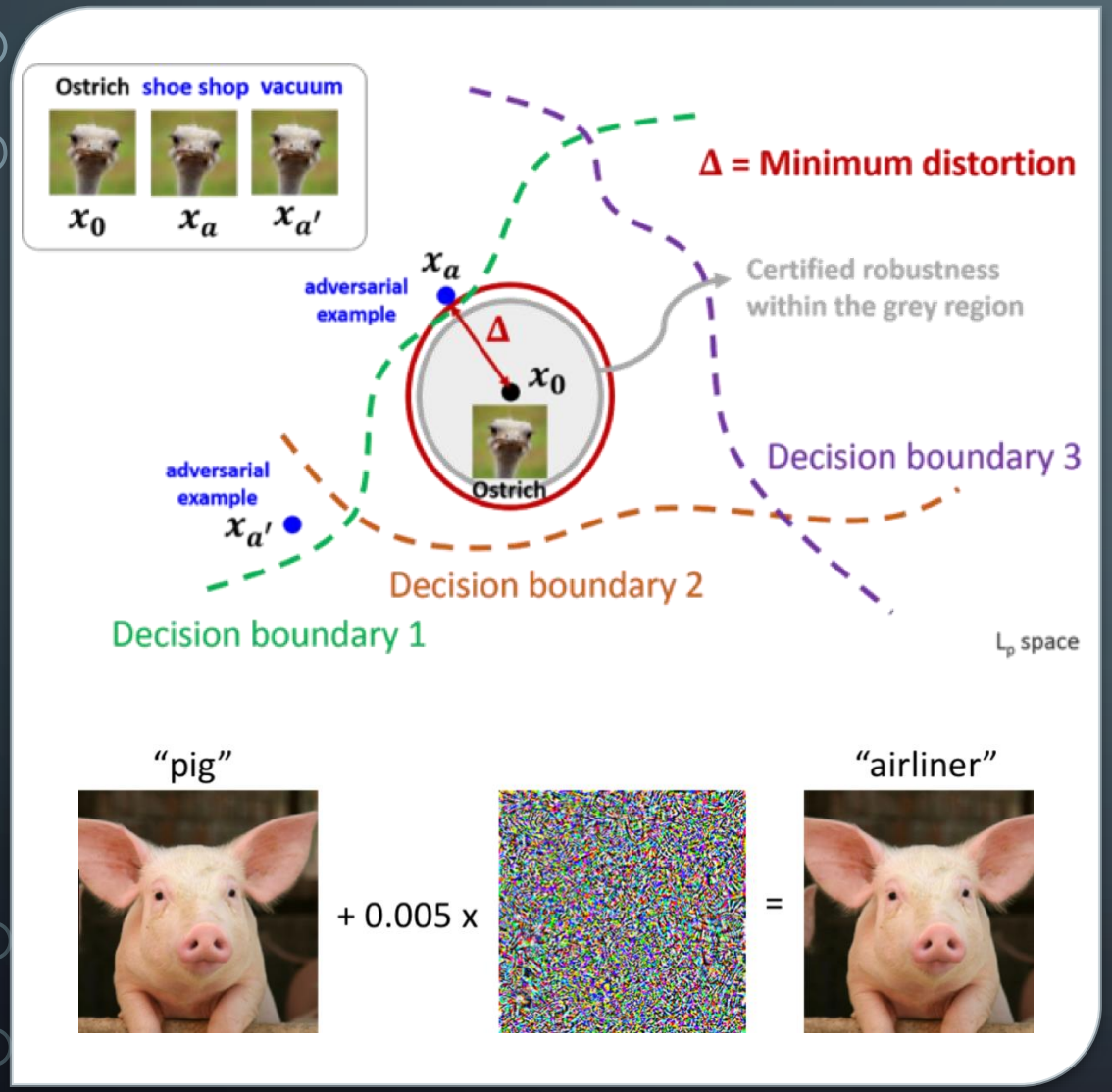
ADVERSARIAL ATTACK: INTERPRETING ADVERSARIAL EXAMPLES VIA CAM

107062548 蔡昀芸
107061518 陳永慶
105060019 楊承諭

OUTLINE

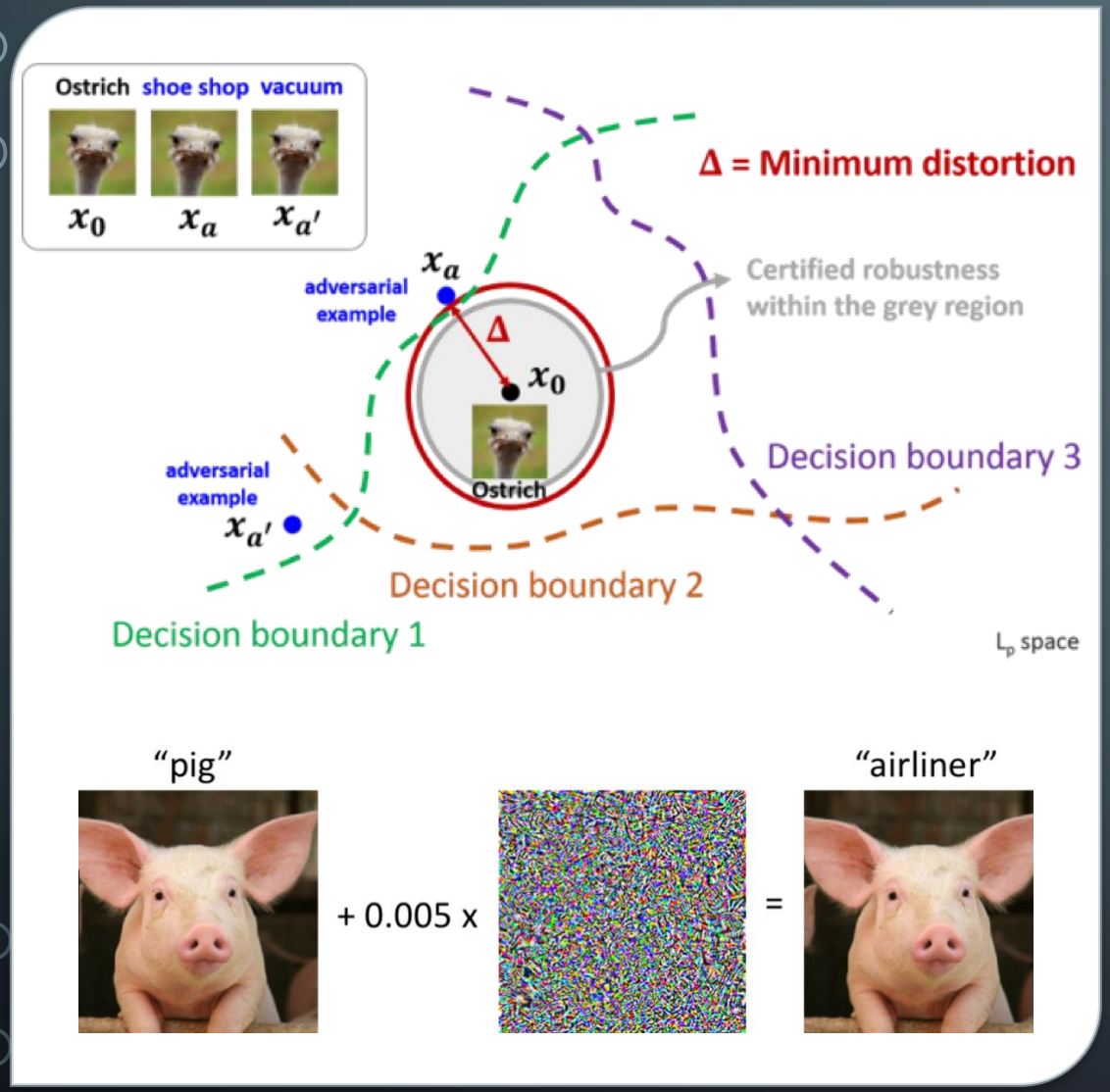
- INTRODUCTION
- METHOD
- EXPERIMENTAL RESULTS
- CONCLUSION

INTRODUCTION -- ADVERSARIAL ATTACK



- This involves carefully crafted perturbations called adversarial examples that, when added to natural examples, lead deep neural network models to misbehave.

