SHIELD: Fast, Practical Defense and Vaccination for Deep Learning using JPEG Compression

Nilaksh Das
nilakshdas@gatech.edu

Madhuri Shanbhogue
madhuri.shanbhogue@gatech.edu

Shang-Tse Chen
schen351@gatech.edu

Fred Hohman
fredhohman@gatech.edu

Siwei Li
robertsiweili@gatech.edu

Li Chen
li.chen@intel.com

Michael E. Kounavis
michael.e.kounavis@intel.com

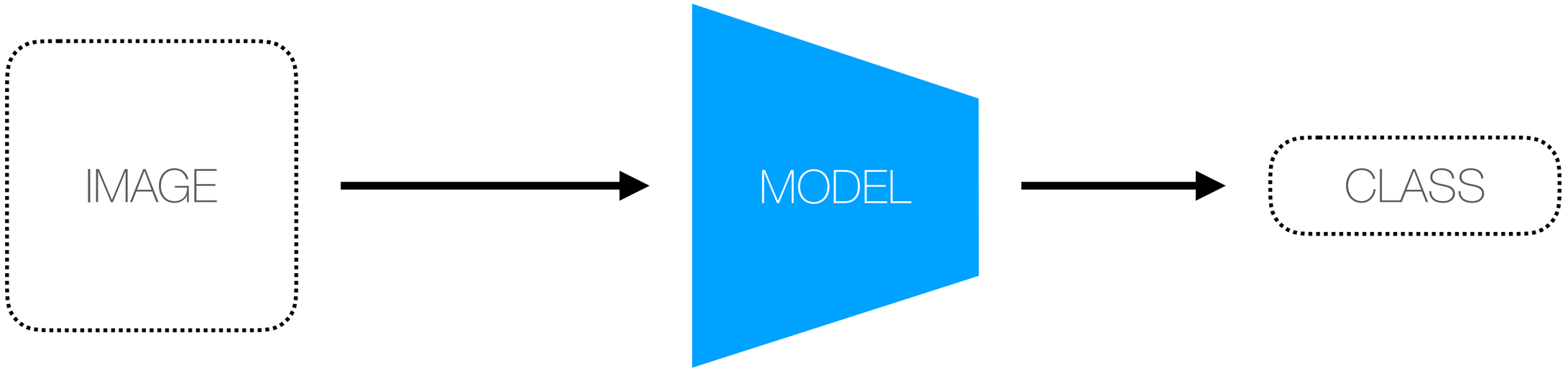
Duen Horng Chau
polo@gatech.edu

KDD

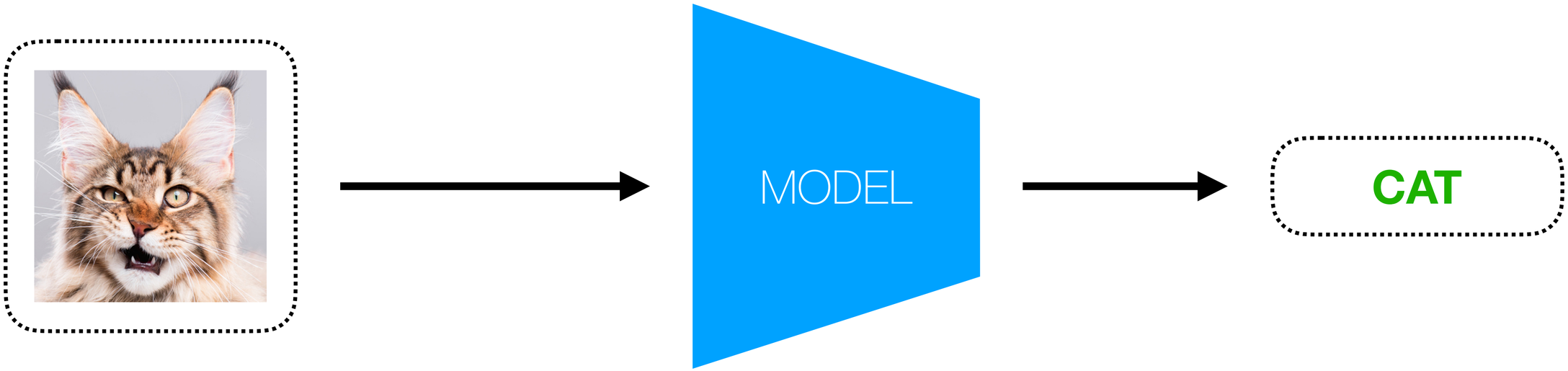
Georgia Tech

intel

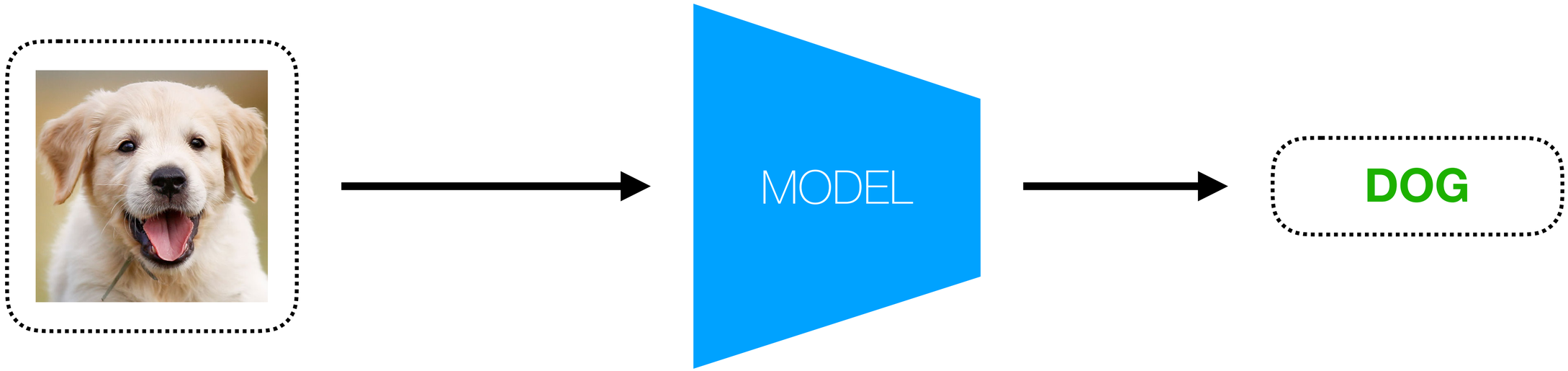
BACKGROUND / Deep Learning for Image Classification



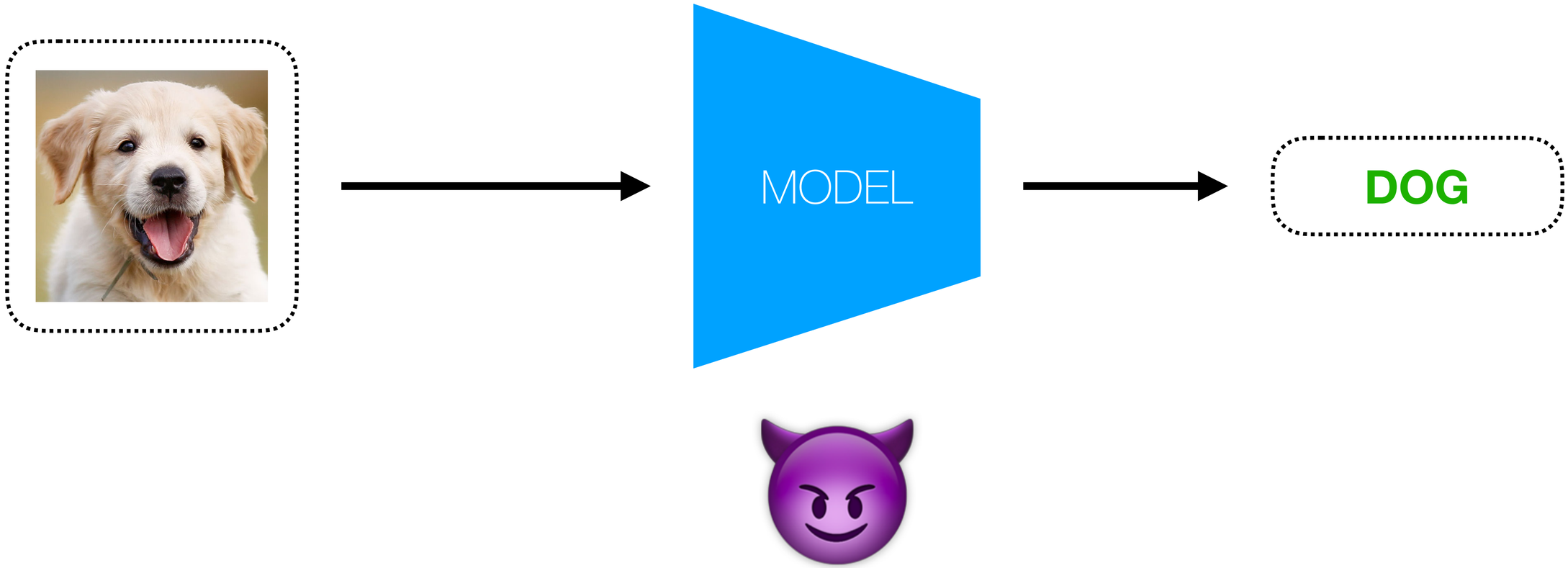
BACKGROUND / Deep Learning for Image Classification



