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

Utku Ozbulak, Maura Pintor, Arnout Van Messem, Wesley De Neve

GHENT UNIVERSITY
University of Cagliari
GHENT UNIVERSITY GLOBAL CAMPUS
LIÈGE université

A NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future

Abstract

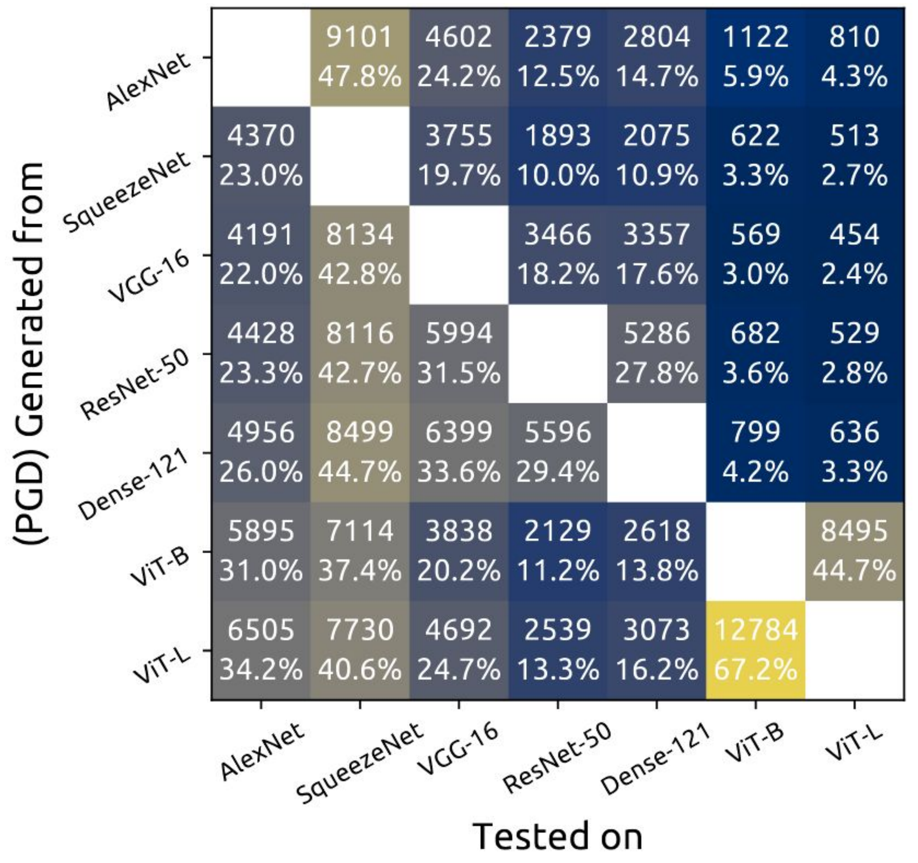
In order to evaluate attacks and defenses in the field of adversarial machine learning, ImageNet remains one of the most frequently used datasets. However, a topic that is yet to be investigated is the nature of the classes into which adversarial examples are misclassified.

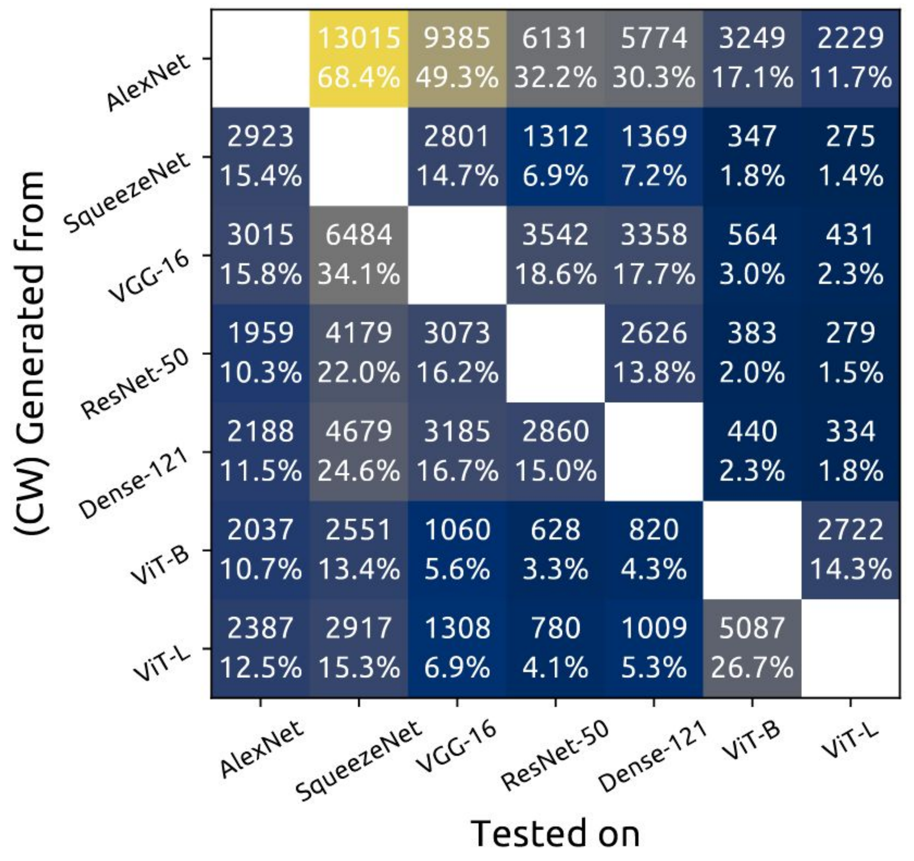
In this work, we perform a detailed analysis of these misclassification classes, leveraging the ImageNet class hierarchy and measuring the relative positions of the aforementioned type of classes in the unperturbed origins of the adversarial examples.

