Evaluating Adversarial Attacks on ImageNet: A Reality Check on Misclassification Classes

Utku Ozbulak[*]
Ghent University, Belgium
utku.ozbulak@ugent.be

Maura Pintor[*]
University of Cagliari, Italy
maura.pintor@unica.it

Arnout Van Messem
University of Liège, Belgium
Arnout.vanmessem@uliege.be

Wesley De Neve
Ghent University, Belgium
wesley.deneve@ugent.be

Abstract

Although ImageNet was initially proposed as a dataset for performance benchmarking in the domain of computer vision, it also enabled a variety of other research efforts. Adversarial machine learning is one such research effort, employing deceptive inputs to fool models in making wrong predictions. To evaluate attacks and defenses in the field of adversarial machine learning, ImageNet remains one of the most frequently used datasets. However, a topic that is yet to be investigated is the nature of the classes into which adversarial examples are misclassified. In this paper, we perform a detailed analysis of these misclassification classes, leveraging the ImageNet class hierarchy and measuring the relative positions of the aforementioned type of classes in the unperturbed origins of the adversarial examples. We find that 71% of the adversarial examples that achieve model-to-model adversarial transferability are misclassified into one of the top-5 classes predicted for the underlying source images. We also find that a large subset of untargeted misclassifications are, in fact, misclassifications into semantically similar classes. Based on these findings, we discuss the need to take into account the ImageNet class hierarchy when evaluating untargeted adversarial successes. Furthermore, we advocate for future research efforts to incorporate categorical information.

1 Introduction

Soon after its release, ImageNet [31] became the de facto standard dataset for performance benchmarking in the field of computer vision, primarily thanks to the diverse set of images and classes it contains. This diversity allowed for research on various vision tasks, including, but not limited to, classification [20, 36], segmentation [1, 23], and localization [14, 30]. Although the tasks put forward during the introduction of ImageNet were considered to be some of the hardest problems to address in the field of computer vision, a number of deep neural networks (DNNs) were, in recent years, able to achieve super-human results on many of these challenges, thus effectively "solving" the aforementioned problems [9]. However, research efforts that make use of ImageNet are not limited to the performance-oriented tasks mentioned before. Indeed, thanks to the diverse set of images it contains, ImageNet enabled a large number of research efforts beyond its initial scope, allowing researchers to experiment with model interpretability [34, 37], model calibration [12], object relations [32], fairness [42], and many other topics.

One research field that was enriched by the availability of ImageNet is the field of study that focuses on adversarial examples. In this context, the term "adversarial examples" refers to meticulously

[*] Equal contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Workshop on ImageNet: Past, Present, and Future.

created data points that come with a malicious intent, aimed at deceiving models that are performing a pre-defined task, steering the prediction outcome in favor of the adversary [2, 40]. Although adversarial examples are a threat for predictive models in domains other than the domain of computer vision [3, 28], the latter is acknowledged to be the one that suffers the most from adversarial examples, since an adversarial example created from a genuine image, through the use of adversarial perturbation, often looks the same as its unperturbed counterpart [10, 25]. This makes it, in most cases, impossible to detect adversarial examples by visually inspecting images.

Although the vulnerability of DNNs to adversarial examples in the image domain was originally mostly evaluated through the usage of two datasets, namely MNIST [22] and CIFAR [19], the authors of [4] revealed that methods derived through the usage of one of these datasets do not necessarily generalize to other datasets. In particular, compared to ImageNet, both of the aforementioned datasets contain images with a smaller resolution and a lower number of classes. As a result, most of the research efforts in recent years started to favor ImageNet over MNIST and CIFAR [7, 11, 39, 41].

From the perspective of adversarial evaluation, ImageNet does not only allow for most, if not all, of the research work that was performed using the previously mentioned datasets, it also enables a wide range of additional research topics in the area of adversariality, such as investigations with regards to regional perturbation [18], color channels [35, 41], and defenses that use certain properties of natural images [13]. However, as demonstrated in this paper, ImageNet has a major shortcoming when it comes to evaluating adversarial attacks, especially in model-to-model transferability scenarios: a large number of synsets/classes in ImageNet are semantically highly similar to one another.

Different from previous research efforts that mostly focus on generating more effective adversarial perturbations or evaluating adversarial defenses, we investigate a topic that is yet to be touched upon: untargeted misclassification classes for adversarial examples. Specifically, with the help of two of the most frequently used adversarial attacks and seven unique DNN architectures, including two recently proposed vision transformer architectures, we present a large-scale study that solely focuses on model-to-model adversarial transferability and misclassification classes in the context of ImageNet, resulting in the following contributions:

- In model-to-model transferability scenarios, we demonstrate that a large portion of adversarial examples are classified into the top-5 predictions obtained for their source image counterparts.