INTRODUCTION -- ADVERSARIAL ATTACK



- Attack lower bound: The least amount of perturbation to a natural example required in order to deceive a classifier.
- Can be viewed as local Lipschitz constant estimation problem.

INTRODUCTION -- TYPES OF ATTACKS

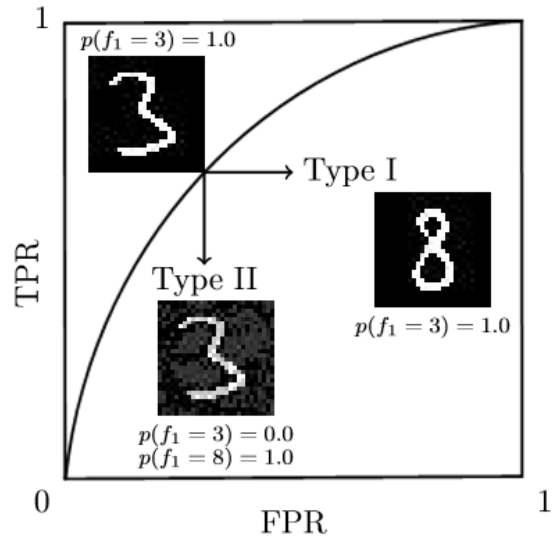


Fig. 1. Relationship between Type I and Type II adversarial attacks on ROC curve of f_1. Through viewing number "3" as true sample and others as false samples, Type II attack aims to decrease the true positive rate (TPR), while Type I attack tries to increase the false positive rate (FPR)

- **Type I** : Generate an adversarial example that is different to the original one in the view of the attacker.

$$From\ x\ Generate\ x' = A(x)$$
$$s.t.\ f_1(x') = f_1(x'),$$
$$d\big(g_2(x), g_2(x')\big) \gg \varepsilon$$

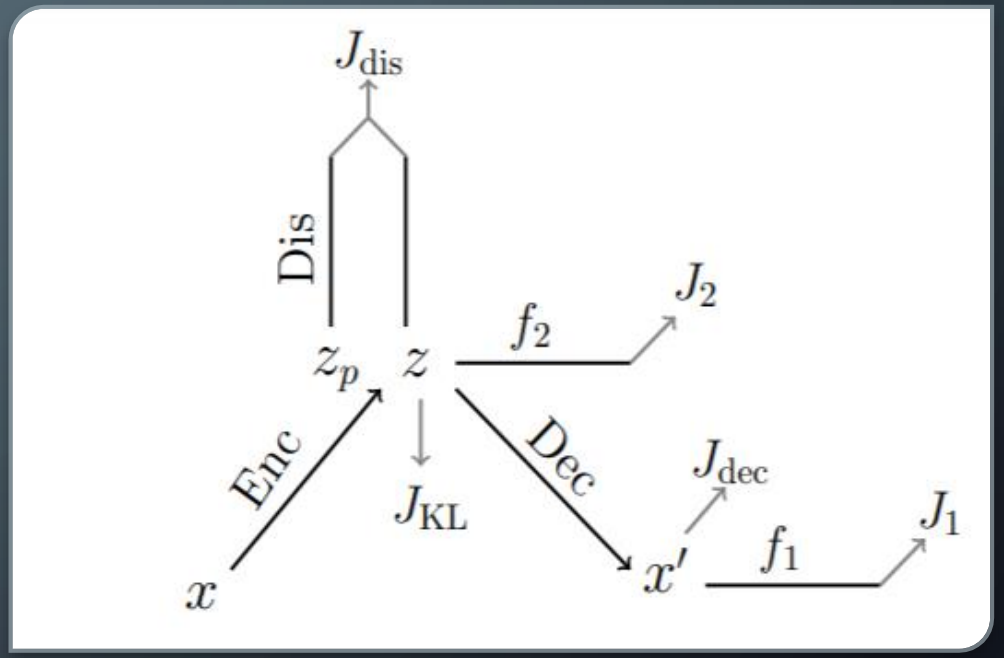
- **Type II** : Generating false negatives examples

$$From\ x\ Generate\ x' = A(x)$$
$$s.t.\ f_1(x') \neq f_1(x'),$$
$$d(g_2(x), g_2(x')) \leq \varepsilon$$

