Interpretability Analysis of Deep Neural Networks Under Adversarial Attacks

Guoqing Xie, *Xi'an Jiaotong University*

Abstract—In recent years, researches on interpretablity of deep neural networks have received more and more attention. In this paper, we introduce a general framework called network dissection, which define interpretablity by the ratio of neurons that align with semantic concepts. Using this method, we quantify interpretablity of three classical convolutional neural networks under adversarial attacks. Based on the analyses, we find that adversarial attacks always decrease the interpretability of networks. Thus, we think there exists inconsistency between the features neurons have learned and semantic concepts. To solve this problem, we also propose an idea about adversarial training with domain adaptation loss.

Index Terms—Deep Neural Networks, Interpretability, Network Dissection, Adversarial Attacks

1 Introduction

With the rapid development of software and hardware, deep learning is playing a more and more important role in our lives. Deep neural networks [1] have been widely used in various fields such as computer vision, language recognition and machine translation. However, due to the lack of understanding and analysis of their internal working mechanism, deep neural networks are usually regarded as opaque [2]. This means that people can only observe the predicted results of the network, but can hardly understand the reasons for the decision-making of the model.

Actually, the stacking of implicit layers and the existence of activation function enable deep neural networks to process data in a non-linear and complex manner. In the training process, deep neural networks can learn the intrinsic rules and implicit representation of sample data, which greatly assist in the interpretation of these trained networks.

1.1 Related Work

1.1.1 Interpretability Of Deep Neural Networks

The interpretability of models has become a key factor in determining whether users can trust these models, especially when we need machines to make predictions and decisions for important tasks related to human life, property security and so on. Because of the diversity of this concept, people can often interpret a network model from different perspectives.

For instance, from the perspective of semantic concept alignment (Bau et al.), deep neural networks can be dissected by evaluating the alignment between individual hidden units (or neurons) and a set of semantic concepts (see Fig. 1). This framework is called Network Dissection [3], which is the classical methods to quantify interpretability of deep visual representations. Furthermore, there are also many ways to improve the interpretability of models during training. For example, Hu et al. [4] propose the architecture

E-mail: abc123abc@stu.xjtu.edu.cn (G. Xie)

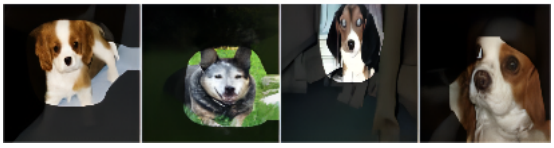


Fig. 1. Dissection of AlexNet in the last conv layer. Here shows part of the results of unit 110, which aligns the concept "dog".

disentanglement methods, which based on the information flow of the reasoning process in the network. And Varshneya et al. [5] propose a novel training methodology: Concept Group Learning (CGL). CGL encourages training of interpretable CNN filters by partitioning filters in each layer into concept groups, each of which is trained to learn a single visual concept.

Although many existing methods have shown that the features learned by units (inside deep neural networks) can be associated with semantic concepts (understood by humans), these methods only use real data for analysis. Dong et al. [2] find some phenomena which are inconsistent with the previous conclusions, by analyzing the feature representation of deep neural network learning with adversarial examples.

1.1.2 Adversarial Examples and Adversarial Attacks

Sometimes, if each pixel of the image is fed into a neural network with a small perturbation, the classifier might misdiagnose it. These perturbations are usually invisible to the human eye, but they can make the network get a false prediction even with high confidence (see Fig.2). This process is called adversarial attack, and the image with this small perturbation is adversarial example.

Based on different manifestations of model prediction errors, adversarial examples can be divided into two categories: the first type is called adversarial examples without targets, which can be incorrectly predicted by the model as any category other than the real category; the second type is called adversarial examples with targets, which will be misclassified by the model into target categories designated by

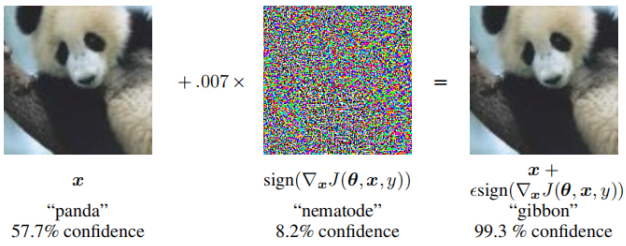


Fig. 2. A demonstration of fast adversarial example generation applied to GoogLeNet on ImageNet. [6]

Fig. 3. The examples after segmentation in Broden dataset.

the attacker, to achieve the specific purpose of the attacker.

