Re-encoding in Neural Machine Translation
J. Baptist

Abstract

Currently, most approaches to neural machine translations use an encoder that produces continuous representations of the words in the input sentence. Next, another component decodes the representations into output words, where a separate attention mechanism typically allows the model to attend to different input words depending on the previous predictions. This approach comes with two potential problems. First, the input representations remain constant during the prediction process, which may not provide enough variance for the model to effectively discriminate between them. Second, a high burden is put on the component responsible for predicting output words (decoder), as translations involves many subtasks: knowing which words to translate next, remembering which words have been translated, producing a fluent and grammatically correct sentence, and more. As a result, many neural models suffer from over-translation (generating unnecessary words), under-translation (forgetting to translate words) and repetition. By extending such models with a re-encoding component, which allows the model to update the input representations depending on the previous predictions, these tasks can naturally be moved away from the decoding component. This thesis investigates two architectures that use re-encoding and compares them to multiple baselines. The qualitative and quantitative results show that re-encoding can potentially improve performance of neural models, especially on longer sentences.