ATTENTION IS ALL YOU NEED

Original Paper by
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.
Presented at NeurIPS 2017

Explained with analogies by
Vishal Dhawal – to make the concepts beginner-friendly and engaging.

Original Paper link- https://arxiv.org/pdf/1706.03762

Table of Contents

1. Abstract

- Traditional Teaching (RNNs & CNNs):
Imagine a classroom where the teacher is explaining a foreign language sentence, like translating from English to German.
- The old-school teacher (**like RNN or CNN)** is very strict — they go one student at a time, starting from the first bench to the last. Each student listens, remembers what was said before them, and waits patiently for their turn. If someone didn’t understand a word, they couldn’t ask others directly — they had to depend on the teacher going back or repeating.
- This **process** was **sequential, slow, and tiring**, especially when the **sentences** were **long.**
- Even if the teacher tried to go faster (by using convolutions in CNNs), it was still hard to capture connections between distant words like:
"The girl who lived in Paris loved croissants"
Here, **"girl"** and **"loved"** are related — but far apart!

- Enter **the Transformer Teacher**:
Now imagine a new-age, cool teacher — let’s call them Professor Transformer.
Prof. Transformer believes in group learning. Instead of explaining one-by-one, everyone in class listens and talks to each other at once. If a student hears a word they don’t understand, they can immediately "look at" or "attend to" another student who knows it better — no need to wait. Everyone can connect ideas, discuss meanings, and understand relationships between words, regardless of their position in the sentence.
For example:
Even if the word "croissants" is at the end, a student at the beginning can immediately attend to it and understand its role in the sentence.
This creates a magical atmosphere:
- **Faster learning** — because students aren’t stuck waiting.
- **Smarter understanding** — because they can focus on the most helpful classmates.
- **Parallel progress** — because everyone is working together at the same time.

2. Introduction

Earlier in classrooms, teachers used a very slow method to teach language:
- The teacher would go **student by student**, from front to back. Each student could only speak when it was their turn. If a student forgot something that was said earlier, they had to **rely on the teacher** to **remember and repeat it.** As **sentences** got **longer**, it became **harder** for the teacher to **remember everything** and harder for students to **stay connected to earlier parts**.
Even though teachers tried improving this method — by being a little faster or more organized — it was still based on the same one-by-one approach, which made learning **slow** and **less effective**.
Then, a new idea came in:
“ What if, instead of relying only on the teacher’s memory, we let students pay attention to other students who already know the answer? ”

This helped a lot! Now, if someone was confused, they could just look around and "**attend**" to the right person in class who had the answer. But still — the class was following the same old structure: one-by-one teaching.
Finally, someone said —
"What if we get rid of the one-by-one teaching completely?"
Now imagine this classroom:
Every student can **talk** to **every other student at the same time**.

If someone is confused, they **immediately focus** on the **right classmate** — no need to raise hands or wait for the teacher. Everyone is learning together, helping each other, in parallel. The **teacher just watches** this and **ensures they all stay on topic.** This new way of learning, where attention is the only tool, is what this paper proposes — and guess what?
This method turned out to be **faster, smarter, and more accurate** than all the older methods.

3. Background

Before this new "**all-attention**" teaching method came in, teachers tried two main upgrades to improve how students learned languages:

Some teachers started using Group Discussions in Circles (**like CNNs)**:
Instead of going **student-by-student**, they made **small circles of students**. Each student could talk to their **nearest neighbors**. It was better than the old one-by-one method, but… if one student on the left corner wanted to talk to someone at the right end? They'd have to pass the message through many students.So, **long-distance understanding was still hard**.

Other teachers tried memory notebooks (**like Neural GPUs or ByteNets**):
These teachers gave students notebooks and told them:
"Jot down whatever you learn, we'll come back to it later."
But again, if the student on Page 1 wrote something useful, the student on Page 100 had to flip through a lot to get there.It helped a bit, but it was still slow and not efficient for far connections.

- The **Problem With Both**:
In both styles, if a student needed to understand something related to a **far-away part of the sentence** — they had to go through **many layers of communication.** So the class couldn't truly focus on the big picture quickly.

