Towards a Better Understanding of Gradient-Based Explanatory Methods in NLP

Qingfeng Du
School of Software Engineering
Tongji University
Shanghai, China
du_cloud@tongji.edu.cn

Jincheng Xu
School of Software Engineering
Tongji University
Shanghai, China
xujincheng@tongji.edu.cn

***Abstract*—To grasp what makes the deep learning models arrive at a particular prediction, gradient-based explanatory methods have been widely used in Natural Language Processing (NLP) recently. While the saliency maps of images can be computed directly in the pixel-level input space, the continuous gradient vector for words has to be reduced to a single value to indicate the word-level importance, and existing methods such as Sensitivity Analysis (SA) and Gradient $\times$ Input (GI) are either tricky or short of a deep investigation. In this paper, we review the family of gradient-based explanatory methods and discuss their practical implications. Specially, we propose the signed version of GI, namely *SignedGI*, while some previous work may have misunderstandings on its signedness. We also show the weakness of SA-based methods. We conduct extensive experiments to evaluate these explanatory methods both qualitatively and quantitatively.**

***Index Terms*—Gradient-based Explanatory Methods; Sensitivity Analysis; Gradient $\times$ Input; Text Classification**

I. Introduction

In the era of Artificial Intelligence (AI), deep learning models have been widely deployed in a variety of applications in Natural Language Processing (NLP), but are often criticized for the inability to explain their decisions. To afford transparency on the nested non-linear structure of the black box and shed light on interpretable AI models, a plethora of explanatory methods have been developed in literature [1] nowadays. Among existing work, gradient-based methods [2][3][4] have been gaining the spotlight recently because they can be easily used in any off-the-shelf neural networks.

It is straightforward to compute the pixel-level gradient in images [5][6], indicating how much the pixel contributes to the final prediction. However, things are different in NLP. Words are usually embedded in a continuous space, and a scalar value rather than a vector of gradients has to be derived for the word-level importance score. Consequently, many variations of gradient-based explanatory methods have been proposed in NLP to compute the scalar value, such as the sum of gradients in raw values [7], the L_1 norm [3], the L_2 norm [8][9][10], or the dot product between the vector of gradients and the word embedding itself [2][11][4]. For brevity, we refer to the first three variations as SA-based (Sensitivity Analysis [6]) methods, and the last one as GI (Gradient $\times$ Input [4]).

Even though we have a sophisticated theory for gradient-based explanatory methods in images, our current understanding on them in NLP is still rudimentary. On the one hand, the SA-based methods are very tricky. The gradient measures the local effect of a particular dimension in the word vector and it does not hold water to sum up the effects along the vector as the word-level importance score, since the true changes in the embedding space are discrete rather than continuous when a word is removed or replaced. On the other hand, the correctness of GI has not been strictly proved yet, especially its *signedness* (A word-level importance score is *signed* if it can distinguish between positive and negative impacts).

In this paper, we shed light on the aforementioned gradient-based explanatory methods. We propose the signed version of GI, namely *SignedGI*, based on the chain rule and the back-propagation algorithm [12]. The SignedGI score is the *opposite* of the dot product, whereas some previous work [2][4] neglect the signedness. Besides, we show the weakness of SA-based method. We conduct extensive experiments to evaluate these explanatory methods both qualitatively and quantitatively.

II. Related Works

There has been a remarkable series of work for explainable artificial intelligence in NLP [1]. In [13], Leave-One-Out (LOO) estimates the word importance by observing the change of the log-likelihood when a particular word is removed. It has been widely used as a black-box explanatory method in NLP [7][4]. In the white-box settings, gradient-based explanatory methods have attracted great interest. As described previously, we mainly consider SA-based methods [3] and GI [2] in this paper. They aim to compute the gradient w.r.t. the word to indicate the word-level importance score. There is another popular gradient-based explanatory method named *Integrated Gradients* [14], which integrates over all gradients on a linear interpolation between the original input and the baseline input masked with zeros. However, the L_p norm in SA-based methods or the dot product in GI are still a prerequisite for the use of *integrated gradients* in NLP [15][16], so we exclude it from the scope of our work.

