Generating Counterfactuals by Optimizing Over Latent Space Representations

Kayden Kehe, Dries Rooryck, Savanna Coffel
CS2822R - Harvard University

December 2024

1 Introduction

This project explores a new method for generating counterfactual images for regression models by optimizing over representations in the latent space. The approach links the output of an autoencoder, which models the input space of a target model, to the target model itself. By optimizing the autoencoder's latent space representation for a specific input, we guide the generated images toward a desired output value, ideally producing a continuous sequence of images which shows a human-interpretable change in the original input. We call our method `LoSoCs`, or Latent Optimization Stream of Counterfactuals. We also propose and implement metrics to quantitatively evaluate the quality of the generated streams of counterfactuals and explore other aspects of the behavior of the process in our latent space.

1.1 Motivation

Our primary motivation for exploring a new method of generating counterfactuals stems from previous work done by Dr. Vineet Raghu et al. in estimating biological age from images of chest radiographs using deep learning [Raghu, Weiss, Hoffmann, et al. 2021]. Here, the authors presented CXR-Age, a CNN-based model that could, given a subject's chest X-ray, predict the subject's biological age. Basic methods of interpretability, such as PCA and feature occlusion, have demonstrated which aspects of a chest X-ray this model assigns the most importance to. However, it seems like most interpretability methods lack the ability to show exactly how the model qualifies its decisions, which could be interesting from a clinical perspective (perhaps the model is processing features a radiologist might not consider). The method described in this paper is an attempt to solve that problem.

Our second primary inspiration was a method for counterfactual generation called Generative Visual Rationales [Seah et al. 2018]. The approach generates altered versions of a chest radiograph on a model trained on latent space representations in a similar way to our work, aiming to explain what features would need to change for the disease not to be present. We wanted to determine if

we could expand this method to produce X-rays with human-understandable changes for a model trained on X-rays instead of latent space representations, potentially generalizing the method to a broader class of model. While we have since deviated working with chest radiographs, and instead focused the verification of our method on data from MNIST, the underlying motivation remains the same.
As we progressed through this project, we sought to answer two key research questions:

1. Can we implement `LoSoCs` successfully and demonstrate its feasibility on an initial data domain?

2. What metrics can we develop to assess the quality of counterfactual sequences `LoSoCs` generates, in terms of how human-interpretable they are?

Insofar as we had a hypothesis for these questions, we hypothesize that `LoSoCs` will successfully generate sequences of counterfactual images that are humanly interpretable, and that our metrics will reflect this.

1.2 Challenges and Adjustments

Although the project was originally formulated as a way of investigating CXR-Age, we found after much failed experimentation that a well-formed latent space is *absolutely crucial* to the success of the method, and acquiring or training a VAE that could generate both high-fidelity X-rays and whose latent space had desirable properties posed a logistical challenge whose solution would involve grueling coding and debugging. This sort of problem-solving seemed against the spirit of the course, so we re-scoped the project and pivoted to MNIST for simplicity.

While our chest X-ray autoencoder succesfully reconstructed and sampled high-quality X-rays, its latent space did a poor job of modeling the spatial relationships between X-rays. It sparsely modeled its input space, so that any movement within the latent space from a sampled point would result in nonsense X-rays. We can clearly see this when we, for example, sample points along the line connecting two X-rays' latent space representations. Instead of a human-interpretable transition, we would observe one X-ray being superimposed onto the other.

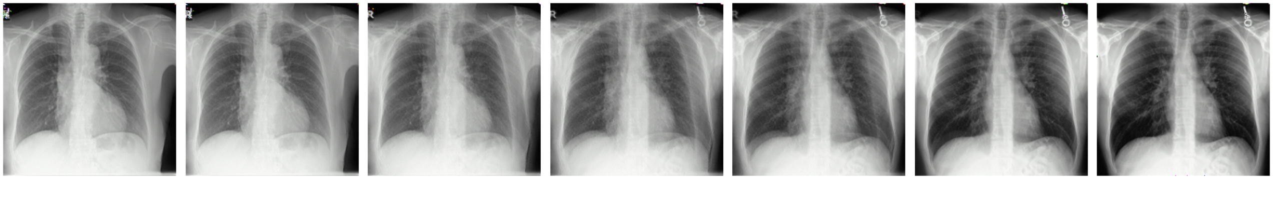


Figure 1: Each plot represents a decoded X-ray along the linear path from the first image to the second. The middle images showcase out-of-distribution images, with one X-ray layered over the other.

