Cascading Explaining - A universal method to improve NLP model

Guo Zikunr

2021225559

Department of Artificial Intelligence, Kyungpook national University, Republic of Korea

gzk798412226@gmail.com

**Abstract—Recently some interpretable methods have been proposed, among which the interpretation method for question answering systems: Integrated Gradients. It is a more popular method. We propose a novel method to improve the performance of the model based on the characteristics of the Integrated Gradients applied in the question answering system: Cascading Explaining. This method is to extract information from the interpretation of the model output, and then let the model predict again. We have used some experiments to prove that this method has successfully passed the sanity check, and has greatly improved the accuracy and probability of model prediction.** **At the same time, this is a general form method: it is not limited by models and interpretation methods. In other words, the performance of any question answering system can be improved by this method.**

**Keywords— Explainable ai, Question Answering System, Integrated Gradients, Sanity Check**

# INTRODUCTION

As machine learning grows in complexity and impact, much hope rests on explanation methods as tools to elucidate important aspects of learned models. Explanations could potentially help satisfy regulatory requirements, help practitioners debug their model, and perhaps, reveal bias or other unintended effects learned by a model. Integrated Gradients is an increasingly popular tool designed to highlight relevant features in the input. Although this method is compared with other methods such as LIME and SHAP, it shows its uniqueness and innovation. But still explaining the precious hard work of the machine learning model faces a methodological challenge: the difficulty of evaluating the scope and quality of the model interpretation. The lack of principled guidelines confuses practitioners when making decisions. Understanding the input and output behavior of a deep network gives us the ability to improve it. This intelligibility is for all computer programs, including machine learning models. Attribution has other applications. They can be used in machine learning-driven products to provide recommended rationale. For example, a deep network based on imaging can help inform the doctor about the part of the image that led to the recommendation. This can help doctors understand and compensate for a person's strengths and weaknesses model. Developers can also use attribution in an exploratory sense. For example, we can use deep networks to extract insights that can be used in rule-based systems.

Our contributions

* We propose a specific, easy-to-implement test to improve model performance : Cascading Explaining Universal method with no limitation by models and interpretation methods
* We show that the interpretation of the bert-based question answering system is not independent of the model. Therefore, to explain this type of model is meaningful and easy to understand

# METHOD

A. Cascading Explaining

The basic goal of Cascading Explaining is well understood. This method is inspired by the attention mechanism of the human brain: when we are doing reading comprehension, we will first focus on the entire article and read it through, and then make a choice based on the goal in our mind: which paragraph may hide what we need Answer. So we read again in the new natural paragraph until we find the answer we hope.

Cascading Explaining is also same attention mechanism of the human brain. Given an input to the model, the contribution of different elements to the model output is calculated through the gradient integration algorithm. Then the high contribution value often hides the answer. Delete or replace the unimportant with meaningless annotations such as delimiters or paragraphs that have no effect on the model, and then re-input this processed sentence into the model. The model will Output the same answer, but the probability of a ground truth has been increased.

In order to ensure that the correct answer is not deleted or replaced. Cascade interpretation was proposed. In the process of processing model interpretation, there may be wrong answers, but because the interpretation is independent of the input, after multiple interpretations of the first range, we make changes to the later based on the interpretation of the initial model and the initial input After the data are classified and compared. Normally, the correct answer is usually accompanied by a possibility. The higher the possibility, the more certain the model is. When the model gives a low probability answer, we need to wonder whether the model really understands the current data. This is the essence of Cascading Explaining. It is precisely because of multiple local adjustments that the misinterpretation caused by mispredictions is reduced.

