Embedding System & Self-Attention – Simple Explanation

This document explains two key parts of how modern language models (like GPT or BERT) work: the embedding system and self-attention. These are the building blocks that allow models to understand and generate human language.

1. What is an Embedding System?

Before a computer can understand language, it needs to convert words into numbers — because models only work with numbers, not text.

Think of it like this:

When you say “dog”, the computer doesn’t know what a dog is. But if we assign it a vector (a list of numbers), then the model can use it for calculations.

Main parts of an embedding system:

Token Embeddings:

Every word (or part of a word) is turned into a unique vector. These vectors capture some meaning — for example, “dog” and “puppy” might have similar vectors.

Positional Embeddings:

Transformers don’t process words in order like humans do. So we add “position info” to each word to show where it appears in the sentence.

Segment Embeddings (optional):

If we give the model two sentences, we can tell it which word belongs to which sentence using these.

Example:

Sentence: "The cat sleeps."

Token Embeddings: ["The", "cat", "sleeps"] → [vectors]

Positional Info: [position 1, 2, 3] → [vectors]

Final Input = Token Vector + Position Vector

2. What is Self-Attention?

Now that each word is a vector, how does the model figure out which words matter most in a sentence?

That’s what self-attention does.

Simple idea:

Self-attention lets the model look at all the words in a sentence at once and decide which ones are important to each other.

Example:

Sentence: "The dog chased the ball because it was fast."

Question: What does “it” refer to?

Self-attention helps the model look back and see that “it” might refer to “dog” or “ball”.

Based on context, it figures out that “ball” is more likely.

3. Multi-Head Attention

Instead of doing this attention once, transformers do it multiple times in parallel, with each “head” looking at the sentence differently.

Some heads might focus on grammar, others on meaning or relationships.

Afterward, all of this information is combined.

4. Transformer Block (Quick Overview)

Here’s a simplified view of what happens inside a transformer model:

Input Text → Embeddings → Self-Attention → Feed-Forward → Output

Each of these steps happens many times (layers), allowing the model to learn deep patterns.

5. Why It Matters

These systems help models:

Understand context (e.g., who “he” or “it” refers to)

Handle long sentences

Generate smart, relevant text

Translate languages, answer questions, summarize documents, and more

6. Real-World Analogy

Imagine you're reading a sentence and trying to understand what it means.

Your brain remembers each word (embedding)

You look back at previous words to understand the current one (attention)

You combine that understanding and move forward (layers)

That’s basically what a transformer model does — just with math and vectors instead of neurons.

7. Summary

Token Embedding turns words into numbers.

Positional Embedding adds word order information.

Self-Attention finds which words are important to each other.

Multi-Head Attention looks at the sentence from multiple perspectives.

Transformer Layers repeat the process to learn deeper relationships.