METHOD FOR TYPE I ATTACK

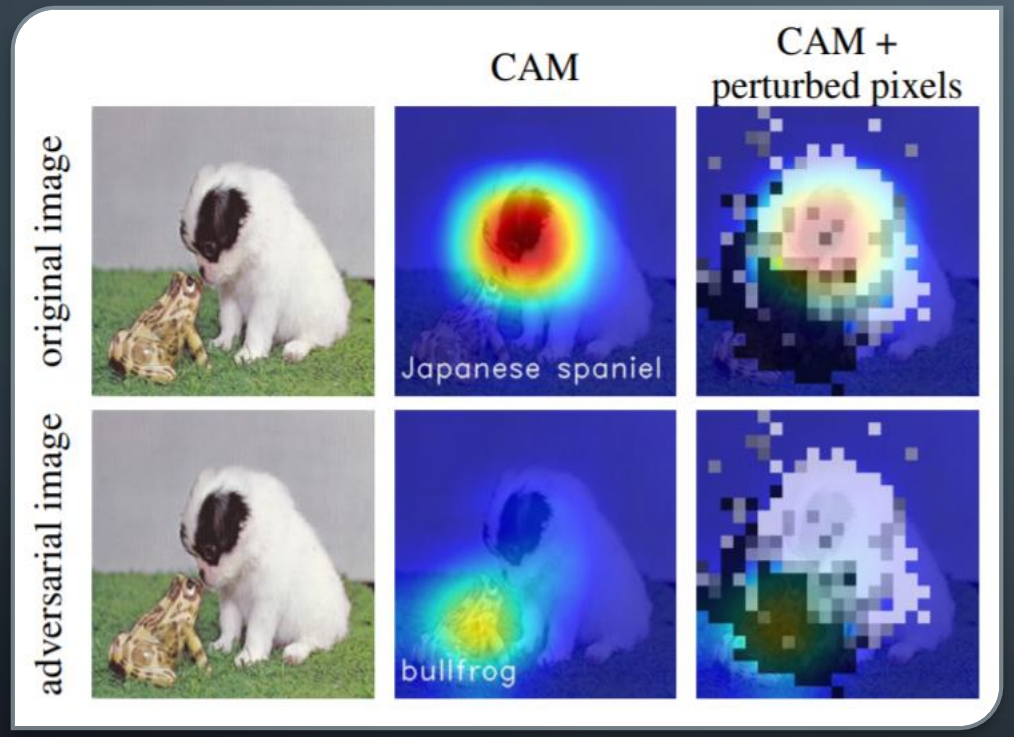
- **Supervised Variational Auto Encoder**

- For attacking, the gradients from f1 propagate to the latent variables z through the decoder.
- Discriminator: preventing the latent variables from lying outside the manifold in the latent space while attacking.



