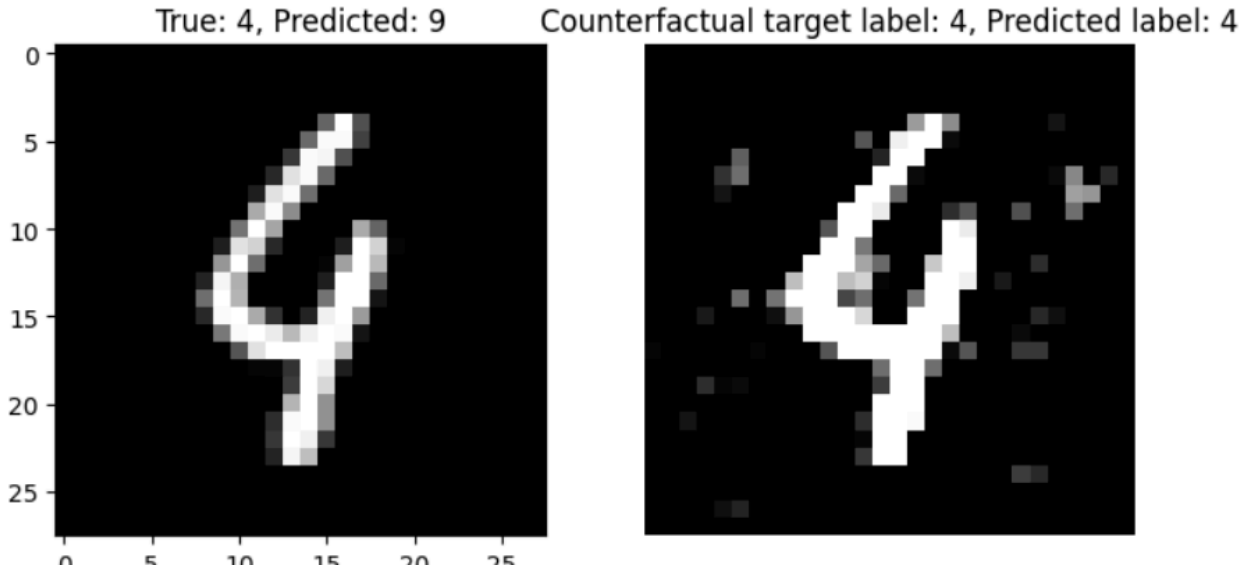
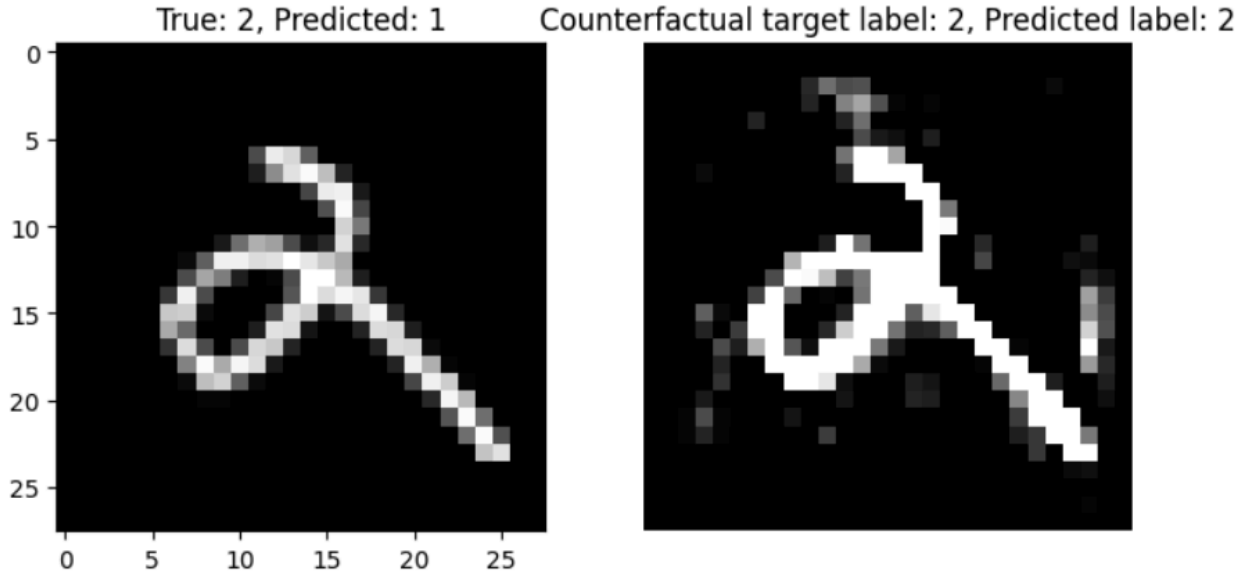
ASSIGNMENT 2
Akanksha Singal 2021008

QUESTION 1:
i) Wachter Method

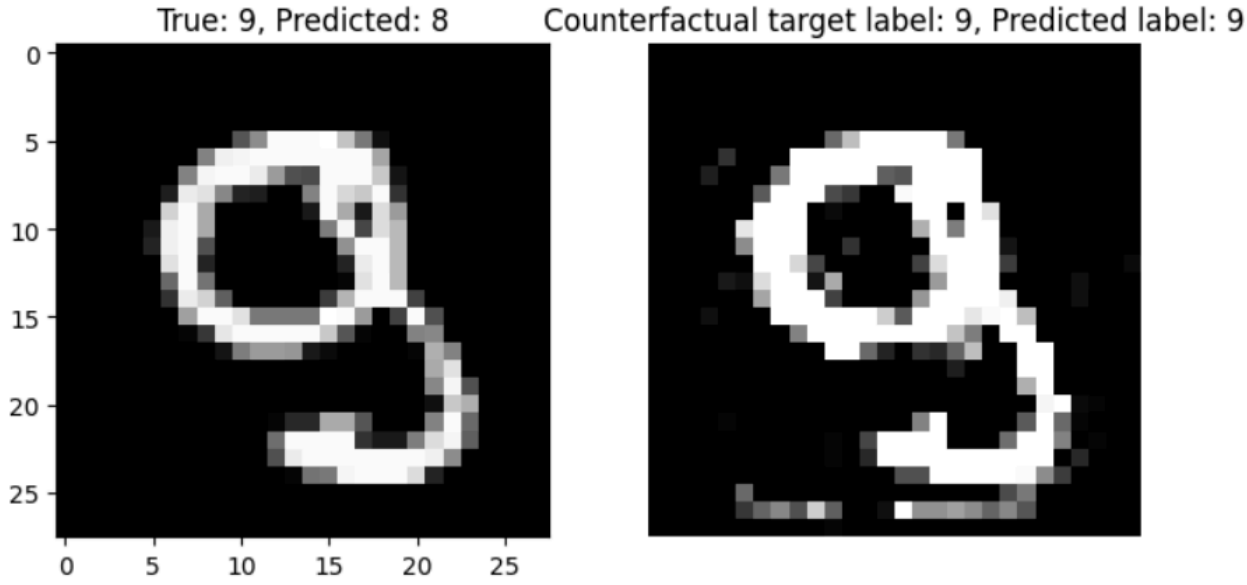
USING L2 NORM:



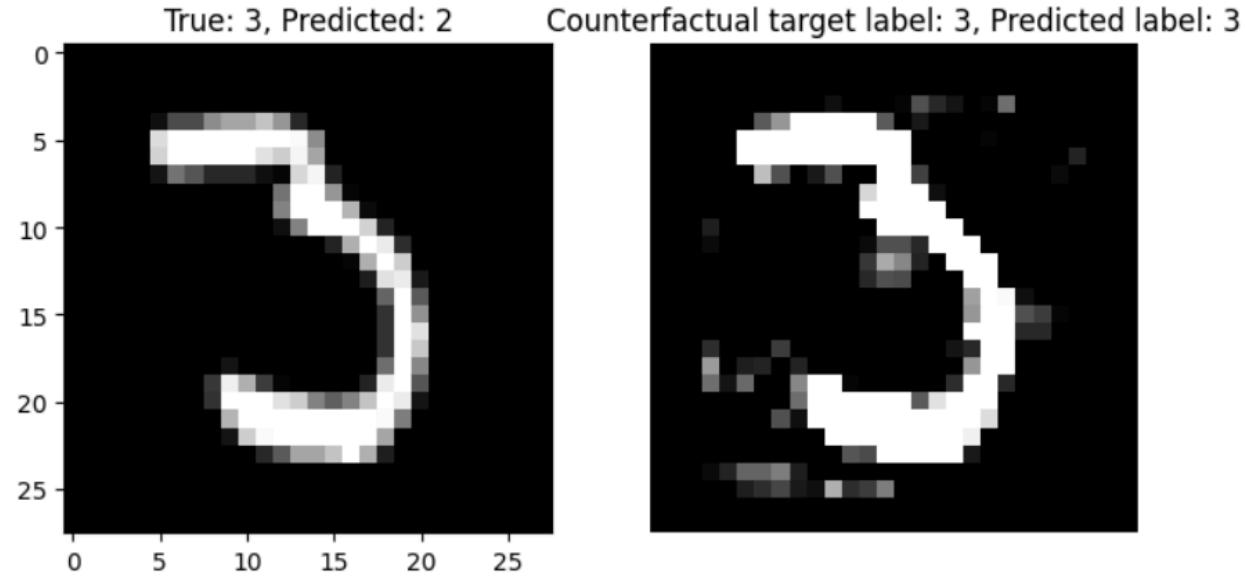
The perturbed pixels in this example, we see that it tries to make 4 more pointed and the the 4 dash longer and the bottom longer is also bold and extended



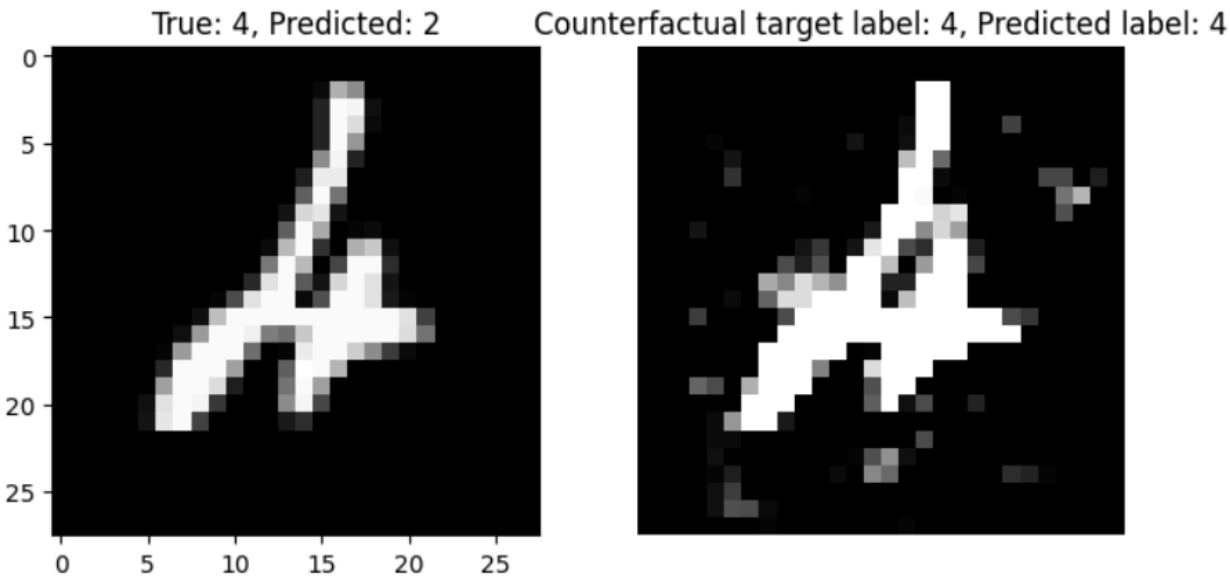
The perturbed pixels in this example, we see that it tries to make 2 more longer in the top portion and make the bottom a little flat similar to the 2 that we write



The perturbed pixels in this example, we see that it tries to make 9 more straighter from the bottom and makes the circle of 9 more bolder

