Learning Explanations from Language Data

Abstract

PatternAttribution is a recent method, introduced in the vision domain, that explains classifications of deep neural networks. We demonstrate that it also generates meaningful interpretations in the language domain.

1 Introduction

In the last decade, deep neural classifiers achieved state-of-the-art results in many domains, among others in vision and language. Due to the complexity of a deep neural model, however, it is difficult to explain its decisions. Understanding its decision process potentially allows to improve the model and may reveal new knowledge about the input. Recently, it was claimed that "popular explanation approaches for neural networks (...) do not provide the correct explanation, even for a simple linear model." They show that in a linear model, the weights serve to cancel noise in the input data and thus the weights show how to extract the signal but not what the signal is. This is why explanation methods need to move beyond the weights, the authors explain, and they propose the methods "PatternNet" and "PatternAttribution" that learn explanations from data. We test their approach in the language domain and point to room for improvement in the new framework.

2 Methodology

Kindermans et al. assume that the data x passed to a linear model $w^T x = y$ is composed of signal (s) and noise (d, from distraction) $x = s + d$. Furthermore, they also assume that there is a linear relation between signal and target y as $= s$ where a_s is a so called signal base vector, which is in fact the "pattern" that PatternNet finds for us. As mentioned in the introduction, the authors show that in the model above, w serves to cancel the noise such that

$$w^T d = 0, w^T s = y. \tag{1}$$

They go on to explain that a good signal estimator $S(x) = \hat{s}$ should comply to the conditions in Eqs. 1 but that these alone form an ill-posed quality criterion since $S(x) = u(w^T u)^{-1} y$ already satisfies them for any u for which $w^T u \neq 0$. To address this issue they introduce another quality criterion over a batch of data x:

$$\rho(S) = 1 - \max_v corr(y, v^T(x - S(x))) \tag{2}$$

and point out that Eq. 2 yields maximum values for signal estimators that remove most of the information about y in the noise. We argue that Eq. 2 still is not exhaustive. Consider the artificial estimator

$$S_m(x) = mx + (1 - m)s = s + md \tag{3}$$

which arguably is a a bad signal estimator for large m as its estimation contains scaled noise, md. Nevertheless, it still satisfies Eqs. 1 and yields maximum values for Eq. 2 since

$$x - S_m(x) = (1 - m)(x - s) = (1 - m)d \tag{4}$$

is again just scaled noise and thus does not correlate with the output y. To solve this issue, we propose the following criterion:

$$\rho'(S) := \max_{v_1} corr(w^T x, v_1^T S(x)) - \max_{v_2} corr(w^T x, v_2^T (x - S(x))). \tag{5}$$

The minuend measures how much noise is left in the signal, the subtrahend measures how much signal is left in the noise. Good signal estimators split signal and noise well and thus yield large $\rho'(S)$. We leave it to future research to evaluate existing signal estimators with our new criterion. For our experiments, the authors equip us with expressions for the signal base vectors as for simple linear layers and ReLU layers. For the simple linear model, for instance, it turns out that $a_s = cov(x, y)/\sigma_y^2$. To retrieve contributions for PatternAttribution, in the backward pass, the authors replace the weights by $w \cdot a_s$.

3 Experiments

To test PatternAttribution in the NLP domain, we trained a CNN text classifier on a subset of the Amazon review polarity data set. We used 150 bigram filters, dropout regularization and a dense FC projection with 128 neurons. Our classifier achieves an F1 score of 0.875 on a fixed test split. We then used PatternAttribution to retrieve neuron-wise signal contributions in the input vector space. To align these contributions with plain text, we summed up the contribution scores over the word vector dimensions for each word and used the accumulated scores to scale RGB values for word highlights in the plain text space. Positive scores are highlighted in red, negative scores in blue. This approach is inspired by similar work. Example contributions are shown in Figs. 1 and 2.

4 Results

We observe that bigrams are highlighted, in particular no highlighted token stands isolated. Bigrams with clear positive or negative sentiment contribute heavily to the sentiment classification. In contrast, stop words and uninformative bigrams make little to no contribution. We consider these meaningful explanations of the sentiment classifications.

5 Related Work

Many of the approaches used to explain and interpret models in NLP mirror methods originally developed in the vision domain. In this paper we implemented a similar strategy. Following Kindermans et al., however, our approach improves upon the latter methods for the reasons outlined above. Furthermore, PatternAttribution is related to work who make use of Taylor decompositions to explain deep models. PatternAttribution reveals a good root point for the decomposition, the authors explain.

6 Conclusion

We successfully transferred a new explanation method to the NLP domain. We were able to demonstrate that PatternAttribution can be used to identify meaningful signal contributions in text inputs. Our method should be extended to other popular models in NLP. Furthermore, we introduced an improved quality criterion for signal estimators. In the future, estimators can be deduced from and tested against our new criterion.