Creating a Benchmark for Investigating Robustness of Pre-Trained Language Models

Supervisor : Lütfi Kerem Şenel

Examiner : Professor Hinrich Schütze

Open to : BSc/Msc

Robustness is a very important property for any deep learning model. Small and insignificant changes in the input should not cause big changes in the model's output. Robustness is usually examined through performing adversarial attacks to a model fine-tuned on a downstream task. However, robustness of a language model can be evaluated on a more general level before fine tuning using probing approaches. Purpose of this project is to create a benchmark for evaluating the robustness of PLMs. For this purpose, we need to construct a comprehensive dataset by systematically making seemingly insignificant semantic (paraphrasing, replacing words with synonyms etc.) and syntactic (reordering words, adding typos, punctuations etc.) alterations on various natural text inputs. Then, stability of various models' predictions will be investigated on this dataset using probing techniques.

Project Takeaways: Gain experience with SOTA NLP models like BERT/GPT-2; gain insights about the strengths and weaknesses of these models.