Adversarial Attack

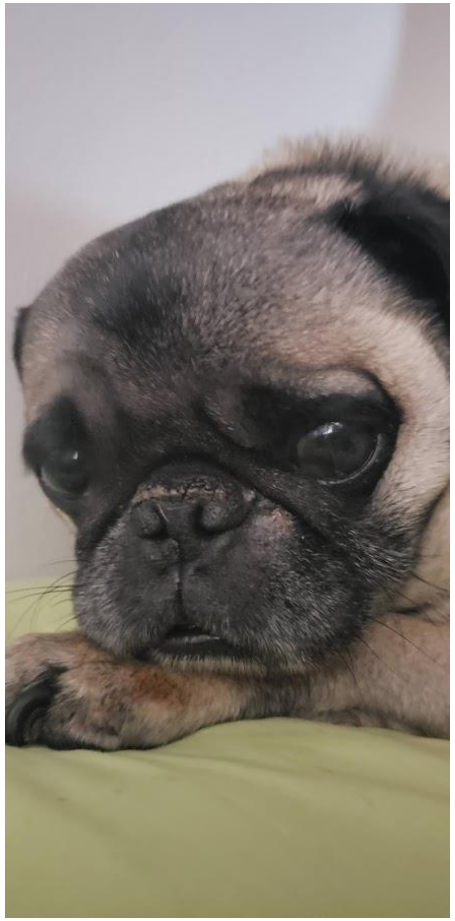
Chen-Kai Tsai

Devon Smart

1127240

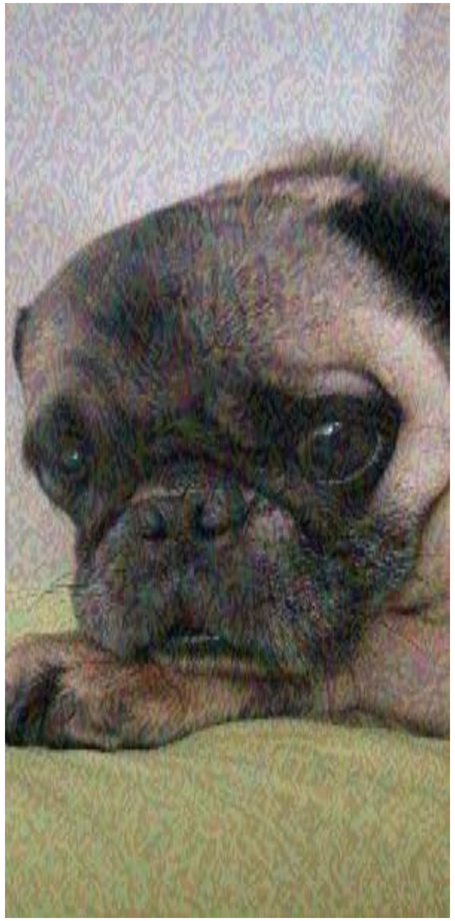
1063474

How to attack a classification model



pug

$$\rightarrow \underset{\text{benign}}{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ \vdots \end{bmatrix}} + \underset{\text{Noise}}{\begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta x_3 \\ \Delta x_4 \\ \Delta x_5 \\ \vdots \end{bmatrix}} \rightarrow$$



Something else

Fast Gradient Sign Method (FSGM)

Loss

$$adv\_x = x + \epsilon * \mathrm{sign}(\nabla_x J(\theta, x, y))$$

A vector of partial derivatives of x

Fast Gradient Sign Method (FSGM)

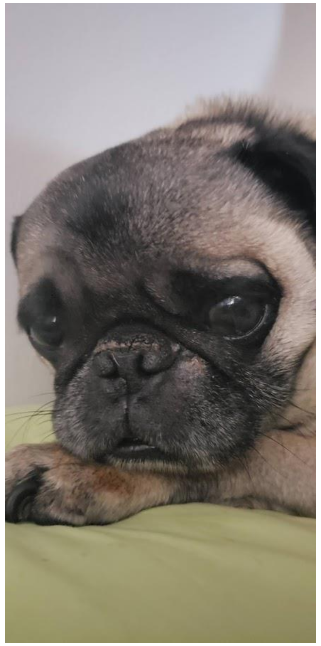
$$adv\_x = x + \epsilon * \underline{\text{sign}(\nabla_x J(\theta, x, y))}$$

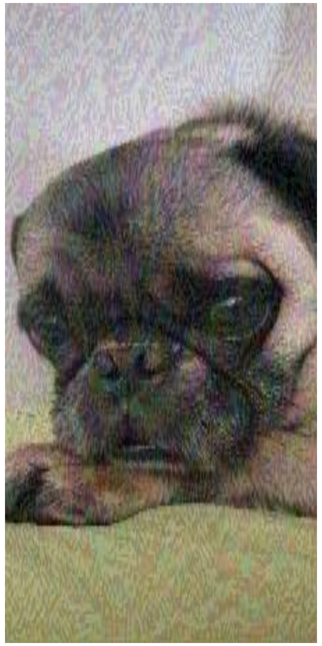
White-box attack
Need to know which model and model’s weight

ImageNet Dataset

- 14 million images
- ImageNet contains more than 20,000 categories
- The possibility of classes from these 20000 categories should add up to one.

