LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

SIEMENS

Analysis of the Attention in Tabular Language Models

Aneta Koleva, Martin Ringsquandl, Volker Tresp

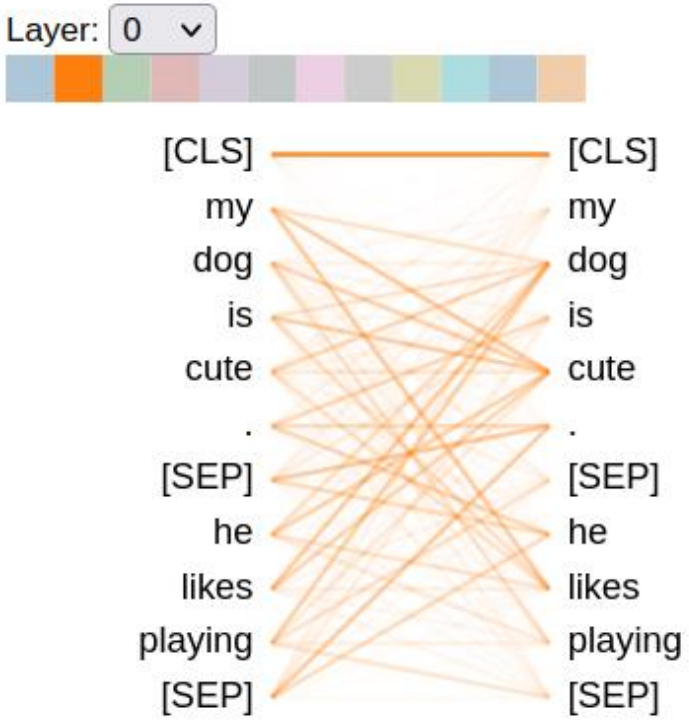
TRL@NeurIPS 2022

Motivation

BERT

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

"My dog is cute. He likes playing."



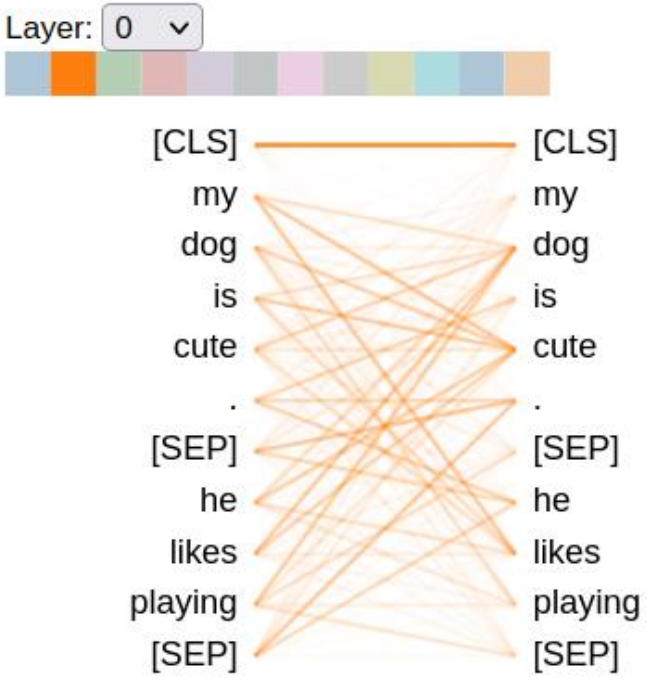
Transformer

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

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Motivation

“My dog is cute. He likes playing.”

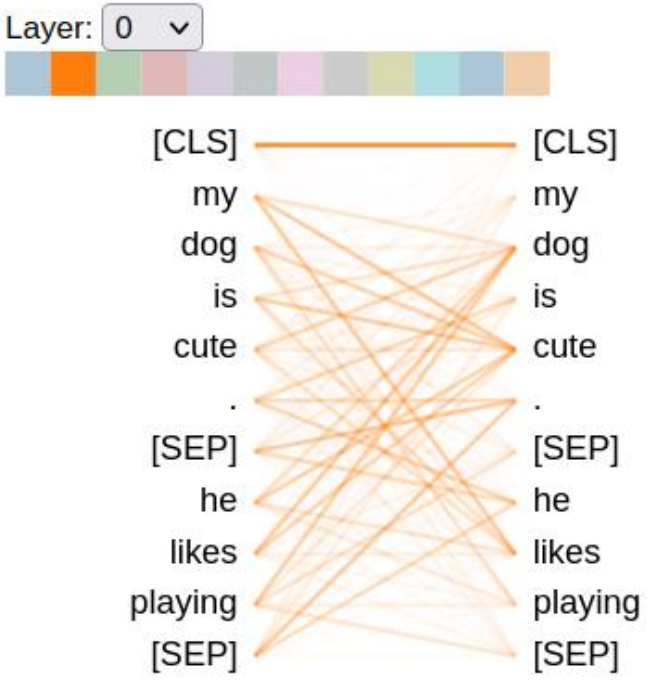


Tabular representation?

Name	Auto Racing Team	Formula 1 Race	Year
Michael Schumacher	Scuderia Ferrari	2005 United States Grand Prix	2005
Niki Lauda	McLaren	1985 Dutch Grand Prix	1985
Sebastian Vettel	Red Bull Racing	2011 European Grand Prix	2011

Motivation

“My dog is cute. He likes playing.”

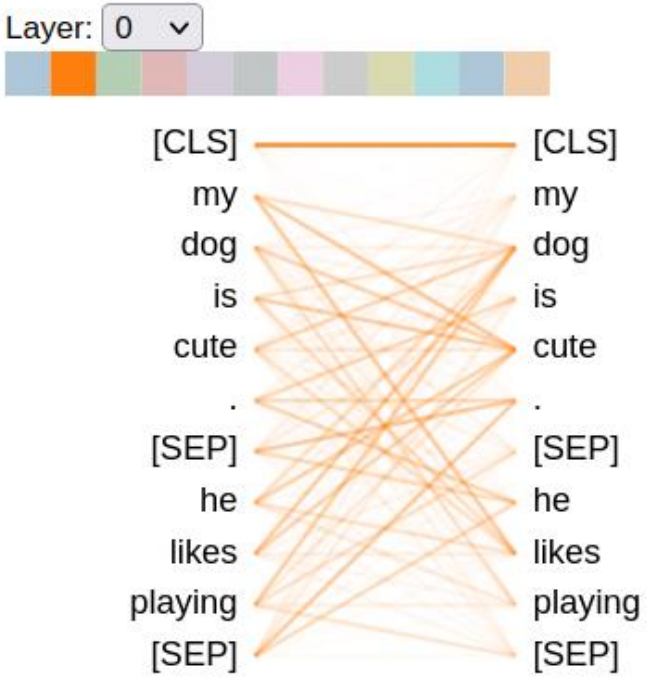


Tabular representation?

