Transformers: Understanding Parallel Intelligence

From Zero to ChatGPT - A BSc Journey

Week 5: Transformers

How Google Reads Your Mind

Try this: Type in Google: "How do transformers..."

Google instantly suggests:

- "...work in machine learning"
- "...process language"
- "...learn from data"

The Mystery:

- Google reads ALL your words at once
- Not word-by-word like old systems
- Understands context instantly

```
How do transformers
_______________________________
...work in machine learning
...process language
...learn from data
...handle attention
```

Question: How does it understand whole sentences simultaneously?

Discovery 1: Words Live in Space

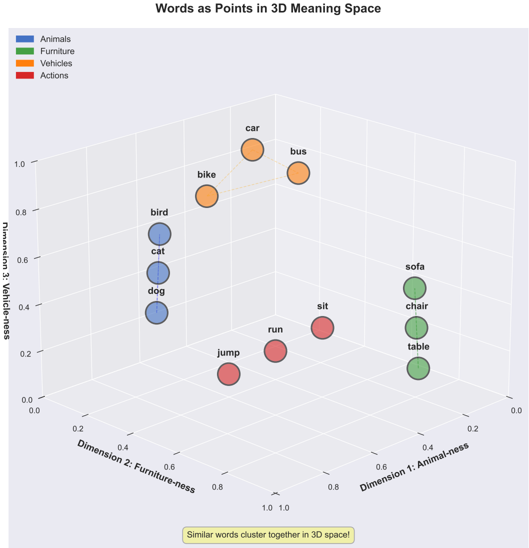
Think about GPS coordinates:

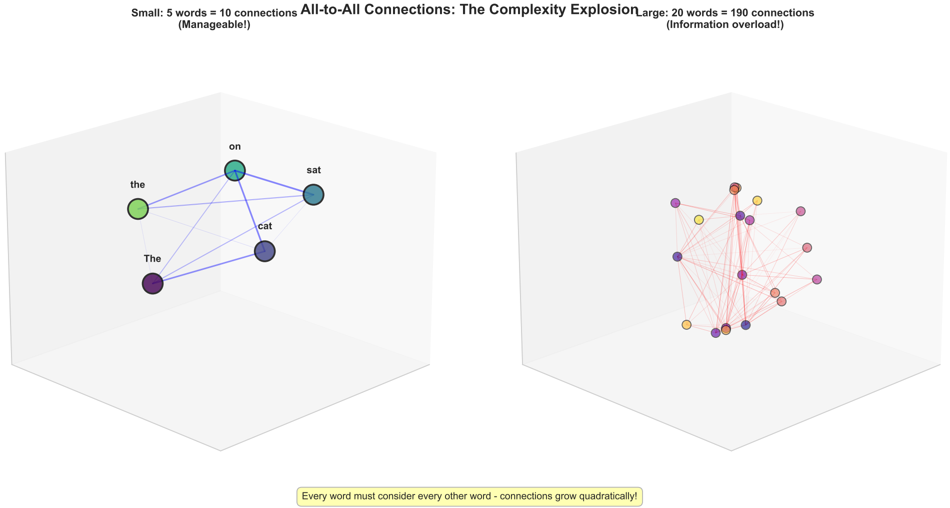
- Paris: (48.8°N, 2.3°E, 35m altitude)
- London: (51.5°N, 0.1°W, 11m altitude)
- Similar cities are nearby in space

Words work the same way!

- "cat": [0.7, 0.2, 0.5] in meaning space
- "dog": [0.8, 0.3, 0.4] (nearby - similar!)
- "car": [0.1, 0.9, 0.2] (far - different!)

This is called: Word Embeddings





In "The cat sat on the mat":

The Explosion:

Computational Explosion:

Words	Connections	Memory	Time
10	45	0.001 GB	0.1 sec
100	4,950	0.1 GB	10 sec
1000	499,500	10 GB	1000 sec
10000	50M	1000 GB	28 hours!

The Challenge: How to find what matters in this chaos?

Forward Question: Can we be selective instead of exhaustive?

Visualization of Growth:
[Exponential explosion chart would go here]

Crisis: Processing everything is impossible at scale!