On NASNetMobile

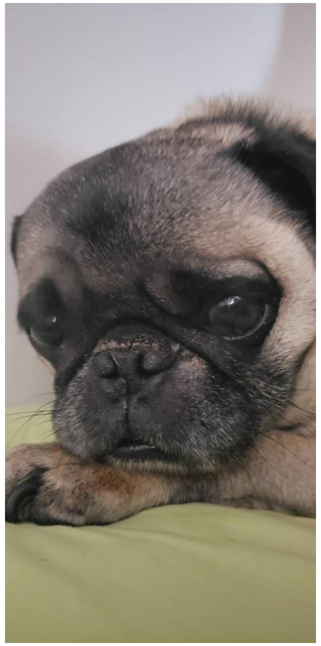


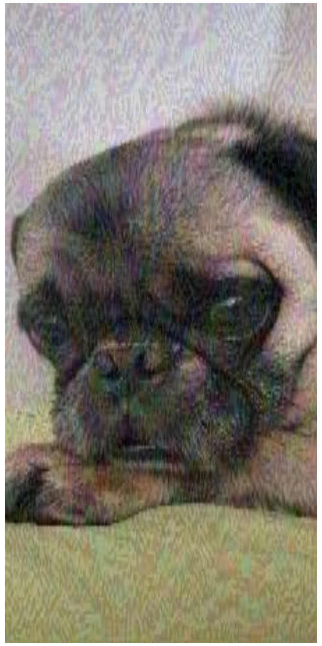


Epsilon 0 0.005 0.010 0.100

pug 80.227304

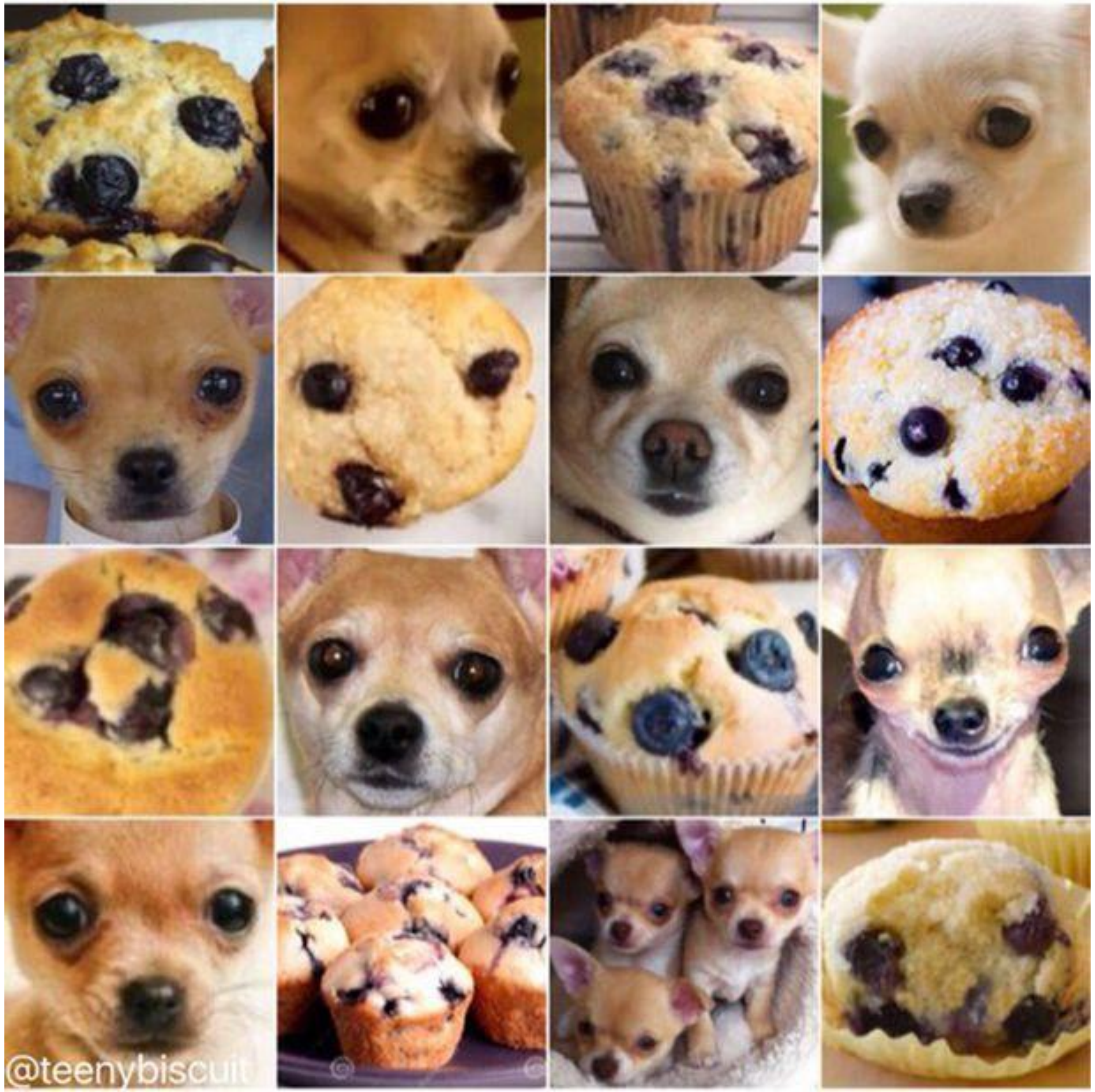
On NASNetMobile



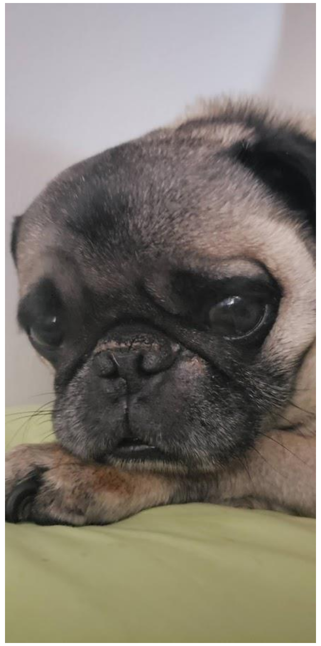


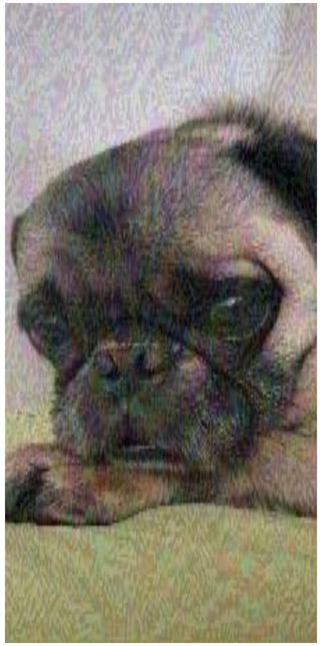
Epsilon	0	0.005	0.010	0.100
	pug 80.227304	Brabancon_griffon 43.02122		

