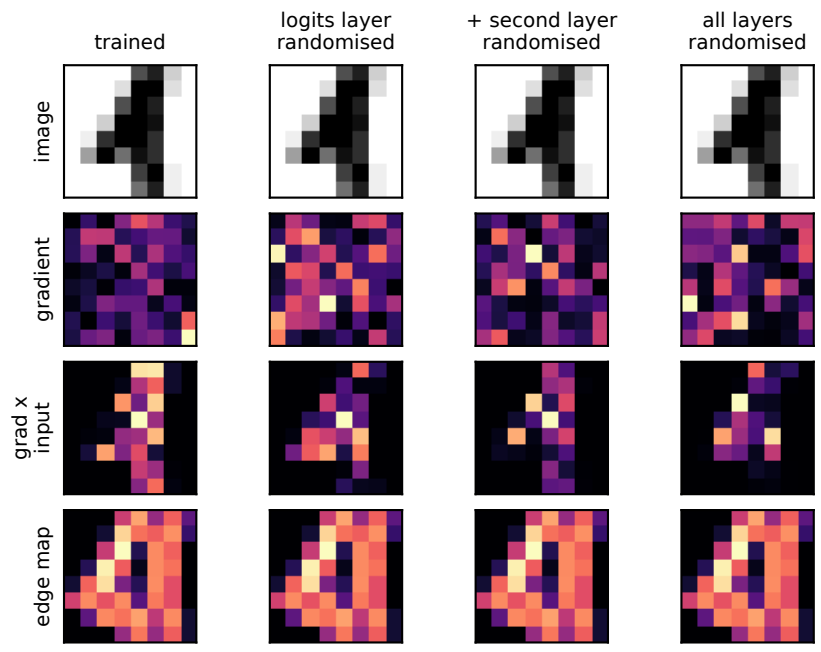
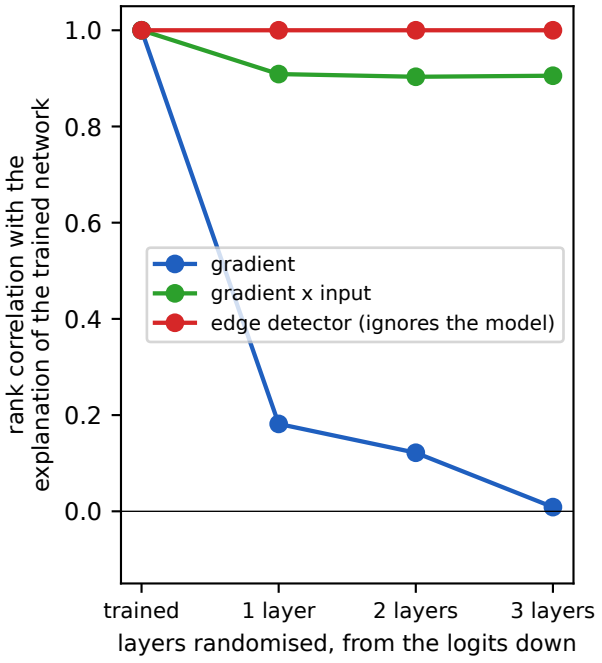

trained
logits layer randomised
+ second layer randomised
all layers randomised
image
gradient
grad x input
edge map


rank correlation with the explanation of the trained network
1.0
0.8
0.6
0.4
0.2
0.0
gradient
gradient x input
edge detector (ignores the model)
trained
1 layer
2 layers
3 layers
layers randomised, from the logits down