When optimizing for higher or lower biological age, the expressivity of the latent space and its lack of ability to properly model traversal resulted in what were essentially adversarial examples.

The autoencoder was so expressive that we lost the benefit of its ability to densely represent important information.

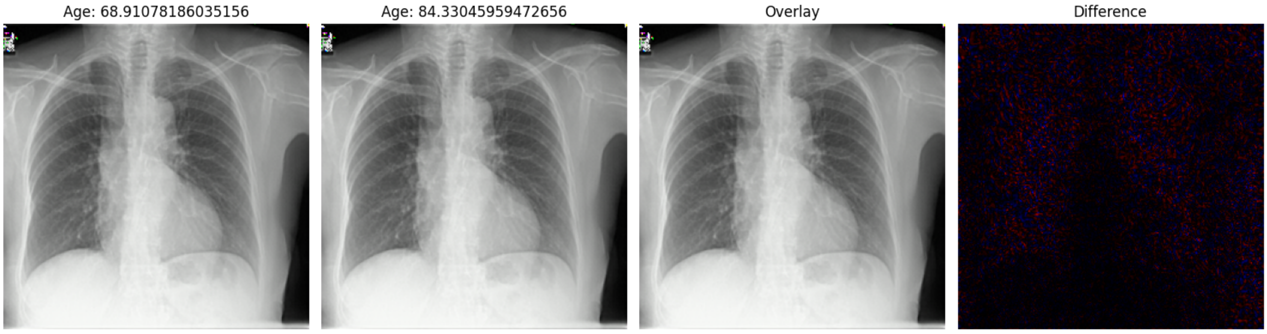


Figure 2: Failed chest X-ray optimization attempt

Despite our many attempts to solve these problem by introducing additional optimization criteria or otherwise alter the optimization process, these issues proved too big a hurdle for us to make progress.

1.3 Evaluation Criteria

Part of the goal for this project was to develop a set of metrics by which we can evaluate streams of counterfactuals. In our selection of the following evaluation metrics, we chose criteria which we believe to be universally applicable and help to isolate changes directly related to the model's decision-making.

- *Linearity of Output* – Measures the degree to which the regression outputs of the counterfactual sequence form a linear relationship. A non-monotonic output would indicate a failure of the method, and nonlinear changes in step size would complicate inference.
- *Continuity* – Assesses whether the generated images represent smooth, incremental changes.
- *Plausibility* – Evaluates whether each generated image remains within the model's input space. We chose not to pursue implementing a metric for plausibility for the sake of this class. However, we will do some qualitative analysis of our optimization process later to assess plausibility.
- *Fundamental Similarity* – Evaluates the preservation of essential characteristics from the original input to the model. For example, in the case of CXR-Age, the generated images should still represent X-rays of the same individual. For datasets like MNIST, we may think of an image being fundamentally similar if it was drawn by the same individual.

2 Methods

2.1 Optimization

To implement LoSocs 🧦 we first trained a simple, fully connected digit classifier and a convolutional VAE with a 16-dimensional latent space on the MNIST dataset. The VAE was adapted from a GitHub page and was trained to minimize reconstruction loss and KL-divergence. We adapted the method for a classifier by optimizing for an increased value in a specific output logit (as opposed to a regression output in the CXR-Age setting). Each iteration of the optimization process follows Algorithm 1:

Algorithm 1 Latent Space Optimization for a Stream of Counterfactuals, LoSoCs

```
Input: Pretrained classifier $C$, pretrained VAE with decoder $D$, target digit $t$, threshold $\tau$
Output: Optimized latent representation $z^*$
Initialize latent space representation $z$
repeat
    Decode latent representation: $x \leftarrow D(z)$
    Pass decoded image through classifier: $\mathbf{p} \leftarrow C(x)$    ▷ $\mathbf{p}$ is the vector of logits from $C$
    Compute loss: $\mathcal{L} \leftarrow -\mathbf{p}[t]$    ▷ Minimizing $\mathcal{L}$ increases the logit for target digit $t$
    Backpropagate $\mathcal{L}$ through $C$ and $D$ to update $z$
until $\text{softmax}(\mathbf{p})[t] > \tau$ (e.g., $> 0.999$)    ▷ Ensure high confidence for target digit
Return: Optimized latent representation $z^*$
```

Because the number of iterations is not fixed, the number of intermediate frames depends on the learning rate, with smoother transitions achieved through smaller steps. This can be considered a hyperparameter to be set by the user.