This work was supported by National Key R&D Program of China (Grant No. 2020YFB2103300)
DOI reference number: 10.18293/SEKE2021-045

The weakness of SA-based methods (in the L_2 norm) has already been noticed in the experiments of some previous work [8][15]. They attribute this observation to the fact that the L_2 norm can only measure the word importance with the inability to distinguish between positive and negative impacts. We will address the deeper cause in the section below, that the overall impact on the loss is uncertain when we mask the word identified by SA-based methods with all-zero paddings. Furthermore, we will display a interesting counterexample model to show the weakness of SA-based methods in the experiments, where the words of the same frequency can have the same gradient vector in the embedding space, albeit with different contributions to the final prediction.

III. Methods

Let $x = (w_1, w_2, \cdots, w_m)$ be the document consisting of m words, where $w = (e_1, e_2, \cdots, e_n)$ is the continuous word representation in the n-dimensional embedding space. Let $l_y : x \to \mathbb{R}^1$ be the loss function w.r.t. the legitimate label y.

A. Sensitivity Analysis

Sensitivity analysis has been a popular method for interpreting non-linear neural networks in images [5][6], where the sensitivity of a particular pixel p for the color channel c can be computed as follows:

$$s_{p,c} = \left(\frac{\partial}{\partial p, c} l_y(x)\right)_{c \in (r,g,b)} \tag{1}$$

Recently this method has been extended to the domain of NLP, and the sensitivity of a particular dimension e in the embedding space can be represented as follows [3]:

$$s_e = \frac{\partial}{\partial e} l_y(x) \tag{2}$$

where the score s_e tells us how much the change in one specific dimension e would exert an influence on the results.

However, words are embedded in the continuous space with more than one dimension. With s_e for each dimension, the L_q norm operation is usually performed to transform the vector into the word-level importance score:

$$s_w = \|(s_{e_1}, s_{e_2}, \cdots, s_{e_n})\|_{L_q} \tag{3}$$

where the norm usually takes the value of $q = 1$ [3][17] or $q = 2$ [8][9][4][10]. We denote the two variations as $|SA|_1$ and $|SA|_2$. As far as we know, there is no current work in NLP use $q = \infty$. However, $q = \infty$ is a common practice in images [5], and we decide to consider $|SA|_\infty$ in our experiments. Apart from the norm, some work directly use the raw value of the gradient [7]. We denote it as $|SA|_{raw}$.

B. Gradient × Input

Gradient × Input (GI) computes the dot product of the word embedding and the gradient of the output w.r.t the embedding itself. It is firstly proposed in [2], where the formal representation of GI is presented with the first-order Taylor expansion of the loss function. However, it only extracts the salient scores without distinguishing between the positive and negative impacts. In this section, we provide the theoretical augments to the deduction of GI from a new perspective, and specially analyze its signedness for SignedGI.

Suppose $E \in \mathbb{R}^{v*n}$ is the embedding layer, where v is the vocabulary size. Let $I_x \in \mathbb{R}^{m*v}$ be the matrix embedding the input document x, where each row $I_{w_i} = (0, \cdots, 1, \cdots, 0)$ is a v-dimensional one-hot vector for the input word w_i. Now, we are interested in the gradient of the value "1" in the one-hot vector, which indicates how much the existence of the word w_i locally affect the network output.

