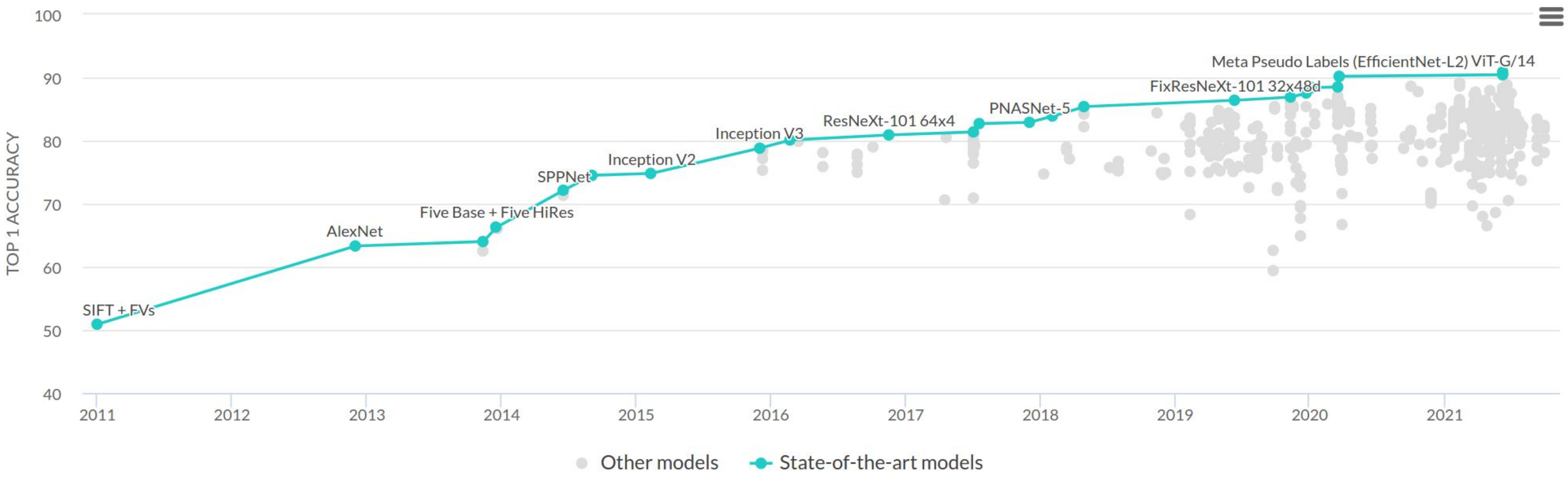
The British Machine Vision Conference (BMVC) - 2021

Selection of Source Images Heavily Influences the Effectiveness of Adversarial Attacks

Utku Ozbulak, Esla Timothy Anzaku, Wesley De Neve, Arnout Van Messem

Impact of deep learning models on computer vision

Deep learning methods drastically improved the state-of-the-art results obtained for different computer vision problems.



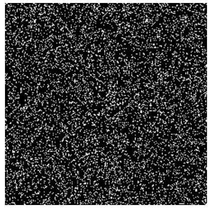
ImageNet validation set Top-1 accuracy for various deep learning models

Adversarial examples and deep learning

Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake.

