Home Work

Erik Norlin

May, 2023

Home work C

3a)

Number of data samples for the autoencoder:
Training samples: 10000
Validation samples: 2000

Latent space: 4 to represent (x, y, r, I)

The larger the latent space the lesser the loss. A latent space of 4 nodes recreates the frame almost perfectly. In theory 4 in latent space represents the position (x, y), the radius r, and the intensity I.

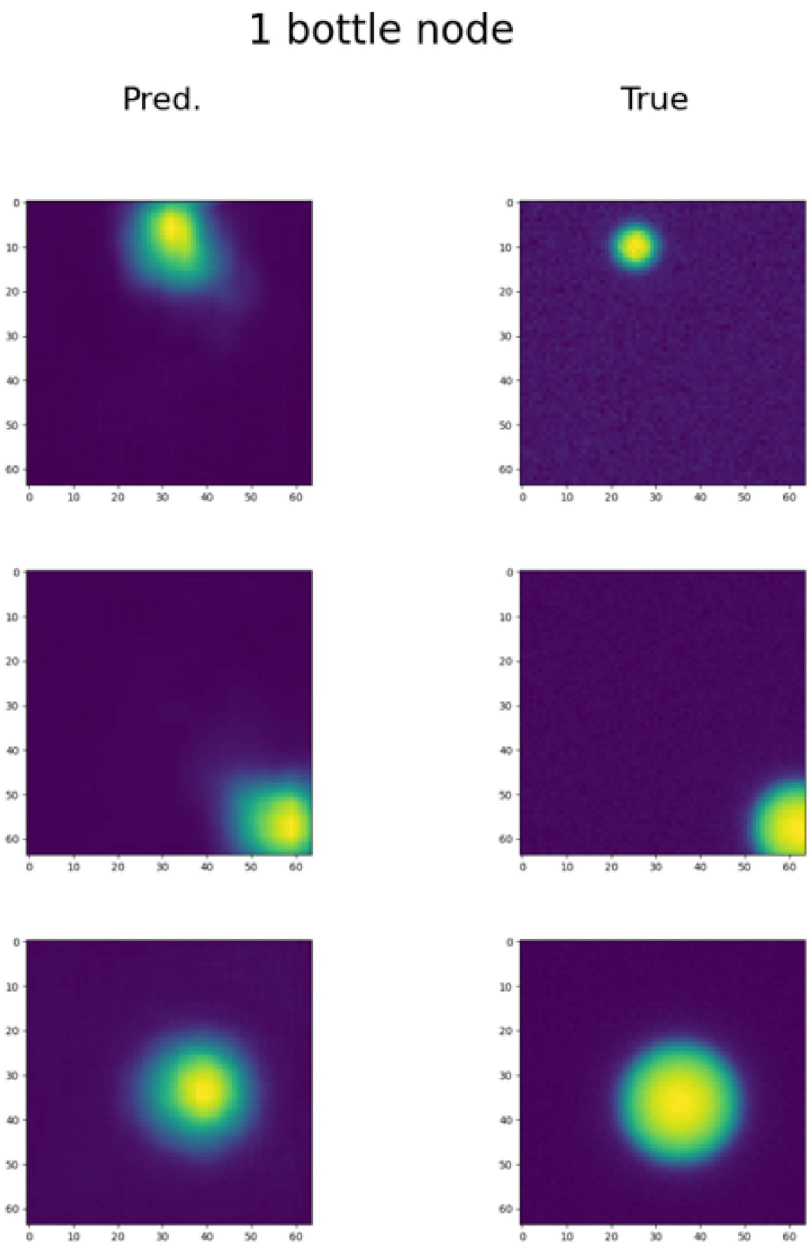

Figure 1: Autoencoder frame prediction for latent space = 1.

