Natural Language Processing has been widely used for tasks such as machine translation, summarisation, and sentence completion. Since Artificial Neural Networks improve their performance through training, it is not immediately clear how they achieve such feats through connectivity. Being able to interpret trained NLP models would not only answer the “how” question, but also open the possibility for targeted modification to improve the performance of the model.

Our approach was inspired by Computer Vision: extraction of the features from RNN (LSTM) and Transformers to answer the questions, whether such models are sufficient for text generation, returning grammatically correct sentence with coherent meanings; we also wanted to find out, what features can be extracted from models and how they compared between LSTM and Transformers. We first trained a Recurrent Neural Network (LSTM variant of RNN) and a Transformer as examples of NLP models, before visualising the activation of neurons following input of example sentences, to explore how the model represents natural language. We hypothesised that later layers may have higher-level representations compared to the earlier layers.

For our model we used a kaggle dataset “190k+ Medium Articles”. From it, we created a custom vocabulary which could be used for Word2Vec technique utilising gensim library. The obtained vectors would then be fed into the NLP model with layers: double LSTM, Linear, Softmax. The output would be compared to the trained vocabulary.

Unfortunately, the model turned out to be very expensive computationally, especially vocabulary building. Given enough time, it should work- based on the partial results. We conclude that building a text generator from scratch and analysis of it is a surprisingly demanding task.

In the future, we could pursue our original hypotheses and explore the properties of the models.