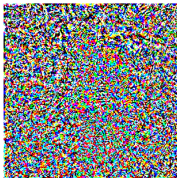


A small
perturbation
to one
training
example:

Label: Fish



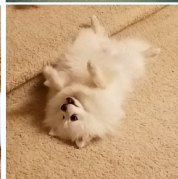
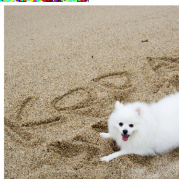
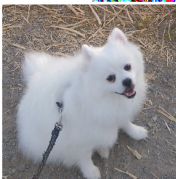
+ ϵ ·



Label: Fish



Can change
multiple **test**
predictions:



Orig (confidence): Dog (97%)
New (confidence): Fish (97%)

Dog (98%)
Fish (93%)

Dog (98%)
Fish (87%)

Dog (99%)
Fish (63%)

Dog (98%)
Fish (52%)

Figure 5. Training-set attacks. We targeted a set of 30 test images featuring the first author’s dog in a variety of poses and backgrounds. By maximizing the average loss over these 30 images, we created a visually-imperceptible change to the particular training image (shown on top) that flipped predictions on 16 test images.