Name	Auto Racing Team	Formula 1 Race	Year
Michael Schumacher	Scuderia Ferrari	2005 United States Grand Prix	2005
Niki Lauda	McLaren	1985 Dutch Grand Prix	1985
Sebastian Vettel	Red Bull Racing	2011 European Grand Prix	2011

Motivation

“My dog is cute. He likes playing.”

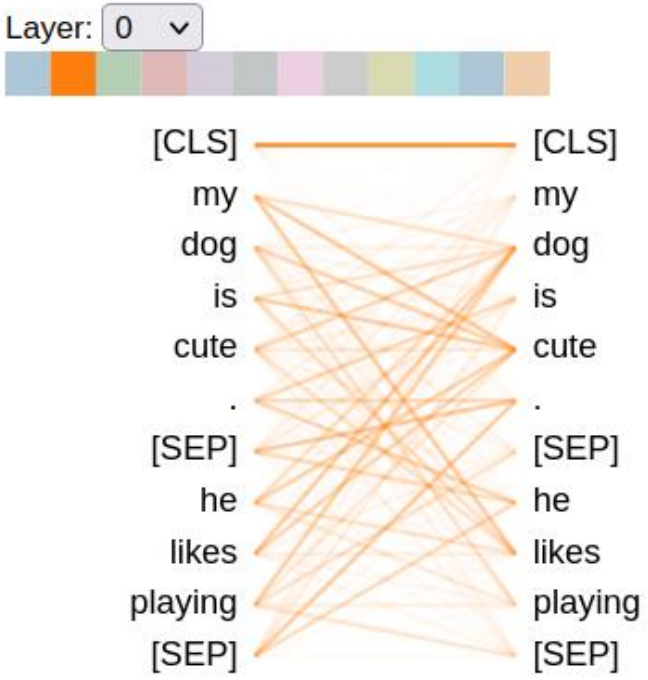


Tabular representation?

Name	Auto Racing Team	Formula 1 Race	Year
Michael Schumacher	Scuderia Ferrari	2005 United States Grand Prix	2005
Niki Lauda	McLaren	1985 Dutch Grand Prix	1985
Sebastian Vettel	Red Bull Racing	2011 European Grand Prix	2011

Motivation

“My dog is cute. He likes playing .”

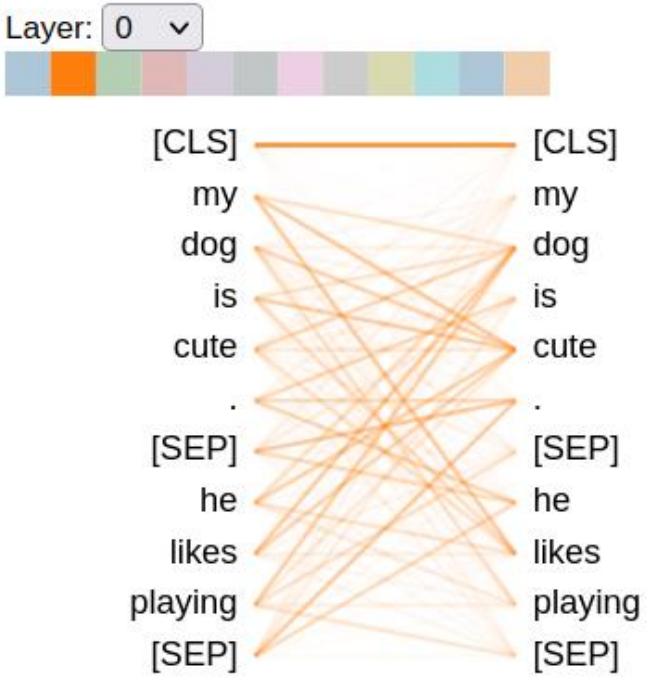


Tabular representation?

Name	Auto Racing Team	Formula 1 Race	Year
Michael Schumacher	Scuderia Ferrari	2005 United States Grand Prix	2005
Niki Lauda	McLaren	1985 Dutch Grand Prix	1985
Sebastian Vettel	Red Bull Racing	2011 European Grand Prix	2011

Motivation

“My dog is cute. He likes playing .”



Tabular representation?

Name	Auto Racing Team	Formula 1 Race	Year
Michael Schumacher	Scuderia Ferrari	2005 United States Grand Prix	2005
Niki Lauda	McLaren	1985 Dutch Grand Prix	1985
Sebastian Vettel	Red Bull Racing	2011 European Grand Prix	2011

- 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)

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

TaBERT

Name	Auto Racing Team	Formula 1 Race	Year
Michael Schumacher	Scuderia Ferrari	2005 United States Grand Prix	2005
⋮	⋮	⋮	⋮

Row linearization

R_1 [CLS] [SEP] **Name** | *text* | Michael Schumacher [SEP] … [SEP] **Year** | *real* | 2005 [SEP]

Input

BERT

Cell level representations

$x_{r_{1,0}}$ $x_{r_{1,1}}$ … $x_{r_{1,3}}$

Michael Schumacher | Scuderia Ferrari | … | 2005

Column level representations

$x_{r_{*0}}$ $x_{r_{*1}}$ … $x_{r_{*3}}$

Vertical Attention

$x_{r_{1*}}$ ⋮ $x_{r_{n*}}$

Yin et al. TABERT: Pretraining for Joint Understanding of Textual and Tabular Data. ACL 2020

TaBERT

Name	Auto Racing Team	Formula 1 Race	Year
Michael Schumacher	Scuderia Ferrari	2005 United States Grand Prix	2005

Row linearization

R_1 [CLS] [SEP] **Name** | *text* | Michael Schumacher [SEP] … [SEP] **Year** | *real* | 2005 [SEP]