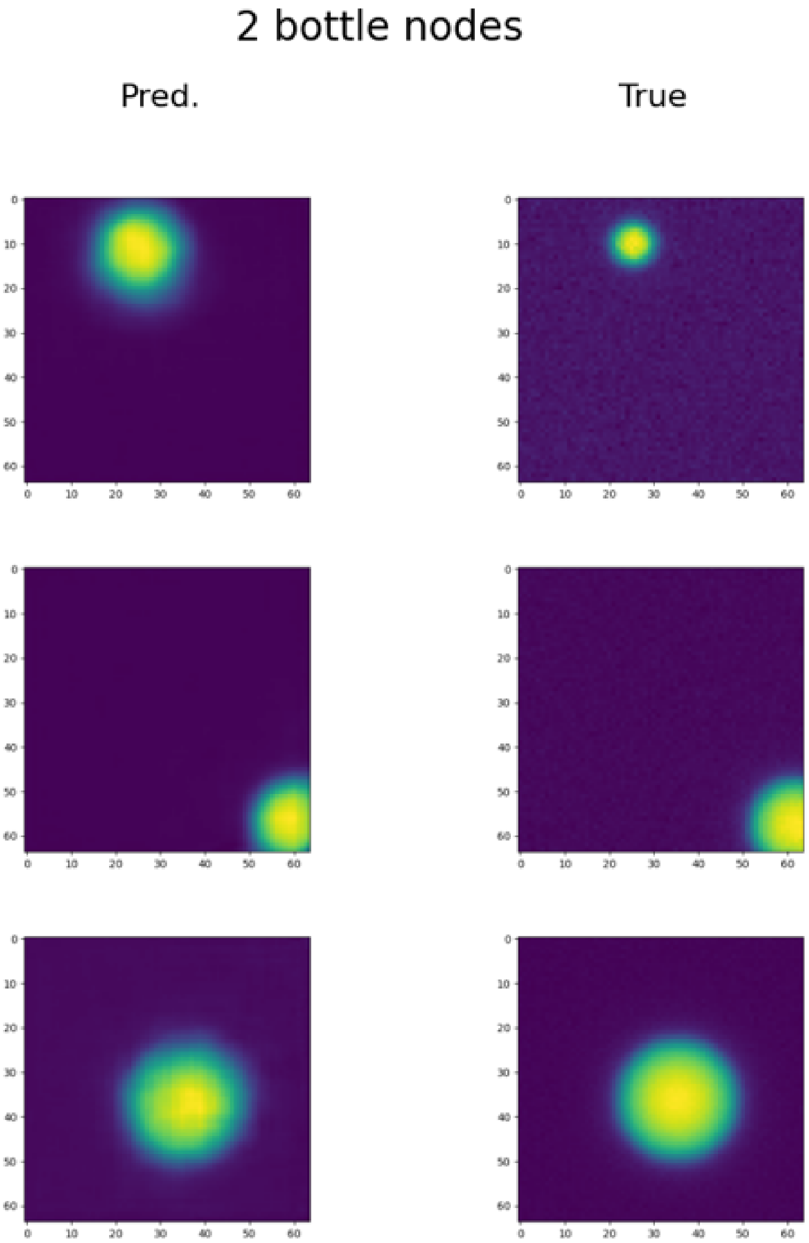

Figure 2: Autoencoder frame prediction for latent space = 2.

