Fooling image copy detection algorithms with GANs

Anton Ponomarev aponom22@stanford.edu
Github link - https://github.com/ant-po/CS229project
Youtube link - https://youtu.be/p2dZFIYB4Ww

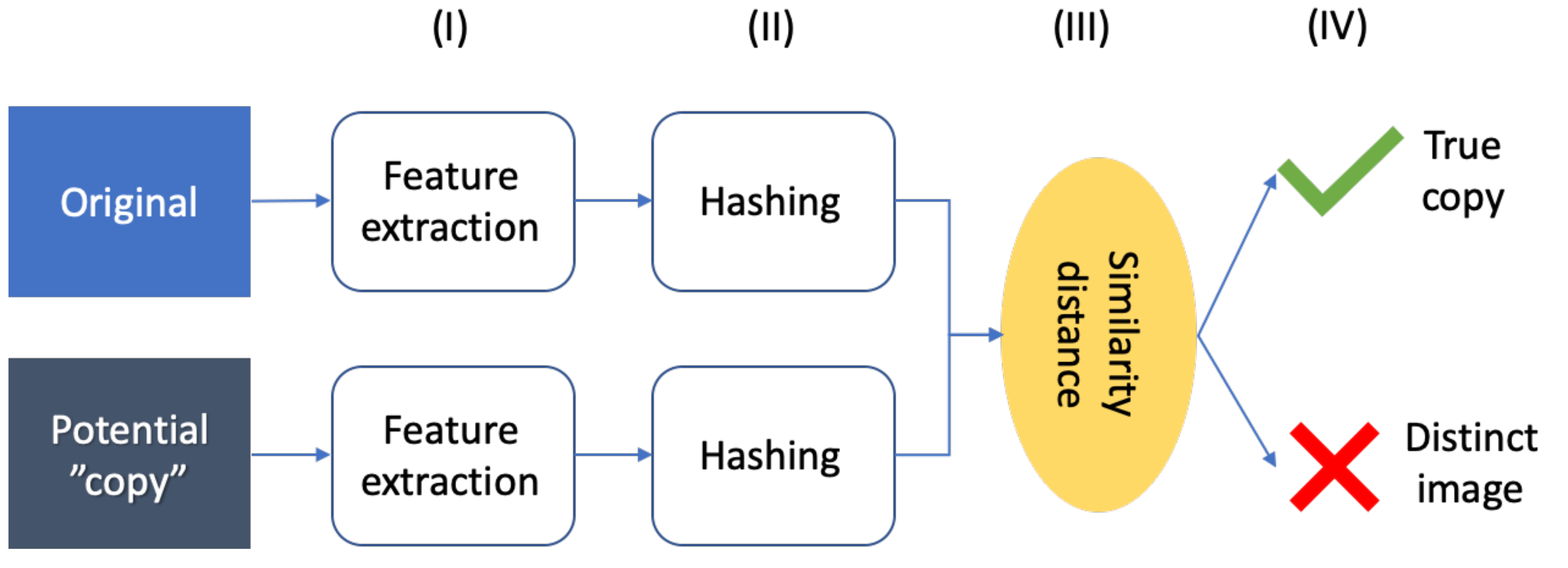
Motivation

Unrelenting digitization of data over the recent years is challenging our perception of authenticity and originality in the context of digital content. Increasingly more human decisions are being made solely relying on one's comprehension of visual online content. However, same advances in deep learning and computer vision that power our daily lives have also put powerful tools in the hands of malicious actors. With near-zero barrier to duplication and modification, even the simple task of determining whether two images are carbon copies of each other becomes non-trivial. In this project, we investigate whether state-of-the-art image copy detection algorithms are susceptible to being fooled by purposefully synthesized adversarial content. Using Generative Adversarial Networks (GANs), the intention is to leverage insights in order to make these commonly used algorithms more resilient to such attacks.

Data & Set up

To demonstrate the approach, we use images from the MNIST dataset: they are appropriately sized (28x28 pixels), already in grayscale and the content is consistent. To simulate a typical image copy detection algorithm, we are making use of perceptual image hashing. Given two images - original and potential copy:

