Robust Image Classification

Pravin Ravishanker
University of California, Berkeley
pravin.ravishanker@berkeley.edu

Harshayu Girase
University of California, Berkeley
harshayugirase@berkeley.edu

Rahul Gupta
University of California, Berkeley
rahul.gupta2020@berkeley.edu

Abstract

In this work, we aim to develop a robust image classification system, and validate our models on the Tiny ImageNet dataset, a subset of the ImageNet dataset. We fine-tune a Wide ResNet50 (pretrained on ImageNet) via transfer learning on an augmented dataset with real world perturbations and various adversarial attacks. Our extensive experimentation tests various combinations of models and datasets, exploring training networks for classifying normal images, augmented (perturbed) images, and images simulating adversarial attacks such as the Fast Gradient Sign Method (FGSM) (4) and the Basic Iterative Method (BIM) (5). Furthermore, we experiment with disentangling different adversarial attacks by training adversarial attack-specific image classifiers.

Ultimately, we concluded that **model hardening** was the best approach to developing robust classification models: our strongest classification model was trained on all possible input image types (normal, augmented, and adversarial). We observed that models only trained on images generated through specific adversarial attacks may not be able to generalize to normal images. We did not get satisfactory results when trying to classify the specific types of adversarial attacks that certain input images were subjected to, thus making robust classification via adversarial attack-specific classifiers infeasible.

1 Problem Statement and Background

Image classification is a highly researched field in the machine learning community, especially since the ImageNet Large Scale Visual Recognition Challenge (3) was released in 2009. The ImageNet challenge aimed to improve the performance of AI models on vital tasks such as image classification (object category classification), single-object localization, and object detection (bounding boxes for determining multiple object positions). (3) Image classification has a wide array of computer vision applications pertaining to robotics, autonomous driving, medical imaging, and training reinforcement learning agents.

In the past few years, research has focused on developing robust computer vision systems, as data in the real world is often susceptible to perturbations and shifts. Additionally, in order to protect against potential attacks, these classification systems also need to consider adversarial inputs, which confound output predictions of neural networks. Adversarial examples are often hard to defend against because it is difficult for machine learning model designers to create a model of all adversarial example creation processes. A ML model designer might be able to block one kind of adversarial attack, but might not be able to defend against another adaptive attacker aware of the ML model's limitations. Robust image classification is constantly evolving to account for new types of adversarial attacks and real-world data perturbations which may affect the distribution of test time data.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

1.1 Tiny ImageNet Dataset

In this work, we aim to develop a robust classifier and validate our models on the Tiny ImageNet dataset, a subset of the ImageNet dataset.

The Tiny ImageNet Dataset (12) consists of 200 classes, with each class having 500 training images, 50 validation images, and 50 test images. For the purpose of this project, the test dataset is unknown; however, images in the test dataset are known to have a different distribution from those in the train and validation sets. In particular, the test dataset contains noisy images with several unknown perturbations and adversarial inputs.

2 Data and Methodology

2.1 Baseline ResNet models

We examined the existing computer vision literature to survey various convolutional neural network architectures for the purposes of image recognition, including LeNet-5, AlexNet, VGG-16, Inception, ResNet, Xception, ResNeXt, and Wide ResNet CNN architectures.

In this project, we investigated different ResNet-based deep learning architectures for image classification.

ResNets, in particular, have had massive success, ever since residual networks won the ImageNet and the COCO competitions in 2015 and achieved state of the performance in object classification, object detection and segmentation, and image classification. (11) ResNets have been proven to show better generalization during transfer learning and have faster convergence due to residual links.

Several computer vision researchers have come up with neural architectures that build up upon ResNets, such as ResNeXts (Aggregated Residual Transformations for Deep Neural Networks) by Xie et al. (11) and Wide ResNets (Wide Residual Networks) by Zagoruyko et. al. (13) In ResNeXt neural architectures, each ResNeXt block has multiple paths of convolutions, with the number of paths determined by the cardinality hyperparameter. (11)

In Wide ResNet neural architectures, research has shown that the widening of residual blocks provides a more effective way of improving ResNet performance compared to increasing depth. (13) Zagoruyko et. al noticed that Wide ResNets can achieve similar performance as regular ResNets that are 50 times more deeper (in terms of number of layers), while being 2 times faster. (13) Wide ResNets have 2 parameters, the deepening factor L (number of convolutional layers per residual block in the Wide Resnet) and the widening factor K (which multiplies the numbers of features in convolutional layers). The following diagram shows how a Wide Residual Block is different from a normal residual block.

