Inverting MNIST Neural Networks for Image Generation

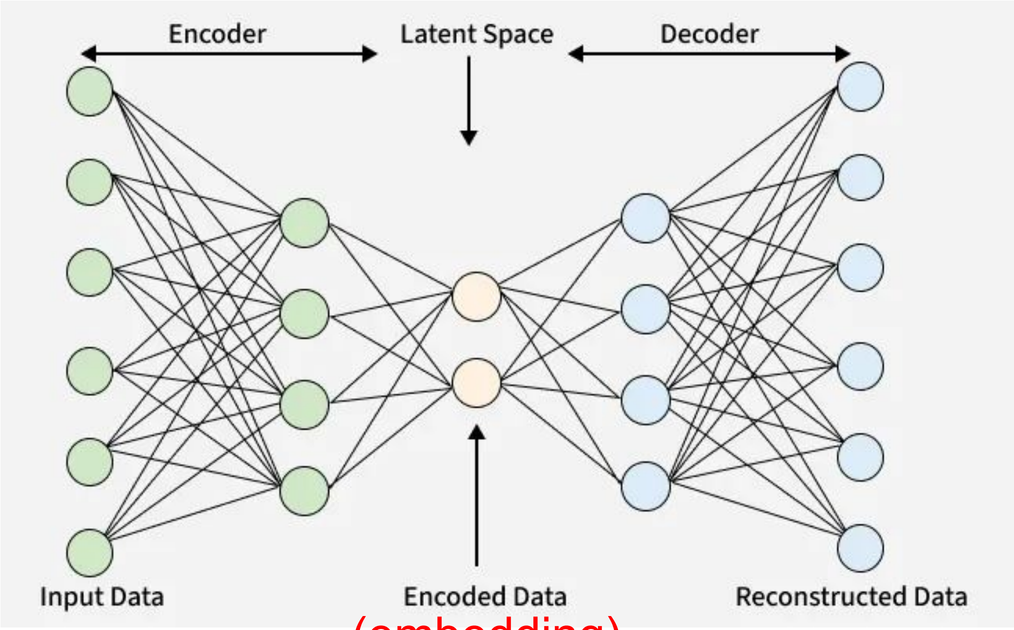
Eric Guo, Yihao Wang

Introduction

- Image generative AI has exploded in popularity
- Too computationally expensive & unintuitive
- Neural networks can be inverted for image generation
- We will invert a NN trained on the MNIST dataset

Related Work: Autoencoders

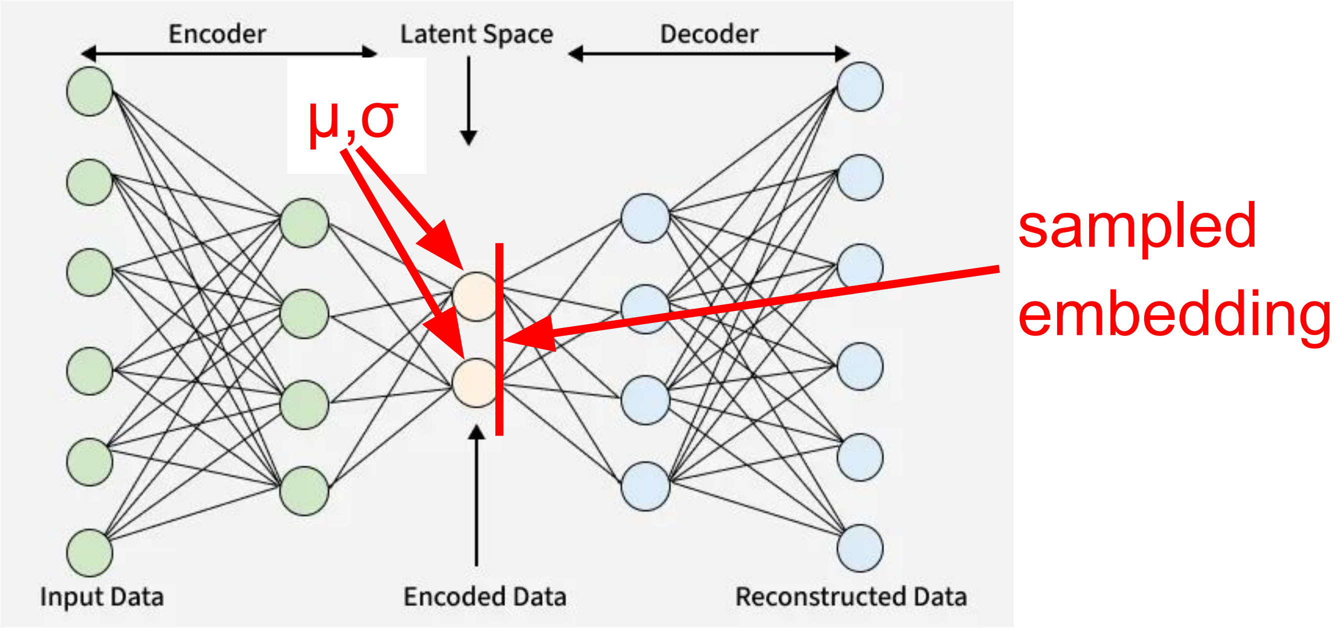
- Creating “embeddings” for high-dimensional data
- New embeddings can generate new data