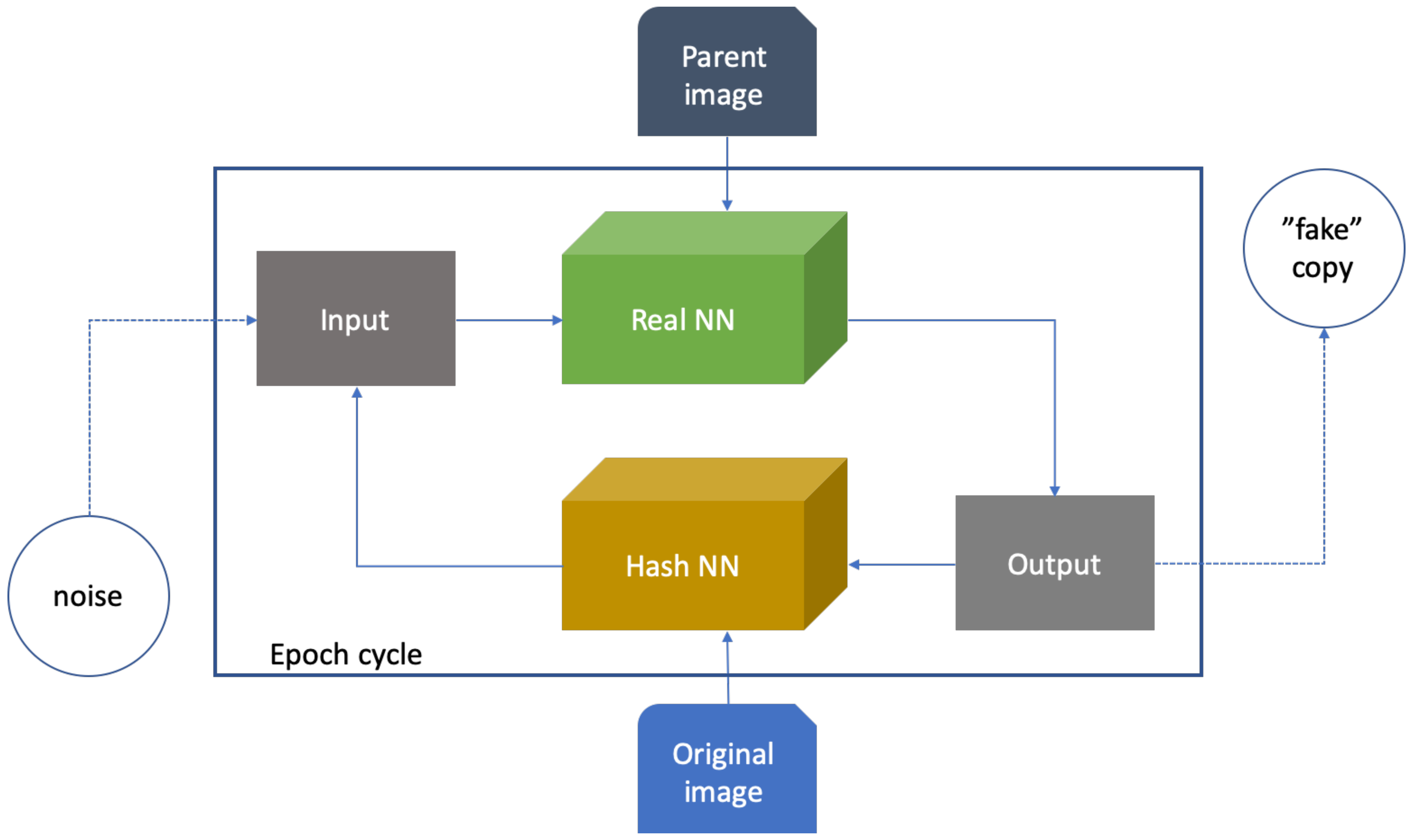
I. Features are extracted from both images using methods like block averaging, Discrete Cosine Transform, Wavelets
II. Perceptual 64-bit hashes are created from the features
III. Hashes are compared to each other by computing the Hamming distance (the number of differing bits)