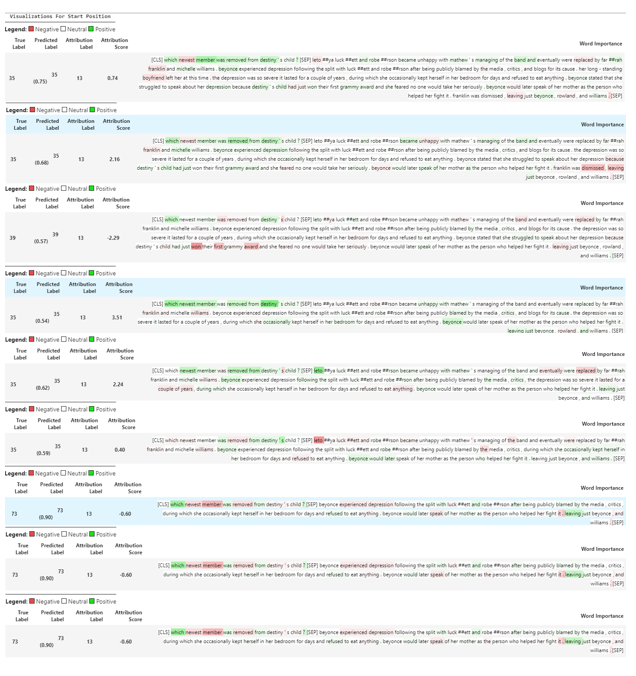


Fig. 1 Example of an Cascading Explaining, This example only provides the first 9 treatments, After deleting the tokenziers with low contribution to the result, different interpretations are obtained, and then iteratively processed, and finally the initial input text is extracted into a short sentence. This sentence is the sentence containing the answer. According to Figure 2, it can also be easily found The shortest model will provide higher possibilities.

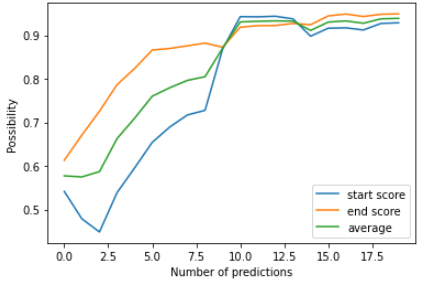


Fig. 2 The X-axis corresponds to the number of loop predictions, and the y-axis uses the probability of the answer. This curve clearly shows that the processed input will get a higher probability.

B. Sanity Check

In our formal setup, an input is a vector x ∈ R d . A model describes a function S : R d → R C , where C is the number of classes in the classification problem. An explanation method provides an explanation map E : R d → R d that maps inputs to objects of the same shape. We now briefly describe some of the explanation methods we examine. The supplementary materials contain an in-depth overview of these methods. Our goal is not to exhaustively evaluate all prior explanation methods, but rather to highlight how our methods apply to several cases of interest. The gradient explanation for an input x is Egrad(x) = ∂S ∂x [22, 23, 8]. The gradient quantifies how much a change in each input dimension would a change the predictions S(x) in a small neighborhood around the input.

The data randomization test compares a given saliency method applied to a model trained on a labeled data set with the method applied to the same model architecture but trained on a copy of the data set in which we randomly permuted all labels. If a saliency method depends on the labeling of the data, we should again expect its outputs to differ significantly in the two cases. An insensitivity to the permuted labels, however, reveals that the method does not depend on the relationship between instances (e.g. images) and labels that exists in the original data. Speaking more broadly, any explanation method admits a set of invariances, i.e., transformations of data and model that do not change the output of the method. If we discover an invariance that is incompatible with the requirements of the task at hand, we can safely reject the method. As such, our tests can be thought of as sanity checks to perform before deploying a method in practice.

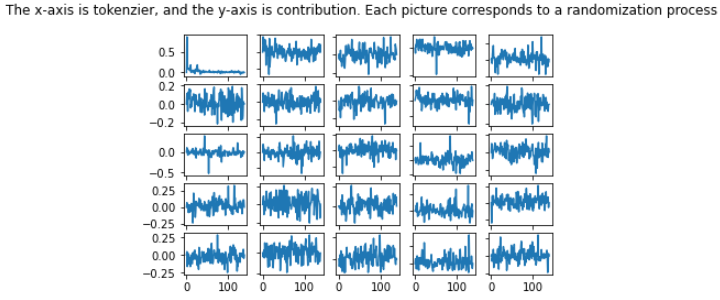


Fig. 3 Cascading Randomization randomize the weights of a model starting from the top layer, successively, all the way to the bottom layer. Then plot the contribution value of the obtained tokenizer.