- **A Better Way:** Students Looking Around Freely (**Self-Attention**)
Now, imagine a setup where each student is allowed to look around the entire classroom freely. If they hear something confusing, they just find the right person — whether they're sitting next to them or across the room. Every student creates their own understanding by observing how others react. This idea is called "**self-attention**" in the paper — but in our class, it just means students being aware of everyone else.
This concept had already been used in other tasks like:
  - Reading comprehension
  - Answering questions
  - Summarizing stories

But no one had yet said:
"Let's make this the only way of teaching."

- And Then It Happened…
One day, this classroom decided to drop all old teaching styles — **no circles, no notebooks, no strict turn-taking**. Instead, they said:
"Let's build the entire classroom system on just looking around and paying attention to the right people."
That's what the **Transformer** is — A class where **everyone learns by freely attending** to each other, without the need for a central teacher controlling every step. And it worked amazingly well.

4. <u>Model Architecture</u>

The Transformer classroom follows the classic two-team setup:
One team, called the **Encoders,** reads and understands the **original sentence** (like English).
The other team, the **Decoders**, writes the **translated version** (like German), one word at a time.
But what's special? Instead of using memory-based teaching (like RNNs), this class relies only on attention — students helping each other by looking around.

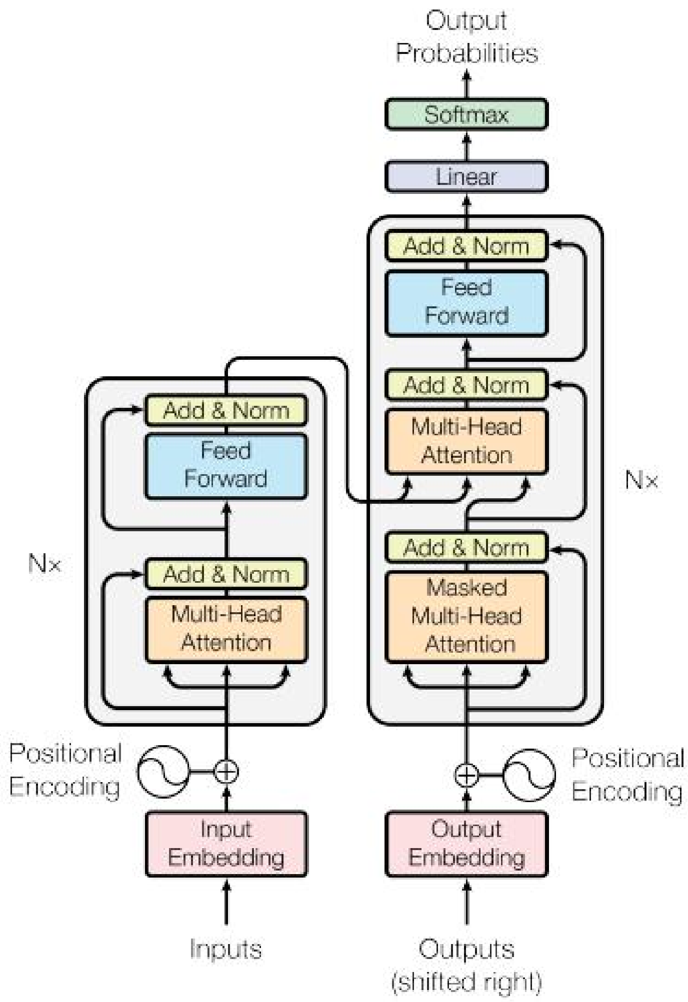

Figure 1: The Transformer - model architecture.

4.1 Encoder and Decoder Teams (Stacks)

Encoder Team – The Readers
The encoder group has 6 layers.
In each layer, students do two things:

Look around at all other students in their team to understand the meaning of the words — this is called **multi-head self-attention**. Think individually after gathering input — that's the feed-forward brain-work. After these two steps, they add their original thoughts back in and normalize things — like double-checking their own notes before passing them on. Each student has a "brain size" of 512 units — enough to hold deep language insights.

Decoder Team – The Writers
The decoder team also has 6 layers.
They do the same two steps: **look around within their team** and **think independently**.
But there's one extra thing they do:

They also **pay attention** to what the **Encoder team figured out** — to help them translate the sentence correctly.
There's one rule though — they can't look at future classmates' answers while writing. They're only allowed to look backwards to avoid cheating. This is done using **masking** — a way of hiding future words during training.

