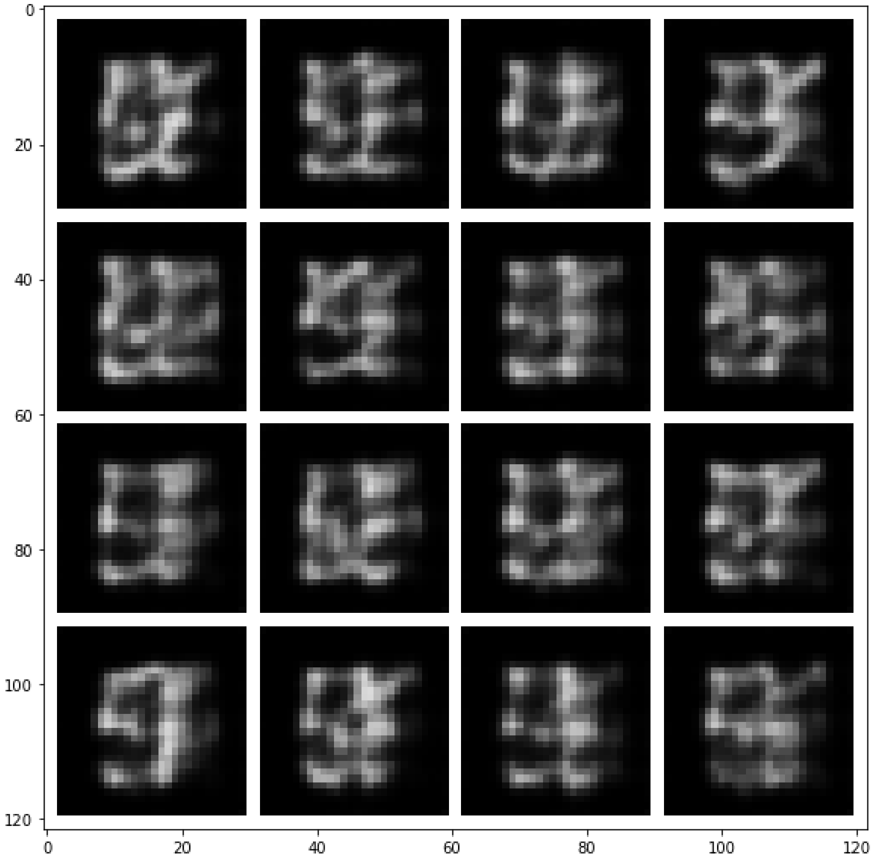
CSCI 566 – Assignment 2 – Problem 1

1. Auto-encoder samples

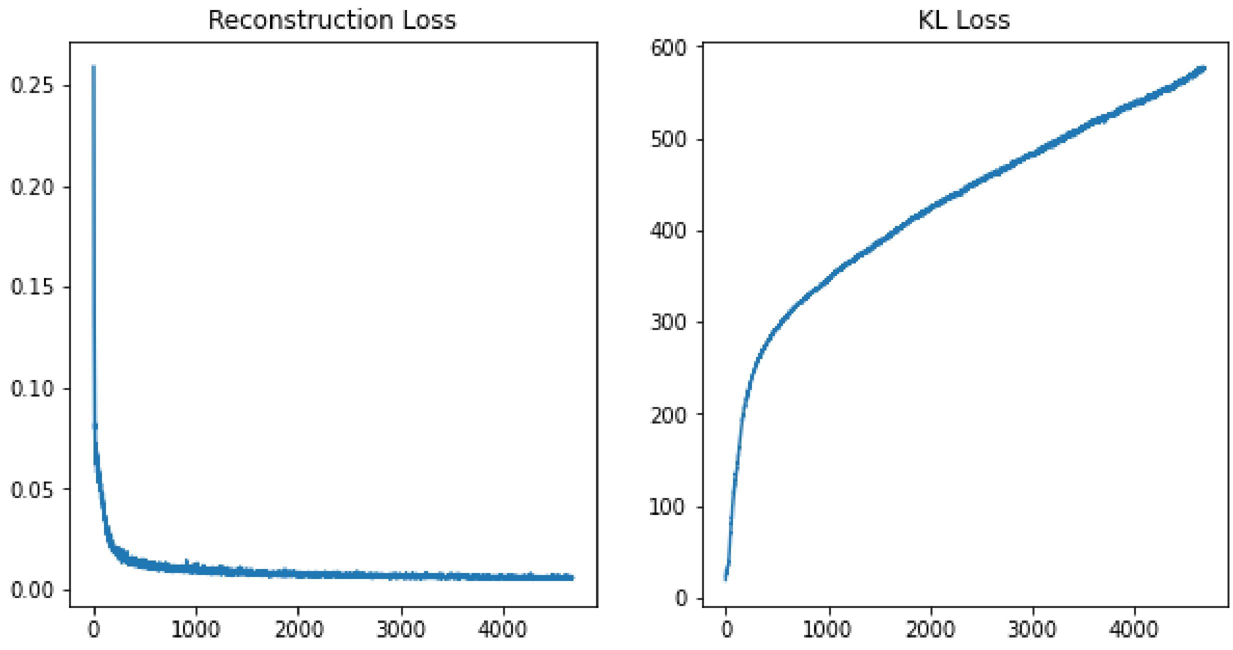


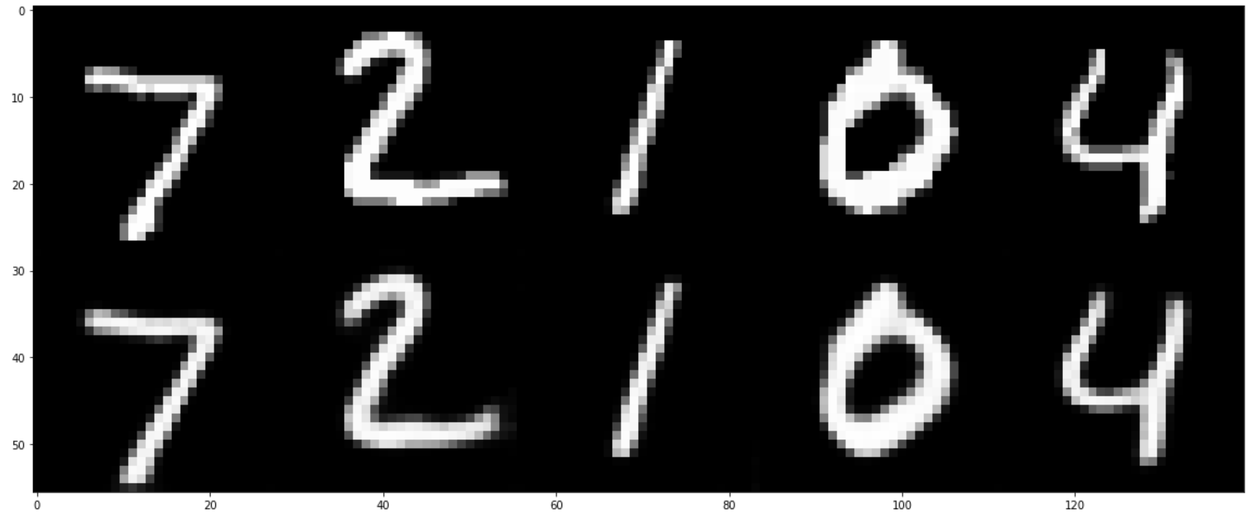
AE sampling inline question answer

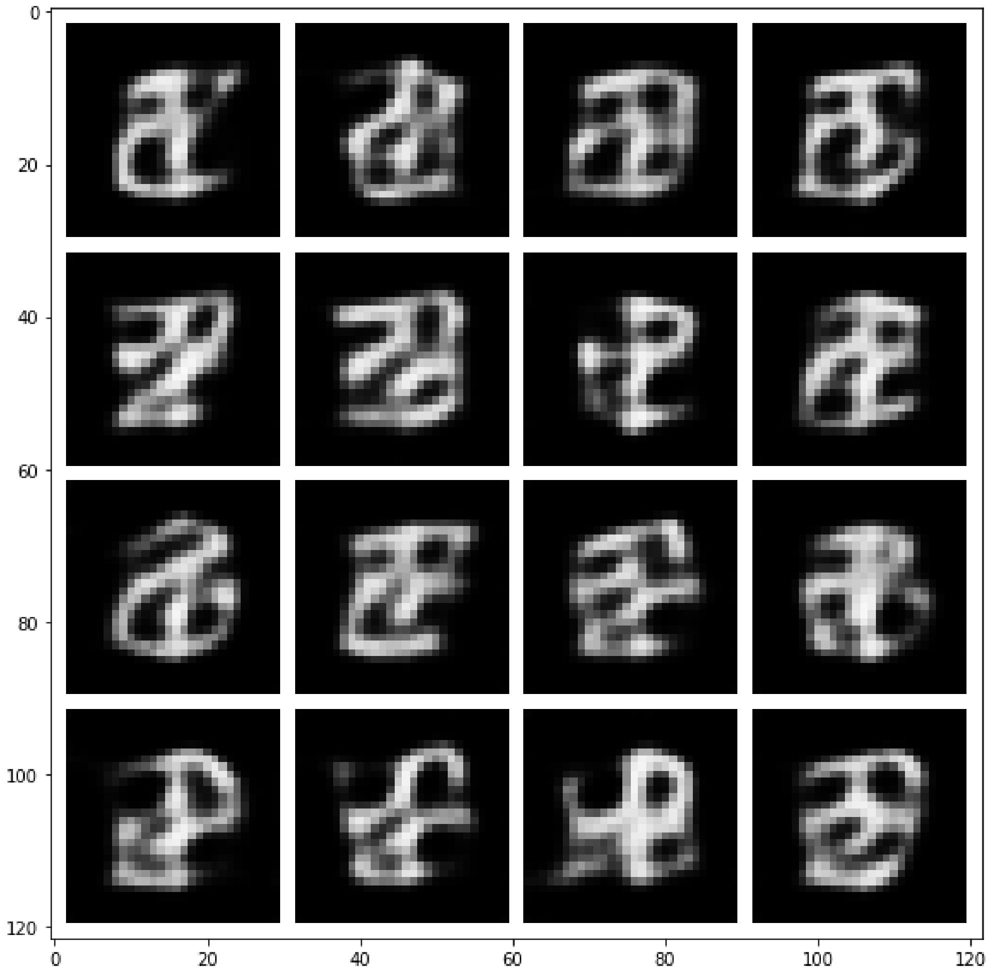
Answer: Our reconstruction of the data in the 'Verifying Reconstructions' section is nearly the same which indicates that our network is able to create a lower level representation of the images at the bottleneck and decode it back to almost the original image. The output plotted above after decoding the sample embeddings from a diagonal unit Gaussian model are hard to recognize as handwritten digits. In my opinion, this has happened due to the embeddings which were sampled from the prior distribution not being representative of the lower level representation that the decoder network would have expected or similar to what it learned during training and has thus led to generating blurry images which do not make sense. The latent space that autoencoders convert their inputs to and where the embeddings lie may not be continuous. If the space has discontinuities and we sample from there, then the decoder generates an output which does not make sense because during training it did not encounter any embeddings from that region of the latent space.

