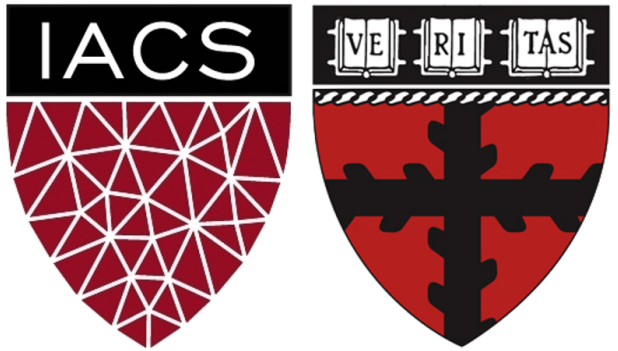
Lecture 12: Interpretation Machine Learning Model for Text Data:

BertViz, Captum, and Latam

AC295

Advanced Practical Data Science

Pavlos Protopapas



Outline

1: Communications

2: BertViz >> Demo

3: Attention with captum >> Demo

Communications

- **Milestone 3 (due 12/03)**
  - Submit code with EDA + baseline model.
  - A short writeup that explains what you have done so far and what you plan to achieve in next ~week.
- **Milestone 4 (due 12/11 & 12/15)**
  - Submit video presentation (~6 mins per group) by 12/11
  - Submit code and medium post and peer review by 12/15. We will have class 12/15 10:30 am where we will discuss few projects.

Outline

1: Communications

2: BertViz >> Demo

3: Attention with captum >> Demo

Model to Interpret Text Data

In just the last few years, a quiet revolution has been taking place in the field of Natural Language Processing (NLP). This was enabled by:

1. New Deep Learning models that dramatically improved the ability of machines to understand language.

2. Perhaps the most well-known of these models is BERT that builds on two recent trends in the field of NLP:

   i. Transformer model that - unlike traditional recurrent networks processes – forms direct connections between individual elements through attention.

   ii. Transfer learning that leverage the knowledge acquired in one domain to improve the model's performance in another one.

Attention for Language Model Review

Let's drill deeper into BERT's **attention mechanism**:

- Attention is a way for a model to assign weight to input features based on their importance to some task.
- For example, a language model trying to complete a sentence may want to pay more attention to the subject because it is more important for predicting the next word than knowing where the subject is.
- Attention can also be used to form connections between words.

Attention for Language Model Review <cont>

Attention (self attention here) is a function that takes **X** as input and returns another sequence **Y** of the same length, composed of vectors of the same length of those in **X**.

Each vector in Y is simply a weighted average of the vectors in X.