- With the help of the ImageNet class hierarchy, we show that adversarial examples created from certain synset collections are mostly misclassified into classes belonging to the same collections (e.g., a dog breed is misclassified as another dog breed).

- Interestingly, we can make the two aforementioned observations consistently for all of the evaluated models, as well as for both adversarial attacks. As a result, we discuss the necessity of evaluating misclassification classes when experimenting with adversarial attacks and untargeted misclassification in the context of ImageNet.

2 Adversarial attacks

Given an M-class classification problem, a data point $\boldsymbol{x} \in \mathbb{R}^k$ and its categorical association $\boldsymbol{y} \in \mathbb{R}^M$ associated with a correct class k ($y_k = 1$ and $y_m = 0 \,, \forall\, m \in \{0, \ldots, M\} \backslash \{k\}$) are used to train a machine learning model represented by θ. Let $g(\theta, \boldsymbol{x}) \in \mathbb{R}^M$ represent the prediction (logit) produced by the model θ and a data point $\boldsymbol{x}$. This data point is then assigned to the class that contains the largest output value $G(\theta, \boldsymbol{x}) = \arg\max(g(\theta, \boldsymbol{x}))$. When $G(\theta, \boldsymbol{x}) = \arg\max(\boldsymbol{y})$, this prediction is recognized as the correct one. For the given setting, a perturbation Δ bounded by an L_p ball centered at $\boldsymbol{x}$ with radius ϵ is said to be an *adversarial perturbation* if $G(\theta, \boldsymbol{x}) \neq G(\theta, \boldsymbol{x} + \Delta)$. In this case, $\hat{\boldsymbol{x}} = \boldsymbol{x} + \Delta$ is said to be an *adversarial example*.

Adversarial examples can be highly *transferable*: an adversarial sample that fools a certain classifier can also fool completely different classifiers that have been trained for the same task [6, 8, 29]. This property, which is called transferability of adversarial examples, is a popular metric for assessing the effectiveness of a particular attack. Let θ_1 and θ_2 represent two DNNs and let $\boldsymbol{x}$, k, and $\hat{\boldsymbol{x}}_1$ be a genuine image, the correct class of this image, and a corresponding adversarial example, respectively, with the adversarial example generated from this genuine image using an attack that targets a class c by leveraging the DNN represented by θ_1. If $G(\theta_1, \hat{\boldsymbol{x}}_1) = G(\theta_2, \hat{\boldsymbol{x}}_1) = c$ and $G(\theta_{\{1,2\}}, \boldsymbol{x}) = k$, then the adversarial example is said to have achieved *targeted adversarial transferability* to the model θ_2. If $G(\theta_1, \hat{\boldsymbol{x}}_1) = c$ but $G(\theta_2, \hat{\boldsymbol{x}}_1) \notin \{c, k\}$, the adversarial example in question is classified into a class

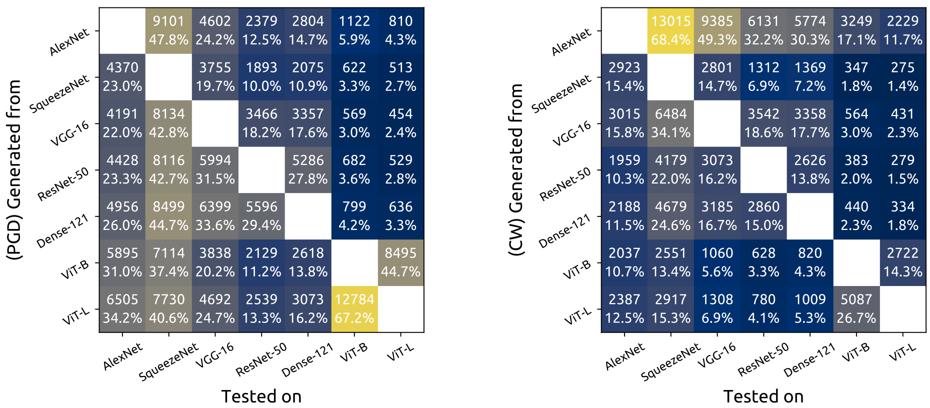

Figure 1: Number (percentage) of source images that became adversarial examples with PGD (*left*) and CW (*right*). Adversarial examples are generated by the models listed along the y-axis and tested by the models listed along the x-axis.

that is different than the targeted one (c) and the correct one (k). In cases like this, an adversarial example is said to have achieved *untargeted adversarial* transferability.