Fail with dignity


@teenybiscuit

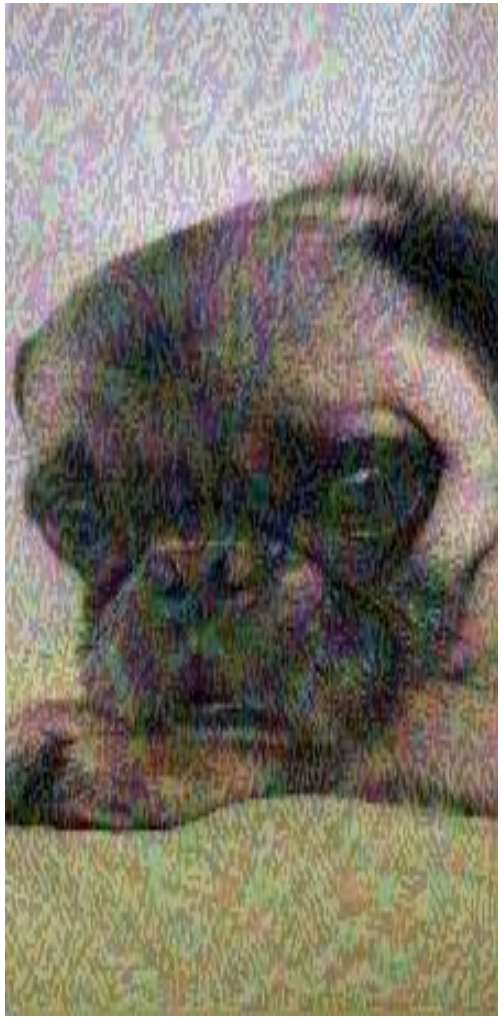
On NASNetMobile





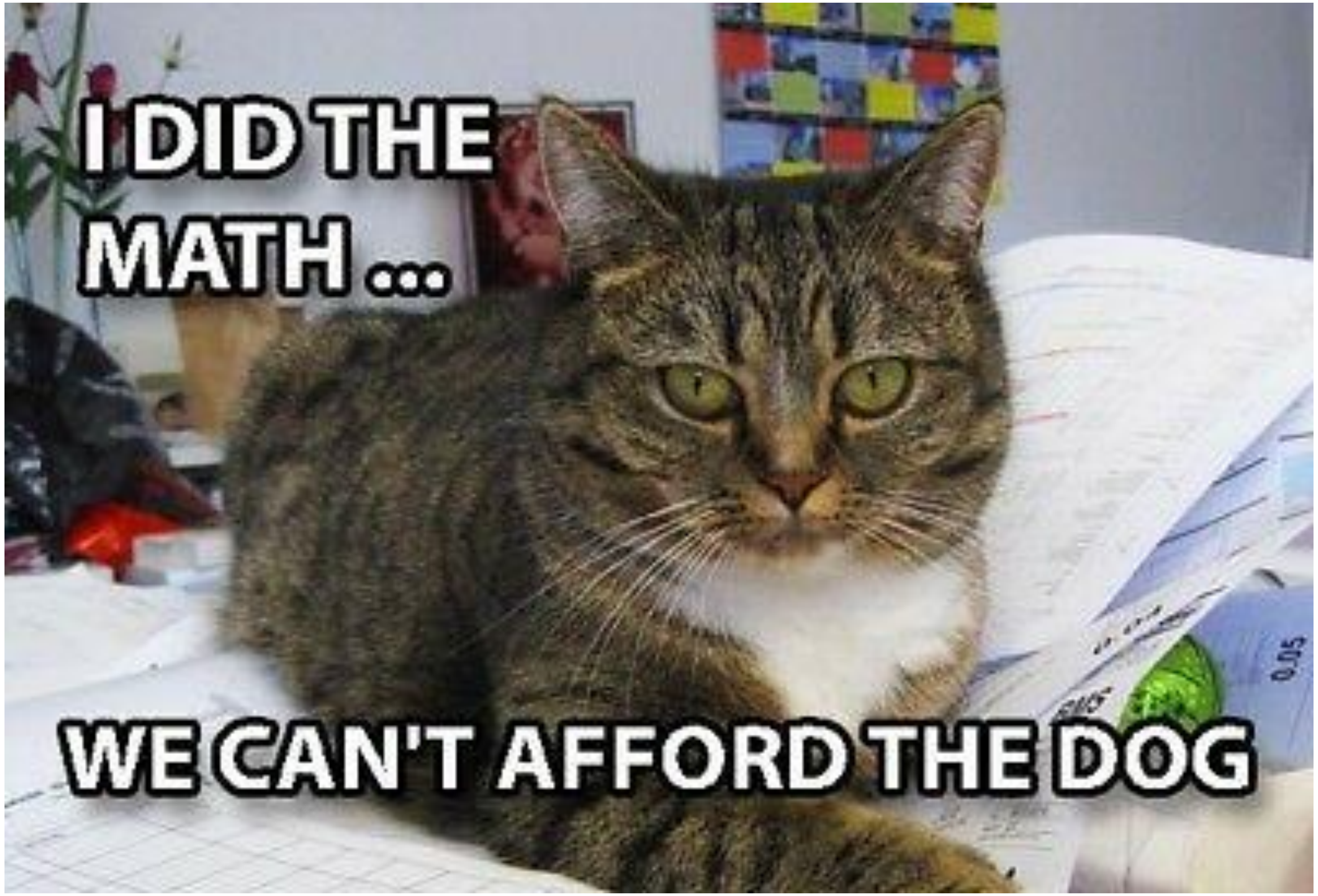
Epsilon	0	0.005	0.010	0.100
	pug 80.227304	Brabancon_griffon 43.02122	Persian_cat 47.073898	Persian_cat 91.48571

What if we have bigger epsilon?