Given a real example x, its real category is y. Meanwhile, given a classifier $f_\theta(\cdot)$ based on deep neural network, which θ represents the parameters of the classifier. An adversarial sample x^* will usually be searched in the neighborhood of the real example x, making it look no different from the real example, but will be misclassified by the model. Based on the optimization method, the adversarial attacks need to solve:

$$\arg\min_{x^*} \left\{ L(f_\theta(x^*), y) \right\}, \qquad s.t. \|x - x^*\|_\infty \leq \varepsilon, \tag{1}$$

or to solve:

$$\arg\min_{x^*} \left\{ \|x - x^*\|_2^2 - \alpha L(f_\theta(x^*), y) \right\}. \tag{2}$$

Eq.(1) is used in Fast gradient sign method (FGSM) [6], while Eq.(2) is used in Basic iterative method (BIM) [7].

2 Methods

In this part, we firstly introduce the network dissection [3], a method to quantify interpretability of trained network with Broden dataset. Then we generate adversarial examples of Broden dataset, mainly using the Fast gradient sign method (FGSM). We wonder if the adversarial attacks will influence the interpretability of deep neural networks. Finally, to improve the interpretability of networks in adversarial examples, we propose a adversarial training method based on the work in [2], and change its feature consistency loss into domain adaptation loss.

2.1 Quantifying Interpretability

Before quantifying Interpretability of deep neural networks, we need a clear definition of the concept of interpretability. A classic definition is given by Bau et al., which mainly contains two aspects [3]. On the one hand, at the unit level, it defines the interpretability by whether the unit in networks is aligned with the visual semantic concept (if a unit can be aligned, it is called a detector); on the other hand, in the whole network or at a certain level, it measures the strength of the interpretability via the number of various semantic detectors.

Based on this definition, it is necessary to define the category of semantic concepts and build semantic concept dataset.The Broden Dataset was derived from Network Dissection [3]. It divides semantic concepts into six categories from low level to high level: color, texture, material, part, object and scene. In addition, it performs pixel-level segmentation of various semantic concepts for a large number of images based on the semantic concepts. The examples after segmentation are shown in Fig.3. The average number of images of each semantic concept is shown in Tab.1.As we can see, the Broden dataset suffers from an imbalance of categories, which should better be taken into account in data enhancements.

TABLE 1
Statistics of each label type included in the data set.

Category	Color	Texture	Material	Part	Object	Scene
Classes	11	47	32	234	584	468
Avg sample	59250	140	1703	854	491	38

In network dissection, the method to determine whether a unit is aligned with a semantic concept is shown in Fig.4, which is mainly divided into three steps:

- First, the image x in the Broden dataset is input into the network to be dissected, and the feature map $A_k(x)$ of a internal convolutional unit k is extracted.
- Then, if the size of the feature map is inconsistent with the input size, upsampling or downsampling is required. Bilinear interpolation method is used to acquire the activation map $S_k(x)$ from $A_k(x)$, which is scaled up to the mask resolution.
- Next, we set the top quantile level T_k for each unit k, such that $P(a_k > T_k) = 0.005$ over every spatial location of the activation map in the dataset. Thus, $S_k(x)$ can be thresholded into a binary segmentation $M_k(x)$ by Eq.3:

$$M_k(x) \equiv S_k(x) \geq T_k. \tag{3}$$

- Finally, we calculate the IOU score between $M_k(x)$ and its segmentation of semantic concepts of category c:

$$IoU_{k,c} = \frac{\sum |M_k(x) \cap L_c(x)|}{\sum |M_k(x) \cup L_c(x)|}. \tag{4}$$

This IOU score describes the magnitude of the coincidence of $M_k(x)$ and $L_c(x)$. When its IOU score is above a certain threshold ϵ, the unit k is considered aligned with the semantic concept.

