

1. Before the Transformer: How did computers read sentences?

Imagine a computer wants to understand a sentence:

“The cat sat on the mat.”

There were two older methods:

A. RNN (Recurrent Neural Network)

Think of RNN like reading a book one word at a time, slowly.

It reads word 1 → then word 2 → then word 3 → etc.

Like this:

The → cat → sat → on → the → mat

Problems:

Can't read words in parallel

Forget what happened earlier in the sentence

If word 1 is connected to word 10 → too far, it forgets

Very slow to train

B. CNN (Convolutional Neural Network)

CNNs look at **groups of words at once**, not one by one.

Example:

[The cat] [cat sat] [sat on] [on the] [the mat]

Better than RNN, but still:

Needs many layers to understand long sentences

Still slower than needed

Still limited

Summary so far:

Method	How it reads	Problems
RNN	one word at a time	slow, forgets far words
CNN	groups of words	needs many layers

Both methods were too slow and bad at long sentences.

C. Then comes ATTENTION (Simple Meaning)

Attention means:

“Look at the whole sentence at once and focus on the important words.”

Example sentence:

“The animal didn’t cross the road because it was tired.”

Question:

What does “it” refer to?

A normal model struggles.

Attention does this:

“it” → looks at → “animal”

Not → “road”

So attention means:

- ✓ Look anywhere
 - ✓ Jump to important words instantly
 - ✓ No need to go word-by-word
 - ✓ Very fast on GPUs
-

THEN THE PAPER SAYS:

“Why don't we build a whole model using ONLY attention and remove RNNs and CNNs completely?”

This new model is the **Transformer**.

2. Now the paper will make sense (Simple Explanation)

★ The Transformer has two parts:

1. Encoder

Reads the input sentence.

2. Decoder

Creates/output the translation (word by word).

Just like Google Translate.

What does the Encoder do?

Imagine you give it the English sentence:

“The cat sat on the mat.”

The encoder:

- reads the whole sentence at once
 - looks at connections between ALL words
 - understands meaning
 - outputs “smart representations” of the sentence
-

What does the Decoder do?

When translating to French:

“Le chat s’est assis sur le tapis.”

The decoder:

- generates one French word at a time
 - but uses **attention** to look back at the English sentence
 - example: when generating “assis” → look at “sat”
-

What makes Transformers special?

1. They read all words at the same time (parallel)

= very fast to train

2. They understand long sentences easily

= because any word can look at any other word instantly

3. No RNN, no CNN

= simpler, faster, better

This one idea changed EVERYTHING

All modern AI systems (ChatGPT, BERT, Bard, Llama) use the Transformer.

3. Now let me explain the important parts of the paper in beginner English

A. Self-Attention (the main idea)

Self-attention means:

A word decides which other words are important.

Example:

Sentence:

“The dog that lived next door barked loudly.”

When understanding “barked,”
the model focuses on “dog,”
not “door.”

This focus = attention.

B. Multi-Head Attention (simple meaning)

Instead of having 1 way to look at words...

The transformer uses **8 ways at the same time**.

Each “head” looks for something different:

- Head 1 → who is subject of verb
- Head 2 → who owns what
- Head 3 → which words describe a noun
- Head 4 → time relations
- Head 5 → locations
- ...

All heads together = powerful understanding.

C. Positional Encoding (simple meaning)

Attention reads all words at once...
but it does NOT know order.

So we add a signal that represents:

- This is word 1
- This is word 2
- This is word 3
- ...

Like giving each word a tag:

Word: "cat"

Position: 2

This helps the model understand grammar.

D. Feed Forward Network (simple meaning)

After attention mixes information,
a small neural network "cleans up" the representation.

Think: polish the understanding.

E. Masking (simple meaning)

In translation, the model generates next word one by one.

So when predicting the 3rd word,
it should not look at the 4th word.

Masking hides the future.

F. Training tricks

They used:

- Adam optimizer (just a way to adjust learning)
- Learning rate that goes up, then down
- Dropout (randomly turning off some neurons to prevent overfitting)
- Label smoothing (don't be too confident)

These make training stable.

G. Results

The Transformer beat all previous models in:

- translation quality
- speed
- training time

This is why it became the standard.

FINAL SUPER-SIMPLE SUMMARY OF THE WHOLE PAPER

The Transformer is a new model that:

- Reads all words at once
- Lets each word focus on other important words
- Uses many “attention heads” to learn different relationships
- Faster and better than old models
- Achieves state-of-the-art translation

- No RNN, no CNN, only attention

This paper changed AI forever.