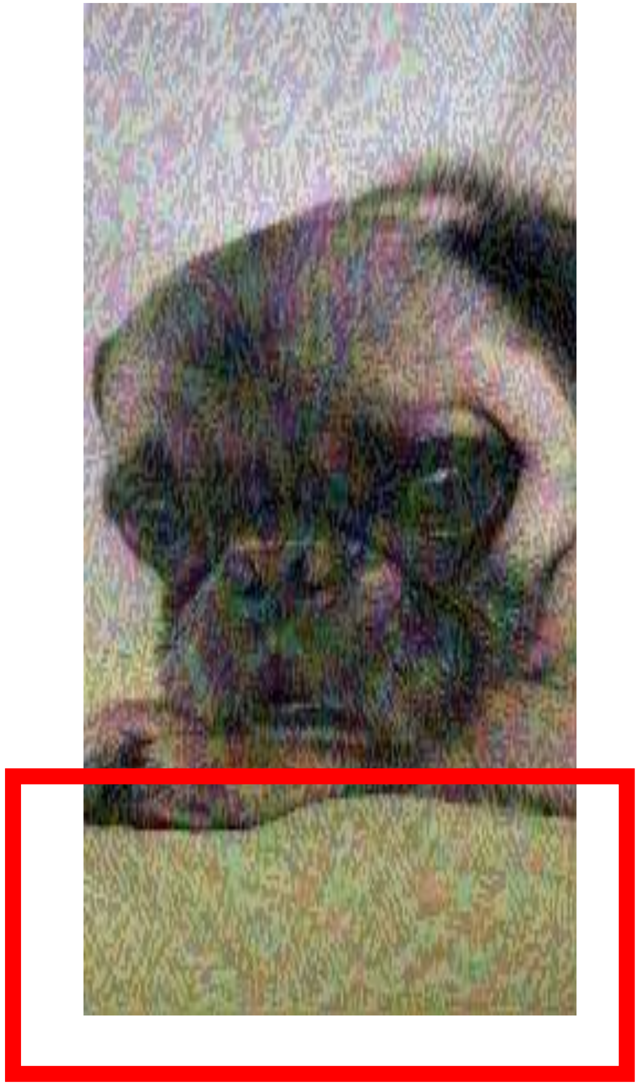


Epsilon = 0.150

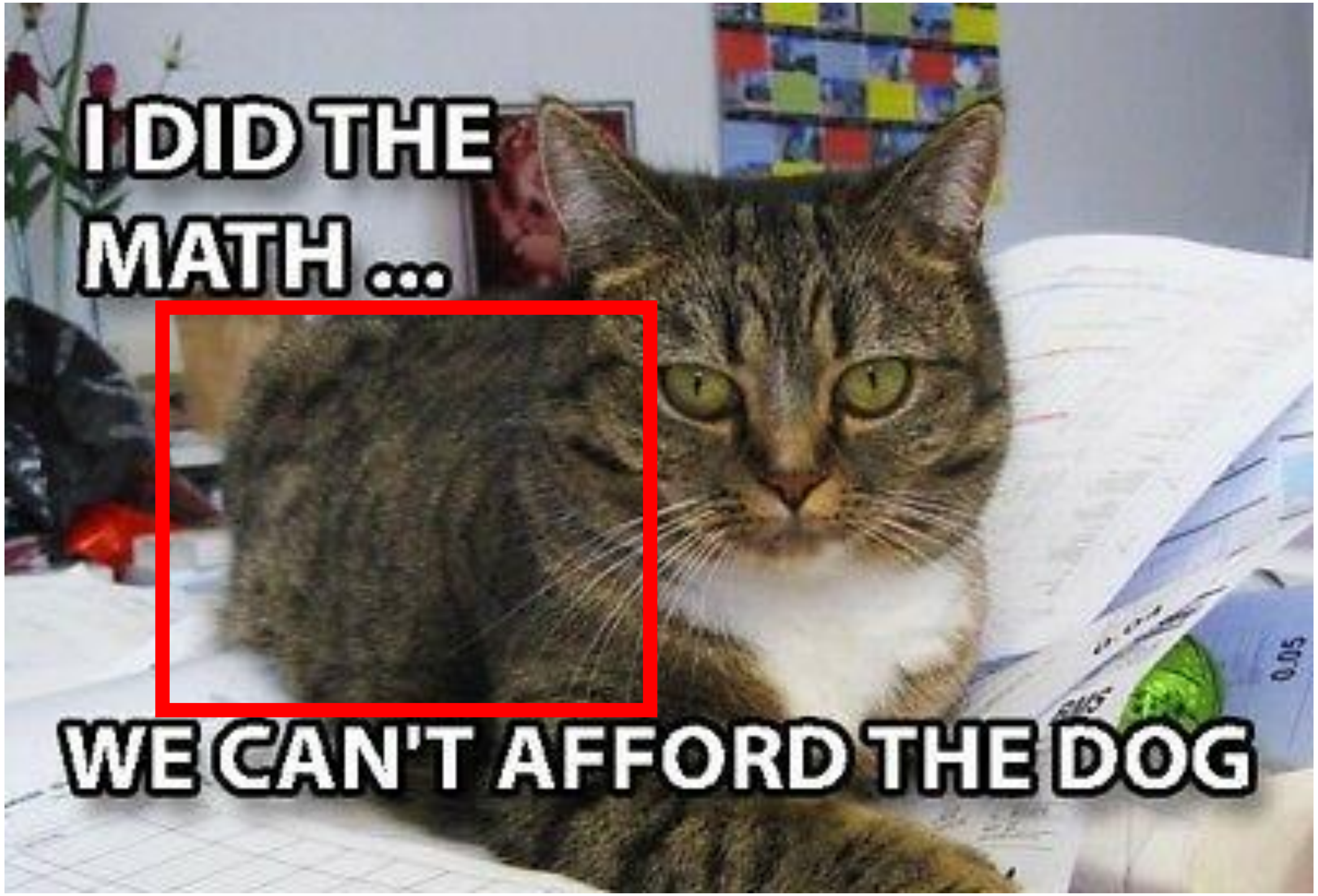
tabby 21.512717



Sorry, I cannot find a good picture contain both tabby cat and pug



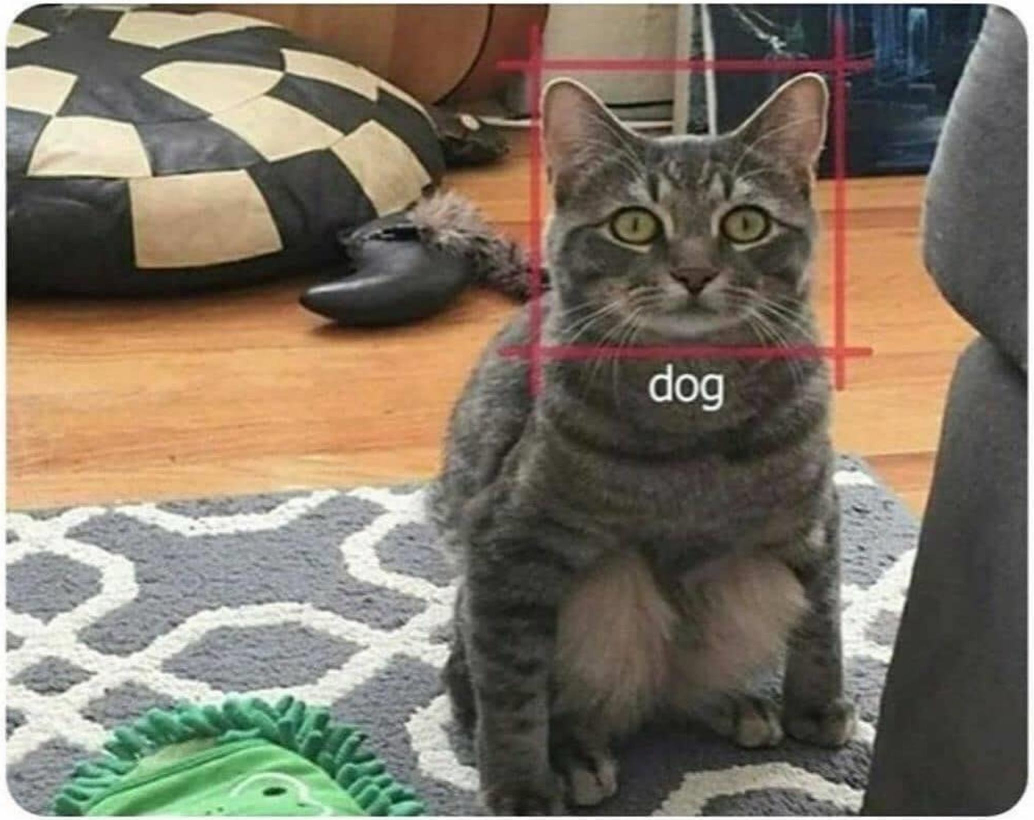
