Polysemy and Multi-Head Attention

Problem: The Word "Bank" in Different Contexts

Sentence 1: "He went to the bank to deposit money."
Sentence 2: "He sat on the bank of the river."

Expected High Attention Weights

Sentence 1: "bank" = financial institution

High attention from "bank" to:

- **"deposit"** - directly related to financial transactions
- **"money"** - confirms financial context
- **"went"** - action associated with going to a place/building

Reasoning: These words disambiguate "bank" as a financial institution through semantic context.

Sentence 2: "bank" = riverbank/shore

High attention from "bank" to:

- **"river"** - directly defines the type of bank
- **"sat"** - physical action associated with a surface/location
- **"on"** - preposition indicating physical location

Reasoning: These words disambiguate "bank" as a geographical feature through spatial/physical context.

How Multi-Head Attention Handles Polysemy

Different Heads Learn Different Relationships

Head 1 might specialize in:

- **Financial/transactional relationships**
- Attends: "bank" → "deposit", "money"
- Captures: commercial/business semantics

Head 2 might specialize in:

- **Spatial/physical relationships**
- Attends: "bank" → "river", "on", "sat"
- Captures: location/geography semantics

Head 3 might specialize in:

- **Syntactic dependencies**
- Attends: "bank" → "the", "went"/"sat"
- Captures: grammatical structure

Mechanism

1. **Different projection matrices** (W_Q^h, W_K^h, W_V^h) in each head create different representation subspaces
2. **Each head independently computes attention** based on its learned projections
3. **Heads specialize** during training to capture different types of relationships
4. **Outputs are concatenated** and combined, allowing the model to integrate multiple perspectives

Result

- In Sentence 1: Financial-focused heads dominate, pulling representation toward "financial institution"
- In Sentence 2: Spatial-focused heads dominate, pulling representation toward "riverbank"
- **Context-dependent representation** emerges from combining all heads

Key Insight

Multi-head attention allows the model to simultaneously consider:

- **Multiple types of context** (semantic, syntactic, positional)
- **Multiple relationship types** (financial, spatial, temporal)
- **Parallel disambiguation strategies**

This is why transformers handle polysemy better than single-context models - they can attend to different disambiguating cues simultaneously.