In the context of ImageNet, the success of targeted transferability for adversarial examples is known to be abysmally lower compared to the success of untargeted transferability [38]. As a result, many studies that propose a novel attack or perform a large-scale analysis of model-to-model transferability use untargeted transferability when showcasing the effectiveness of attacks, without evaluating the classes that adversarial examples are classified into [7, 11, 41]. Therefore, in this work, we investigate the success of untargeted adversarial transferability and the characteristics of misclassification classes.

3 Methodology

Models – In order to evaluate a variety of model-to-model adversarial transferability scenarios, we employ the following architectures: AlexNet [20], SqueezeNet [17], VGG-16 [36], ResNet-50 [15], and DenseNet-121 [16], as well as two recently proposed vision transformer architectures, namely ViT-Base$/16-224$ and ViT-Large$/16-224$ [9].

Data – For our adversarial attacks (see further in this section), we use images from the ImageNet validation set as inputs. Hereafter, these unperturbed input images will be referred to as *source images*. In order to perform a trustworthy analysis of adversarial transferability, we ensure that all source images are correctly classified by all employed models. To that end, we filter out all images incorrectly classified by at least one model, leaving us with $19,025$ source images to work with.

ImageNet hierarchy – Classes in ImageNet are organized according to the WordNet hierarchy [26, 31], grouping classes into various collections depending on their semantic meaning. We use the aforementioned hierarchy in order to measure intra-collection adversarial misclassifications. In that respect, an intra-collection misclassification is when an adversarial example created from a source image that belongs to a class under a collection is misclassified into a class under the same collection (e.g., an image belonging to a cat breed misclassified as another breed of cat is an intra-collection misclassification for the *Feline* collection). More details about the ImageNet hierarchy are given in the supplementary material (see Figure I).

Attacks – We use the adversarial examples generated for our previous study [27], where those adversarial examples are generated using two of the most commonly used attacks: Projected Gradient Descent (PGD) [24] and Carlini & Wagner's attack (CW) [5].

PGD can be seen as a generalization of L_∞ attacks [10, 21], aiming at finding an adversarial example $\hat{\boldsymbol{x}}$ that satisfies $||\hat{\boldsymbol{x}} - \boldsymbol{x}||_\infty < \epsilon$. The adversarial example is iteratively generated as follows:

$$\hat{\boldsymbol{x}}^{(n+1)} = \Pi_\epsilon \Big(\hat{\boldsymbol{x}}^{(n)} - \alpha \operatorname{sign}\big(\nabla_x J(g(\theta, \hat{\boldsymbol{x}}^{(n)})_c)\big) \Big) \,, \tag{1}$$

with $\hat{\boldsymbol{x}}^{(1)} = \boldsymbol{x}$, c the selected class, and $J(\cdot)$ the cross-entropy loss. We use PGD with 50 iterations and set ϵ to $38/255$. We adopt this constraint as the maximum perturbation-size bound in order to be able to produce a large number of adversarial examples that achieve model-to-model transferability.

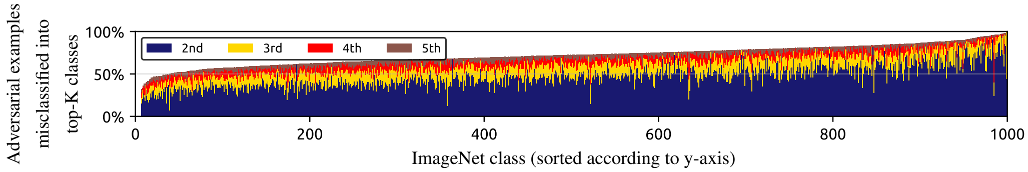

Figure 2: Number of adversarial examples, given per class, that are classified into the top$-\{2, 3, 4, 5\}$ classes predicted for their underlying source images.