tabby 21.5



tabby 71.87

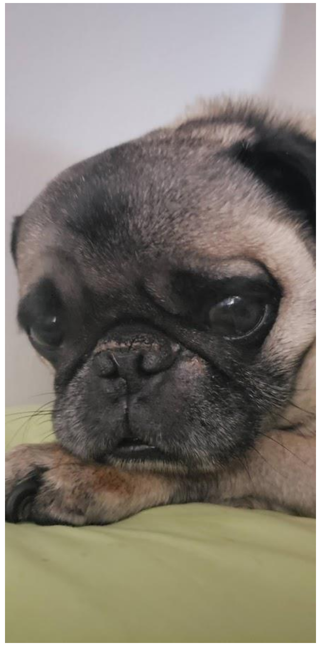
90's Media: AI WILL DESTROY THE WORLD IN A DECADE

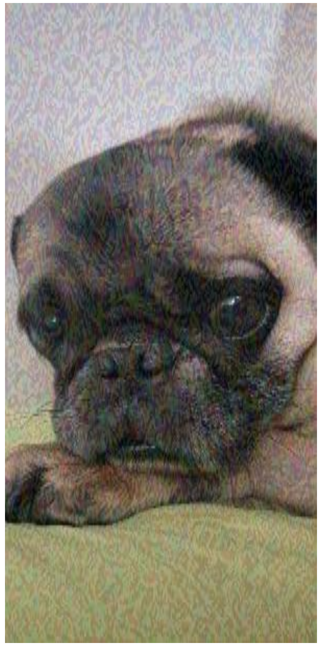
That AI today:



What about other models?

