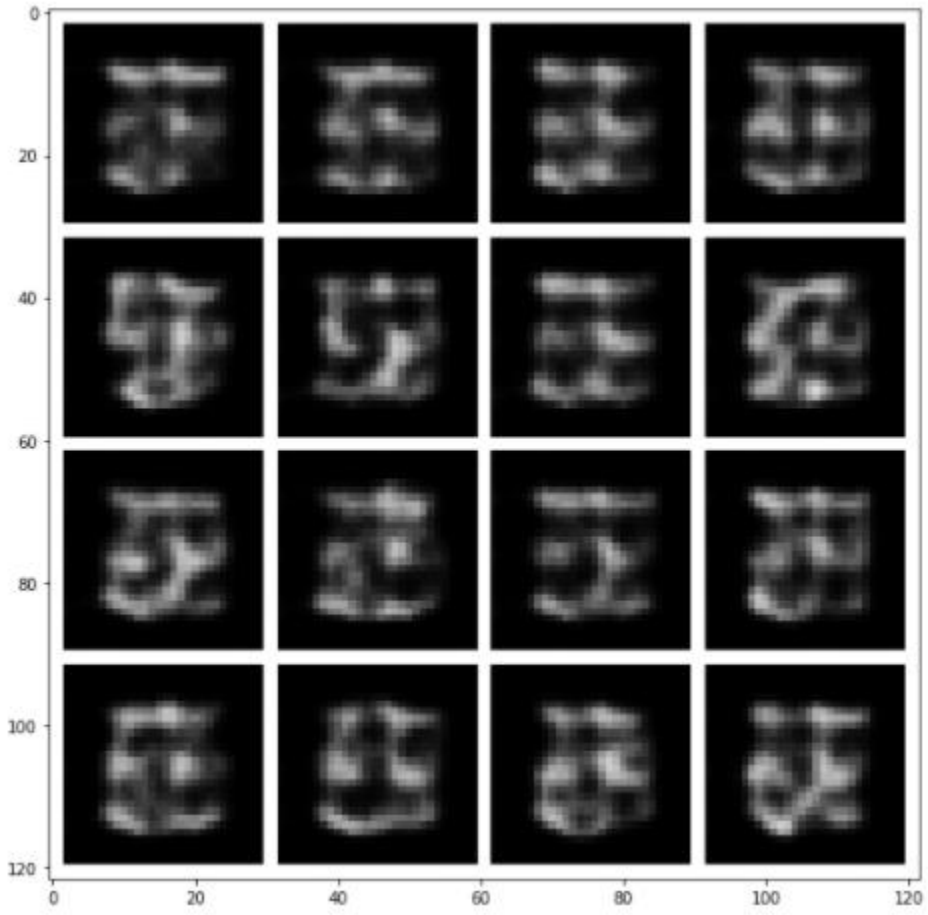
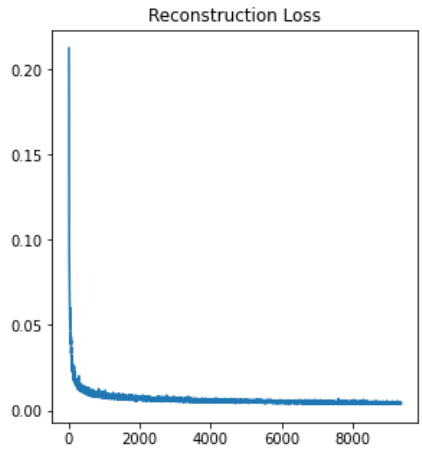
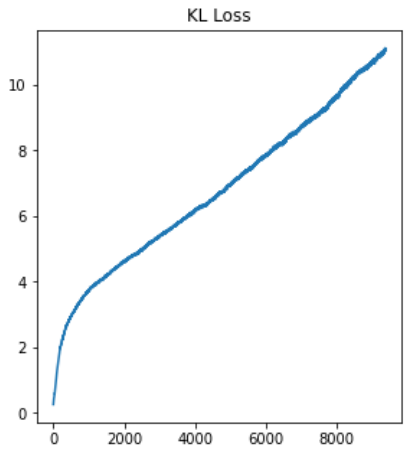
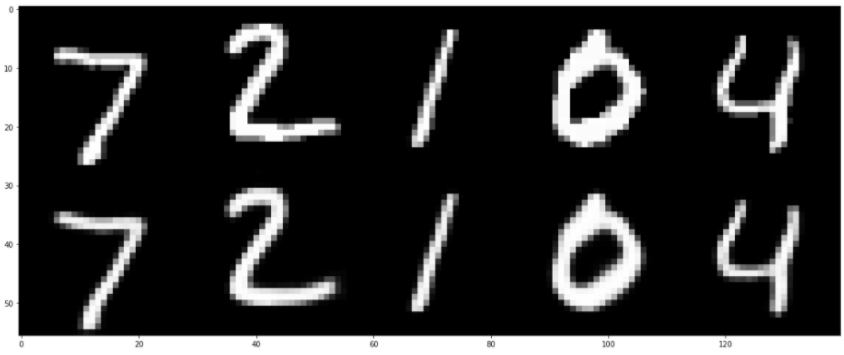
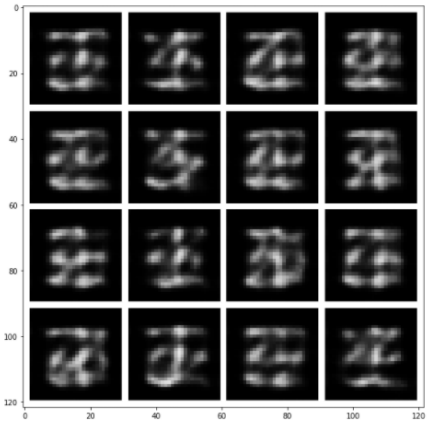