(embedding)

Related Work: Variational Autoencoders (VAEs)

- Encoder outputs probabilistic distributions
- Reduces overfitting & increases stability



Related Work: Diffusion Models

- Trained by slowly adding noise to images
- Model learns to reverse noise

Dataset and Features

- MNIST dataset from torchvision
- Instances: 28x28 black-and-white images
- Labels: integers

Label = 4

Methods

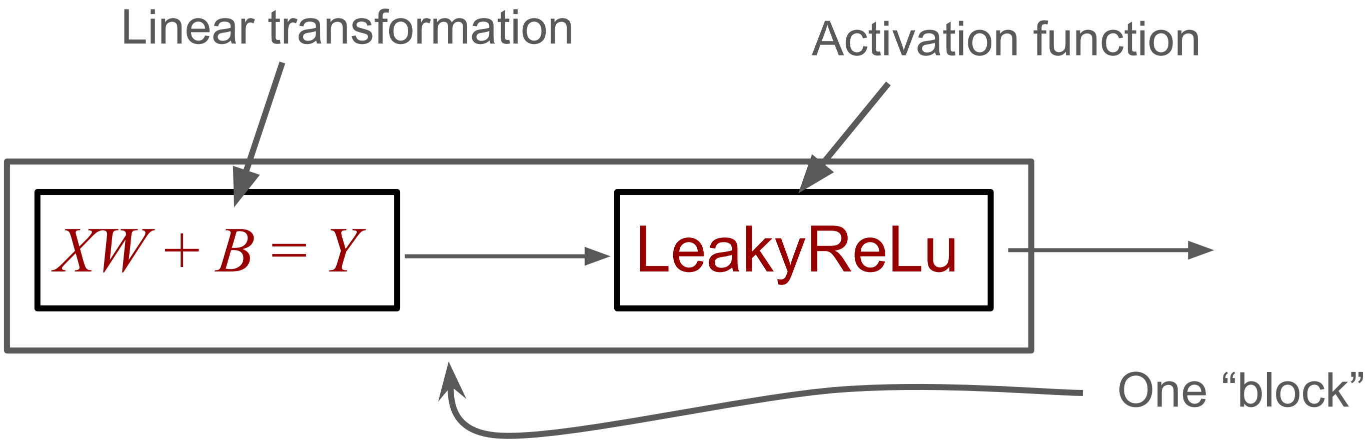
Neural network layers can be represented as

