CS470 Fall Semester

Applying XAI methods for Transformer model to analyze reviews

[TEAM 21] 20170154 김정원 20170717 한승우 20180237 박민규 20200506 이중훈 20224591 이현석

Introduction

We are placed in many moments of choice in our daily lives. In this situation, most of us make decisions based on reviews. Reviews are a valuable source of detailed information from the point of view of consumers, sellers and companies. However, the problem is that the volume is too large to read all the reviews. Some of the reviews may lead us biased and making the wrong choice. So, our team is going to develop a service summarizes reviews and provides it.

The biggest challenge in summarizing reviews is how to find key contents in each review. We should be able to distinguish main point and unimportant contents from long reviews. If rating system is included, we should look for contents related to the rating in the reviews. This is especially important for the sellers because they want to find the reasons of low rating and improve it. This is also the limitation of classical review analysis. Classical review analysis works by learning world embeddings or clustering reviews, which fail to analyze the causes of low rating.

XAI – What is XAI and why do we use it?

To solve these problems, we propose a way to use Explainable AI, XAI. XAI is a method of interpreting a model like activation visualization we learned in the lecture. We can find which features of the data are most influenced by using XAI. There are XAI techniques that can be applied to LSTM or transformer models, and by using them, we can apply XAI to natural language processing deep learning models.

a worthy entry into a very difficult genre .
it 's a good film -- not a classic , but odd , entertaining and authentic .
it never fails to engage us .

Overall Procedure

Dataset : Amazon Review Dataset

Preprocessing
remove duplicate contents and balance labels.

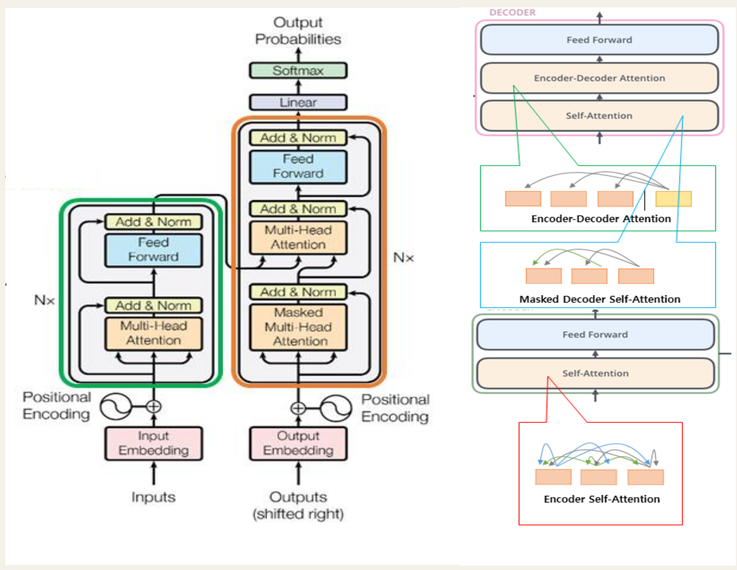
Fine-tuning
fin-tuned transformer(BERT) model.
We tried two ways
1. Binary classification (positive or negative)
2. Regression (rating)

Generating explanation
Using fine-tuned model, generate explanation for each review. Then we get scores for each word that represents how important each word is.

Summarization
Choose top-N most important words Based on occurrences, ratings and generated scores.

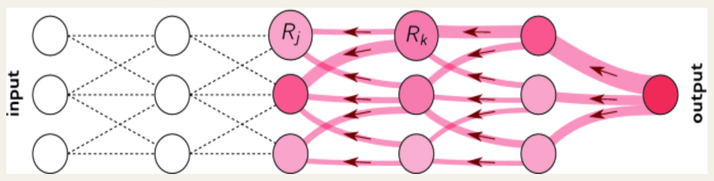
Transformer

Transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. It is used primarily in the fields of NLP and CV. Like RNN, transformer is designed to process sequential input data, such as natural language, with applications towards tasks such as translation and text summarization. However, unlike RNN, transformer process the entire input all at once. The attention mechanism provides context for any position in the input sequence. Below figure shows the architecture of transformer.



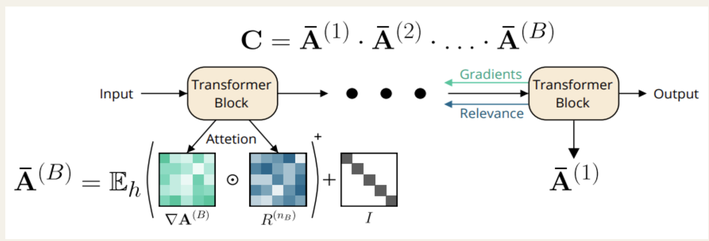
Layer-wise Relevance Propagation (LRP)

For individual prediction, LRP calculates how much each feature attributes to prediction.
Consider adjacent layer j and k. $\sum_j a_j w_{jk} = a_k$
Then it can be interpreted as "a_j attributes a_k as much as $a_j w_{jk}$ " Likewise, back-propagate output value to input features



$R_j = \sum_k \frac{a_j w_{jk}}{\sum_j a_j w_{jk}} R_k$, $\sum R_j = \sum R_k$ = output

