Explaining answers given by neural question answering systems

AI2100 | Deep Learning

Pranav(AI20BTECH11004) Chirag (AI20BTECH11006)
Dishank(AI20BTECH11011) Adhvik(AI20BTECH11015)
Yashas(AI20BTECH11027)

IIT Hyderabad

May 6, 2022

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Contents

1. Abstract
2. Initial Literature Review
   - Question Answering and Explainable AI
   - Methods in Question Answering
   - Methods in Explainable AI
3. Further Literature Review
4. Implementation
5. Approach
6. Results
   - Verification of some of the previous works
   - Some of our (interesting) findings
   - Verification of the claims using adversarial examples
7. References

Abstract

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. Hence, we explored various explainability methods in the context of neural question answering through this project and verified some of their results. We also report our findings and frame adversarial examples to support our claims.

Initial Literature Review

Question Answering

- Question answering (QA):
    - It is the task of providing accurate answers to questions posed in natural language.
    - QA can be broadly studied under two categories: Open domain and Closed domain
    - **Closed domain QA** is provided with a question and a corresponding text which may have the answer to the question.
        - It provides a measure for how well a system understands text. Some popular datasets for this task are SQuAD 1.1 and SQuAD 2.0 datasets
    - **Open domain QA** is not provided a corresponding text but has to perform additional information retrieval operations on public knowledge bases like Wikipedia or internet.
        - Such QA systems are of importance as they are already being integrated into search engines and voice assistants.

Methods in Question Answering

- In the early days, the question answering systems such as BASEBALL [1], SHRDLU [13] were procedural, their knowledge bases were targeted at specific domain, the data needed to be carefully prepared by experts in the domain and they were not resilient to syntactic variations.
- Around 1990s RNNs were popular for language inference tasks, but they have their own limitations of vanishing gradient. While, LSTM [2] maintain constant error propagation, they tend to be computationally inefficient, easy to overfit
- Transformer [11] models build on the model of self-attention, and it introduces multi-head self-attention. This makes transformers parallelizable, and faster.
- BERT which is used in Google search engine is a transformer based model. It is pre-trained on large corpus of data including that from Wikipedia. It can be fine-tuned on small dataset for specific tasks.

Explainable AI

- Explainable AI (XAI) seeks to understand how machine learning models arrive at their predictions. What might consist an explanation is debatable.
- XAI systems must follow four principles [5] set out by the National Institute of standards and technology:
  - **Explanation**: Systems deliver accompanying evidence or reason(s) for all outputs
  - **Meaningful**: Systems provide explanations that are understandable to individual users
  - **Explanation Accuracy**: The explanation correctly reflects the system ' s process for generating the output
  - **Knowledge Limits**: The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output
- Explanations for machine learning models are important as they greatly help in building trust in the models. Such explanations can also pave the way to build better models in the future.

Methods in Explainable AI

- Intrinsic and post hoc explanations, are the two common types of explanations in XAI
- Intrinsic explanations rely on the interpretability of the model in use.
    - ex: linear models and decision trees are considered inherently explainable.
    - Not suited for deep learning models as they are not easily interpretable.
- Post hoc explanations try to provide explanations for individual predictions of the model.
    - ex: which part of the input was important for getting the output or what representations did the model learn for a particular input.
    - Some approaches like LIME [7], counterfactual examples [12] are model-agnostic.
    - Approaches like Saliency maps [8], Integrated Gradients [9] are specific to deep learning models.

Further Literature Review

Further Literature Review

- We focused our further literature review on explainability of BERT for QA.
- Although transformers are believed to be moderately interpretable through the inspection of their attention values, [3] shows that this may not be the case always.
- [6] attempts to interpret layer-wise functionality in BERT for RCQA 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.

Further Literature Review Contd.

- [10] proposes that the hidden states contain valuable information that can be leveraged to get interpretable results.
  - They applied 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.