2.2 Evaluation Metrics

Linearity of Output: Computed as the coefficient of determination (R^2) between the target logits and sequence indices from each iteration in LoSocs. Higher R^2 values indicate a more linear relationship.

Continuity: Measured by the coefficient of variation (CV) of absolute differences between the pixel values of consecutive frames. Lower CV values indicate smoother transitions.

Linearity and continuity are computed over a cohort of 1000 LoSoCs instances, each involving a random digit being transformed into a random digit.

Fundamental Similarity

Global evaluation: Analyzing convergence patterns in the latent space across multiple inputs. If paths traveled through the latent space tend to converge to a common local area, meaning the optimization process tends to result in very similar outputs, it is clear that fundamental similarity is not being maintained. For a cohort of 500 images each, we converted 4 digits to 7 digits, random digits to 7 digits, 4 digits to random digits, and random digits to random digits. From there, we computed the pairwise Euclidean distances between the initial images in the latent space and the distances between optimized images in the latent space for analysis.

Local evaluation: Assessing whether fundamental characteristics were preserved by seeing if an external classifier can consistently predict for one optimized image, what original image it was a result of. Our implementation of this was to train a Siamese network, which takes in two images processed in parallel by two independent sub-networks with the same parameters to produce a feature map. The siamese network tries to classify this pair as similar or not similar based on the Euclidean distance between the two images' extracted feature maps. In our case, we use a classifier with a simple LeNet feature extractor, and we train on a generated dataset of 8000 pairs of images and 1 or 0 labels. To generate this dataset, we start with 5000 pairs of images of 4-digits optimized to various other digits. On 4000 of these pairs, we create 4000 pairs so that one image is an original 4-digit and the other is the image of the digit it was optimized to (1 labels), and 4000 pairs with a 4-digit image and an image that was optimized from a different starting 4-digit (0 labels). This gives us a balanced training dataset of 8000 pairs. On the remaining 1000 pairs, we apply the same process to get a balanced hold-out test dataset of 2000 pairs. *The Siamese network should learn to extract features of the image that remain after our optimization process.* The metric associated with this exploration is how much more accurately than a naive baseline the Siamese network can predict which pairs were 1-labeled in the test dataset.

3 Results

3.1 Optimization

The method succeeds in generating streams of counterfactual images.

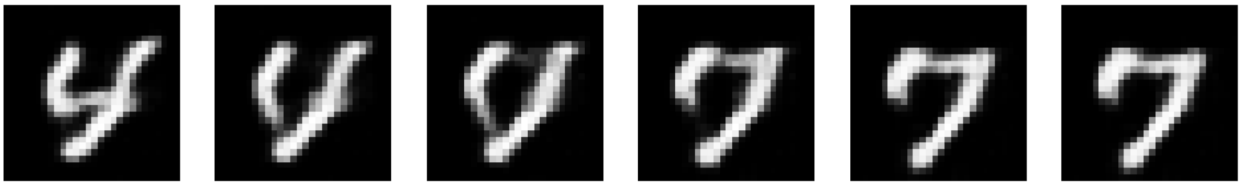


Figure 3: An example of a reasonably successful optimization of a 4-digit to a 7-digit.

3.2 Linearity and Continuity

The results from the cohort of 1000 images:

Table 1: Summary Statistics for Linearity and Continuity

Metric	**Mean**	**Standard Deviation**
R^2	0.58	0.40
Coefficient of Variation (CV)	2.52	0.75

3.3 Fundamental Similarity

Latent Space Convergence

For each of the following histograms, the left plot shows the pairwise Euclidean distances in the latent space between initial images, and the right plot shows the distances between final optimized images.

