ABSTRACT

Adversarial attacks with improved transferability – the ability of an adversarial example crafted on a known model to also fool unknown models – have recently received much attention due to their practicality. Nevertheless, existing transferable attacks craft perturbations in a deterministic manner and often fail to fully explore the loss surface, thus falling into a poor local optimum and suffering from low transferability. To solve this problem, we propose Attentive-Diversity Attack (ADA), which disrupts diverse salient features in a stochastic manner to improve transferability. Primarily, we perturb the image attention to disrupt universal features shared by different models. Then, to effectively avoid poor local optima, we disrupt these features in a stochastic manner and explore the search space of transferable perturbations more exhaustively. More specifically, we use a generator to produce adversarial perturbations that each disturbs features in different ways depending on an input latent code. Extensive experimental evaluations demonstrate the effectiveness of our method, outperforming the transferability of state-of-the-art methods. Codes are available at https://github.com/wkim97/ADA.

Index Terms— Adversarial examples, Black-box, Transferability, Attention, Diversity

1. INTRODUCTION

While deep neural networks (DNNs) have achieved impressive performance on numerous vision tasks [1, 2, 3], recent studies [4, 5] have revealed their vulnerability against adversarial examples, which are crafted by adding a maliciously designed perturbation to the image. Such adversarial attack is categorized as either white-box or black-box depending on the knowledge of the model owned by the attacker, and recent works have focused on more challenging black-box attacks. Query-based attacks [6, 7, 8] use the query outputs to estimate the gradients of an unknown model, but the excessive number of required queries limits their practicality. Instead, transfer-based attacks that rely on transferability, which is the ability of an adversarial example crafted on a white-box surrogate model to fool black-box target models, have received more attention.

However, traditional attacks (e.g., BM [9], etc.) easily overfit to the surrogate model and exhibit poor transferability. To solve this issue, some have proposed more advanced optimization algorithms. Dong et al. [10] applied momentum to avoid poor local optimum, Xie et al. [11] applied random transformations to the image, Dong et al. [12] proposed a translation-invariant attack, and Wang et al. [13] applied variance tuning for more stable momentum. Based on findings that different models learn similar features, Zhou et al. [14] maximized the distance between the features of the original image and the adversarial image. However, classifiers tend to also learn model-specific features [15], and more recent works perturbed salient features; Wu et al. [16] disrupted the attention heatmaps, and Wang et al. [17] proposed aggregated gradients to perturb object-aware features.

Nevertheless, these attacks rely on a gradient-based method [9] that generates perturbations in a deterministic manner. At each iteration, they update a perturbation only in a single, specific direction that maximizes the given loss function, and with lack of stochasticity in this process, they often fail to fully explore the entire loss surface. Thus, as shown in Fig. 1, with low diversity, adversarial examples crafted by these attacks easily fall into a poor local optimum and overfit to the surrogate model, suffering from low transferability.

To solve this problem, we propose Attentive-Diversity Attack (ADA), which improves the transferability of adversarial examples by disrupting salient features in a diverse manner. Primarily, based on recent findings [16, 17] that disrupting salient features boosts transferability, we perturb the image attention, which highlights features that are responsible for model decision and are likely to be shared across different models. Greedily corrupting the attention using a gradient-based method, however, may lead to deterministic adversarial examples that easily fall into poor local optima. To avoid such local optimum, for the first time, we propose to disrupt these features in a diverse and stochastic manner. More specifically, we guide a generator to craft diverse perturbations that disrupt the attention differently depending on an input latent code. As shown in Fig. 1, in that way, we can explore the search space of transfer-
able adversarial examples more exhaustively, and the generator can learn to craft diverse perturbations that are located outside the poor local optimum. These adversarial examples effectively fool the target models, while those crafted by existing deterministic methods become overfitted to the surrogate model.

In summary, our contributions are as follow:

- For the first time, we introduce stochasticity to adversarial examples in the feature level to improve transferability.
- We propose Attentive-Diversity Attack (ADA), an effective generator-based adversarial attack framework that perturbs image attention in a diverse, non-deterministic manner.
- Extensive experiments exhibit the superior transferability of our method as compared to existing state-of-the-art methods.