Original image
Prediction: Cat
Confidence: 96%

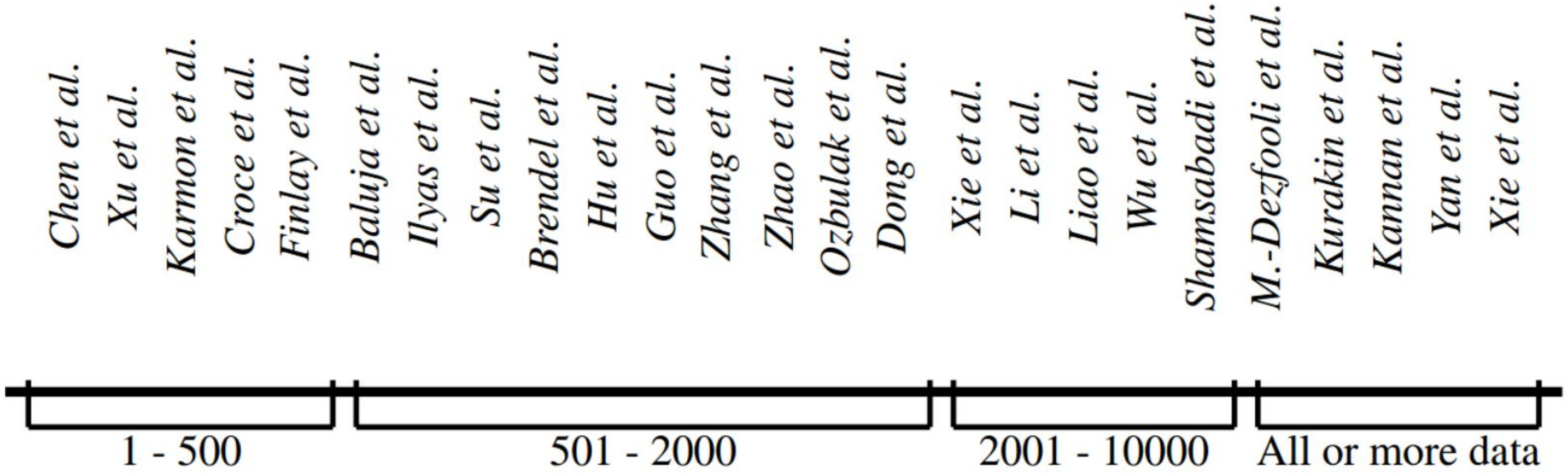
+Perturbation



Adversarial image
Prediction: Plane
Confidence: 99%

Computational cost of creating adversarial examples

Many research labs cannot utilize a large number of images from ImageNet for a detailed investigation on adversarial attacks due to computational limitations.



Number of source images used from ImageNet to create adversarial examples

Terminology

A *source image* is an (unperturbed) image used to create adversarial example(s).

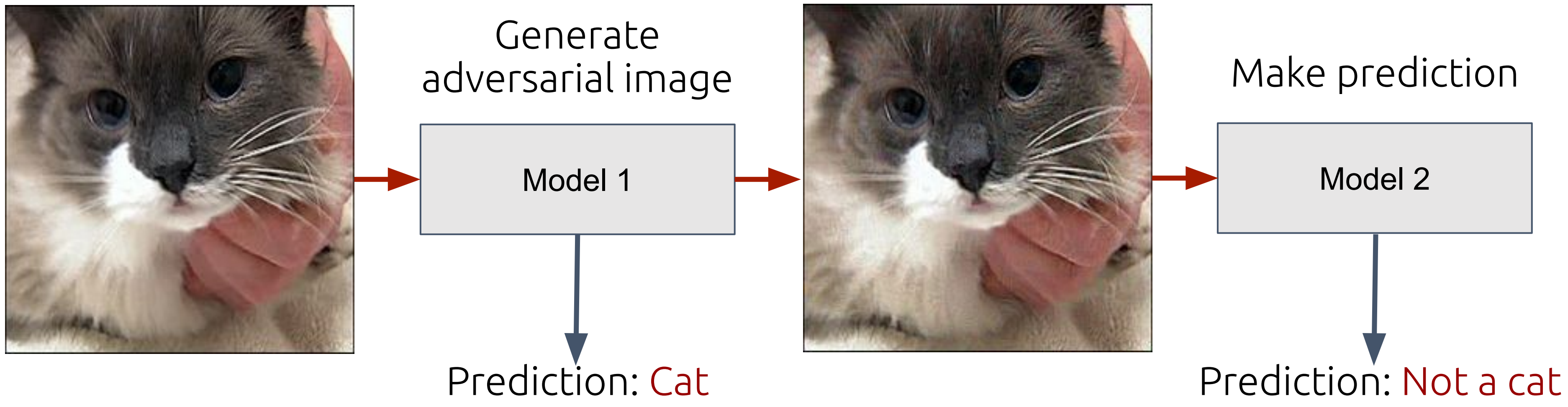
An *adversarial perturbation* is the noise added to a source image in order to create an adversarial example.

A *source model* is the model used to calculate the adversarial perturbation.

A *target model* is the model that a perturbed image is tested on to observe whether or not it became an adversarial example.

Terminology

(Untargeted) model-to-model transferability success: when an adversarial example created by a model (Model 1) is also misclassified by another model (Model 2).



