Step 6 – Stack Layers (Nx)

Each encoder and decoder in the Transformer is created by **repeating the same structural block N times**. This layering strategy makes the model **deeper, more powerful, and capable of understanding complex patterns** in language and data.

Understanding the "Nx" Notation

What Does Nx Mean?

The notation "Nx" indicates that we **repeat the same architectural block N times**, stacking identical structures vertically to create depth.

Each repeated block contains four key components that work together:

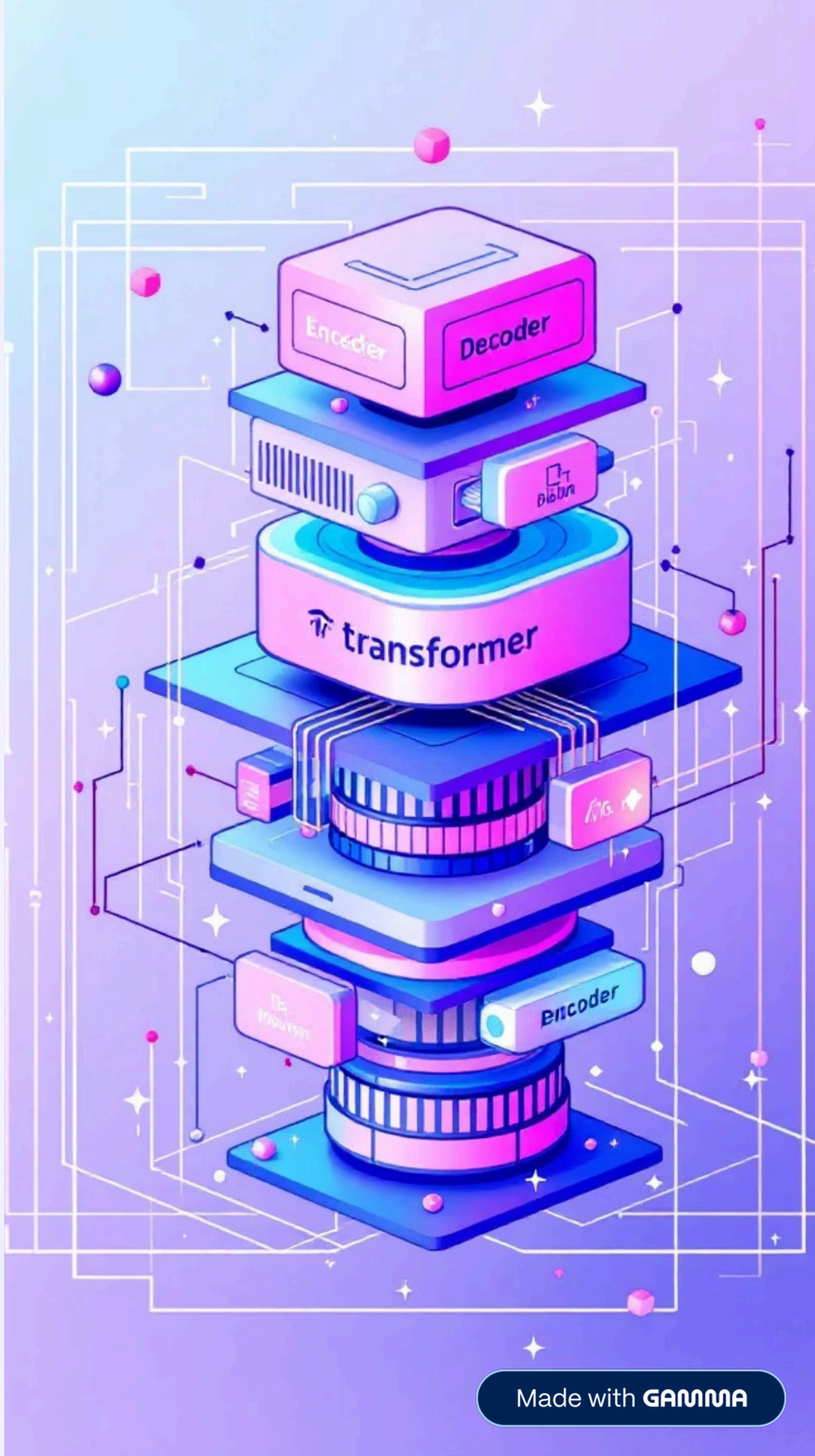
- Multi-Head Attention mechanism
- Add & Norm layer
- Feed Forward Network
- Add & Norm layer (applied again)

Common Layer Counts

Transformer Base Model: Uses N = 6, creating 6 encoder layers and 6 decoder layers for balanced performance.

Advanced Models: Larger architectures like BERT, GPT-3, and modern LLMs use significantly more layers:

- BERT-Base: 12 layers
- GPT-3: 96 layers
- Some models: Up to 24+ layers



Why Do We Repeat Layers?

Each successive layer in the stack **refines and deepens the understanding** developed by previous layers. This hierarchical learning mirrors how humans process complex information—starting with basics and building towards sophisticated insights.

1

Early Layers

Learn: Basic word meanings and grammatical structures

Example: Identifying simple phrases like "The cat sat"

2

Middle Layers

Learn: Relationships between words and syntactic dependencies

Example: Understanding subject-verb connections and word order

3

Deeper Layers

Learn: Abstract reasoning, contextual nuances, and semantic meaning

Example: Comprehending complex relationships like "who trusted whom"

Key Insight: More layers enable progressively deeper understanding of language structure, context, and meaning.

Analogy: Processing Like Human Reasoning

> "The professor who the student who the dean met trusted left the university."

01

Layer 1: Basic Parsing

Identifies fundamental grammar—nouns, verbs, and simple phrase structures within the sentence.

02

Layer 2: Relationship Mapping

Determines connections between clauses, understanding nested relationships like "who trusted whom".

03

Layer 3+: Complete Understanding

Constructs the full semantic meaning by resolving all dependencies and building a coherent interpretation.

Each layer represents **one step in a chain of reasoning**—similar to how humans parse complex sentences by breaking them into manageable pieces and gradually assembling the complete picture.

Stacking ≠ Recursion

How Layer Stacking Works

Although the architectural structure is repeated N times, it's crucial to understand that **each layer is fundamentally independent**.

Every layer in the stack has:

- **Unique learnable weights** that are optimized during training
- **Independent parameters** (not shared or reused across layers)
- **Dedicated processing** of the output from the previous layer

The Chain of Experts

Think of stacked layers as a **chain of specialized experts**, where each expert receives input from the previous one and adds their own refined perspective.

This design allows each layer to learn different aspects of the data, progressively refining the representation at each step.

Quick Summary

Nx Notation

Represents repeating the same encoder or decoder block **N times** to create a deep architecture.

Typical N Value

The Transformer Base model uses **N = 6 layers** for both encoder and decoder stacks.

Core Purpose

Enables **deeper reasoning and progressively better understanding** of input data and context.

Layer Independence

Each layer has **unique weights and transformations**—not recursive or parameter-shared.

Final Takeaway: More layers create context-aware, high-level representations that make Transformers remarkably intelligent!