CMSC-636: Annotated Bibliography (Individual)

Shubhashis Roy Dipta, ID: JS93659
Team 2: SeeBel

[1] J. Vig, "A Multiscale Visualization of Attention in the Transformer Model," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019, pp. 37–42.

In Natural Language Processing (NLP), attention based models have achieved state-of-the-art results in different domains of NLP tasks. But the interpretability of how they work remains a mystery. In this work, the authors have used different idioms to increase the interpretability of attention based models. Authors have provided three interactive idioms, (1) attention-head view (Fig 1), (2) model view (Fig 2) and (3) neuron view (Fig 3). In the attention-head view, authors have explored the self-attention weights for one or multiple heads for each input. Subsequently, in the model view the authors have presented a high level overview of the whole model's attention heads. And in the neuron view, authors have drawn an idiom to show the individual neuron weights with respect to query, key, value. All of the views are interactive with the option to choose layers and heads (1, 3) and zoom in on the individual head (2).

Even if our proposal is in a different domain (Computer Vision) than this work, this paper motivated us to pursue our goal to make the vision models more interpretable by using different visualizations. Another main difference is that, in our work we have used visualization to explore both the training and prediction stage rather than only focusing on prediction model.

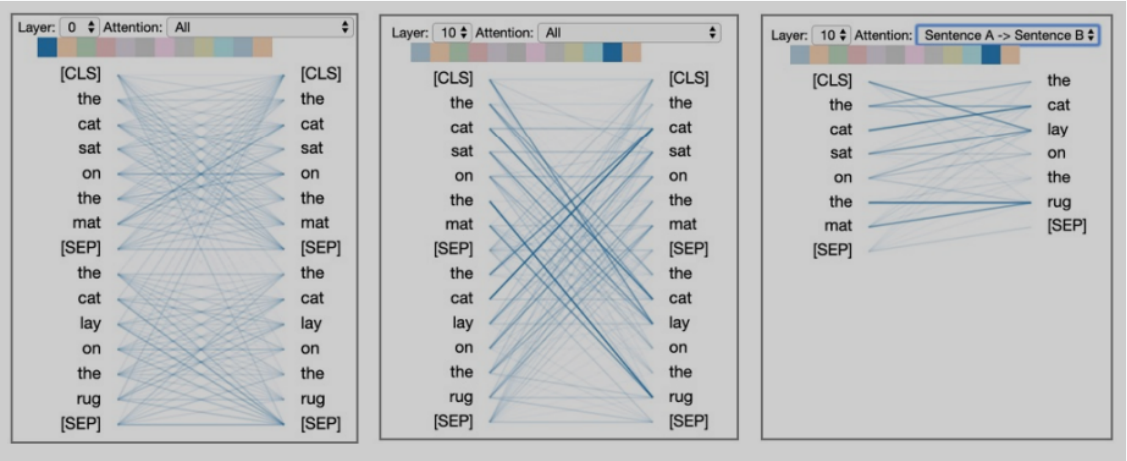

Fig 1: Attention-head view for BERT model with input as pair of sentences. The left and center figure depicts the self-attention weight for different layers, whereas the right one depicts the attention weights between the sentences. All of them can be changed based on the dropbox list to select Layer, Attention inputs and Attention heads by the color **[1]**

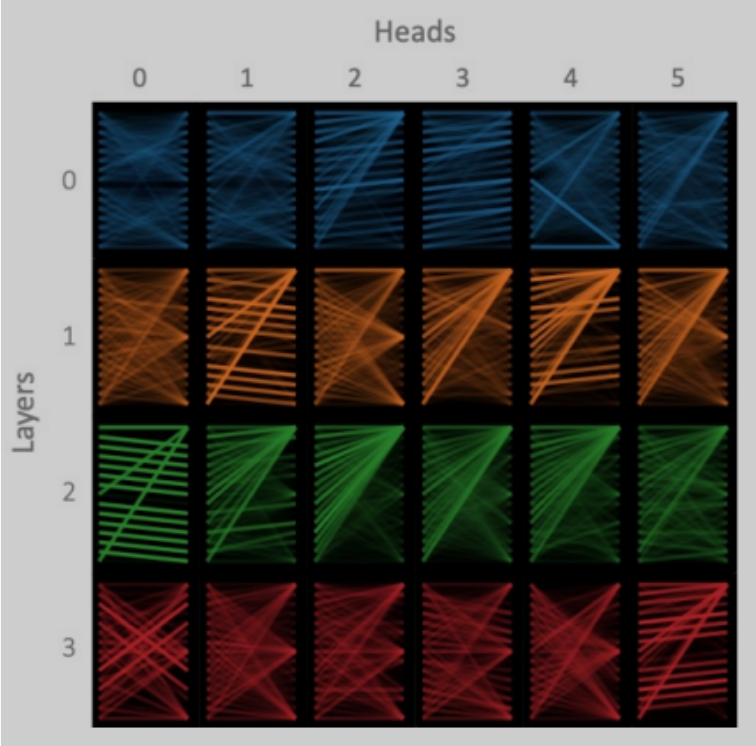

Fig 2: Model view of BERT for same input of Fig 1,
Layers 4-11 and Heads 6-11 are filtered for better visualization.
All the boxes are interactive based on selection and hovering. [1]

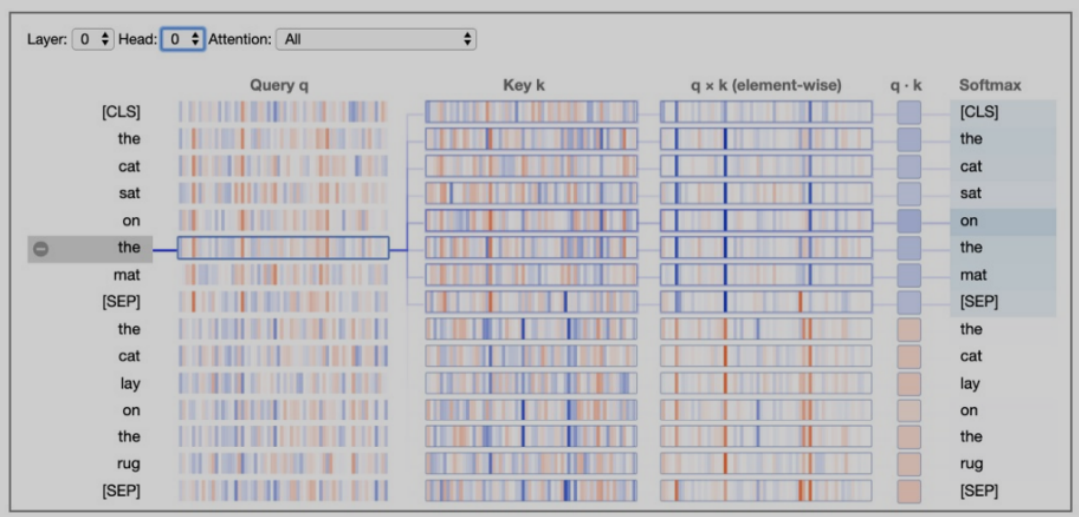

Fig 3: Neuron view of BERT for the same input as Fig 1 for layer 0 and head 0.
Positive and negative values are marked with blue and orange, respectively.
Color satuation channel is used based on the attention weights.
View can be changed based on the dropbox list of Layer, Head and Attention Inputs. [1]