Mohamed Sameh Saafan

Neural machine translation with Attention paper summary

Introduction

- Neural Machine Translation (NMT) achieves state-of-the-art results in large-scale translation.
- Minimal domain knowledge required; trained end-to-end.
- Attention mechanisms improve translation quality by focusing on relevant parts of the source sentence.

Neural Machine Translation Basics

- Consists of encoder and decoder; models conditional probability p(y/x).
- Decoder typically uses RNNs (LSTM or GRU).
- Attention introduced to overcome fixed-length context vector limitations.

Attention Mechanisms

- Global Attention: attends to all source words; more computational cost.
- Local Attention: attends to a subset (window) of source words; efficient and nearly differentiable.
- Input-feeding: past attention vectors fed into next steps to retain coverage awareness.

Experiments and Results

- Tested on WMT English-German tasks (2014 & 2015).
- Local attention with input-feeding and predictive alignment achieved highest BLEU scores.
- Ensemble model scored 25.9 BLEU (new SOTA) on WMT'15 English-German.

Analysis

- Attention improves learning, alignment quality, and performance on long sentences.
- Dot-product and general alignment functions performed better than location and concatinate.
- Attention-based models can align named entities and complex phrases accurately.

Conclusion

- Global and Local attention models significantly enhance NMT.
- Input-feeding boosts model’s awareness of past decisions.
- Results confirm superiority of attention over non-attentional baselines.