Input

BERT

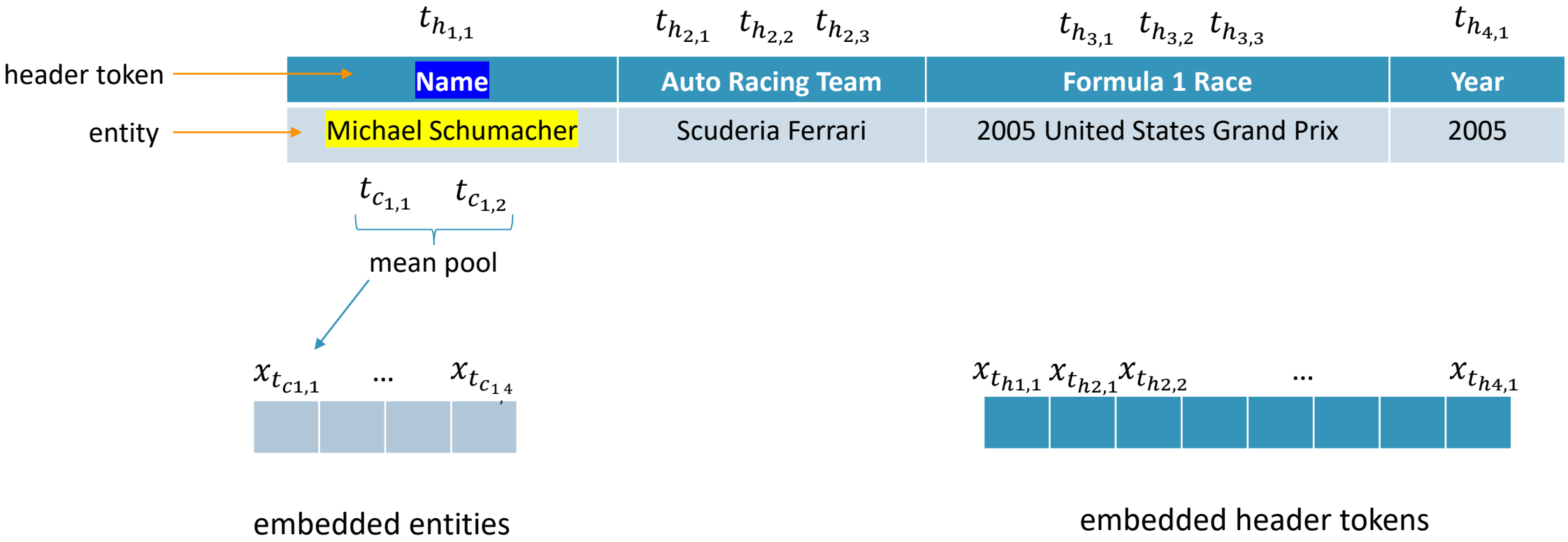
Cell level representations

$x_{r_{1,0}}$ $x_{r_{1,1}}$ … $x_{r_{1,3}}$

Michael Schumacher | Scuderia Ferrari | … | 2005

Yin et al. TABERT: Pretraining for Joint Understanding of Textual and Tabular Data. ACL 2020

TURL



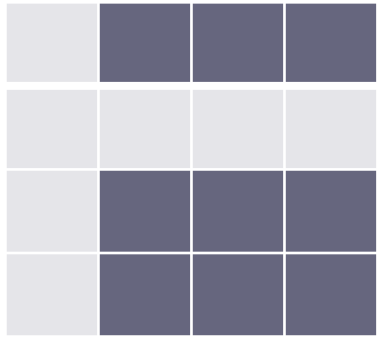
Deng et al. TURL: Table Understanding through Representation Learning. VLDB 2021

TURL

Name	Auto Racing Team	Formula 1 Race	Year
Michael Schumacher	Scuderia Ferrari	2005 United States Grand Prix	2005

