Analysis of the Attention in Tabular Language Models

Aneta Koleva, Martin Ringsquandl, Volker Tresp
TRL@NeurIPS 2022
**Motivation**

**BERT**
- Pre-trained on large corpus of text
- Masked Language Modeling (MLM)
- Next Sentence Prediction (NSP)

> “My dog is cute. He likes playing.”

**Transformer**
- Encoder Block
  - Multi-layer with self-attention heads

\[
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V
\]
Motivation

“My dog is cute. He likes playing.”

Tabular representation?

<table>
<thead>
<tr>
<th>Name</th>
<th>Auto Racing Team</th>
<th>Formula 1 Race</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Schumacher</td>
<td>Scuderia Ferrari</td>
<td>2005 United States Grand Prix</td>
<td>2005</td>
</tr>
<tr>
<td>Niki Lauda</td>
<td>McLaren</td>
<td>1985 Dutch Grand Prix</td>
<td>1985</td>
</tr>
</tbody>
</table>
Motivation

“My dog is cute. He likes playing.”

Tabular representation?

<table>
<thead>
<tr>
<th>Name</th>
<th>Auto Racing Team</th>
<th>Formula 1 Race</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Schumacher</td>
<td>Scuderia Ferrari</td>
<td>2005 United States Grand Prix</td>
<td>2005</td>
</tr>
<tr>
<td>Niki Lauda</td>
<td>McLaren</td>
<td>1985 Dutch Grand Prix</td>
<td>1985</td>
</tr>
</tbody>
</table>
Motivation

“My dog is cute. He likes playing.”

Tabular representation?

<table>
<thead>
<tr>
<th>Name</th>
<th>Auto Racing Team</th>
<th>Formula 1 Race</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Schumacher</td>
<td>Scuderia Ferrari</td>
<td>2005 United States Grand Prix</td>
<td>2005</td>
</tr>
<tr>
<td>Niki Lauda</td>
<td>McLaren</td>
<td>1985 Dutch Grand Prix</td>
<td>1985</td>
</tr>
</tbody>
</table>
Motivation

“My dog is cute. He likes playing.”

Tabular representation?

<table>
<thead>
<tr>
<th>Name</th>
<th>Auto Racing Team</th>
<th>Formula 1 Race</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Schumacher</td>
<td>Scuderia Ferrari</td>
<td>2005 United States Grand Prix</td>
<td>2005</td>
</tr>
<tr>
<td>Niki Lauda</td>
<td>McLaren</td>
<td>1985 Dutch Grand Prix</td>
<td>1985</td>
</tr>
</tbody>
</table>
Motivation

“My dog is cute. He likes playing.”

Tabular representation?

<table>
<thead>
<tr>
<th>Name</th>
<th>Auto Racing Team</th>
<th>Formula 1 Race</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Schumacher</td>
<td>Scuderia Ferrari</td>
<td>2005 United States Grand Prix</td>
<td>2005</td>
</tr>
<tr>
<td>Niki Lauda</td>
<td>McLaren</td>
<td>1985 Dutch Grand Prix</td>
<td>1985</td>
</tr>
</tbody>
</table>

• Best task-agnostic approach?
• Architecture biased towards the table structure?
Overview of attention in TaLMs

Table Language Models (TaLMs)

- Pre-trained on large corpus of tables
- Masked Language Modeling (MLM)
- Masked Column Prediction (MCP)
- Masked Entity Recovery (MER)

<table>
<thead>
<tr>
<th>Model</th>
<th>Attention</th>
<th>L</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAPAS</td>
<td>Transformer attention</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>TaBERT</td>
<td>Transformer attention + vertical on the columns</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>TURL</td>
<td>Restricted to entities in the same column/row + header</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>TUTA</td>
<td>Joint bi-tree based. Focused on spatial and hierarchical info</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>MATE</td>
<td>Column/row restricted attention heads</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>TableFormer</td>
<td>Added attention-bias to each attention head</td>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>
TaBERT

<table>
<thead>
<tr>
<th>Name</th>
<th>Auto Racing Team</th>
<th>Formula 1 Race</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Schumacher</td>
<td>Scuderia Ferrari</td>
<td>2005 United States Grand Prix</td>
<td>2005</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Row linearization

\[ R_1 \quad [\text{CLS}] \quad [\text{SEP}] \quad \text{Name} \mid \text{text} \mid \text{Michael Schumacher} [\text{SEP}] \ldots [\text{SEP}] \quad \text{Year} \mid \text{real} \mid 2005 [\text{SEP}] \]