$$XW + B = Y$$

$$X = (Y - B)W^{-1}$$

- Activation functions must be invertible
- Use LeakyReLU
- W must be square

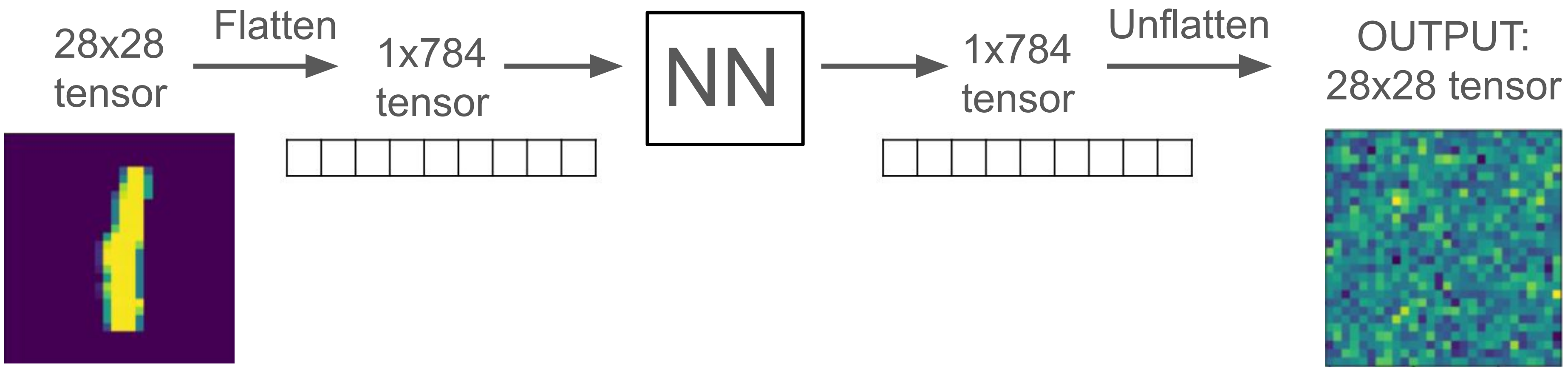
Methods



Train 3 models: Base, Small, NoActivations

- Base model: 4 blocks
- Small model: 1 block
- NoActivations model: 4 linear transformations

Methods

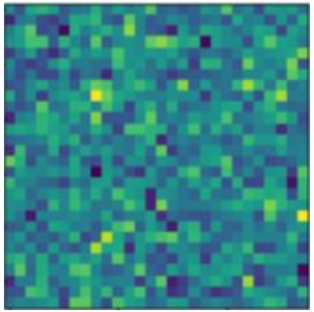

28x28
tensor
Flatten
1x784
tensor
NN
1x784
tensor
Unflatten
OUTPUT:
28x28 tensor

Methods

All pixels appear sampled from a Gaussian distribution (normal distribution)

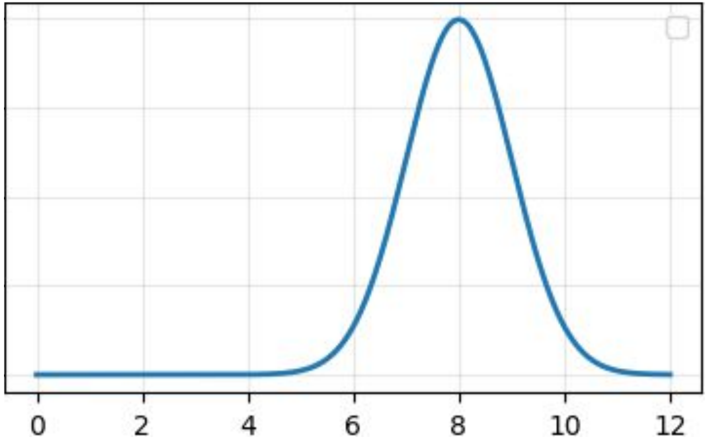
- Mean: μ = label * 2
- Variance: $\sigma^2 = 1$

OUTPUT:
28x28 tensor



ex. 4 image

$\mu = 2*4=8; \sigma^2=1$



Methods

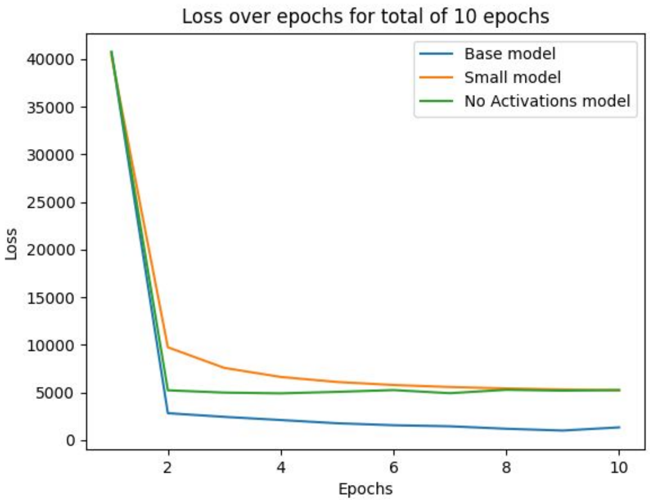
Loss function needs to consider both **mean** and **variance**