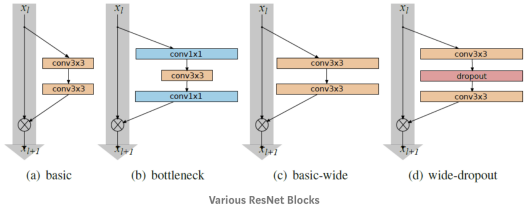

Figure 1: Structure of Wide ResNet blocks (13)

The Pytorch deep learning toolbox was utilized to compare the performance of three 50 layer ResNet models (ResNet-50, ResNeXt-50, Wide ResNet-50) pretrained on the ImageNet dataset and transfer learned on the Tiny ImageNet dataset. Two extra fully connected layers were added to the ResNet model, one with 512 neurons and a second with 200 neurons (equal to the number of output classes in Tiny ImageNet) to output classification predictions.

In the end, during transfer learning, the Wide ResNet-50 model had the highest validation accuracy on the validation set, so this model was utilized for the rest of the project.

2.2 Data Augmentation

To account for real world perturbations in images, we compared the performance of a Wide ResNet-50 model trained on the base training dataset to the performance of a Wide ResNet-50 model trained on data augmented with several common perturbations.

The base training set (100,000 images) was augmented with the following transformations to create an Augmented Training Dataset consisting of 100,000 images:

1. Affine Transforms: Translations, scaling, shearing
2. Grayscale: RBG image to grayscale image
3. Horizontal Flips
4. Vertical Flips
5. Small Rotations
6. Random Perspective Changes
7. Color Jitter: brightness, contrast, saturation, hue

We also applied the same data augmentations to the original validation data set (10,000 images), to create an Augmented Validation Dataset consisting of 10,000 images.

2.3 Adversarial Images

The recent success in image classification is primarily due to the advances made in learning-based approaches. However, several studies have investigated potential issues with deep networks in image classification. In (1), the authors demonstrate how networks are susceptible to various adversarial attacks, which can fool networks with perturbations that are unnoticeable to even humans. Recently Su et al. showed that with network parameter information adversaries can fool the network by simply changing one pixel (8).

The goal of evasion attacks in adversarial machine learning is to modify the input example to a classifier such that the classifier misclassifies that input example, in such a way that the modification is as small as possible. Untargeted attacks aim to misclassify the modified input without any constraints on what the new class should be, while targeted tasks aim to modify the input so that the new class that is predicted is determined exactly by the attacker.

For the purposes of adversarial defence, model hardening is utilized to augment the training data of the classifier with adversarial examples.

To make our model robust to adversarial inputs, we investigate 2 popular, easily generated white-box adversarial attacks described below:

1. Fast Gradient Sign Method
2. Basic Iterative Method

2.3.1 Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method can be applied in both targeted and untargeted attacks and aims to generate perturbations in original images with the smallest L_∞ norm.

In targeted settings (in which the attacker wants to force the classifier to classify the modified input with a particular class y), the adversarial perturbation $\psi(x, y)$ generated by the FGSM attack is given by:

$$\Psi(x, y) = -\epsilon * sign(\nabla_x \mathcal{L}(x, y)) \tag{1}$$

where $\epsilon > 0$ is the attack strength and y is the target class specified by the attacker. The adversarial sample is later generated with the following equation, where x_{min} and x_{max} indicate minimum and maximum clippings:

$$x_{adv} = clip(x + \Psi(x, y), x_{min}, x_{max}) \tag{2}$$

The untargeted version of the Fast Gradient Sign Method attack is as follows, where C(x) represents the model's current prediction on input x and $\rho(x)$ refers to the adversarial perturbation that is later added to the original image x.

$$\rho(x) = \epsilon * sign(\nabla_x \mathcal{L}(x, C(x))) \tag{3}$$

In the untargeted case, we aim to increase the classifier's loss function when the model continues to output the prediction C(x) for the input x. In this manner, the classifier is incentivized to output a different prediction that is not equal to C(x).