Input

BERT

Cell level representations

\[ x_{r_{1,0}} \quad x_{r_{1,1}} \quad \ldots \quad x_{r_{1,3}} \]

Column level representations

\[ x_{r_{*0}} \quad x_{r_{*1}} \quad \ldots \quad x_{r_{*3}} \]

Vertical Attention
### TaBERT

<table>
<thead>
<tr>
<th>Name</th>
<th>Auto Racing Team</th>
<th>Formula 1 Race</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Schumacher</td>
<td>Scuderia Ferrari</td>
<td>2005 United States Grand Prix</td>
<td>2005</td>
</tr>
</tbody>
</table>

Row linearization:

\[
R_1 = [\text{[CLS]} \text{[SEP]} \text{Name} \mid \text{text} \mid \text{Michael Schumacher} \text{[SEP]} \ldots \text{[SEP]} \text{Year} \mid \text{real} \mid 2005 \text{[SEP]}]
\]

Input:

```
BERT
```

Cell level representations:

\[
x_{r_{1,0}}, x_{r_{1,1}}, x_{r_{1,3}}
\]

- Michael Schumacher
- Scuderia Ferrari
- 2005
<table>
<thead>
<tr>
<th>Name</th>
<th>Auto Racing Team</th>
<th>Formula 1 Race</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Schumacher</td>
<td>Scuderia Ferrari</td>
<td>2005 United States Grand Prix</td>
<td>2005</td>
</tr>
<tr>
<td>Name</td>
<td>Auto Racing Team</td>
<td>Formula 1 Race</td>
<td>Year</td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------------</td>
<td>---------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Michael Schumacher</td>
<td>Scuderia Ferrari</td>
<td>2005 United States Grand Prix</td>
<td>2005</td>
</tr>
</tbody>
</table>

Deng et al. TURL: Table Understanding through Representation Learning, VLDB 2021
<table>
<thead>
<tr>
<th>Name</th>
<th>Auto Racing Team</th>
<th>Formula 1 Race</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Schumacher</td>
<td>Scuderia Ferrari</td>
<td>2005 United States Grand Prix</td>
<td>2005</td>
</tr>
</tbody>
</table>

**Visualization**

- **TaBERT**
  - [CLS]
  - [SEP]
  - name
  - text
  - michael schumacher
  - [SEP]
  - auto
  - racing
  - team
  - [SEP]
  - formula
  - [SEP]
  - race
  - year
  - [PAD]

- **TURL**
  - name
  - auto
  - racing
  - team
  - formula
  - race
  - year
  - [PAD]

**Header Tokens**

- name
- auto
- racing
- team
- formula
- race
- year
- [PAD]

**Entities**

- michael schumacher
- scuderia ferrari
- united states grand prix
- 2005
TaBERT – model view

- generated with BertViz
TaBERT – model view
Vig and Belinkov. Analyzing the Structure of Attention in a Transformer Language Model. BlackboxNLP 2019
Aggregate Attention Analysis

Datasets

• T2D [Ritze et al. Matching HTML Tables to Dbpedia. WIMS 2015]
  • Textual

• GitTables [Hulsebos et al. GitTables: A Large-Scale Corpus of Relational Tables. arXiv2021]
  • Numeric

• Sampled rows
  • $n = 1, 3, 5$

• Analysis
  • Attention to special tokens
  • Attention to header-body tokens
  • Attention Entropy
Special tokens

- TaBERT - Attention to [SEP] and [CLS]
  - GitTables
  - n = 1
Special tokens

- TURL - Attention to [PAD]
  - n = 1
Header-body attention

- GitTables
- $n = 1$

(a) TaBERT

(b) TURL
Attention entropy

- GitTables
- $n = 1$

(a) TaBERT

(b) TURL
Conclusion

• First work focused on analyzing the attention in TaLMs
• Heterogenous space of attention mechanisms in TaLMs
• Input matters
• TURL - more attention to the header, TaBERT - more attention to special tokens
• Do we need 12 layers with 12 attention heads in TaLMs?