Transferability of Adversarial Examples to Attack Cloud-based Image Classification Service

Dou Goodman(兜哥) @ Baidu X-Lab

Our Team X-Lab

AI Security Research @  Baidu X-Lab

Open Source Projects:

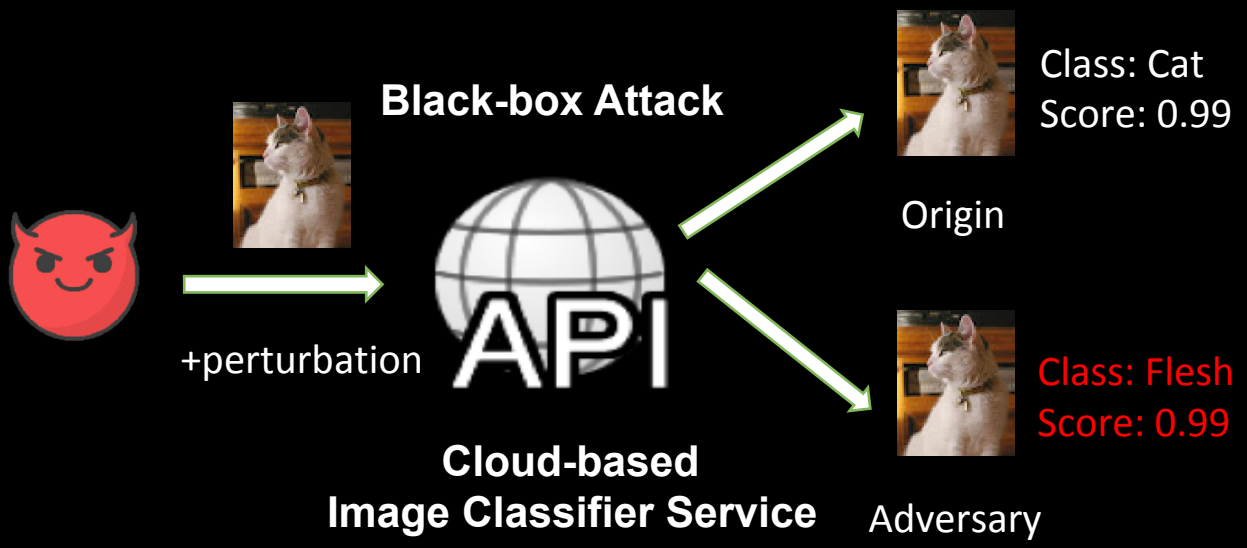








Today's Topics



Demo:Fool Google Image Search



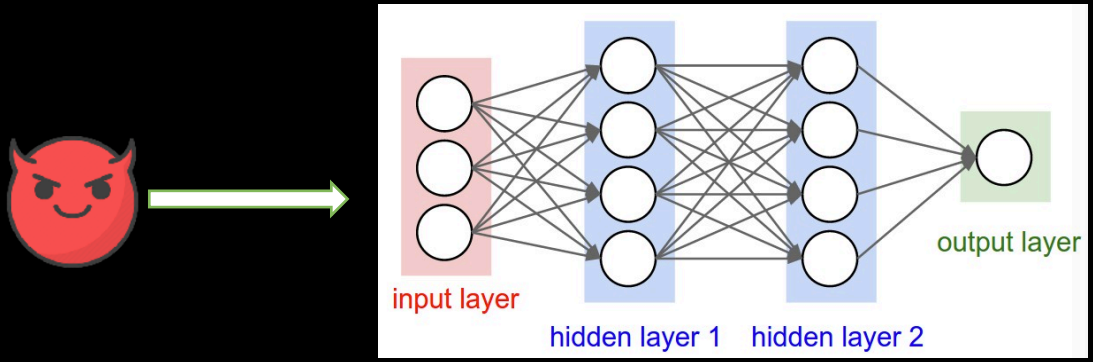
Demo:Fool Google Image Search



Attacks Overview

- We propose Fast Featuremap Loss PGD(FFL-PGD) untargeted attack based on Substitution model ,which achieve a high evasion rate with a very limited number of queries.
- Instead of millions of queries in previous studies, our method find the adversarial examples using average only one or two of queries.