We find that a large portion of 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.

Experimental Approach

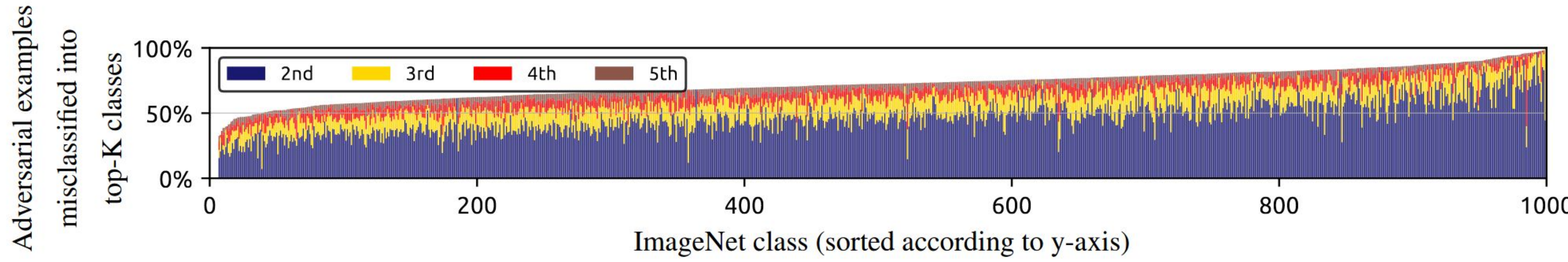
- ➔ Select 7 deep neural networks to evaluate adversarial model-to-model transferability.
  - AlexNet[1] , SqueezeNet[2], VGG-16[3], ResNet-50[4], DenseNet-121[5], ViT-B[6], and ViT-L[6].
- ➔ Filter (unperturbed) source images from ImageNet that are correctly classified by all selected models.
  - Result: 19,025 source images.
- ➔ Generate adversarial examples with the two most commonly used attacks: PGD[7] and CW[8].
  - Result: 289,244 adversarial examples.
- ➔ Evaluate model-to-model transferability success using the aforementioned 7 models.





Key Findings

➔ Most of the adversarial examples that achieve (untargeted) model-to-model transferability (i.e., adversarial examples misclassified by the target model) are misclassified into one of the top-{2,3,4,5} categories of its own (unperturbed) source image.



➔ When we analyze the misclassifications in detail with the help of the ImageNet class hierarchy, **we observe that a large portion of our adversarial examples are misclassified into classes that are in the same ImageNet collection as their (unperturbed) source image**, even for collections that are highly granular (e.g., types of animals).

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%

- 84% of the adversarial examples created from **dog** images are misclassified as another dog breed.
- 71% of the adversarial examples created from **bird** images are misclassified as another type of bird.
- 57% of the adversarial examples that are created from **insect** images are misclassified as another type of insect.
- 56% of the adversarial examples that are created from **vehicle** images are misclassified as another type of vehicle.
- 41% of the adversarial examples that are created from **structure** images are misclassified as another type of structure.

➔ In the context of ImageNet, most of the misclassifications made by deep neural networks for adversarial examples that achieve model-to-model adversarial transferability are genuine misclassifications that semantically make sense.

➔ Adversarial examples are not only misclassified into categories that are within the same collection in the ImageNet hierarchy, those categories are also, more-often-than-not, within the top-3/5 predictions obtained for the (unperturbed) source image counterparts.

Workshop

Code and resources

References

[1] A. Krizhevsky et al. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 2012.
[2] F. N. Iandola et al. Squeezenet: Alexnet-level Accuracy with 50x Fewer Parameters and< 0.5 mb Model Size. CoRR,abs/1602.07360, 2016.
[3] K. Simonyan et al. Very Deep Convolutional Networks For Large-Scale Image Recognition. International Conference on Learning Representations, 2015.
[4] K. He et al. Deep Residual Learning For Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[5] G. Huang et al. Densely Connected Convolutional Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[6] A. Dosovitskiy et al. An Image is Worth 16 x 16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations, 2021.
[7] A. Madry et al. Towards Deep Learning Models Resistant To Adversarial Attacks. International Conference on Learning Representations, 2018.
[8] N. Carlini et al. Towards Evaluating The Robustness of Neural Networks. IEEE Symposium on Security and Privacy, 2017.