The Fast Gradient Sign Method requires only one gradient evaluation, and this attack can easily be applied to a batch of image inputs in the training set or the validation set. In this manner, FGSM can quickly generate a large batch of adversarial image examples.

2.3.2 Basic Iterative Method (BIM)

The Basic Iterative Method is an iterative variant of FGSM, as the name suggests. Instead of applying a single gradient update to the original image, this process is repeated for multiple iterations k. In our work, we set $k = 5$. Another difference between BIM and FGM is that the perturbation is projected onto an l_p-ball with radius ϵ, a technique known as PGD. As stated in Madry, et al. (7), in optimization literature, standard PGD uses the exact gradient while this PGD uses the steepest descent with respect to the L_∞ norm. As a result, PGD is considered the universal "first-order adversary", the strongest attack utilizing the local first-order information about the deep neural network.

The parameter ϵ is a knob that determines the strength of the attack. A higher epsilon results in a perturbation that is more clearly visible to the human-eye. In this work, ϵ = 0.2 was used for both FGSM and BIM adversarial image generators. This choice was determined to be a good middle ground that balanced strength and evasiveness, with the goal being to emulate life-like image inputs.

2.3.3 Engineering Image Classification Models that are Robust to Adversarial Attacks

After a Wide ResNet-50 was trained on 100,000 images from the Normal Training Set of Tiny ImageNet, the following adversarial images were generated:

1. 100,000 adversarial training images were generated using the FGSM method from the 100,000 normal training images, using model gradients from the Wide ResNet-50.
2. 100,000 adversarial training images were generated using the BIM method from the 100,000 normal training images, using model gradients from the Wide ResNet-50.
3. 10,000 adversarial validation images were generated using the FGSM method from the 10,000 normal validation images, using model gradients from the Wide ResNet-50.
4. 10,000 adversarial validation images were generated using the BIM method from the 10,000 normal validation images, using model gradients from the Wide ResNet-50.

Next, the following five Wide ResNet-50 deep learning models were trained:

1. Wide ResNet-50 Trained on Normal Images (Without Data Augmentation)
2. Wide ResNet-50 Trained on Augmented Images (described in Section 2.2)
3. Wide ResNet-50 Trained on FGSM-generated adversarial images
4. Wide ResNet-50 Trained on BIM-generated adversarial images
5. Wide ResNet-50 Trained on Normal Training Images, Augmented Images, and all Adversarially Generated Images (FGSM + BIM)

The performance of each of these 5 image classification models was tested on 4 validation sets:

1. Normal Validation Set (without data augmentation)
2. Augmented Validation Images (described in Section 2.2)
3. Adversarial FGSM-generated Validation Images
4. Adversarial BIM-generated Validation Images

Validation Accuracy on 4 Validation Sets				
Classification Model	Validation Images (No Augmentation)	Validation Images (With Augmentation)	FGSM Adversarially Generated Validation Images	BIM Adversarially Generated Validation Images
Wide ResNet (Trained on Normal Images)	.6489	0.4224	0.198	0.1067
Wide ResNet (Trained on Augmented Images)	**0.6747**	**0.596**	0.2898	0.2166
ResNet (Trained on FGSM Attack Adversarial Inputs)	0.4714	0.3007	0.4045	0.4141
ResNet (Trained on BIM Attack Adversarial Inputs)	0.4891	0.3063	0.4306	0.4505
Wide ResNet (Trained on Normal + Augmented + FGSM Adversarial + BIM Adversarial Images)	0.6189	0.5494	**0.4768**	**0.5205**

Table 1: This table depicts our various experimented models' accuracy on different validation datasets. Particularly, we experiment on normal, augmented, FGSM adversarial, and BIM adversarial images. This table shows that the model trained on all types of perturbations and adversarial attacks has the best overall performance.

2.4 Evaluation Metrics

For all models, we evaluate performance on validation datasets with the following 2 metrics:

1. Accuracy: (num correct samples / total samples)
2. Accuracy@k: (num correct samples in top k predictions / total samples)

2.5 Summary of Approach

Overall, our methods use the **model hardening** approach, in which a single robust classifier model is trained on all types of inputs (normal images, augmented (perturbed) images, and adversarial images generated by the FGSM and BIM algorithms). This robust classifier model can then handle any type of input image (normal, perturbed, or adversarial). Several papers have shown the effectiveness of training a single end-to-end model for multiple tasks such as image detection, segmentation, and tracking (9).

