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

$$e_{ts} = v_a^\top \tanh(W_q s_t + W_k h_s)$$
$$\alpha_{ts} = \frac{\exp(e_{ts})}{\sum_j \exp(e_{tj})}$$

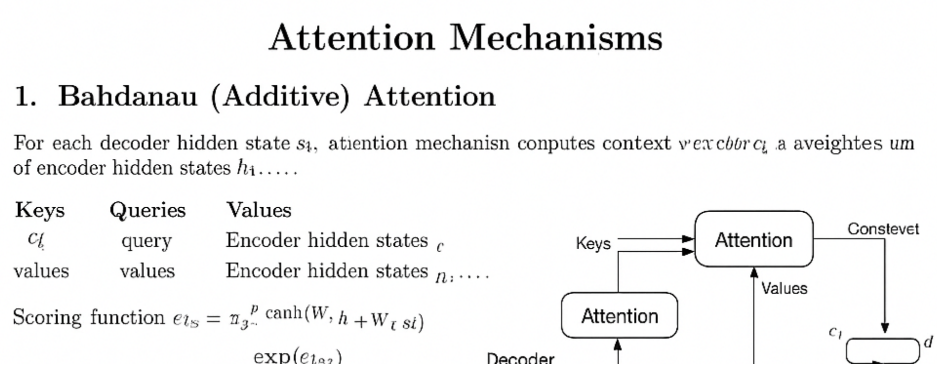

Figure 1: Bahdanau Attention flow diagram.

2. Luong (Multiplicative) Attention

The "global" variant uses the decoder hidden state from the previous timestep.
Keys: Encoder hidden states h_s
Queries: Decoder hidden state s_{t-1}
Values: Encoder hidden states h_s

$$e_{ts} = h_t^\top h_s \quad \text{or} \quad e_{ts} = h_t^\top W h_s$$

$$\alpha_{ts} = \frac{\exp(e_{ts})}{\sum_j \exp(e_{tj})}$$

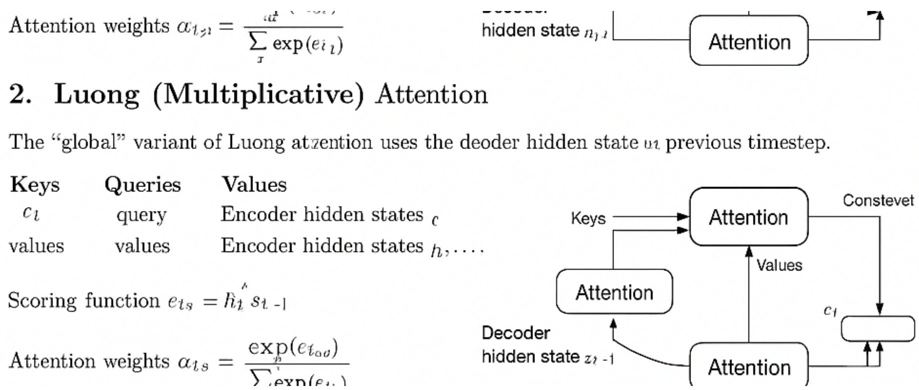

Figure 2: Luong Attention flow diagram.

3. Multi-Head Attention

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

$$Q_h = QW_h^Q, \quad K_h = KW_h^K, \quad V_h = VW_h^V$$

$$A_h = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}}\right), \quad O_h = A_h V_h$$

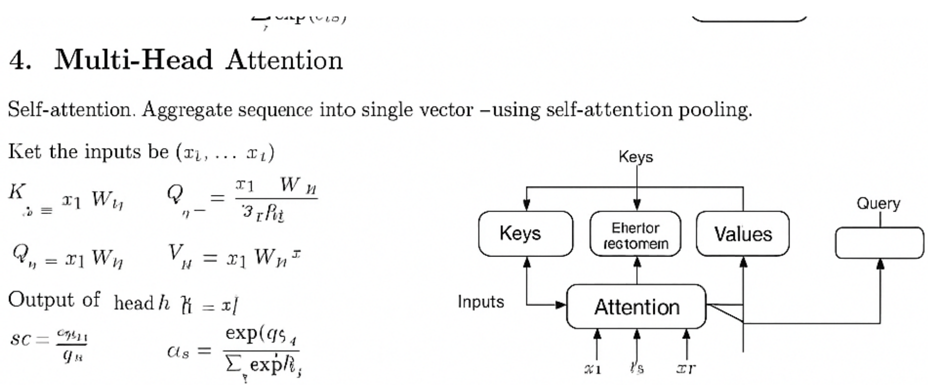

Figure 3: Multi-Head Attention computation from the provided diagram.

4. Self-Attention

Aggregates a sequence into a single vector via scaled dot-product.

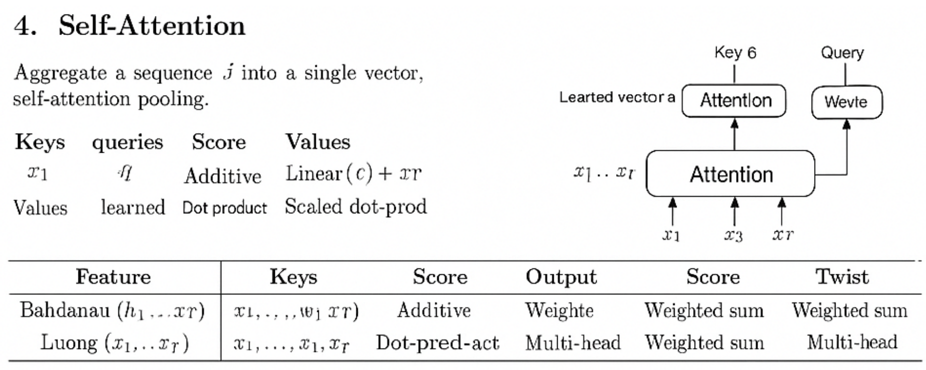

Figure 4: Self-Attention flow from the provided diagram.

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