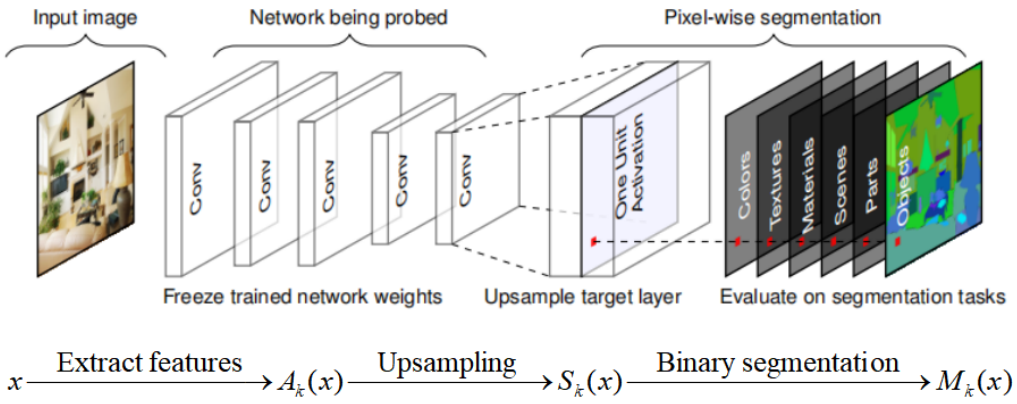


Fig. 4. The main framework of Network Dissection. [3]

2.2 Generating Adversarial Examples

Previous work suggested that deep neural networks could learn the feature representation of decoupling of image content [3], in which neurons detect semantic concepts understood by humans so that the overall network (or layer) can be interpreted. However, it is shown by Dong et al. [2] that neurons that can detect semantic concepts (such as objects or physical components) can be easily fooled by adversarial examples, showing inconsistencies between feature that neurons learned and semantic concepts. Here, we wonder if the adversarial attacks will influence the interpretability of deep neural networks.

There are many methods to generate adversarial examples. Here we choose the fast gradient sign method (FGSM), which produces adversarial examples through a one-step gradient iteration. Eq.5 is the white box attack without target by FGSM:

$$x^* = x + \varepsilon \cdot sign(\nabla_x L(f_\theta(x), y)), \tag{5}$$

where ε is the noise scale of the perturbation, and $sign(\cdot)$ represents the sign function. It can be seen as a solution of Eq.1. If we want achieve the adversarial attack with target, we can rewrite Eq.5 as:

$$x^* = x - \varepsilon \cdot sign(\nabla_x L(f_\theta(x), y^*)), \tag{6}$$

where y^* is the attack target, namely the label we want the model misdiagnose.

2.3 Adversarial Training

Actually, Dong et al. [2] have done a similar work, showing that the neurons do not have the ability to detect higher-level semantic information (objects or parts) in the images, but produce stronger responses to the specific category predicted by the model . It means whether the prediction is correct or not, the relationship between neurons and labels is unchanging.

To solve this problems, Dong et al. [2] propose a adversarial training method with a consistent loss of feature representation. And their experiments show this loss can truly improve the consistency between the feature representation learned by deep neural networks and the semantic concepts understood by humans during the training process. The consistent loss of feature representation is defined as Eq.7:

$$L_{con}(\theta, x) = d(\phi_\theta(x), \phi_\theta(x^{FGSM})), \tag{7}$$

where $\phi_\theta(x)$ returns the feature representation vector of the network for input x, and d is squared Euclidean distance. As we can see, this loss is actually to enhance the interpretability strength (mainly about high-level semantic concepts) of the network for the adversarial examples.

However, this consistent loss of feature representation is 'individual to individual', which means it can only reflect the alignment of features between individual samples. We want change this 'individual to individual' loss into 'whole to whole' loss by domain adaptation.