First Attempt: Connect Everything

The Naive Idea:

- Connect every word to every other
- More connections = better understanding?
- Like everyone in a room shouting at once

Implementation:

- Compute all pairwise relationships
- Store in giant matrix
- Hope for the best

What Actually Happens:
[Chaos visualization would go here]

Computing All Relationships

For "The cat sat":

	The	cat	sat
The	1.0	0.3	0.2
cat	0.3	1.0	0.7
sat	0.2	0.7	1.0

Each number = relationship strength

- "cat" - "sat" = 0.7 (strong!)
- "The" - "sat" = 0.2 (weak)

Matrix grows quadratically:

- 3 words = 3×3 matrix
- 100 words = 100×100 matrix
- 1000 words = 1,000,000 numbers!

Complete matrix for every sentence!

SUCCESS! (On Simple Cases)

Works Great For:

- "The cat" $\rightarrow$ predicts "sat" ✓(95%)
- "Water is" $\rightarrow$ predicts "wet" ✓(92%)
- "Birds can" $\rightarrow$ predicts "fly" ✓(89%)
- "Coffee tastes" $\rightarrow$ predicts "good" ✓(91%)

Why it works:

- Few connections to track
- Clear patterns visible
- No information overload yet

Celebration!
We can predict words!

The approach seems valid!
Let's scale it up!

FAILURE: Signal Lost in Noise

Performance Collapse:

Length	Signal	Noise	Accuracy
10 words	3	7	85%
50 words	5	45	42%
100 words	8	92	18%
500 words	15	485	3%

The Pattern: More words = More noise!

What Goes Wrong:

- Important connections drowned out
- 95% of connections irrelevant
- Can't find what matters
- Like finding needle in haystack

Diagnosis: We need to be SELECTIVE, not exhaustive!

How Do Humans Actually Read?

Try this experiment:
Read: "The black cat sat on the soft red mat"

When reading "mat", did you:

- Look at EVERY word equally? $\times$
- Or focus on specific words? ✓

You actually focused on:

- "on the" (35%) - location pattern
- "sat" (20%) - what's happening
- "cat" (15%) - who's doing it
- Ignored "black", "soft", "red" (5% each)

Key Realization:
Humans SELECTIVELY FOCUS!

We don't process everything equally.

We spotlight what matters!

The Insight: What if computers could learn WHERE to look?

The Hypothesis: Selective Attention

Breaking It Down: Attention as Percentages

For "The cat sat on the ___":

When predicting next word, look at:

- "on": 35% attention
- "the": 25% attention
- "sat": 20% attention
- "cat": 15% attention
- "The": 5% attention

Key Properties:

- Percentages sum to 100%
- Higher % = more important
- Learned from data

Visualization:
[Pie chart of attention distribution]

These percentages are called **Attention Weights**

The Math: How Similar Are Two Words?

Remember: Words are vectors!

- Query: "What follows 'on the'?"
- Key: "I am a furniture word"

Dot Product = Similarity:

- Query vector: [0.8, 0.2]
- Key vector: [0.6, 0.4]
- Dot product: $0.8 \times 0.6 + 0.2 \times 0.4 = 0.56$

Higher number = More relevant!

Geometric Intuition:
[Vector angle diagram]

Key insight:

- Similar direction = High dot product
- Opposite direction = Low dot product
- Same principle in 512 dimensions!

Every word asks three questions:

1. Query (Q): "What am I looking for?"

- Word "mat" asks: "Need location info"

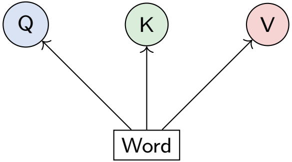
2. Key (K): "What do I contain?"

- Word "on" says: "I have location info"

3. Value (V): "What info do I provide?"

- Word "on" gives: "Preposition pattern"

The Process:



Transform to 3 spaces

Q and K determine attention weights
V provides the actual information

Step-by-Step: Computing Attention

Example: "mat" attending to all words

Step 1: Compute relevance (Q·K)

- Q("mat") · K("on") = 0.8
- Q("mat") · K("the") = 0.6
- Q("mat") · K("sat") = 0.4
- Q("mat") · K("cat") = 0.3
- Q("mat") · K("The") = 0.1

Step 2: Convert to percentages (softmax)