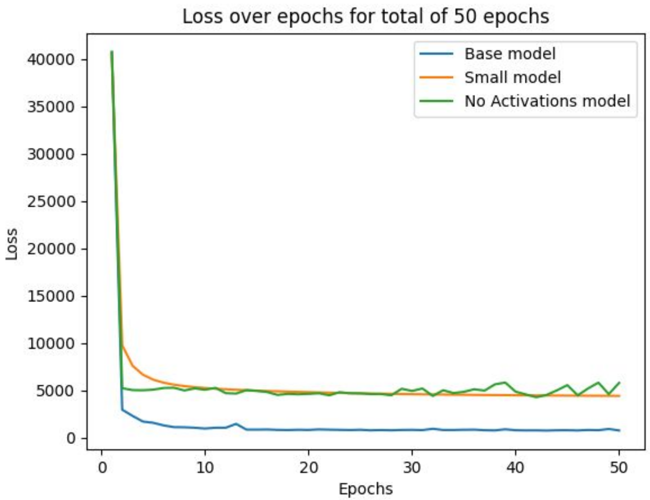
$$\text{Loss(N)} = \boxed{\frac{|N|\ln(2\pi)}{2} + \frac{|N|}{2}\sum (n-2a)^2} + \boxed{(\text{Variance}(N) - 1)^2}$$

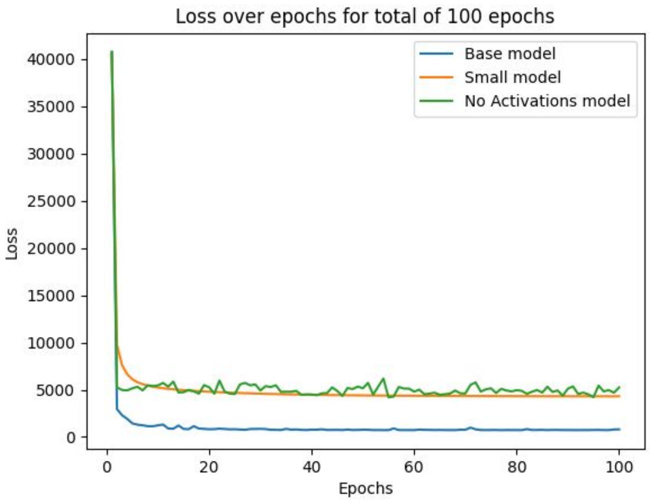
Negative Log Likelihood (NLL)
Ensure μ = 2*label
& output is noisy

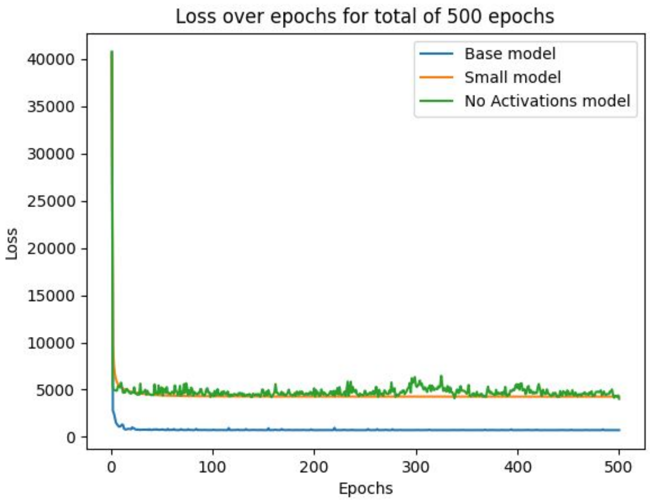
Ensure outputs vary &
don't collapse,
variance converges to 1