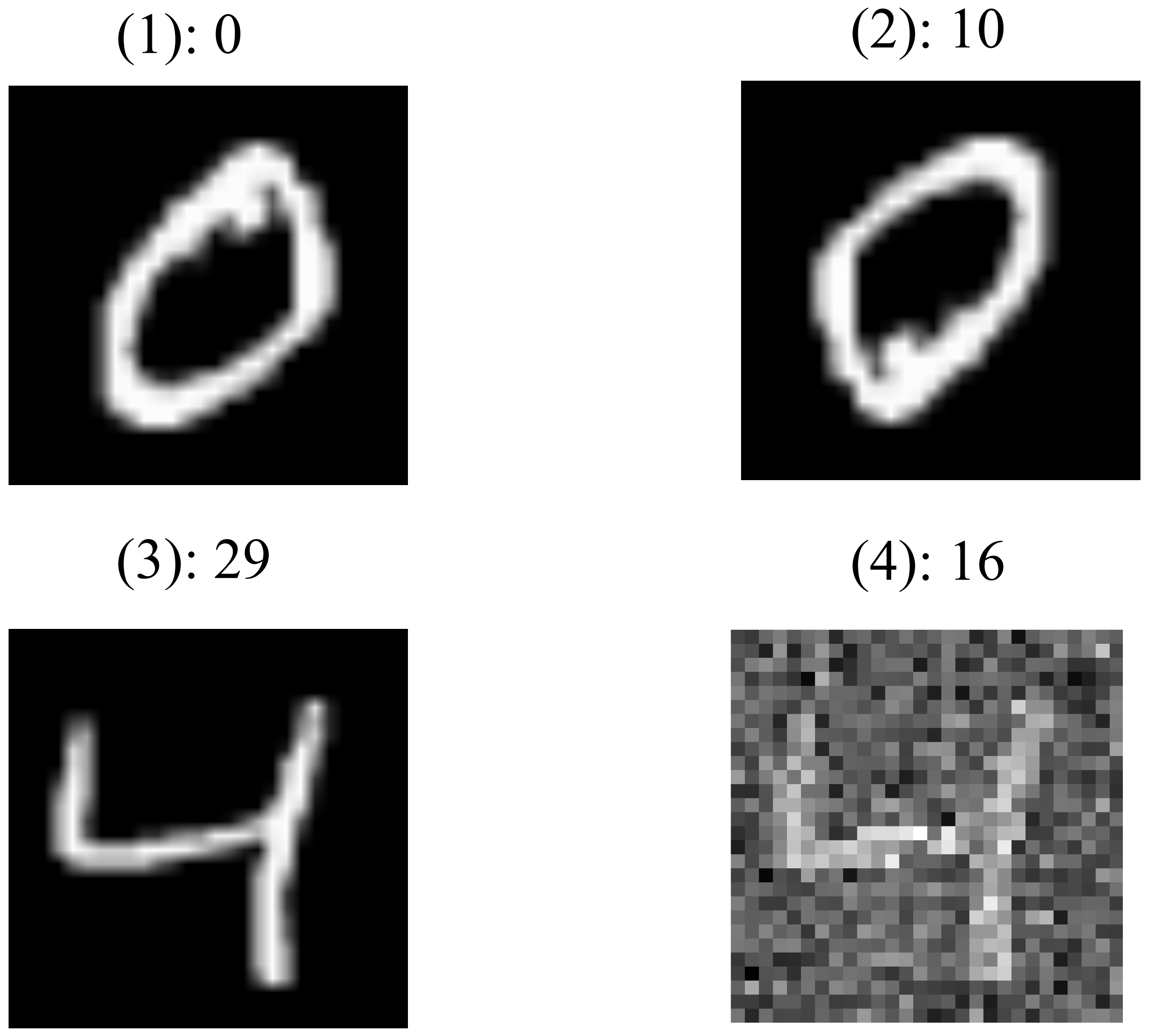
IV. If the similarity distance is below a certain threshold then the potential copy is considered a *true copy*. Otherwise it is labelled as a *distinct image*.

Model

We consider a family of models with the GAN-like architecture in order to generate a synthetic image with two qualities: 1) it can be easily identified by a human eye as distinct from the original but still realistic, 2) the Hamming distance relative to the original should be small. In effect, we aim to create a "fake" copy of the original that, if processed by the algorithm described above, would result in a false positive classification. The diagram below explains one of the structures considered in this project. Two neural networks are competing with each other: "Real NN" seeks to minimize the error between the input and the parent image, while "Hash NN" then minimizes the Hamming distance between the relevant perceptual hashes. The process is iterated over multiple epochs, until both networks converge.



Here is an illustration of the result from experiment. Four images are presented together with the corresponding Hamming distance relative to the original: (1) original, (2) "true" copy (rotated 180º), (3) "true" distinct, and (4) "fake" copy (generated by the model)



Observations

- Experiments suggest that it is totally possible to "fool" a common image copy detection algorithm with modest compute required
- The more complex the hashing mechanism used in the image copy detection, the harder it was to implement as it often resulted in the loss function being non-differentiable
- Relative dominance between two networks proved difficult to balance, alternative architecture with a joint loss objective exhibited more stable behavior
- Training the networks on multiple examples within the same MNIST class, has improved the performance as more robust features were learnt by the networks

Future research

- Expand the analysis to images with RGB channels and higher resolution
- Consider using CNN architecture instead