Analyzing the Structure of Attention in a Transformer Language Model

parc

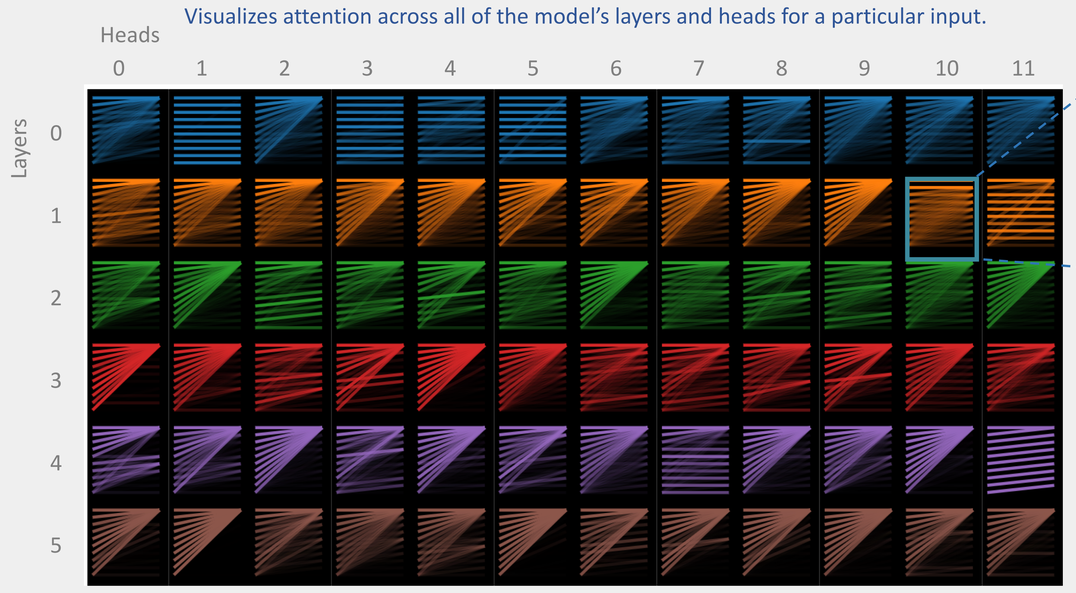
Jesse Vig
jesse.vig@parc.com

Yonatan Belinkov
belinkov@seas.harvard.edu

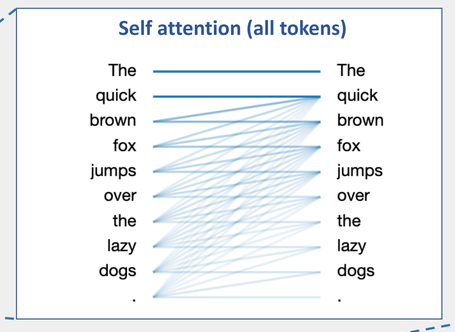
Visualizing Attention in Individual Instances