4.2 Attention — The Superpower of This Classroom

Every student in this class uses **attention** to figure things out:
They start with a question (**query**),Look around at who might have useful info (**keys**), And collect the answers (**values**).
They combine all these answers into their final understanding — a weighted average where more useful classmates get more attention.

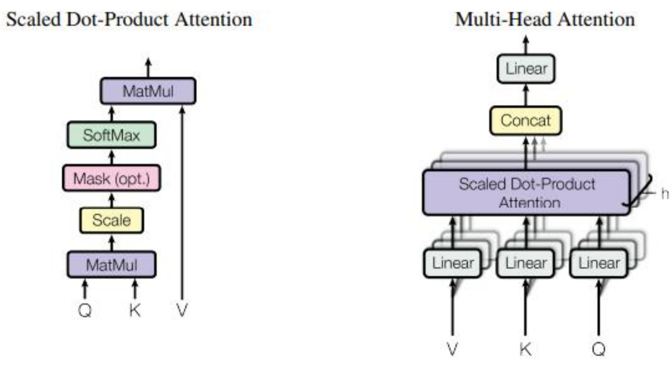

Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

4.2.1 Scaled Dot-Product Attention

Let’s say a student asks a question and listens to everyone’s answers.
To decide how much to listen to each classmate, the student:
Scores how **similar** their question is to each classmate’s knowledge (**query × key**),
Divides the **score by a scaling factor** (so scores don’t explode), Applies **softmax** to turn **scores** into **attention weights** (like “who to trust more”).
They then take a weighted average of everyone’s answers (**values**) — and that’s what they learn.

4.2.2 Multi-Head Attention

Instead of listening with just one mindset, each student listens in 8 different ways at once.
Think of it like 8 different “thinking caps” — each head focuses on different types of patterns.
After listening through all 8 heads, they combine everything into one final understanding.
This allows students to look at **multiple aspects of language at the same time — grammar, word meaning, sentence flow,** etc.

4.2.3 Where Attention Is Used

The Transformer class uses attention in three main places:
Encoder Self-Attention:
Students in the encoder team look around within their team to better understand the sentence.
Decoder Self-Attention:
Students in the decoder team look at their own team — but only at earlier students (no future peeking!).
Encoder-Decoder Attention:
Decoder students also look at the Encoder team to understand what they’re supposed to translate.

4.3 Feed-Forward Thinking (After Attention)

After each attention round, students don’t just stop there — they process what they learned on their own. Each student passes their **current thoughts through a two-step** mini brain module:
Think and transform (using math layers),
Apply ReLU (to make it smart and non-linear),

Output the refined thought.
This helps them go beyond just what others said — they develop their own processed version of the idea.

4.4 Embeddings and Word Predictions

Before all this attention starts:
Every word is turned into a vector (like assigning a brain to each word).
These embeddings have **512 dimensions** — enough to capture meaning.
At the end, once the decoder has thought everything through, it predicts the next word using a final **softmax** — like picking the most likely next word from a word-list.

Bonus trick: The model shares the same weight matrix for input and output word mappings — to save memory and learn smarter.

4.5 Positional Encoding – Knowing the Order

Now here's the problem:
Since students are just looking around freely, they might forget who said what and when. So we give them a sense of position — like telling them where each student sits in the row. To do this, we add a pattern of waves (sine and cosine functions) to each word's brain — this helps them figure out the order of words.This way, even without step-by-step teaching, the model still understands order — like who came before whom in the sentence.That's how the Transformer classroom works — by letting students think together, focus smartly, and write creatively — all powered by attention, not memory.

5. Self-Attention

Now imagine you have to choose the best way to run a language class — and you have three goals:

Goal 1: Keep it Fast
Goal 2: Let Everyone Work in Parallel
Goal 3: Help Students Connect Distant Ideas Easily

Let's see how different classroom styles compare:

Old-Style: One-by-One Teaching (Recurrent)

The teacher goes from one student to the next, one at a time. Students have to wait their turn to speak or ask questions. If a student wants to connect something from 10 students earlier, it takes time. Slow and no parallel work. Works fine for short sentences but struggles with long ones.

Group Circles (Convolutional Layers)

Students talk to a small group of neighbors. To connect with a far-away student, the message has to go through many middle students. It's better than the one-by-one method, but still takes effort to connect distant thoughts. Also costly when trying to go deep.