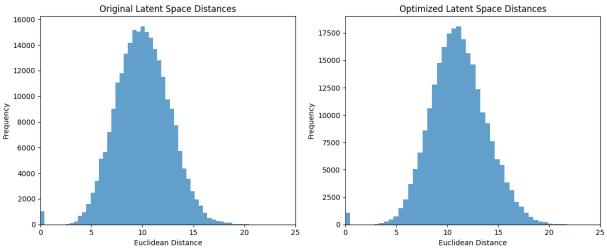


Figure 4: 4 → 7

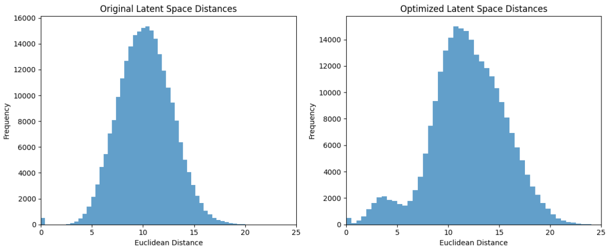


Figure 5: 4 → random

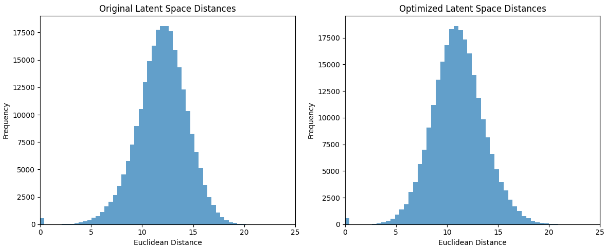


Figure 6: random → 7

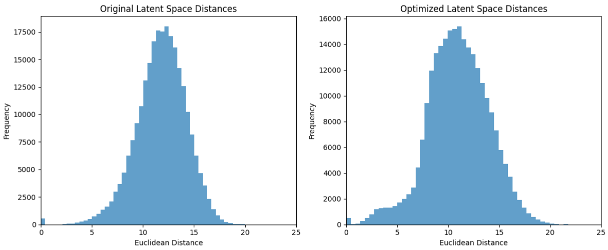


Figure 7: random → random

Metric	**4 → 7**	**4 → random**	**random → 7**	**random → random**
Original Distance Mean	10.02	10.21	11.95	11.95
Final Distance Mean	11.09	11.83	11.07	11.01
Original Distance STD	2.60	2.58	2.45	2.41
Final Distance STD	2.74	3.69	2.52	3.02

Table 2: Metrics for the pairwise distances. Each row represents a metric and stage in the optimization process, each column represents a specific optimization beginning digit and target digit.

Classifier

The table below shows classification metrics of our Siamese Network in associating original images with their optimized outputs.

Dataset	Accuracy	Precision	Recall	F1-Score
Test Set	0.818	0.8171	0.8268	0.8219
5-Fold Cross-Validation	0.7925	0.8112	0.7626	0.7850

Table 3: Performance metrics of the Siamese Network in associating original images with their optimized versions. The 5-fold cross-validation accuracy is measured by training a Siamese Network on 4/5ths of the data, 5 times, and averaging metrics, whereas test set accuracy is of a measure using a re-trained siamese network on the full training data over the 2000 hold-out test pairs.

We also create four distinct pairs of images from the below four distinct 4-digits as 'candidate original images' and 1 image, the result of optimizing one of these 4-digits to a 3-digit. The below image is a qualitative result of in a typical forward pass, how the Siamese model typically assigns distances. We also soft-max the negative of these distances to get probabilities as a proxy of what the model's confidence is in saying each of the four candidates are the true original image.

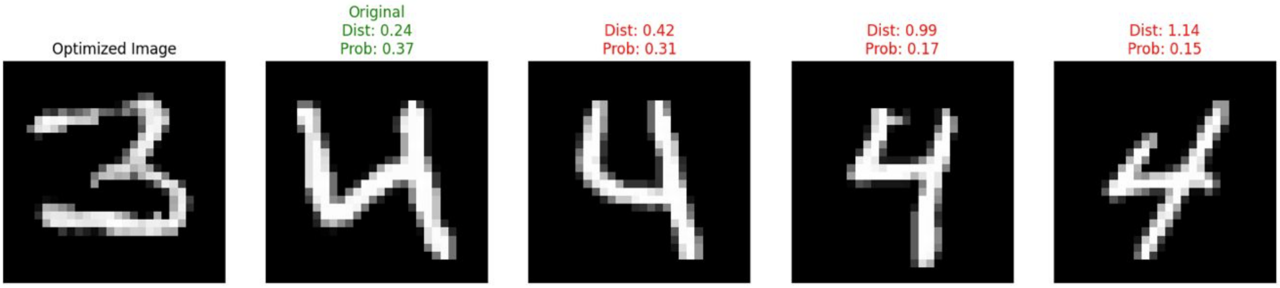


Figure 8: Comparison of the optimized image with the original image (green) and distractors (red). Each subplot includes the distance (Dist) as computed by the Siamese Network, and probabilities (Prob) based on these.