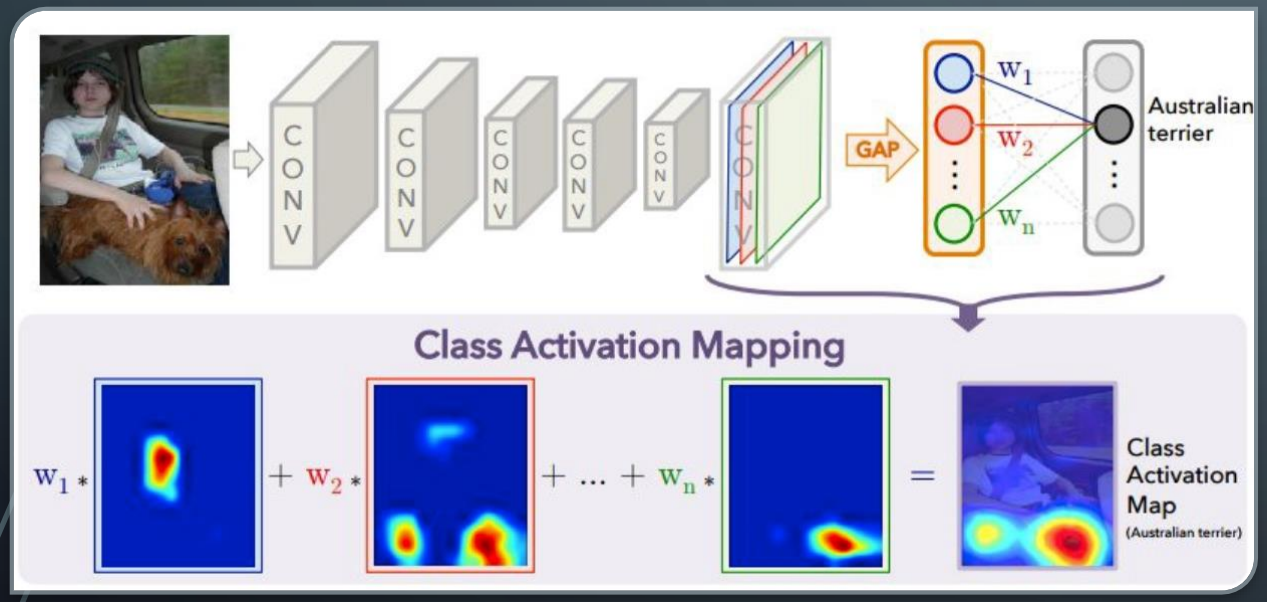
$$
\begin{aligned}
J &= -\mathrm{KL}[q(z|x)||p(z)] + E_{z\sim q(z|x)}[\log(p(y|z))] \\
&\quad + E_{z\sim q(z|x)}[\log(p(x|z))] \qquad (6) \\
&\triangleq -(J_{\mathrm{KL}} + J_2 + J_{dec}),
\end{aligned}
$$

CLASSIFICATION ACTIVATION MAP (CAM)



- Using a **global average pooling (GAP)** layer at the end of neural networks instead of a fully-connected layer resulted in excellent localization, which gives us an idea about where neural networks pay attention.

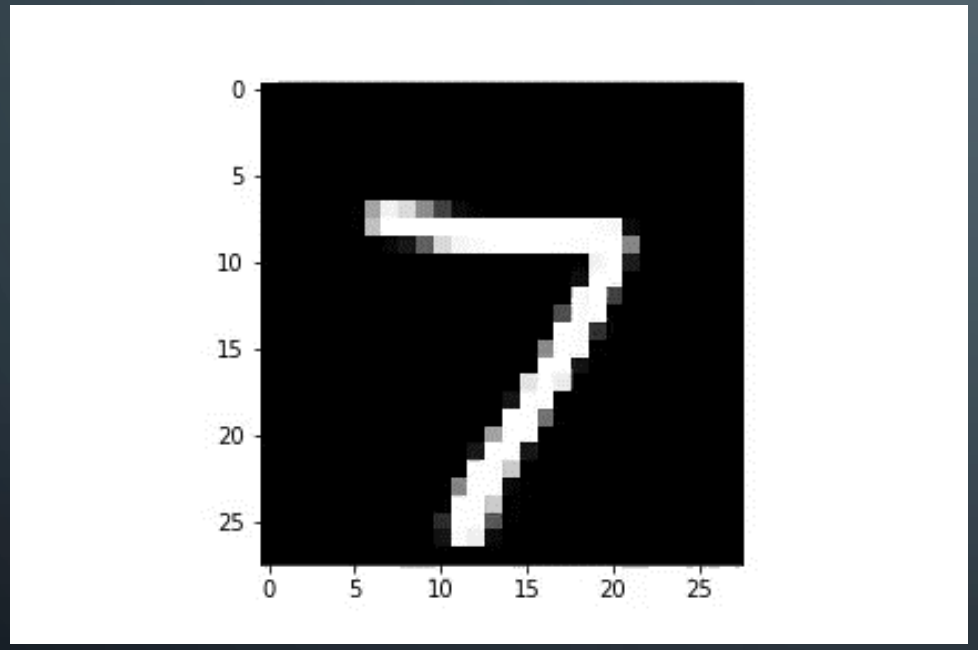
CLASSIFICATION ACTIVATION MAP (CAM)