The Self-Attention Class (Transformer Way)

Everyone can instantly look at everyone else, no matter where they sit.
A student can easily connect the first and last word of a sentence in just one step.
All students think at the same time — super parallel! Much faster and smarter.

More Transparent Thinking

In the Self-Attention classroom, we can actually see which student looked at whom while answering — like watching thought bubbles.
This helps us understand:

Who influenced the decision,
What part of the sentence mattered most,
And how meaning was formed.

So not only is it fast and smart, but it's also more interpretable — like watching learning in action.

6. Training the Transformer

Once the classroom (model) was designed, it had to be trained — like preparing students for exams. So here's how the teachers trained their classroom of smart, attention-powered students:

6.1 Training Data and Batching — "Preparing Practice Worksheets"

To help students learn, the teachers gave them millions of bilingual sentence pairs:
For **English-German**, about **4.5 million sentence pairs**.
For **English-French**, around **36 million sentence pairs**!
But before handing them out: The sentences were broken into small word pieces (like syllables) using **byte-pair** encoding, so students could handle rare and new words better.

Then, to make training fair and efficient, worksheets were grouped such that sentences in each batch were roughly the same length — easier to manage together.Each batch had about 25,000 input words and 25,000 output words.

6.2 Hardware and Schedule — "How Long Did Training Take?"

The training was like daily coaching sessions for the class, held on 8 GPUs.
The **base classroom** (smaller model) learned everything in about **12 hours** (100,000 steps).The **big classroom** (larger model) trained longer — **for 3.5 days** (300,000 steps) — but came out sharper and more accurate.
Each step was fast:

Just 0.4 seconds per step for the base model,
Around 1 second for the big model.

6.3 Optimizer — "Smart Coaching Strategy"

Instead of using a fixed learning pace, teachers used a smart training schedule with the Adam optimizer. Here's how it worked: At the beginning, students learned slowly, gradually speeding up — like warming up. After a point (4000 steps), the learning slowed down again, so they could fine-tune their knowledge. This warmup-cooldown cycle helped avoid burnout or overfitting, and kept learning steady and stable.

6.4 Regularization — "Avoiding Overconfidence & Laziness"

To keep the class in shape and avoid bad habits, teachers used two key tricks:
Dropout (Random Surprise Test):
Sometimes, during practice, some neurons were randomly dropped — like hiding clues from students so they don't get too dependent.

This helped prevent overfitting.

Label Smoothing (Soft Grading):
Instead of giving absolute answers during training, the teachers allowed a little uncertainty — like saying,
"The correct answer is most likely this... but we're a bit open to other options too."

This made students more adaptable and improved accuracy in real-world testing.

Final Outcome — "How Well Did the Class Perform?"

On their final test — the WMT 2014 translation exams:
The Transformer base model scored higher than previous students — and it learned much faster.

The big Transformer model outperformed even the best previous ensemble methods.
And it did all this with less training cost (fewer GPU hours).

Model	EN-DE BLEU	EN-FR BLEU	Cost (FLOPs)
ByteNet	23.75	—	—
GNMT + RL	24.6	39.92	Very high
ConvS2S	25.16	40.46	Very high
Transformer (base)	27.3	38.13	3×10^{18}
Transformer (big)	28.4	41.8	2.3×10^{19}

So in short — the Transformer class studied smarter, faster, and outperformed older methods.

7. Results

After all the training and regular practice sessions, it was time for the Transformer students to take their exams — in multiple subjects. Let's see how they did.

7.1 Machine Translation – "The Language Test"

The students took their final test on translating English to German and French, and they aced it! The big Transformer model scored **28.4 BLEU** on English-to-German — 2 points higher than any past model. Even the base model, trained faster and cheaper, beat many older, heavier models. On English-to-French, the big Transformer scored **41.0 BLEU**, better than all previously known single models — and did so using only ¼ the effort (training cost).
Exam Conditions:

> Students' answers were selected using beam search (like picking the best possible set of words). For more stable results, final answers were averaged over the last few best-performing checkpoints.

Key takeaway: The Transformer class showed that attention-based learning is faster, cheaper, and smarter than older RNN- and CNN-based classrooms.

7.2 Model Variations – "Trying Different Classroom Setups"

The teachers wanted to explore: "What happens if we change the number of students or their learning style?" So they ran mini-experiments by tweaking classroom settings:

A. Changing Heads

Using just 1 attention head made performance worse.
Sweet spot: 8 heads — students could focus on multiple things at once without overload.