Implementation

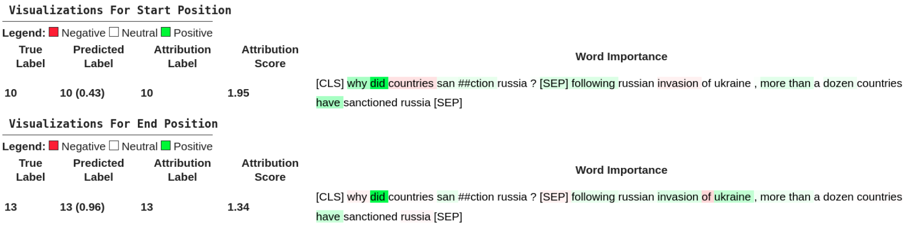
Implementation

We have used an BERT-base-uncased 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:

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

Implementation Contd.

We obtained the following result



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: Visualization of attribution scores for each input word.

Predicted Answer : russian invasion of ukraine

Approach

Approach

We analysed BERT model for the task of RCQA using integrated gradients. We fine-tuned BERT-base-uncased for 3 epochs on SQuAD-v1 and saved the fine-tuned model. To understand what the fine-tuned BERT is looking at, and to explain the functionality of its layers, we propose to use the embeddings that are produced after each layer. A few previous works look at attention scores produced by BERT to conclude facts about interpretability. But, more recent work [3] shows that attention, atleast in some cases, is not suited for explainability.

Approach Contd.

Hence we chose to work on the embeddings which carry some information about what BERT has learnt, from each layer. We analyse them using two approaches:

1. Visualisations of integrated gradients of embeddings, after each layer
2. t-SNE visualisations of the embeddings, after each layer.

In the following section, we verify some of the results which previous works have found. Also, we outline our findings for the fine-tuned model.

1. Using Integrated gradient technique, we came to the conclusion that Bert model uses certain heuristics to answer question.
2. To further verify this claim, we built adversarial examples on which these heuristics would not work. We found that Bert model did fail on these adversarial examples.

Results

Verification of some of the previous works

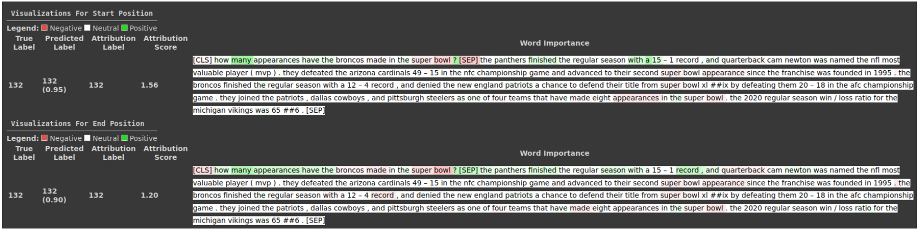
1. [6] claims that initial layers of BERT focus on words common to the question and the passage. In the later layers, the focus on such words decreases, and more focus is on supporting words that surround the answer. Their approach uses Integrated Gradients. Proceeding on similar lines, we confirm their claims.

We calculated the attribution scores for embeddings of words after each layer using integrated gradients approach. We leveraged the captum library to achieve this. We visualised these using the inbuilt-captum functionality, as well as by plotting the heatmaps. We indeed get similar results 2 which the paper claims. Initial layers focus more on words common to the question and the context. Later layers, try to look at supporting context words. Finally, the model focuses on the answer in the last layers.

Verification Contd.

Consider the context-question pair given below.
Context: The Panthers finished the regular season with a 15–1 record, and quarterback Cam Newton was named the NFL Most Valuable Player (MVP). They defeated the Arizona Cardinals 49–15 in the NFC Championship Game and advanced to their second Super Bowl appearance since the franchise was founded in 1995. The Broncos finished the regular season with a 12–4 record, and denied the New England Patriots a chance to defend their title from Super Bowl XLIX by defeating them 20–18 in the AFC Championship Game. They joined the Patriots, Dallas Cowboys, and Pittsburgh Steelers as one of four teams that have made eight appearances in the Super Bowl. The 2020 regular season win/loss ratio for the Michigan Vikings was 656.
Question: How many appearances have the Broncos made in the super bowl?
Predicted Answer: eight

Verification Contd.