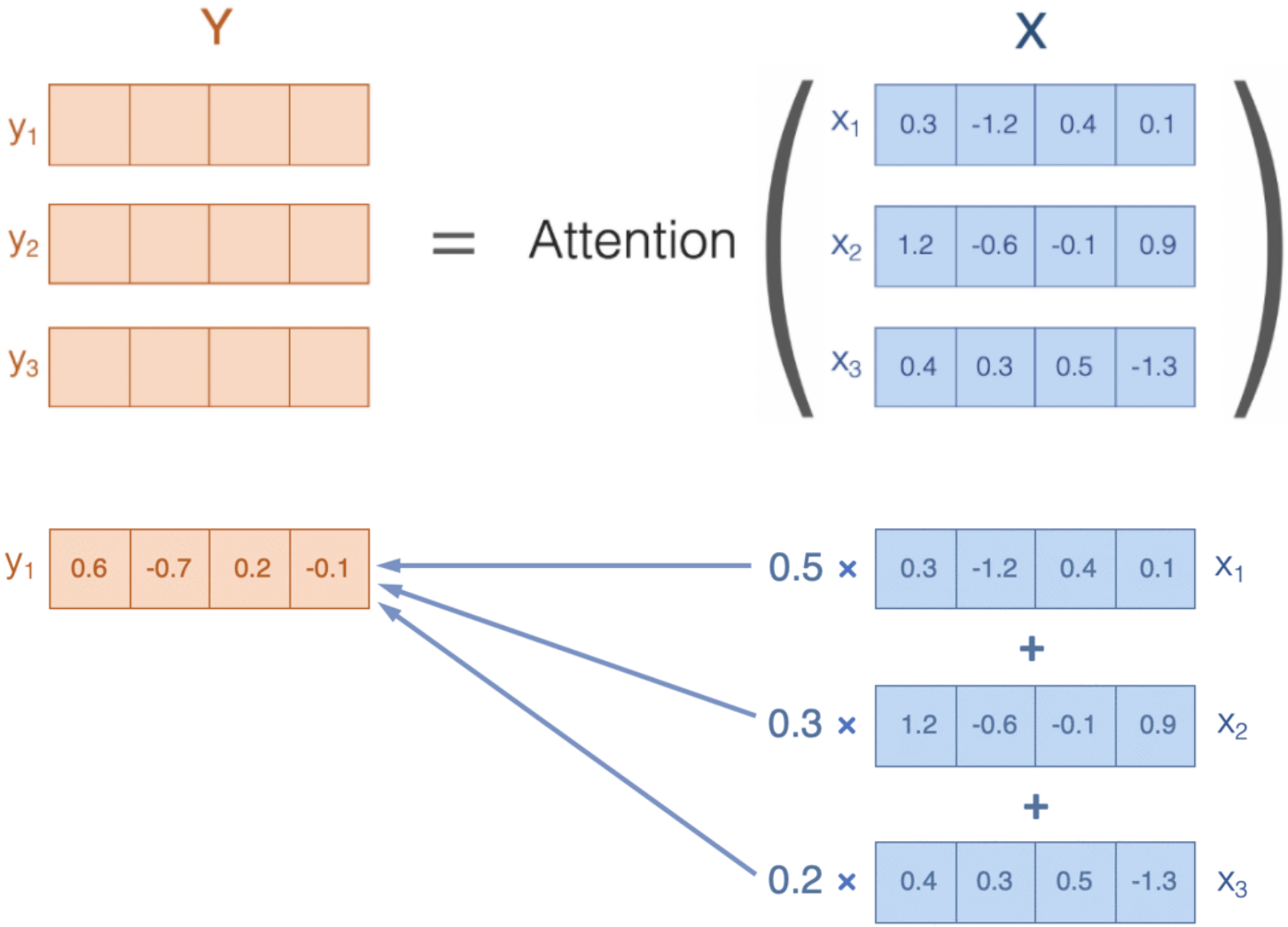


Image from Jesse Vig , *Deconstructing `BERT [Part 2]*, 2020

Attention for Language Model Review <cont>

Well, suppose that **X** represents a sequence of words like "*the dog ran*". By applying attention to the word embeddings in **X**, we have produced composite embeddings (weighted averages) in **Y**.

Why is this important? To fully comprehend language, it is not sufficient to understand the individual words that make up a sentence, the model must understand how the words relate to each other in the context of the sentence.

The attention mechanism enables the model to do this, by forming composite representations that the model can reason about.

Jesse Vig for Toward Data Science, *Deconstructing BERT [Part 2]*, 2020

BertViz: Visualizing Attention

Attention can be used as a lens through which we can see how BERT forms **composite representations to understand language.**

- BertViz is an interactive tool that visualizes attention in BERT from multiple perspectives. The attention is induced by a sample input text.
- BertViz visualizes attention as lines connecting the word being updated (left) with the word being attended to (right):
  - Uses different color intensity to reflect the attention weight
  - Users may highlight a particular word to see the attention from that word
  - It is based on the Tensor2Tensor visualization tool from Llion Jones