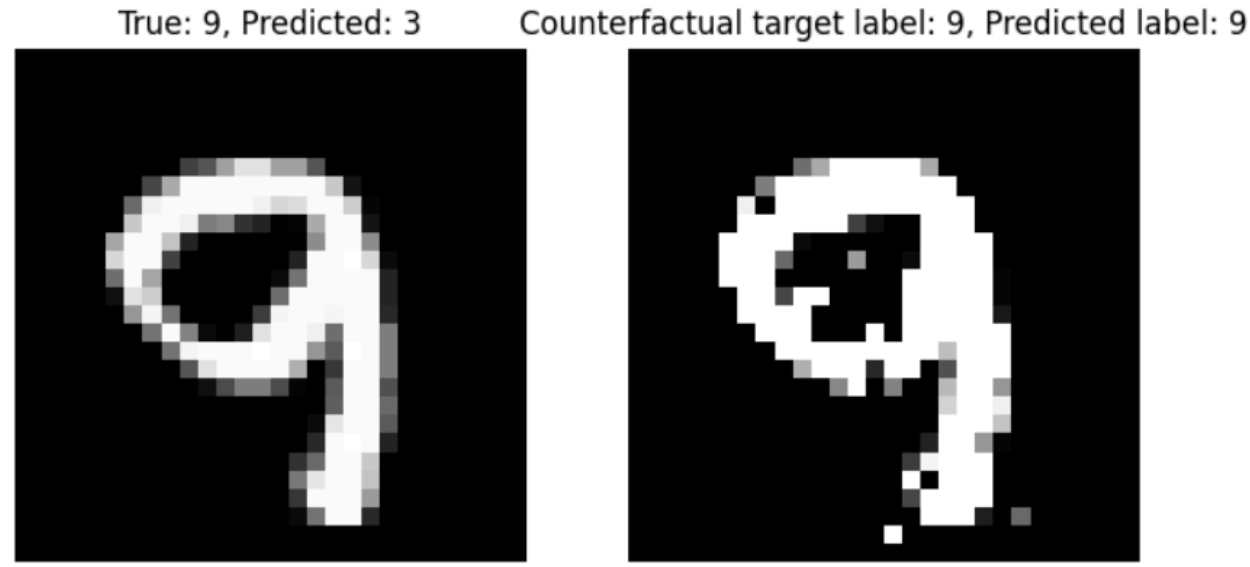


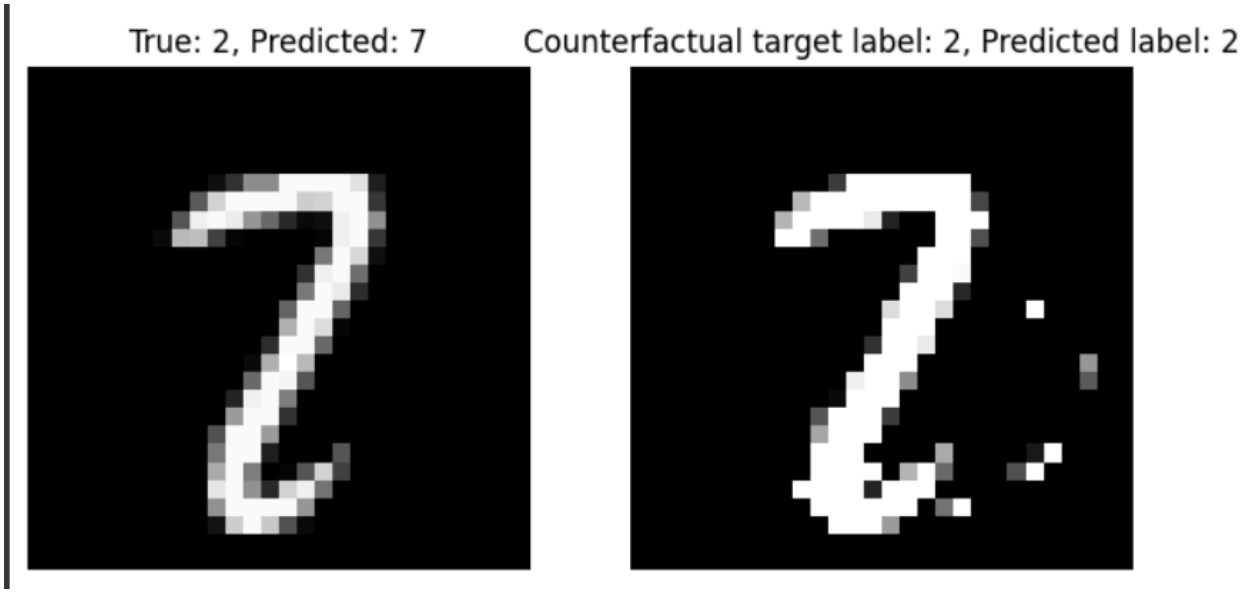
The perturbed pixels in this example, we see that it tries to make the middle dash in 3 longer which is a characteristic of number 3 and makes the rounds at the 3 more circular



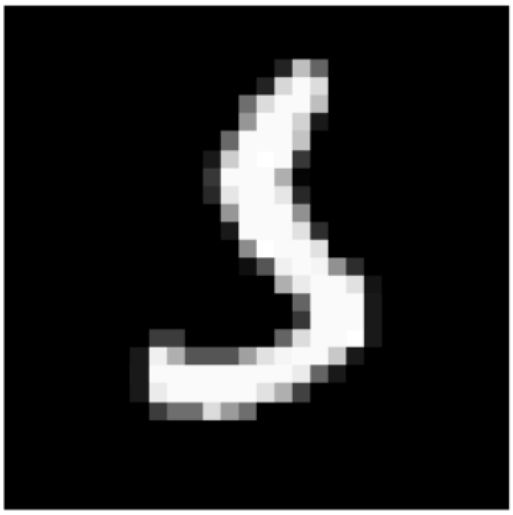
The perturbed pixels in this example, we see that it tries to make 4 more pointed and the the 4 dash longer and the bottom longer

Using Manhattan distance:

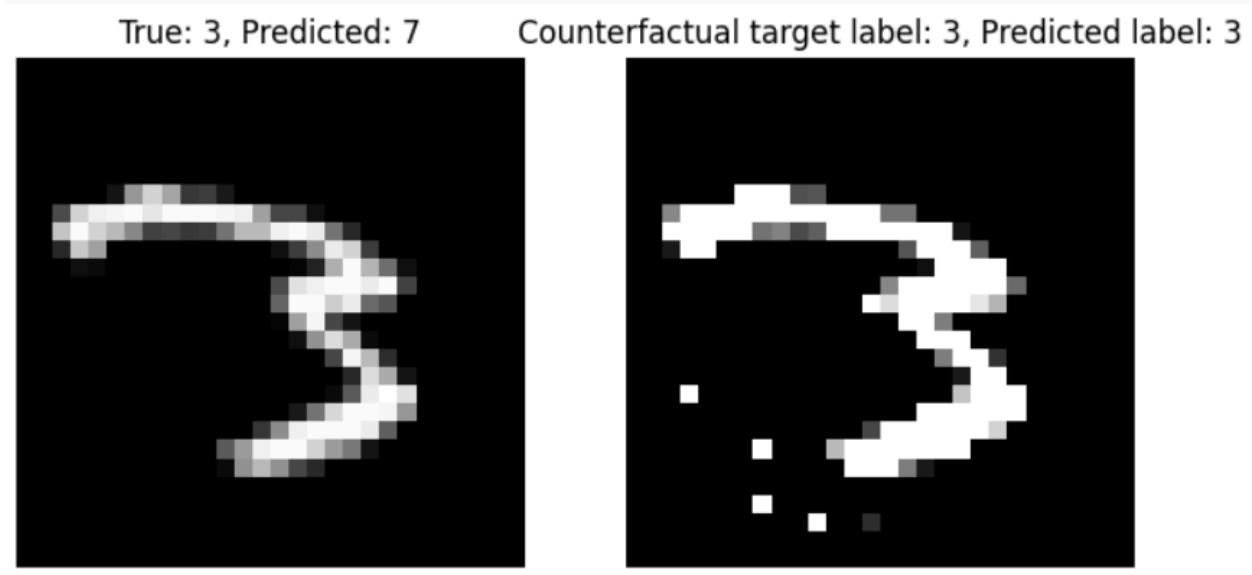

True: 9, Predicted: 3
Counterfactual target label: 9, Predicted label: 9

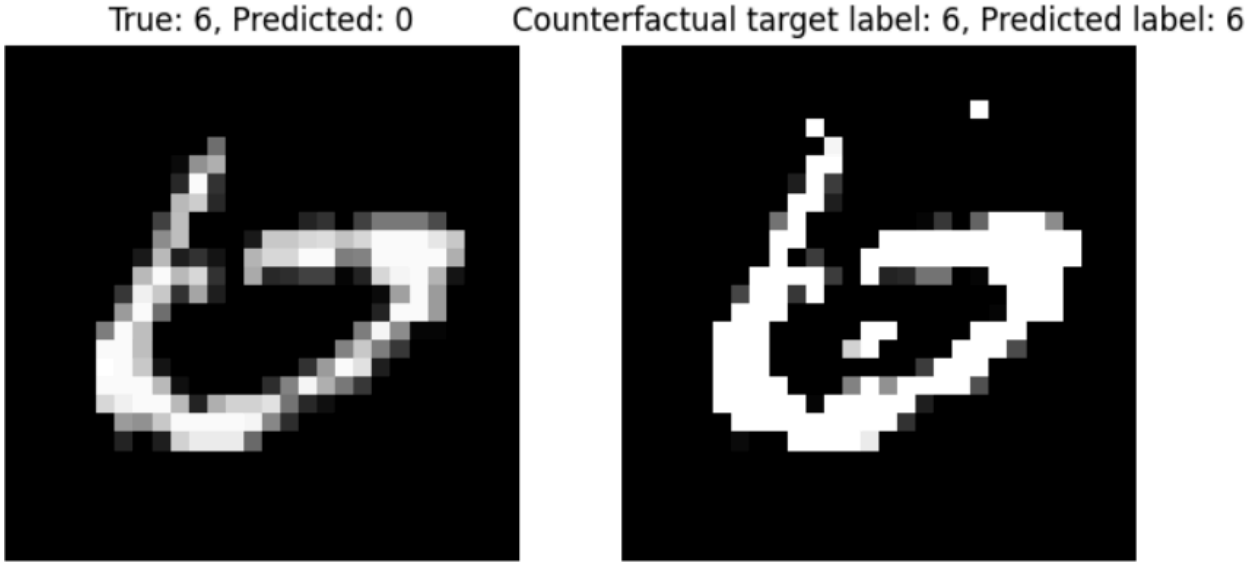

True: 2, Predicted: 7
Counterfactual target label: 2, Predicted label: 2

True: 5, Predicted: 3



Counterfactual target label: 5, Predicted label: 5





We see the same results for wachter method using manhattan distance

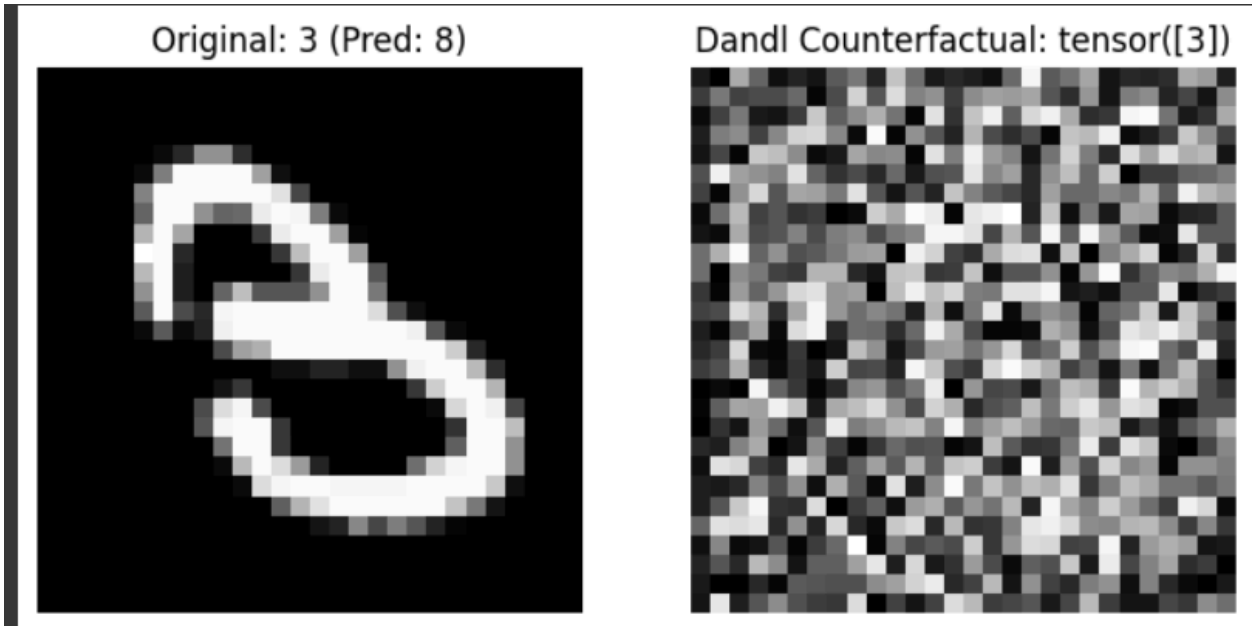
Updates to get counterfactual examples:

1. Pixel value clipping
2. Cross entropy loss between predicted and target predictions
3. Implemented with both L2 norm and manhattan distance with smoothening

DANDL'S METHOD

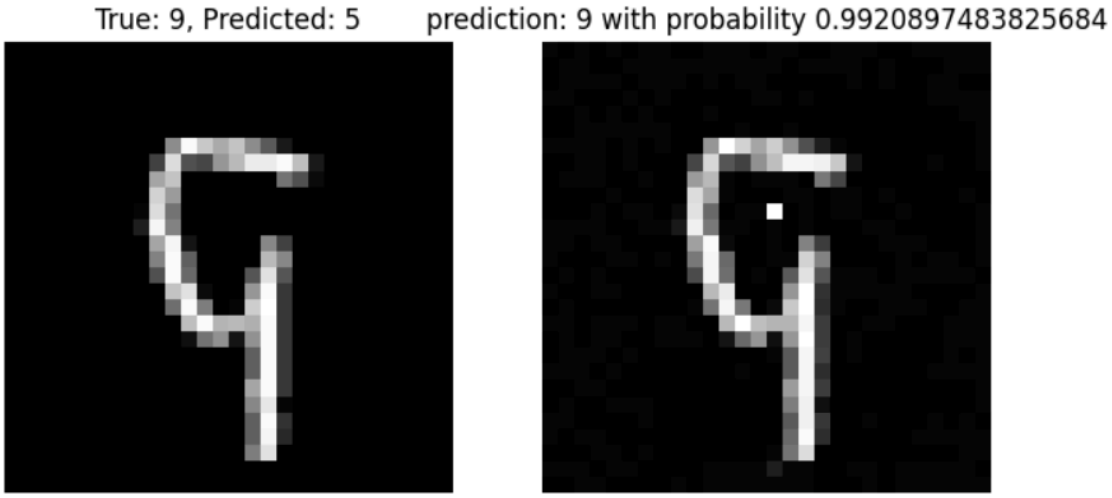
Methods used:

1. Used cross entropy loss
2. Smoothening
3. Used adam optimizer



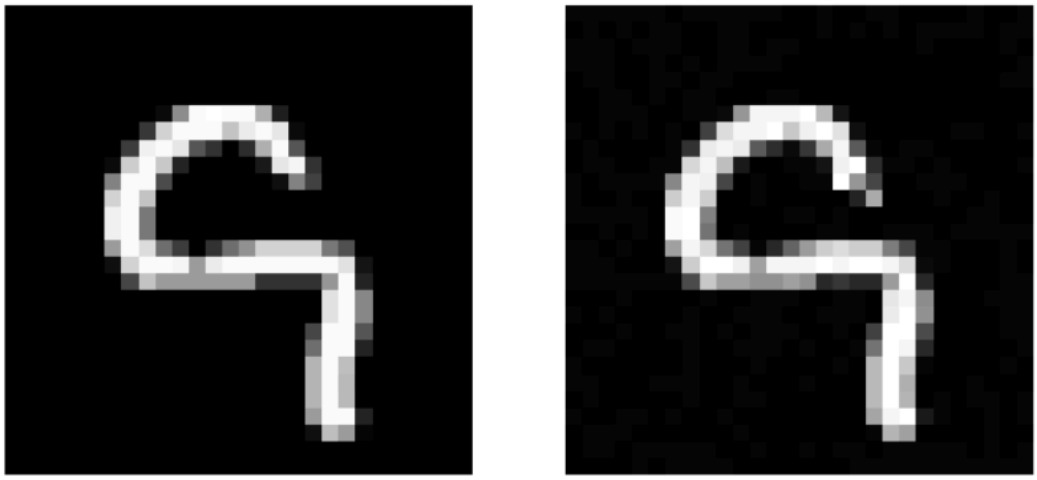
For this example, I was using NSGA2 method and was getting out of distribution counterfactual example

Using library:



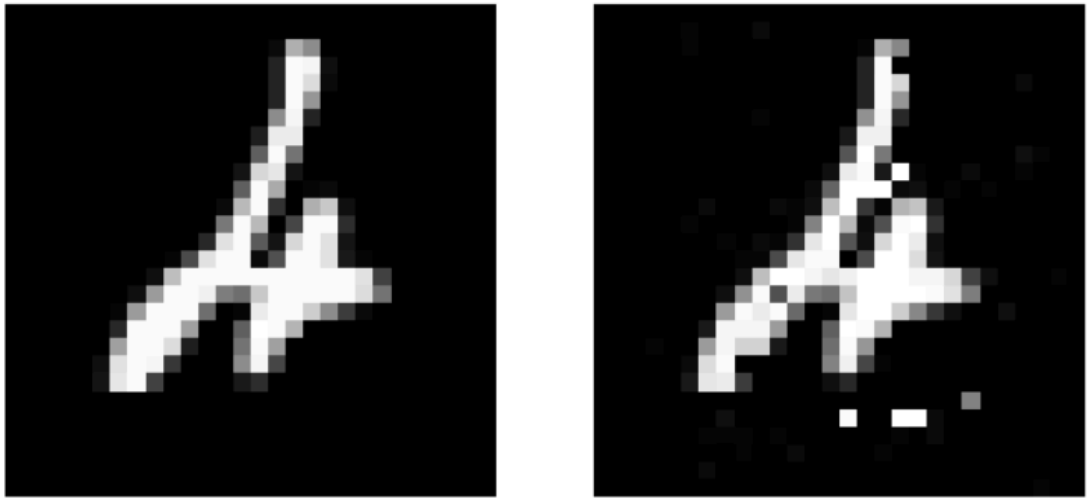
The example tries to join the gap in 9 to improve prediction.

