Explaining answers given by neural question answering systems

AI2100 Mid-Term Report

Paper ID 47

Abstract

Question answering has been a fundamental challenge in NLP with vast applications. State of the art neural question answering systems are able to beat humans at this task when evaluated on carefully prepared datasets. However it is difficult to explain in human comprehensible manner, how these models reach the answers that they do. Such an explanation is often important in high stakes situations such as healthcare or business applications. Thus we propose to explore explainability methods in the context of neural question answering through this project.

1. Further Literature Review

To widen our understanding of the chosen problem, we did a further review of literature, this time, about explainability in NLP, with major emphasis on BERT model for Question Answering.

[3] proposes to use Integrated Gradients (IG) [7] for visual question answering, tabular question answering as well as reading comprehension to find input attribution scores. The model used for reading comprehension is an attention based model. This method guarantees that uninfluential variables will not get any attribution and helps us to identify weaknesses of the model which can be used to improve it.

BERT and its variants have achieved state-of-the-art performance on various NLP tasks, such as sentiment analysis, question answering, etc. Although transformers are believed to be moderately interpretable through the inspection of their attention values, [2] shows that this may not be the case always.

[8] proposes that the hidden states contain valuable information that can be leveraged to get interpretable results. They inspect how QA models transform token vectors, to find the answer. To that end, they apply a set of general and QA specific probing tasks to get an insight into internal representations in transformer layers. They also performed visualization of hidden representations, using dimensionality reduction techniques like PCA [4] to get further insight. They show that transformations within BERT go through phases that are related to traditional pipeline tasks. These phases are also analogous to human reasoning processes. Also, they reveal that fine-tuning has little impact on models' semantic abilities and prediction errors can be identified in representations of early layers itself.

A similar work, [6] attempts to interpret layer-wise functionality in BERT specifically for Reading Comprehension Question Answering (RCQA) task, using Integrated Gradients . They found that, the initial layers of BERT focus on query-passage interaction, while the later layers focus more on task specific functionalities such as contextual understanding, improving the answer predictions, etc. They also observed that, BERT focuses on confusing words in the later layers, but still manages to give the right answers.

The above two works show that, the initial layers of BERT capture general semantic meanings of words, while the middle layers extract the contextual relations, eventually helping the final layers to come up with the answers to the questions.

2. Results reproduced from literature

We fine-tuned a pre-trained Bert-base-uncased model from the huggingface library. We used the SQuAD 1.1 [5] dataset for fine-tuning. In this dataset, there are 87599 examples for training and 10570 examples for evaluation. Each example consists of a question and a corresponding context.

The answer to each question is a contiguous phrase from the context itself. Therefore every answer can be represented as a start-position and an end-position with respect to the context. Thus it can be easily seen that we have reduced the task of question answering to the task of regression of two variables. In order to do this regression, a linear layer with two outputs is added on top of the Bert model. This model has already been implemented in huggingface library as BertForQuestionAnswering and was used from there.

The loss function used is the average of cross-entropy losses of start and end positions. Fine-tuning was done for 2 epochs. Total time taken was 4 hours on Tesla T4 GPU in Google colab. The optimizer used was AdamW [1]. The final training loss was 0.779 and final validation loss was 1.016.

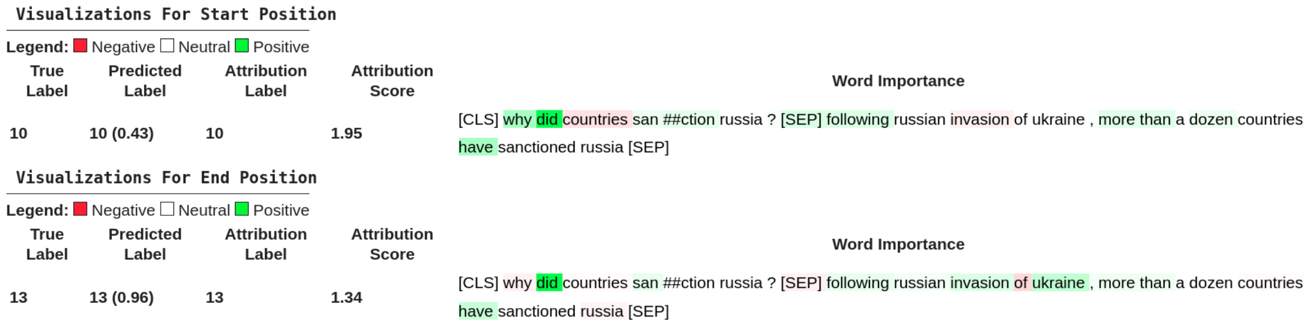
Huggingface APIs were taken full advantage of. The code can be found in the attached notebook: ”bert-fine-tune.ipynb”.

