URSI GASS 2023, Sapporo, Japan, 19 – 26 August 2023

A model local interpretation routine for deep learning based radio galaxy classification

Hongming Tang*(1), Shiyu Yue(2), Zijun Wang(2), Jizhe Lai(3), Leyao Wei(2), Yan Luo(2), Chuni Liang(2), Jiani Chu(1), Dandan Xu(1)

(1) Department of Astronomy, Tsinghua University, Beijing 100084, China; e-mail: hongmingt@mail.tsinghua.edu.cn

(2) School of Physics and Astronomy, Sun Yat-sen University, 2 Daxue Road, Zhuhai 519082, China; e-mail: yueshy5@mail2.sysu.edu.cn

(3) School of Physics, Sun Yat-sen University, No. 135 Xingang Xi Road, Guangzhou 510275, P.R. China

Abstract

Radio galaxy morphological classification is one of the critical steps when producing source catalogues for large-scale radio continuum surveys. While many recent studies attempted to classify source radio morphology from survey image data using deep learning algorithms (i.e., Convolutional Neural Networks), they concentrated on model robustness most time. It is unclear whether a model similarly makes predictions as radio astronomers did. In this work, we used Local Interpretable Model-agnostic Explanation (LIME), an state-of-the-art eXplainable Artificial Intelligence (XAI) technique to explain model prediction behaviour and thus examine the hypothesis in a proof-of-concept manner. In what follows, we describe how **LIME** generally works and early results about how it helped explain predictions of a radio galaxy classification model using this technique.[1]

1 Introduction

Radio galaxy morphological classification is highly valued in radio astronomy as it reveals both the evolution process of a radio galaxy and how it interacted with the local environment [1]. Motivated by the rapidly growing radio source sample number produced by large-scale radio continuum surveys (i.e., [6, 7]), people started to face the data challenge using machine learning. In recent years, multiple deep learning algorithms have been developed to either find and classify radio galaxy morphology (i.e., [8]) or do classification alone (i.e.,[1, 2, 5]). Most of them have achieved human-comparable classification accuracy.

Besides the robust model performance these algorithms achieved, their model interpretability received less attention. Whether a deep learning algorithm is predicting radio galaxy morphology in the same way we radio astronomers did remains an ongoing question to answer. The latest effort to address this problem introduced a self-attention mechanism to their models, which enabled people to explain model prediction behaviour by looking at reasonably static image features from generated model attention maps [2]. However, when explaining many state-of-the-art radio galaxy classification algorithms, post-hoc model explanation methods that do not require model architecture manipulation would still be necessary.

Though radio galaxy classification system has become so complicated [10], Fanaroff and Riley binary classification (FR classification hereafter) system remains popular and used widely [3] since 1974. A radio galaxy would be identified as either edge-brightened sources (FR II) or edge-darkened sources (FR I) [2]. In this work, we tried to explain a deep learning model developed for FR classification using an eXplainable Aritificial Intelligence technique called Local Interpretable Model-agnostic Explanation (LIME)[2]. In order to perform model interpretation, we made used of FR-DEEP v2, a machine learning dataset for FR classification[3] to train a Convolutional Neural Network (CNN) based FR classification algorithm, reaching $\sim 91\%$ model general accuracy in a testset of 130 radio sources. Since we here focus on model interpretation, detailed model training and evaluation process would be shown in Tang and Yue et al. (2023, in prep.) instead. For the rest of this work, we would introduce the **Felzenszwalb** image segmentation method and how it combines with **LIME** in Section 2 and 3. Section 4 would shows our early results of how **LIME** help explaining model predictions, and we summarize our conclusion in Section 5.

2 Felzenszwalb

Felzenszwalb [4] is an graph-based image segmentation method. By seeing each image pixel as a **vertice**, neighbouring vertices are connected by **edges** along with weights measuring the dissimilarity of each vertice pair [4]. For any two neighbouring image segmented components C1 and C2, they can only stay independent to each other if their pairwise comparison **D(C1,C2)** satisfy:

$$\mathrm{D(C1,C2)} = \text{True if } \mathrm{Dif(C1,C2)} > \mathrm{Mint(C1,C2)} \quad (1)$$

[1] Hongming Tang and Shiyu Yue have equally contributed to this work.