We show how to compute the word-level gradient in I_x now. For brevity, we assume that I_x is a document consisting of only one word (or $m = 1$), and the one-hot representation is $I_x = (t_1, \cdots, t_v) \in \mathbb{R}^{1*v}$ where $t_i = 0$ or 1. Assume WLOG that $t_1 = 1$ in I_x, then t_2 to t_v are all zeros. In other words, we assume that the only one word in I_x corresponds to the first word in the vocabulary. Then, the output of the embedding layer O_E in a neural network can be computed as follows, where $O_E = (e_1, \cdots, e_n) \in \mathbb{R}^{1*n}$ is the embedding for the only one word in I_x:

$$O_E = I_x \times E \tag{4}$$

Let $l_y : x \to \mathbb{R}^1$ be the loss function w.r.t. the legitimate label y. We compute the word-level gradient in I_x using the chain rule and the back-propagation algorithm [12]:

$$\begin{aligned}
\frac{\partial l_y}{\partial I_x} &= \frac{\partial l_y}{\partial O_E}\frac{\partial O_E}{\partial I_x} \\
&= (\frac{\partial l_y}{\partial e_1}, \cdots, \frac{\partial l_y}{\partial e_n})(\frac{\partial O_E}{\partial t_1}, \cdots \frac{\partial O_E}{\partial t_v}) \\
&= (\frac{\partial l_y}{\partial e_1}, \cdots, \frac{\partial l_y}{\partial e_n}) \begin{vmatrix} \frac{\partial e_1}{\partial t_1} & \cdots & \frac{\partial e_1}{\partial t_{v-1}} & \frac{\partial e_1}{\partial t_v} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{\partial e_n}{\partial t_1} & \cdots & \frac{\partial e_n}{\partial t_{v-1}} & \frac{\partial e_n}{\partial t_v} \end{vmatrix} \\
&= (\frac{\partial l_y}{\partial e_1}, \cdots, \frac{\partial l_y}{\partial e_n}) \begin{vmatrix} e_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ e_n & \cdots & 0 & 0 \end{vmatrix} \\
&= (\sum_{i=1}^{n} \frac{\partial l_y}{\partial e_i} * e_i, \cdots, 0, 0) \in \mathbb{R}^{1*n}
\end{aligned} \tag{5}$$

The word-level gradient of t_1 in I_x is exactly the dot product between the word embedding and the gradient w.r.t. the embedding itself. The proof here can be easily extended to the case where I_x consists of more than one word.

Let the gradient of the word w be g_w. The SignedGI scores are defined as the *opposite* of g_w, or more formally:

$$s_w = -g_w = -\sum_{i=1}^{n} \frac{\partial l_y}{\partial e_i} * e_i \tag{6}$$

Now we explain the reasons. According to the algorithm of gradient descent, the parameters should move in the direction of steepest descent as defined by the negative of the gradient to minimize the loss. Here, the parameter t_1 can only take the value of 1 or 0, indicating the existence or non-existence of the word w respectively. Hence, in the case of $g_w > 0$, the change of t_1 from 1 to 0 follows the gradient descent direction

and thus decreases the loss l_x. In other words, the removal of the word w has a positive impact on the performance, or, *the existence of the word w has a negative impact on the performance when $g_w > 0$*. Similarly, *the existence of the word w has a positive impact on the performance when $g_w < 0$*. By convention, the word-level importance score should be positive if the word contributes to the current prediction, so we arrive at the final representation for SignedGI as in Equation 6. Note that the signedness of GI has been neglected in its original publication [2], and been misunderstood in the subsequent work [4].

We can also analyze the SA-based methods in a similar way. Let us take $|SA|_2$ as an example. The removal of a word identified by $|SA|_2$ can be interpreted as masking its word embedding with all-zero paddings. In the original embedding space, the word is represented as a n-dimensional vector, and the values in various dimensions can be larger than 0 or smaller than 0. The mask of 0 will make them either follow or go against the gradient descent direction. As a result, the overall impact on the loss becomes uncertain, which leads to the weakness of SA-based methods. In summary, SA-based methods compute the gradients in the embedding level and the sum of them to indicate the word importance is inaccurate. In contrast, SignedGI computes the gradient in the word level directly, so it should be more faithful than SA-based methods.

IV. Experiments

A. Preliminaries

Datasets We use two publicly available text classification datasets: (1) *AG's News*: A topic classification dataset consisting of four categories, including World, Sports, Business, and Sci/Tech. (2) *Internet Movie Database (IMDB)*: A binary sentiment analysis dataset on movie reviews.

