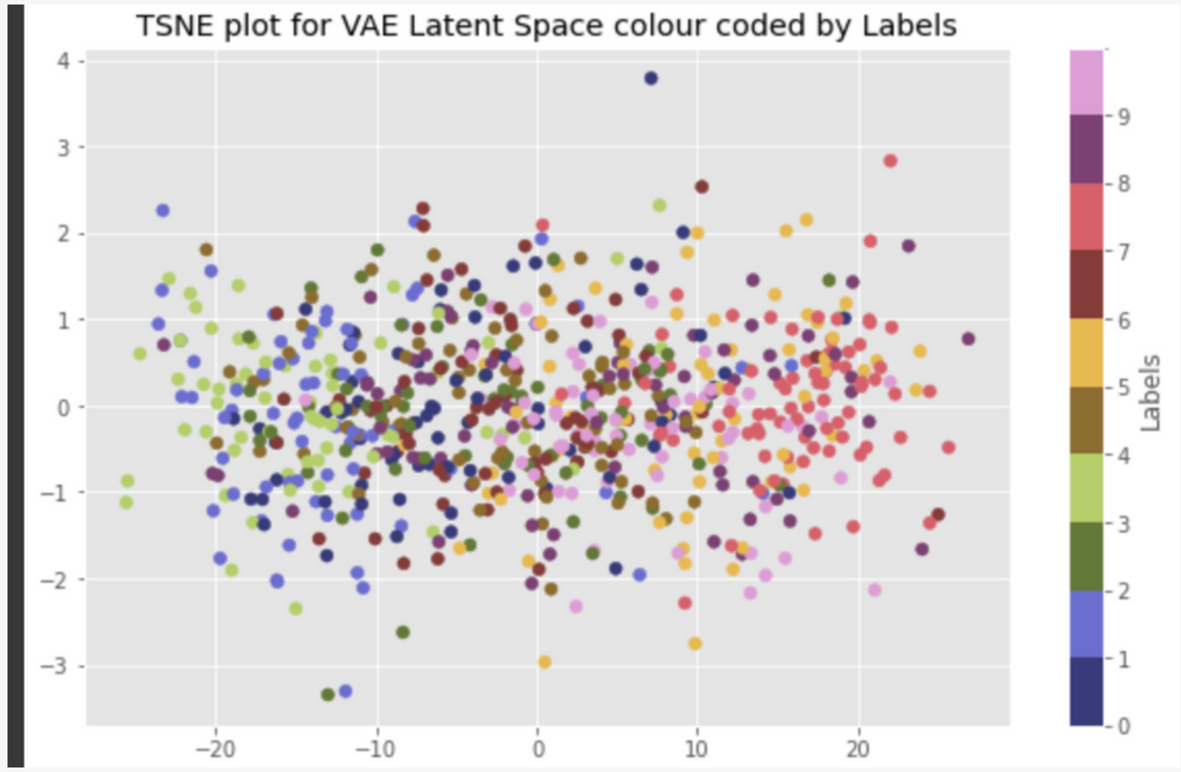
Disentanglement in VAEs

Yash Manish Shimpi
IIT (BHU), Varanasi

Overview

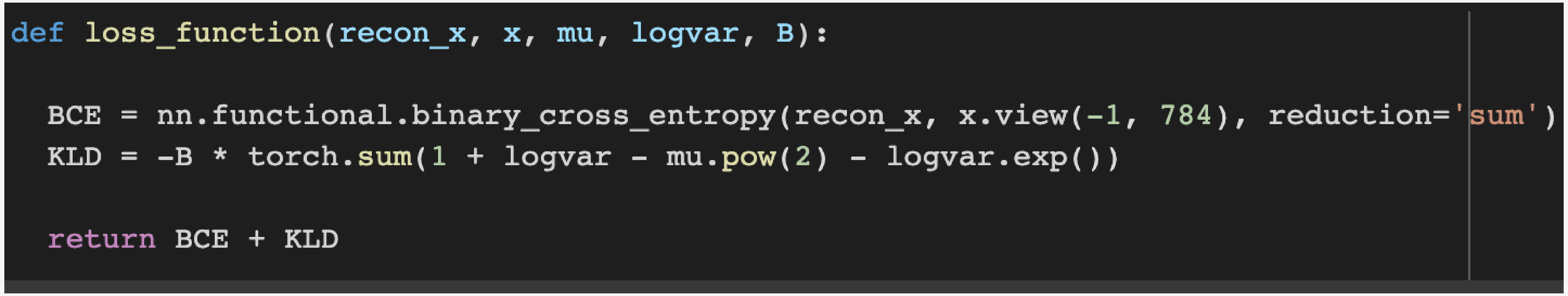
- In VAEs, latent space variables get entangled; changing one variable can affect other dimensions.
- Disentanglement allows to independently change a dimension in the latent space vector without compromising the reconstruction quality of the VAE.
- Disentanglement can be done by modifying the KLD loss.
- Our aim is to disentangle these dimensions such that there shouldn't be over overlapping of clusters



