on just classification task while decoders h, h^* reconstruct the original input images from the adversarially perturbed latent features.

Observations. We reconstruct the latent features of adversarial attacks on both models to visualize the effect of the adversarial training. We used the MNIST dataset that contains 60,000 training and 10,000 test images. We use PGD attack($L_\infty(\epsilon = 0.3)$) to generate perturbation images. The structure of our model is shown in Table 3. The reconstructions of adversarial images with respect to two models are shown in Figure 6. While the reconstruction from natural model f, g, h still contains the noise in the background, the reconstruction from the adversarially robust model f^*, g^*, h^* is more similar to the original image. From this result, we hypothesize that the adversarially trained model filters the adversarial noise information in the background of the image.

Hypothesis and Proposed Method. From the observations, we hypothesize that the reason for poor performance of the natural model on adversarial examples is the distribution mismatch between

Table 3: The architecture of the model to interpret adversarial training

Encoder(f)	Decoder(h)	FC classifier(g)
Input(28*28 grey image)	deconv5-64	FC(3136 * 1024)
conv5-32	deconv5-32	FC(1024 * 10)
conv5-64	Output(28*28 grey image)	Label

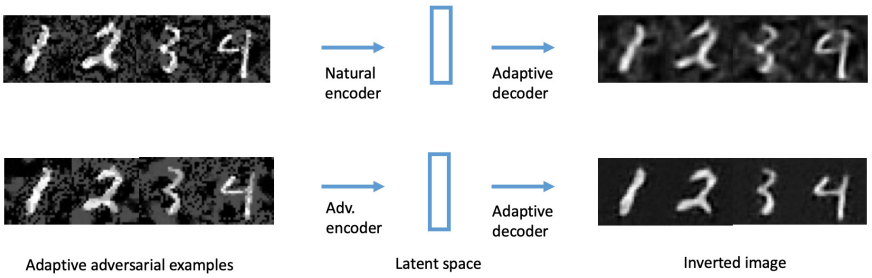


Figure 6: Comparison on inverted image between natural network and adversarially trained network.

adversarial and natural images. That is, the adversarial examples do not exist in the natural image manifold. We show the pixel distribution for further support our hypothesis. In the natural MNIST test dataset, most pixels(80.72%) are black pixels. On the contrary, in the adversarial MNIST test dataset, only 30.38% of pixels are black. In addition, Figure 7 indicates even the distributions of grey pixels are quite different in the two datasets.

In order to project the adversarial images back to a natural image manifold, we propose a thresholding function (equation 6) to binarize the MNIST dataset so that the transformed image gets closer to the natural manifold.

$$T(x)_{i,j} = \begin{cases} 0 & x_{i,j} \leq \mu \\ 1 & o.w. \end{cases} \tag{6}$$

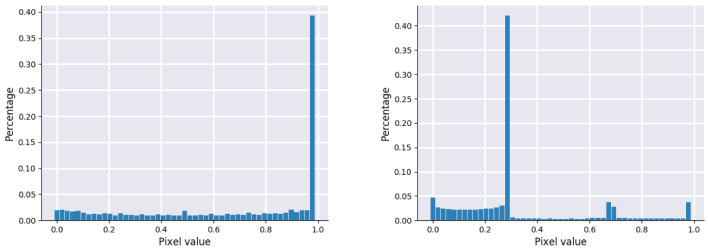
Since the piecewise function is non-differentiable at transition points, we use the $\frac{\partial L}{\partial T}$ as the gradient to perform the attack.

Experimental Results. We evaluate the performance of the input thresholding against adversarial attacks. We again used the MNIST dataset and the architecture as in Table 3. We evaluate our defense model on PGD attack with L_∞ bound $\epsilon = 0.1$. We choose the thresholding parameter $\mu = 0.38$ by doing a grid search so that there are least pixels whose value locates in $[\mu - \epsilon, \mu + \epsilon]$. In this way, we shrink the adversarial manifold or the number of possible adversarial attacks by the most.

We get 96.16% accuracy against the adversarial dataset which is comparable to the state-of-the-art 96.4% [6]. The reason behind the high accuracy is that the input thresholding projects the MNIST dataset to a binarized space. This projection also reduce the size of the adversarial manifold. Within a smaller manifold, it becomes much harder to find the adversarial examples to fool the network.