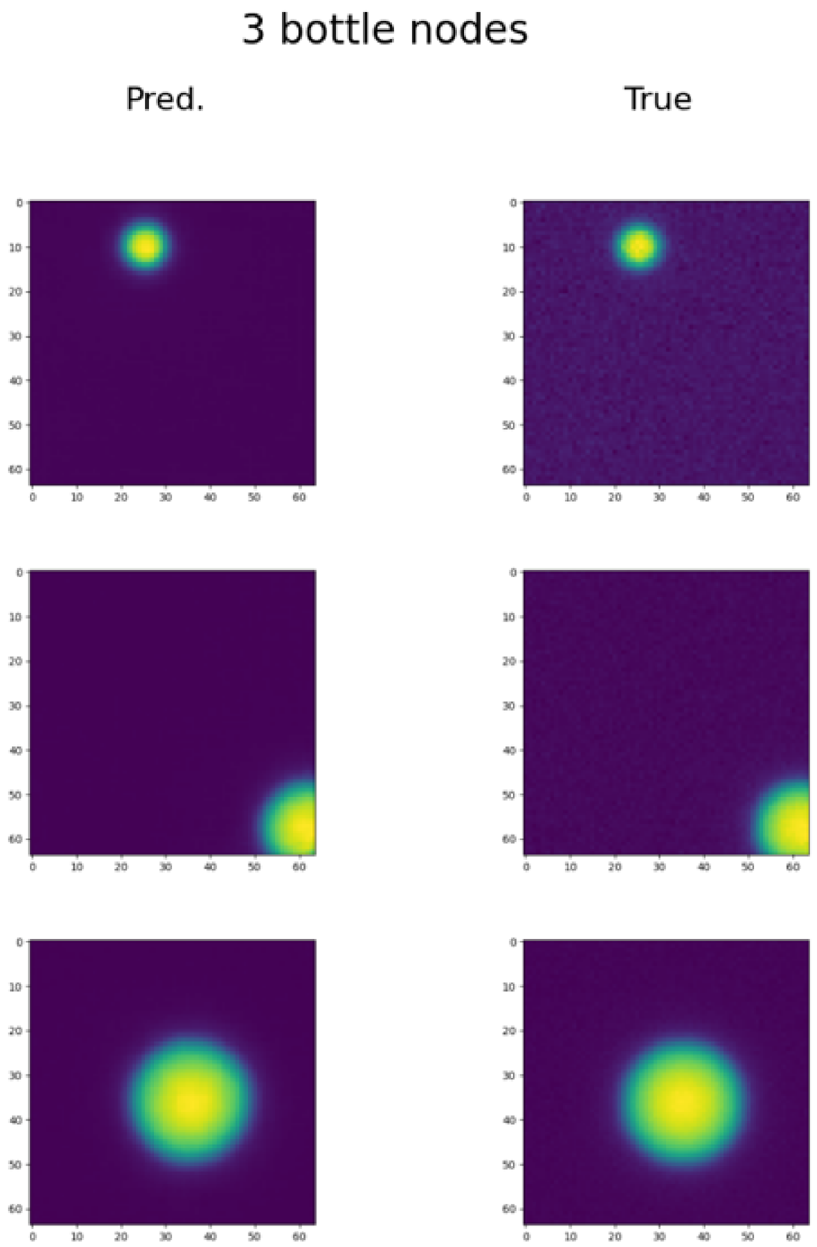

Figure 3: Autoencoder frame prediction for latent space = 3.