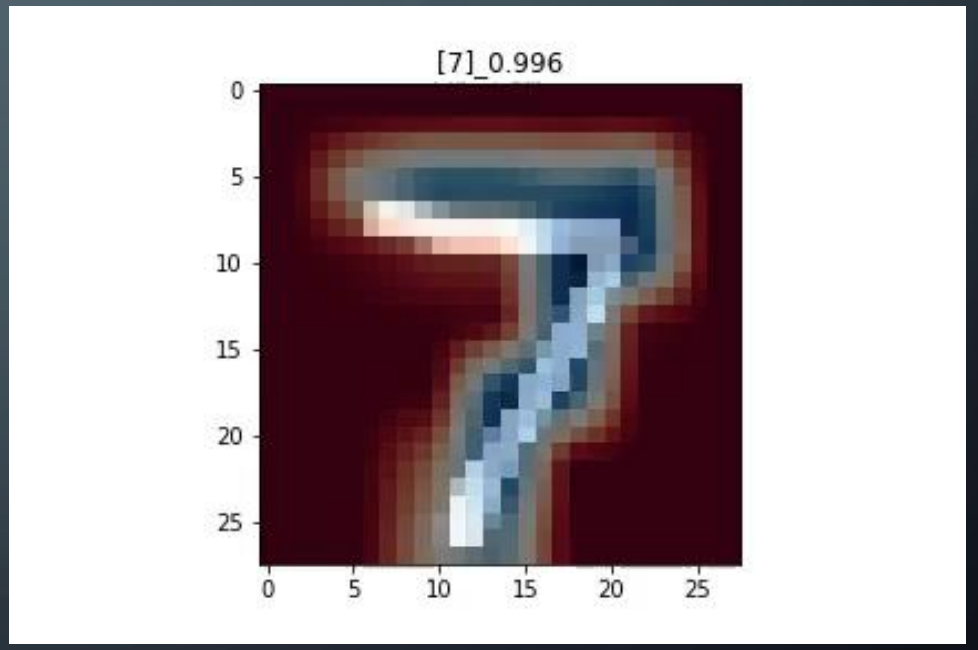
- Get all the weights connected between the fully-connected layer and the softmax class for which we want to predict.
- Take the **feature maps** that are about to be passed through GAP layer.

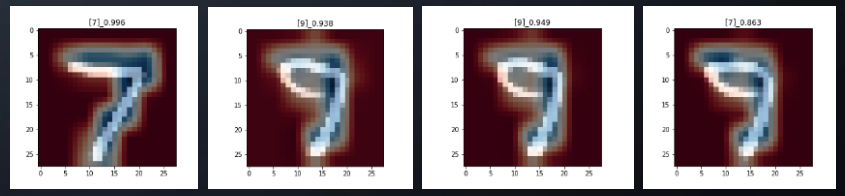
EXPERIMENTAL RESULTS

- Transition attack from digit 7 to 9



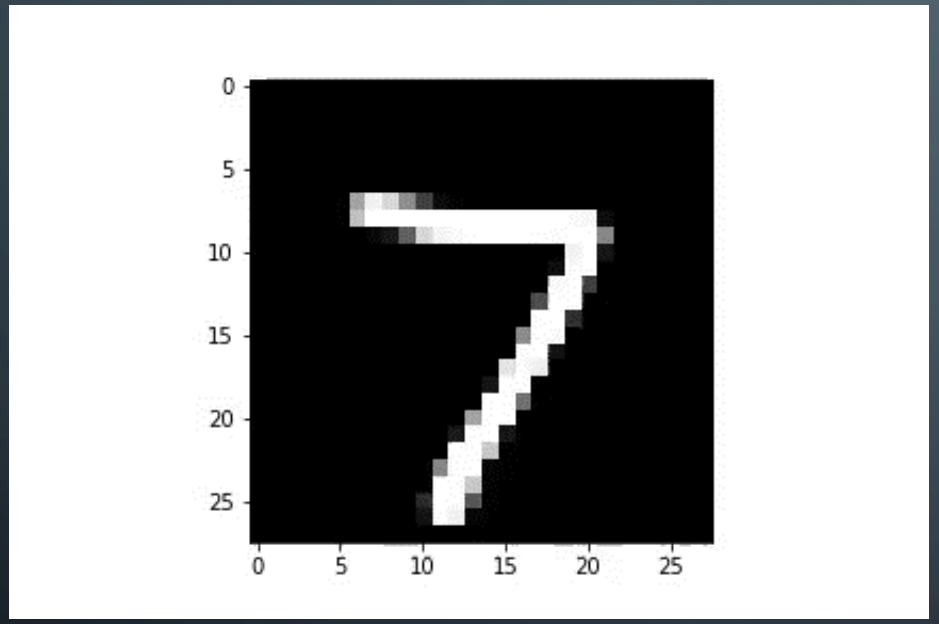
- Transition attack with CAM from digit 7 to 9