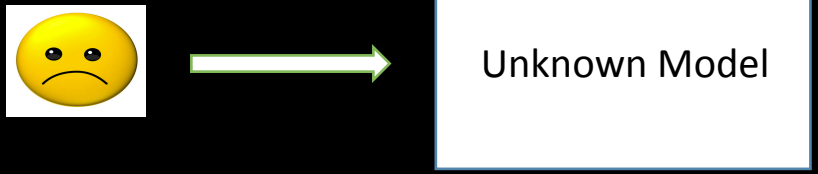
White-box Attack is Easy



The attacker knows the network structure and parameters, and has unlimited access to model input

Black-box Attack is Difficult

Unknown parameters
Unknown network structure

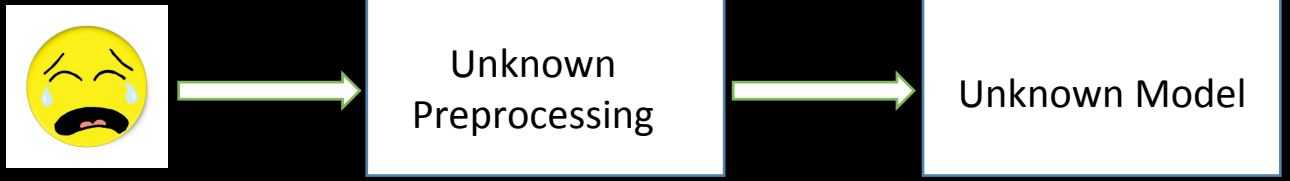


The attacker can unlimited access to model input

Attack Cloud-based Image Classifier Service is More Difficult !

resize,crop,blur,…

Unknown parameters
Unknown network structure



The attacker can only access to model input with unknown preprocessing and limited queries

Keeping model in cloud provides a FALSE sense of security !



(Img from: http://www.sohu.com/a/215163641_115479)

Steps of Our Attack

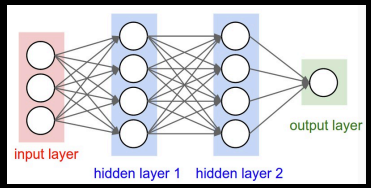


Step1:Substitute Model Training
Step2:Adversarial Sample Crafting

Step 2

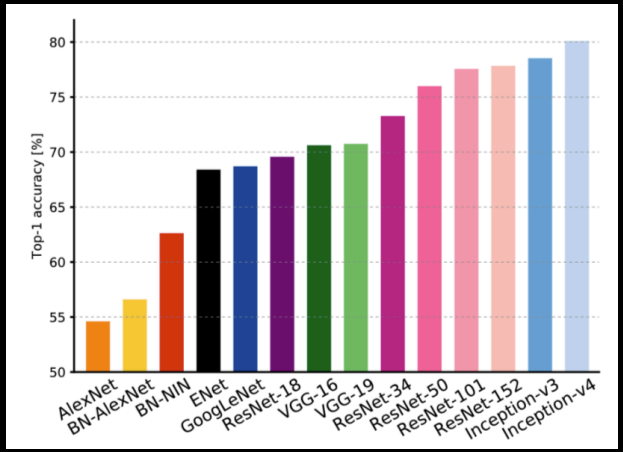


Step 1



Substitute Model Training(1)

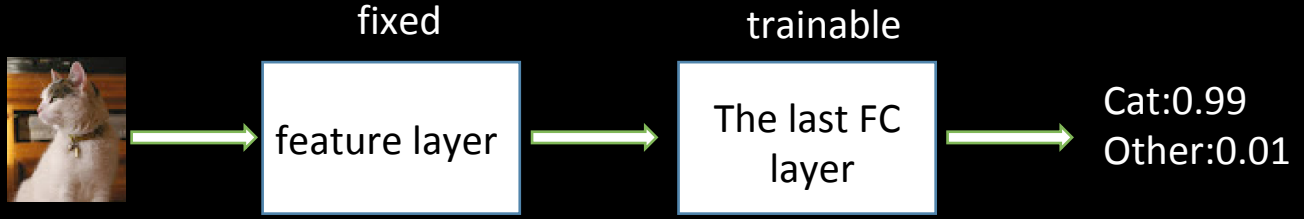
- We can DNNs which pretrained on ImageNet as our substitute model .
- Better top-1 accuracy means stronger feature extraction capability.



Top1 vs. network. Top-1 validation accuracies for top scoring single-model architectures (Img from

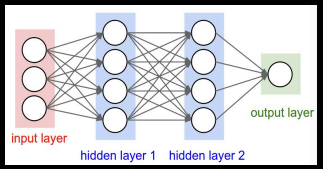
Substitute Model Training(2)

- We simplify untargeted attack into binary classification problem :Cat or not?
- We fix the parameters of the feature layer and train only the full connection layer of the last layer.