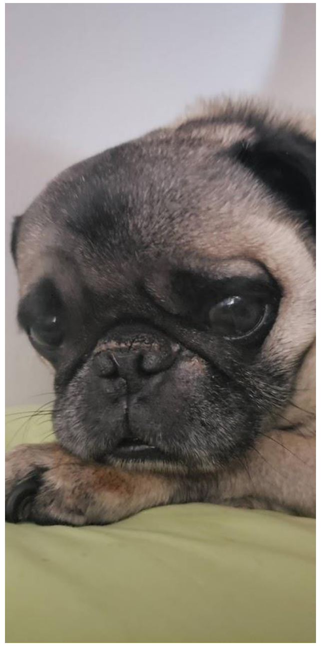
On InceptionV3

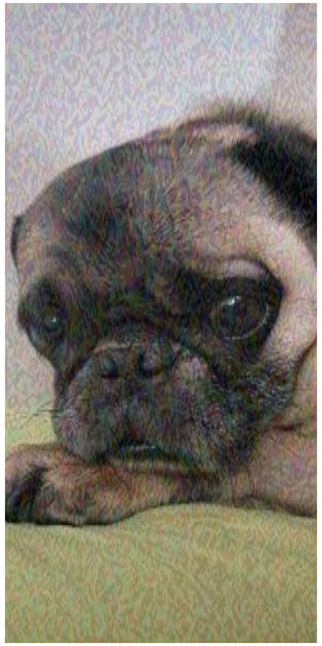




Epsilon	0	0.005	0.010	0.100
	pug 84.87882	pug 8.665805		

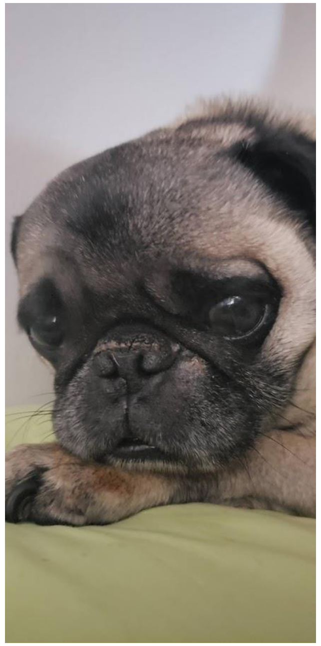
On InceptionV3

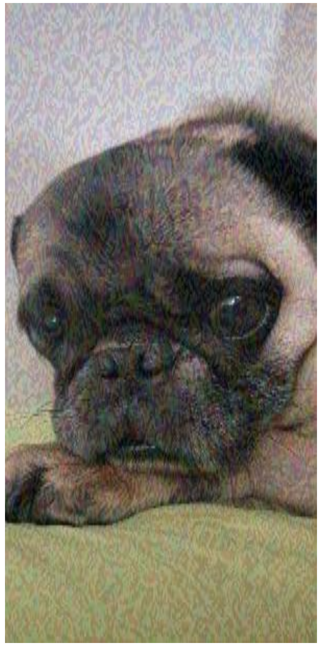




Epsilon	0	0.005	0.010	0.100
	pug 84.87882	pug 8.665805	pug 10.113112	

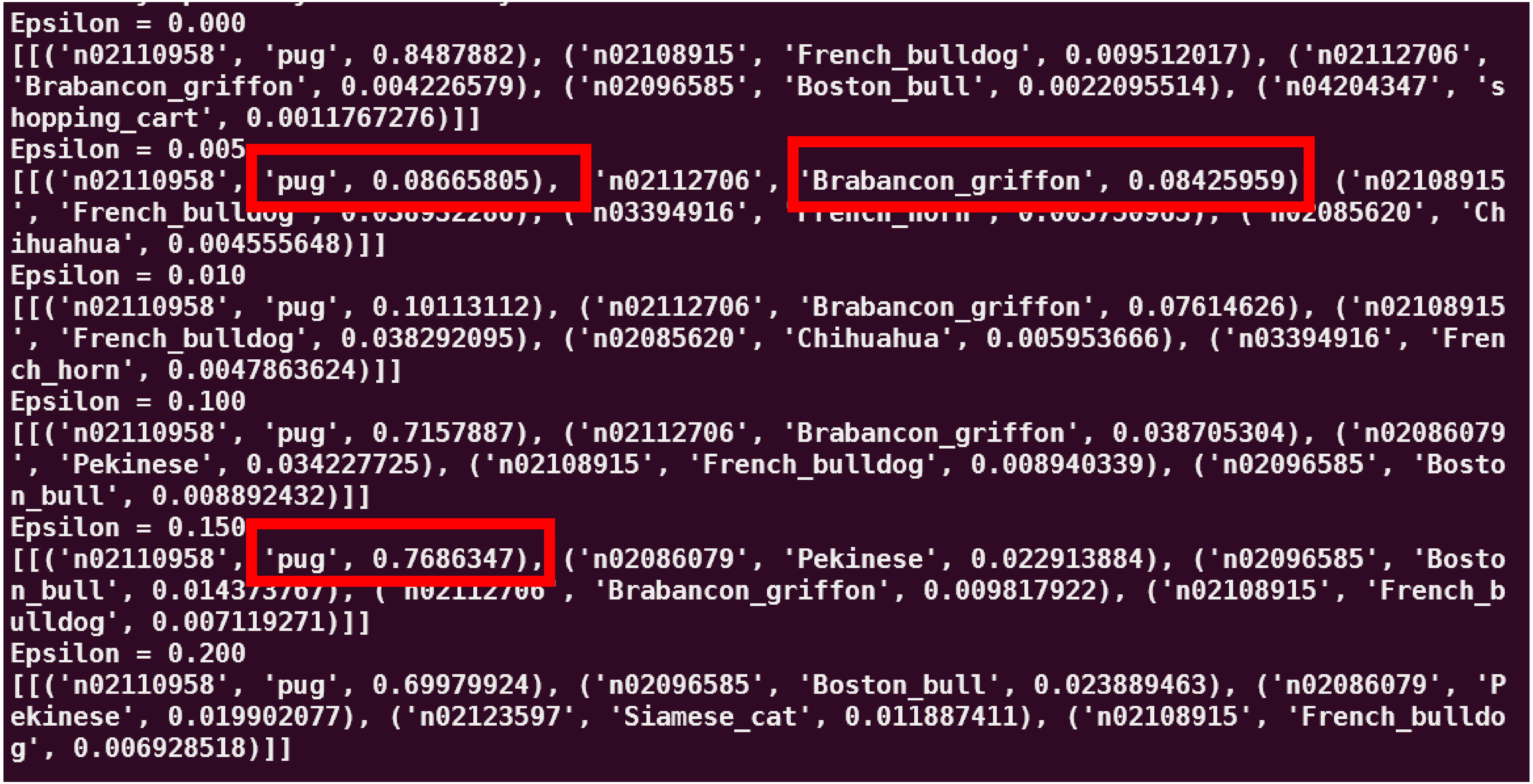
On InceptionV3





Epsilon	0	0.005	0.010	0.100
	pug 84.87882	pug 8.665805	pug 10.113112	pug 71.57887