The code for all experiments and methods of `LoSoCs` is available on GitHub: `LoSoCs GitHub Repository`.

4 Discussion

4.1 Linearity and Continuity

Based on our findings, it seems like the changes in our target logits are fairly non-linear from frame to frame, with an R-squared of around 0.58. This is somewhat surprising, but somewhat explainable by the understanding we got of latent space movement from the next section. Second, we see relatively small changes from image to image, as evidenced by the CV of 2.52.

However, after further consideration, and having computed the metrics, we think *excluding* linearity and continuity as good, informative, and universally applicable metrics may be the better option. Broadly speaking, the key information conveyed by our approach is the information shown during the transformation (this is less the case in the MNIST example but more clearly the case for something like CXR-Age), and neither of these metrics influence our ability to understand and reason about those transitions. If we see the size of the heart increase as biological age increases, for example, the speed at which it increases in size is unimportant. Both of these metrics essentially measure the speed at which the size of the heart increases, which highlights another issue - these metrics convey somewhat overlapping information. Continuity is also not universally desirable. It may be the case that faster changes for specific ranges of output showcase something interesting about the model, and it isn't necessarily desirable for those changes to be smoothed out. The metrics are also generally uninteresting and difficult to interpret in isolation. Having computed these metrics provided evidence against, and gave us reason to think that our hypothesis, that these would be useful metrics, was wrong.

4.2 Fundamental Similarity

Distances, Path Traversal

The results from our analysis of fundamental similarity were some of the most surprising of the project.

Firstly, the results from analyzing the distances between original and optimized latent space representations shows that the process doesn't tend to converge, and in fact, the distances generally stay the same for both pre- and post-optimized digits regardless of whether the initial or target digits are fixed or randomly selected. This is counterintuitive, since we'd expect digits to cluster together and therefore generally be closer together than digits selected totally at random, but this can likely be explained by the high-dimensional nature of the latent space compared to the number of digits, since each additional dimension allows more spatial overlap between the clusters. We chose to analyze the distance for both random and fixed digits to see whether that affected the distances, but the difference seems negligible (when the starting digit is fixed to 4, the initial distance is slightly smaller on average). Also of note, the histogram of distances from 4 to a random digit has a smaller peak before the overall peak, and it appears this peak is mostly made up of transformations to 0. Perhaps the 0 cluster is slightly closer to the 4 cluster than the rest.

At this point, our conjecture to the typical behavior of the optimization method was that it took a linear path through the latent space and stopped the moment it hit the edge of the cluster (i.e., essentially the moment the predicted probability of the target digit reached about 1.00).

We successfully found that the optimization process for MNIST takes fairly linear paths. We can see this visually in Figure 9.