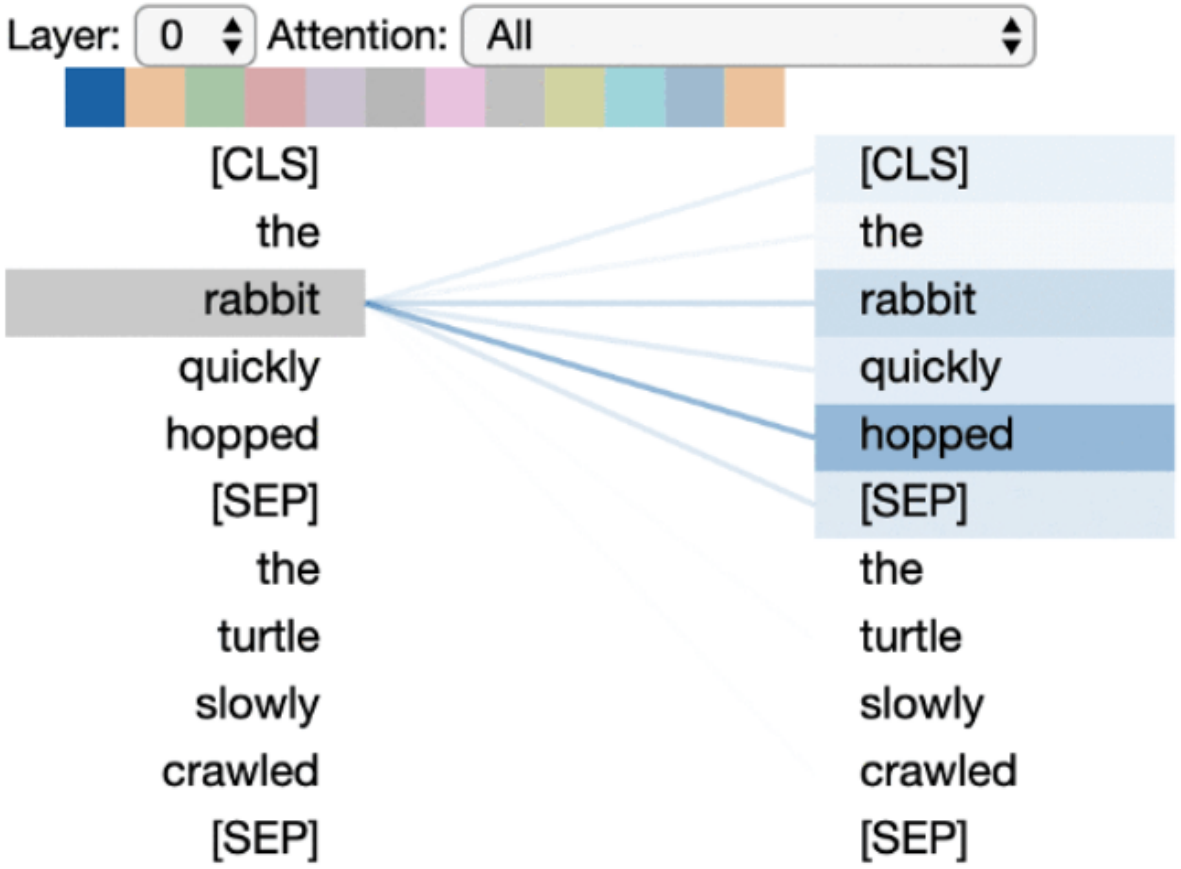
Jesse Vig , *Deconstructing BERT [Part 2]*, 2020

Visualizing Attention: Attention-head View

The model seems to understand that it should relate words to other words in the same sentence to understand their context best.

Some specific word pairs have higher attention weights than others.

In this example, understanding the relationship between these words might help the model determine that this is a description of a nature scene as opposed to a foodie's review.