2. METHODS

Preliminaries. Let \( f_v \) be a target classifier. The objective of an untargeted adversarial attack is to create an adversary \( x^{adv} \) of an image \( x \) in class \( t \) such that it leads to a misclassification on the target classifier (i.e., \( f_v(x^{adv}) \neq t \)). In this paper, we consider a black-box attack where we do not have access to the target classifier. Instead, we employ an accessible surrogate model \( h_0 \) that shares the same output space with \( f_v \) but has different architectures and/or parameters. We then generate a transferable adversarial example on the surrogate model as follows:

\[
\arg \max_{x^{adv}} L_\theta(x^{adv}, t), \quad \text{s.t.} \quad \|x - x^{adv}\|_\infty \leq \epsilon, \quad (1)
\]

where \( L_\theta(\cdot, \cdot) \) is the classification loss on the surrogate model \( h_0 \), and \( \epsilon \) is a constraint set on the magnitude of perturbation.

The success of such transfer-based attack highly depends on its transferability. Nevertheless, existing attempts rely on a gradient-based method to optimize Eq. 1 and thus greedily find a deterministic solution that maximizes \( L_\theta(x^{adv}, t) \). Such deterministic property can overfit the adversarial examples to the surrogate model and thus lower the chance of black-box attack being successful.

To address this challenge, we propose Attentive-Diversity Attack (ADA) (Fig. 2), which significantly improves the transferability of adversarial examples by using an attack generator guided by attention perturbation and feature diversification.

Attack Generator. Instead of a gradient-based method that crafts perturbations in an iterative and deterministic manner, we use a generator to parameterize the adversary with a DNN. Given an image \( x \) and a latent code \( z \) sampled from a Gaussian distribution, the generator \( g \) learns to output an adversarial perturbation that is dependent on the latent code. We then form an adversarial image \( x^{adv}_z \) as follows:

\[
x^{adv}_z = C\text{lip}_{\alpha}(x + \epsilon \cdot g(x, z)), \quad (2)
\]

where \( C\text{lip}_{\alpha} \) clips the perturbation in a per-pixel manner so that it is bounded to \( \epsilon \)-ball of \( L_\infty \) norm.

Attention Perturbation. In order to boost transferability, we disrupt the image attention, which highlights features that are responsible for model decision and are likely to be relevant to the main objects of the image. As different classifiers universally rely on these object-related features to make decisions, perturbations on these features will effectively transfer to other models.

Based on Grad-CAM [18], we define attention \( A \) as a weighted representation of features \( F \), which we set as the output from the last convolutional layer (e.g., \( \text{Mixed}_3 \) for Inc-v3), as follows:

\[
A(x; t) = \alpha t F = GAP \left( \frac{\partial g(y)}{\partial F} \right) F, \quad (3)
\]

The weight \( \alpha t \) denotes the importance of the feature \( F \) given the ground truth class \( t \). It is obtained by taking the gradient of \( y_t \) – the prediction for class \( t \) – with respect to \( F \) and applying global average pooling (\( \text{GAP}(\cdot) \)) over the spatial dimension. To prevent the generator from perturbing only the few channels with the highest magnitudes, we further apply channel-wise normalization on \( A(x; t) \).

Then, the generator learns to maximize the distance between the attention representations of the original image and the adversarial image by maximizing the following attention loss \( L_{\text{attn}} \):

\[
L_{\text{attn}} = \|A(x^{adv}_z; t) - A(x; t)\|_2. \quad (4)
\]

While Wu et al. [16] have similarly disrupted the attention heatmaps extracted using the techniques of Grad-CAM [18], our method differs from their approach on that we additionally apply channel normalization. Without channel normalization, the generator perturbs only the few feature channels with highest magnitudes, which limits the diversity of perturbations it can generate. To prevent this, we normalize each feature channel and enable the generator to disrupt more diverse features.

Feature Diversification. Without any guidance, the generator may learn to ignore the input latent code and greedily maximize the attention loss in a deterministic manner just like gradient-based methods. Thus, we guide the generator to explore and corrupt diverse features in a stochastic manner. We train it to disturb the attention representations differently for two distinct input latent codes \( z_1 \) and \( z_2 \), each sampled from a Gaussian distribution, by applying a diversity regularization [19] and maximizing the following diversity loss \( L_{\text{div}} \):

\[
L_{\text{div}} = \|A(x^{adv}_{z_1}; t) - A(x^{adv}_{z_2}; t)\|_2 \quad (5)
\]

We craft two adversarial examples \( x^{adv}_{z_1} \) and \( x^{adv}_{z_2} \) each by passing \( z_1 \) and \( z_2 \), respectively, into the generator (Eq. 2) and obtain their respective attention representations \( A(x^{adv}_{z_1}; t) \) and \( A(x^{adv}_{z_2}; t) \) (Eq. 3). Then, by maximizing the distance between the two representations, we force the generator to craft semantically diverse perturbations.