Entangled Latent space using t-SNE plot

Approaches

- Beta-VAE varies the KLD loss by multiplying it with a Beta term that modifies the weightage of KLD loss in the overall loss term of our VAE model.
- As the weightage of KLD Loss increases, it compels the model to learn each dimension properly.
- Other approaches include KLD Clipping, clustering, correlation factoring, adversarial training by introducing a discrimination model, etc.



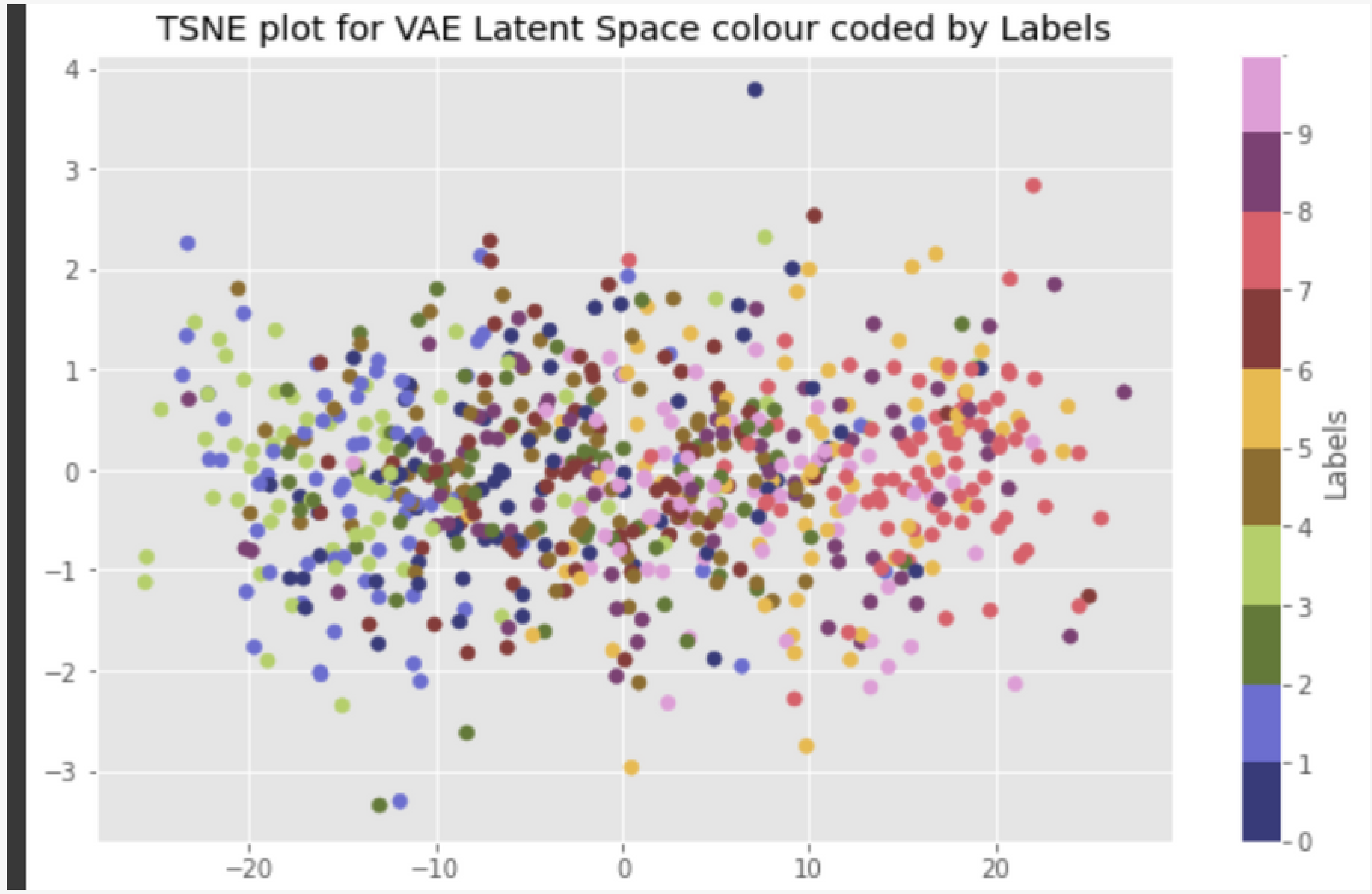
```
def loss_function(recon_x, x, mu, logvar, B):

  BCE = nn.functional.binary_cross_entropy(recon_x, x.view(-1, 784), reduction='sum')
  KLD = -B * torch.sum(1 + logvar - mu.pow(2) - logvar.exp())

  return BCE + KLD
```