The problem

Given that a large number of studies use a limited number of source images to create adversarial examples, how representative are the results obtained when using a subset of source images in terms of:

- Creating adversarial examples;
- Adversarial model-to-model transferability success;
- Required perturbation to achieve model-to-model transferability.

Experimental setup: deep learning models

Models	*ImageNet Accuracy (Top-1 / Top-5)*
(2013) AlexNet	56 % / 79 %
(2016) SqueezeNet	58 % / 80 %
(2014) VGG-16	71 % / 90 %
(2015) ResNet-50	76 % / 92 %
(2016) DenseNet-121	74 % / 91 %
(2020) ViT Base-16/224	80 % / 97 %
(2020) ViT Large-16/224	82 % / 97 %

Experimental setup: source images

We only use source images from the ImageNet validation set that are correctly classified by all models, effectively eliminating images that are hard to classify for at least one model.

19,025 source images, corresponding to 38% of the validation set.

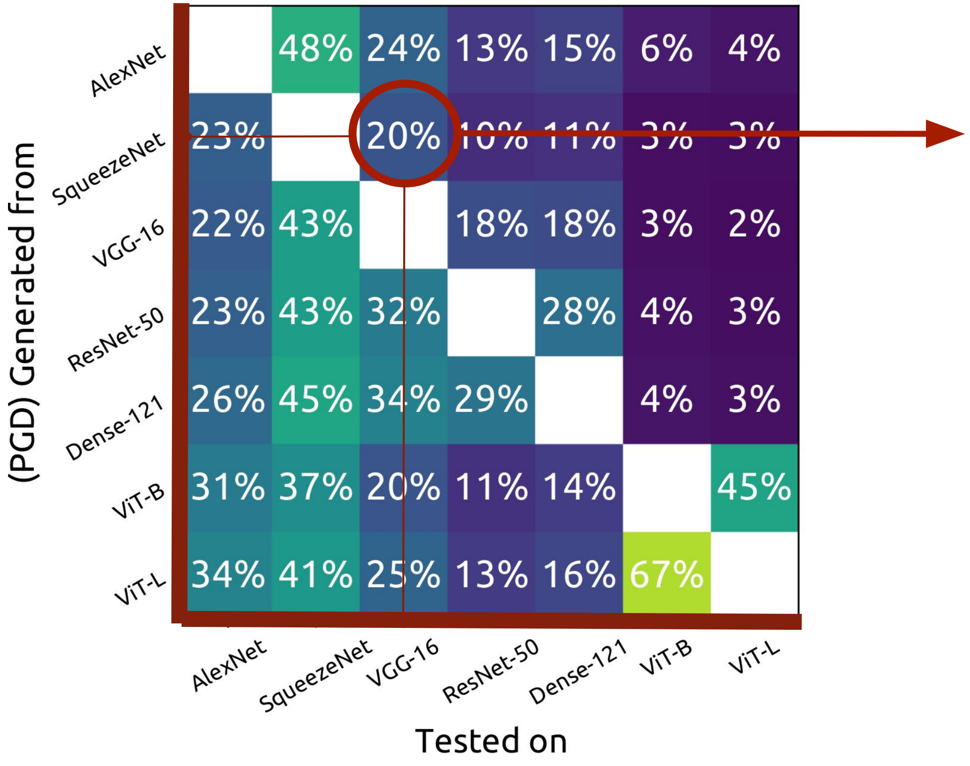
Experimental setup: adversarial attacks

Adversarial attacks:

- Projected Gradient Descent (PGD);
- Carlini & Wagner's Attack (CW);
- Momentum Iterative Fast Gradient Sign (MI-FGSM).