2. **VAE training curves, reconstructions and samples for:**

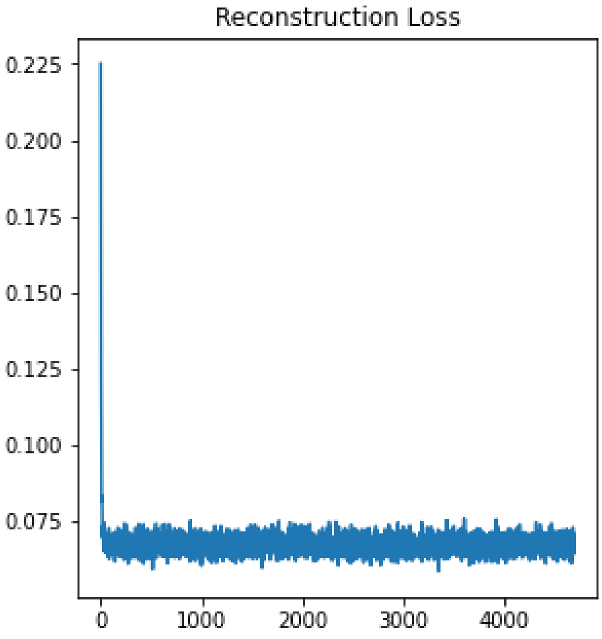
   a. β=0

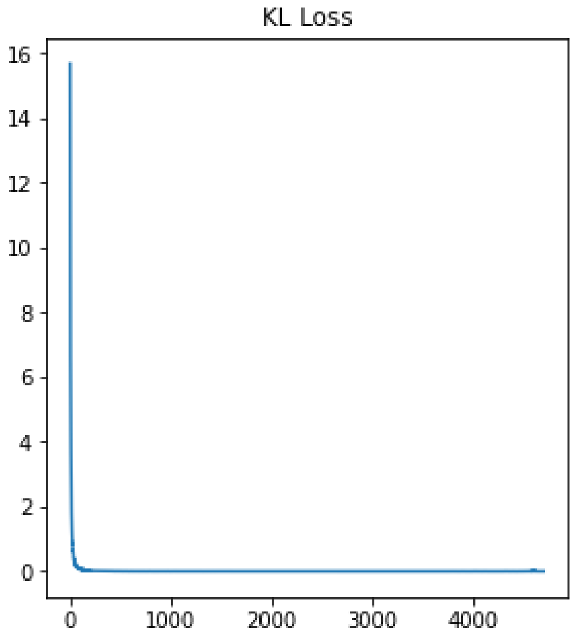


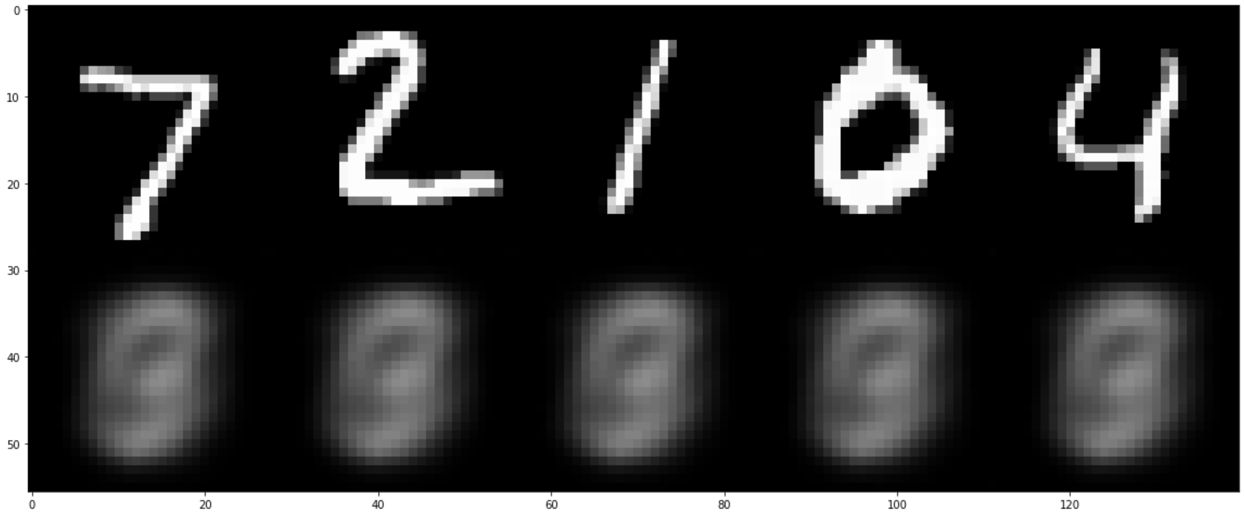