3. Preliminary results

We have used an uncased BERT model from the Hugging face library which is fine-tuned on the SQuAD-v1 to analyse the layer attributions. We used **Integrated Gradients** from **Captum** (an explainable AI toolkit) to compute the attribution scores for each word. Integrated Gradients for a model F and input x_i is calculated as follows:

$$IG(x_i) = \int_0^1 \frac{\partial F(\tilde{x} + \alpha(x_i - \tilde{x})}{\partial x_i} d\alpha \tag{1}$$

We obtained the following results,



Visualizations For Start Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
10	10 (0.43)	10	1.95	[CLS] why did countries san ##ction russia ? [SEP] following russian invasion of ukraine , more than a dozen countries have sanctioned russia [SEP]

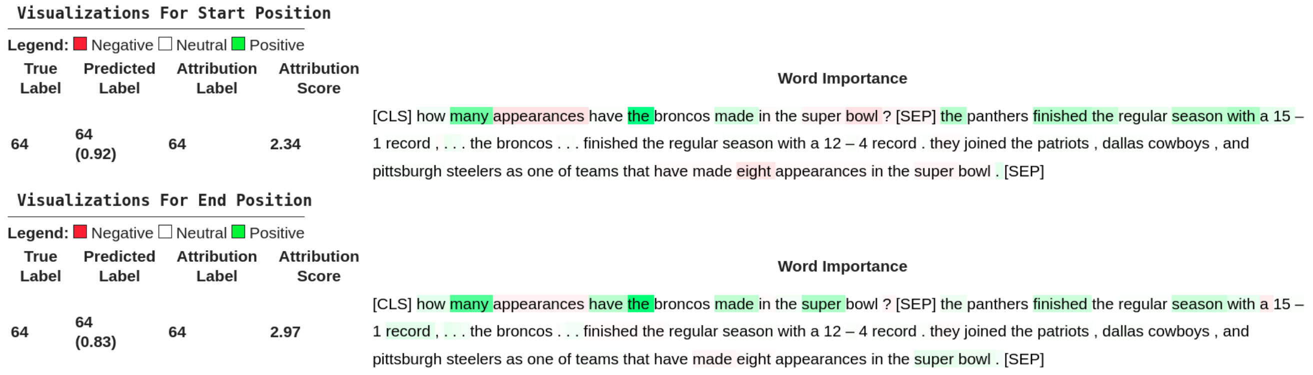
Visualizations For End Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
13	13 (0.96)	13	1.34	[CLS] why did countries san ##ction russia ? [SEP] following russian invasion of ukraine , more than a dozen countries have sanctioned russia [SEP]

Figure 1. Visualization of attribution scores for each input word.

Predicted Answer : russian invasion of ukraine



Visualizations For Start Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
64	64 (0.92)	64	2.34	[CLS] how many appearances have the broncos made in the super bowl ? [SEP] the panthers finished the regular season with a 15 – 1 record , . . . the broncos . . . finished the regular season with a 12 – 4 record . they joined the patriots , dallas cowboys , and pittsburgh steelers as one of teams that have made eight appearances in the super bowl . [SEP]

Visualizations For End Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
64	64 (0.83)	64	2.97	[CLS] how many appearances have the broncos made in the super bowl ? [SEP] the panthers finished the regular season with a 15 – 1 record , . . . the broncos . . . finished the regular season with a 12 – 4 record . they joined the patriots , dallas cowboys , and pittsburgh steelers as one of teams that have made eight appearances in the super bowl . [SEP]

Figure 2. Visualization of attribution scores for each input word.

Predicted Answer : eight

From the above results we can see that for predicting the start position of the answer, the model focuses more on the question and for predicting the end position of the answer, the model starts focusing on the context. The code for visualizing the attributions can be found in the attached notebook: ”Attribution_visualization.ipynb”

References

[1] Frank Hutter Ilya Loshchilov. Decoupled weight decay regularization. *ICLR*, 2019. 1

[2] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019. 1

[3] Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. Did the model understand the question? 1

[4] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901. 1

[5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250, 2016. 1

[6] Sahana Ramnath, Preksha Nema, Deep Sahni, and Mitesh M. Khapra. Towards interpreting BERT for reading comprehension based QA. *CoRR*, abs/2010.08983, 2020. 1

[7] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1

[8] Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. How does BERT answer questions? A layer-wise analysis of transformer representations. *CoRR*, abs/1909.04925, 2019. 1