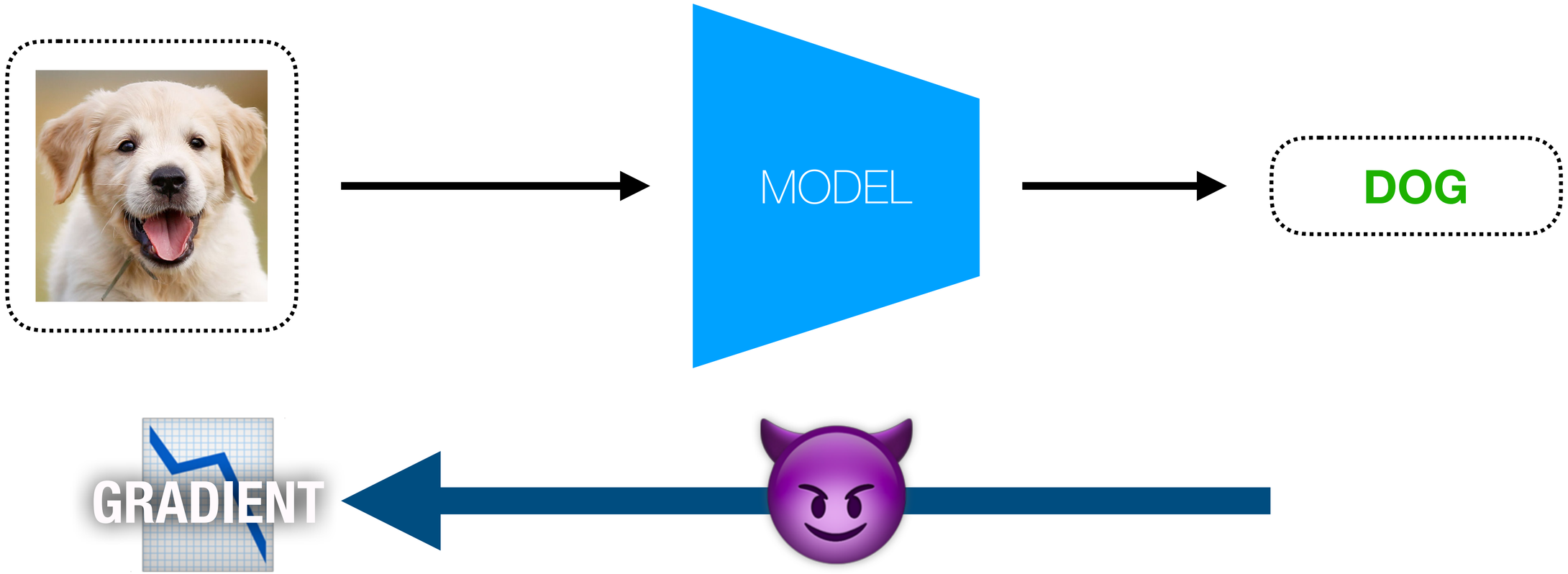
BACKGROUND / Deep Learning for Image Classification



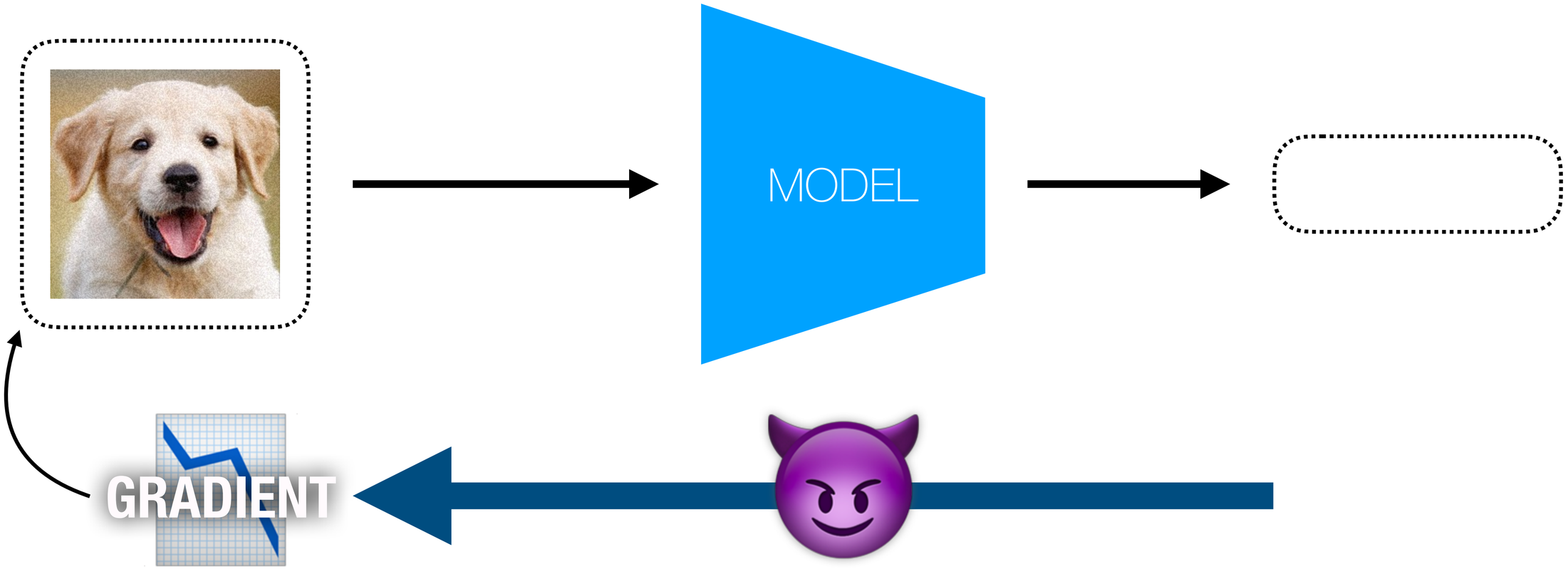
BACKGROUND / *Adversarial Attack on Deep Learning*



