Transformers are uninterpretable with myopic methods: a case study with bounded Dyck grammars

Kaiyue Wen, Yuchen Li, Bingbin Liu, Andrej Risteski

Tsinghua University & Carnegie Mellon University

Transformers in Real-World


NEWS | 08 December 2022

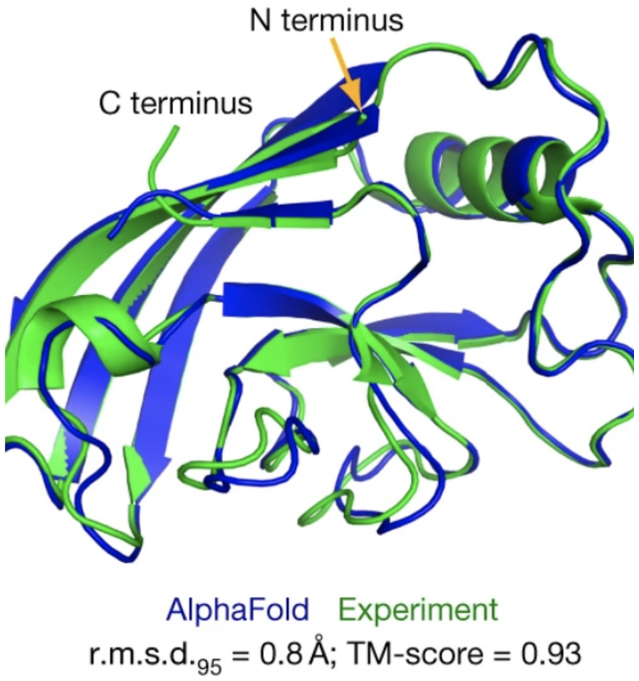
Are ChatGPT and AlphaCode going to replace programmers?

OpenAI and DeepMind systems can now produce meaningful lines of code, but software engineers shouldn't switch careers quite yet.

Davide Castelvecchi

Natural & programming languages

Computer vision



Scientific domains

Photos: nature.com; Alexey Dosovitskiy et al. An image is worth 16x16 words

Transformers in Real-World

Reliability → **Interpretability**

NEWS | 08 December 2022

Are ChatGPT and AlphaCode going to replace programmers?

OpenAI and DeepMind systems can now produce meaningful lines of code, but software engineers shouldn't switch careers quite yet.