Visualizations For Start Position

Legend: 🟥 Negative ⬜ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
132	132 (0.95)	132	1.56	[CLS] how many appearances have the broncos made in the super bowl ? [SEP] the panthers finished the regular season with a 15 – 1 record , and quarterback cam newton was named the nfl most valuable player (mvp) . they defeated the arizona cardinals 49 – 15 in the nfc championship game and advanced to their second super bowl appearance since the franchise was founded in 1995 . the broncos finished the regular season with a 12 – 4 record , and denied the new england patriots a chance to defend their title from super bowl xl ##ix by defeating them 20 – 18 in the afc championship game . they joined the patriots , dallas cowboys , and pittsburgh steelers as one of four teams that have made eight appearances in the super bowl . the 2020 regular season win / loss ratio for the michigan vikings was 65 ##6 . [SEP]

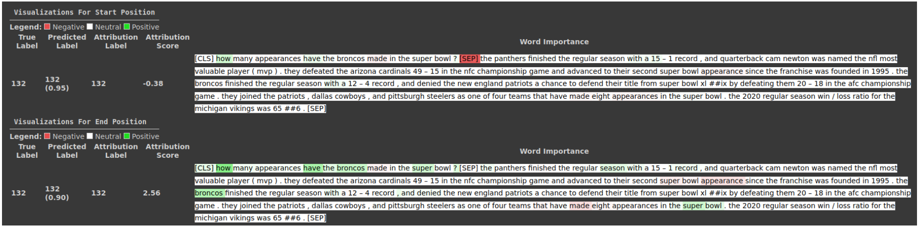
Visualizations For End Position

Legend: 🟥 Negative ⬜ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
132	132 (0.90)	132	1.20	[CLS] how many appearances have the broncos made in the super bowl ? [SEP] the panthers finished the regular season with a 15 – 1 record , and quarterback cam newton was named the nfl most valuable player (mvp) . they defeated the arizona cardinals 49 – 15 in the nfc championship game and advanced to their second super bowl appearance since the franchise was founded in 1995 . the broncos finished the regular season with a 12 – 4 record , and denied the new england patriots a chance to defend their title from super bowl xl ##ix by defeating them 20 – 18 in the afc championship game . they joined the patriots , dallas cowboys , and pittsburgh steelers as one of four teams that have made eight appearances in the super bowl . the 2020 regular season win / loss ratio for the michigan vikings was 65 ##6 . [SEP]

Figure: Visualisation of Integrated gradients of word embedding after layer 0.

Verification Contd.



Visualizations For Start Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
132	132 (0.95)	132	-0.38	[CLS] how many appearances have the broncos made in the super bowl ? [SEP] the panthers finished the regular season with a 15 – 1 record , and quarterback cam newton was named the nfl most valuable player (mvp) . they defeated the arizona cardinals 49 – 15 in the nfc championship game and advanced to their second super bowl appearance since the franchise was founded in 1995 . the broncos finished the regular season with a 12 – 4 record , and denied the new england patriots a chance to defend their title from super bowl xl ##ix by defeating them 20 – 18 in the afc championship game . they joined the patriots , dallas cowboys , and pittsburgh steelers as one of four teams that have made eight appearances in the super bowl . the 2020 regular season win / loss ratio for the michigan vikings was 65 ##6 . [SEP]

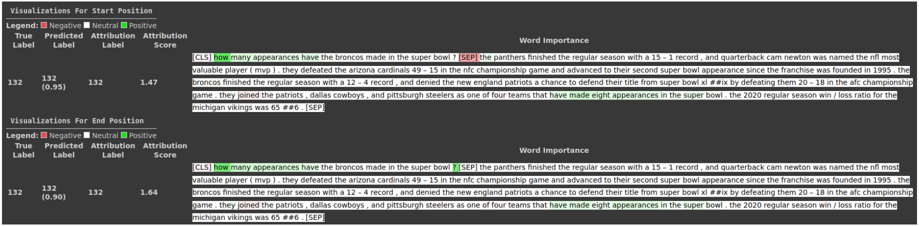
Visualizations For End Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
132	132 (0.90)	132	2.56	[CLS] how many appearances have the broncos made in the super bowl ? [SEP] the panthers finished the regular season with a 15 – 1 record , and quarterback cam newton was named the nfl most valuable player (mvp) . they defeated the arizona cardinals 49 – 15 in the nfc championship game and advanced to their second super bowl appearance since the franchise was founded in 1995 . the broncos finished the regular season with a 12 – 4 record , and denied the new england patriots a chance to defend their title from super bowl xl ##ix by defeating them 20 – 18 in the afc championship game . they joined the patriots , dallas cowboys , and pittsburgh steelers as one of four teams that have made eight appearances in the super bowl . the 2020 regular season win / loss ratio for the michigan vikings was 65 ##6 . [SEP]

Figure: Visualisation of Integrated gradients of word embedding after layer 3.