3 Results

Table 1 shows the predictive accuracy of the 5 image classification models on the 4 validation data sets (Normal Validation Images, Augmented Validation Images, FGSM Adversarially Generated Validation Images, and BIM Adversarially Generated Validation Images).

Our Data Augmentation procedures are effective, since Data Augmentation on the Training Set improves performance on the Validation Set. Figure 2 shows how the Top-K prediction accuracy on Normal Validation datasets and Augmented Validation datasets is higher for Augmented Training Models compared to Unaugmented Training Models.

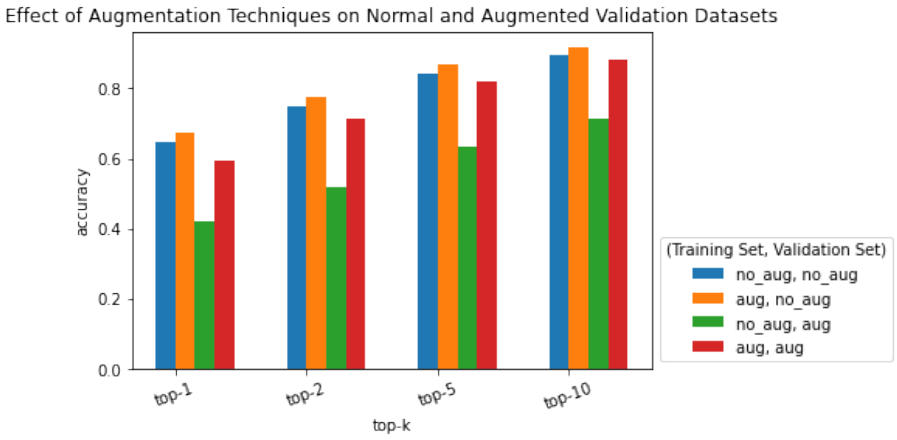

Figure 2: This figure illustrates the performance gain of training on an augmented dataset. We provide results on both augmented and non-augmented validation data.

Moreover, on Adversarially Generated Validation Images (FGSM + BIM), models trained on FGSM Attack Adversarial Images and on BIM Attack Adversarial Images outperform models trained on normal images or augmented images. Figure 3 shows that Adversarial Image Pretraining improves performance (Top K accuracy) on Adversarial Validation Datasets.

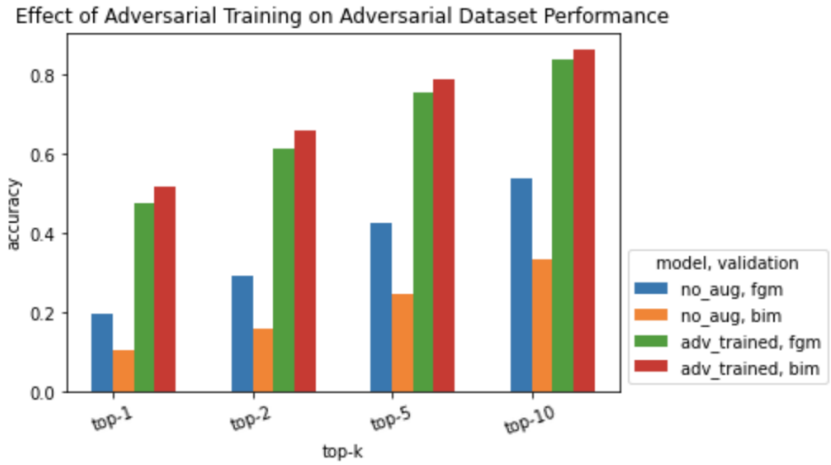

Figure 3: Here we observe that training a model specifically designed to classify adversarial images definitely helps, meaning it is possible to account for attacks by training on adversarial images.

Furthermore, we notice that Adversarial Training significantly improves performance on the Augmented Image Validation Datasets, as shown in Figure 4. This shows that training on adversarially attacked FGM and BIM examples allows us to handle augmented images that are subjected to random perturbations.

4 Discussion

The goal of this work was to develop a generalizable vision classifier that is robust to data perturbations and adversarial attacks that could occur in real world datasets.