Davide Castelvecchi

Natural & programming languages

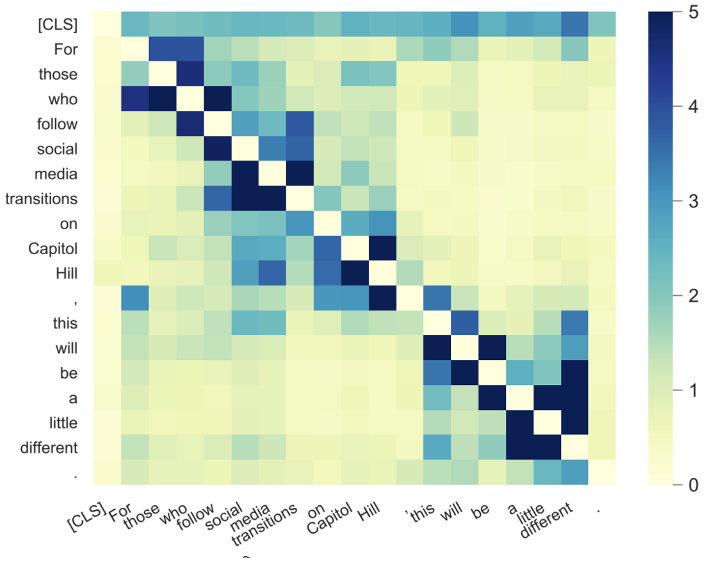
Input Attention

Computer vision

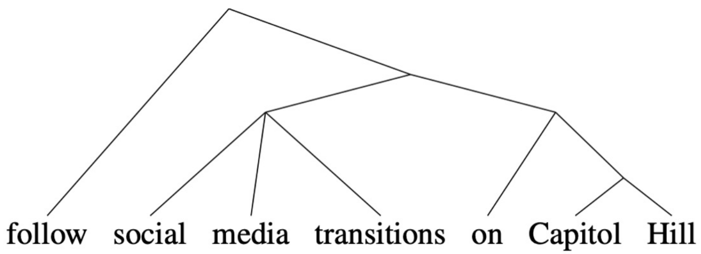
Scientific domains

Photos: nature.com; Alexey Dosovitskiy et al. An image is worth 16x16 words

Interpreting Transformers



attention map → syntactic trees



From “A Primer in BERTology” (Rogers et al. 20)

Pitfalls

- Can be misleading[1].
- Lack formal understanding.

1. Jain & Wallace, 2019; Serrano & Smith, 2019; Rogers et al., 2020; Brunner et al., 2020; Prasanna et al., 2020; Meister et al., 2021; ...

Interpreting Transformers

Question: Can we reliably interpret the algorithm implemented by a Transformer by *looking at individual components*?

"Individual" 1) attention patterns and 2) single weight components.

"myopic methods"

Answer: Transformers may *not* be interpretable by inspecting *individual parts*.

Approach: theoretical and empirical investigation on **Dyck**.

Background: the Dyck language

valid []()[()]

invalid [)(][(])

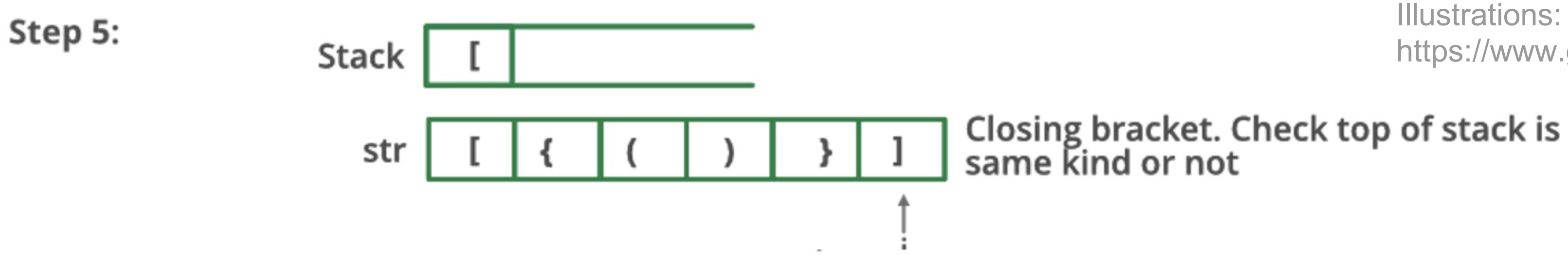
Definition: the language of **balanced parentheses**

- ***Depth*** of a bracket = number of unclosed brackets before it.

Task: predict the **type and openness** of the next bracket.

- Most naturally processed by maintaining a stack.

Illustrations: https://www.geeksforgeeks.org/

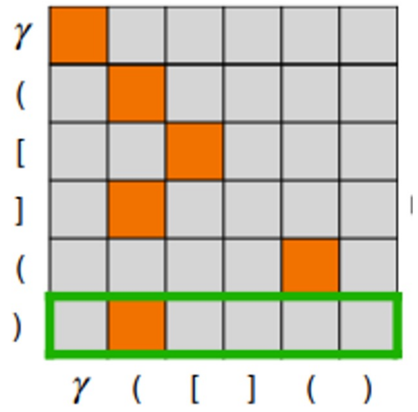


Question: how do Transformers process this Dyck language?

How do Transformers process Dyck?

Prior work [Ebrahimi et.al, Yao et.al]: Transformers learn Dyck with highly ***stack-like*** attention patterns.