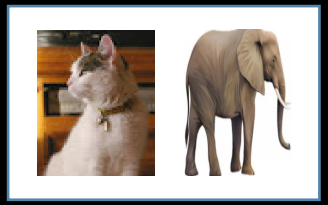


Substitute Model Training(3)

Key point:
we use images which we will attack as our training set



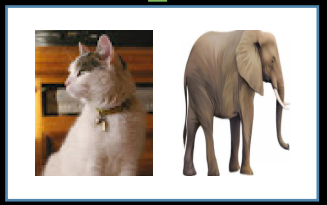
Train DNNs on S'



Initial training set S



Label the substitute training set



Labeled training set S'

Adversarial Sample Crafting(1)

We propose Fast Featuremap Loss PGD attack which has a novel loss function to improve the success rate of transfer attack.
The loss function L is defined as:

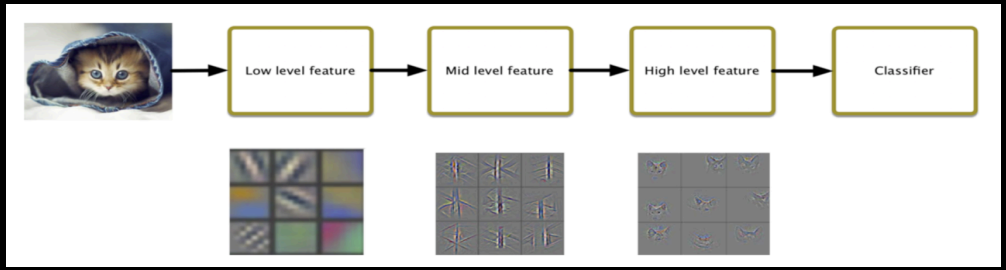
$$L = class_loss + \beta * FeatureMaps_loss$$

Adversarial Sample Crafting(2)

- Class Loss makes the result of classification wrong
- FeatureMap Loss which is the output of the last convolution layer of the substitute model, represents the highest level of semantic features of the convolution layer and improves transferability of adversarial sample

Adversarial Sample Crafting(3)

Illustration of cat recognition, the first convolution layer mainly recognizes low level features such as edges and lines. In the last convolution layer, it recognizes high level features such as eyes and nose.



Adversarial Sample Crafting(4)

We assume the original input is O, the adversarial example is ADV , and the featuremap loss is:

$$FeatureMap_loss(ADV, O) = \|L_n(ADV) - L_n(O)\|_2$$

Datasets and Preprocessing(1)

- 100 cat images and 100 other animal images are selected from the ImageNet val set.
- Images are clipped to the size of 224×224×3
- Image format is RGB

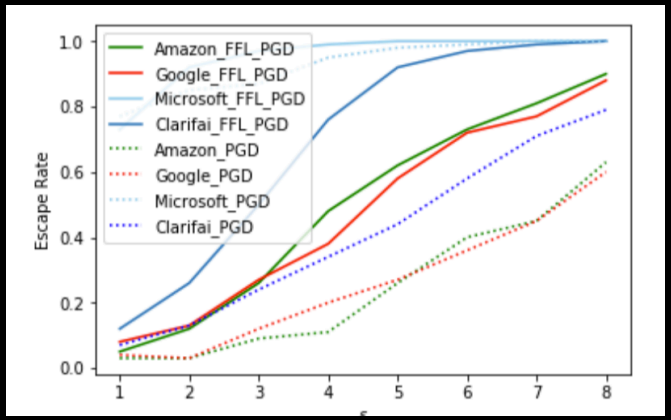
Datasets and Preprocessing(2)

- We use these 100 images of cats as original images to generate adversarial examples and make a black-box untargeted attack against real-world cloud-based image classification services .
- We count the number of top-1 misclassification to calculate the escape rate.

Attack Evaluation

- We choose ResNet-152 as our substitute model
- We launche PGD and FFL-PGD attacks against our substitute model to generate adversarial examples.
- We compare FFL-PGD with PDG and ensemble-model attack , which are considered to have good transferability .