As a matter of fact, we can combine standard adversarial training with domain adaptation [8], to minimize the domain gap between clean domain D and adversarial domain A. We introduce an adaptation layer into the network, and add the domain adaptation loss into the last feature layer of the deep neural network. The domain adaptation loss contains the loss of covariance distance:

$$L_{CORAL}(D, A) = \frac{1}{k^2} \left\| C_{\phi(D)} - C_{\phi(A)} \right\|_1, \tag{8}$$

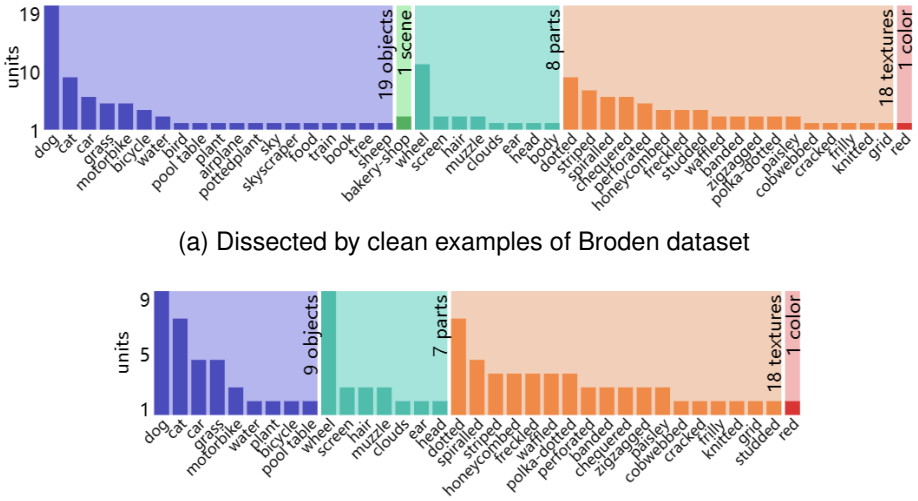
and the loss of maximum mean deviation:

$$L_{MMD}(D, A) = \frac{1}{k} \left\| \frac{1}{|D|} \sum_{x \in D} \phi(x) - \frac{1}{|A|} \sum_{x^* \in A} \phi(x^*) \right\|_1. \tag{9}$$

Eq.8 and Eq.9 both belong to unsupervised domain adaptation (UDA) loss.

3 Experiments

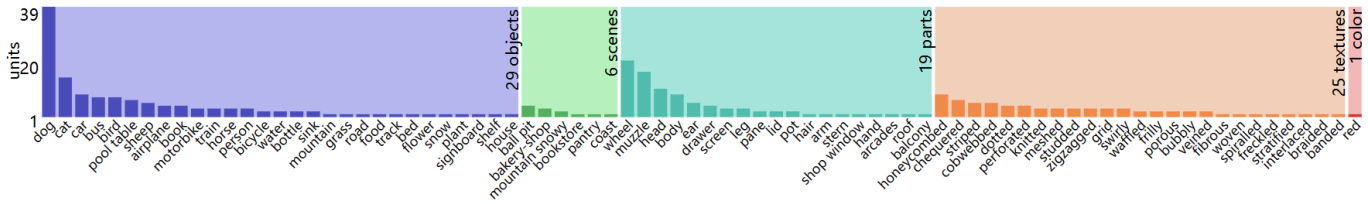
In this section, we first generate the adversarial examples for Broden dataset using FGSM. Then we quantify interpretability of deep neural networks in these adversarial examples by network dissection, and compare the results in clean examples. Finally, we try to improve the consistency of feature representation between adversarial examples and clean examples by the domain adaptation loss. However, the results of the last experiment are not shown here due to some problems in the code and the limited time. We will provide a more complete version of this paper as soon as possible.