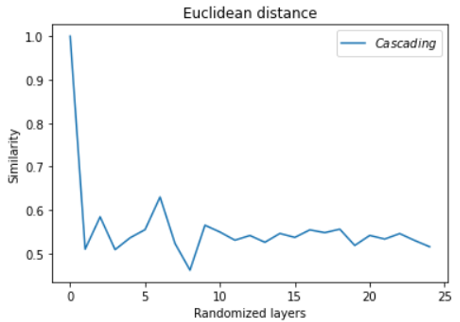


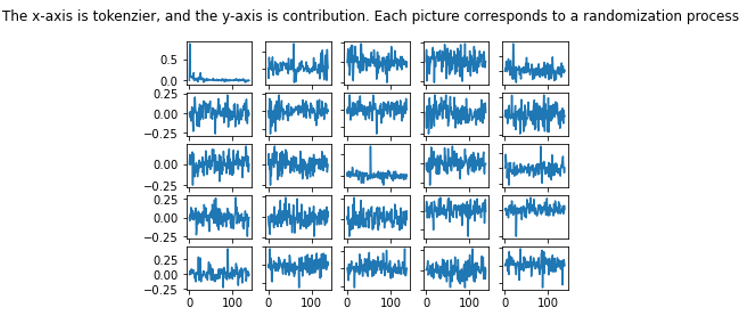
Fig. 4 Calculate different random layers and calculate the contribution of each layer's tokenzer, use Euclidean distance to calculate the similarity of the contribution.

Fig. 5 Independent Randomization the conduct an independent layer-by-layer randomization with the goal of isolating the dependence of the explanations by layer.

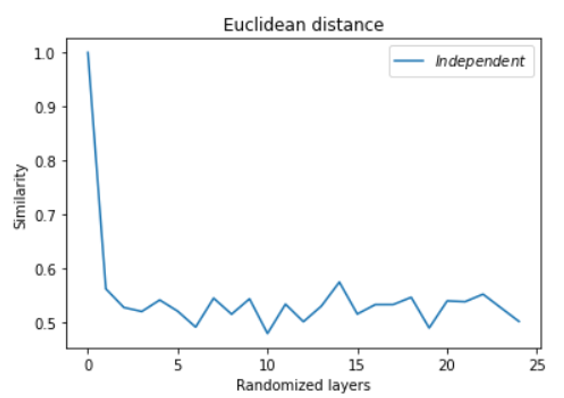


Fig. 6 Independent Random layers and calculate the contribution of each layer's tokenzer, use Euclidean distance to calculate the similarity of the contribution.

