# Transformers learning Graphs

Becsó Gergely

May 2024

# In this presentation...

I would like to talk about:

- Motivation of interpretability and interpretability in neural networks
- Basic concept of transformers
- Positional Encoding
- Application on graph problems
- Results

# Interpretability of transformers

Today transformers are

- the most researched models in AI,
- the basis of LLMs, with huge impact on everyday life
- still not well understood.

Interpretability researches are focusing on the inner mechanisms.

# Current Approach

Big models with huge parameters space are harder to work with and are complex: let's use toy models.

NLP problems are less understood than many math problems: let's use graph problems - shortest path.

**Toy models** with limited parameter space, **well understood problem**: hopefully easier to interpret results

# Network Architecture

Keeping the essence of a transformer network:

- embedder
- positional encoding
- attention
- aggregation layer

# Positional Encoding

Classical way: sinusoidal encoding

Cost ineffective way: concatenating to the embedding

RoPE: Rotary Positional Embedding - partitioning the embedding space into the product of 2D spaces and rotating in each.

# Data Generation

Shortest Path Problem tokenized:

$G = (V, E),\ V = \{1, \ldots, n\}$

Question of $G, u, v$ without context:

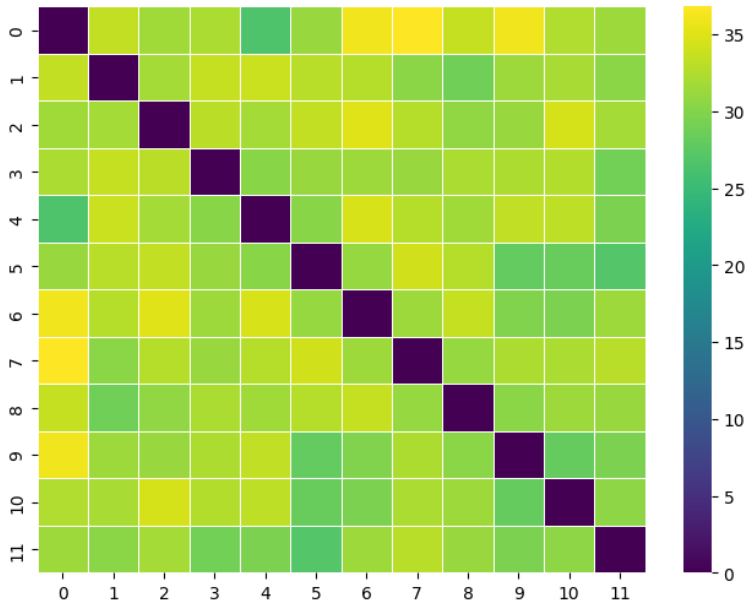
$$u, 0, v, 0,\ \textit{n noise symbols}$$

Expected answer:

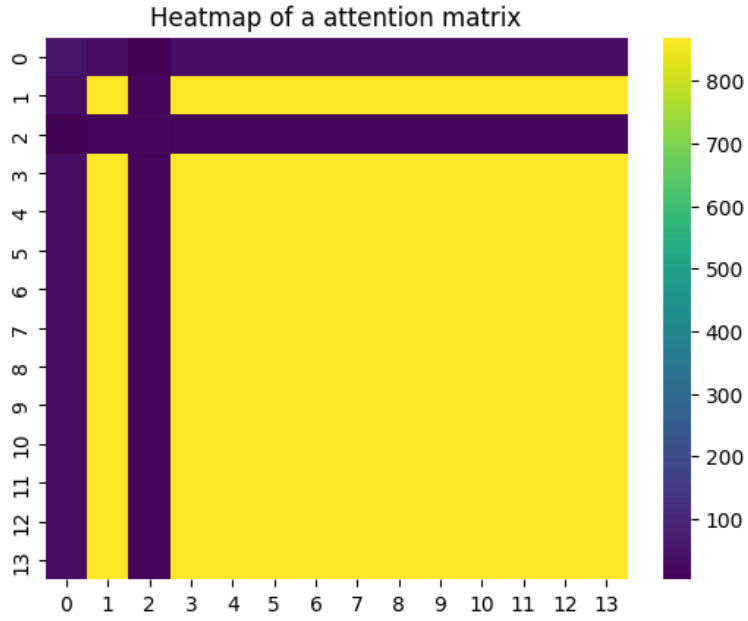
$$u, 0, v, 0,\ \textit{shortest path nodes},\ \textit{noise symbols}$$

Context can be given by listing the nodes and then the edges in front of the question separated with 0s.

# Experiments

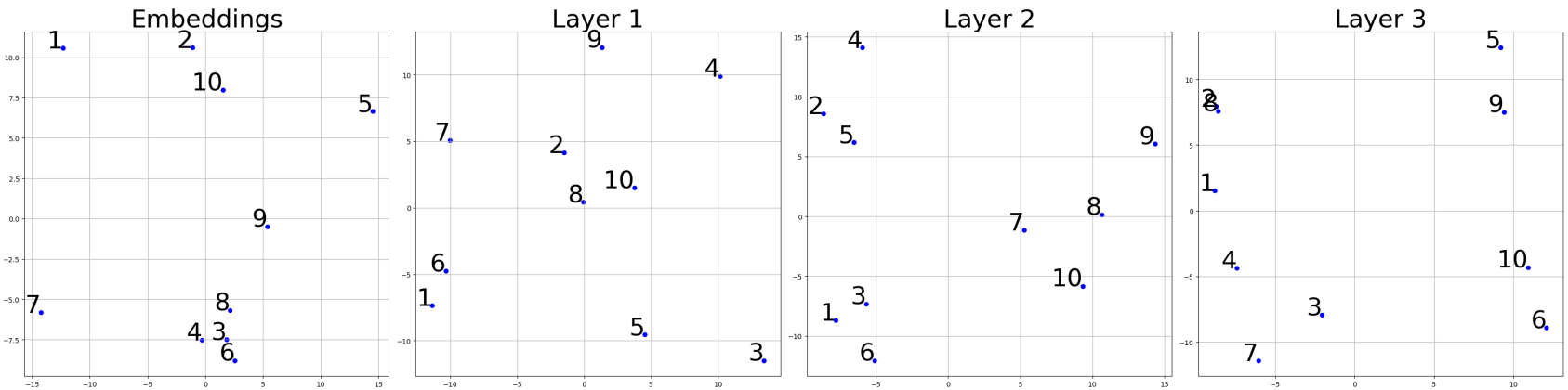



(a) Distance heatmap of token embeddings




(b) Attention matrix without graph context

# Experiments




Embeddings of vertice tokens through a model

# Experiments

Generalization on randomly split datasets

| Train Accuracy | Correct Tokens | All Tokens | Test Accuracy | Correct Tokens | All Tokens |
|---|---|---|---|---|---|
| 0.95 | 213 | 224 | 0.87 | 244 | 280 |
| 1.0 | 224 | 224 | 0.91 | 255 | 280 |
| 1.0 | 224 | 224 | 0.89 | 250 | 280 |
| 1.0 | 224 | 224 | 0.90 | 252 | 280 |
| 1.0 | 224 | 224 | 0.87 | 246 | 280 |

Table of accuracies

# Future Plans

Further tests of the limitation of the current model

Usage of linear probes

Dataset of networks

Thank you for the attention!