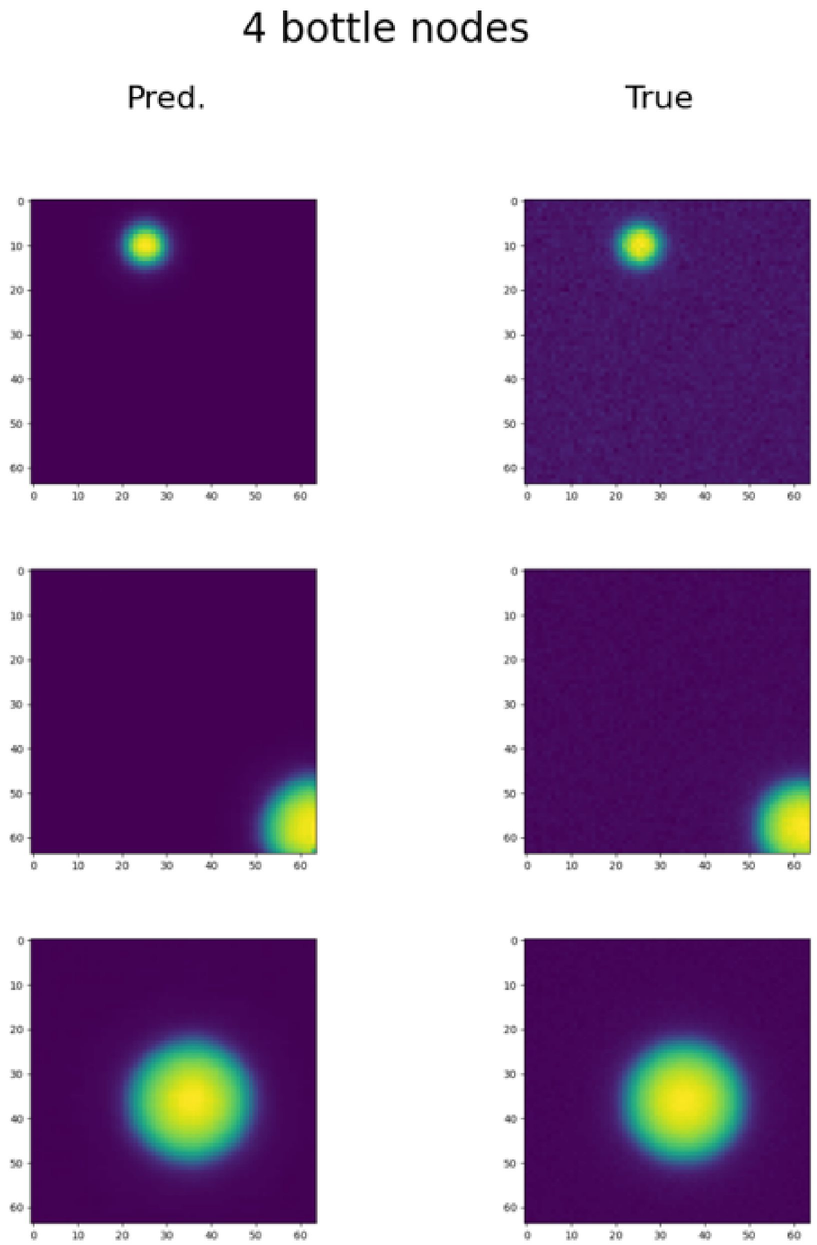

Figure 4: Autoencoder frame prediction for latent space = 4.