embedded tokens and entities

Visibility matrix



Input

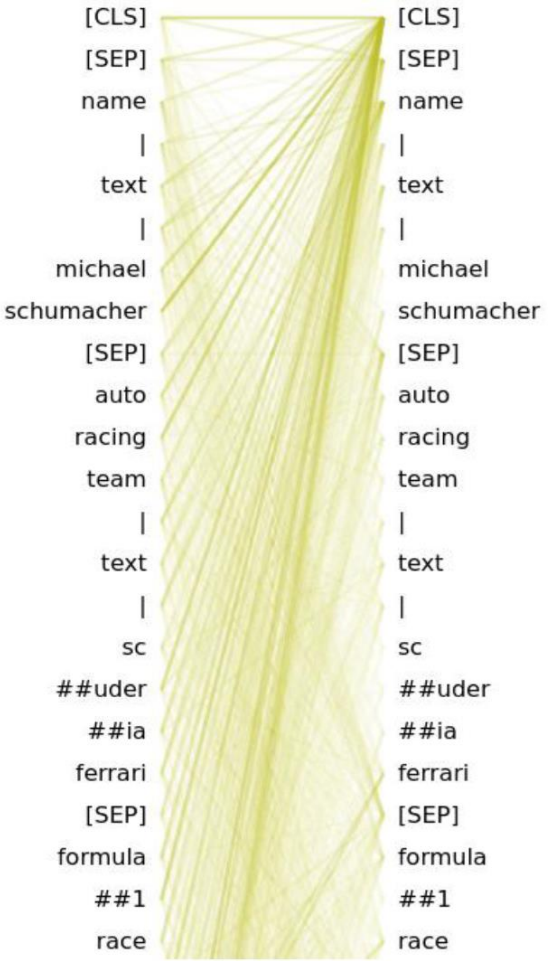
Transformer Encoder

Deng et al. TURL: Table Understanding through Representation Learning. VLDB 2021

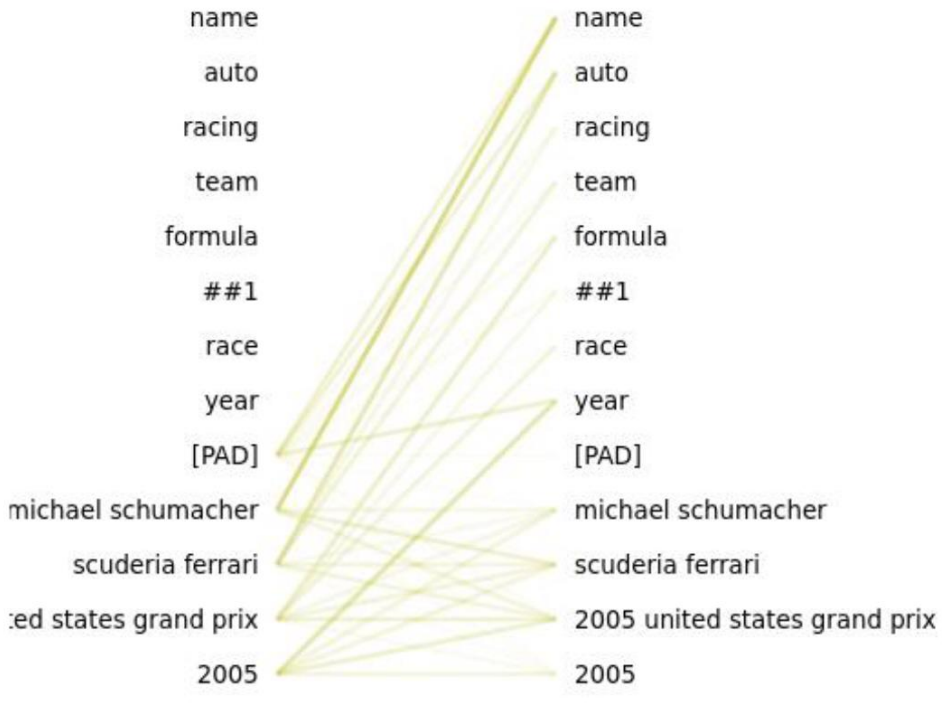
Visualization

Name	Auto Racing Team	Formula 1 Race	Year
Michael Schumacher	Scuderia Ferrari	2005 United States Grand Prix	2005