While Yang et al. [19] originally proposed the diversity regularization, their applications have been limited to pixel or feature levels. Diversity on these levels, however, may not necessarily translate to diversity on the attention space and may fail to guide our generator to disrupt the salient features in a diverse manner. To explicitly guide it to perturb the meaningful features in a diverse manner, unlike existing approaches, we apply diversity on the attention level.

Overall, we learn the attack generator \( g \) to maximize:

\[
L = L_{\text{cls}} + \lambda_{\text{attn}} \cdot L_{\text{attn}} + \lambda_{\text{div}} \cdot L_{\text{div}}. \quad (6)
\]
where $L_{cls}$ is the cross-entropy loss between the adversarial image and the ground truth label, and $\lambda_{att}$ and $\lambda_{div}$ control the weights of the attention loss ($L_{att}$) and the diversity loss ($L_{div}$), respectively.

There have been several attempts to craft diverse adversarial examples. Jang et al. [20] and Dong et al. [21] modeled diverse perturbations from a single image, but their approaches are limited to pixel-level diversity and improving adversarial robustness. Xie et al. [11] boosted transferability by crafting perturbations on randomly transformed images, but their approach can only implicitly perturb features in a diverse manner as a result of pixel-level transformations. In contrast, for the first time, we craft semantically diverse perturbations by explicitly disrupting diverse features. As a result, we effectively avoid poor local optimum, improving transferability as also shown by the experiment results (Table 1).

### 3. EXPERIMENTS

#### Experiment Setup

For the attack generator, we employ a U-Net [22, 23] based convolutional encoder-decoder consisting of three encoding blocks and three decoding blocks. Each encoding and decoding block consists of a convolutional layer and a transposed convolutional layer, respectively, followed by a batch normalization layer and a ReLU layer. At each encoding block, the latent code $z$ is spatially expanded and concatenated to the input of the block. The generator is trained for 100 epochs with learning rate of 1e-4, batch size of 8, and an Adam optimizer [24] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and weight decay 1e-5.

We use 10,000 images randomly selected from the ImageNet validation set [25] for train data and 1,000 images from the NeurIPS 2017 adversarial competition [26] for test data. We test our method on Inception-v3 (Inc-v3) [2], Inception-v4 (Inc-v4) [27], Inception-ResNet-V2 (IncRes-v2) [27], ResNet-V2 (Res-v2) [28], and VGG16 (Vgg-16) [29]. We compare our method with various state-of-the-art attacks – MI-FGSM [10], DIM [11], VMI-FGSM [13], TAP [14], and FIA [17] – for which we set the number of iterations $T = 10$, the step size $\alpha = 1.6$, and the rest of the hyperparameters as specified in their respective references. The maximum perturbation constraint $\epsilon$ is set to 16 under $L_{\infty}$ norm. For our method, we use $\lambda_{att} = 10$, $\lambda_{div} = 1000$, and 16 for the length of latent code $z$.

#### Comparison of Transferability

To compare the transferability of our method and the baselines, we craft adversarial examples on four surrogate models – Inc-v3, Inc-v4, IncRes-v2, and Res-v2 – and measure their attack success rates (ASR), or the misclassification rate of a model [13], on five target models (Table 1). The ensemble denotes ASR on the ensemble of all target models excluding the white-box model (i.e., target model is different from the surrogate model), whose prediction is the average probability output of all models. The results indicate that our method fools black-box target models with higher ASR than existing methods in most cases, outperforming FIA by average of 6.4%p on the ensemble model. Also, while our method generally shows lower ASR on the white-box target models, it shows much higher ASR on black-box target models, showing that it overfits less to the surrogate model and generalizes well to unknown models.

An interesting observation is that our method outperforms existing methods by a larger margin when the architectures of the target models are more different from that of the surrogate model. For example, while our attack crafted on Inc-v3 outperforms FIA by 4.7%p on a more similarly structured Inc-v4 [30], it outperforms FIA by a larger margin of 9.6%p on a more differently structured Vgg-16 [30]. While existing methods overfit to the surrogate model and show low transferability on models with more distinct structures, our method shows high transferability regardless of the model structure.

#### Generation of Diverse Perturbations

Recent findings [20, 21] have shown that the robustness of adversarial training model improves when trained against diverse adversarial examples. Based on this idea, we test the robustness of a classifier trained against adversarial examples crafted by ADA to show that our method indeed generates diverse perturbations. We train Inc-v3 for 30 epochs using batch size of 8, an SGD optimizer with learning rate of 0.001, momentum of 0.9, and weight decay of 5e-4 on Caltech101 dataset [31] with 8,681 images and 101 classes randomly split into 7,332/1,349 for training/test set.