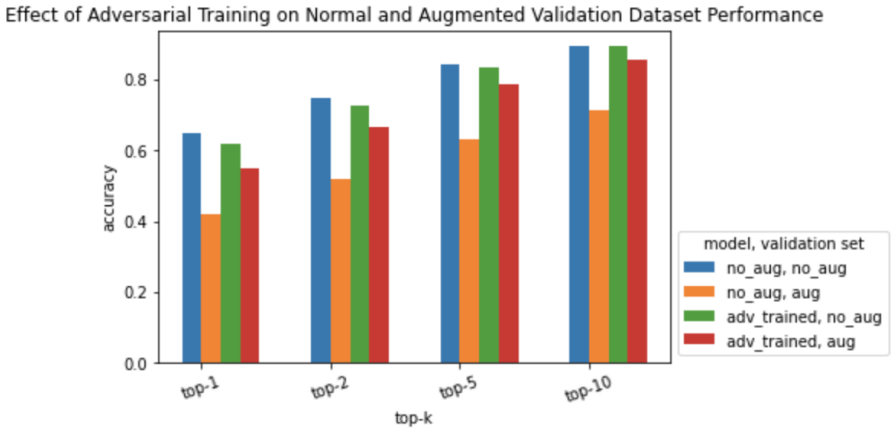

Figure 4

Real-world images often are collected at a variety of angles, camera orientations, and external environmental conditions. As a result, computer vision engineers can model real-world perturbations by considering Image Augmentations like Affine Transforms (Translation/Scaling/Shearing), Grayscaling, Horizontal Flips, Vertical Flips, Small Rotations, Perspective Changes, and Color jitters to brightness, contrast, saturation, and hue. In this project, we showed that training Convolutional Neural Networks on Augmented Image Data boosted Image Classification performance.

Sometimes, an adversary might try to attack our computer vision models by providing ingeniously designed input images that resemble the original class, but consist of special perturbations that force the model to output the wrong class. To solve the problem of adversarial attacks, machine learning designers must perform **model hardening**, in which robust models are trained on normal images and adversarial inputs.

Although we were not able to consider all possible adversarial attacks in this project, we considered two of the most prominent algorithms for adversarial input generation, the Fast Gradient Sign Method and the Basic Iteration Method. By inspecting the loss gradients of our base Wide ResNet models, these algorithms generate training and validation adversarial images that can later be used to develop robust computer vision models. In this project, we showed that training ResNets on Adversarial Images increased image classification performance on randomly perturbed Augmented Images and Adversarially Attacked Images.

The ResNet trained on BIM Adversarial Inputs had higher validation predictive accuracies than the ResNet trained on FGSM Adversarial Inputs, proving that BIM is a stronger adversarial attack than FGSM. This observations makes sense, for the Basic Iterative Method conducts multiple gradient steps in order to create an adversarial image. Given that each step's perturbation is norm-limited, having multiple gradient steps leads to multiple parts of the image being perturbed.

Runtime detection of adversarial inputs is an orthogonal technique to model hardening. We hypothesize that due to the large number of potential adversarial attacks, another potential approach is to train attack-specific classifier models that are capable of handling adversarial images generated from specific types of attacks. This **runtime detection** approach involves detecting specific adversarial attacks in input images via a deep learning classifier, and later calling the correct attack-specific classification model to handle the malicious input image.

We were curious if it was possible to be more fine-grained and classify an image based on the type of adversarial attack. We approached this problem by training a ResNet on an equally partitioned dataset of normal images, FGSM-generated images, and BIM-generated images. Our hypothesis rested on results from Xie and Carlini (10) (see Future Work). However, this model performed no better than random (33 percent accuracy) at distinguishing between natural, FGSM-generated, and BIM-generated images. So, we abandoned the **runtime detection** approach and used the **model hardening** approach in this project.

Overall, simply training a large network on all data types (normal images, augmented/perturbed images, adversarial images) had the best performance.

