Detailed Analysis of the Encoder

Single Encoder Layer

1

Multi-Head Self-Attention

Captures relationships between all positions

+ Residual → Layer Norm

2

Feed-Forward Network

Position-wise transformation

+ Residual → Layer Norm

This structure is **repeated 6 times** (stacked layers)
Each layer has **different parameters**

Residual Connections

Around **each sub-layer** to help gradient flow

Layer Normalization

Applied **after each sub-layer** for stability

Same Structure

All layers share **identical structure** but have different parameters

Output dimension **preserved** throughout:
d_model = 512