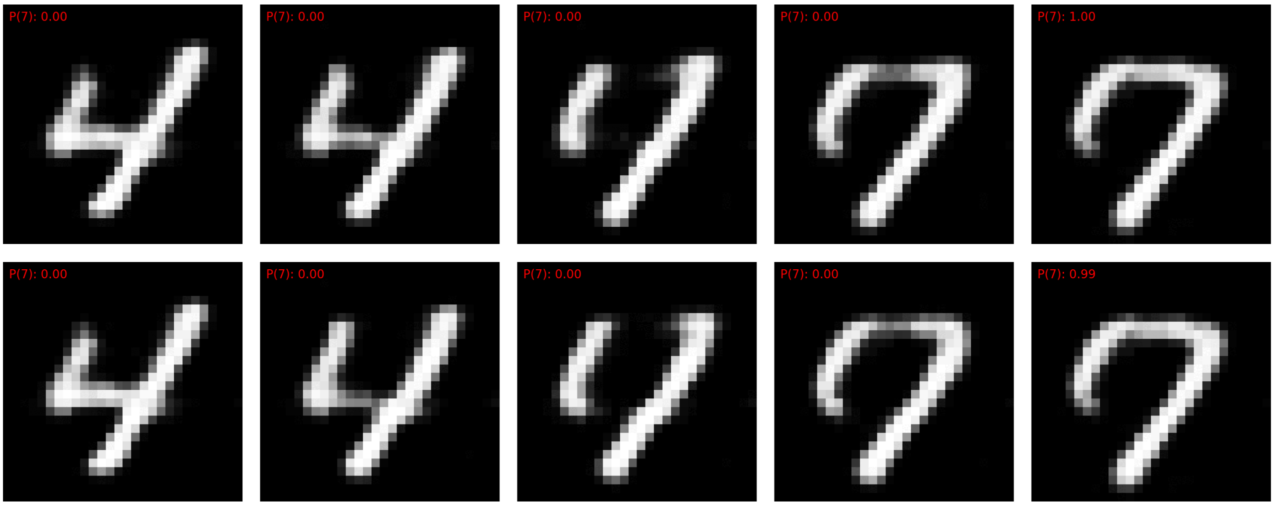


Figure 9: Two `LoSoCs` streams from a 4 to a 7. The top shows a linear traversal through the latent space, the bottom shows our optimization process.

To quantify this, for each optimization in a cohort of optimizations from random digits to random digits, we calculated the coefficient of determination over each dimension paired with its timestep, providing the histogram in Figure 10. Globally, averaging over each optimization and dimension, we saw an R^2 value of about 0.93, showing a fairly strong linear relationship between the values in each dimension over time.

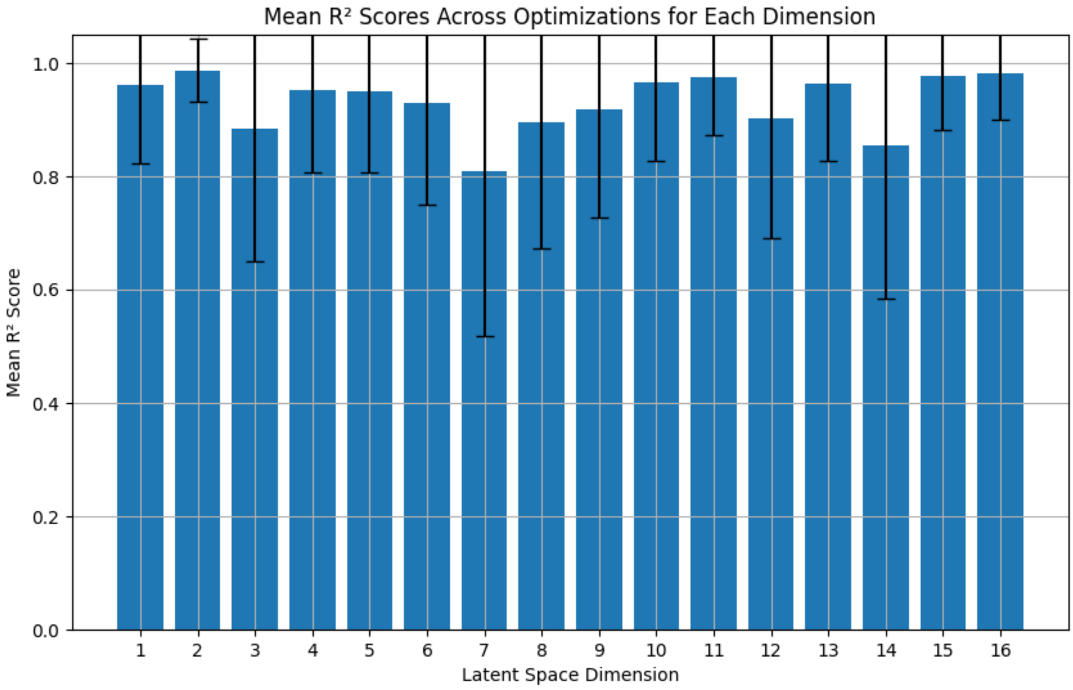


Figure 10: Mean R^2 scores for each dimension

To assess our conjecture about whether the optimization process halts at cluster boundaries, we assessed the absolute difference between successive images before and after reaching a predictive probability of 0.99 for a cohort of 1000 images, beginning as random digits with random target digits.