CW, on the other hand, is a complex attack that incorporates L_2 norm minimization:

$$\text{miminize} \quad ||\boldsymbol{x} - (\boldsymbol{x} + \Delta)||_2^2 + \; f(\boldsymbol{x} + \Delta)\,. \tag{2}$$

In the paper introducing CW [5], multiple loss functions (i.e., f) are discussed. However, in later works, the creators of CW prefer to make use of the loss function that is constructed as follows:

$$f(\boldsymbol{x}) = \max\big(\max\{g(\theta, \boldsymbol{x})_i : i \neq c\} - g(\theta, \boldsymbol{x})_c, -\kappa\big)\,, \tag{3}$$

where this loss compares the predicted logit value of target class c with the predicted logit value of the next-most-likely class i. The constant κ can be used to adjust the *strength* of the produced adversarial examples (for our experiments, we use $\kappa = 20$ and the settings described in [5] and [27]).

We keep executing the attacks until a source image becomes an adversarial example or until the attacks reach a maximum number of iterations. At each iteration, we examine whether or not the images under consideration became adversarial examples for the aforementioned models.

4 Experiments

Leveraging the attacks described above and through the usage of $19,025$ source images that are correctly classified by the models employed, we create $289,244$ adversarial examples, where $173,549$ of those adversarial examples are generated with PGD and $115,695$ with CW. Detailed untargeted model-to-model transferability successes of those adversarial examples can be found in Figure 1.

To investigate misclassifications made into semantically similar classes, we first have a look at the adversarial examples that are misclassified into classes that lie in the top-5 positions of their source image predictions, where the four remaining classes, apart from the first one, are the classes that were deemed to be the most-likely prediction classes by the model under consideration, with the first one being the correct classification. Doing so, we provide Figure 2, with this figure displaying, for each class, the percentage of adversarial examples that had their predictions changed into one of the top-5 classes as described above. Specifically, we observe that $215,717$ (approximately 71%) adversarial examples are predicted into one of the top-5 predictions of their unperturbed source images, where these classes in the top-5 are often highly similar to the correct predictions for the source images the adversarial examples are generated from (see Figure II in the supplementary material).

Although this graph hints that a large portion of untargeted adversarial transferability successes are (plausible) misclassifications rather than adversarial successes, on its own, it does not provide enough evidence to make such a claim. In order to solidify this observation, we expand on misclassifications and utilize the ImageNet class hierarchy. In Table 1, we provide the count and the percentage of adversarial examples that are originating from a number of collections and their intra-collection misclassification rates for a number of collections under the *Organism* branch of the hierarchy. Table 1 represents the aforementioned measurements for all adversarial examples that achieved adversarial transferability to any of the models and with any attack.

Naturally, the larger the collection, the higher the intra-collection misclassification rate will be. For example, a source image taken from the *Organism* collection has 409 other classes that may contribute to intra-collection misclassification. However, even for smaller, more granular collections such as the *Bird* collection, which only contains 59 classes, we observe that adversarial examples are more-often-than-not misclassified into the classes in the same collection. Furthermore, a number of collections such as *Canine*, *Bird*, *Reptilian*, and *Arthropod* stand out among other collections for having remarkably high intra-collection misclassification rates. For example, 84% of all adversarial examples that originate from a canine (i.e., dog) image are misclassified as another breed of canine.

Table 1: For the adversarial examples that achieved model-to-model transferability, intra-collection misclassifications and misclassifications into the top-{3,5} prediction classes in the target models are provided. The results for the adversarial examples are grouped into collections according to the classes of their source image origins.

Hierarchy	Collection	Classes in collection	Source images in collection	Adversarial examples originating from collection	Intra-collection misclassifications		Misclassification into top-K classes	
					Count	%	Top-3	Top-5
	All	1000	19,025	289,244	289,244	100.0%	59.6%	71.1%
1	Organism	410	9,390	147,621	132,865	**90.0%**	61.2%	72.8%
1.1	Creature	398	9,009	143,996	130,409	**90.6%**	61.4%	73.1%
1.1.1	Domesticated animal	123	2,316	50,036	41,978	**83.9%**	63.4%	75.6%
1.1.2	Vertebrate	337	7,692	126,913	112,828	**88.9%**	61.3%	73.2%
1.1.2.1	Mammalian	218	4,665	89,004	76,351	**85.8%**	61.4%	73.5%
1.1.2.1.1	Primate	20	475	9,333	5,301	**56.8%**	58.9%	70.4%
1.1.2.1.2	Hoofed mammal	17	419	6,206	2,751	44.3%	58.4%	71.6%
1.1.2.1.3	Feline	13	319	3,895	1,998	**51.3%**	64.3%	75.9%
1.1.2.1.4	Canine	130	2,502	53,294	45,089	**84.6%**	63.5%	75.7%
1.1.2.2	Aquatic vertebrate	16	366	5,355	2,383	44.5%	65.0%	75.6%
1.1.2.3	Bird	59	1,937	22,402	15,993	**71.4%**	59.8%	71.3%
1.1.2.4	Reptilian	36	547	7,635	4,795	**62.8%**	63.8%	75.2%
1.1.2.4.1	Saurian	11	188	2,416	1,050	43.5%	58.4%	71.1%
1.1.2.4.2	Serpent	17	223	3,202	1,700	**53.1%**	67.0%	77.1%
1.1.3	Invertebrate	61	1,317	17,083	10,698	**62.6%**	61.9%	72.3%
1.1.3.1	Arthropod	47	1,018	13,200	8,863	**67.1%**	63.1%	73.5%
1.1.3.1.1	Insect	27	652	7,850	4,468	**56.9%**	59.9%	70.5%
1.1.3.1.2	Arachnoid	9	189	2,824	1,476	**52.3%**	69.7%	79.5%
1.1.3.1.3	Crustacean	9	137	2,035	955	46.9%	70.0%	80.1%