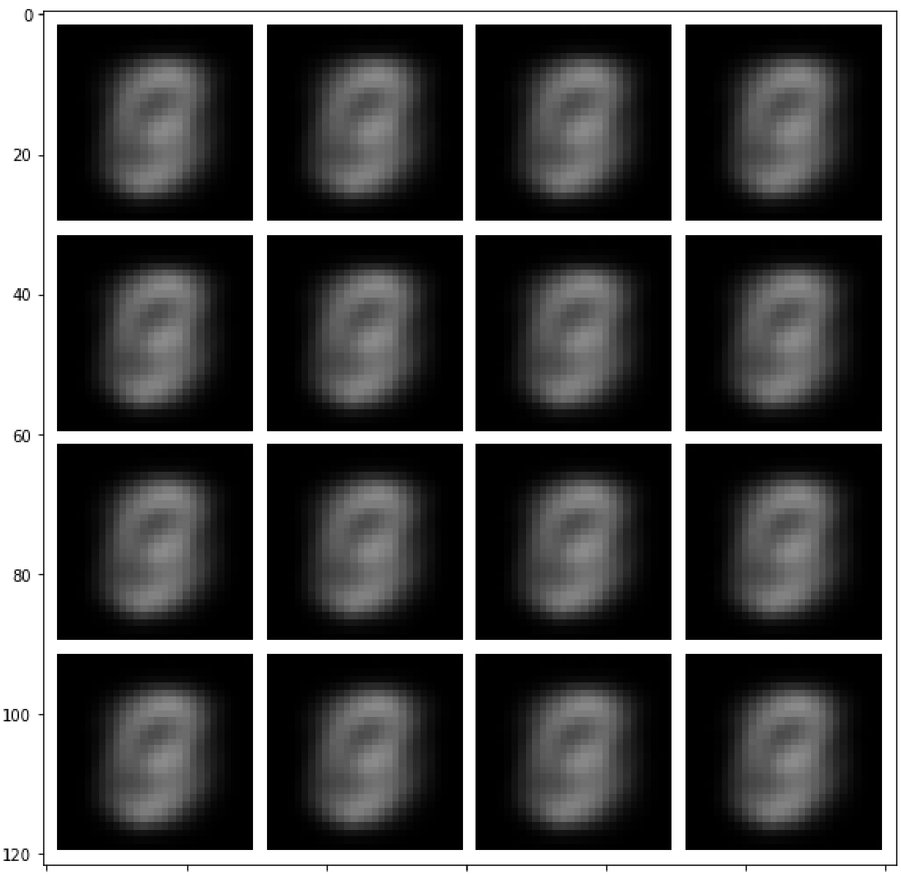

0
20
40
60
80
100
120
0
20
40
60
80
100
120

b. β=10

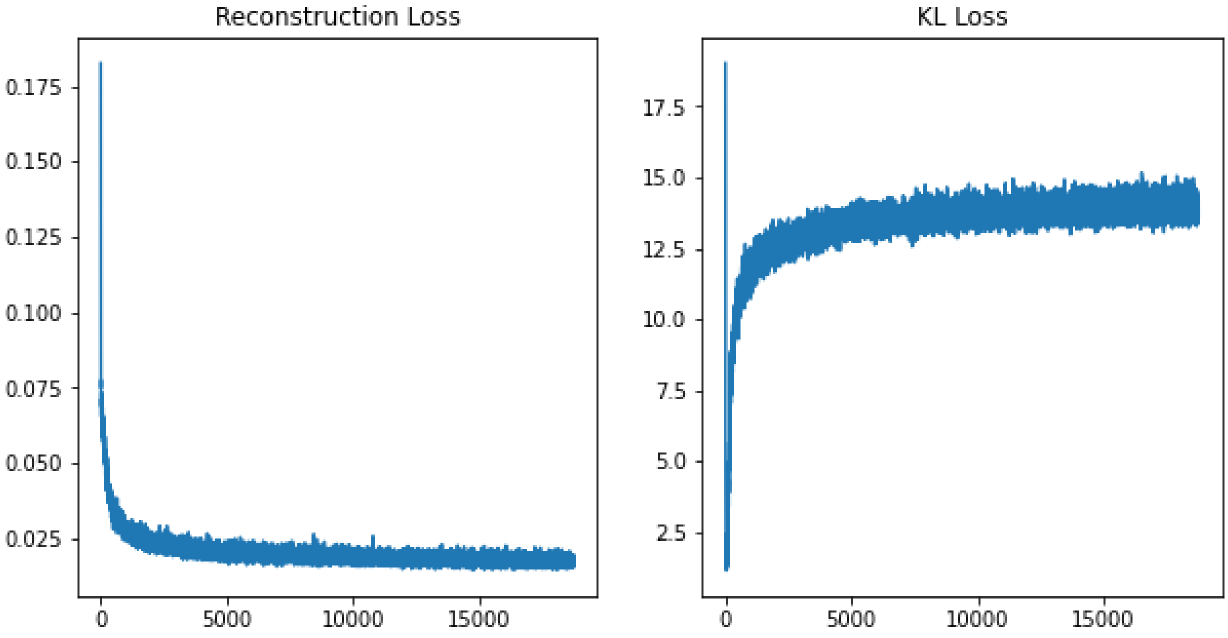

Reconstruction Loss
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0
1000
2000
3000
4000

