Class Notes: Attention Mechanisms in Neural Networks

1. Bahdanau (Additive) Attention

For each decoder hidden state s_t, the attention mechanism computes a context vector c_t as a weighted sum of encoder hidden states h_s.
Keys: Encoder hidden states h_s
Queries: Decoder hidden state s_t
Values: Encoder hidden states h_s

$$e_{t,s} = v_a^\top \tanh(W_q s_t + W_k h_s), \qquad \alpha_{t,s} = \frac{\exp(e_{t,s})}{\sum_j \exp(e_{t,j})}, \qquad c_t = \sum_s \alpha_{t,s} h_s.$$

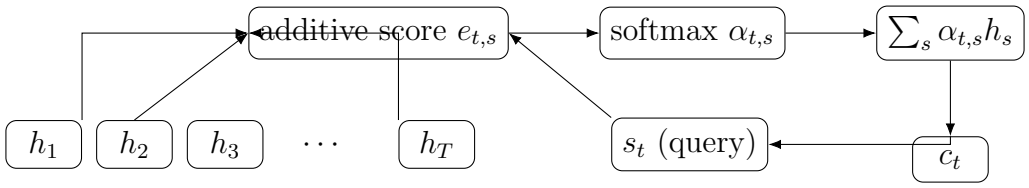

Figure 1: Bahdanau additive attention.

2. Luong (Multiplicative) Attention

The "global" variant uses the current (or previous) decoder state as query and a dot/general product for scoring.
Keys: Encoder hidden states h_s
Queries: Decoder hidden state s_t (or s_{t-1})
Values: Encoder hidden states h_s

$$e_{t,s} = s_t^\top h_s \quad \text{(dot)} \qquad \text{or} \qquad e_{t,s} = s_t^\top W h_s \quad \text{(general)}.$$

3. Multi-Head Attention (Parallel)

Self-attention where Q, K, V are projected into h subspaces, processed in parallel, then concatenated.

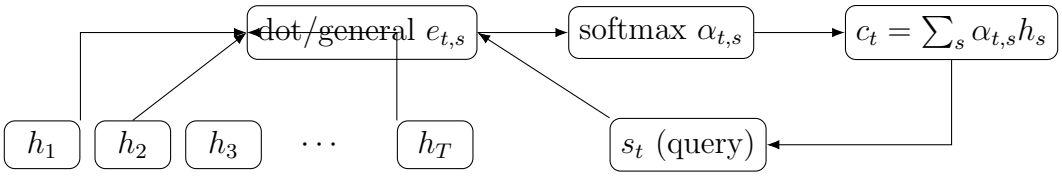

Figure 2: Luong multiplicative attention.

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V, \qquad \text{head}_i = \text{softmax}\left(\frac{QW_i^Q(KW_i^K)^\top}{\sqrt{d_k}}\right)VW_i^V.$$

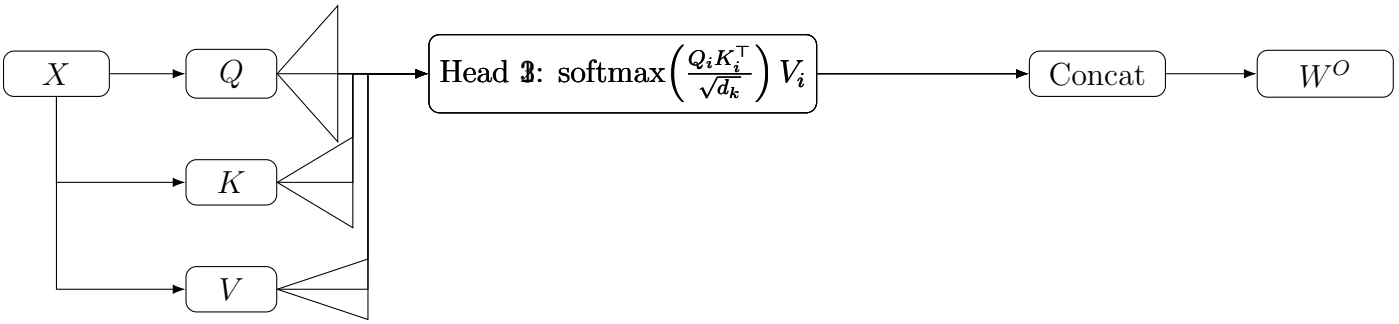

Figure 3: Multi-head attention: parallel heads over projected Q/K/V.

4. Self-Attention

Aggregates a sequence into updated token representations using scaled dot-product attention where Q, K, V come from the same source.

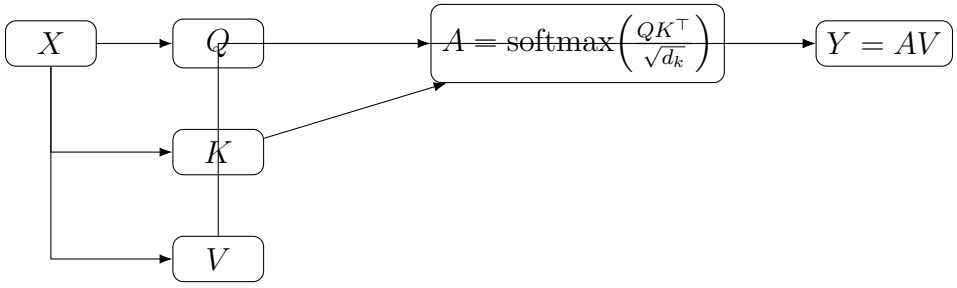

Figure 4: Self-attention with shared source for Q/K/V.

5. Summary Table

Variant	Q Source	K Source	Scoring	Use Case
Bahdanau	Decoder RNN	Encoder RNN	Additive MLP	Seq2Seq
Luong	Decoder RNN	Encoder RNN	Dot/General	Seq2Seq
Self-Attn	Same sequence	Same sequence	Scaled Dot	Transformers
Multi-Head	Same sequence	Same sequence	Scaled Dot (multi)	Rich relations