Explaining answers given by neural question answering systems

AI2100 preliminary project 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. Introduction

The proposed problem statement sits at the intersection of two fields: Question Answering and Explainable AI. Thus we will be briefly discussing each of them next. Individual discussion on both these topics will make it ample clear why they need to be pursued together.

1.1. Question Answering

Question answering (QA) 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 (also known as reading comprehension) is the task where the QA system is provided with a question and a corresponding text which may have the answer to the question. The QA system has to correctly find the answer from the text. Some popular datasets for this task are SQuAD 1.1 and SQuAD 2.0 datasets. Closed domain QA is of particular theoretical interest as it provides a measure for how well a system 'understands' text.

Open domain QA is more general, in the sense that the QA system is not provided a corresponding text but has to perform additional information retrieval operations on public knowledge bases like Wikipedia or internet to find the relevant information. Such QA systems are of importance as they are already being integrated into search engines and voice assistants.

1.2. Explainable AI

Explainable AI (XAI) seeks to understand how machine learning models arrive at their predictions. What might consist an explanation is debatable. However, it is popularly understood that XAI systems must follow four principles [7] 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.

2. Literature Review

2.1. Methods in QA

In the early days, the question answering systems such as BASEBALL [4], SHRDLU [14] 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. Feedforward RNNs were popular for tasks requiring language inference in 1990s. Feedforward recurrent networks fail at encoding long term memory because of exploding or vanishing gradients. LSTM [5] on the other hand maintains constant error backpropagation, and it generalises well with syntactic variations in the input. However, LSTMs are computationally inefficient, easy to overfit, and they tend to perform poorly on unseen words. Unseen words are usually common, since several word embedding implementations require a word to appear a minimum number of times owed to incorrect spellings in the corpus.

BiDAF [9] handles the case of unseen words using character level embedding, which is a 1D-CNN used to find the vector representation of the unseen word by character decomposition. BiDAF also uses attention mechanism to focus on the words that might be important in the task. However like LSTM, BiDAF also suffers from being computationally inefficient.

Transformer [12] models build on the model of self-attention and get rid of RNN based models such as LSTM and replace it by multi-headed self-attention. This makes transformers parallelizable, and faster.

BERT [3] is a transformer based deep-learning model, which is currently being used in Google search queries. BERT is a bidirectional, self-supervised language representation model. It is pre-trained on a large corpus of plain-text, which is computationally expensive task. It can be fine tuned on small datasets for specific tasks, several models with fine tuning are also available in public domain such as RoBERTa [6].

Recently, there have been some progress in using symbolic artificial intelligence with pre-trained deep learning models, the results of which outperform both the symbolic models, and the deep learning models in the context of Natural language inferencing [1].

2.2. Methods in Explainable AI

There are two common types of explanations given in XAI: intrinsic explanations and post hoc explanations.

Intrinsic explanations rely on the interpretability of the model in use. For example, linear models and decision trees are considered inherently explainable. However, often deep learning models are not easily interpretable. Thus post hoc explanations tend to be more popular for deep learning models.

Post hoc explanations try to provide explanations for individual predictions of the model. They try to identify which part of the input was important for getting the output or what representations did the model learn for a particular input. For example, [2] talks about how BERT's attention heads attend to linguistic notions like syntax and coreference, how attention heads in similar layer exhibit similar behaviour, etc.

Some post hoc explainability approaches are model agnostic, and can be used for wide variety of ML models. LIME [8] being one of them, tries to provide explanations by locally approximating the original model with surrogate interpretable model. Another set of explainability methods based on counterfactual examples [13] give explanations by coming up with an example input which is close to an original input data but resulting in a different output.

Other post hoc explainabilty approaches specific to deep learning models try to compute input attribution by calculating gradients of the model's output with respect to input features. Saliency maps [10] approximate the network as first order Taylor expansion, where the gradients are coefficients of features in the linear representation of model. Integrated Gradients [11] is another attribution method primarily based on two axioms of Sensitivity and Implementation Invariance. It computes input attribution by taking integral of gradients along the path from a given baseline to input.

3. Conclusion

As alluded to in section 1.1, QA is often seen as a test of how much a model understands text. If a model truly understands a text and then answers questions based on the text, we expect it to be able to explain the reasons for the answers. This would serve as a better test of understanding. This reason in particular among other benefits of XAI encourages us to pursue explainability in QA.

References

[1] Zeming Chen, Qiyue Gao, and Lawrence S. Moss. Neural-log: Natural language inference with joint neural and logical reasoning. *CoRR*, abs/2105.14167, 2021. 2

[2] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019. 2

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. 2

[4] 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. 1

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 2

[7] 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. 1

[8] 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. 2

[9] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603, 2016. 2

[10] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. 2

[11] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 2

[12] 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. 2

[13] 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. 2

[14] Nigel Ward. *SHRDLU*. 01 2006. 1