Attention Mechanisms as Explanations: Transformer Visualization

Self-Attention Matrix (BERT Example)

	This	movie	is	great	!	[SEP]
This	**0.42**	0.18	0.08	0.15	0.05	0.02
movie	0.20	**0.35**	0.10	0.22	0.08	0.05
is	0.08	0.15	0.18	**0.45**	0.12	0.02
great	0.10	0.25	0.18	**0.38**	0.07	0.02
!	0.05	0.08	0.05	0.28	**0.48**	0.06
[SEP]	0.12	0.10	0.08	0.15	0.10	0.25

🔍

Self-Attention

Shows token-to-token importance in context

→ Query-Key-Value mechanism

→ Context-aware weights

→ Bidirectional relationships

🧩

Multi-Head Attention

Different heads capture different aspects

→ Parallel attention layers

→ Diverse relationship types

→ Aggregate for full picture

🔄

Attention Methods

→ **Rollout:** Aggregate across layers

→ **Flow:** Trace information path

🟥 High (>0.35) 🟥 Med-High 🟥 Medium 🟦 Low ⬜ Very Low

Token Attention Example: "movie" attending to:

This **movie** is **great** !

→ **Gradient:** Combine with gradients

⚠ Important Limitation

Attention ≠ Explanation
Recent research shows attention weights don't always reflect true model reasoning. Combine with gradient-based methods for better insights.