### Table 1: Attack success rates of different attacks against various target models.

| Attack     | Inception V3 | Inception V4 | Inception-ResNet V2 | ResNet V2 | VGG16 | Ensemble | Rank |
|------------|--------------|--------------|---------------------|-----------|-------|----------|------|
| MI-FGSM    | (97.9%)      | 42.9%        | 39.9%               | 41.2%     | 53.1% | 83.1%    | 35.7%| 6     |
| DIM        | (98.0%)      | 68.3%        | 61.9%               | 53.1%     | 68.6% | 68.2%    | 58.2%| 5     |
| VMI-FGSM   | (97.9%)      | 69.6%        | 66.7%               | 57.6%     | 70.0% | 61.8%    | 4    | 4     |
| TAP        | (100.0%)     | 77.9%        | 73.5%               | 53.1%     | 70.6% | 69.1%    | 3    | 3     |
| FIA        | (98.5%)      | 84.2%        | 80.1%               | 69.3%     | 85.6% | 77.6%    | 2    | 2     |
| Ours       | (96.1%)      | 88.9%        | 82.9%               | 82.4%     | 95.2% | 85.3%    | 1    | 1     |
| MI-FGSM    | 59.4%        | (98.9%)      | 44.5%               | 47.8%     | 63.6% | 44.9%    | 6    | 6     |
| DIM        | 75.5%        | (97.9%)      | 66.5%               | 60.5%     | 74.9% | 66.5%    | 5    | 5     |
| VMI-FGSM   | 76.6%        | (98.5%)      | 70.0%               | 65.1%     | 76.8% | 68.9%    | 4    | 4     |
| TAP        | 75.6%        | (100.0%)     | 70.2%               | 59.7%     | 82.6% | 72.2%    | 3    | 3     |
| FIA        | 83.3%        | (99.0%)      | 78.5%               | 74.6%     | 85.2% | 78.3%    | 2    | 2     |
| Ours       | 85.2%        | (97.7%)      | 67.6%               | 79.0%     | 80.5% | 79.5%    | 1    | 1     |
| MI-FGSM    | 58.0%        | 52.6%        | (99.4%)             | 46.9%     | 63.5% | 47.0%    | 7    | 7     |
| DIM        | 73.0%        | 70.6%        | (94.8%)             | 57.7%     | 70.4% | 65.6%    | 4    | 4     |
| VMI-FGSM   | 78.4%        | 77.4%        | (99.3%)             | 64.5%     | 74.2% | 71.6%    | 3    | 3     |
| TAP        | 74.1%        | 66.8%        | (95.2%)             | 64.7%     | 63.0% | 61.8%    | 5    | 5     |
| FIA        | 81.6%        | 77.1%        | (88.6%)             | 66.9%     | 81.9% | 74.8%    | 2    | 2     |
| Ours       | 85.2%        | 89.4%        | (93.0%)             | 80.7%     | 80.4% | 85.9%    | 1    | 1     |
| MI-FGSM    | 54.3%        | 47.9%        | 44.1%               | (99.6%)   | 61.4% | 44.4%    | 6    | 6     |
| DIM        | 75.3%        | 70.6%        | 68.8%               | (99.1%)   | 73.9% | 70.0%    | 3    | 3     |
| VMI-FGSM   | 72.7%        | 67.4%        | 64.7%               | (97.6%)   | 72.3% | 65.3%    | 4    | 4     |
| TAP        | 51.8%        | 44.3%        | 44.5%               | (92.4%)   | 68.2% | 52.6%    | 5    | 5     |
| FIA        | 81.2%        | 76.7%        | 74.5%               | (99.9%)   | 82.9% | 74.1%    | 2    | 2     |
| Ours       | 79.7%        | 90.9%        | 71.9%               | (94.0%)   | 93.1% | 80.6%    | 1    | 1     |

*These models are publicly available at: https://github.com/Cadene/pretrained-models.pytorch
Fig. 3: Analysis of our method with all adversarial examples crafted on Inc-v3 as the surrogate model. (a) PCA visualization on surrogate model (Inc-v3) and target model (Inc-v2) for features of adversarial examples crafted by FIA (blue) and our method (red), (b) comparison of attack success rates (ASR) on the ensemble of black-box target models with varying perturbation constraint $\epsilon$, (c) ASR on varying weights for attention loss $\lambda_{\text{attn}}$ when $\lambda_{\text{div}} = 1000$, and (d) ASR on varying weights for diversity loss $\lambda_{\text{div}}$ when $\lambda_{\text{attn}} = 10$.