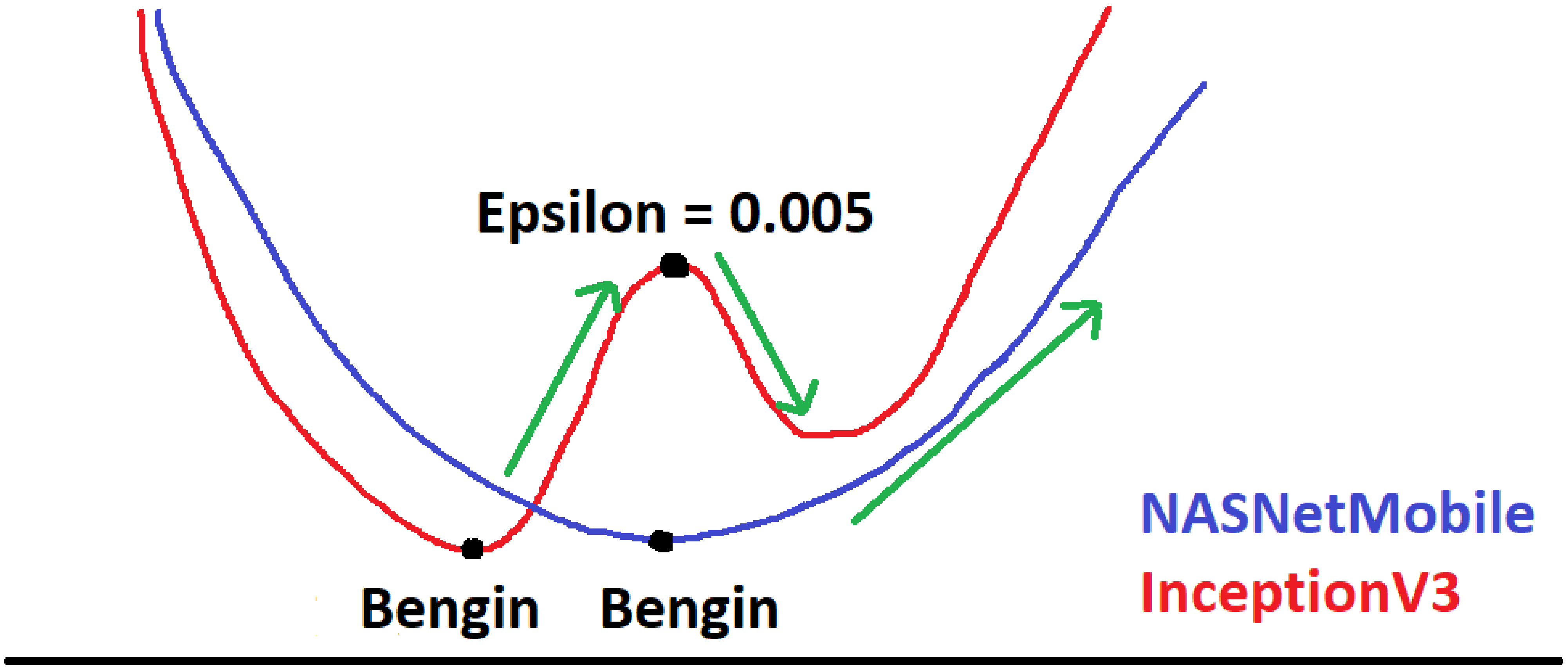
Top 5 labels for different Epsilon



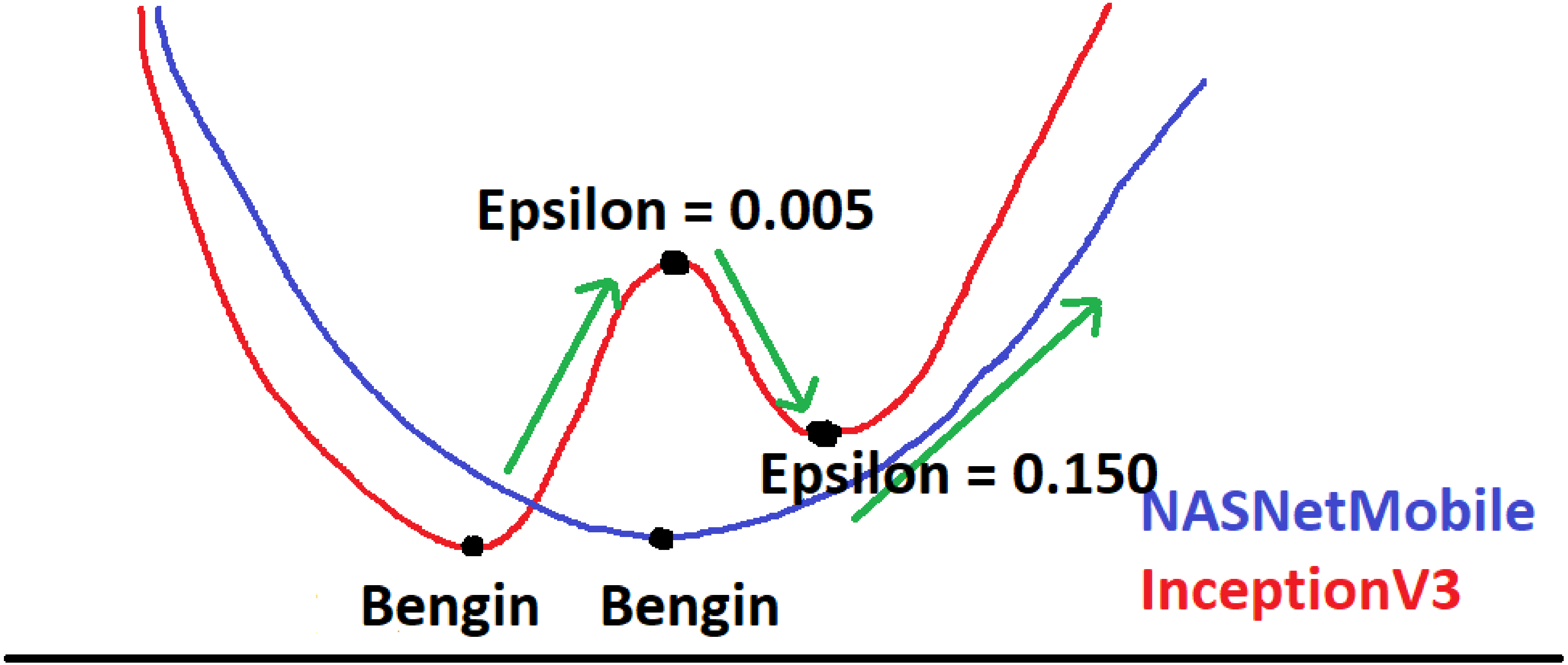
```
Epsilon = 0.000
[[('n02110958', 'pug', 0.8487882), ('n02108915', 'French_bulldog', 0.009512017), ('n02112706',
'Brabancon_griffon', 0.004226579), ('n02096585', 'Boston_bull', 0.0022095514), ('n04204347', 's
hopping_cart', 0.0011767276)]]
Epsilon = 0.005
[[('n02110958', 'pug', 0.08665805), ('n02112706', 'Brabancon_griffon', 0.08425959), ('n02108915
', 'French_bulldog', 0.038932286), ('n03394916', 'French_horn', 0.005750963), ('n02085620', 'Ch
ihuahua', 0.004555648)]]
Epsilon = 0.010
[[('n02110958', 'pug', 0.10113112), ('n02112706', 'Brabancon_griffon', 0.07614626), ('n02108915
', 'French_bulldog', 0.038292095), ('n02085620', 'Chihuahua', 0.005953666), ('n03394916', 'Fren
ch_horn', 0.0047863624)]]
Epsilon = 0.100
[[('n02110958', 'pug', 0.7157887), ('n02112706', 'Brabancon_griffon', 0.038705304), ('n02086079
', 'Pekinese', 0.034227725), ('n02108915', 'French_bulldog', 0.008940339), ('n02096585', 'Bosto
n_bull', 0.008892432)]]
Epsilon = 0.150
[[('n02110958', 'pug', 0.7686347), ('n02086079', 'Pekinese', 0.022913884), ('n02096585', 'Bosto
n_bull', 0.014373767), ('n02112706', 'Brabancon_griffon', 0.009817922), ('n02108915', 'French_b
ulldog', 0.007119271)]]
Epsilon = 0.200
[[('n02110958', 'pug', 0.69979924), ('n02096585', 'Boston_bull', 0.023889463), ('n02086079', 'P
ekinese', 0.019902077), ('n02123597', 'Siamese_cat', 0.011887411), ('n02108915', 'French_bulldo
g', 0.006928518)]]
```