TaBERT

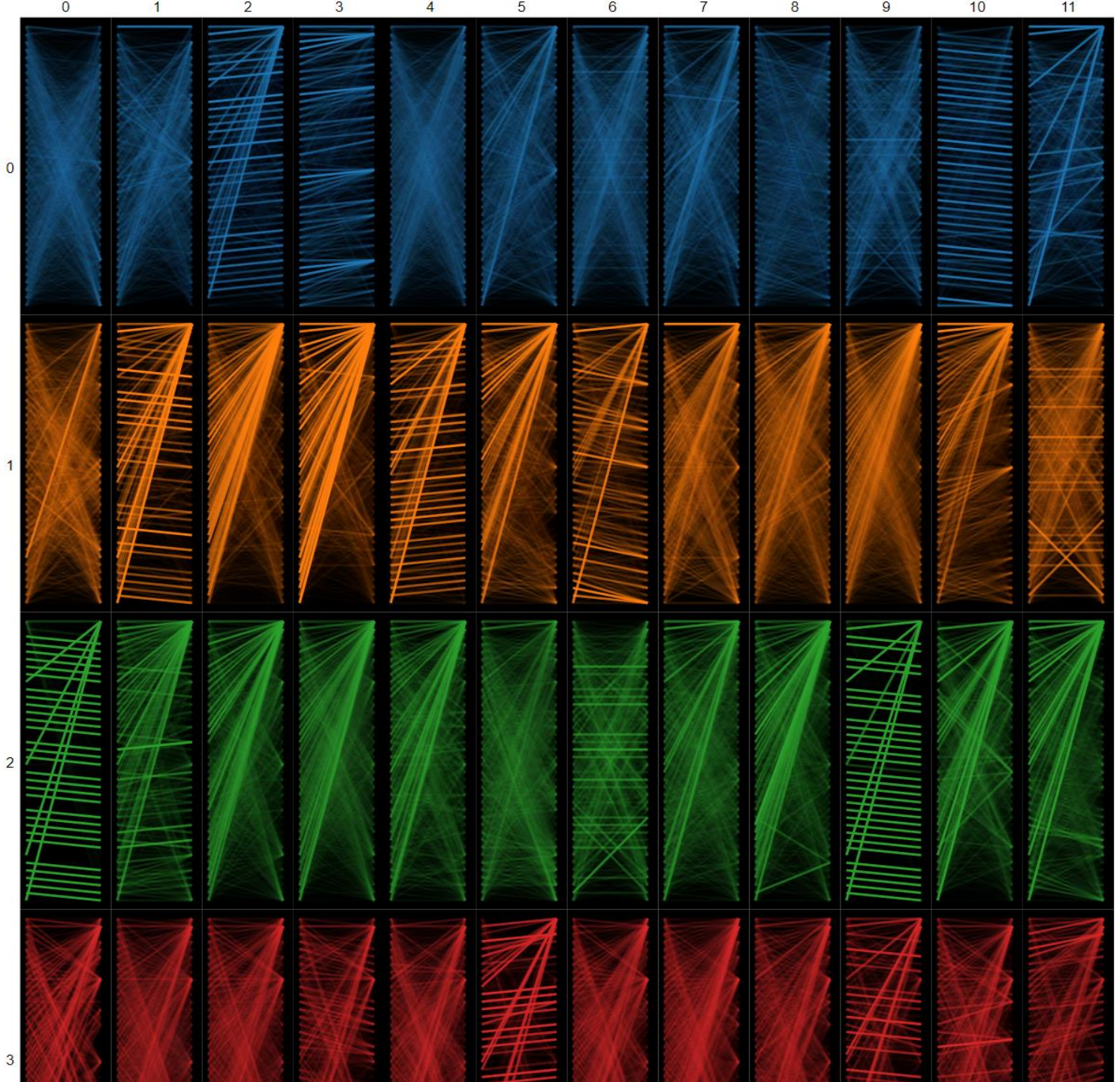


TURL



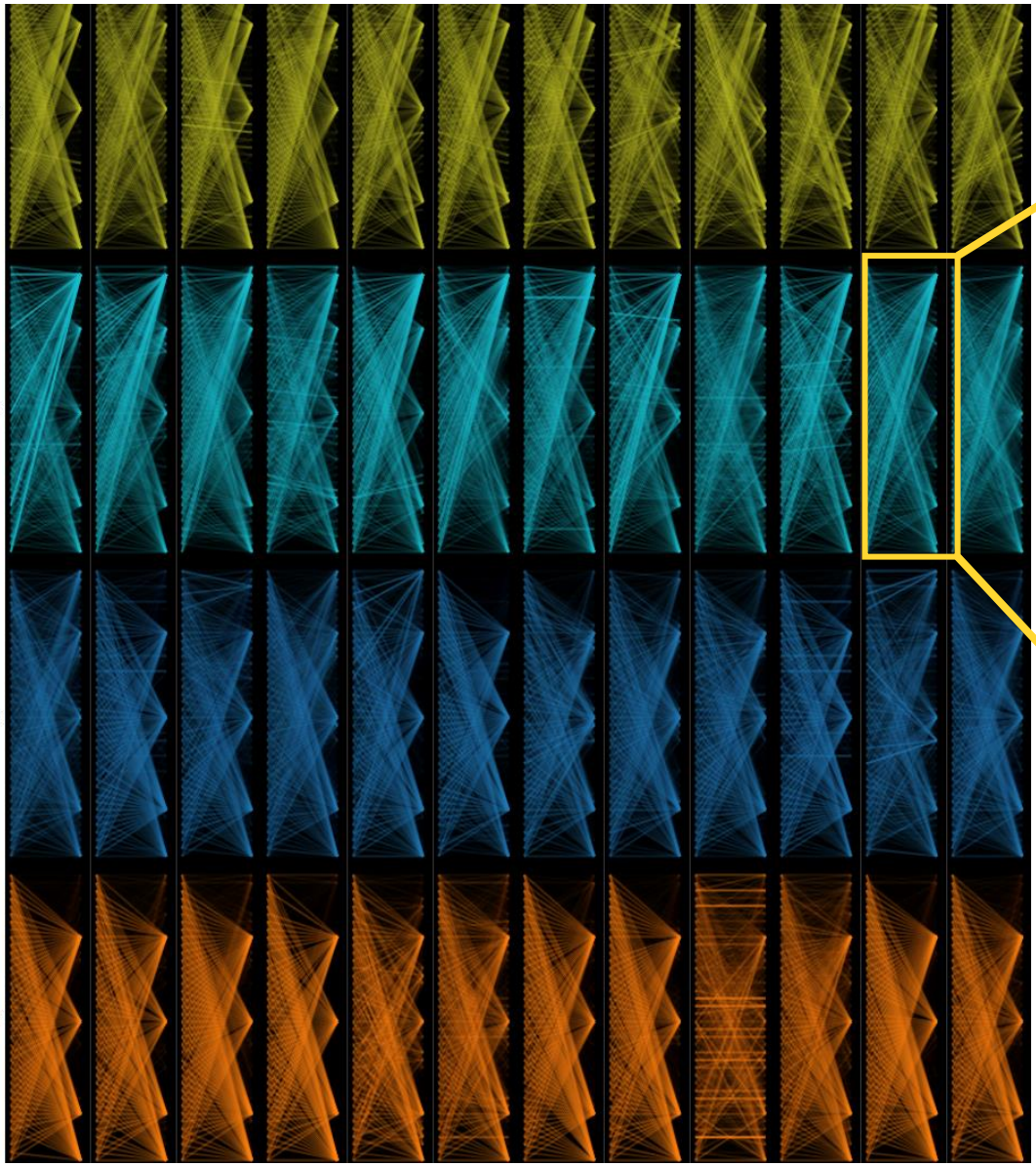
TaBERT – model view

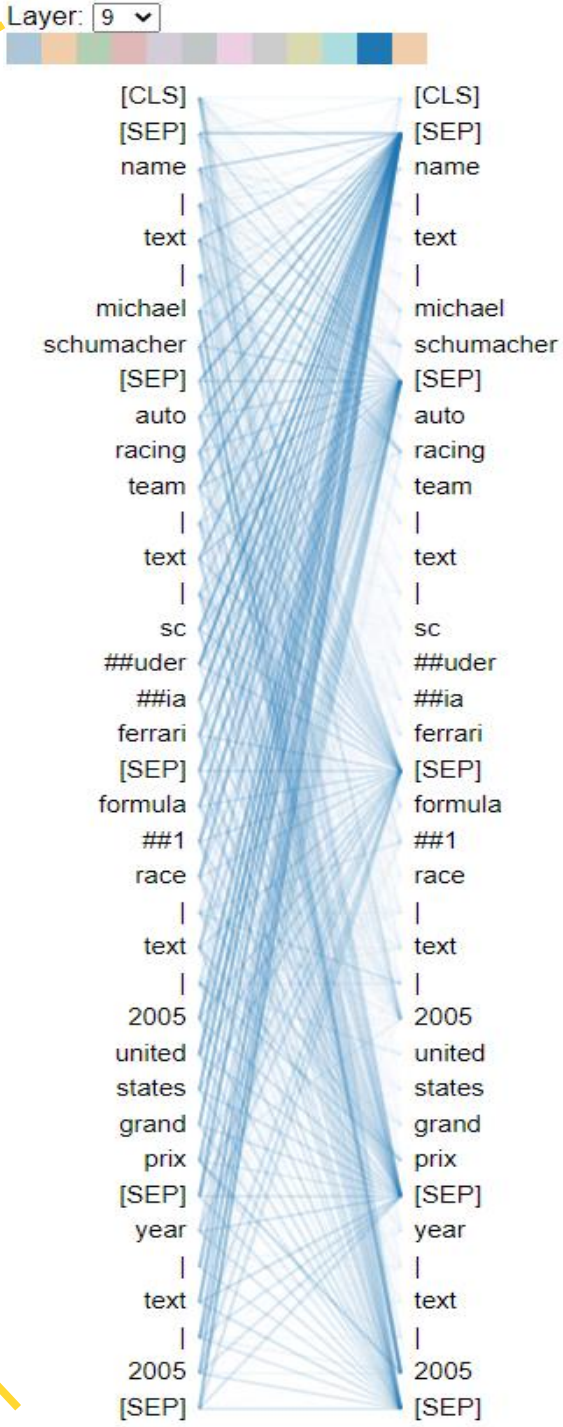
- generated with BertViz



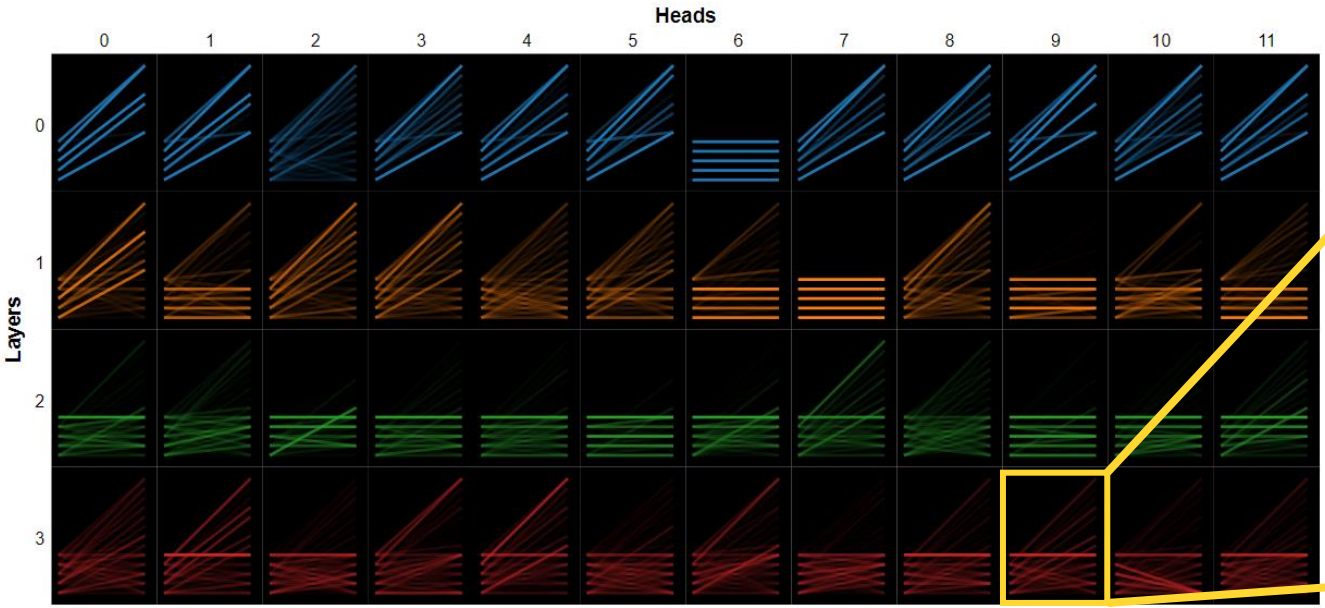
Vig and Belinkov. Analyzing the Structure of Attention in a Transformer Language Model. BlackboxNLP 2019

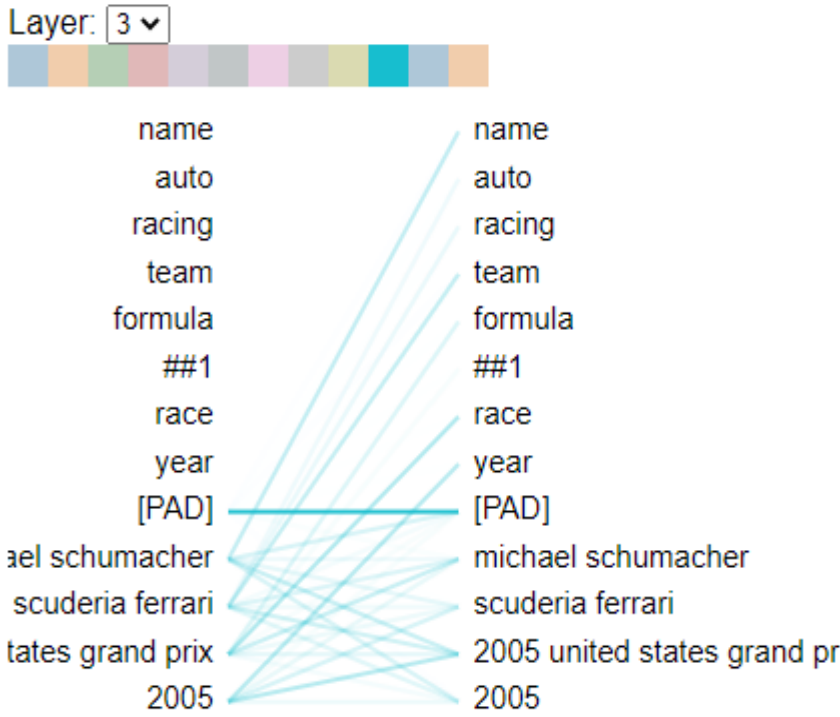
TaBERT – model view