Table 2: Classification accuracy of adversarial training models under various attacks. Leftmost column and uppermost row represent the attacks used for training and the attacks used for evaluation, respectively. Adversarial examples used for evaluation are crafted on a classifier trained with original images (marked †). Best results are highlighted in bold.

Table 4: Original/adversarial images (top row) crafted by our attack and their attention (bottom row). Our attack crafts adversarial examples that disrupt the attention and final predictions in a diverse manner.

4. CONCLUSION

In this paper, we have proposed Attentive-Diversity Attack (ADA) that generates highly transferable adversarial examples. ADA relies on a generator to generate perturbations that disrupt image-salient features in a non-deterministic manner. Consequently, the crafted adversarial examples avoid falling into poor local optima and become less overfitted to the surrogate model. Exhaustive experiments validate the superior performance of ADA against state-of-the-art methods and the effectiveness of its individual components.
5. REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in CVPR, 2016.

[2] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, “Rethinking the inception architecture for computer vision,” in CVPR, 2016.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell, “Fully convolutional networks for semantic segmentation,” in CVPR, 2015.

[4] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, “Explaining and harnessing adversarial examples,” in ICLR, 2015.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, “Intriguing properties of neural networks,” in ICLR, 2014.

[6] Wieland Brendel, Jonas Rauber, and Matthias Bethge, “Decision-based adversarial attacks: Reliable attacks against black-box machine learning models,” in ICLR, 2018.

[7] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin, “Black-box adversarial attacks with limited queries and information,” in ICML, 2018.

[8] Jonathan Usato, Brendan O’donoghue, Pushmeet Kohli, and Aaron Oord, “Adversarial risk and the dangers of evaluating against weak attacks,” in ICLR, 2018.

[9] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, “Adversarial examples in the physical world;” in ICLR Workshop, 2017.

[10] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li, “Boosting adversarial attacks with momentum,” in CVPR, 2018.

[11] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille, “Improving transferability of adversarial examples with input diversity,” in CVPR, 2019.

[12] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu, “Evaiving defenses to transferable adversarial examples by translation-invariant attacks,” in CVPR, 2019.

[13] Xiaosen Wang and Kun He, “Enhancing the transferability of adversarial attacks through variance tuning,” in CVPR, 2021.

[14] Wen Zhou, Xin Hou, Yongjun Chen, Mengyin Tang, Xiangqi Huang, Xiang Gan, and Yong Yang, “Transferable adversarial perturbations,” in ECCV, 2018.

[15] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, “Adversarial examples are not bugs, they are features,” in NeurIPS, 2019.

[16] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai, “Boosting the transferability of adversarial samples via attention,” in CVPR, 2020.

[17] Zhbo Wang, Hengchang Guo, Zhiwei Zhang, Wenzin Liu, Zhan Qin, and Kui Ren, “Feature importance-aware transferable adversarial attacks,” in ICCV, 2021.

[18] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in ICCV, 2017.

[19] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee, “Diversity-sensitive conditional generative adversarial networks,” in ICLR, 2019.

[20] Yunseok Jang, Tianchen Zhao, Seunghoon Hong, and Honglak Lee, “Adversarial defense via learning to generate diverse attacks,” in ICCV, 2019.

[21] Yinpeng Dong, Zhijie Deng, Tianyu Pang, Hang Su, and Jun Zhu, “Adversarial distributional training for robust deep learning,” in NeurIPS, 2020.

[22] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie, “Generative adversarial perturbations,” in CVPR, 2018.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention, 2015.

[24] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in ICLR, 2015.

[25] 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,” in International Journal of Computer Vision, 2015.

[26] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al., “Adversarial attacks and defences competition,” in The NIPS’17 Competition: Building Intelligent Systems, 2018.

[27] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in AAAI, 2017.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Identity mappings in deep residual networks,” in ECCV, 2016.

[29] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” in ICLR, 2015.

[30] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao, “Is robustness the cost of accuracy?-a comprehensive study on the robustness of 18 deep image classification models,” in ECCV, 2018.

[31] Li Fei-Fei, Rob Fergus, and Pietro Perona, “One-shot learning of object categories,” in TPAMI, 2006.

[32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, “Towards deep learning models resistant to adversarial attacks,” in ICLR, 2018.