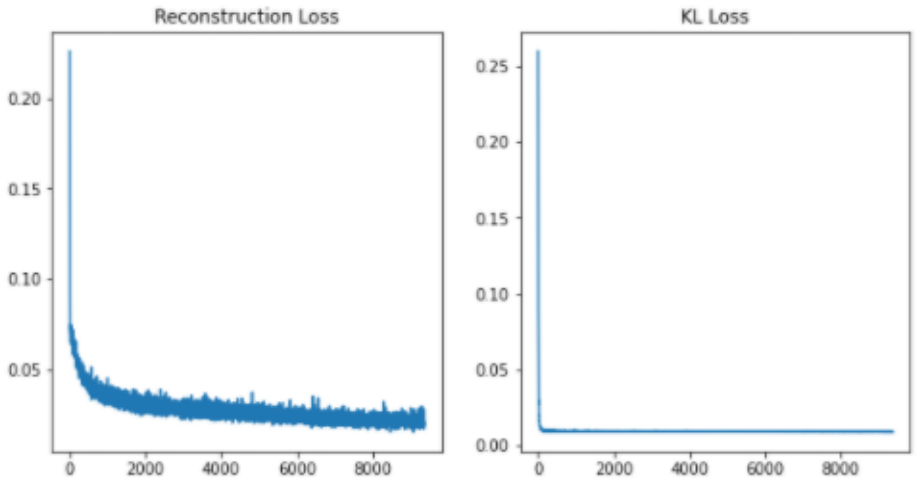
Inline Question: Describe your observations, why do you think they occur? [2pt]
(please limit your answer to <150 words)
**Answer: An interesting fact is when I increased the number of epoch, the images become more and more similar and hard to recognize or even unrecognizable. When the number of epoch is 10, you can tell nothing from the output image. I guess it's due to the limitation of prior distribution, overfitting and the points distribute are more closed when the reconstruction loss gets smaller, thus generative model will unlikely to generates point close to these points.