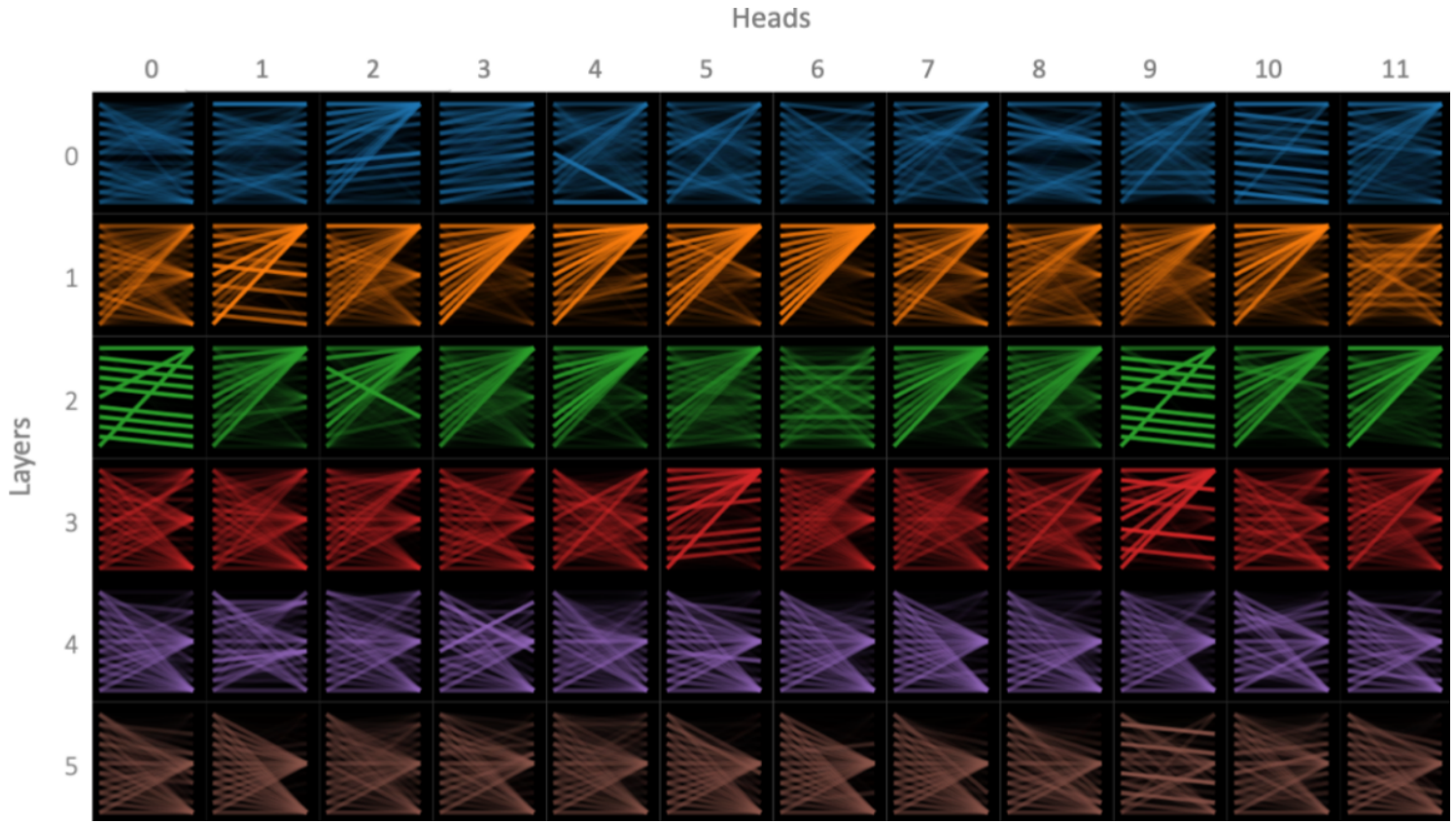


Jesse Vig for Toward Data Science, *Deconstructing BERT [Part 2]*, 2020

Visualizing Attention: Model View

BERT learns multiple attention mechanisms, called *heads*, which operate in parallel to one another.

BERT also stacks multiple layers of attention, each of which operates on the output of the layer that came before.



The version of BERT (base) considered has 12 layers and 12 heads (144 distinct attention mechanisms). Only 5 layers are plotted.