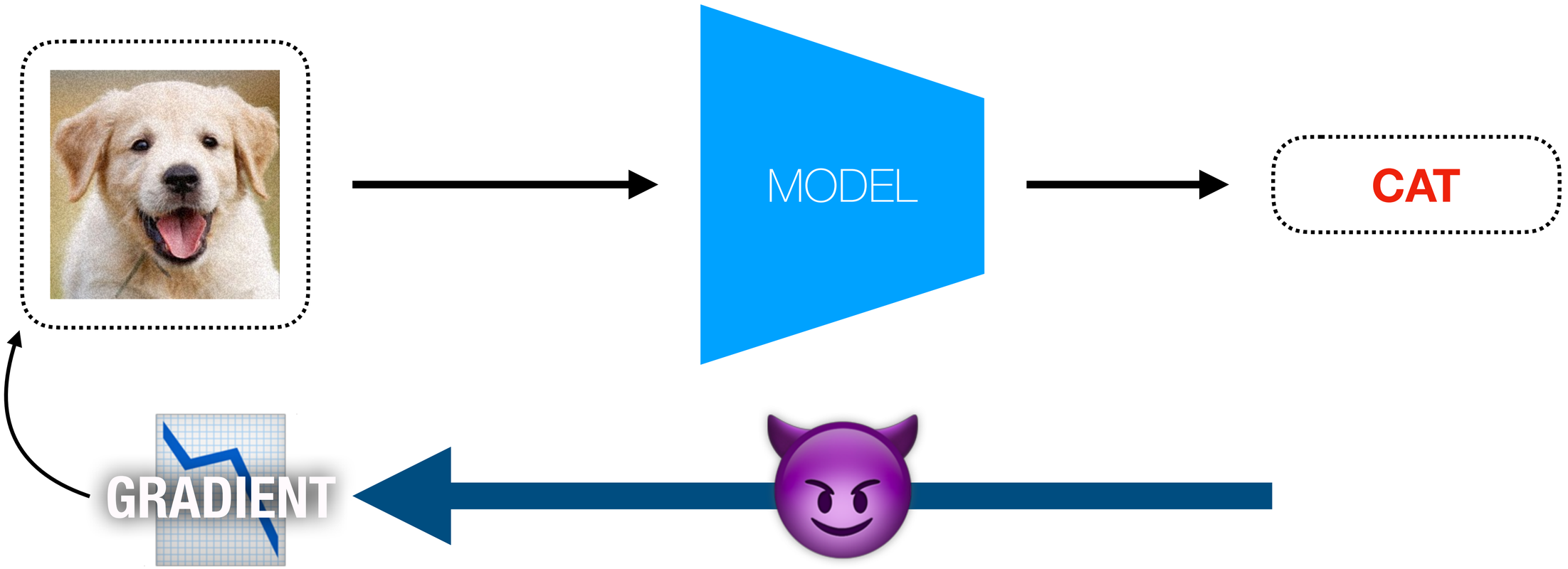
BACKGROUND / *Adversarial Attack on Deep Learning*