Verification Contd.



Visualizations For Start Position

Legend: 🟥 Negative ⬜ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
132	132 (0.95)	132	1.47	[CLS] how many appearances have the broncos made in the super bowl ? [SEP] the panthers finished the regular season with a 15 – 1 record , and quarterback cam newton was named the nfl most valuable player (mvp) . they defeated the arizona cardinals 49 – 15 in the nfc championship game and advanced to their second super bowl appearance since the franchise was founded in 1995 . the broncos finished the regular season with a 12 – 4 record , and denied the new england patriots a chance to defend their title from super bowl xl ##ix by defeating them 20 – 18 in the afc championship game . they joined the patriots , dallas cowboys , and pittsburgh steelers as one of four teams that have made eight appearances in the super bowl . the 2020 regular season win / loss ratio for the michigan vikings was 65 ##6 . [SEP]

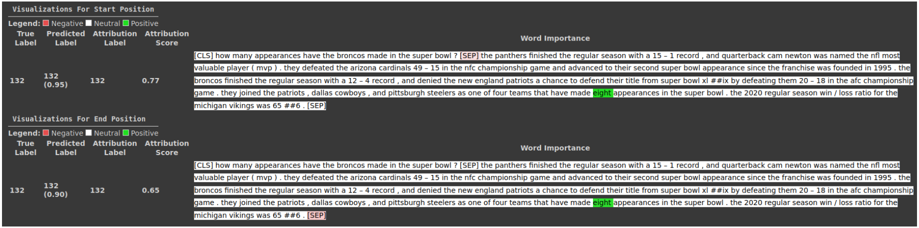
Visualizations For End Position

Legend: 🟥 Negative ⬜ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
132	132 (0.90)	132	1.64	[CLS] how many appearances have the broncos made in the super bowl ? [SEP] the panthers finished the regular season with a 15 – 1 record , and quarterback cam newton was named the nfl most valuable player (mvp) . they defeated the arizona cardinals 49 – 15 in the nfc championship game and advanced to their second super bowl appearance since the franchise was founded in 1995 . the broncos finished the regular season with a 12 – 4 record , and denied the new england patriots a chance to defend their title from super bowl xl ##ix by defeating them 20 – 18 in the afc championship game . they joined the patriots , dallas cowboys , and pittsburgh steelers as one of four teams that have made eight appearances in the super bowl . the 2020 regular season win / loss ratio for the michigan vikings was 65 ##6 . [SEP]

Figure: Visualisation of Integrated gradients of word embedding after layer 7.

Verification Contd.



Visualizations For Start Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
132	132 (0.95)	132	0.77	[CLS] how many appearances have the broncos made in the super bowl ? [SEP] the panthers finished the regular season with a 15 – 1 record , and quarterback cam newton was named the nfl most valuable player (mvp) . they defeated the arizona cardinals 49 – 15 in the nfc championship game and advanced to their second super bowl appearance since the franchise was founded in 1995 . the broncos finished the regular season with a 12 – 4 record , and denied the new england patriots a chance to defend their title from super bowl xl ##ix by defeating them 20 – 18 in the afc championship game . they joined the patriots , dallas cowboys , and pittsburgh steelers as one of four teams that have made eight appearances in the super bowl . the 2020 regular season win / loss ratio for the michigan vikings was 65 ##6 . [SEP]

Visualizations For End Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
132	132 (0.90)	132	0.65	[CLS] how many appearances have the broncos made in the super bowl ? [SEP] the panthers finished the regular season with a 15 – 1 record , and quarterback cam newton was named the nfl most valuable player (mvp) . they defeated the arizona cardinals 49 – 15 in the nfc championship game and advanced to their second super bowl appearance since the franchise was founded in 1995 . the broncos finished the regular season with a 12 – 4 record , and denied the new england patriots a chance to defend their title from super bowl xl ##ix by defeating them 20 – 18 in the afc championship game . they joined the patriots , dallas cowboys , and pittsburgh steelers as one of four teams that have made eight appearances in the super bowl . the 2020 regular season win / loss ratio for the michigan vikings was 65 ##6 . [SEP]

Figure: Visualisation of Integrated gradients of word embedding after layer 11.