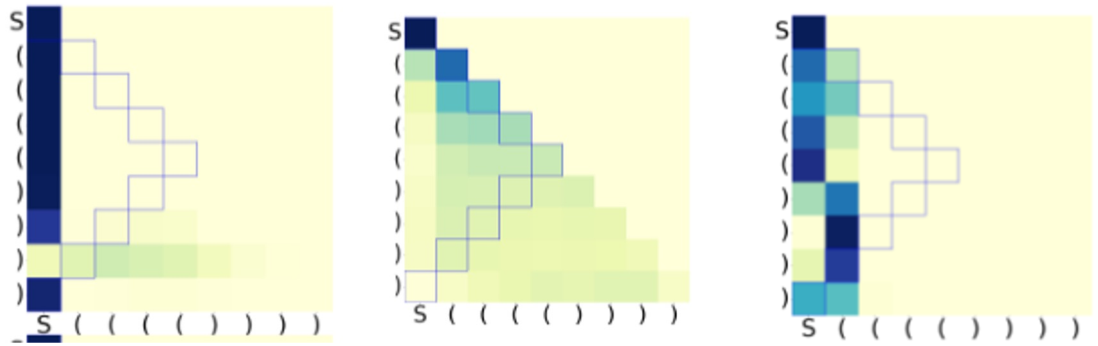
- Predict by focusing on the last unclosed bracket.



stack-like attention [Yao et.al]

Our results: Transformers learn ***diverse*** attention patterns on Dyck.

- Both in theory and in practice.
- All models reach high accuracy.



our findings: diverse attentions

Transformer model architecture

- l –th layer of a Transformer

attention pattern

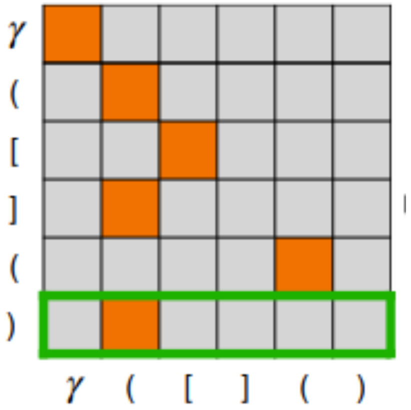
$$f_l(X) = g^{(l)}\left(\mathrm{LN}\left(W_V^{(l)} X \boxed{\sigma\left(C + \left(W_K^{(l)} X\right)^{\top}\left(W_Q^{(l)} X\right)\right)}\right) + X\right)$$

- σ: column-wise softmax operation

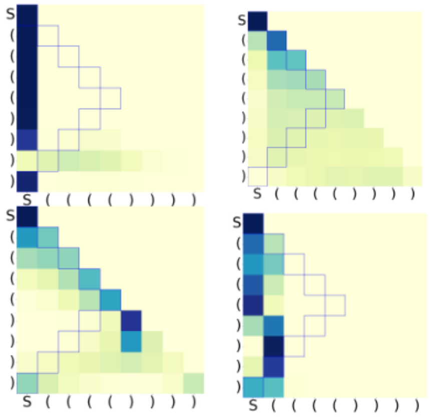
$$\sigma(A)_{i,j} = \frac{\exp(A_{i,j})}{\sum_{k=1}^{N} \exp(A_{k,j})}$$

- Full model: predicts the next token

$$T(X) = W_{\mathrm{HEAD}}\left[f_L\left(f_{L-1}(\cdots f_1(X))\right)\right]_{:,N+1}$$



stack-like attention
[Yao et.al]



empirical diverse attention

Training objective: next token prediction

- Prefix: ([ ] (___
- Continuation: ([ ] ( )

prediction $\hat{y} =$

)	0.4
[	0.3
(	0.2
]	0.05
END	0.03
START	0.02

label $y =$)

loss at that position $l(\hat{y}, y)$

training loss $\sum l(\hat{y}, y)$ for all positions

Uninterpretable Attention Patterns

Minimal first layer: the outputs $\{ e_{t,d} \}$ only depend on the bracket type t and depth d. ... independent of anything else, e.g. the position[1]

- Sequence: [ ] { < ...
- $\{ e_{t,d} \}$: type [depth 1, type] depth 0, type { depth 1, type < depth 2, ...