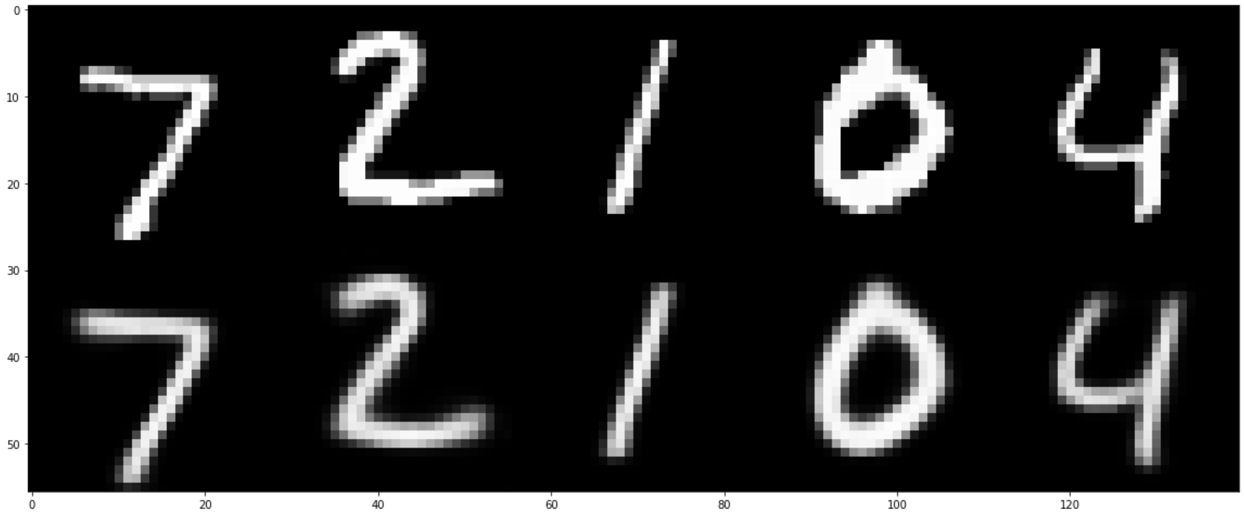

KL Loss
16
14
12
10
8
6
4
2
0
0
1000
2000
3000
4000

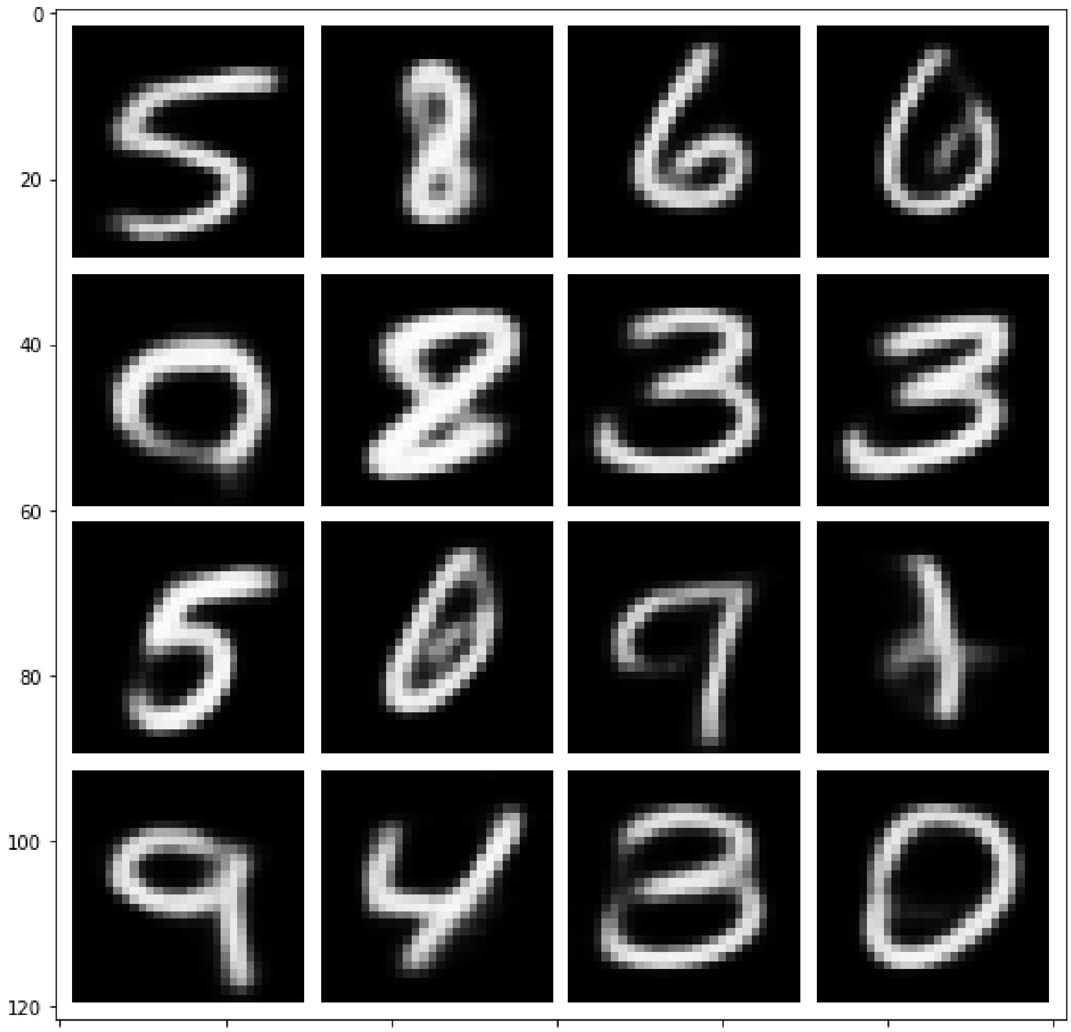

0
10
20
30
40
50
0
20
40
60
80
100
120


0
20
40
60
80
100
120
0
20
40
60
80
100
120

c. β=0.001


Reconstruction Loss
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0
5000
10000
15000
KL Loss
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
5000
10000
15000


0
10
20
30
40
50
0
20
40
60
80
100
120


0
20
40
60
80
100
120
0
20
40
60
80
100
120

3. Answers to all inline questions in VAE section

   a. **Answer:** With β=0, we can observe that the reconstructions are extremely similar to the original image however the samples do not look like any handwritten digits and look randomly generated. It essentially reduces the variational autoencoder loss to the original autoencoder loss by ignoring the KL loss component and only considers the reconstruction loss. This theory is supported by the reconstructions and samples which we observe. These look alike to the autoencoder samples generated earlier. Since we only consider the reconstruction loss and ignore the regularization term, we move away from the prior distribution and thus the KL loss increases as well.

   b. **Answer**: With β=10, it can be observed that the reconstructions and samples are blurred and noisy. They do not look like any handwritten digits. The reason this happens is because of the tradeoff between reconstruction and regularization in the variational autoencoder loss. With such a high beta value the network will decrease the value of the KL loss, however it loses out on the reconstruction quality of the images. A very high β value introduces a stronger constraint on the latent bottleneck and encourages stronger disentanglement in the generative model. This limits the overall representation capacity of z thus causing reconstruction issues for the VAE.