Models We consider three popular text classification models, including a linear classifier *FastText* [18], a convolutional neural network *TextCNN* [19] and a bi-directional recurrent neural network *BiLSTM*.

Baselines Apart from the gradient-based explanatory methods as mentioned previously, we introduce two more baselines, namely *Random (RD)* and *Leave-One-Out (LOO)* [13]. The first baseline simply generates a random permutation of words to simulate the decreasing order of word importance. It can be considered as a very uninformative approach. The second baseline estimates the importance scores by erasing each word from the input and tracking the effect. The variations of LOO can be found in [13][20]. In our implementation, we compute the difference in loss:

$$s_w = l_y(x) - l_y(x_{|w=0}) \tag{7}$$

where $l_y(x)$ is the original loss and $l_y(x_{|w=0})$ is the loss when masking the word embedding of w with all-zero paddings. LOO is very similar to the perturbation experiment itself (which will be introduced later). Similar baselines have also been set up in [7][4]. With the *possible* upper bound and lower bound on the explanatory ability, we can show the results of gradient-based methods in a more intuitive way. The word "possible" means LOO may not produce the best explanatory ability among existing methods, but it is faithful enough. Taking it as an upper bound is helpful for us to see the difference between the results of gradient-based methods and faithful explanations. So it is with RD.

Metrics In order to evaluate the explanatory ability of different methods, existing work usually perform the perturbation-based experiment [7][4], which perturbs the original input in a word level (e.g., the mask of zero paddings, or the deletion operation), and subsequently measures the changes on the performance (e.g., the changes on accuracies, probabilities, or losses). The word importance increases monotonically with the change. Based on this observation, an objective quality measure, *AOPC*, is proposed in [6] to evaluate ordered collections of features quantitatively. While originally designed for images, AOPC can be easily extended to NLP [7]:

$$AOPC = \frac{1}{K+1}\left\langle \sum_{k=0}^{K} f_y(x^{(0)}) - f_y(x^{(k)}) \right\rangle_{avg} \tag{8}$$

where $x^{(k)}$ is the perturbed input with the most important k words masked with zero paddings, $f_y(x)$ is the probability of the legitimate label y, K is the cut-off point w.r.t. the top-K important words, and $\langle\cdot\rangle_{avg}$ represents the average over all the documents. The perturbation of the most important words implies a steep decreases of $f_y(x)$, so the method with the better explanatory ability has a larger AOPC.

In fact, AOPC values measure the absolute word importance. We can also describe the word importance in a relative way. We take LOO as a well-established benchmark because it provides the possible upper bound on the explanatory ability, and we report the *Pearson correlation coefficient* between the results of LOO and gradient-based methods.

$$\rho = \frac{\sum_{i=1}^{m}(a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{m}(a_i - \overline{a})^2}\sqrt{\sum_{i=1}^{m}(b_i - \overline{b})^2}} \tag{9}$$

where $a = (a_1, a_2, \cdots, a_m)$ and $b = (b_1, b_2, \cdots, b_m)$ are the score vectors. $\overline{a}$ and $\overline{b}$ denote the average operation. If ρ is close to 1, there is a strong positive linear association between a and b, indicating that the estimated word importance in b is as faithful as a and vice versa. On the contrary, if ρ is close to -1, the relationship is strongly negative.

Others When computing the word importance with explanatory methods, existing work use either the *predicted* class [7] or the *legitimate* class [4] as the target class. Since we already have the ground-truth labels, we use the latter approach.

B. Quantitative Comparison of Explanatory Methods

We quantitatively compare the explanatory methods and compute their AOPC values varying the cut-off point K from 0 to the maximum document length. Fig. 1 illustrates the results.

