From Sequence to Attention; Search for a Compositional Bias in Sequence-to-Sequence Models
K.W. Korrel

Abstract

Recent progress in deep learning has sparked a great, renewed interest in the field of artificial intelligence. This is in part because of achieved superhuman performance on several problems, and great versatility. A trained deep learning model, however, can typically only be applied in a very narrow domain as they only excel on test data that is drawn from the same distribution as the training data. This is exemplified by research on adversarial examples that shows how deep learning models respond on valid and perturbed data. However, even when test data comes from a significantly different distribution than the train data, it may be valid in a compositional sense. Recent research on systematic compositionality has provided evidence that deep learning models generally lack a compositional understanding of the domains that they are trained on.

Compositionality is a feat that is often attributed to humans that allows quick few-shot learning and easy generalization to new domains and problem instances. Such an understanding is also crucial in natural language. In short, the principle of semantic compositionality means that the semantic meaning of a complex expression can be explained by the meaning of its constituents and the manner in which they are combined.

In this thesis we show that although deep learning models are potentially capable of having such an understanding, they typically do not converge on such an solution with regular training techniques. We propose two new techniques that aim to induce compositional understanding in sequence-to-sequence networks with attention mechanisms. Both are founded on the hypothesis that a salient, informative attention pattern helps in finding such a bias and in countering the use of spurious patterns in the data. The first of these methods, Attentive Guidance, guides a model in finding correct alignments between input and output sequences. It is a minor extension to existing sequence-to-sequence models and is intended to confirm the aforementioned hypothesis. The second method, the sequence-to-attention architecture, involves a more rigorous overhaul of the sequence-to-sequence model with the intention to further explore and exploit this hypothesis. We use existing data sets to show that both methods perform better on tasks that are assumed to correlate with systematic compositionality.