   c. **Answer**: A good result for a well-tuned VAE with the appropriate β value would have the following characteristics:

      i. For the reconstructions - we can expect to observe better reconstructed outputs which look extremely similar to the original image.

      ii. For the samples - we can expect to observe new images generated which resemble handwritten digits. A well-tuned β value would ensure that the embeddings which the model learns are not too far away from the prior distribution and can also generate images based on sampled embeddings from the distribution.

> **Inline Question: What can you observe when setting $\beta = 0$? Explain your observations! [3pt]**
> (please limit your answer to <150 words)
> **Answer:** With $\beta = 0$, we can observe that the reconstructions are extremely similar to the original image however the samples do not look like any handwritten digits and look randomly generated. It essentially reduces the variational autoencoder loss to the original autoencoder loss by ignoring the KL loss component and only considers the reconstruction loss. This theory is supported by the reconstructions and samples which we observe. These look alike to the autoencoder samples generated earlier. Since we only consider the reconstruction loss and ignore the regularization term, we move away from the prior distribution and thus the KL loss increases as well.

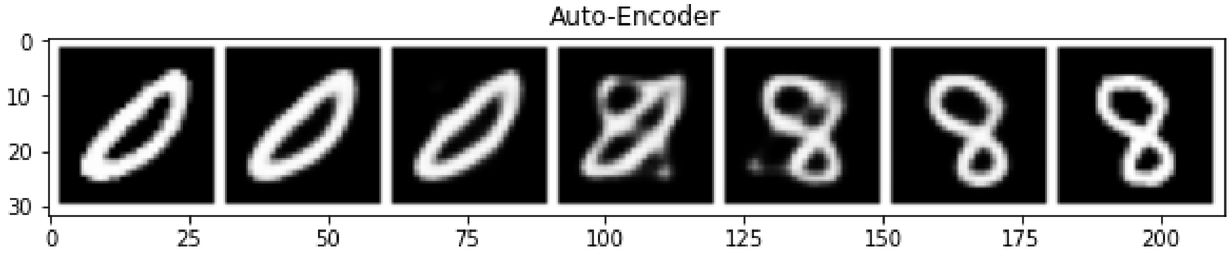
Let's repeat the same experiment for $\beta = 10$, a very high value for the coefficient. You can modify the β value in the cell above and rerun it (it is okay to overwrite the outputs of the previous experiment, but **make sure to copy the visualizations of training curves, reconstructions and samples for $\beta = 0$ into your solution PDF** before deleting them).

