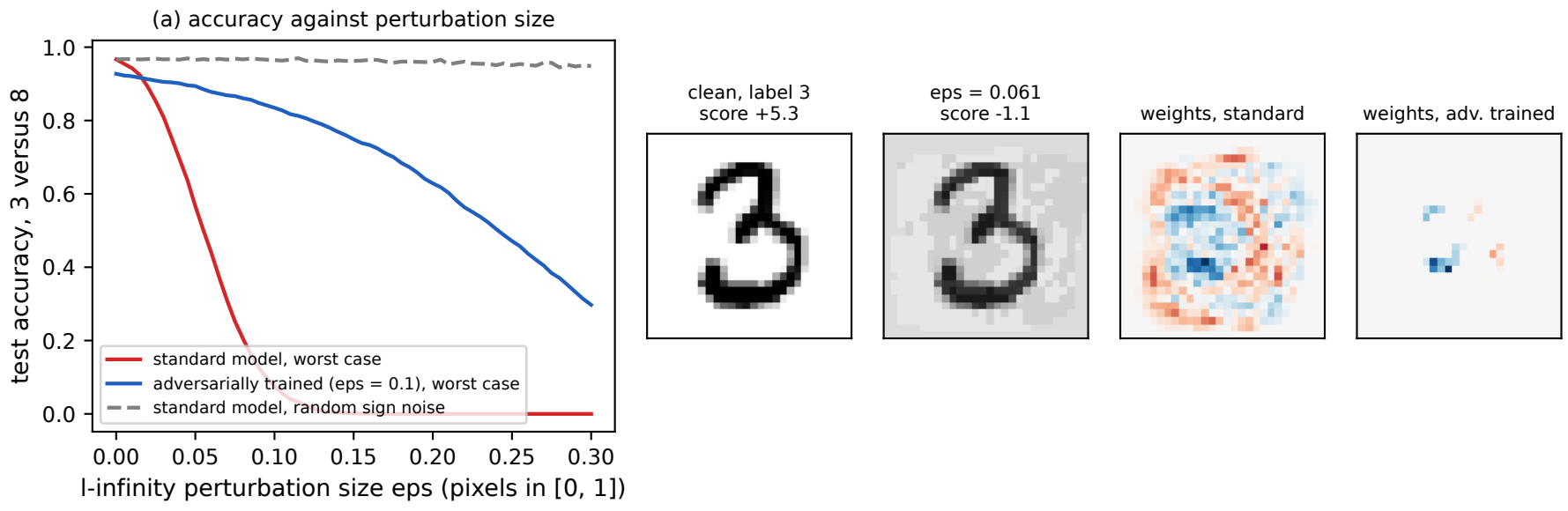

(a) accuracy against perturbation size
test accuracy, 3 versus 8
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
l-infinity perturbation size eps (pixels in [0, 1])
standard model, worst case
adversarially trained (eps = 0.1), worst case
standard model, random sign noise
clean, label 3
score +5.3
eps = 0.061
score -1.1
weights, standard
weights, adv. trained