Verification Contd.

2. [10] claims that initial layers look at the semantic meanings of each word, and subsequent layers look at relations between the words in context. The final layers filter out the relevant information, and come up with the answer. Their approach uses PCA visualisation of embeddings. Proceeding on similar lines, we confirm their claims.

We perform t-SNE visualisations of embeddings after each layer, instead of PCA, in order to preserve more information, which might have possibly lost during PCA. We indeed get similar results 6 which the paper claims. Initial layers look at semantic meanings of words (words like white, black or loves, like, or cat, dog occur together). Later layers, learn relations in context (white, cat, black, dog become together). Finally, all the important information is filtered out (apples, is reasonably seperated).

Verification Contd.

Consider the context-question pair given below.
Context: Alfred is a white cat and likes apples. Bobby is a black dog which loves mangoes.
Question: What does Alfred like?
Predicted Answer: Apples

Verification Contd.

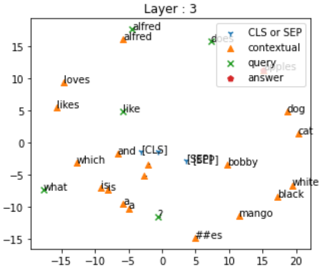


Figure: t-SNE visualisations of word embeddings after layer 3.

Verification Contd.

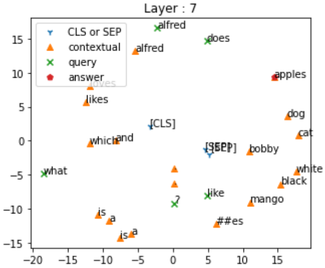


Figure: t-SNE visualisations of word embeddings after layer 7.

Verification Contd.

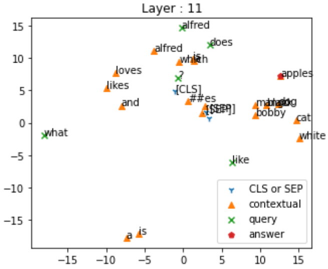


Figure: t-SNE visualisations of word embeddings after layer 11.

Some of our (interesting) findings

Analysing the heat-maps for different examples, we observed few patterns. First was that initially the model was focusing greatly on question words like "Who, What, When, Where". The model was also focusing on possible answers to such nouns and numbers in the query.
Consider the context-question pair given below 9
Context: 93 players have been awarded the Fewest Prized Team distinction for the Champ Bowl.
Question: How many players have been awarded the Fewest Prized Team distinction for the Super Bowl?
Second was that as mentioned previously, the model was focusing on words which were same in the question and the context. Then the model was looking at neighbours of the words it focused on initially, as mentioned above. (Look at section 1).

Verification Contd.

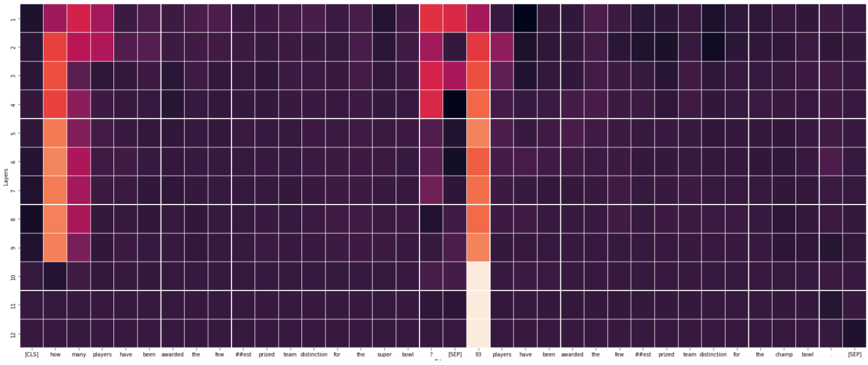


Figure: Example demonstrating importance in initial layers

Findings contd.