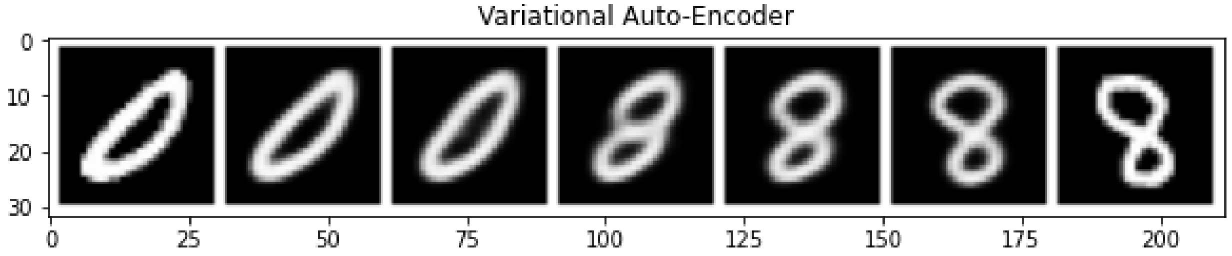
> **Inline Question: What can you observe when setting $\beta = 10$? Explain your observations! [3pt]**
> (please limit your answer to <200 words)
> **Answer**: With $\beta = 10$, it can be observed that the reconstructions and samples are blurred and noisy. They do not look like any handwritten digits. The reason this happens is because of the tradeoff between reconstruction and regularization in the variational autoencoder loss. With such a high beta value the network will decrease the value of the KL loss, however it loses out on the reconstruction quality of the images. A very high β value introduces a stronger constraint on the latent bottleneck and encourages stronger disentanglement in the generative model. This limits the overall representation capacity of z thus causing reconstruction issues for the VAE.

Now we can start tuning the beta value to achieve a good result. First describe what a "good result" would look like (focus what you would expect for reconstructions and sample quality).

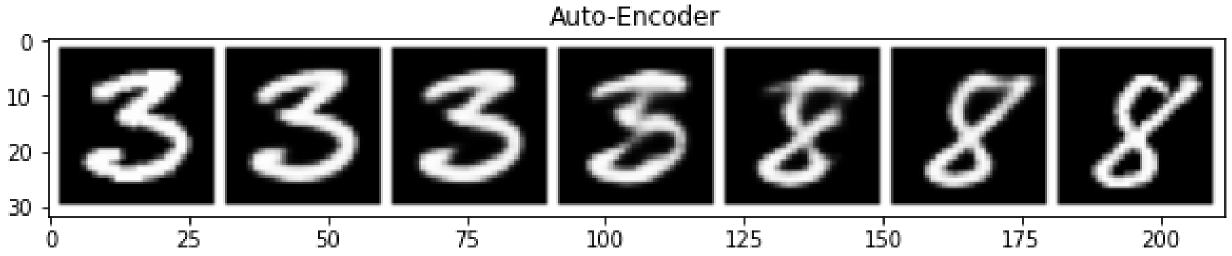
> **Inline Question: Characterize what properties you would expect for reconstructions (1pt) and samples (2pt) of a well-tuned VAE! [3pt]**
> (please limit your answer to <200 words)
> **Answer**: A good result for a well-tuned VAE with the appropriate β value would have the following characteristics:
>
> 1. For the reconstructions - we can expect to observe better reconstructed outputs which look extremely similar to the original image.
> 2. For the samples - we can expect to observe new images generated which resemble handwritten digits. A well-tuned β value would ensure that the embeddings which the model learns are not too far away from the prior distribution and can also generate images based on sampled embeddings from the distribution.

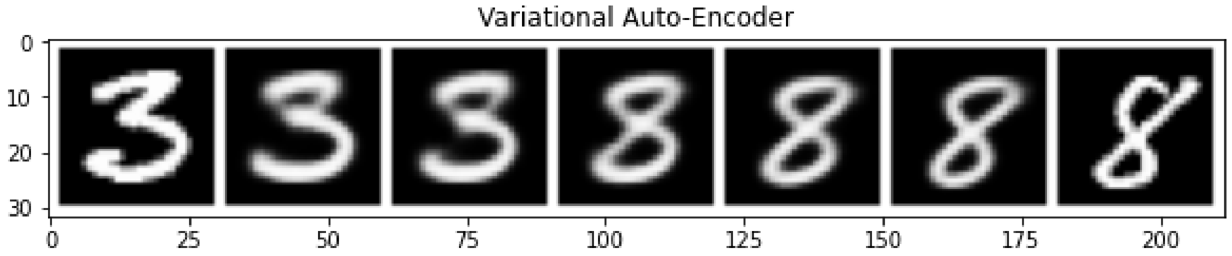
4. **Three representative interpolation comparisons that show AE and VAE embedding interpolation between the same images.**
   a. START_LABEL = 0
   END_LABEL = 8