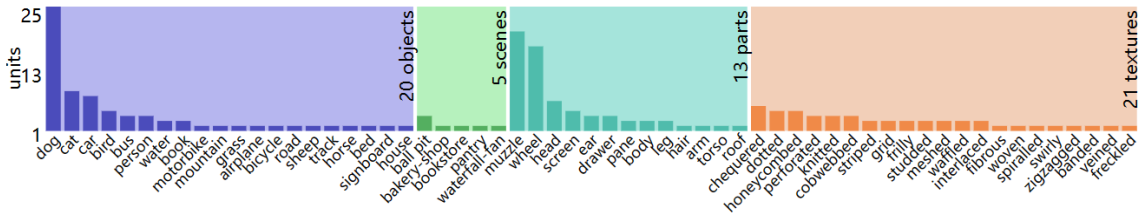
(a) Dissected by clean examples of Broden dataset

(b) Dissected by adversarial examples of Broden dataset

Fig. 5. Network dissection results of AlexNet in the feature layer

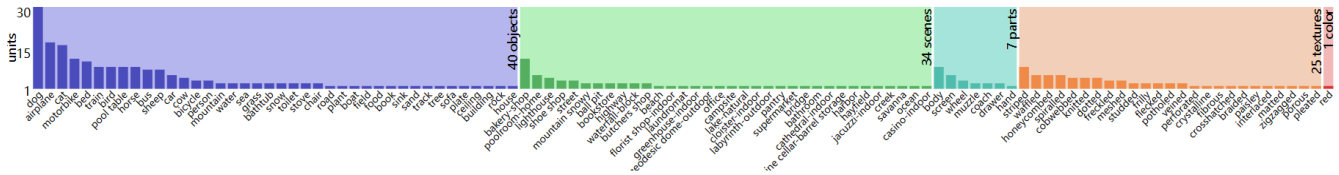


(a) Dissected by clean examples of Broden dataset

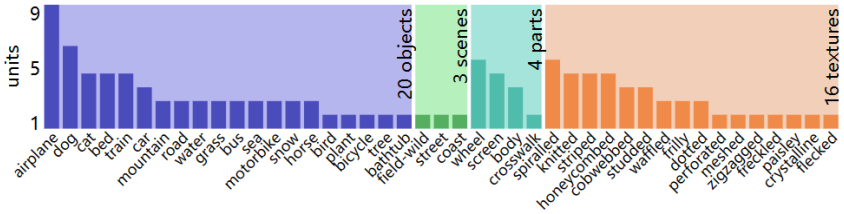


(b) Dissected by adversarial examples of Broden dataset

Fig. 6. Network dissection results of VGG-11 in the feature layer



(a) Dissected by clean examples of Broden dataset

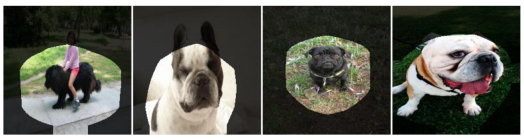


(b) Dissected by adversarial examples of Broden dataset

Fig. 7. Network dissection results of ResNet-18 in layer 4

3.1 Implementation Details

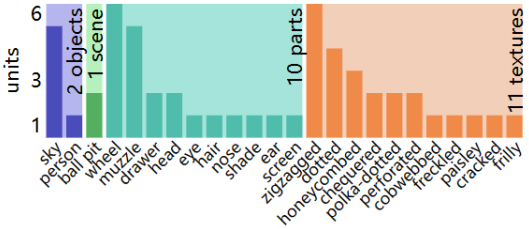
Networks In this paper, we choose three classical convolutional neural networks: AlexNet [9], VGG [10] and ResNet [11], to quantify interpretability of their feature layers. All the networks have be pretrained in the ImageNet dataset.



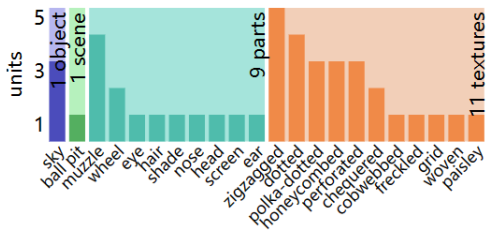
(a) Unit 137 (object: dog) with IoU 0.11 in clean examples of Broden dataset.

(b) Unit 137 (object: dog) with IoU 0.05 in adversarial examples of Broden dataset.

Fig. 8. Visualization of dissection results of ResNet-18 in layer 4



(a) Dissected by clean examples of Broden dataset



(b) Dissected by adversarial examples of Broden dataset

Fig. 9. Network dissection results of ResNet-18 in layer 3

TABLE 2
The ratio (%) of neurons that align with semantic concepts for each model (with clean examples or adversarial examples of Broden dataset).