Results and Discussion

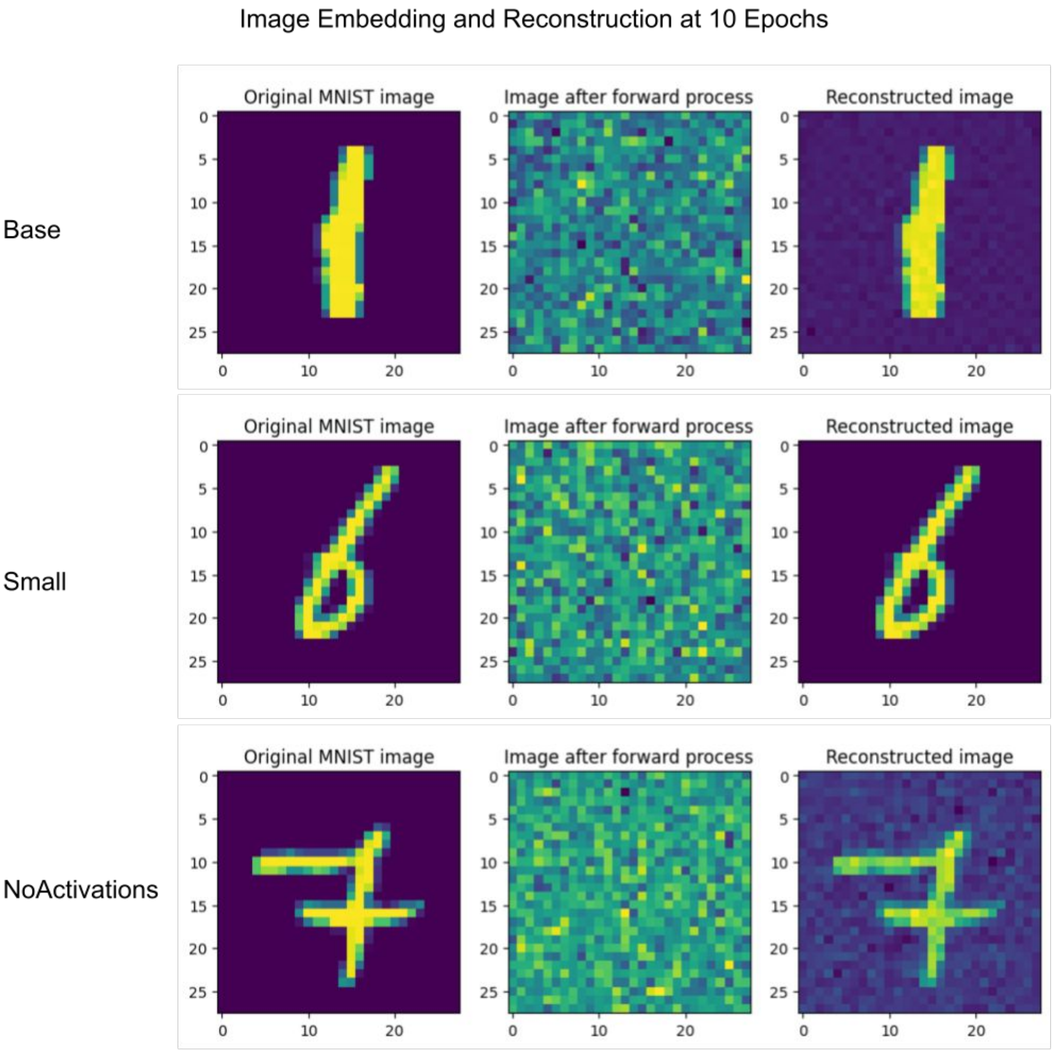








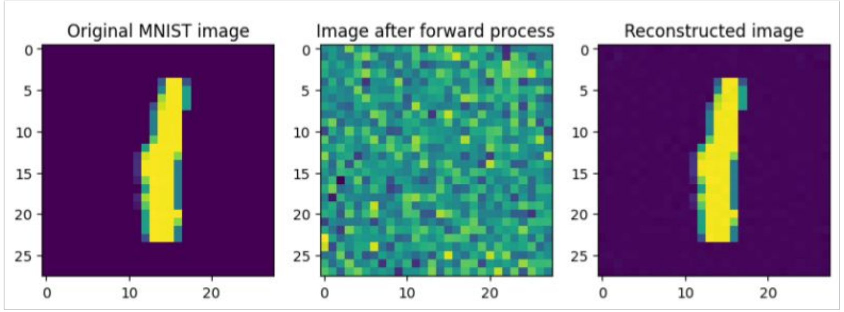
Results and Discussion



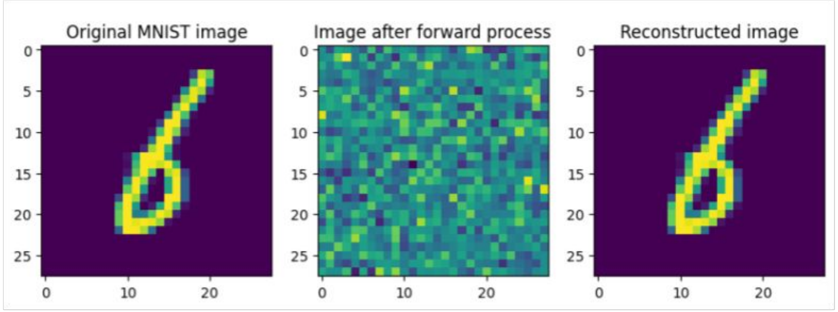
Results and Discussion

Image Embedding and Reconstruction at 50 Epochs

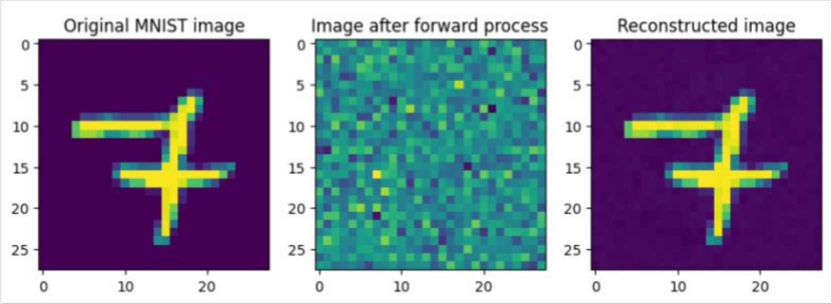
Base



Small



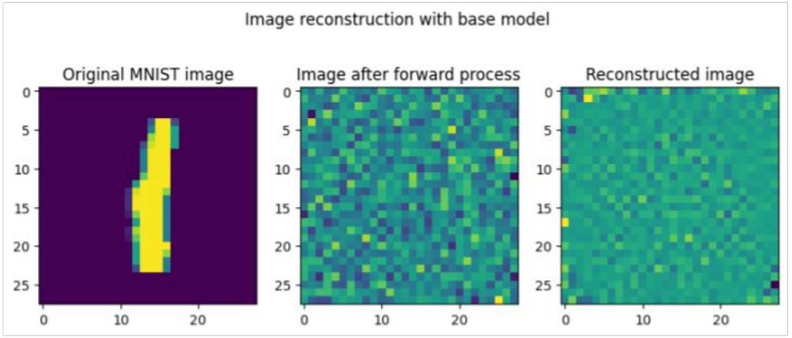
NoActivations



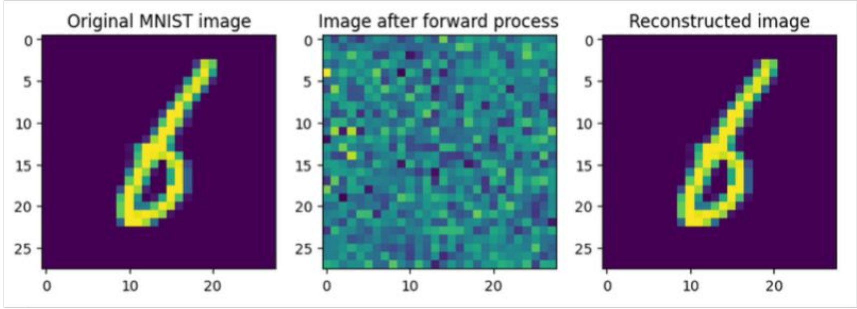
Results and Discussion

Image Embedding and Reconstruction at 100 Epochs

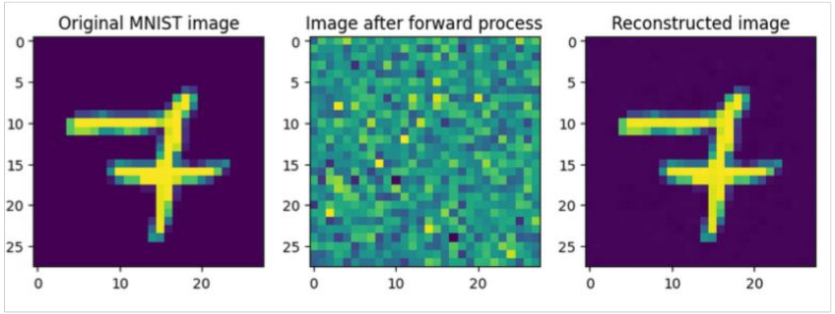
Base



Small



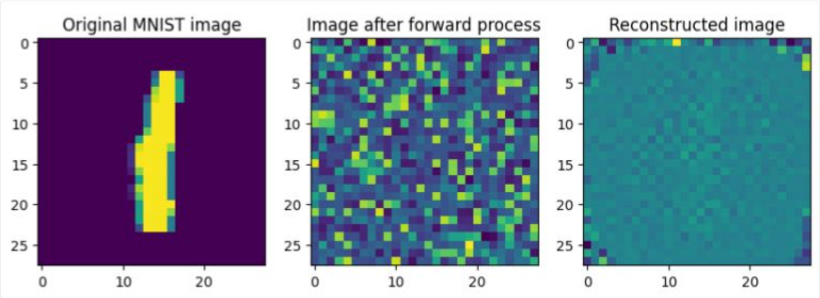