Experimental results: model-to-model transferability

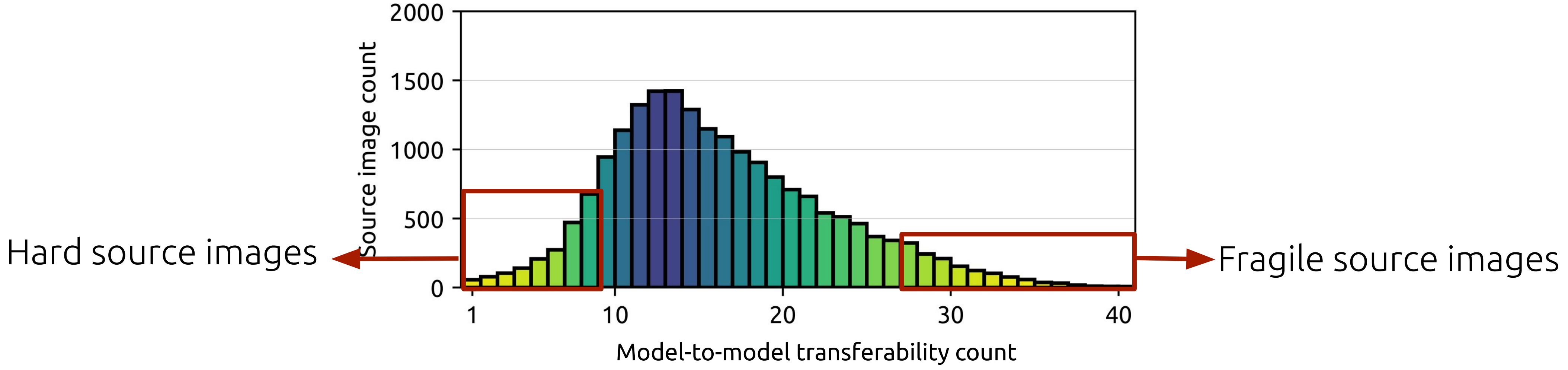
We attempt to create adversarial examples and report the success rate for each attack for each model-to-model scenario.



PGD was able to convert 20% (3,755) of the source images to adversarial examples that achieve model-to-model transferability from SqueezeNet to VGG-16.

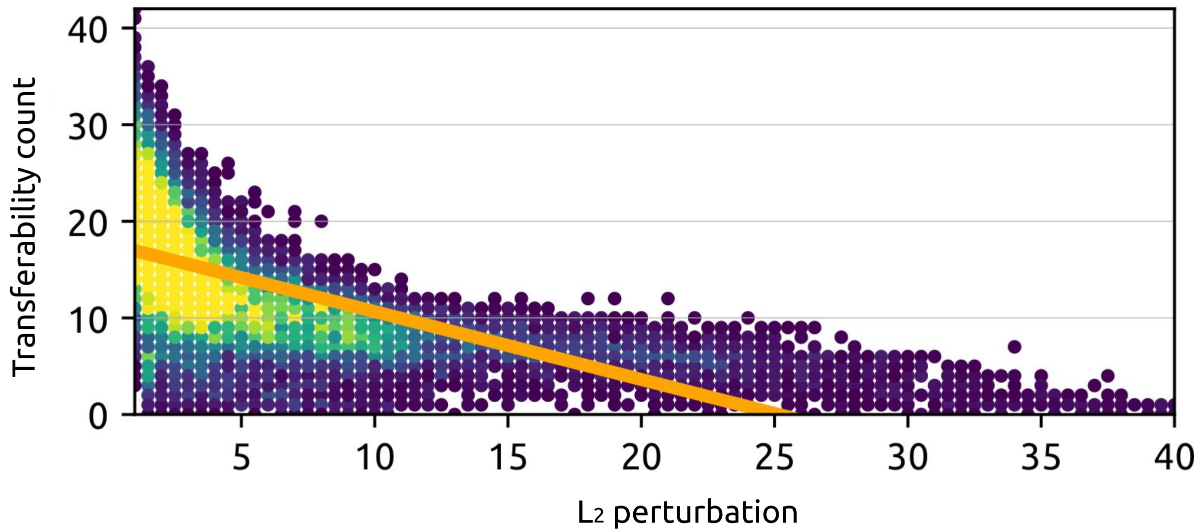
Experimental results: transferability count per source image