TURL – model view





Vig and Belinkov. Analyzing the Structure of Attention in a Transformer Language Model. BlackboxNLP 2019

Aggregate Attention Analysis

Datasets

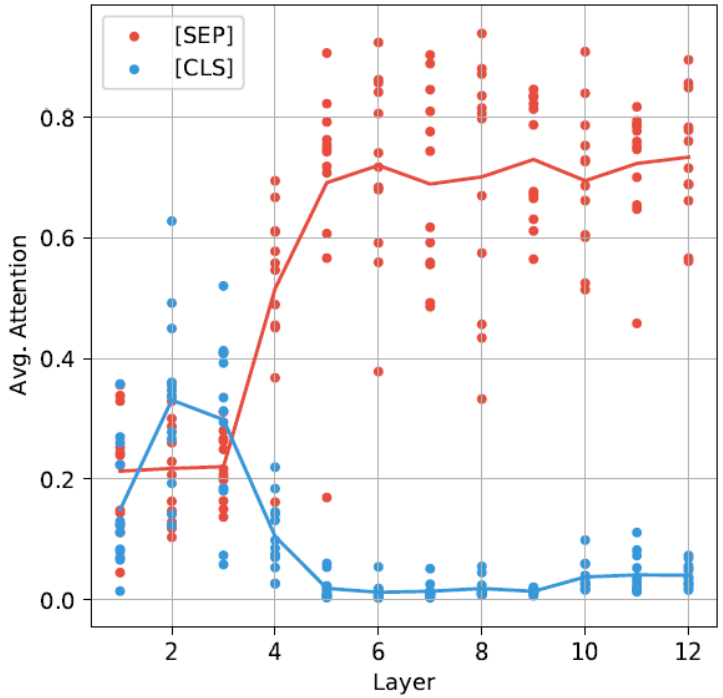
- T2D *[Ritze et al. Matching HTML Tables to Dbpedia. WIMS 2015]*
  - Textual
- GitTables *[Hulsebos et al. GitTables: A Large-Scale Corpus of Relational Tables. arXiv2021]*
  - Numeric