NoActivations



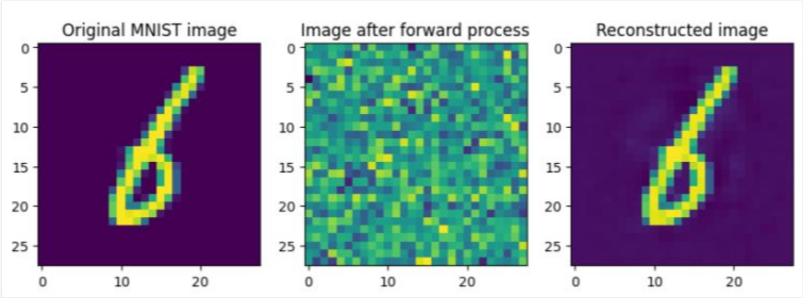
Results and Discussion

Image Embedding and Reconstruction at 500 Epochs

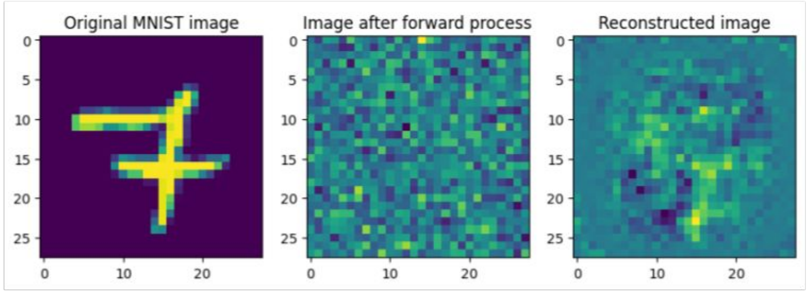
Base



Small



NoActivations



Conclusion & Future Work

- Inverting neural networks for image generation is infeasible
- NN constraints and sensitivity limit image generation

Future research should focus on

- slowly producing coherent images from noise
- inherently probabilistic models
- better loss function, e.g. KL-Divergence

References

[1] ErSE 222 Machine Learning in Geoscience Course, “Dimensionality reduction — autoencoders,” https://dig-kaust.github.io/MLgeoscience/lectures/13dimred/autoencoders, 2025, accessed: 2026 − 01 − 20.

[2] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” arXiv preprint arXiv:1312.6114, 2013.

[3] ErSE 222 Machine Learning in Geoscience Course, “Generative modelling and variational autoencoders,” https://dig-kaust.github.io/MLgeoscience/lectures/14v ae/, 2025, accessed: 2026 − 01 − 20.

[4] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” arXiv preprint arXiv:2006.11239, 2020, accessed: 2026-01-20. [Online]. Available: https://arxiv.org/abs/2006.11239