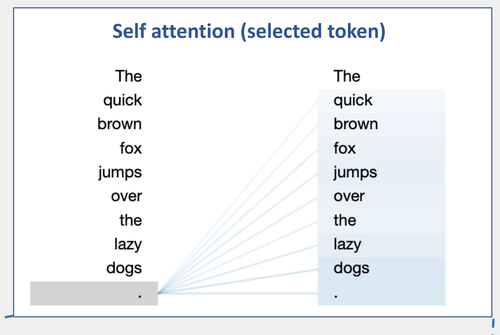
Model View

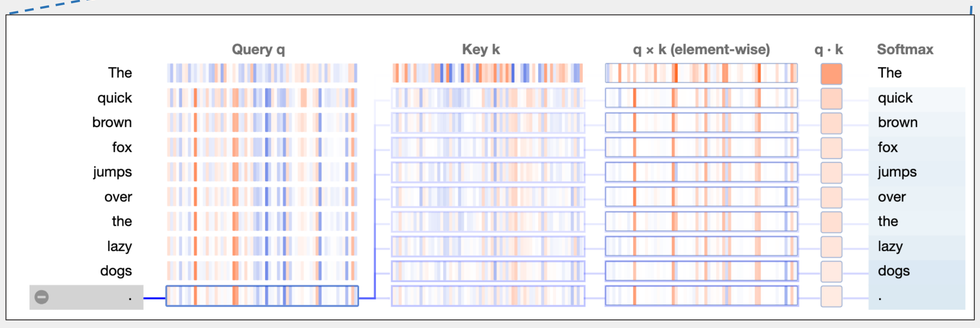
Visualizes attention across all of the model's layers and heads for a particular input.



https://github.com/jessevig/bertviz







Attention-Head View*

Visualizes attention in one or more heads for a given layer.

*Based on Jones [1,2]

Neuron View

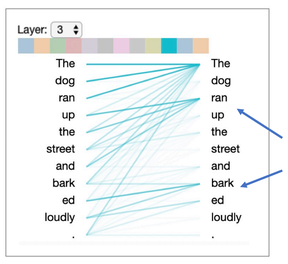
Shows how attention is computed from query and key vectors.

Analyzing Attention in Aggregate

What are the characteristics of attention?

Alignment with syntax

How does attention correlate to syntactic properties?

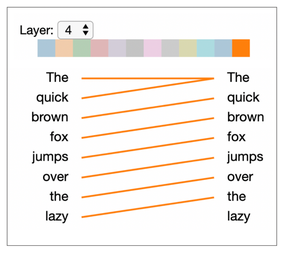


Example

Targets verbs

Distance

Does attention capture short-range or long-range relationships?

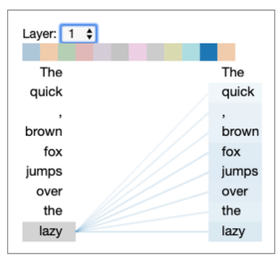


Example

Short-range (~1 token)

Entropy

Is attention dispersed broadly over many tokens or focused on a few?



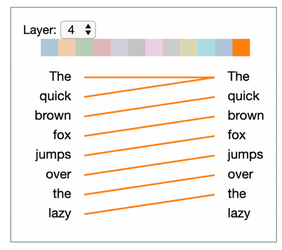
Example

High entropy (dispersed)

Variability

How does attention vary over inputs?

High variability → content-dependent
Low variability → position-dependent



Example

Low variability: focuses on previous token for all input sentences regardless of content

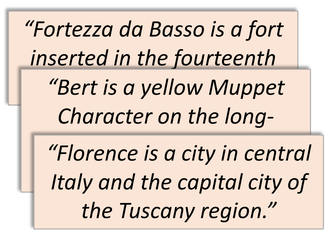
Exemplars

Which sentences/tokens most strongly induce attention in a particular head?