Generally speaking, all the SA-based methods suffer performance decline definitely compared to SignedGI. The curves of $|SA|_1$, $|SA|_2$ and $|SA|_\infty$ are almost overlapped and indicate the similar explanatory abilities. This is mainly because the

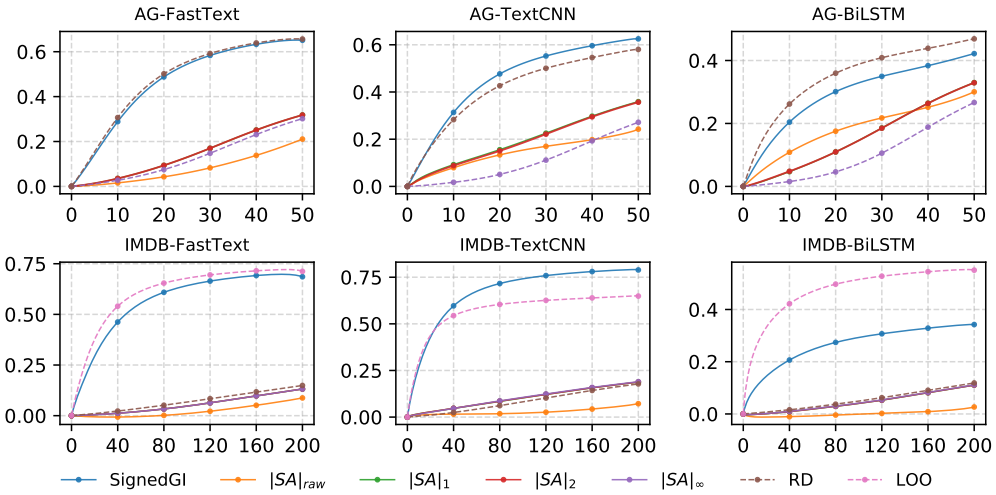


Fig. 1. Comparison of the considered explanatory methods in terms of AOPC values. The horizontal axis represents the cut-off point K, and the vertical axis represents the AOPC value. For each K, a larger AOPC value indicates the better explanatory ability of the top-K important words.

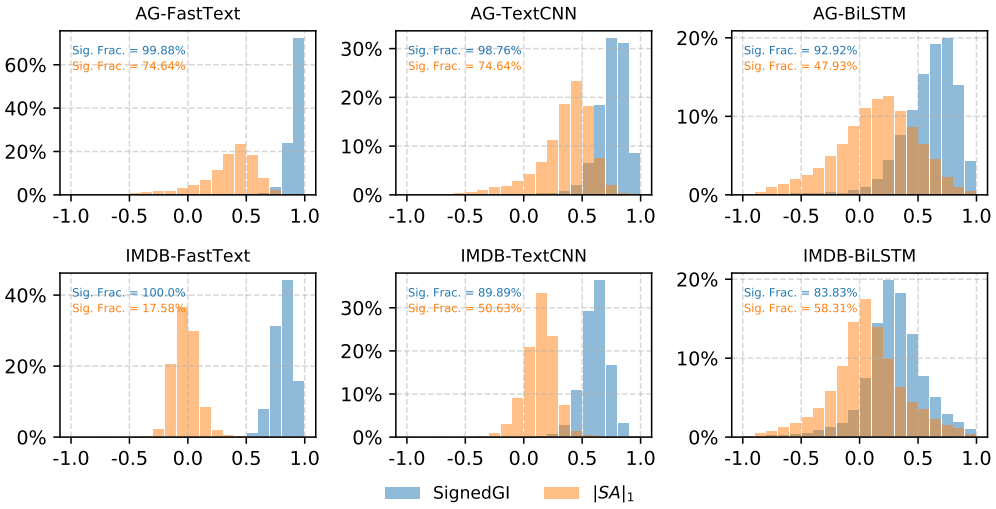
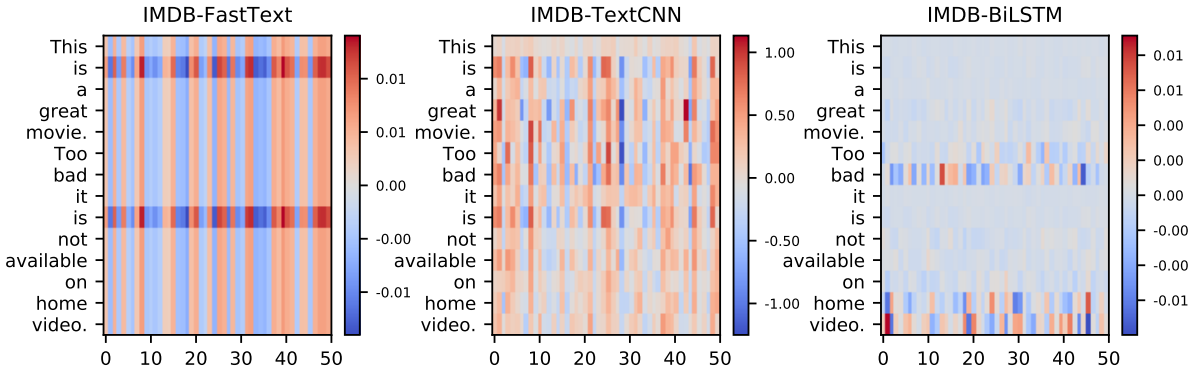