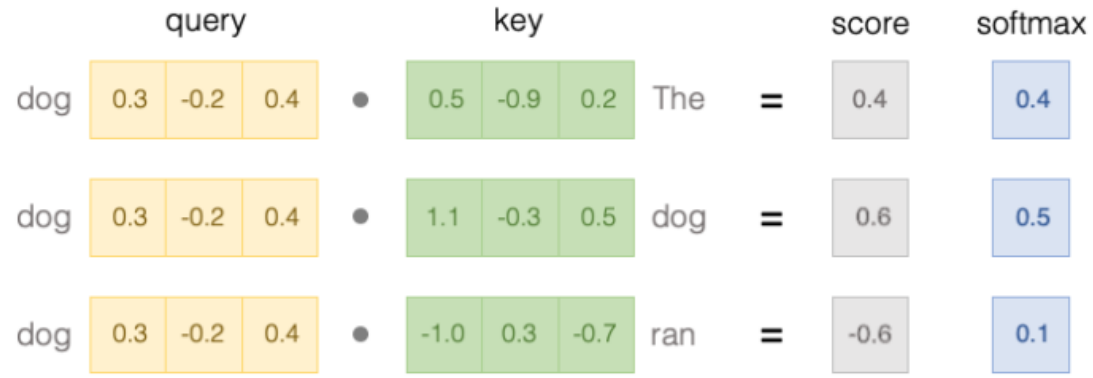
Jesse Vig for Toward Data Science, *Deconstructing BERT [Part 2]*, 2020

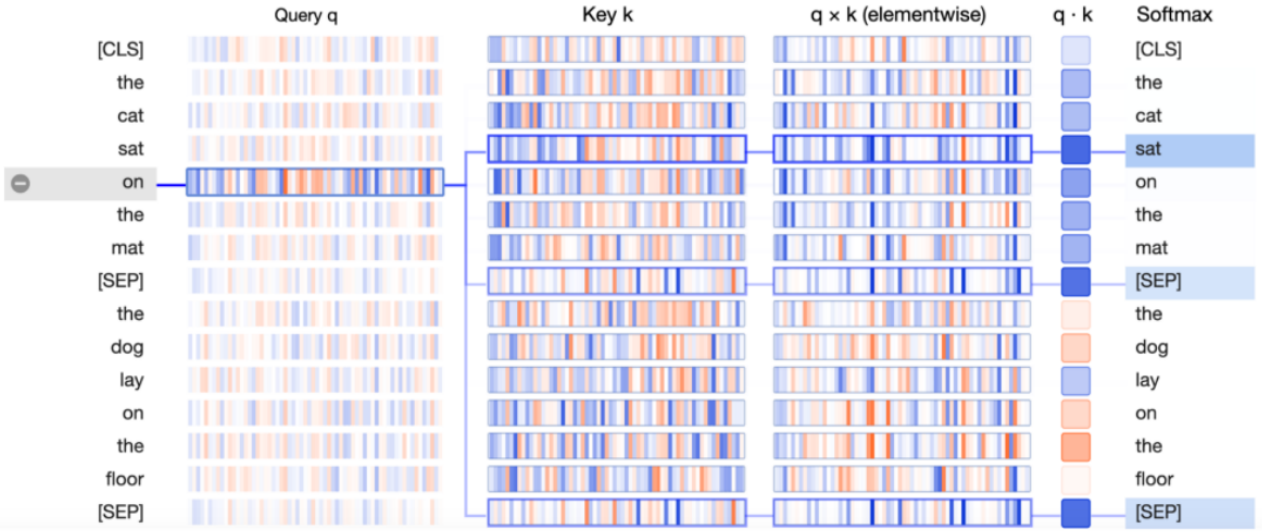
Visualizing Attention: Neuron View

The Neuron View visualizes how attention weights are computed from query and key vectors (top image show the math, bottom VizBert).

The Neuron View show the computation of attention from the selected word on the left to the complete sequence of words on the right.

Positive values are colored blue and negative values orange.





Jesse Vig for Toward Data Science, *Deconstructing BERT [Part 2]*, 2020

To the notebooks!

Attention-head View [Link]

Model View [Link]

Neuron View [Link]

Outline

1: Communications

2: BertViz >> Demo

3: Attention with captum >> Demo [LINK]

Reading

- LSTMViz
- Deconstructing BERT
- BERTVIZ
- Introduction to Captum

THANK YOU

AC295 Advanced Practical Data Science

Pavlos Protopapas