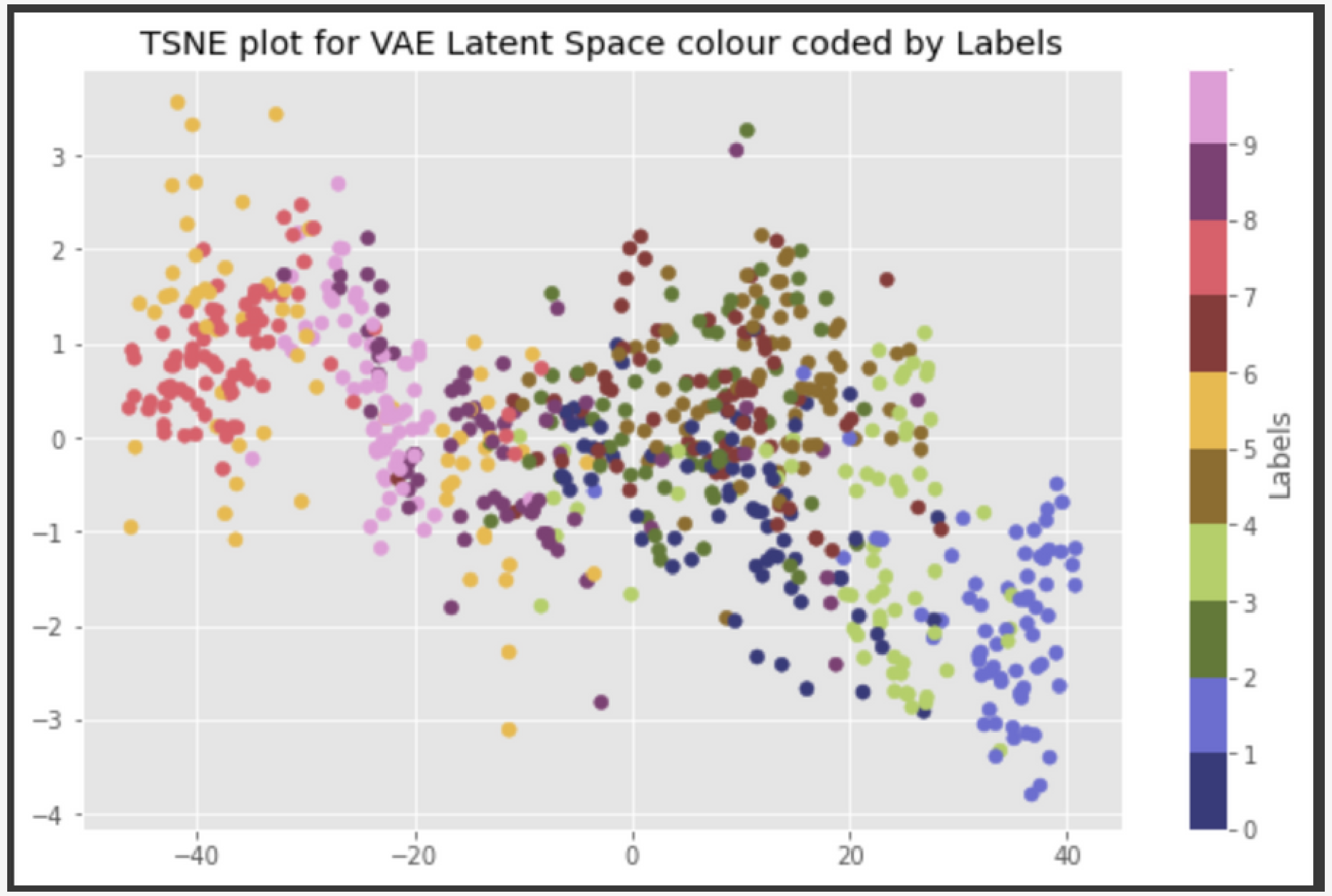
Adding Beta term to the KLD loss function