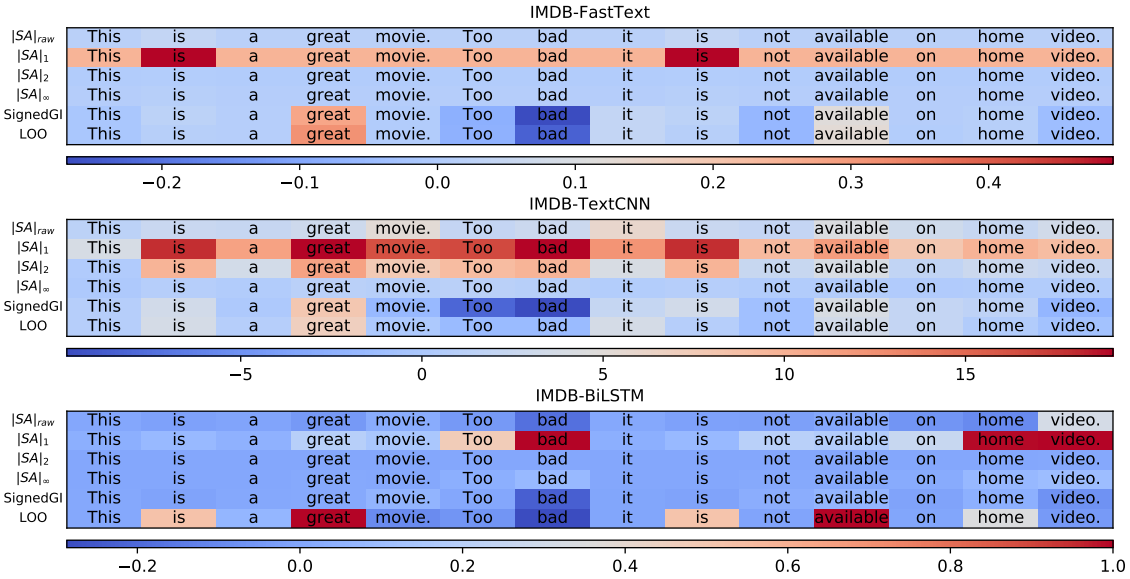
Fig. 2. Histogram of Pearson correlation coefficient ρ between LOO and gradient-based methods. The horizontal axis represents ρ and the vertical axis represents the distribution. $|SA|_1$ is selected to represent SA-based methods since the results of others are similar. The fraction of instances whose correlation is statistically significant (p-value $<= 0.05$) has been reported in the top-left corner. Note that the p-value here largely depends on the document length, so the correlation is prone to be weak.

highest gradient magnitude in a particular dimension that decides the value of $|SA|_\infty$ usually contributes most to the value of $|SA|_1$ and $|SA|_2$. These three methods outperform the RD baseline by a moderate margin on AG-TextCNN and AG-BiLTSM, but only tie RD in the other cases. It seems that their performance are heavily affected by the model or the dataset. $|SA|_{raw}$ performs worse than other SA-bsed methods. Sometimes even the RD baseline can beat $|SA|_{raw}$.