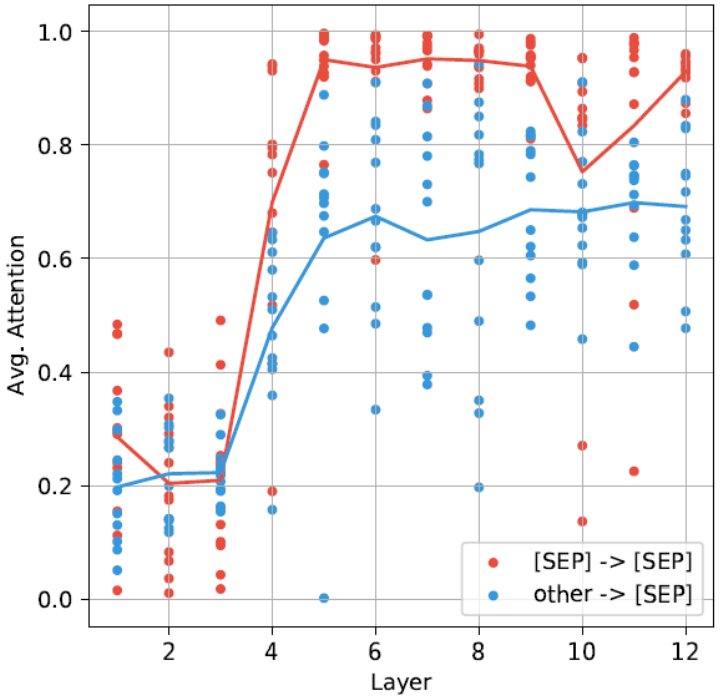
- Sampled rows
  - n = 1, 3, 5

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

Special tokens

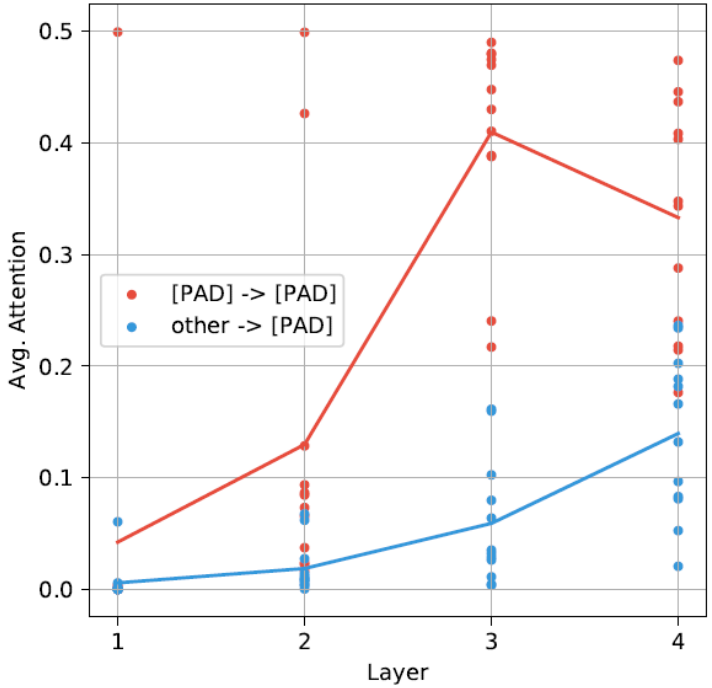
- TaBERT - Attention to [SEP] and [CLS]
  - GitTables
  - n = 1