Thm 1. Any 2-layer Transformer with a min first layer need to satisfy the ***balance condition**** to be optimal on Dyck:

$$\left(e_{[,d} - e_{],d-1}\right)^{\top} (W^K)^{\top} W^Q \left(e_{\},d_1} - e_{>,d_2}\right) = 0$$

Intuition: embeddings for <u>matching pairs</u> of brackets should cancel out.

- similar to the pumping lemma for regular languages.

1. Inspired by the construction in Yao et al, 2021 .

Uninterpretable Attention Patterns

Thm 1 (Balance condition)

$$\left(e_{[,d} - e_{],d-1}\right)^{\top}(W^K)^{\top}W^Q(e_{\},d_1} - e_{>,d_2}) = 0$$

Remark 1: balanced != interpretable.

- **Cor 1**: Dyck can be solved by uniform attention – not reflecting task structure.

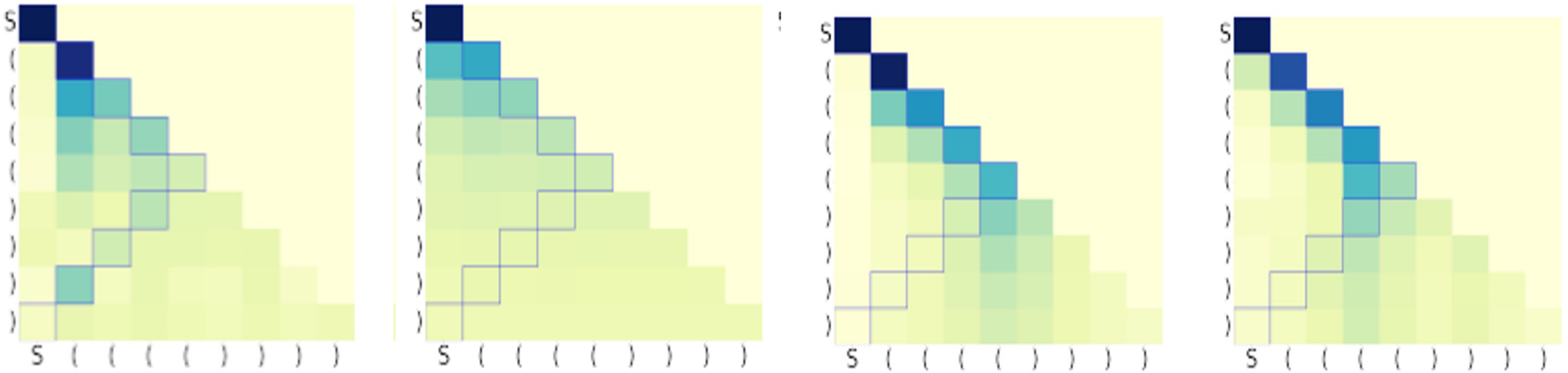
Remark 2: extension to an **approximate condition**.

- Thm 2: approximate balance from finite samples.
    - *Intuition*: the deviation from perfect balance needs to be bounded.

Empirical evidence

Balance condition is *a very weak constraint* on the attention patterns.

- Setup: freezing minimal first layers; train the rest till convergence.
- Results: high-acc models with diverse and **non-stack-like** attention patterns.