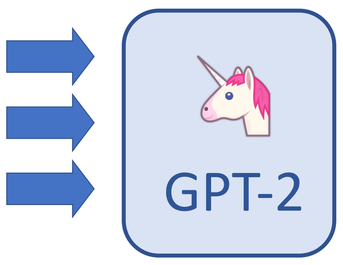
Example

Highest attention weights for this head connect acronyms to phrases

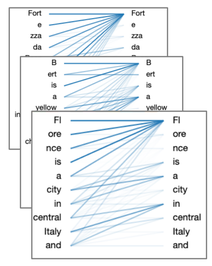
Experiment



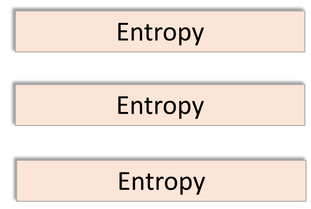
Sample sentences



Run GPT-2 on each sentence



Extract attention



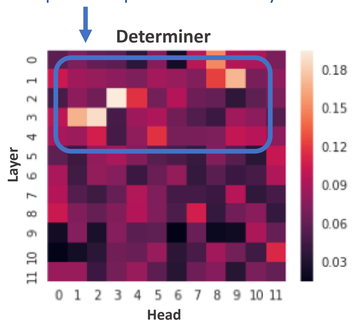
Compute assorted metrics

Aggregate over corpus

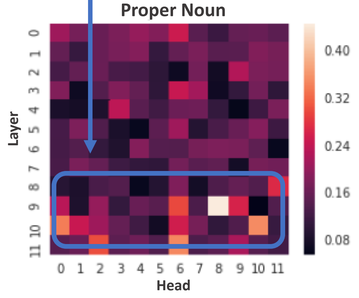
Key Results

Attention heads specialize in different parts of speech at different layer depths.

Heads attend more to lower-level parts of speech in lower layers

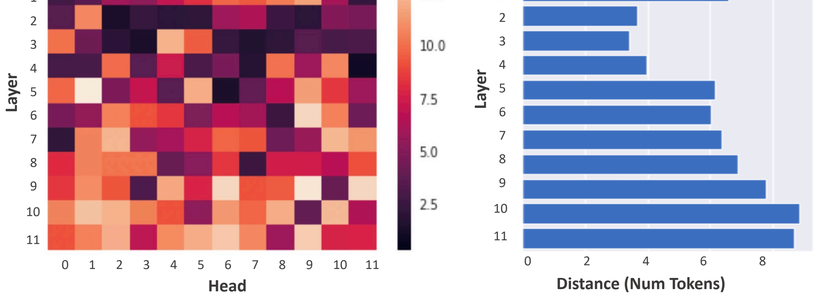


Heads attend more to higher-level parts of speech in deeper layers



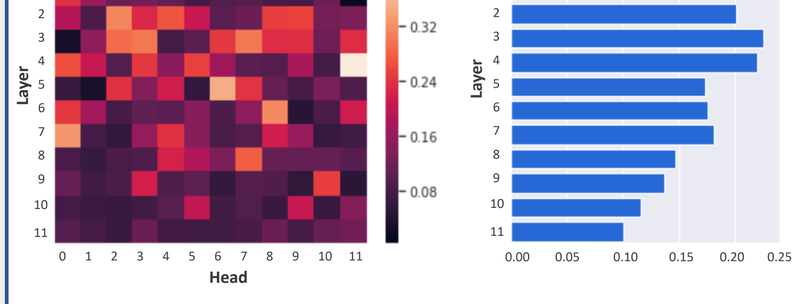
Each heatmap shows the proportion of total attention focused on the respective part of speech, broken out by layer (vertical axis) and head (horizontal axis).

The deepest layers capture the longest-range relationships.



Mean attention distance is the mean distance (in number of tokens) between all pairs of tokens, weighted by the attention between the tokens.

Attention aligns with syntactic dependencies most strongly in the middle layers of the model.



Dependency alignment is the proportion of attention that connects tokens that are also in a dependency relation with one another. This metric is inversely correlated with attention distance (left).

Attention heads target very specific lexical patterns

For each attention head, we identified the sentences (exemplars) that most strongly induced attention in that head. Below we show the top exemplar for each of 3 heads, along with the arcs with maximum attention. Other exemplars for each head followed similar patterns.

Layer 10 / Head 10: *The Australian search and rescue service is provided by Aus S AR , which is part of the Australian Maritime Safety Authority (AM S A).* — Connects acronym to associated phrase (likely for predicting next acronym piece)

Layer 11 / Head 2: *After the two prototypes were completed , production began in Mar iet ta , Georgia , where over 2 , 300 C - 130 s have been built...* — Connects comma to preceding place name (likely for predicting following place name)

Layer 11 / Head 10: *... same scale as in World War I , the prospects of Anglo – American assistance in another war with Germany appeared to be doubtful ...* — Connects end of noun phrase to head word (likely for predicting following verb)

1. Llion Jones. 2017. Tensor2tensor transformer visualization. https://github.com/tensorflow/tensor2tensor/tree/master/tensor2tensor/visualization
2. Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for Neural Machine Translation.