We used explanation technique that combines LRP and attention

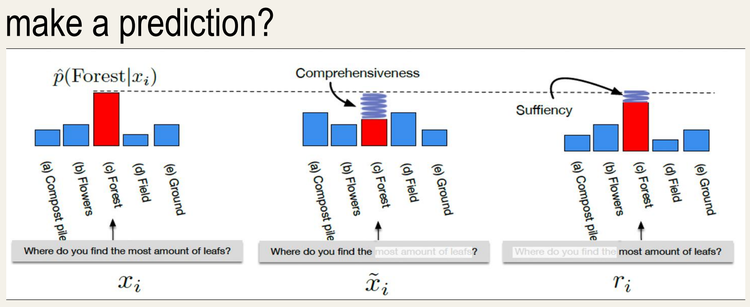


Metric

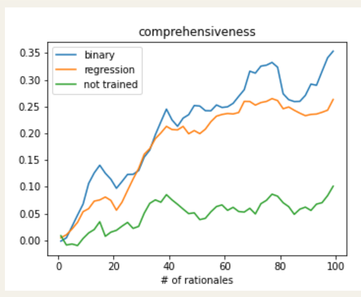
Intersection Over Union (IOU)
IOU = area of overlap / area of union

Comprehensiveness = $m(x_i)_j - m(x_i \backslash \mathrm{r}_i)_j$
Are all features (words) needed to make a prediction selected?

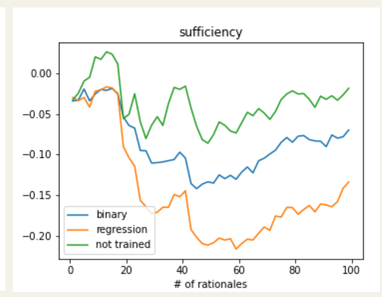
Sufficiency = $m(x_i)_j - m(r_i)_j$
Do the extracted rationales contain enough signal to make a prediction?



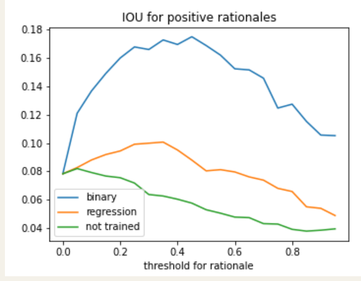
Result



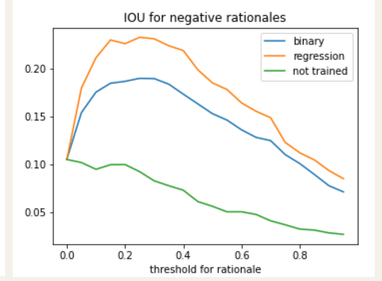
Figure(a). Comprehensiveness



Figure(b). Sufficiency



Figure(c). IOU - Positive



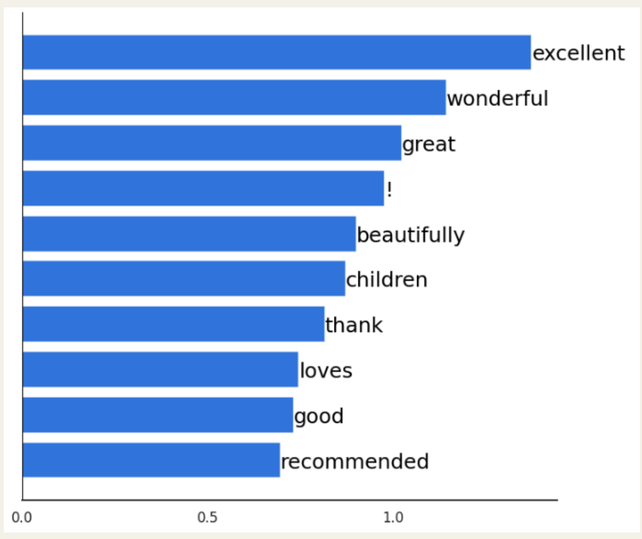
Figure(d). IOU - Negative

Figure(a),(b): After training, the two models show higher comprehensiveness score and than lower sufficiency score then untrained model

Figure(c),(d): Compare the rationale selected by the person and the rationale selected by the model. After training, models choose rationales more like a human.

there are many editions of robert louis stevenson ' s a child ' s garden of verses . this one is notable because it is illustrated by eve garnett , who has filled this edition with most delightful pencil sketches and drawings . next to tasha tudor ' s illustrations for the poems she selected for her collection , wings from the wind , i like these lovely (on a smaller scale) drawings by eve garnett . garnett was the first british writer of children ' s stories to write wonderful stories of working class families (in london) (the family terrible , blurry , faded reproduction of original book . this product deserves no starts . amazon shouldn ' t sell it at all .

the kindle version of the everyman ' s library children ' s classics is poorly formatted . some of the poems which originally appeared within illustrations are only in the illustration , and the illustrations are small and not very high quality , so those poems are difficult to read .



Conclusion and Suggestion

Our study proposed a new method using XAI. We used LRP to Transformer model and found the best XAI method based on metrics such as comprehensiveness and sufficiency. Positive and negative rationale were found based on review analysis. We suggest that summary performance can be improved by finding non-trivial rationale or using the text rank and our relevance score. Also, we can use other dataset than the Amazon dataset. Finally, we can provide our review analysis as a web API.

References

[1] Bach S. et. al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PLoS ONE, 2015
[2] Montavon, G. et al. "Explaining nonlinear classification decisions with deep Taylor decomposition." Pattern Recognition, 2017
[3] Hila Chefer, Shir Gur, Lior Wolf, "Transformer Interpretability Beyond Attention Visualization," CVPR, 2021
[4] Jay DeYoung, et al. "Eraser: A benchmark to evaluate rationalized nlp models." arXiv preprint arXiv:1911.03429. 2019
[5] Jianmo Ni, Jiacheng Li, Julian McAuley, "Justifying recommendations using distantly-labeled reviews and fined-grained aspects" EMNLP, 2019, dataset: https://nijianmo.github.io/amazon/