Results



Before

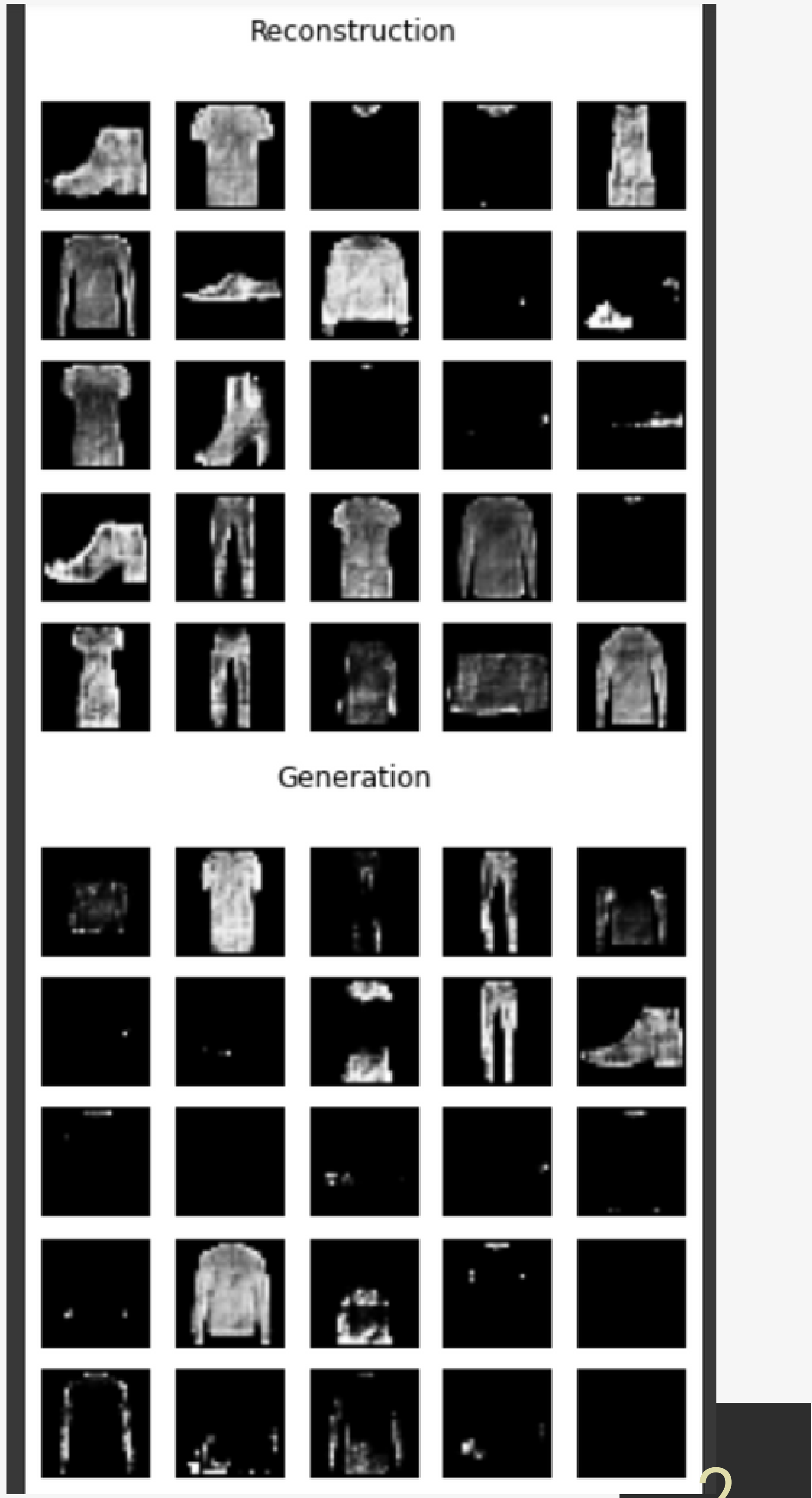
B=4



After

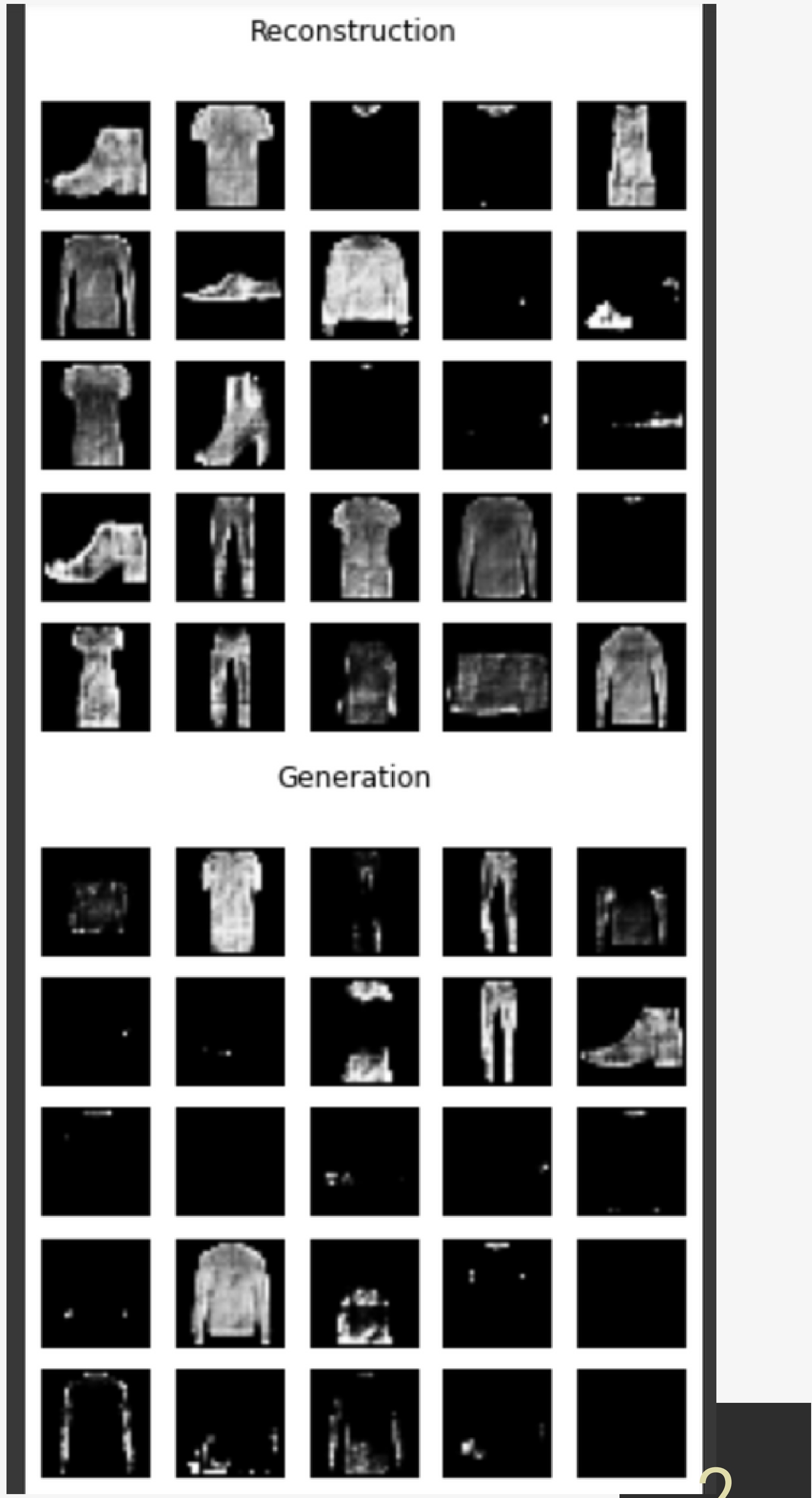
Work for future

- Improving the architecture of VAE by using Convolutional layers instead of linear layers.
- Experimenting on how this disentanglement is actually affecting the latent space by adding epsilon term to the latent space vector.
- Modifying the KLD loss equation specifically by use case and dataset.
- Trying to combine two techniques like, clipping and adversarial training together, etc.



Work for future

- Improving the architecture of VAE by using Convolutional layers instead of linear layers.
- Experimenting on how this disentanglement is actually affecting the latent space by adding epsilon term to the latent space vector.
- Modifying the KLD loss equation specifically by use case and dataset.
- Trying to combine two techniques like, clipping and adversarial training together, etc.



Thank you