B. Reducing Attention Size

Making the students' focus range smaller (low dk) hurt performance — they couldn't judge connections properly.

C. Scaling the Brain (Model Size)

Larger brains (more neurons) helped — smart students performed better when given more thinking power.

D. Dropout Settings

No dropout led to overfitting — students got overconfident.
A little dropout improved generalization.

E. Position Knowledge
Whether students were told their positions via sine waves or learned embeddings, the results were nearly the same. So, the Transformer classroom is flexible — but works best with a balanced setup of heads, size, and regularization.

7.3 English Constituency Parsing – "The Grammar Test"

After proving themselves in translation, the teachers wondered:
"Can our students handle grammar analysis too?"
They gave the class a grammar structure test called Constituency Parsing, where the model had to break down sentence structures like subject, verb, object, etc.

Test details:

Only 40K training examples (a small dataset).
No special tuning — just reused most of the translation model settings.

Results:

Even with only 4 layers, the Transformer did better than most past models.
In semi-supervised learning (with extra data), it reached **92.7 F1** score, outperforming many well-known **parsers**.

What's impressive?

The Transformer didn't need complex grammar rules or big changes — it generalized well, just by using attention and parallel processing.

Final Report Card Summary:

Task	Transformer Result	Outperformed?
EN→DE Translation	28.4 BLEU (Big model)	All previous models
EN→FR Translation	41.0 BLEU (Big model)	All previous singles
Grammar Parsing (WSJ)	91.3–92.7 F1	Most older models

So overall, the Transformer class didn't just top their main subject (translation), but also excelled in side subjects (grammar parsing) — and did so efficiently, flexibly, and impressively.

Attention Visualizations

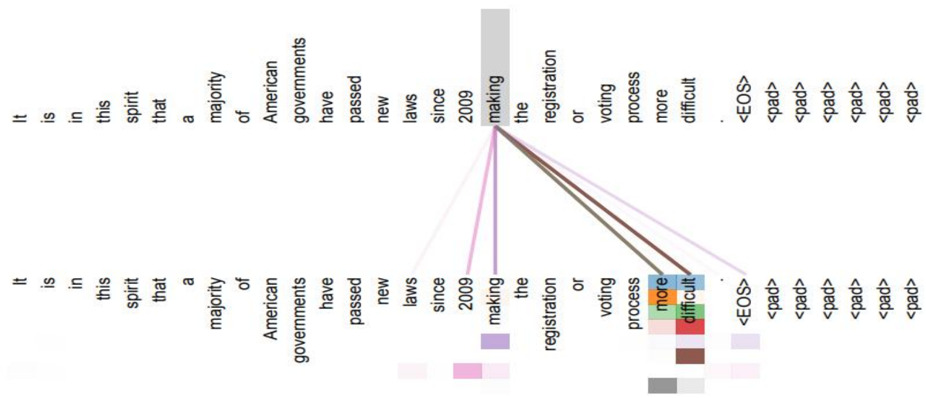

Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word 'making'. Different colors represent different heads. Best viewed in color.

8. <u>Conclusion</u>

So, how did the Transformer classroom change everything?
Previously, most classrooms (models) relied on step-by-step memory techniques — like RNNs and CNNs — where students learned one at a time or in small neighbor groups.
But the Transformer flipped the teaching style completely:

> No more slow memory-passing. Just pure attention.

Now, every student could:
Look at all other students instantly (self-attention), think independently but in sync (parallel processing), and work together through multi-headed perspectives (multi-head attention).

What Did It Achieve?

Faster training than older models — saving time and computing power.
Top scores on language translation tests (English-German and English-French).
Even beat past group-taught ensemble models — all by itself!

What's Next?
The teachers (researchers) are now dreaming bigger:
- Applying this attention-based learning to other subjects like images, videos, and audio.
- Exploring local attention — like letting students focus only on nearby seats for longer documents or big data.
- Making text generation faster by reducing how much it depends on past steps.

Final Thought

The Transformer showed that when students are given the freedom to look around, collaborate, and think in parallel, they learn faster, better, and smarter.

Special thanks to the AI research community for making papers like "Attention Is All You Need" accessible to everyone.

This analogy-based explanation was inspired by the desire to simplify, share, and spark curiosity in others.