Explanation of neural language models using SHAP

Emese Vastag

Abstract

Explainability in machine learning is becoming increasingly important. In more and more cases working with an explainable algorithm is at least as important as a good model performance. There are several existing approaches as to how we can explain a machine learning model, model-agnostic and model-specific versions as well.

In this thesis, we examined SHAP values that asssign importance to each feature in the data, based on how much they contributed to the final prediction. This method is based on game theory results from Lloyd Shapley, who first introduced the Shapley values. These values show that in an n-person cooperative game how much the players contributed to the overall gain while also satisfying a group of desirable axioms.

We tested SHAP values on natural language models, for that we used an LSTM model which is an improved version of recurrent neural networks. We tested an own implementation of SHAP as well as the official SHAP implementation in Python. The results are mostly consistent with human intuitions, however we can see differences between different approaches to compute SHAP values.