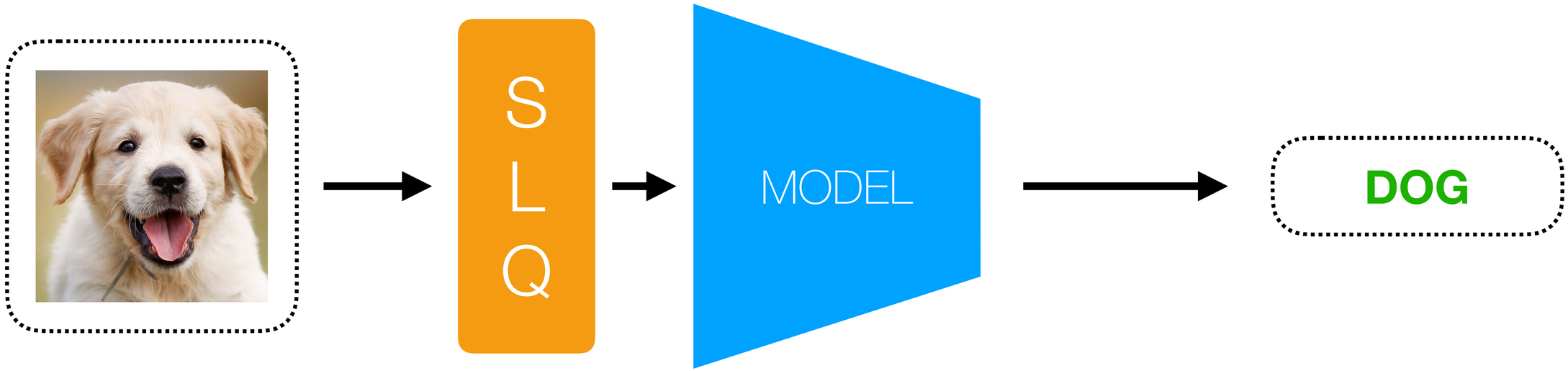
BACKGROUND / *Adversarial Attack on Deep Learning*



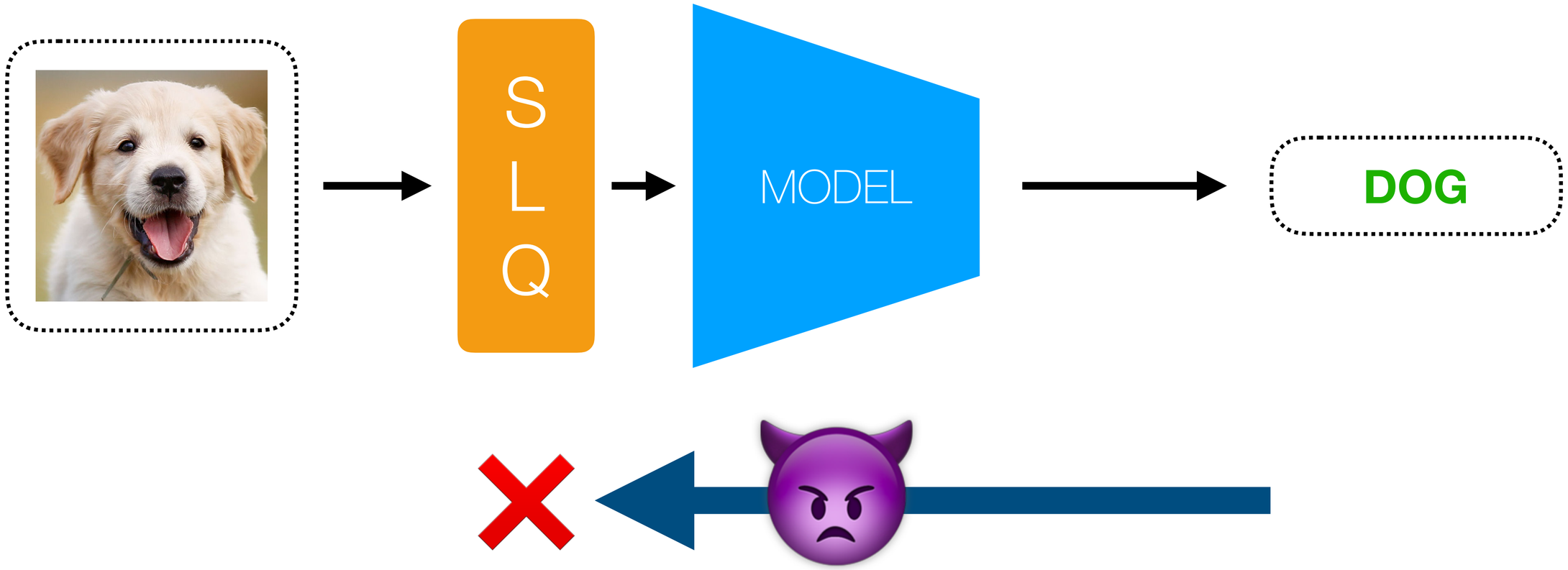
BACKGROUND / *Adversarial Attack on Deep Learning*