5 Future Work

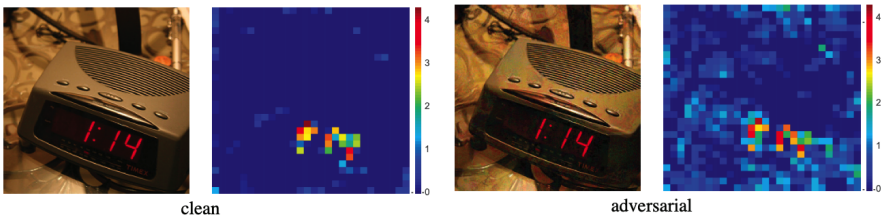

Figure 5: Feature Map Comparison for Adversarial Image Input Detection (10)

Xie, et al. (10) showed that the feature maps of adversarially-generated inputs are more information dense. Accordingly, we were surprised that a trained ResNet was not able to differentiate between normal images, FGSM adversarial images, and BIM adversarial images (effectively all images were predicted as benign). Future approaches could be training for more epochs or extracting the feature map from the res_3 block and training directly on this.

Another approach would be to try different model architectures for detecting adversarial image inputs. Bayesian neural network uncertainty is a detection method that relies on network randomness to distinguish adversarial images. The idea is that natural images will have the same label even under network randomness while adversarial images will be viewed inconsistently. Binary classification is done with an uncertainty threshold; Carlini's work (2) showed a 96% accuracy on MNIST with a 1% false positive rate. We would extend this work to a multi-class setting where multiple adversarial attacks are possible.

On the data front, one major limitation we had for this project was compute and time – each model took 10 hours to train. With longer training, we believe there is potential for more performance gains.

6 Extra Credit: Explainable AI

Explaining image classification predictions is possible through unsupervised image captioning. Feng et. al (6) in Unsupervised Image Captioning writes, "Deep neural networks have achieved great successes on the image captioning task. However, most of the existing models depend heavily on paired image-sentence datasets...Instead of relying on manually labeled image-sentence pairs, our proposed model merely requires an image set, a sentence corpus, and an existing visual concept detector."

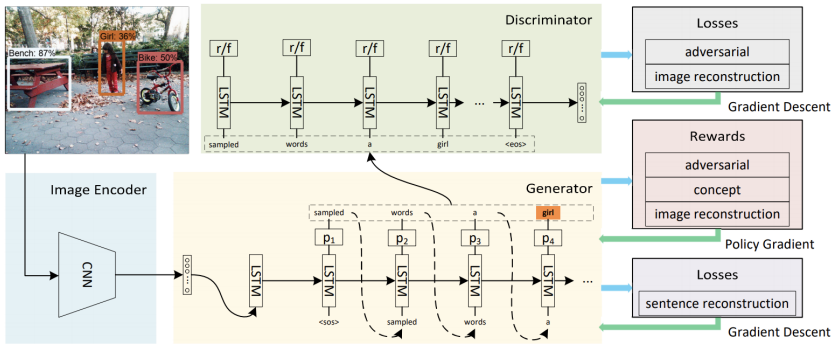

Figure 6: Unsupervised Captioning, taken from Feng et. al (6)

7 Contributions

Pravin Ravishanker wrote up over 90 percent of the Final Project Report. Pravin wrote the Abstract, Introduction, Approach, Results, and Discussion sections of the Final Report. Moreover, Pravin wrote the PyTorch model training code for training baseline Wide ResNet-50 models. I, Pravin Ravishanker, did around 33 PERCENT of the Final Project. Pravin is taking CS182 for a letter grade.

Harshayu Girase wrote the PyTorch code for generating Adversarially Attacked FGSM and BIM images, as well as the model training code for training ResNet models on Augmented and Adversarially Attacked Validation Datasets. Harshayu did around 33 PERCENT of the Final Project. Harshayu is taking CS182 for a letter grade.

Rahul Gupta wrote the PyTorch code for runtime detection of adversarial examples. Rahul also wrote the BIM algorithm explanation in the Approach section, and the Future Work section. Rahul did around 33 PERCENT of the Final Project. Rahul is taking CS182 for Pass / No Pass.

Our fourth member of our team, David Wang, dropped the class, and did not contribute to the project at all. David dropped CS 182.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

[2] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *arXiv preprint arXiv:1705.07263*, 2017.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[5] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

[6] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. *arXiv preprint arXiv:1811.10787*, 2018.

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

[8] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[9] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7942–7951, 2019.

[10] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. *arXiv preprint arXiv:1812.03411*, 2018.

[11] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[12] Leon Yao and John Miller. Tiny imagenet classification with convolutional neural networks. *CS 231N*, 2(5):8, 2015.

[13] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.