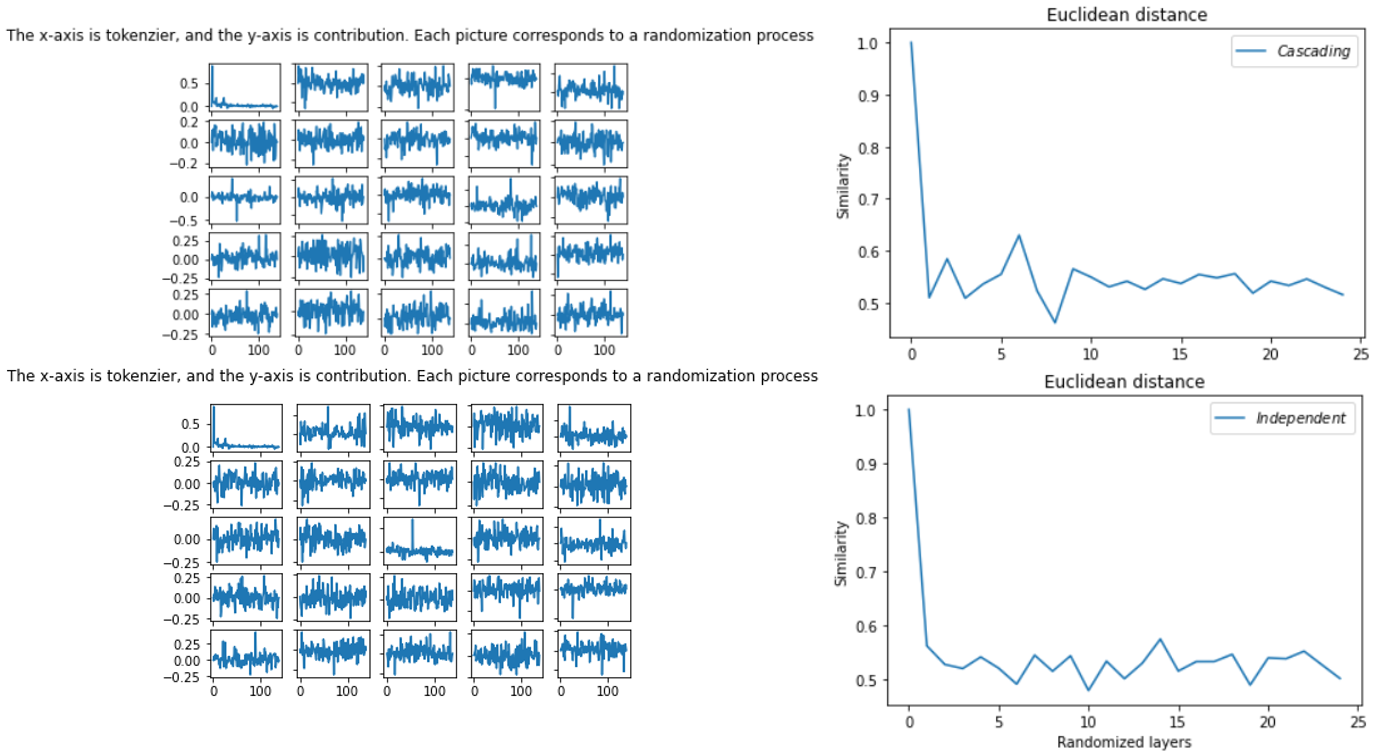


Fig. 3 Cascading Randomization & Independent Randomization for each layer, we fix the weights of other layers to their original values, and randomize one layer at a time.

Actionable methodology to evaluate what kinds of explanations a given method can and cannot provide. Reliance, solely, on assessment can be misleading. some existing saliency methods are independent both of the model and of the data generating process. Consequently, methods that fail the proposed tests are inadequate for tasks that are sensitive to either data or model, such as, finding outliers in the data, explaining the relationship between inputs and outputs that the model learned, and debugging the model.

Your paper must use a page size corresponding to A4 which is 210mm (8.27") wide and 297mm (11.69") long. The margins must be set as follows:

* Top = 19mm (0.75")
* Bottom = 43mm (1.69")
* Left = Right = 14.32mm (0.56")

Your paper must be in two column format with a space of

4.22mm (0.17") between columns.

# RELATED WORK

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

1. Text Font of Entire Document

The entire document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes.

Recommended font sizes are shown in Table 1.

1. Title and Author Details

Title must be in 24 pt Regular font. Author name must be in 11 pt Regular font. Author affiliation must be in 10 pt Italic.

Email address must be in 9 pt Courier Regular font.

# TABLE I

FONT SIZES FOR PAPERS

|  |  |  |  |
| --- | --- | --- | --- |
| Font Size | Appearance (in Time New Roman or Times) | | |
| Regular | Bold | Italic |
| 8 | table caption (in Small Caps), figure caption, reference item |  | reference item (partial) |
| 9 | author email address  (in Courier), cell in a table | abstract body | abstract heading (also in Bold) |
| 10 | level-1 heading (in Small Caps), paragraph |  | level-2 heading, level-3 heading, author affiliation |
| 11 | author name |  |  |
| 24 | title |  |  |

All title and author details must be in single-column format and must be centered.

Every word in a title must be capitalized except for short minor words such as “a”, “an”, “and”, “as”, “at”, “by”, “for”, “from”, “if”, “in”, “into”, “on”, “or”, “of”, “the”, “to”, “with”.

Author details must not show any professional title (e.g. Managing Director), any academic title (e.g. Dr.) or any membership of any professional organization (e.g. Senior Member IEEE).

To avoid confusion, the family name must be written as the last part of each author name (e.g. John A.K. Smith).

Each affiliation must include, at the very least, the name of the company and the name of the country where the author is based (e.g. Causal Productions Pty Ltd, Australia).

Email address is compulsory for the corresponding author.

1. Section Headings

No more than 3 levels of headings should be used. All headings must be in 10pt font. Every word in a heading must be capitalized except for short minor words as listed in Section III-B.

* 1. Level-1 Heading: A level-1 heading must be in Small Caps, centered and numbered using uppercase Roman numerals. For example, see heading “III. Page Style” of this document. The two level-1 headings which must not be numbered are “Acknowledgment” and “References”.
  2. Level-2 Heading: A level-2 heading must be in Italic, left-justified and numbered using an uppercase alphabetic letter followed by a period. For example, see heading “C. Section Headings” above.
  3. Level-3 Heading: A level-3 heading must be indented, in Italic and numbered with an Arabic numeral followed by a right parenthesis. The level-3 heading must end with a colon. The body of the level-3 section immediately follows the level3 heading in the same paragraph. For example, this paragraph begins with a level-3 heading.