Based on these observations, we claim that the Bert model uses heuristics to predict the start and end tokens. These heuristics work as follows:

1. Recognize the question type by focusing on the question words. Focus on candidate answers.
2. Focus on words that are same in both the question and the answer.
3. Look at neighbouring words of the words on which we focused in the above steps and predict start and end tokens from these.

How exactly the model picks start and end token during heuristic 3 is still a mystery.

Verification of the claims using adversarial examples

We constructed adversarial examples by assuming that the model uses these heuristics. We expected that the model to fail on these examples and that is exactly what happened.

To test heuristic 1, consider the context-question pair given below

Context: Yes, a person can jump from a hill and still be alive.

Question: After jumping from a hill, can a person be alive?

Predicted Answer: still be alive

Verification Contd.

Clearly, the model fails here. To test heuristic 2, we constructed examples where there were no similar words or where there were more than one similar word. Consider the context-question pairs given below

Context: Trump has a Trump card having his photo.

Question: What has Trump's photo?

Predicted Answer: Trump has a Trump card

Context: Ram was drawing with pencils. Teacher was picking her nose with pencil. Ram observed that pencils can be used to pick nose.

Question: What was Ram's conclusion?

Predicted Answer: None

Verification Contd.

In the last example, if ”conclusion” is replaced with ”observation” in the query, the model gives correct output. This further shows the necessity of same words. This also makes the model brittle and easy to fool because there is a lack of language understanding.

Verification Contd.

To test heuristic 3, we constructed an example in which the correct answer was much farther than words on which model is focusing using heuristic one and two. Consider the context-question pair given below 10

Context: Ram had a big white great beautiful majestic creative cute cat and bad dog.

Question: What did Ram have that was bad?

Predicted Answer: a big white great beautiful majestic creative cute cat and bad dog

Verification Contd.

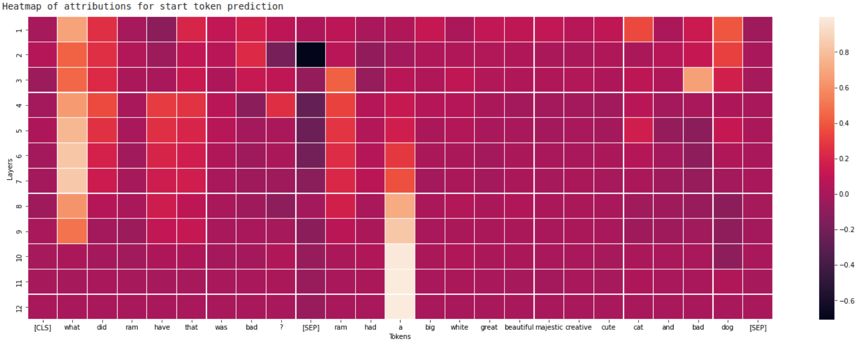


Figure: Adversarial example

Verification Contd.

The heat-map for the last example shows us the full story of how the model predicts the answer. First the model focuses on "cat" and "dog" as they are candidate answers for the question word "what". The model also focuses on the word "what". Then the model finds the word "Ram" to be same in both the question and the context. Thus it focuses on "Ram" in the context. Then it looks at it's neighbouring word: "a" and predicts it for start token. The model also focuses on the word "bad" as it is also same in context and question. However, in step 3 of the heuristics, the model chooses to not "go ahead" with the word "bad".

References

[1] Bert F. Green, Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. Baseball: An automatic question-answerer. In *Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference*, IRE-AIEE-ACM '61 (Western), page 219–224, New York, NY, USA, 1961. Association for Computing Machinery.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

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

[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.

[5] P. Jonathan Philips, Catrina A. Hahn, Peter C. Fontana, David A. Broniatowski, and Mark A. Przybocki. Four principles of explainable artificial intelligence, 2020. National Institute of Standards and Technology, U.S. Department of Commerce.

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

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[8] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

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

[10] 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.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[12] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[13] Nigel Ward. *SHRDLU*. 01 2006.

Thank you!