Title: Learning Capabilities of the Transformer Neural Networks

Author: Dušan Variš

Department: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Ondřej Bojar, Ph.D.,
Institute of Formal and Applied Linguistics

Abstract:

Although the contemporary neural networks, inspired by biological neurons, were able to reach human-like performance on many tasks in recent years, their optimization (learning) process is still very far from the one observed in humans. This thesis investigates various aspects of learning in the current state-of-the-art Transformer neural networks, the dominant architecture in the current neural language processing. Firstly, we measure the level of generalization in Transformers using several probing experiments based on the idea of adversarial evaluation. Secondly, we explore their potential for incremental learning when combined with regularization using the elastic weight consolidation approach. Lastly, we propose a modular extension of the existing Transformer architecture enabling subnetwork selection conditioned on the intermediate hidden layer outputs and analyze the attributes of this network modularization. We investigate our hypotheses mainly within the scope of neural machine translation and multilingual translation showing the limitations of the original Transformer and the elastic weights consolidation regularization while presenting promising results of the novel modular Transformer architecture.

Keywords: neural machine translation, catastrophic forgetting, modular neural networks, incremental learning, generalization