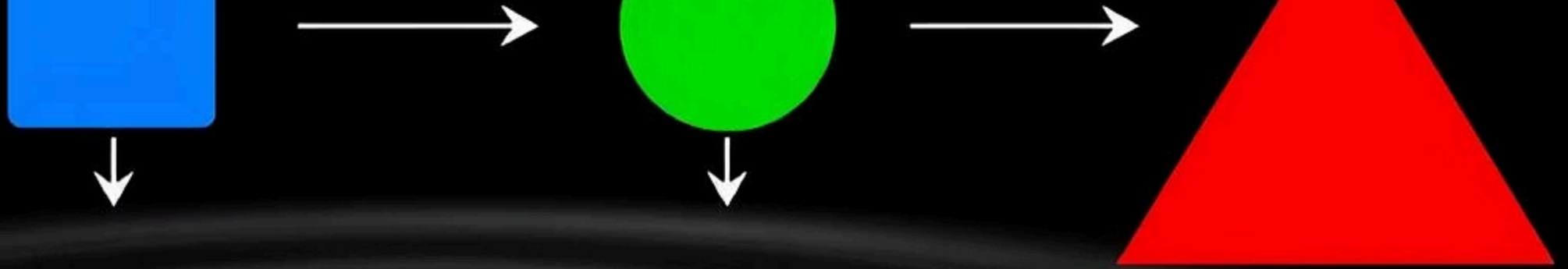


Unveiling Sequential Models: From RNNs to Transformers

Understanding how machines process sequential data is key.

Text, speech, and time-series data present unique AI challenges.

This presentation covers the evolution of smarter models, step-by-step.



Recurrent Neural Networks (RNNs): The First Step



Sequential Processing

RNNs process sequences one element at a time.



Hidden State Memory

They maintain a "hidden state" capturing past information.



Shared Weights

Weights are shared consistently across time steps.



RNN Disadvantages: Short Memory & Instability

Vanishing/Exploding Gradients

Difficulty learning long-term dependencies. Information fades or explodes.

Short-Term Memory

Struggles with sequences exceeding 50-100 steps.

Sequential Bottleneck

Must process data one by one, leading to very slow operation.

Long Short-Term Memory (LSTMs): Enhancing Memory

Introduce Gates

LSTMs use gates to control information flow.

Output Gate

Determines the final output.



Cell State Highway

A "cell state" acts as a data highway.

Forget Gate

Decides what information to discard.

Input Gate

Controls what new information is added.

LSTM Advantages & Remaining Limitations

Advantages

Mitigate vanishing gradients effectively.

Better long-term memory, handling hundreds of steps.

Widely used in speech recognition and machine translation.

Limitations

Still process information fundamentally sequentially.

Cannot effectively capture very long-range global dependencies.

Scalability remains a challenge for extremely long texts.

Attention Mechanism: Focusing on Relevance



Focus Concept

Model focuses on important input parts.