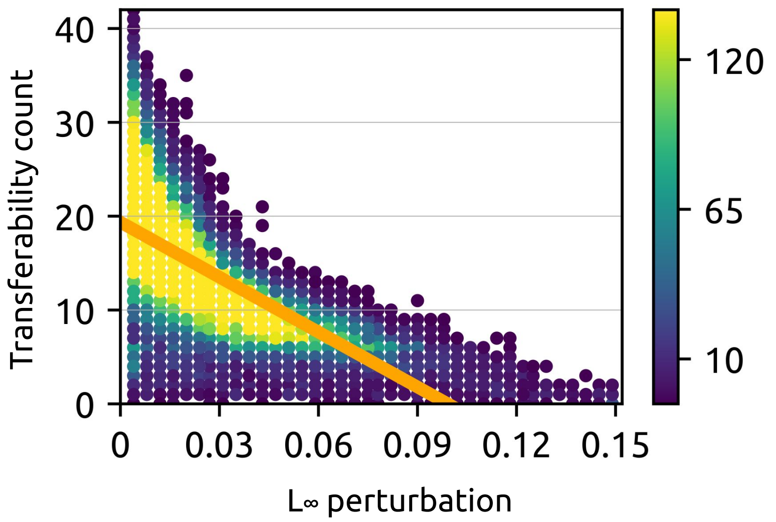
When we investigate the model-to-model transferability count for each source image, we observe a large discrepancy between fragile and hard source images.



Experimental results: transferability and perturbation

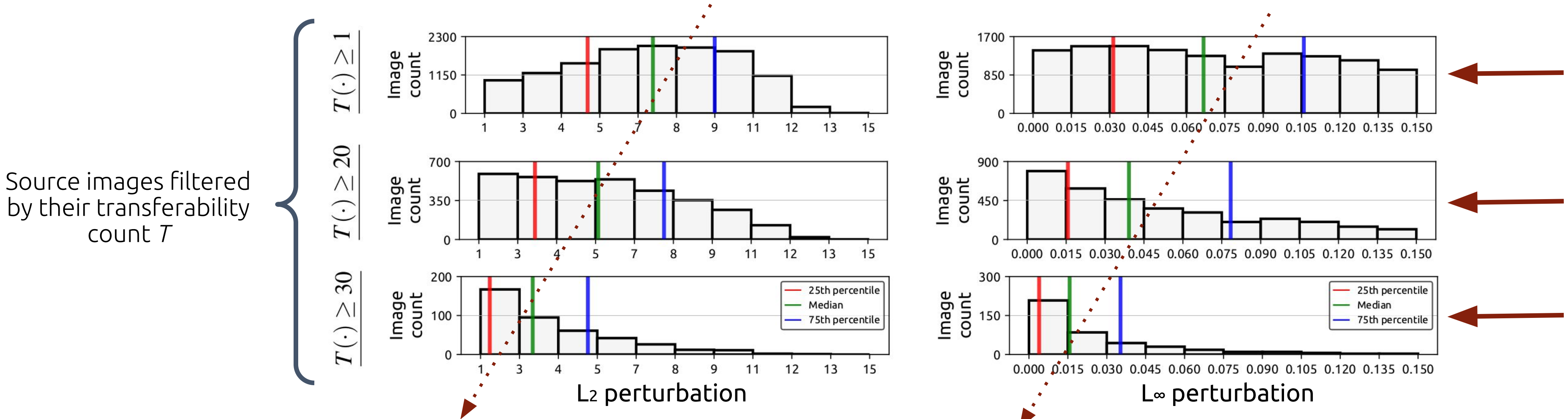
Source images that have high transferability counts are more likely to achieve model-to-model transferability with less perturbation when the perturbation is measured with L_2 or L_∞ norms.





Experimental results: transferability and perturbation

When we progressively filter source images based on their transferability count, we once again observe that the source images with high transferability counts have less perturbation.



Source images that achieved adversarial transferability to ViT-B are selected based on transferability count.

Experimental results: identifying fragile source images

Instead of devising a large-scale transferability scenario, we aim to identify whether or not a source image is fragile based on its prediction confidence.

Correlation of various error estimates based on prediction confidence with transferability and $L_{\{2,\infty\}}$ norms of perturbation.

Error measurement	PGD			CW			MI-FGSM		
	$T(\cdot)$	$d_2(\cdot)$	$d_\infty(\cdot)$	$T(\cdot)$	$d_2(\cdot)$	$d_\infty(\cdot)$	$T(\cdot)$	$d_2(\cdot)$	$d_\infty(\cdot)$
$Q(P(\theta,\boldsymbol{x}))$	0.58	**−0.64**	**−0.58**	**0.57**	**−0.59**	**−0.66**	0.42	−0.54	−0.54
$1-\max(P(\theta,\boldsymbol{x}))$	**0.61**	−0.60	−0.57	**0.57**	−0.54	−0.63	**0.43**	**−0.58**	**−0.57**
$\text{MSE}(P(\theta,\boldsymbol{x}),\boldsymbol{y})$	0.56	−0.57	−0.53	0.56	−0.51	−0.61	0.37	−0.51	−0.53
$\text{WD}(P(\theta,\boldsymbol{x}),\boldsymbol{y})$	0.33	−0.35	−0.37	0.33	−0.32	−0.37	0.29	−0.38	−0.38

Q(P(θ, x))	Ratio of predictions: second largest to the largest
1-max(P(θ, x))	Prediction error made for the correct class
MSE(P(θ,x), y)	Mean squared error
WD(P(θ,x), y)	Wasserstein distance

Experimental results: filtering source images based on Q(.)

