28/05/2020

Robust Rationalization of Ratings from Reviews

Description:

Automated predictions require explanations to be interpretable by humans. However, neural methods generally offer little transparency, and interpretability often comes at the cost of performance. In many domains, providing the underlying reasons for such a prediction has a real impact. However, what is meant by an explanation in natural language processing is ambiguous. Therefore, in this project, we will be working on reviews for rating predictions, where the ambiguity of a justification is minimal. **We define a rationale/mask as one or more text fragments from the input text that alone suffice for the prediction.**

Many works have been applied to **single-aspect sentiment analysis for reviews**. At each training, a model infers a **binary rationale** for **one aspect**. There are different paradigms to infer rationales:

1) A generator-predictor framework for rationalization; a co-operative game that maximizes the mutual information between the selected rationales and labels [2]. [3,7] identifies factual & counter-factual masks simultaneously with a game-theoretical approach. [4] is a variant robust to spurious correlation (i.e., words related to another aspect to predict the rating of the target) based on a prior *environment* (i.e., some brands are better at the appearance).
2) Using reinforcement learning and a trained model [5].
3) Integrating rationale annotation during training and constraint the attention to be close to the rationale [6].

On the one hand, a strict assignment of words to aspects might lead to ambiguities that are difficult to capture with a binary mask. In the text ”The room was large, clean and close to the beach.”, ”room” refers to the aspects Room, Cleanliness, and Location. On the other, collecting human rationales at scale is expensive and impractical.

Recently, we propose MAM [1], a model addressing **multi-aspect sentiment analysis** and generating a **probabilistic multi-dimensional mask** (one dimension per aspect). In terms of performance, it achieves higher results than the prior work [2] and attention mechanisms, without any assumption on the data[1]. We believe that this vectorial representation of a mask help to reduce spurious correlations.

Different directions to improve the model: the generative process, the masker component, the loss, the integration of the masks to the predictor.

According to you, what would be your propositions to improve the model? (no worries I have ideas)

Let’s work together!

Prerequisites:

- **Understand details of prior work [1,2] and general concepts of [3,4,7]**
- Strong Knowledge in Machine Learning and NLP.
- Efficient in Tensorflow or Python.

Supervisor: Diego Antognini (diego.antognini@epfl.ch)

References:

[1] https://www.dropbox.com/s/x1vp0u65r3hkpef/Multi_Dimensional_Explanation_of_Ratings_from_Reviews.pdf
[2] https://people.csail.mit.edu/taolei/papers/emnlp16_rationale.pdf
[3] https://arxiv.org/pdf/1910.12853.pdf
[4] https://arxiv.org/pdf/2003.09772.pdf
[5] https://arxiv.org/pdf/1612.08220.pdf
[6] https://arxiv.org/pdf/1808.09367.pdf
[7] https://www.aclweb.org/anthology/D19-1420.pdf
[8] https://arxiv.org/pdf/1907.04907.pdf
[9] http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[1] [2,3,4,7] assume only decorrelated subsets of reviews and consider only three aspects out of five.