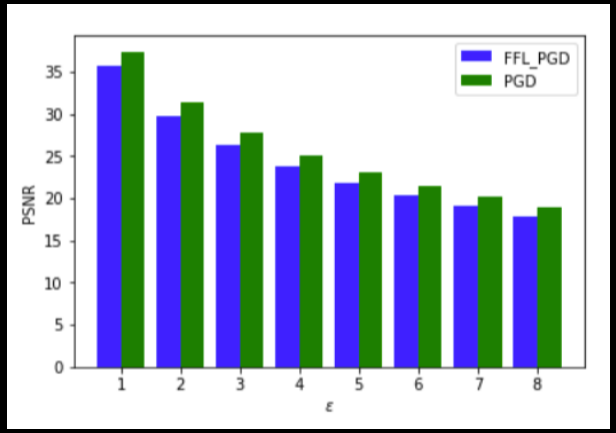
Attack Evaluation: Escape Rates



We increase step size ε from 1 to 8, the figure records the escape rates of PGD and FFL-PGD attacks

- FFL-PGD attack has a success rate over 90% among different cloud-based image classification services .
- Our FFL-PGD has a better transferability than PGD

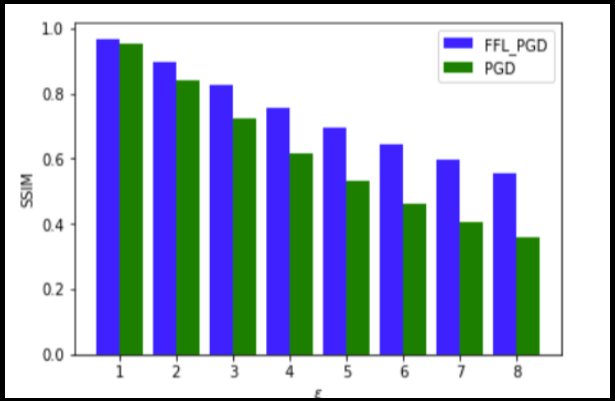
Attack Evaluation:PSNR



The figure records the PSNR of PGD and FFL-PGD attacks

PGD has a higher PSNR ,which is considered as better image quality .But both of them higher than 20dB when ε from 1 to 8, which means both of them are considered acceptable for image quality.

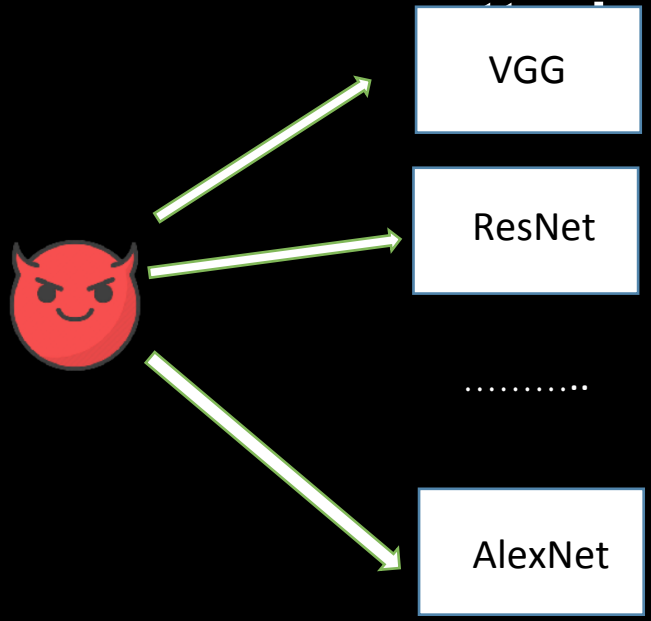
Attack Evaluation:SSIM



The figure records the SSIM of PGD and FFL-PGD attacks

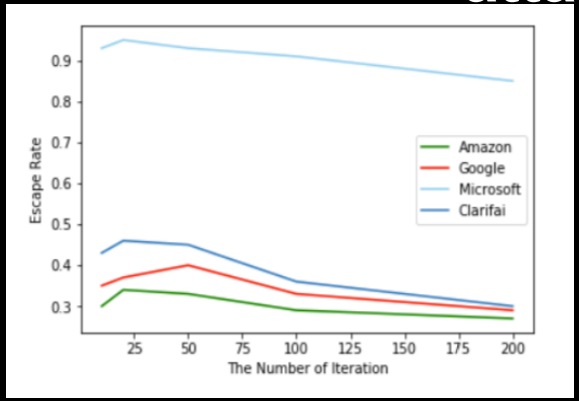
FFL-PGD has a higher SSIM ,which is considered as better image similarity

Attack Evaluation: Ensemble-model



Ensemble-model attack a lot of DNNs to generate adversarial examples which can fool them at once

Attack Evaluation: Ensemble-model attack



The figure records the escape rate of ensemble-model attack

- The escape rates of Amazon, Google and Clarifai are below 50%
- The transferability decreases in the face of the pre-processing of cloud services,such as resizing,cropping

Conclusion

- Keeping model in cloud provides a FALSE sense of security
- Our FFL-PGD attack have a success rate over 90% among different cloud-based image classification services using only two queries per image