True: 9, Predicted: 5 prediction: 9 with probability 0.9930877089500427

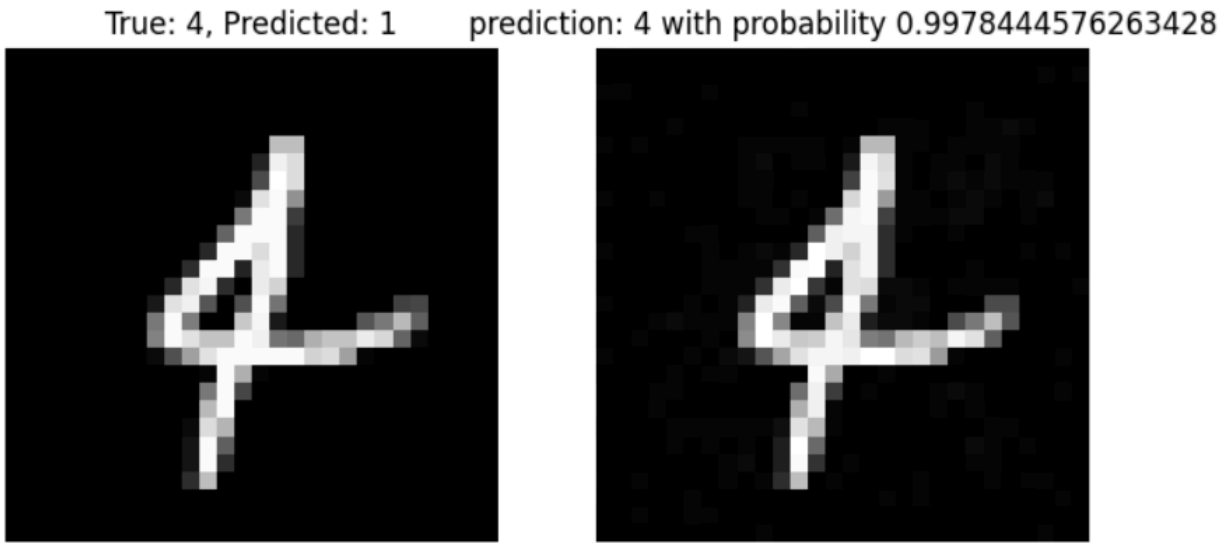


The example again tries to join the gap between the circle and the line to improve prediction to 9.

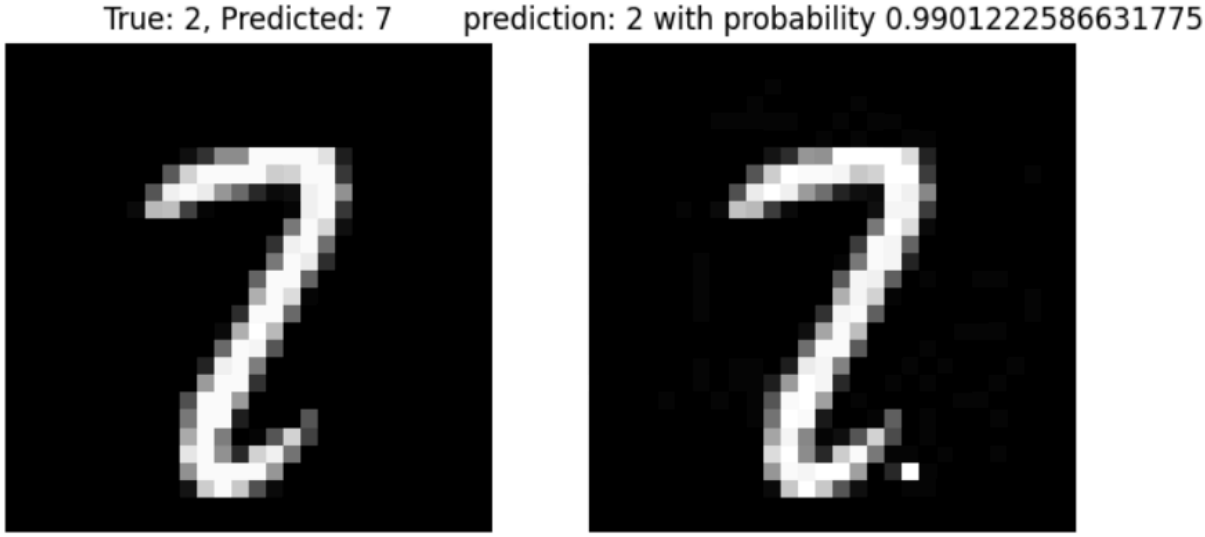
True: 4, Predicted: 6 prediction: 4 with probability 0.9916581511497498



In this example it changes the prediction by increasing the length of the line in 4

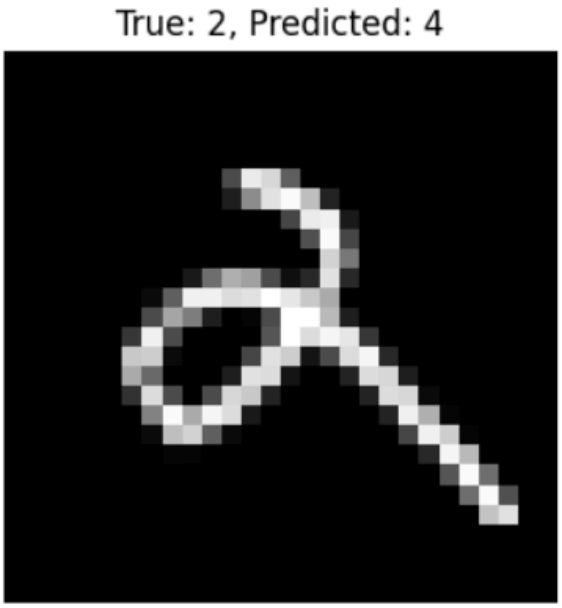


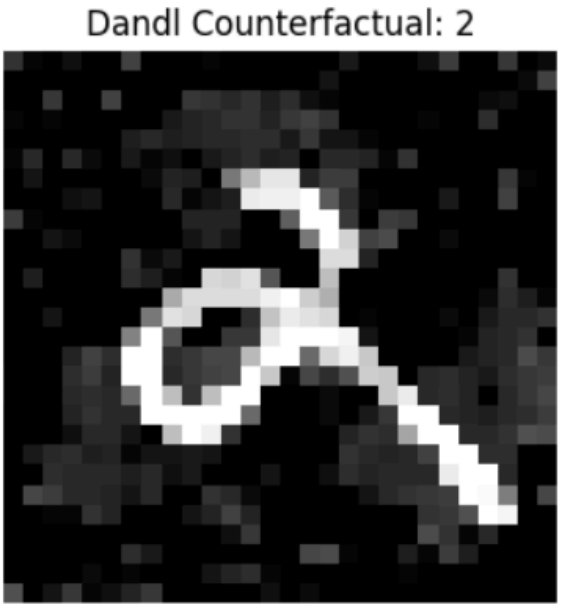
This example the lines are making the lines brighter and the pointy tip longer.