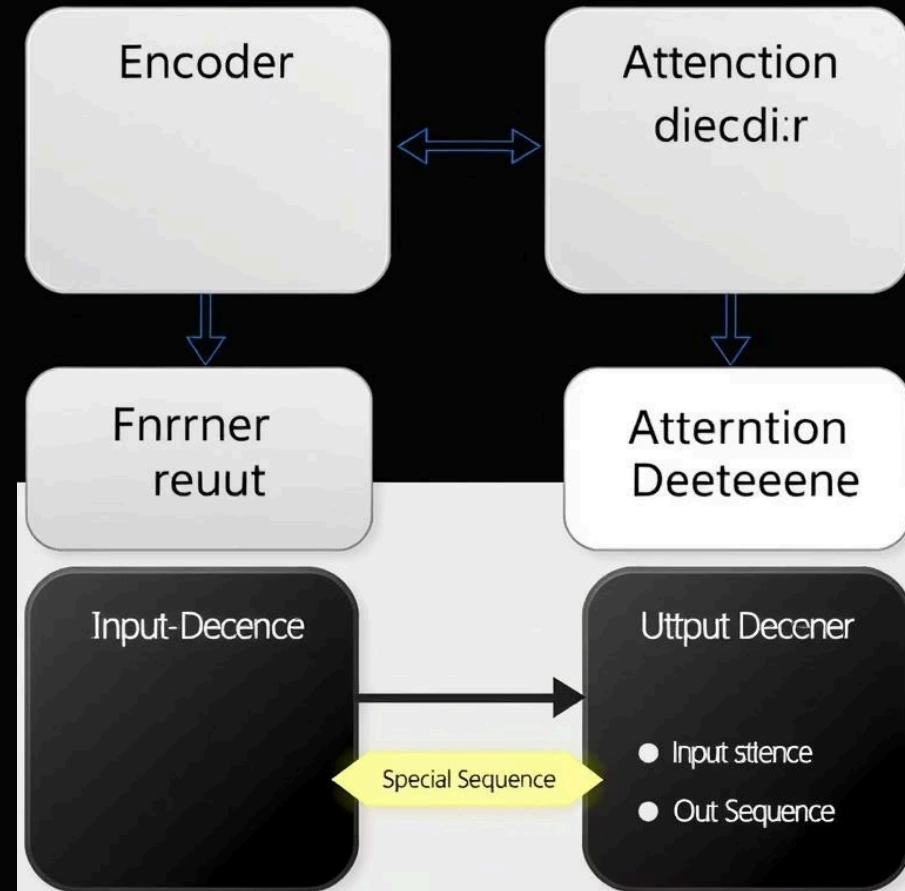
Assigns Weights

Assigns importance weights to elements.



