Challenge 2: Molecule Generation

Name: Lukas Nöbauer
ID: k11702439

For the challenge 2, I first revisited the slides of the corresponding lectures and did some online research regarding molecule generation with Machine Learning. The best performing models for these type of tasks have shown to be LSTMs, that why my first and final approach was also a LSTM.

Luckily I already had some experience with LSTMs, because I did Sepp Hochreiter's LSTM course last semester. In the exercises we also had to implement a CharNN, which I also used as the basis for this challenge.

The first thing I set up was a proper dataloader within my Google Colab environment, which encoded the characters contained in the smiles into tokens. These tokens where then one-hot encoded and used as the inputs for the LSTM. The outputs were logits with the same shape as the inputs. These output logits were then passed to the Cross Entropy Loss function, which afterwards was used to initialize the backward pass through the network.

I the beginning I trained the model on ~500.000 smile strings with a sequence length of 30 to keep training times short, to do some exploration to get a feeling how the network performs with the given settings.

What has shown to yield good results where 2 LSTM (Pytorch) cells along with Xavier initialization. For the optimizer I used Adam. To validate my model I used 10% of my data as a validation set. Furthermore, I also tried Dropout techniques, which finally showed to be not too beneficial for the outcome.

The best results I was able to achieve was with a hidden size of 1024, batch size of 124. With my first submissions I had problems with the FCD score, but after adapting the top_k parameter to 5, I was able to improve my FCD score significantly to around 1.

Although I was already familiar with CharNNs I was not aware, that these networks work so well for molecule generation. In general, it was really fun doing this challenge and I am happy that I was able to achieve such good results!