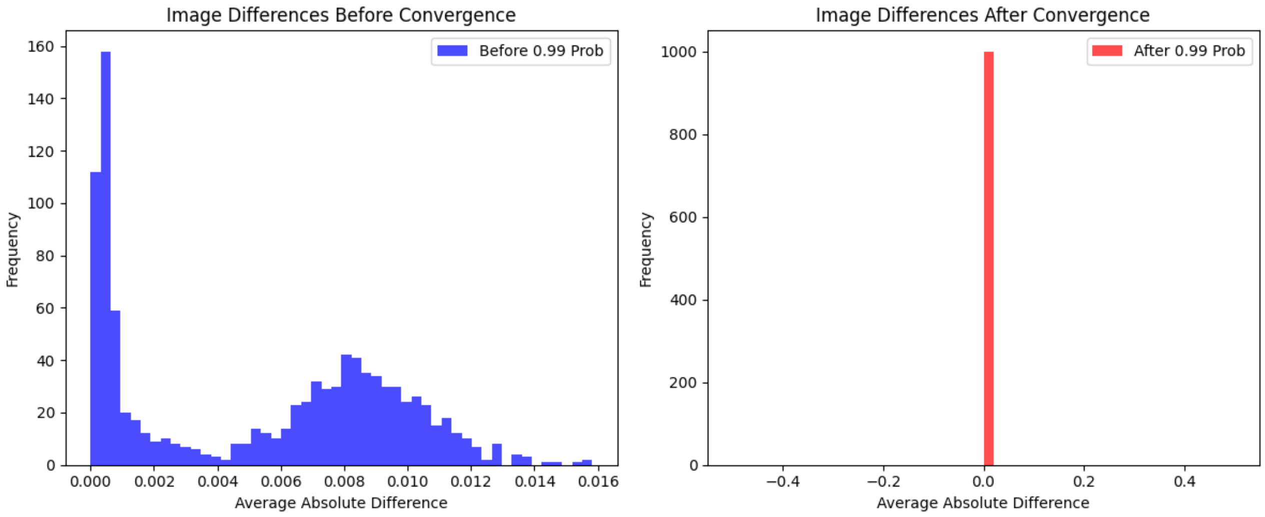


Figure 11: Distribution of absolute difference in images before and after a predicted probability of 0.99

As you can see from comparing the two distributions in figure 11, the images stop changing as soon as they converge to the correct class. Thus, we have a clear picture of the optimization process for MNIST - we begin at the latent representation of a starting digit, and take the shortest path to the boundary of the target digit before halting. This behavior of halting at the cluster boundaries may also be partially responsible for our implausible digits.

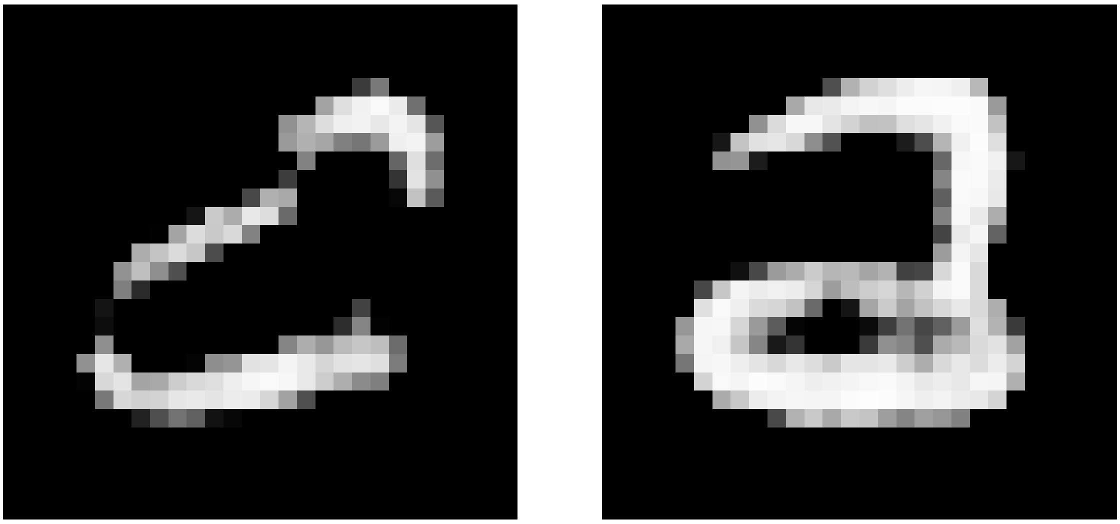
Classifier

The success of our classifier indicates that some information is being maintained throughout the optimization process. Since our dataset is balanced and our precision and recall scores are close to our accuracy scores, the model seems to be well calibrated across classes. The similarity between cross-validation and test scores goes to show that the feature extractor is not overfit on training data. And being in the vicinity of 80% accuracy means that the classifier does much better than the 50 % naive baseline of randomly guessing if a pair of images is the result of optimization or not. This was a surprising result, and it suggests that the optimization process is conserving some conceptual information about images! We were not able to prove any conjectures about what the Siamese network 'sees'. We suspect that due to LoSoCs producing close to linear paths in the latent space, and MNIST being a simple space of images, the LeNet feature extractor may be able to reconstruct what a linear path looks like in our own latent space.