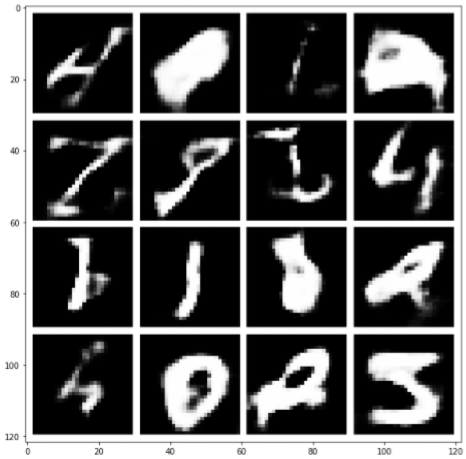




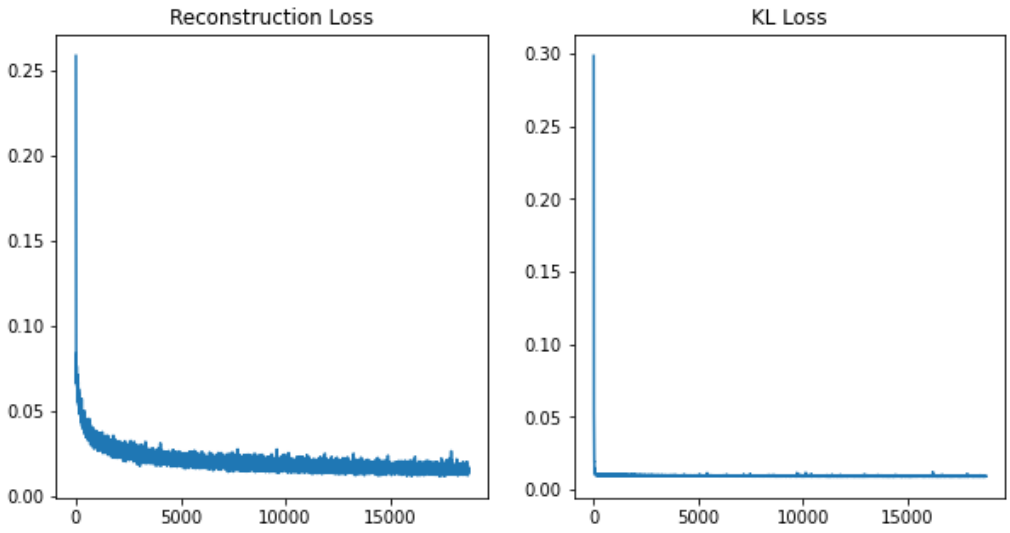




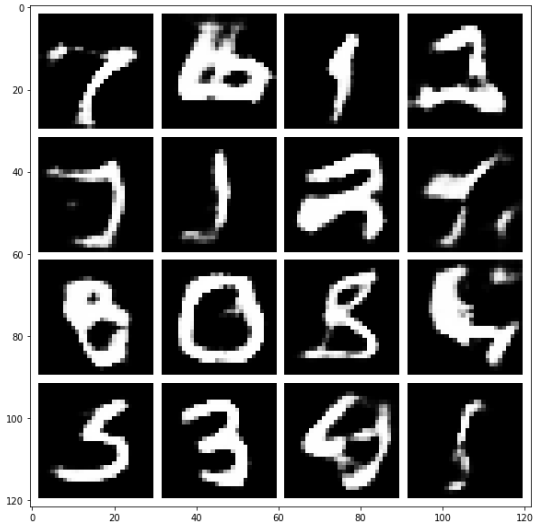

Reconstruction Loss
KL Loss
0.20
0.15
0.10
0.05
0.25
0.00
0
2000
4000
6000
8000



Best-tune beta = `6.77`




Press Esc to exit full screen



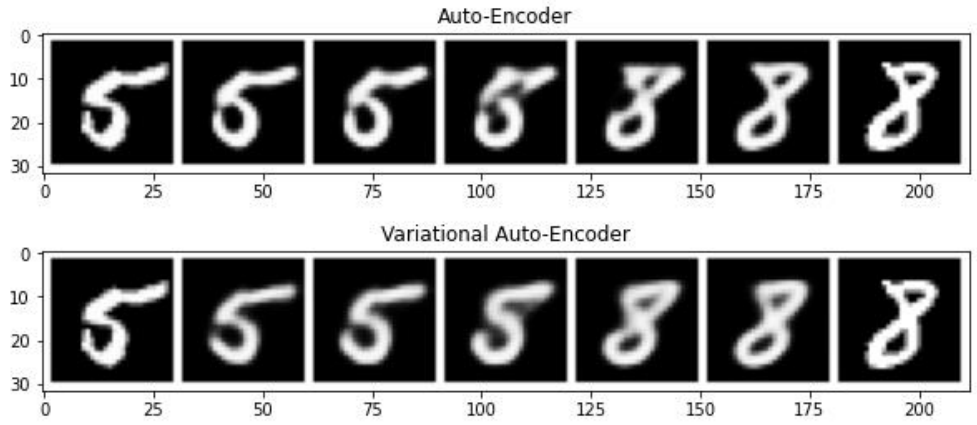
> **Inline Question: What can you observe when setting $\beta = 0$? Explain your observations! [3pt]**
> (please limit your answer to <150 words)
> **Answer: When $\beta = 0$, though the reconstruction loss gets really small very quickly, KL loss quickly increases. Our reconstruction result is almost perfect because our reconstruction loss really small. But the sampling result is unrecognizable, mainly because KL divergence tells us how well the probability distribution Q approximates the probability distribution P, if we set beta too low, KL is unconstrained, and distribution Q will not approximate P and thus our image cannot be good.