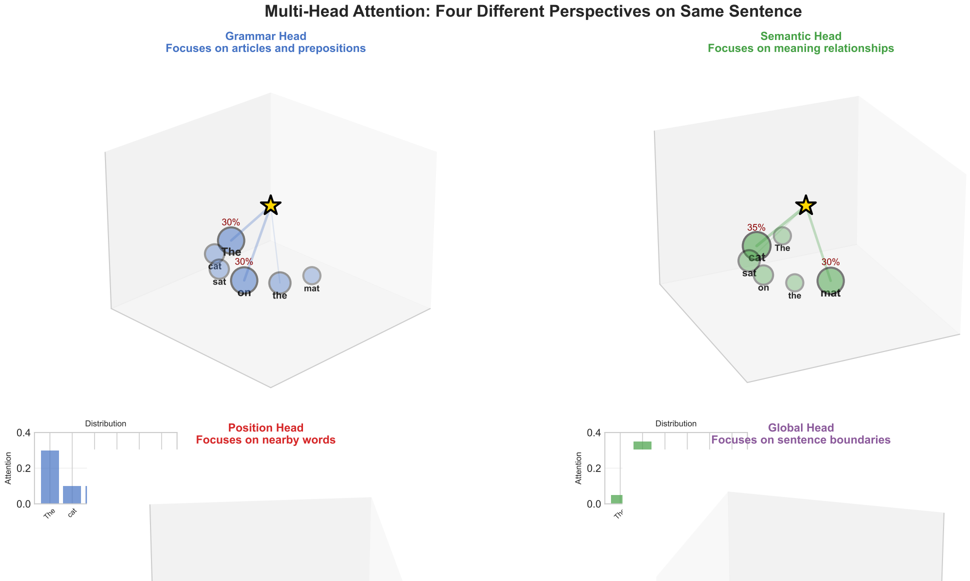
- "on": 35%
- "the": 27%
- "sat": 18%
- "cat": 14%
- "The": 6%

Step 3: Weighted combination

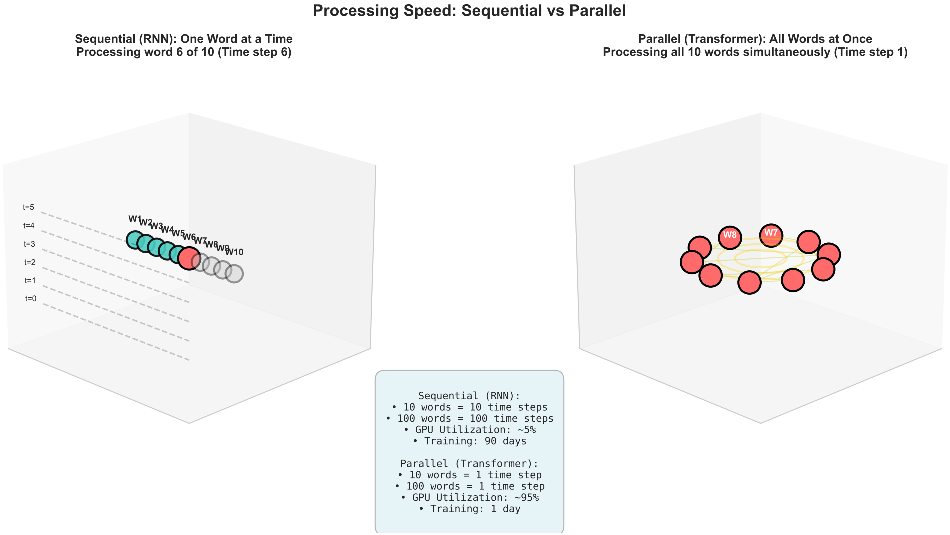
$$\begin{aligned}\text{Output} = {} & 0.35 \times V(\text{“on”}) + \\ & 0.27 \times V(\text{“the”}) + \\ & 0.18 \times V(\text{“sat”}) + \\ & 0.14 \times V(\text{“cat”}) + \\ & 0.06 \times V(\text{“The”})\end{aligned}$$

Result: Context-aware representation that knows "mat" likely follows "on the"!

Multiple Perspectives: 4 Different Experts



The Speed Revolution: Everything at Once



Preserving Order: Where Words Live

The Problem:

- Parallel processing loses order
- "cat sat" same as "sat cat"?
- Need position information