We filter source images based on quartiles of Q(P(θ, x)) and observe the difference in model-to-model transferability success and the perturbation.

Properties of adversarial examples created from randomly sampling 1, 000 source images 10, 000 times. Adversarial examples are generated from DenseNet-121 and adversarially transferred to ResNet-50.

			All images	Hard images		Easy (fragile) images	
			$\mathbb{S}$	$\mathbb{S}_{Q<10}$	$\mathbb{S}_{Q<25}$	$\mathbb{S}_{Q>90}$	$\mathbb{S}_{Q>75}$
Source images in set:			19,025	1,904	4,758	1,904	4,758
Transferability	PGD	Low	23.9%	5.2%	6.9%	65.8%	50.1%
		Avg	29.4%	7.4%	9.8%	69.2%	55.8%
		High	35.2%	9.8%	13.1%	72.8%	61.2%
	CW	Low	10.3%	0.8%	1.6%	43.8%	29.0%
		Avg	15.0%	1.7%	3.2%	48.6%	33.7%
		High	19.8%	2.8%	5.2%	52.5%	39.2%
Perturbation (L_2 / L_∞)	PGD	Low	6.41 / 0.06	7.50 / 0.08	7.47 / 0.08	5.28 / 0.04	5.86 / 0.05
		Avg	6.97 / 0.07	8.01 / 0.09	8.10 / 0.09	5.54 / 0.05	6.25 / 0.06
		High	7.50 / 0.08	8.53 / 0.10	8.65 / 0.10	5.78 / 0.06	6.49 / 0.06
	CW	Low	2.77 / 0.07	2.95 / 0.08	2.97 / 0.8	2.42 / 0.05	2.68 / 0.07
		Avg	3.21 / 0.08	3.41 / 0.09	3.58 / 0.9	2.59 / 0.06	2.91 / 0.07
		High	3.66 / 0.10	3.89 / 0.10	4.36 / 0.11	2.75 / 0.07	3.18 / 0.08

Takeaway messages

- Not all source images are equal when generating adversarial examples.
- Adversarial examples created from a subset of fragile source images achieve unnaturally high model-to-model transferability.
- Fragile source images also become adversarial examples with considerably less perturbation.
- The prediction confidence of a source image, combined with various error estimations, is a decent baseline indicator for detecting fragile source images.

Thank you!

- If you have any queries regarding our research, don't hesitate to send an email to: utku.ozbulak@ugent.be

- The code for this research is released at the following repository: github.com/utkuozbulak/imagenet-adversarial-image-evaluation

Arms race between adversarial attacks and defenses

For each defense that prevents an adversarial attack, a novel attack that bypasses that defense is discovered.