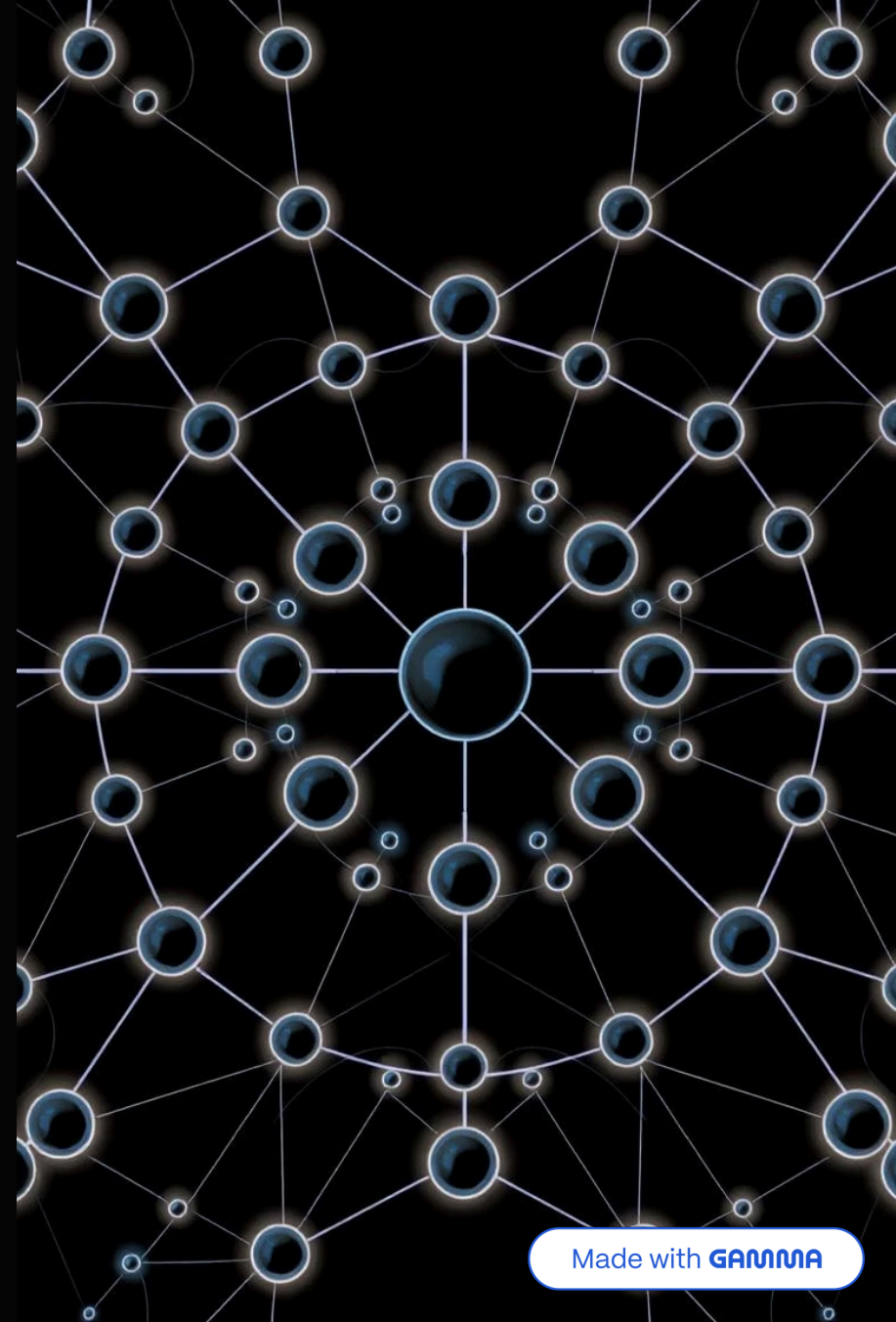
Seq2Seq Improvement

Introduced for improving Seq2Seq models.



How Attention Overcomes Limitations

- Breaks Bottleneck
No longer compresses all info into one vector.
- Direct Access
Connects output to all input elements directly.
- Parallel Processing
Enables looking at input parts simultaneously.
- Long-Range Solution
Solves long-range dependency for specific task pairs.



Transformers: Attention is All You Need

1

Revolutionary

Completely removes recurrence.

2

Self-Attention

Relies solely on self-attention.

3

Parallel Processing

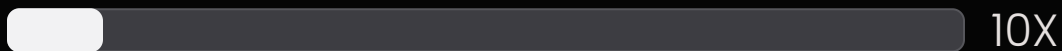
Processes entire sequences in parallel.

4

Positional Encoding

Adds information about token order.

Transformer Advantages: Speed & Scale



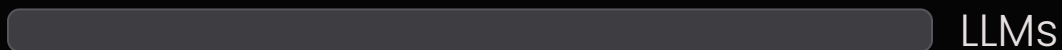
Faster Training

Full parallelization enables rapid training.



Longer Sequences

Excels at very long-range dependencies.



Basis for LLMs

Powers BERT, GPT-3, and other large models.



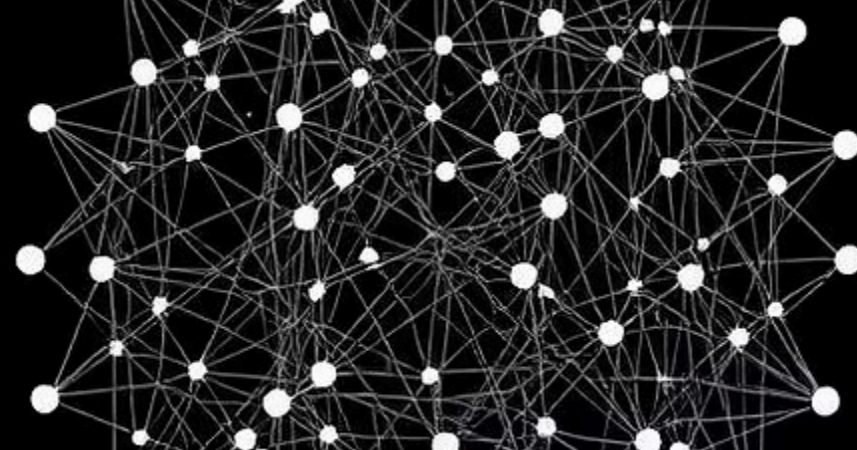
Dominates NLP

Leading the way in natural language tasks.



1986:
Backpropagationt

1997.
LSTM



Conclusion: The Evolution of Sequence Models

Models evolved from sequential (RNNs, LSTMs) to parallel (Transformers).

Each iteration successfully addressed critical limitations.

Transformers enable unprecedented scale and performance breakthroughs.

Future research will focus on efficiency and expanded capabilities.