Multi-Granularity Self-Attention for Neural Machine Translation

Yang Wei 🐷 51184506043

godweiyang@gmail.com

godweiyang.com

Computer Science and Technology

East China Normal University

Multi-Granularity Self-Attention for Neural Machine Translation

Jie Hao*
Florida State University
haoj8711@gmail.com

Xing Wang
Tencent AI Lab
brightxwang@tencent.com

Shuming Shi
Tencent AI Lab
shumingshi@tencent.com

Jinfeng Zhang
Florida State University
jinfeng@stat.fsu.edu

Zhaopeng Tu
Tencent AI Lab
zptu@tencent.com

Motivations

- SANs generally focus on disperse words and ignore continuous phrase patterns, which have proven essential in both SMT and NMT.
- The power of multiple heads in SANs is not fully exploited.
- Thus this paper (MG-SA) assigns several attention heads to attend over phrase fragments at each granularity.

Framework

- word-level → phrase-level memory:

$$H_g = F_h(H)$$

- single head self-attention:

$$\begin{aligned} Q^h, K^h, V^h &= HW_Q^h, H_g W_K^h, H_g W_V^h \\ O^h &= \mathrm{ATT}(Q^h, K^h)V^h \end{aligned}$$

- final output of MG-SA:

$$\text{MG-SA}(H) = [O^1, \ldots, O^N]$$