1. Figures and Tables

Figures and tables must be centered in the column. Large figures and tables may span across both columns. Any table or figure that takes up more than 1 column width must be positioned either at the top or at the bottom of the page.

Graphics may be full color. All colors will be retained on the CDROM. Graphics must not use stipple fill patterns because they may not be reproduced properly. Please use only SOLID FILL colors which contrast well both on screen and on a black-and-white hardcopy, as shown in Fig. 1.

Fig. 1 A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy

Fig. 2 shows an example of a low-resolution image which would not be acceptable, whereas Fig. 3 shows an example of an image with adequate resolution. Check that the resolution is adequate to reveal the important detail in the figure.

Please check all figures in your paper both on screen and on a black-and-white hardcopy. When you check your paper on a black-and-white hardcopy, please ensure that:

* the colors used in each figure contrast well,
* the image used in each figure is clear, • all text labels in each figure are legible.

1. Figure Captions

Figures must be numbered using Arabic numerals. Figure captions must be in 8 pt Regular font. Captions of a single line (e.g. Fig. 2) must be centered whereas multi-line captions must be justified (e.g. Fig. 1). Captions with figure numbers must be placed after their associated figures, as shown in Fig. 1.



Fig. 2 Example of an unacceptable low-resolution image



Fig. 3 Example of an image with acceptable resolution

1. Table Captions

Tables must be numbered using uppercase Roman numerals. Table captions must be centred and in 8 pt Regular font with Small Caps. Every word in a table caption must be capitalized except for short minor words as listed in Section III-B. Captions with table numbers must be placed before their associated tables, as shown in Table 1.

1. Page Numbers, Headers and Footers Page numbers, headers and footers must not be used.
2. Links and Bookmarks

All hypertext links and section bookmarks will be removed from papers during the processing of papers for publication. If you need to refer to an Internet email address or URL in your paper, you must type out the address or URL fully in Regular font.

1. References

The heading of the References section must not be numbered. All reference items must be in 8 pt font. Please use Regular and Italic styles to distinguish different fields as shown in the References section. Number the reference items consecutively in square brackets (e.g. [1]).

When referring to a reference item, please simply use the reference number, as in [2]. Do not use “Ref. [3]” or “Reference [3]” except at the beginning of a sentence, e.g. “Reference [3] shows …”. Multiple references are each numbered with separate brackets (e.g. [2], [3], [4]–[6]).

Examples of reference items of different categories shown in the References section include:

* + example of a book in [1]
  + example of a book in a series in [2]
  + example of a journal article in [3]
  + example of a conference paper in [4]
  + example of a patent in [5]
  + example of a website in [6]
  + example of a web page in [7]
  + example of a databook as a manual in [8]
  + example of a datasheet in [9]
  + example of a master’s thesis in [10]
  + example of a technical report in [11]
  + example of a standard in [12]

# CONCLUSIONS

The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are self-contained. Causal Productions has used its best efforts to ensure that the templates have the same appearance.

Causal Productions permits the distribution and revision of these templates on the condition that Causal Productions is credited in the revised template as follows: “original version of this template was provided by courtesy of Causal Productions (www.causalproductions.com)”.

# ACKNOWLEDGMENT

The heading of the Acknowledgment section and the References section must not be numbered.

Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

# REFERENCES

1. S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
2. J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.
3. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.
4. M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, “High resolution fiber distributed measurements with coherent OFDR,” in Proc. ECOC’00, 2000, paper 11.3.4, p. 109.
5. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digitalto-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997.
6. (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
7. M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available:

http://www.ctan.org/texarchive/macros/latex/contrib/supported/IEEEtran/ [8] FLEXChip Signal Processor (MC68175/D), Motorola, 1996.

1. “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.
2. A. Karnik, “Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP,” M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
3. J. Padhye, V. Firoiu, and D. Towsley, “A stochastic model of TCP Reno congestion avoidance and control,” Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
4. Wireless LAN Medium Access Control (MAC) and Physical Layer

(PHY) Specification, IEEE Std. 802.11, 1997.