Why?

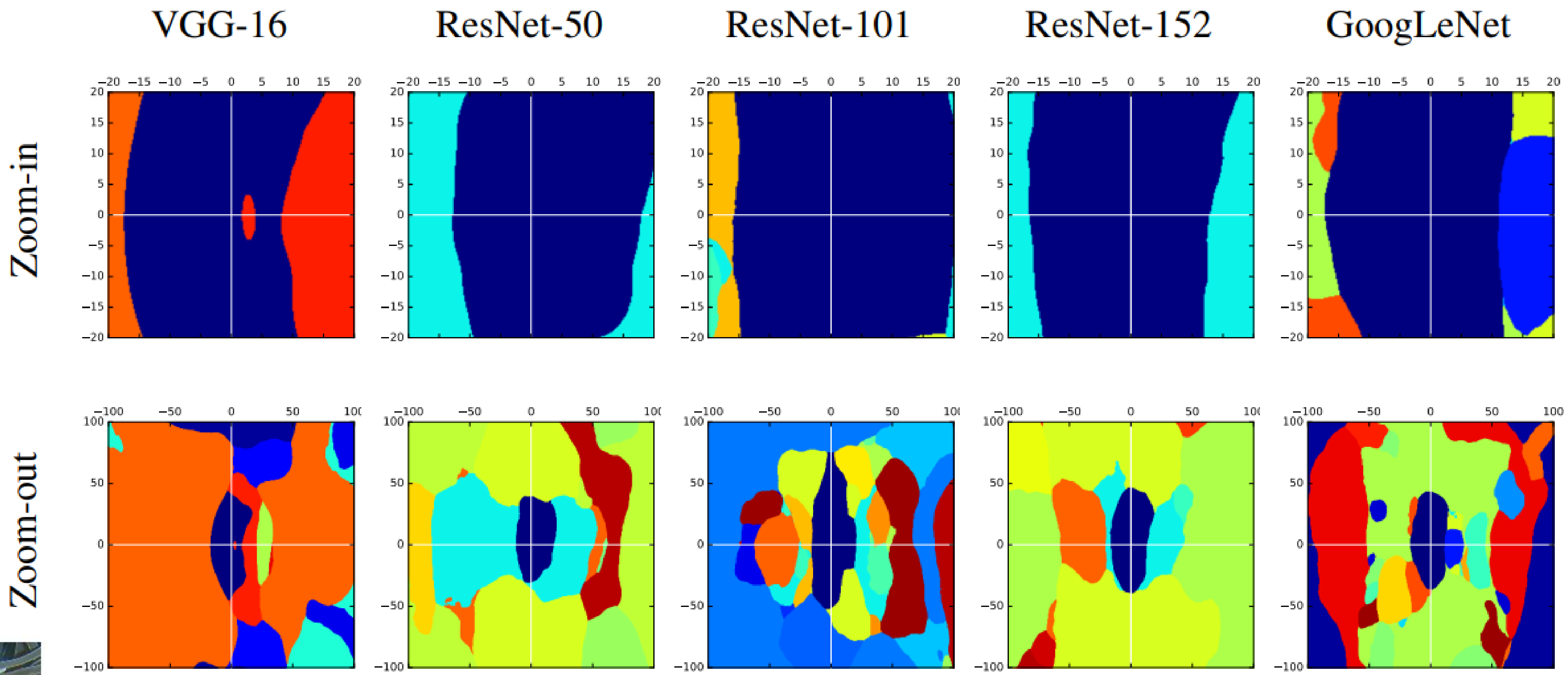
Project the high dimension loss function to the 2D graph



Project the high dimension loss function to the 2D graph



Is Black-box attack possible ? Yes! & Why?



About dataset

https://arxiv.org/pdf/1611.02770.pdf

Thank you