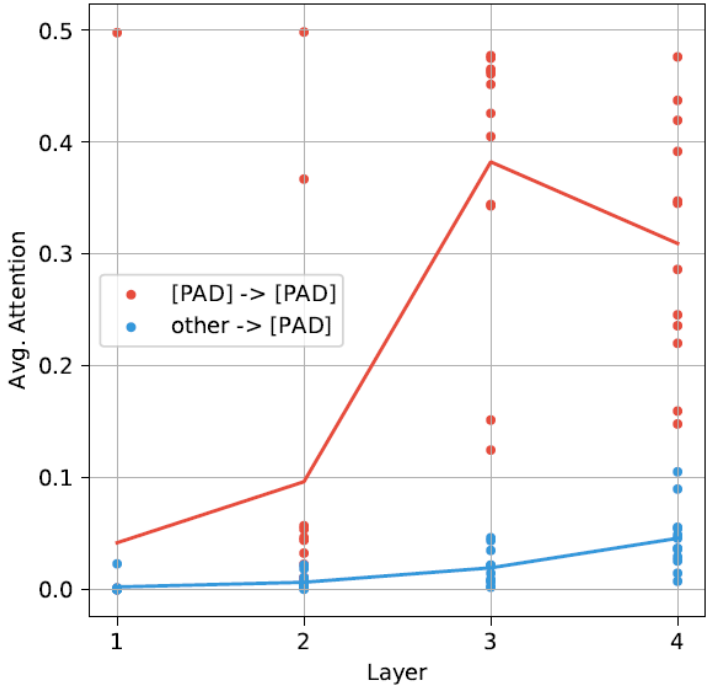


Special tokens

- TURL - Attention to [PAD]
  - n = 1



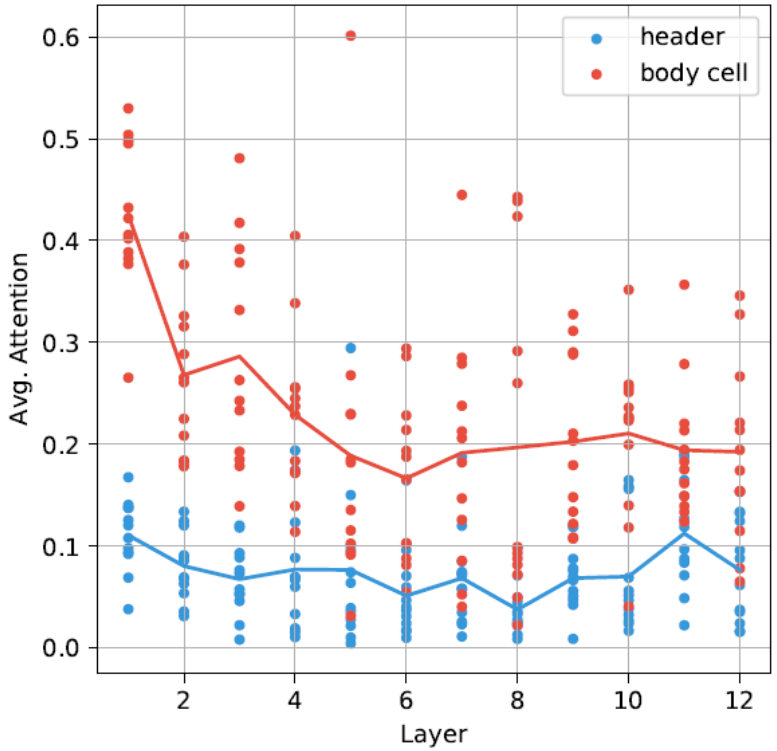
(a) GitTables



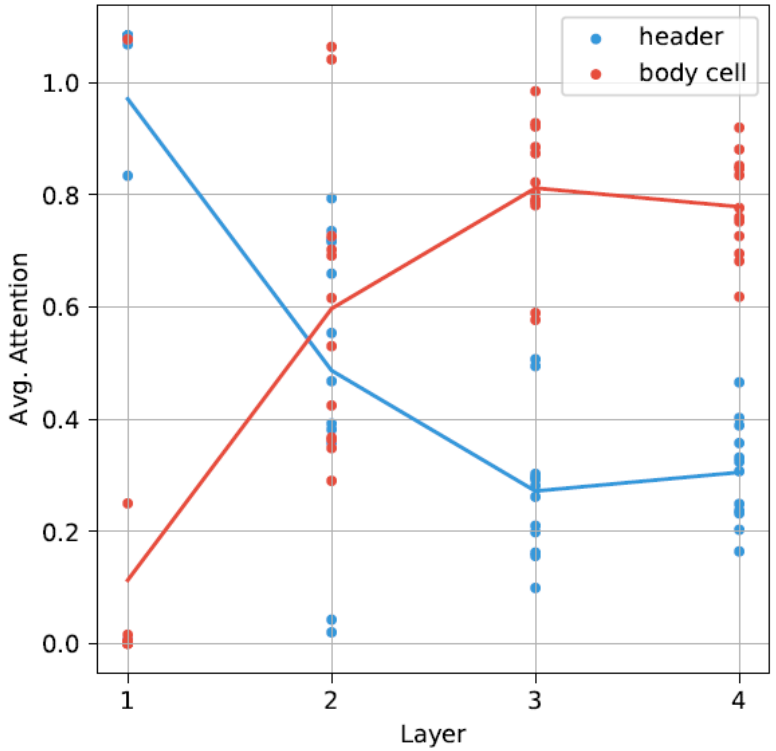
(b) T2D

Header-body attention

- GitTables
- n = 1



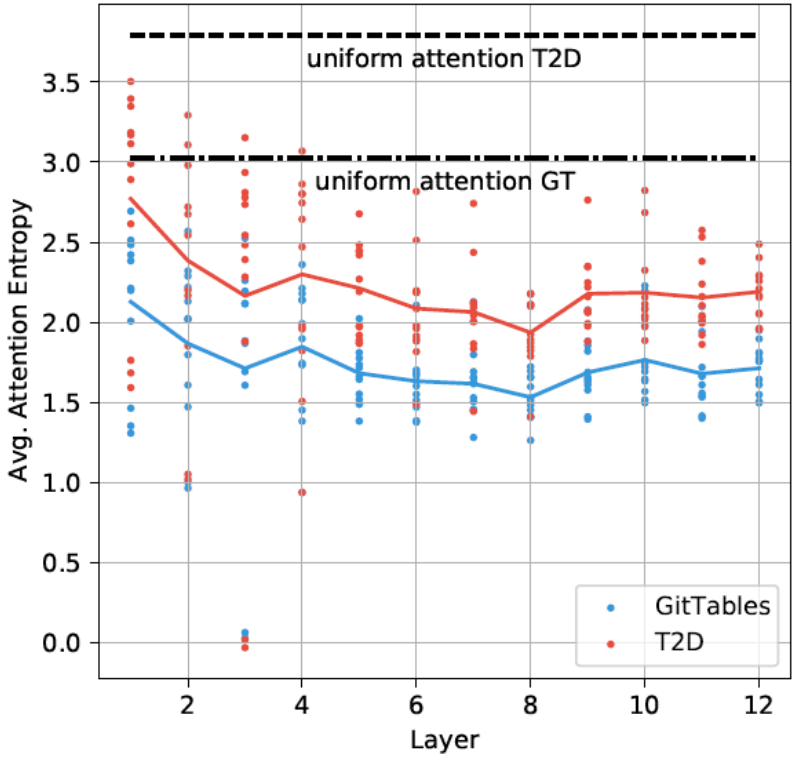
(a) TaBERT



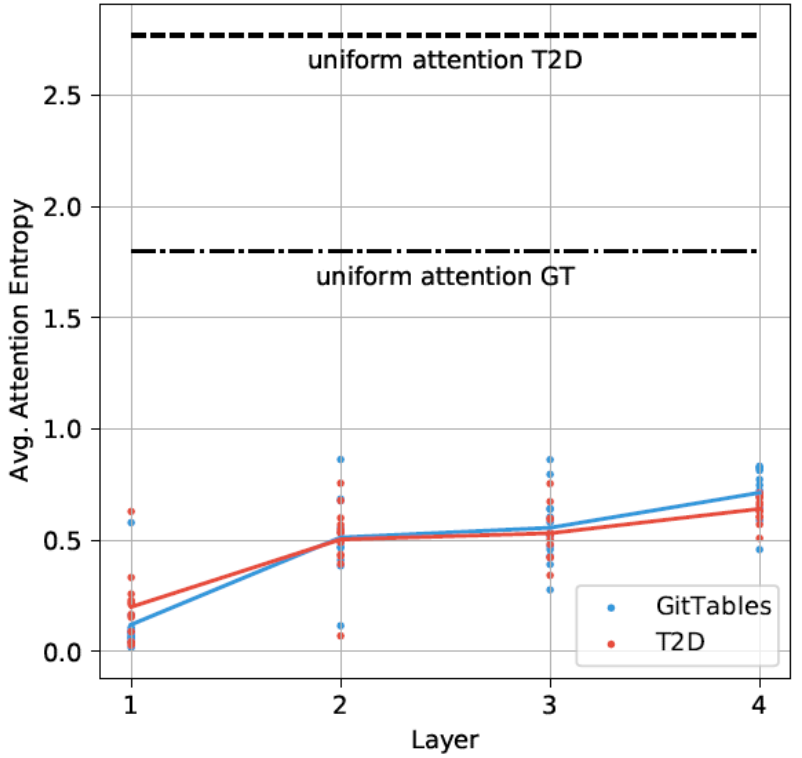
(b) TURL

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?