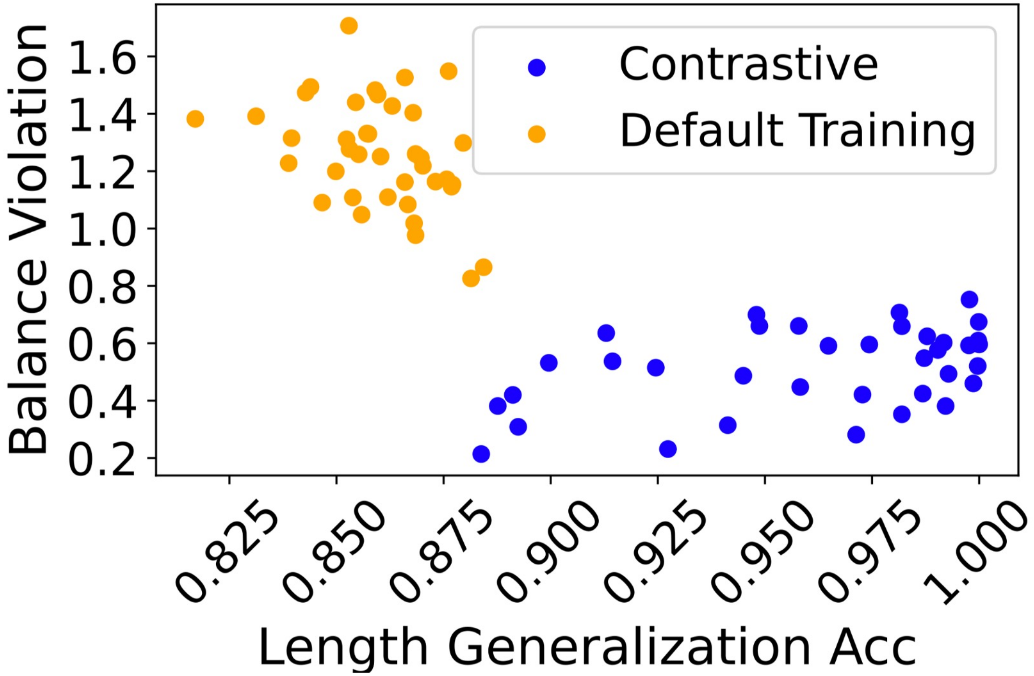


Empirical evidence

Balance condition can substantially improve out-of-distribution (length) generalization.

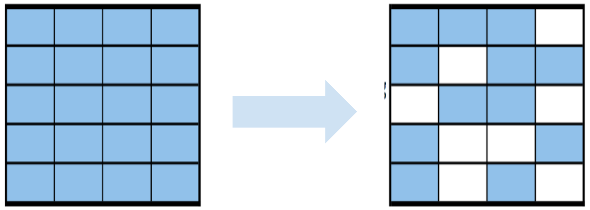
- A **contrastive objective** that penalizes balance violation.
- *Intuition*: optimal models should be ***balanced***.

more balanced⟶ better generalization?



Single Component Indistinguishability

Nonstructural pruning: zero out some entries in weight matrices.



Thm 3. Consider any given Transformer $\boldsymbol{T}$, and a polynomially larger Transformer $\boldsymbol{T_L}$ with random weights.

Then, $\boldsymbol{T}$ can be approximated by a non-structural pruning of $\boldsymbol{T_L}$ w.h.p.

Proof sketch: similar to repeated applications of the **lottery ticket hypothesis**.

- Each layer is approximated by a pruning of 4 random layers.

Remark: Uninterpretability of single weight matrix

- **Cor 2**: There exist functionally different Transformers $\boldsymbol{T_1}$, $\boldsymbol{T_2}$ that coincide with the non-structural pruning of **any single component** of $\boldsymbol{T_L}$.

Takeaway

Transformers are not interpretable via myopic methods.

- **Dyck as testbed**: fully controllable; theory-friendly.
- **Uninterpretable attention patterns**: balanced condition.
  - Little restriction on attention patterns (e.g. uniform attention)
  - Contrastive objective: reduced balance violation → better generalization.
- **Uninterpretable weight matrix**: lottery ticket hypothesis.