Multi-Head Attention Advantage

Key Advantage: Parallel Attention to Different Relationship Types

Multi-head attention allows the model to **simultaneously focus on different types of relationships** in the same input.

Example

In the sentence "The quick brown fox jumps over the lazy dog":

- **Head 1** might attend to **syntactic structure** (subject-verb-object relationships)
- **Head 2** might attend to **semantic associations** (adjective-noun pairs like "quick fox", "lazy dog")
- **Head 3** might attend to **positional patterns** (nearby words or long-range dependencies)

Why This Matters

Single-head attention is forced to compromise - it must blend all these relationship types into one attention pattern, losing specificity.

Multi-head attention lets each head specialize, then combines their outputs. This creates a **richer, more nuanced representation** that captures multiple aspects of the input simultaneously.

Bottom Line

Instead of one "blurry" attention pattern trying to capture everything, you get multiple specialized patterns that together provide a more complete understanding of the input.

Concrete Example: Different Heads, Different Patterns

Sentence: "The small dog quickly chased the cat"

Query word: "dog"

Head 1: Syntactic Relationships

Attends to: "chased" (80%), "dog" (15%), others (5%)

- Focuses on verb that "dog" performs
- Captures subject-verb dependency

Head 2: Semantic Attributes

Attends to: "small" (70%), "dog" (25%), others (5%)

- Focuses on adjectives describing "dog"
- Captures descriptive properties

Head 3: Action Context

Attends to: "cat" (60%), "chased" (30%), "dog" (10%)

- Focuses on object of the action
- Captures agent-patient relationship

Result

All three perspectives combine to give "dog" a rich representation that includes:

- What it does (chases)
- What it's like (small)
- What it interacts with (cat)

Single-head attention would have to average these, losing the distinct patterns each head captures.