LOO and SignedGI always demonstrate the larger AOPC values. In other words, they better identify the important words. Even though we assume that LOO provides the upper bound on the explanatory ability in our perturbation-based experiments, SignedGI outperforms LOO in TextCNN interestingly. We attribute this observation to the weakness of LOO, that it computes the contribution of words independently without considering their mutual effects. Hence, it might result in a sub-optimal explanation [7]. In most cases, the AOPC values of SignedGI and LOO are generally comparable, except for IMDB-BiLSTM where LOO surpasses SignedGI by a large margin. A possible reason is that the long document length of IMDB causes the vanishing gradient problem in the recurrent structure, which exerts a negative impact on the SignedGI performance. Note that a deep investigation into the performance difference between LOO and SignedGI is not the



(a) Visualization of gradient information. Each row represents the continuous word representation, where each cell is the gradient of a particular dimension in the embedding space.



(b) Visualization of instance-wise explanations. Each row displays the word-level importance scores for the instance, and the used explanatory method is indicated on the left.

Fig. 3. Qualitative comparison of explanatory methods. The instance is selected from the test set of IMDB and the target class is "positive". The red color indicates a positive score, and the blue color indicates a negative score. The maximum value of IMDB-BiLSTM has been reduced from 5.0 to 1.0 for better visualizations, otherwise the negative color bar will be overwhelmed.

scope of our paper, and the important thing is that SignedGI can achieve a good explanatory ability

C. Correlation Between LOO and Gradient-based Methods

The full distributions of the Pearson correlation coefficient ρ have been illustrated in Fig. 2. The general observation from the figures is that SignedGI does tend to have a strong positive association with LOO, and a statistically significant correlation can be consistently established. The association on IMDB-BiLSTM seems to be less strong, which is consistent with the results in Fig. 1 that the difference between SignedGI and LOO is a bit pronounced on on IMDB-BiLSTM. On the other hand, the centrality of densities for $|SA|_1$ lies in the range of 0.2$\sim$0.5 on the dataset of AG, which shows a very weak positive association with LOO. On the dataset of IMDB, the centrality hovers around 0.0, indicating almost no association. The results here further support the fact than SignedGI shows better explanatory ability compared to SA-based methods.

D. Qualitative Comparison of Explanatory Methods

In this section, we illustrate the gradient information in the embedding space in Fig. 3 (a), and we visualize the instance-wise explanations in Fig. 3 (b).

Now we take a look at the results of IMDB-FastText in Fig. 3 (a). Interestingly, the gradients keep the same in each dimension, except for the word "is", whose gradient values are exactly twice as much as other words. In fact, in the architecture of FastText, the word embeddings are averaged into an internal representation, followed by the output layer directly. As a result, the gradient value of a specific dimension

in the embedding space is always proportional to the word frequency in the input document. As we can see, *FastText is a pretty compelling counterexample to the effectiveness of SA-based methods.* The words of the same frequency will always be assigned with the same SA-based scores, but they are embedded in various continuous representations and go through the same linear layer, meaning that their contributions to the final prediction are different in reality.

In Fig. 3 (a), the results does not have a clear focus in the heatmap of IMDB-TextCNN. In IMDB-BiLSTM, the words "bad", "home" and "video" stands out, but the target label is "positive" and the model attaches almost zero emphasis on the positive sentiment word "great". Note that the explanatory methods only reflect the model's own "view" on the model prediction rather than human reasoning, so it is possible that "great" does not play an important role in the binary sentiment analysis task. However, as we will show later, "great" does have a strong positive impact in this model, and the heatmap here indeed fails to capture the relevant information.