Model-Layer	Clean examples						Adversarial examples					
	C	T	S	P	O	All	C	T	S	P	O	All
AlexNet-Feature layer	0.4	20.3	0.8	7.8	22.3	**51.6**	0.4	16.4	0.0	7.0	11.7	35.5
VGG11-Feature layer	0.2	13.3	2.3	16.2	24.6	**56.6**	0.0	8.6	1.4	12.3	12.9	35.2
ResNet18-Layer4	0.2	12.1	11.9	4.5	35.0	**63.7**	0.0	7.0	0.6	2.5	10.4	20.5
ResNet18-Layer3	0.0	9.4	0.8	8.2	2.3	**20.7**	0.0	9.8	0.4	5.1	1.2	16.4

Dataset We choose the Broden dataset [3] based on network dissection. And a variant of FGSM attack [12] is used to generate an adversarial example x^* from the clean example x by Eq.10:

$$x^* = x + \varepsilon \cdot sign(\nabla_x L(f_\theta(x), y_{pred})), \tag{10}$$

where y_{pred} denotes the predicted class of the model. Eq.10 can effectively avoid the label leaking effect in FGSM.

3.2 Results

We compare the network dissection results of AlexNet, VGG-11 and ResNet-18. The number of detectors with different semantic concepts is shown here, as a significant index to quantify interpretability of these networks. Fig.5, Fig.6, Fig.7 show the network dissection results of AlexNet (feature layer), VGG-11 (feature layer) and ResNet-18 (layer 4).

As we can see, the number of detectors of all these networks will decrease under adversarial examples, which means adversarial attacks truly influence the interpretability of deep neural networks. Tab.2 shows the quantitative results. Thus, when we try to align the neurons with semantic concepts, we must know if the relationship still exists under adversarial attacks. Otherwise, we can hardly determine whether neurons are associated with semantic concepts or specific categories.

Fig.8 also proves this idea. The unit 137 in layer 4 of ResNet-18 is aligned with the concept dog whether existing adversarial attacks or not. However, we can see that the IOU score decrease from 0.11 to 0.05, and some visualization regions are wrong. In fact, the adversarial target cause the network to misidentify these adversarial images as dog, which means this unit is associated with specific categories (dog) instead of semantic concepts.

Furthermore, comparing Fig.7 and Fig.9, it seems that the adversarial examples mainly influence the deeper layer, which is closer to the classifier in network. The ratio of neurons that align with semantic concepts decrease from 63.7 to 20.5 in layer 4 of ResNet-18, but only decrease from 20.7 to 16.4 in layer 3.

4 Conclusion

This paper quantifies interpretability of the classical convolutional neural networks based on network dissection method, and compares the results in clean examples with adversarial examples. In our experiments, we measure the ratio of neurons that align with semantic concepts for these networks. All the results show that the rate decreases in adversarial examples, which means adversarial attacks truly influence the interpretability of networks. Therefore, we

think there exists inconsistency between the features neurons have learned and semantic concepts. In addition, to solve this problem, we put forward a adversarial training method with domain adaptation loss.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] S. H. Dong Yin-Peng and Z. Jun, "Interpretability analysis of deep neural networks with adversarial examples," *Acta Automatica Sinica*, vol. 48, no. 1, pp. 75–86, 1 2022.

[3] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6541–6549.

[4] J. Hu, L. Cao, T. Tong, Q. Ye, S. Zhang, K. Li, F. Huang, L. Shao, and R. Ji, "Architecture disentanglement for deep neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 672–681.

[5] S. Varshneya, A. Ledent, R. A. Vandermeulen, Y. Lei, M. Enders, D. Borth, and M. Kloft, "Learning interpretable concept groups in cnns," *arXiv preprint arXiv:2109.10078*, 2021.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[7] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[8] C. Song, K. He, L. Wang, and J. E. Hopcroft, "Improving the generalization of adversarial training with domain adaptation," *arXiv preprint arXiv:1810.00740*, 2018.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.