In Table 1, we also provide misclassifications into the top-3 and the top-5 classes for adversarial examples that are originating from source images taken from individual collections. As can be seen, the observations we made when evaluating all adversarial examples also hold true for individual collections, where most of the adversarial examples in those collections have a misclassification rate of about 60% and 70% for the top-3 and the top-5 classes, respectively. To make matters worse, we can even see trends similar to the aforementioned observations when we filter adversarial examples for individual attacks and when we investigate misclassifications on a model-to-model basis, demonstrating that our observations are not specific to a single model or to one of the attacks. Extended results covering more collections and individual models/attacks can be found in the supplementary material (Table I to Table V).

5 Conclusions and outlook

In the context of a classification problem, what differentiates an adversarial success from a plausible misclassification? If an adversarial example is misclassified into a class that is highly similar to the class of its unperturbed origin, should it still be considered an adversarial success? In this case, how should we measure the similarity between the classes? The aforementioned questions are not trivial to answer, and different answers may find different logical explanations depending on the context of the evaluation performed. However, given that the threat of adversarial examples is evaluated from the perspective of security, does a semantically similar misclassification that has been made in the context of ImageNet (e.g., a brown dog breed misclassified as another brown dog breed) carry the same weight as a lethal misclassification in the context of self-driving cars (e.g., a road sign misclassification leading to an accident)?

Finding answers to the questions presented above requires meticulous investigations on the topic of misclassification classes, where these investigations should involve various threat scenarios, similar to the work presented in [33, 43, 44]. In this paper, we took one of the first steps in analyzing misclassification classes in the context of ImageNet, with the help of large-scale experiments and the ImageNet class hierarchy, showing that a large number of untargeted adversarial misclassifications in model-to-model transferability scenarios are, in fact, plausible misclassifications. In particular, we observe that categories under the *Organism* branch have considerably high intra-collection misclassifications compared to classes in the *Artifact* branch. To aid future work on this topic in the context of ImageNet, we share an easy-to-use class hierarchy of ImageNet, as well as other resources, in the following repository: https://github.com/utkuozbulak/imagenet-adversarial-image-evaluation.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A Deep Convolutional Encoder-Decoder Architecture For Image Segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[2] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion Attacks Against Machine Learning At Test Time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2013.

[3] N. Carlini and D. Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018.

[4] N. Carlini and D. A. Wagner. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017.

[5] N. Carlini and D. A. Wagner. Towards Evaluating The Robustness of Neural Networks. *2017 IEEE Symposium on Security and Privacy*, 2017.

[6] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu. Improving Black-box Adversarial Attacks with a Transfer-based Prior. In *Advances in Neural Information Processing Systems*, 2019.

[7] F. Croce and M. Hein. Sparse and Imperceivable Adversarial Attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[8] A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, and F. Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 321–338, Santa Clara, CA, Aug. 2019. USENIX Association.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[10] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations*, 2015.

[11] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger. Simple Black-box Adversarial Attacks. *International Conference on Machine Learning*, 2019.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, 2017.

[13] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten. Countering Adversarial Images Using Input Transformations. *CoRR*, abs/1711.00117, 2017.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[15] 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.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[17] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. SqueezeNet: AlexNet-level Accuracy with 50x Fewer Parameters and< 0.5 MB Model Size. *CoRR*, abs/1602.07360, 2016.

[18] D. Karmon, D. Zoran, and Y. Goldberg. Lavan: Localized and Visible Adversarial Noise. *International Conference on Machine Learning*, 2018.