In Fig. 3 (b), let us focus on the results of LOO firstly. It can be clearly seen that "great" has a large positive score across all three models, meaning that "great" contributes a lot to the target label. On the contrary, "bad" is always assigned with a large negative score, indicating that it has a negative impact to the current prediction. The results agree with human observations, that "great" has a positive impact and "bad" has a negative impact on the prediction of positive sentiments. Not surprisingly, the results of SignedGI are very close to LOO, except for the case on BiLSTM where SignedGI misses the word "great", which also agrees with previous results that SignedGI works less well on BiLSTM. Nevertheless, SignedGI still filters out the important words in the qualitative experiments and provides the reasonable signed explanations. On the other hand, the results of SA-based methods are less focused. Sometimes they cannot select the important words correctly, or cannot distinguish between negative and positive impacts.

V. Conclusion

Gradient-based explanatory methods have been widely used in NLP nowadays. In this paper, we review existing methods and discuss their practical implications. We propose the signed version of GI, namely *SignedGI*, and show the weakness of SA-based methods. We conduct comprehensive experiments to evaluate different methods, and the empirical results demonstrate that SignedGI significantly outperforms SA-based methods in explanatory ability. We hope our work helps researchers to obtain the more accurate instance-wise explanations via gradient-based explanatory methods in NLP.

References

[1] A. Alishahi, G. Chrupala, and T. Linzen, "Analyzing and interpreting neural networks for NLP: A report on the first blackboxnlp workshop," *CoRR*, vol. abs/1904.04063, 2019.

[2] M. Denil, A. Demiraj, and N. de Freitas, "Extraction of salient sentences from labelled documents," *CoRR*, vol. abs/1412.6815, 2014.

[3] J. Li, X. Chen, E. H. Hovy, and D. Jurafsky, "Visualizing and understanding neural models in NLP," in *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016*, 2016, pp. 681–691.

[4] L. Arras, A. Osman, K. Müller, and W. Samek, "Evaluating recurrent neural network explanations," *CoRR*, vol. abs/1904.11829, 2019.

[5] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[6] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, 2017.

[7] D. Nguyen, "Comparing automatic and human evaluation of local explanations for text classification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, 2018, pp. 1069–1078.

[8] L. Arras, F. Horn, G. Montavon, K. Müller, and W. Samek, "Explaining predictions of non-linear classifiers in NLP," in *Proceedings of the 1st Workshop on Representation Learning for NLP, Rep4NLP@ACL 2016*, 2016, pp. 1–7.

[9] L. Arras, G. Montavon, K. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, 2017, pp. 159–168.

[10] E. Wallace, J. Tuyls, J. Wang, S. Subramanian, M. Gardner, and S. Singh, "Allennlp interpret: A framework for explaining predictions of NLP models," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, 2019, pp. 7–12.

[11] S. Feng, E. Wallace, A. G. II, M. Iyyer, P. Rodriguez, and J. L. Boyd-Graber, "Pathologies of neural models make interpretation difficult," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018*, 2018, pp. 3719–3728.

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 696–699, 1988.

[13] J. Li, W. Monroe, and D. Jurafsky, "Understanding neural networks through representation erasure," *CoRR*, vol. abs/1612.08220, 2016.

[14] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, 2017, pp. 3319–3328.

[15] N. Pörner, H. Schütze, and B. Roth, "Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, 2018, pp. 340–350.

[16] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, 2019, pp. 3543–3556.

[17] H. Liu, Y. Zhang, Y. Wang, Z. Lin, and Y. Chen, "Joint character-level word embedding and adversarial stability training to defend adversarial text," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 8384–8391.

[18] E. Grave, T. Mikolov, A. Joulin, and P. Bojanowski, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 427–431.

[19] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1746–1751.

[20] Á. Kádár, G. Chrupala, and A. Alishahi, "Representation of linguistic form and function in recurrent neural networks," *Computational Linguistics*, vol. 43, no. 4, 2017.