These two findings from the exploration of paths in the latent space, and of our Siamese Network, in tandem point to the optimization procedure succeeding in maintaining *some* notion of fundamental similarity for MNIST images, though this may be abstract to humanly understand. This goes to show that fundamental similarity by itself is not sufficient for human-understandable streams of counterfactuals. It is, however, an encouraging result for us to expand and streamline our approach towards CXR-Age on chest X-rays.

4.3 Method Practicality

It is clear from our experimentation that the foremost determining factor for the success of the optimization technique is the quality of the latent space of the autoencoder. In our first experimentation with chest X-rays, as we discussed previously, moving in the latent space produced nonsense X-rays, rendering the strategy useless. In the MNIST domain, our VAE was successful in that it allowed for local traversal around the latent space while still generally producing sensible digits, but it didn't *always* succeed in producing plausible digits. Further exploration of a metric for plausibility could greatly enhance our method, as it could be used not only as an evaluative metric, but as a penalty term during the `LoSoCs` process. Below we show two examples of out-of-distribution `LoSoCs`-outputs, which we like to call *demon numbers*:



(a) Generated 5-digit from `LoSoCs`.

(b) Generated 2-digit from `LoSoCs`.

Figure 12: Examples of out-of-distribution LoSoCs-outputs generated from the latent space of our VAE.

While the method succeeded in its goal of producing streams of counterfactual images, the process of finding or creating a VAE that allows the method to be successful is a task which requires great effort. We believe the method may be useful, but it would best be reserved for cases where we stand to potentially gain information from the model, like in the CXR-Age case, where understanding how the model makes its decisions could result in interesting clinical insights.

4.4 Future Work

We would like to continue this research in a structured way beyond the scope of this class in preparation for a potential workshop paper. To this end, we want to continue to:

1. Consider the plausibility metric further, perhaps by carefully training a classifier to detect in-distribution decoded images versus out-of-distribution decoded images. The implementation for such a plausibility metric poses an interesting problem. We theorized that we could implement this by, say, manually labeling points sampled from the autoencoder as in- and out-of-distribution and training a classifier to predict based on those labels.

2. More work on metrics for essential similarity would also be fruitful. While our metrics give us some idea about whether information is retained or lost during the optimization process, we do not know *what* information is retained. In a practical data domain, it could be that `LoSoCs` is retaining information that is not desirable. For example, in the chest X-ray setting, we may see positive results from an essential similarity classifier, but the X-rays produced do not look like they are from the same person and instead, say, have similar bone densities.

3. We would like to conduct a literature review of interpretability work on plausibility and metrics for plausibility.

4. We also wish to train a VAE for chest radiographs that will allow us to port `LoSoCs` to the `CXR-Age` domain. We think it would also be helpful if there were quantitative measures of the 'quality' of a latent space for this optimization process, so we could avoid the trial and error or even optimize for the specific qualities we are looking for during the VAE training process.

4.5 References

References

Raghu, Vineet, Jordan Weiss, Udo Hoffmann, et al. (Nov. 2021). "Deep Learning to Estimate Biological Age From Chest Radiographs". In: *JACC: Cardiovascular Imaging* 14.11, pp. 2226–2236. DOI: 10.1016/j.jcmg.2021.01.008. URL: https://doi.org/10.1016/j.jcmg.2021.01.008.

Seah, Jarrel et al. (2018). *Generative Visual Rationales*. arXiv: 1804.04539 [cs.CV]. URL: https://arxiv.org/abs/1804.04539.

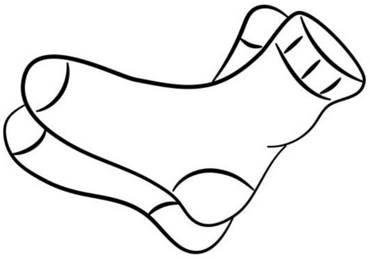


Figure 13: Visual of `LoSoCs` philosophy.