Stochastic Local Quantization (SLQ)



Stochastic Local Quantization (SLQ)



SLQ leverages JPEG compression

SLQ leverages JPEG compression

JPEQ Quality 80

JPEQ Quality 60

JPEQ Quality 40

JPEQ Quality 20

SLQ leverages JPEG compression

JPEQ Quality 80

JPEQ Quality 60

JPEQ Quality 40

JPEQ Quality 20

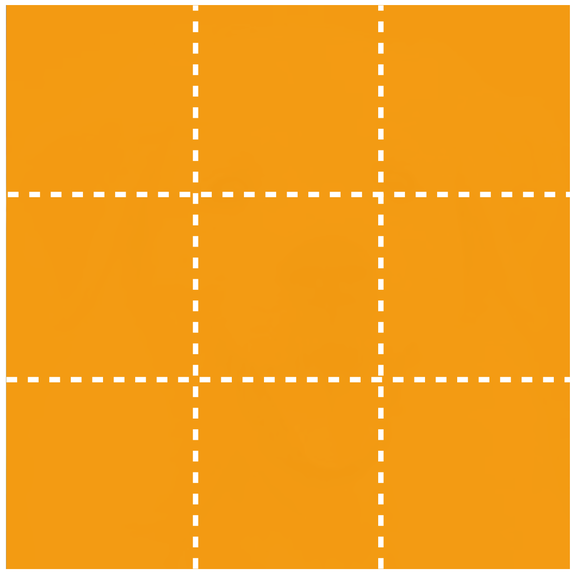
SLQ leverages JPEG compression

JPEQ Quality 80

JPEQ Quality 60

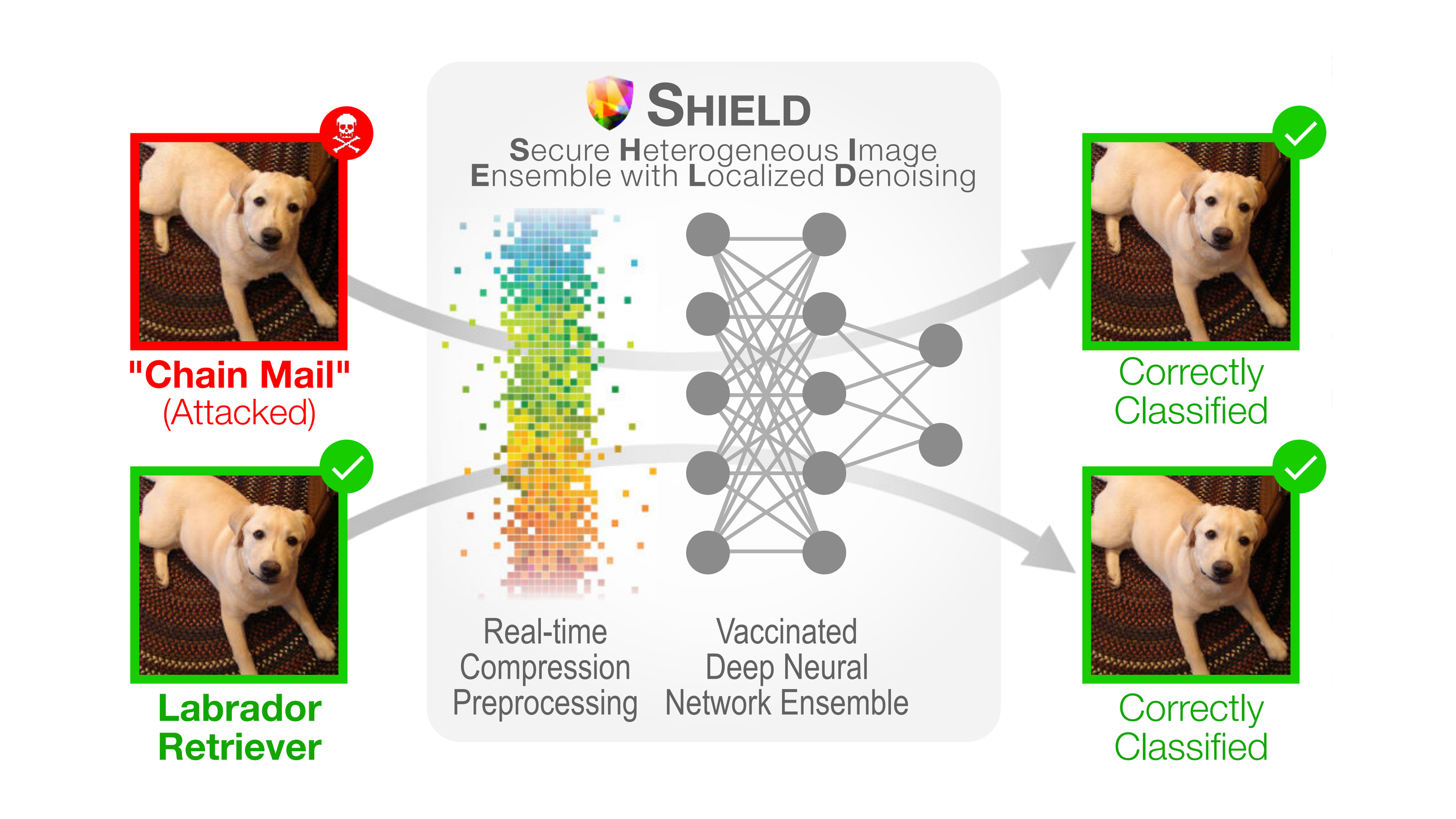