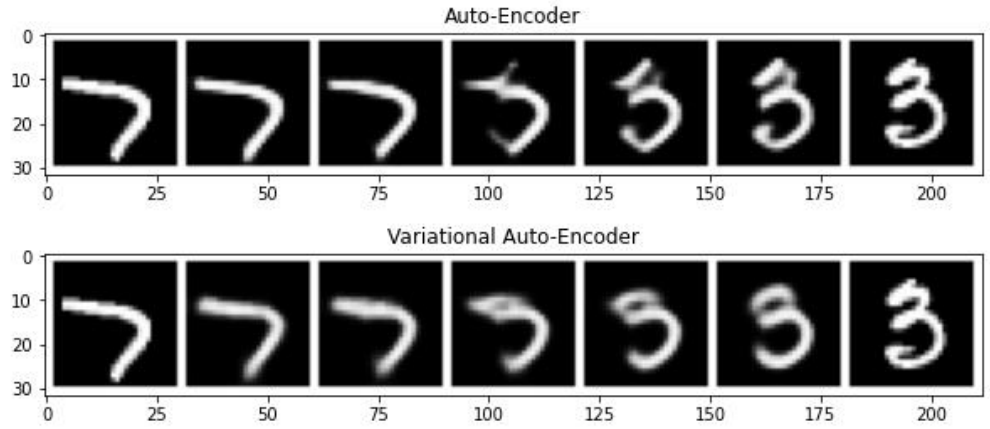
Let's repeat the same experiment for $\beta = 10$, a very high value for the coefficient. You can modify the β value in the cell above and rerun it (it is okay to overwrite the outputs of the previous experiment, but **make sure to copy the visualizations of training curves, reconstructions and samples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

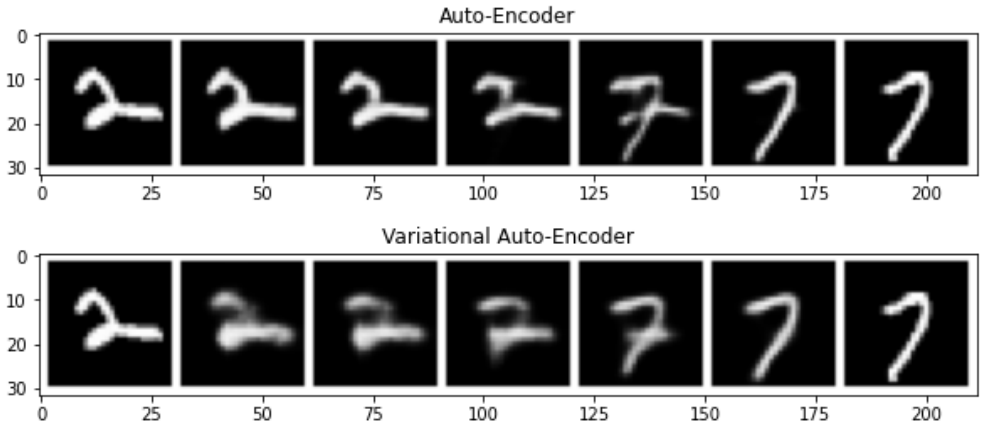
> **Inline Question: What can you observe when setting $\beta = 10$? Explain your observations! [3pt]**
> (please limit your answer to <200 words)
> **Answer**: Once we make $\beta = 10$, both our reconstruction loss and kl loss decrease very quickly. The difference between reconstruction loss comparing to $\beta = 0$ is our reconstruction loss has a higher variance after reconstruction loss is below 0.05. Our reconstruction image is almost perfect and our sampling result is much better, we can recognize most of them with high probablity. The main reason is our reconstruction and kl loss are both really low.

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**: I think a "good reasult" should have a sampling result that is easily recognizable with high confidence and without too much blurred and a perfect reconstruction result.







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:

1. Some digits in AE feel like they are seperated by some cuts and some digits in VAE contains are kind of blurred.
2. Since VAE has KL divergence loss, it has better proximate distribution. That's why digits in AE may seems seperated by some cuts and digits in VAE are more like a real digits. However, VAE will have some higher variance for reconstruction loss, which explains the blurredness in the VAE images.