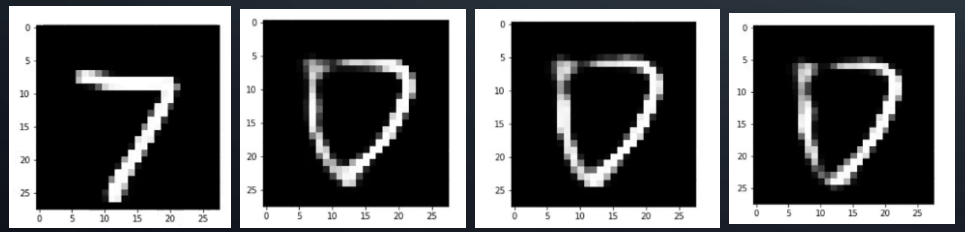




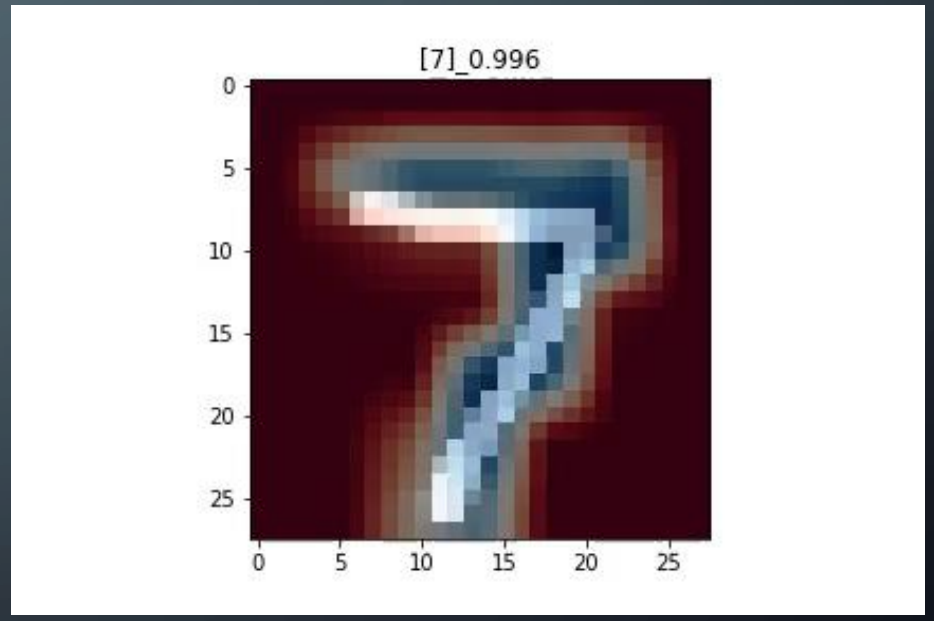
EXPERIMENTAL RESULTS

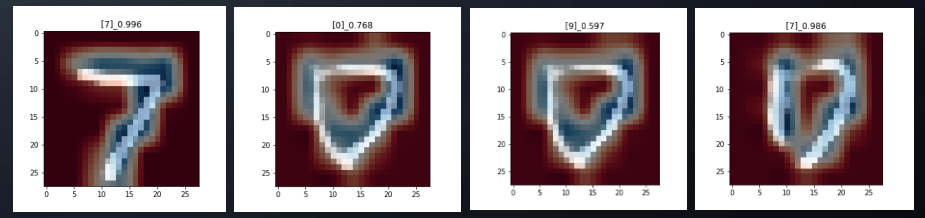
- Transition attack from digit 7 to 0





- Transition attack with CAM from digit 7 to 0





CONCLUSION

- We discovered that CAM was not reliable, it could easily fool by adversarial attack.
- What the neural network have seen couldn't represent what classification decisions it made.
- We would like to figure out other more robust interpretations.

THANK YOU