Class Notes: Attention Mechanisms in Neural Networks

1. Motivation

Traditional RNNs compress entire sequences into a single vector, creating an information bottleneck. Attention mechanisms allow models to dynamically focus on different parts of the input, improving performance on tasks like translation, summarization, and question answering.

2. Core Components

Given hidden states $H = [h_1, \ldots, h_T]$:

- **Keys (K)**: Represent what each position contains.
- **Queries (Q)**: Represent what we are looking for.
- **Values (V)**: Contain the actual content to retrieve.

Projected via learnable matrices:

$$Q = HW^Q, \quad K = HW^K, \quad V = HW^V$$

3. Scoring Functions

(a) **Additive (Bahdanau, 2014)**:

$$e_{ij} = v_a^\top \tanh(W_q q_i + W_k k_j)$$

(b) **Multiplicative (Luong, 2015)**:

$$e_{ij} = q_i^\top k_j$$

or $q_i^\top W k_j$.

(c) **Scaled Dot-Product (Transformer, 2017)**:

$$e_{ij} = \frac{q_i^\top k_j}{\sqrt{d_k}}$$

4. Attention Weights and Context Vector

Softmax normalization:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'=1}^{T} \exp(e_{ij'})}$$

Context vector:

$$c_i = \sum_{j=1}^{T} \alpha_{ij} v_j$$

5. Variants

- **Bahdanau (Additive)**: Q from decoder RNN state; K, V from encoder outputs.
- **Luong (Multiplicative)**: Faster, fewer parameters.
- **Self-Attention**: Q, K, V from same sequence.
- **Multi-Head**: Multiple Q/K/V projections capture different relationships.

6. Training Considerations

- Masking for autoregressive tasks.
- Dropout on attention weights.
- Memory complexity $O(T^2)$; use efficient variants for long sequences.

7. Example: Self-Attention Step

For $H \in \mathbb{R}^{4\times 8}$:

1. Project to Q, K, V with $d_k = d_v = 4$.
2. Compute $QK^\top \in \mathbb{R}^{4\times 4}$.
3. Scale, apply softmax row-wise to get weights.
4. Multiply by V to get new token representations.

8. 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