Adversarial attack → Adversarial training

Multi-staged attacks → Distillation, MagNet

Optimization-based attacks → Thermometer detection

… → …

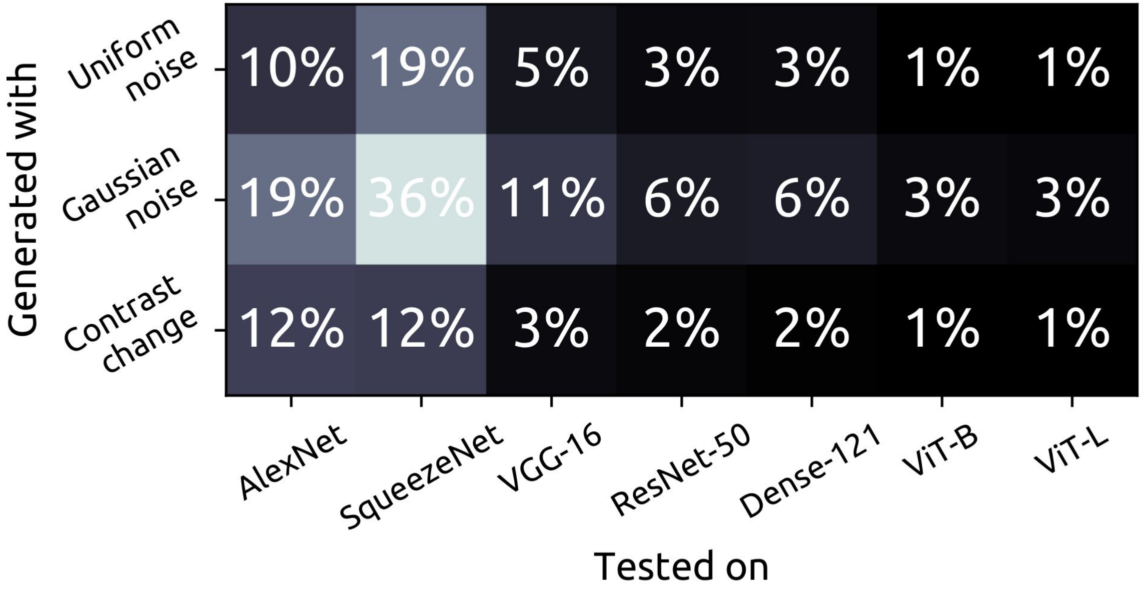
Future work

- Investigate the impact of fragile/hard images on adversarial defenses. Do adversarial examples created from fragile or hard images bypass defenses easier?

- We observe that a large number of adversarial examples are misclassified into categories that are similar to the category of their source image counterparts. Can we quantify this similarity?

Experimental results: non-adversarial noise

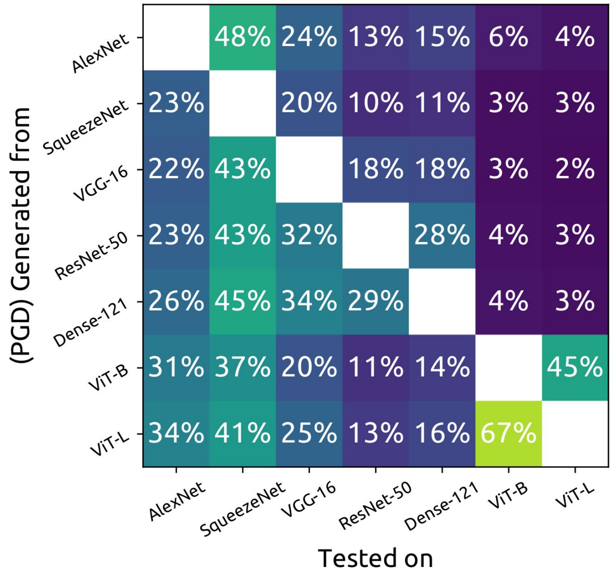
9,615 source images (~50%) have their predictions changed with commonly-used non-adversarial noise generation techniques.



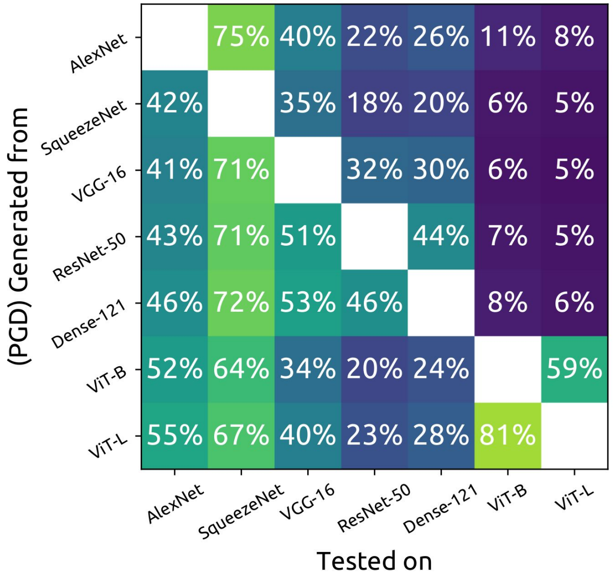
Let us call these images, that change their predictions easily, fragile source images.

Experimental results: fragile and hard source images

All source images



Easy (fragile) source images



Remaining source images