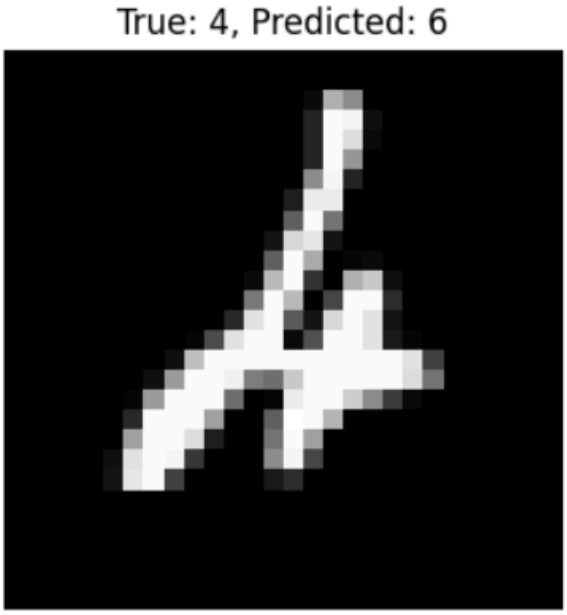
This example tries to extend the length of the baseline in 2 to make it longer.

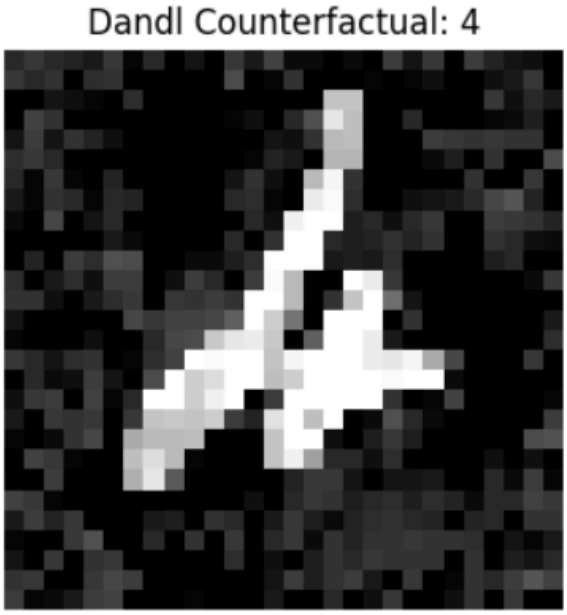
Using Adam Optimizer:



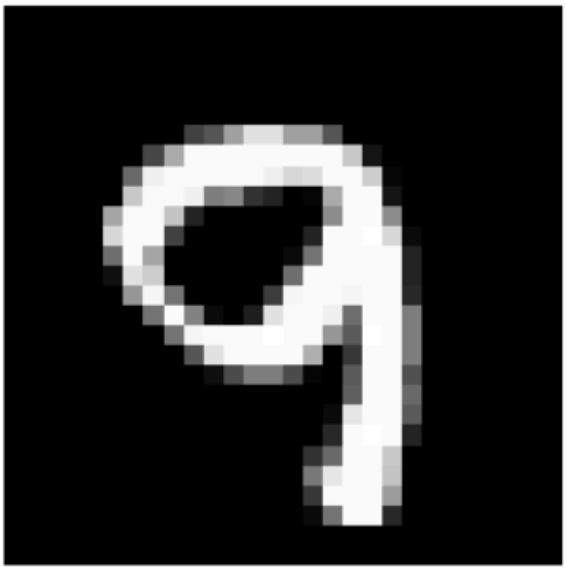


This example tries to improve the height of the upper portion of



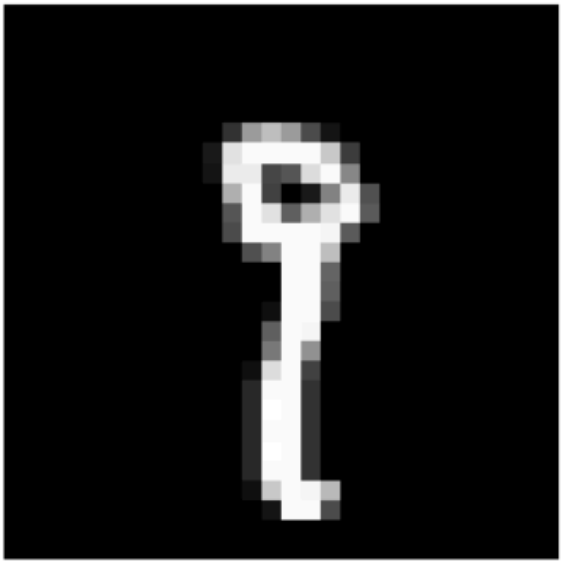


True: 9, Predicted: 3

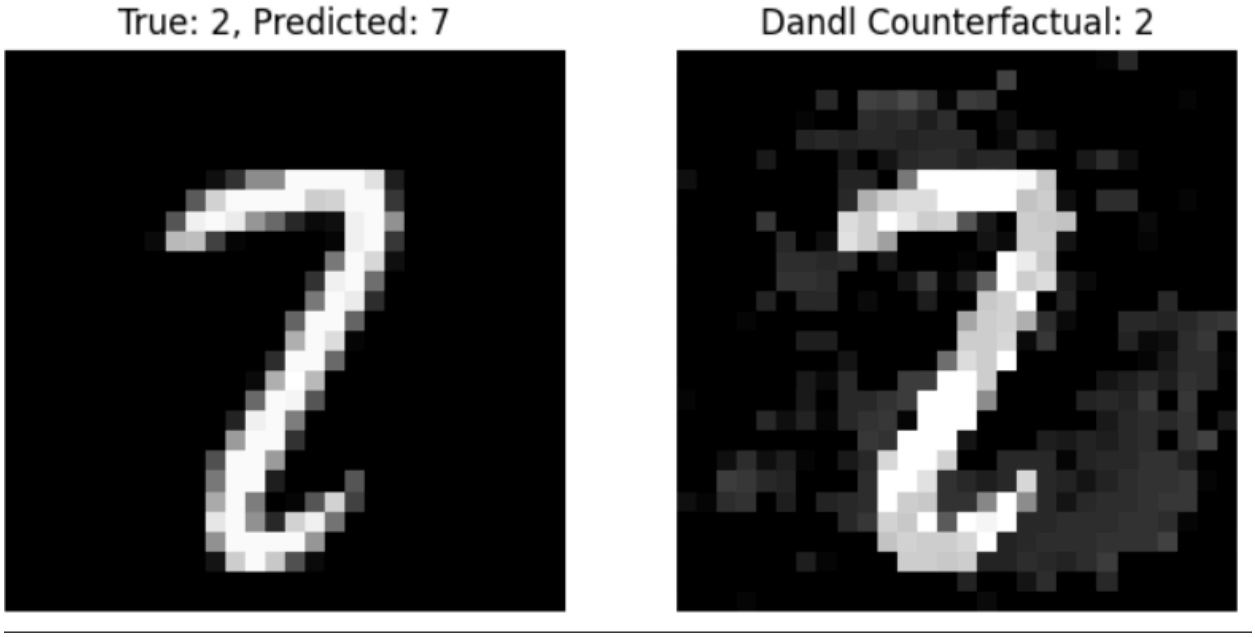


Dandl Counterfactual: 9

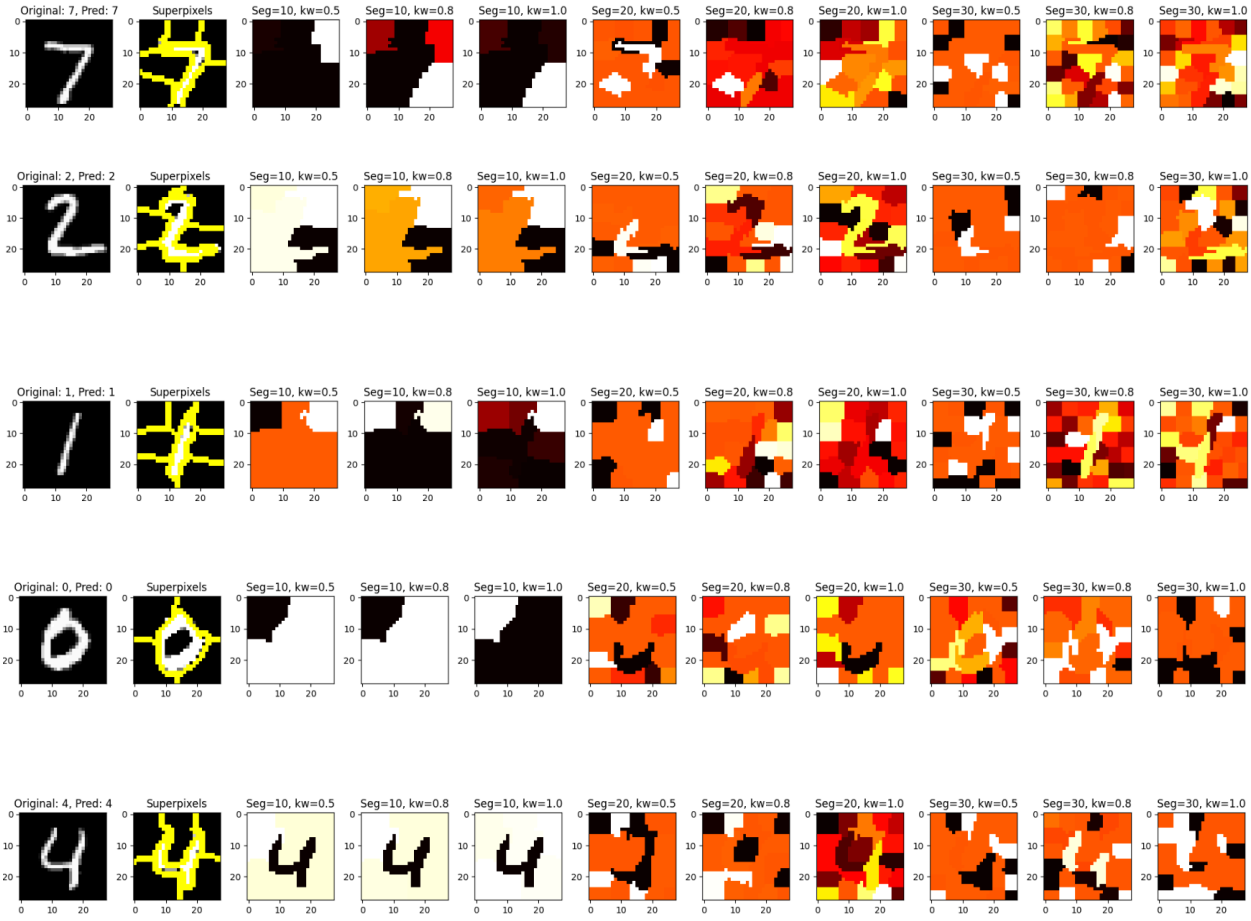
True: 9, Predicted: 7



Dandl Counterfactual: 9


True: 2, Predicted: 7
Dandl Counterfactual: 2

QUESTION 2:



LIME (Local Interpretable Model-agnostic Explanations) gives interpretable representations using 2 main parameters: segment size and kernel width.
The segmentation parameter is used for the number of superpixels generated from the original image.
For small segment size: The image is divided into fewer segments with large areas, but this may lead to missing fine grained details that may help in classification. These are easier to interpret. Larger segment size means the image is divided into more segments and focuses on the specific details as well. The stability reduces as tiny perturbations may change the explanation.

The kernel width is used for weighting of perturbed samples when fitting the local model.

For small kernel width: It gives less weights to perturbed samples close to the original instance. This means the explanations focus more on locality but they may be less stable as even tiny perturbations may change the explanation.

For higher kernel width: It gives higher weights to perturbations for the model to consider a larger neighborhood. It gives smoother explanations with less fluctuation. However, it may filter

out feature importance, leading to less specific insights into why the model made a decision.A **segmentation of 20 with kernel width 0.8 and 1** seems to provide the best trade-off between faithfulness and stability.

Kernel width	Number of super pixels or segment size	Explanation
0.5	10	Small number of segments, explanation is more localized but coarse
0.8	10	Small number of segments, Moderate balance between locality and smoothness.
1	10	Small number of segments, Smoother explanation but may lose specific detail.
0.5	20	Better number of segments with fine details and not over fine grained, Smaller segments, some regions more clearly highlighted.
0.8	20	Better number of segments with fine details and not over fine grained, Moderate superpixel size, balance between detail and stability.
1	20	Better number of segments with fine details and not over fine grained, Larger kernel smooths regions but may introduce artifacts
0.5	30	Very easy changed to perturbations, More segments, better granularity but increased noise.
0.8	30	Very easy changed to perturbations, Increased segmentation, loss of locality in some cases.
1	30	Very easy changed to perturbations, Higher

		segmentation and kernel width may over-generalize.