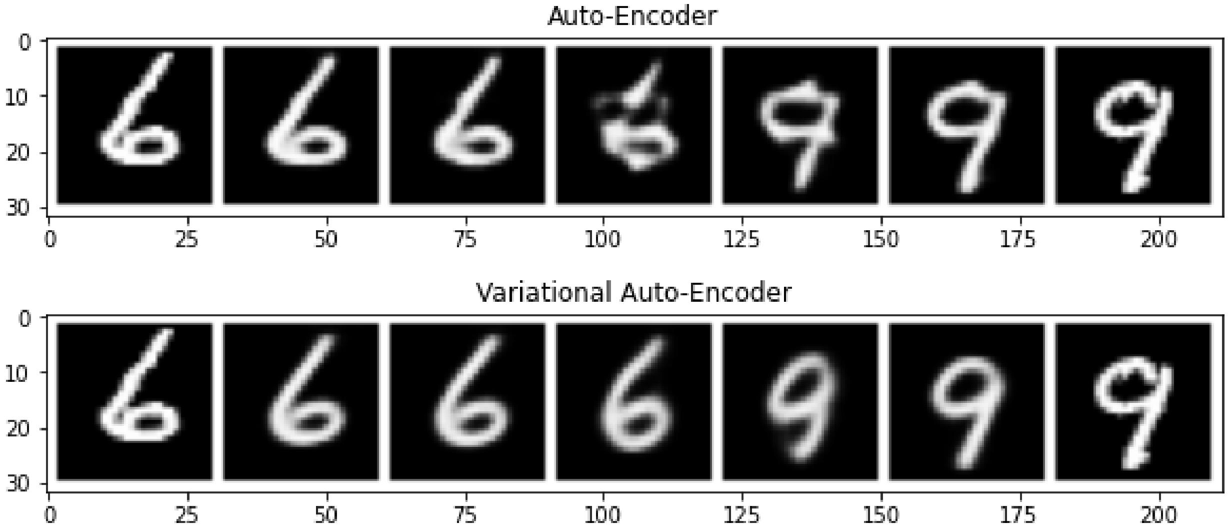


   b. START_LABEL = 3
   END_LABEL = 8





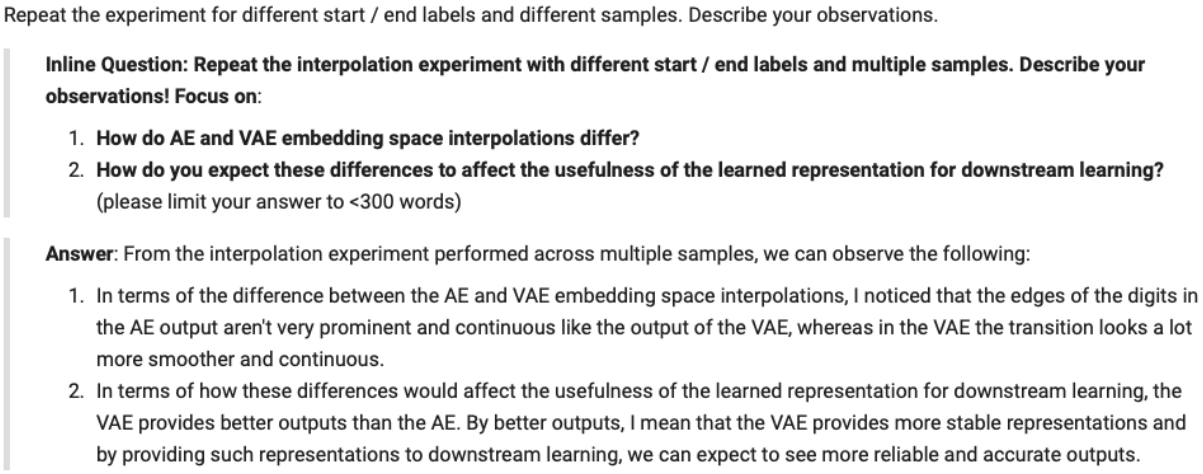
c. START_LABEL = 6
END_LABEL = 9



5. Answer to interpolation inline question

Answer: From the interpolation experiment performed across multiple samples, we can observe the following:

1. In terms of the difference between the AE and VAE embedding space interpolations, I noticed that the edges of the digits in the AE output aren't very prominent and continuous like the output of the VAE, whereas in the VAE the transition looks a lot more smoother and continuous.
2. In terms of how these differences would affect the usefulness of the learned representation for downstream learning, the VAE provides better outputs than the AE. By better outputs, I mean that the VAE provides more stable representations and by providing such representations to downstream learning, we can expect to see more reliable and accurate outputs.



Repeat the experiment for different start / end labels and different samples. Describe your observations.

> **Inline Question: Repeat the interpolation experiment with different start / end labels and multiple samples. Describe your observations! Focus on**:
>
> 1. **How do AE and VAE embedding space interpolations differ?**
> 2. **How do you expect these differences to affect the usefulness of the learned representation for downstream learning?** (please limit your answer to <300 words)

> **Answer**: From the interpolation experiment performed across multiple samples, we can observe the following:
>
> 1. In terms of the difference between the AE and VAE embedding space interpolations, I noticed that the edges of the digits in the AE output aren't very prominent and continuous like the output of the VAE, whereas in the VAE the transition looks a lot more smoother and continuous.
> 2. In terms of how these differences would affect the usefulness of the learned representation for downstream learning, the VAE provides better outputs than the AE. By better outputs, I mean that the VAE provides more stable representations and by providing such representations to downstream learning, we can expect to see more reliable and accurate outputs.