JPEQ Quality 40

JPEQ Quality 20



SLQ applies JPEG compression
of a random quality to each
8 x 8 block of the image

* larger blocks shown for presentation

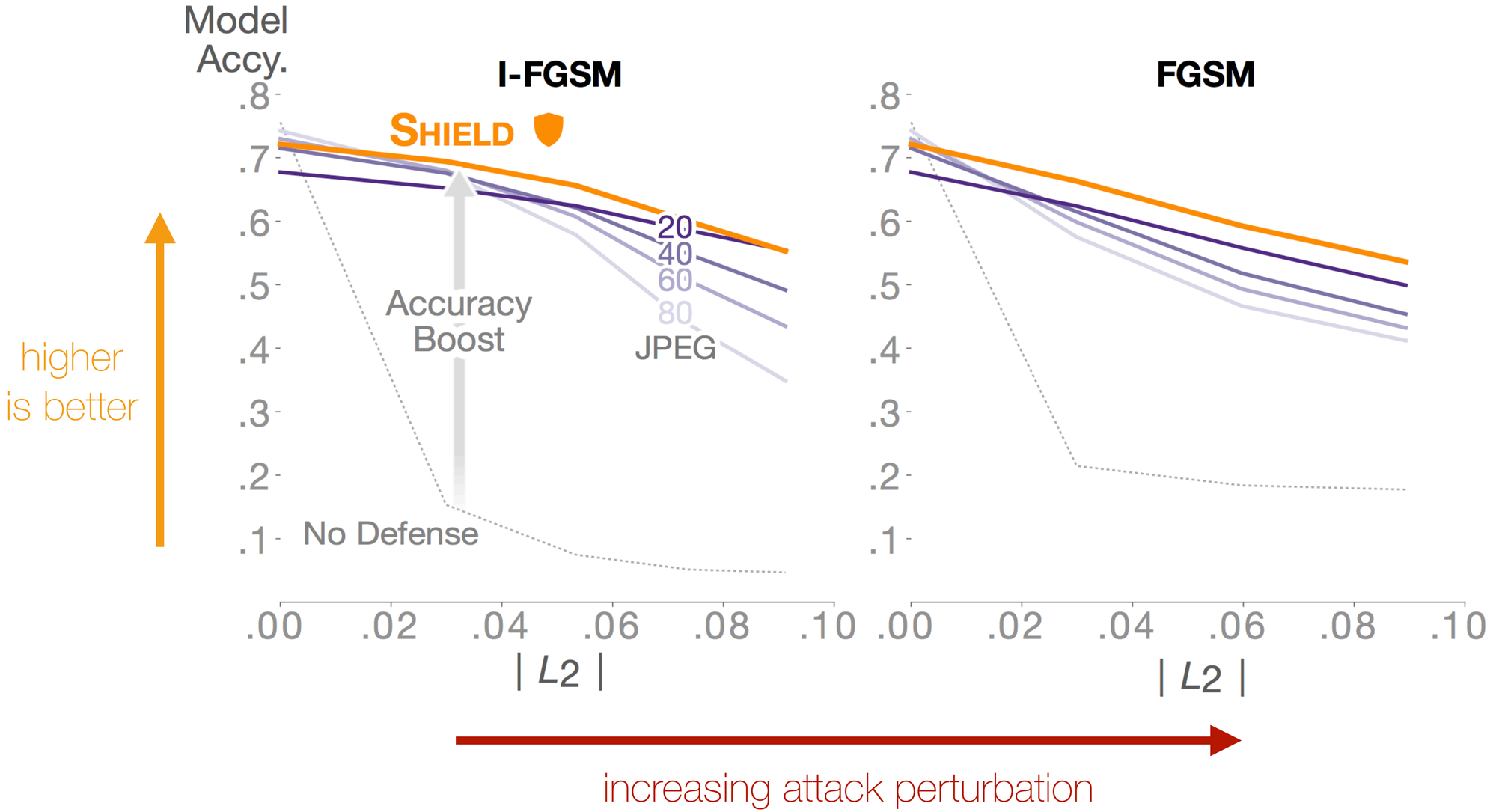

SHIELD
Secure Heterogeneous Image
Ensemble with Localized Denoising
"Chain Mail"
(Attacked)
Labrador
Retriever
Real-time
Compression
Preprocessing
Vaccinated
Deep Neural
Network Ensemble
Correctly
Classified
Correctly
Classified

