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

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

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 types 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 (unperturbed) source image. We also find that a large subset of untargeted misclassifications are, in fact, misclassifications that semantically make sense.

Experimental Approach

- Select 7 deep neural networks to evaluate adversarial model-to-model transferability.
- AlexNet, SqueezeNet, VGG-16, ResNet-50, DenseNet-121, VIT-B, and VIT-L.
- 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 and CW.
- 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 (22, 30, 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).

- 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 another type of the same collection.

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 types 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 (unperturbed) source image. We also find that a large subset of untargeted misclassifications are, in fact, misclassifications that semantically make sense.

Experimental Approach

- Select 7 deep neural networks to evaluate adversarial model-to-model transferability.
  - AlexNet, SqueezeNet, VGG-16, ResNet-50, DenseNet-121, VIT-B, and VIT-L.
  - 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 and CW.
  - 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 (22, 30, 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).

- 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 another type of the same collection.