The Solution: Positional Encoding

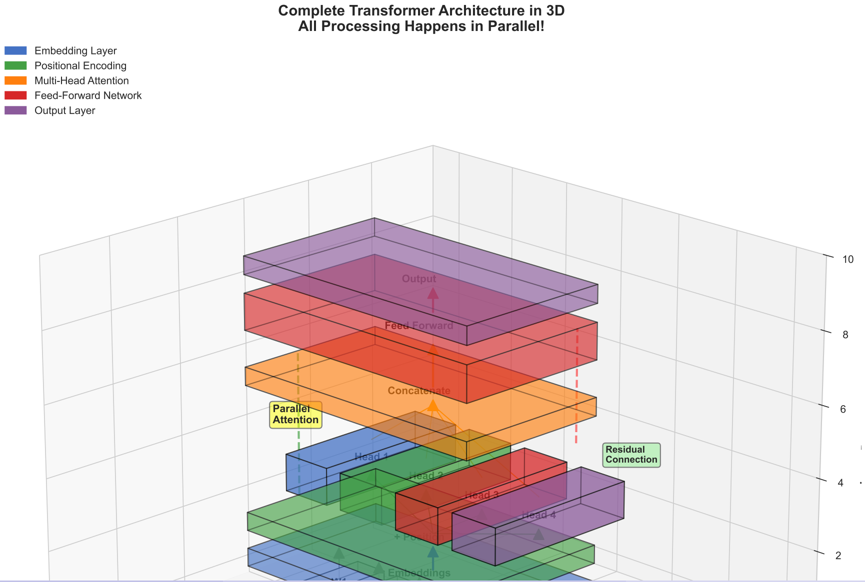
- Add unique wave patterns
- Position 0: Low frequency
- Position 50: Mixed frequency
- Position 100: High frequency

Each position gets unique signature!

Sine/Cosine Waves:
[Positional encoding visualization]

Words know their order without sequential processing!

Everything Together: The Transformer



Proof It Works: Real Results

Performance Comparison:

Length	RNN	Transformer	Gain
5 words	95%	96%	+1%
20 words	67%	89%	+33%
50 words	31%	84%	+171%
100 words	12%	81%	+575%

Pattern: Massive gains on long text!

Why the improvement:

- No information bottleneck
- Direct access to all words
- Parallel computation
- Multiple perspectives

Validation: The hypothesis works!

The Revolution: 2017-2024

Timeline of Innovation:

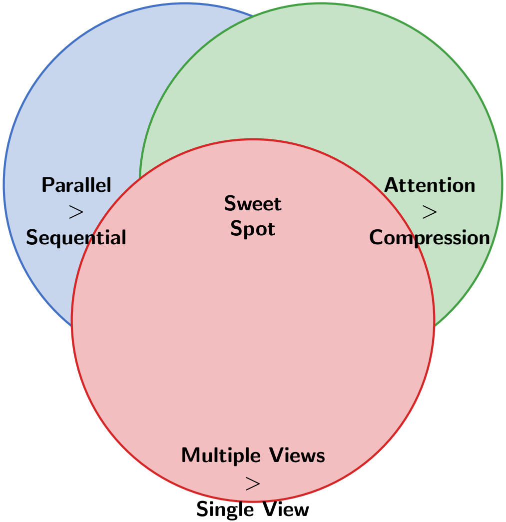
- 2017: Original Transformer paper
- 2018: BERT (understanding text)
- 2019: GPT-2 (generating text)
- 2020: GPT-3 (175B parameters)
- 2022: ChatGPT (conversation)
- 2023: GPT-4 (multimodal)
- 2024: Claude, Gemini, Llama 3

Why it exploded:

- Training 100x faster
- Scales to billions of parameters
- Works on any sequence data
- Same architecture everywhere

One architecture conquered all of AI!

The Three Core Principles



What makes transformers special:

Where You Use Transformers Every Day

Text:

- ChatGPT conversations
- Google search
- Gmail autocomplete
- DeepL translation

Code:

- GitHub Copilot
- Cursor
- Replit AI

Multimodal:

- DALL-E (text to image)
- Whisper (speech to text)
- GPT-4V (vision)
- Sora (text to video)

Science:

- AlphaFold (protein folding)
- Weather prediction
- Drug discovery

All using the same transformer architecture!

Check Your Understanding

You now understand:

- ✓ Words live in high-dimensional space
- ✓ Every word connects to every other
- ✓ Attention selects what's relevant
- ✓ Multiple heads = multiple perspectives
- ✓ Parallel processing enables scale
- ✓ Position encoding preserves order
- ✓ Same architecture powers ChatGPT

Quick Quiz:

1. Why are transformers fast?
   Parallel processing
2. What does attention do?
   Selects relevant information
3. Why multiple heads?
   Different perspectives

Congratulations! You understand the technology behind ChatGPT!
From zero knowledge to transformer expert in 25 slides!

Your Journey Continues

This Week's Lab:

- Build attention mechanism
- Implement multi-head attention
- See the magic happen

Next Week: Pre-training

- How to train on internet scale
- Why size matters
- The emergence phenomenon

Key Takeaway:

Transformers =

Parallel Attention
on All Words
with Multiple Perspectives

Questions?