SHIELD is multi-pronged approach that incorporates

- Stochastic Local Quantization
- Model Vaccination (re-training)
- Ensembling

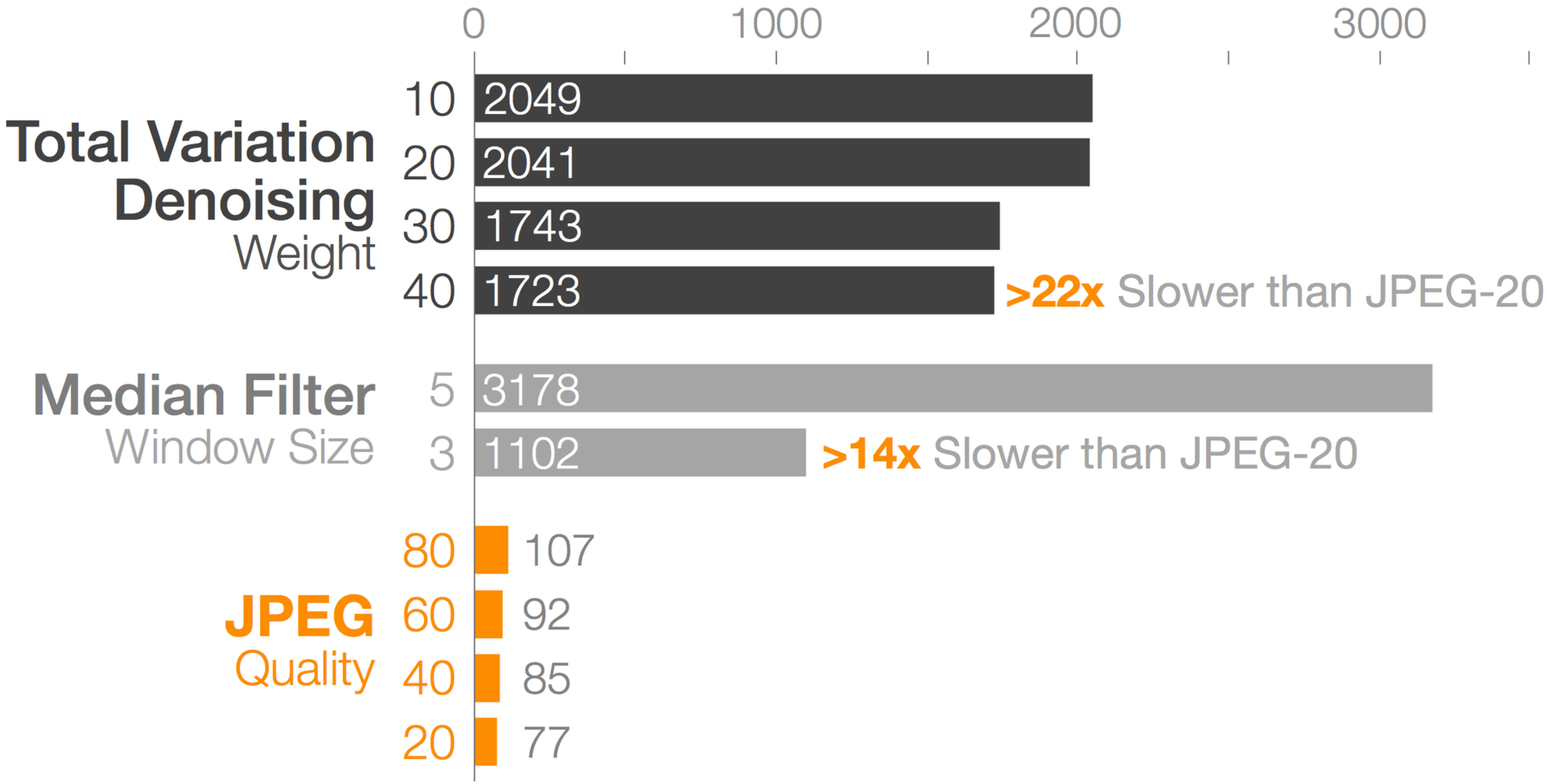
to mitigate adversarial attacks

Results with ResNet-50 v2 (on ImageNet validation set)



Defense Runtime Comparison

(in seconds; shorter is better)



tested on 50,000 images from the ImageNet validation set