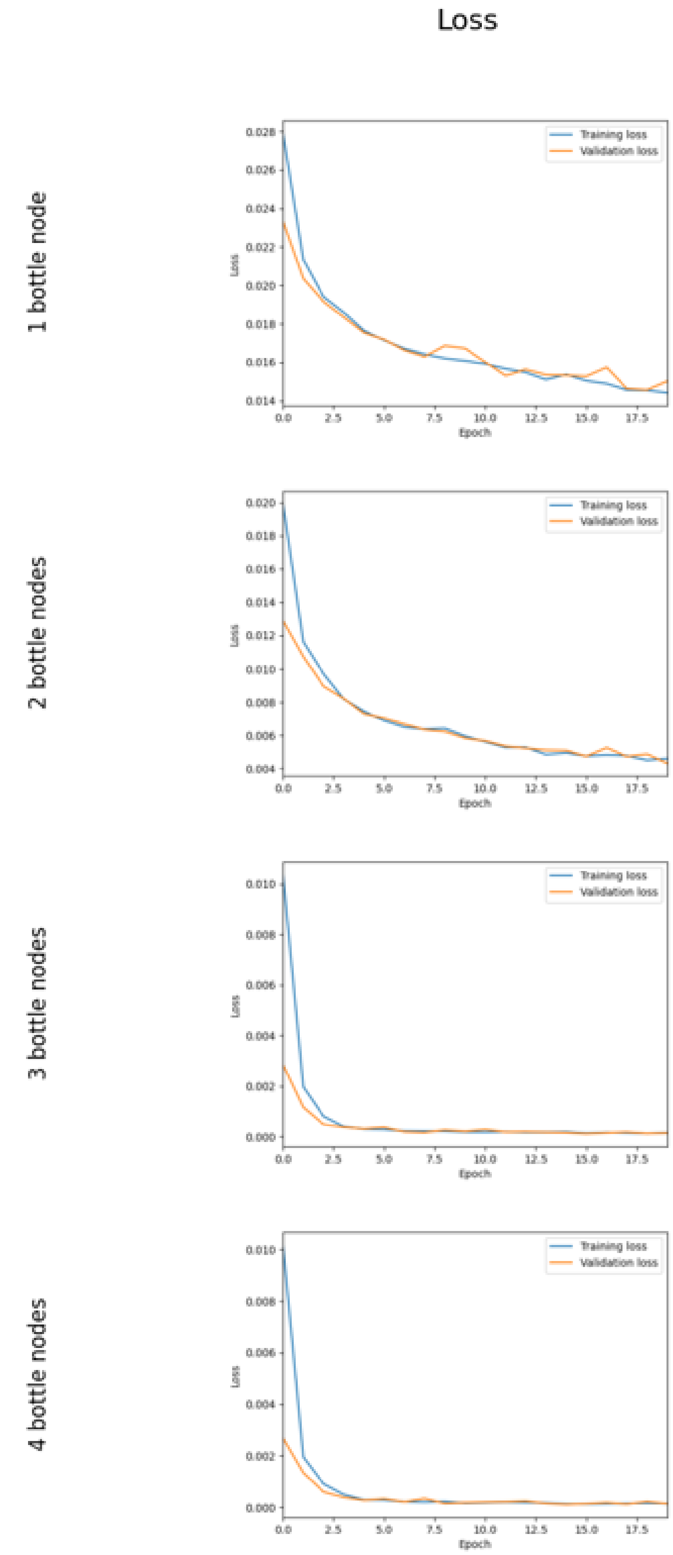

Figure 5: Loss for training the autoencoder for different latent spaces. 4 bottle nodes gives the smallest loss of a minimum of $\approx 9e-5$.

Minimum loss of autoencoder: 6e-5 (latent space: 10), 9e-5 (latent space: 4).

3b)

Number of data samples (sequences) for the transformer:
Sequence length: 10 (input: first 9, label: 10th)
Training samples: 1680
Validation samples: 420

When predicting sequences of frames with the transformer, the latent space was increased to 10 to improve the autoencoders frame prediction (loss improved to $6e-5$). The reason for this is that the autoencoder is not perfect and will always predict with a slight error. As the predicted next frame comes out from the decoder the frame is slightly distorted. This frame goes back into the encoder as the latest frame of the sequence to predict the next frame, becoming even more distorted. This process recursively can distort the frame sequences significantly over many iterations. It is therefore preferable to minimize the loss of the autoencoder as much as possible before training the transformer.

Minimum loss of transformer: 0.03

For the transformer architecture, the number of attention gates and encoders were experimented with. It was found that 2-3 stacked transformer encoders and 3-4 attention gates per multi-headed attention gave the best result of predicting the next frame of a sequence.

Max frame prediction for transformer: 20 identical frames but generally 10-15. At best 45 frames if considering motion and discounting lag.

3c)

Number of data samples (sequences) for the LSTM:
Sequence length: 10 (input: first 9, label: 10th)
Training samples: 1680
Validation samples: 420

The LSTM model was not better at predicting the next frame long term as accurately as the transformer. The transformer also had a lower minimum loss. The LSTM architecture that gave the best results consisted of 4 LSTM layers with 512 nodes per layer.

Minimum loss for LSTM: 0.05

Max frame prediction for LSTM: generally 3-5 frames. At best 20 frames if considering motion and discounting lag.

In conclusion: When predicting the next frame of a time sequence, the transformer on average is able to predict 5 more frames accurately than the LSTM. If only considering motion and discounting lag, the transformer is able to predict at best 30 more frames accurately than the LSTM.