4 Related Works

First adversarial attack and defense were based on the local search of adversary using gradient information. FGSM attack [11] constructs adversary by adding the sign of one gradient of loss. C&W attack [4] also uses the gradient sign for perturbation but with the distance penalty to control the magnitude of perturbation. PGD attack [19] iteratively apply gradient ascent with projection to find the worst adversary. Despite simple and approximate nature of these heuristic attacks, these attacks are effective in fooling regularly trained models (without adversarial training); the model trained on PGD adversary forms strong baseline defense method [21]. Although the models trained under minimax framework become robust against adversarial attacks, humans do not require adversarial training to be robust against adversarial attacks. Therefore, the models trained under minimax framework does not provide the evidence that our model achieves the "real" perception.



(a) Natural MNIST test set (b) Adversarial MNIST test set

Figure 7: The distribution of grey pixels in different datasets

In order to overcome the unnaturality of adversarial training, we require the training process to be more similar to human and the prediction to be more interpretable. Recent works on computer vision indicate that neural network models rely more on texture feautures rather than global object shapes [1, 3, 10]. It was demonstrated that with dataset augmented with style transfer, the trained model show more robustness against common perturbations [10]. However, whether the models trained in this way are robust against adversarial perturbations is unanswered yet.

The study of adversarial attacks in terms of interpretability have attracted interest recently [25, 7, 28, 32, 33]. Most of the methods are based on the visualization of input regions [26, 34, 22, 35, 24] and neurons associated with interpretable high-level concept [2, 5, 20]. Tao et al. [28] identify witness neurons which are strongly coupled with the input landmark and detect adversarial attacks by the discrepency between the orignal model prediction and the witness-strengthened model prediction. Xu et al. [33] induce group sparsity in adversarial perturbation so that perturbed region represents the object of the original class or of the attack target class.

Our approach leverages interpretable deep model to understand the dynamics in attack and defense and to design interpretable defense strategies. This has never been addressed in previous works.

5 Conclusion

In this project, we study the adversarial attacks in connection with interpretability. We investigate into two directions motivated by the effect of adversarial attacks and defense on the reconstructed images. We make two hypotheses and show the experimental results to verify our hypotheses. Our simple method based on the second hypothesis achieves impressive performance on MNIST dataset.

Author Contribution Statement

Haizong Zheng and Junghwan Kim developed the theoretical framework. Haizong Zheng performed the interpretable model experiments and Do June Min performed the style-transfer experiments. Haizong Zheng, Junghwan Kim, Jihum Lim, and Do June Min analyzed the results. Haizong Zheng, Junghwan Kim, Jihum Lim, and Do June Min wrote the article.

References

[1] Pedro Ballester and Ricardo Matsumura Araujo. On the performance of googlenet and alexnet applied to sketches. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017.

[3] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *International Conference on Learning Representations*, 2019.

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

[5] Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: deep learning for interpretable image recognition. *arXiv preprint arXiv:1806.10574*, 2018.

[6] Francesco Croce, Maksym Andriushchenko, and Matthias Hein. Provable robustness of relu networks via maximization of linear regions. *arXiv preprint arXiv:1810.07481*, 2018.

[7] Yinpeng Dong, Hang Su, Jun Zhu, and Fan Bao. Towards interpretable deep neural networks by leveraging adversarial examples. *arXiv preprint arXiv:1708.05493*, 2017.

[8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations*, 2019.

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

[12] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

[13] 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.

[14] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[17] N Kalchbrenner, E Grefenstette, and Philip Blunsom. A convolutional neural network for modelling sentences. In *52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2014.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[20] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[21] 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.

[22] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016.

[23] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pages 3575–3583, 2018.

[24] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*, 2016.

[25] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[26] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[27] 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.

[28] Guanhong Tao, Shiqing Ma, Yingqi Liu, and Xiangyu Zhang. Attacks meet interpretability: Attribute-steered detection of adversarial samples. In *Advances in Neural Information Processing Systems*, pages 7728–7739, 2018.

[29] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017.

[30] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5283–5292, 2018.

[31] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, pages 8410–8419, 2018.

[32] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *International Conference on Learning Representations*, 2018.

[33] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. Structured adversarial attack: Towards general implementation and better interpretability. *International Conference on Learning Representations*, 2019.

[34] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[35] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.