Neural Machine Translation by Jointly Learning to Align and Translate Summary

Paper Link: https://arxiv.org/pdf/1409.0473

This paper fixes a an issue in early translation models. The old approach tried to squeeze a whole sentence into one vector before translating it, which doesn't work well for longer sentences as it makes it lose too much information.

The solution the paper suggested is simple, instead of using one vector for the whole sentence, the model learns to "pay attention" to different parts of the input as it translates. So when it's generating each word in the translation, it focuses only on the relevant words in the source. That attention mechanism is what makes the model so much better.

They call the model **RNNsearch**, and it outperforms the older **RNNencdec**, especially on long sentences. It also does about as well as traditional phrase-based systems, but with a single neural mode.

Another thing they added is a bidirectional RNN for the encoder, which gives each word context from both sides. The model also learns soft alignments between source and target words, which actually look pretty natural when you visualize them — it handles things like word reordering nicely.