[19] A. Krizhevsky and G. Hinton. Learning Multiple Layers Of Features From Tiny Images. Technical report, Citeseer, 2009.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, 2012.

[21] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial Examples In The Physical World. *Workshop Track, International Conference on Learning Representations*, 2016.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied To Document Recognition. *Proceedings of the IEEE*, 1998.

[23] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks For Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[24] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistant To Adversarial Attacks. *International Conference on Learning Representations*, 2018.

[25] P. McDaniel, N. Papernot, and Z. B. Celik. Machine Learning In Adversarial Settings. *IEEE Security & Privacy*, 2016.

[26] G. A. Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[27] U. Ozbulak, E. T. Anzaku, W. D. Neve, and A. V. Messem. Selection of Source Images Heavily Influences the Effectiveness of Adversarial Attacks. *CoRR*, abs/2106.07141, 2021.

[28] U. Ozbulak, B. Vandersmissen, A. Jalalvand, I. Couckuyt, A. Van Messem, and W. De Neve. Investigating the significance of adversarial attacks and their relation to interpretability for radar-based human activity recognition systems. *Computer Vision and Image Understanding*, 2021.

[29] N. Papernot, P. D. McDaniel, and I. Goodfellow. Transferability In Machine Learning: From Phenomena To Black-Box Attacks Using Adversarial Samples. *CoRR*, abs/1605.07277, 2016.

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, 2015.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015.

[32] O. Russakovsky and L. Fei-Fei. Attribute Learning in Large-scale Datasets. In *European Conference of Computer Vision (ECCV), International Workshop on Parts and Attributes*, 2010.

[33] L. Schwinn, R. Raab, A. Nguyen, D. Zanca, and B. Eskofier. Exploring Misclassifications of Robust Neural Networks to Enhance Adversarial Attacks. *CoRR*, abs/2105.10304, 2021.

[34] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-Cam: Why Did You Say That? Visual Explanations From Deep Networks Via Gradient-Based Localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[35] A. S. Shamsabadi, R. Sanchez-Matilla, and A. Cavallaro. Colorfool: Semantic Adversarial Colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[36] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks For Large-Scale Image Recognition. *International Conference on Learning Representations*, 2015.

[37] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving For Simplicity: The All Convolutional Net. *CoRR*, abs/1412.6806, 2014.

[38] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao. Is Robustness the Cost of Accuracy?– A Comprehensive Study on the Robustness of 18 Deep Image Classification Models. In *Proceedings of the European Conference on Computer Vision*, 2018.

[39] J. Su, D. V. Vargas, and K. Sakurai. Empirical Evaluation On Robustness Of Deep Convolutional Neural Networks Activation Functions Against Adversarial Perturbation. In *International Symposium on Computing and Networking Workshops (CANDARW)*. IEEE, 2018.

[40] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing Properties Of Neural Networks. *International Conference on Learning Representations*, 2014.

[41] K. Xu, S. Liu, P. Zhao, P.-Y. Chen, H. Zhang, Q. Fan, D. Erdogmus, Y. Wang, and X. Lin. Structured Adversarial Attack: Towards General Implementation and Better Interpretability. *International Conference on Learning Representations*, 2019.

[42] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. In *Conference on Fairness, Accountability, and Transparency*, 2020.

[43] C. Zhang, P. Benz, T. Imtiaz, and I. S. Kweon. Understanding Adversarial Examples from the Mutual Influence of Images and Perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14521–14530, 2020.

[44] Z. Zhao, Z. Liu, and M. Larson. On Success and Simplicity: A Second Look at Transferable Targeted Attacks. *CoRR*, abs/2012.11207, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [N/A] We did not propose a novel method; we simply performed an analysis using a viewpoint that was not taken before.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Not applicable to our study.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] Not applicable to our study.
   (b) Did you include complete proofs of all theoretical results? [N/A] Not applicable to our study.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] `https://github.com/utkuozbulak/imagenet-adversarial-image-evaluation`
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] The models used in this study are publicly available pretrained models in the PyTorch vision repository.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] Not applicable to our study.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Not applicable to our study.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] All assets used are already publicly available, having permissive licences that enable research.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We only included additional figures and tables in the supplementary material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All assets used are already publicly available, having permissive licences that enable research.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] All assets used are already publicly available, having permissive licences that enable research.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Not applicable to our study.

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Not applicable to our study.

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Not applicable to our study.