[2] https://github.com/marcotcr/lime

[3] https://github.com/HongmingTang060313/FRDEEP_v2.0

where **Dif(C1,C2)** denotes the minimum weight edge connecting C1 and C2. **Mint(C1,C2)** represents the minimum internal difference (**int(C)**; the largest weight in the neighbouring spanning tree of a component C) considering both component C1 and C2:

$$\mathrm{Mint}(\mathrm{C1},\mathrm{C2}) = \mathrm{Min}(\mathrm{int}(\mathrm{C1}) + \frac{\mathrm{k}}{|\mathrm{C1}|}, \mathrm{int}(\mathrm{C2}) + \frac{\mathrm{k}}{|\mathrm{C2}|}) \quad (2)$$

where k is a constant to control the preference of having larger or smaller segments, and $|C1|$ corresponds to the size of component C1 (similar for $|C2|$). The two components will merge if D(C1, C2) equals *False*.

Compared with other segmentation methods, **Felzenszwalb** can generate image segments or **super-pixels** neither not "too coarse" nor "too fine", making it an appropriate tool when objects of interest in an image share modest sizes. In the next section, we shall address the connection between **Felzenszwalb** and **LIME**.

3 LIME

LIME was primarily developed to address the "trusting a prediction" problem, which is vital for decision-making [9]. This is achieved by providing individual model prediction explanations: In terms of image classification, presenting visual artefacts that qualitatively correlate typical patches (in our case, **super-pixel**) of an image and its model prediction [9].

Before talking about **LIME** mechanism under the image classification scheme, we firstly review the definition of the following variables/functions [9]:

- classifier f: a well-trained complex classification model (i.e., CNN)
- x ($x \in R^d$): original representations of an image instance awaited for model explanation.
- x' ($x' \in \{0,1\}^{d'}$): the binary vector for the interpretable representation of x, representing the "presence" (1) or "absence" (0) of image super-pixels formed via image segmentation (**Felzenszwalb** in this work).
- g ($g \in G$): an explanation as a model, where G is a family of interpretable models (i.e., linear models)
- $\Omega(g)$: model complexity of g (i.e., number of non-zero weights if g is a linear model).
- f(x): the probability that x belongs to a typical class.
- $\pi_x(z)$: a proximity measure between instance x and z.
- $L(f,g,\pi_x)$: a measure to quantify the ability of g to approximate f in the π_x defined locality. The smaller the L, the better g has performed.
- z' ($z' \in (0,1)^{d'}$): a perturbed sample containing a fraction of the non-zero elements in x'.

To ensure g is both interpretable to human (low $\Omega(g)$) and faithful (low $L(f,g,\pi_x)$), **LIME** produces its model explanation by [4]:

$$\xi(x) = \underset{g \in G}{argmin} \; L(f,g,\pi_x) + \Omega(g) \quad (3)$$

Since **LIME** aims to perform **model-agnostic** model explanation, it samples instances around x' by randomly drawing (hiding) non-zero elements of x' and gives z'. By recovering z' (weighted by $\pi_x(z)$) in the original representation z, one shall obtain f(z). f(z) then can be seen as a label for explanation model g. These perturbed samples and their corresponding labels could then be used to optimize Equation 3 and finally obtain $\xi(x)$, the model explanation for the instance original representation x. One can then know which super-pixel in an image has positively/negatively contributed to class prediction, and hence evaluate whether the model predicts as humans do qualitatively.

4 Application to Radio Galaxy Classification

Though **LIME** could be a useful model explanation method, there is a loose restriction of this technique: the user are often (not always) required to know what to expect before explaining a model. That is to say, a user should be aware of which features in an image contribute to object classification. Luckily, the FR classification problem we consider here has reasonably well-defined features for each class. In this case, **LIME** can be used to investigate the following questions:

1. Did the model predict the image class mainly according to the central targeted source emission regions?
2. Did the model consider irrelevant emission regions when making predictions?
3. Did the model predict source class in accordance with those field regions responsible for typical source class morphology just as radio astronomers do?

Figure 1 can be seen as an early example of the **LIME** model explanation in our work. It can be seen from the 4_{th} subplot of the figure that for the particular CNN classification algorithm in this work, the network has correctly identified 3C 248 as a FR II source, with both of its radio lobes contributing to FR II class prediction. On the top left of

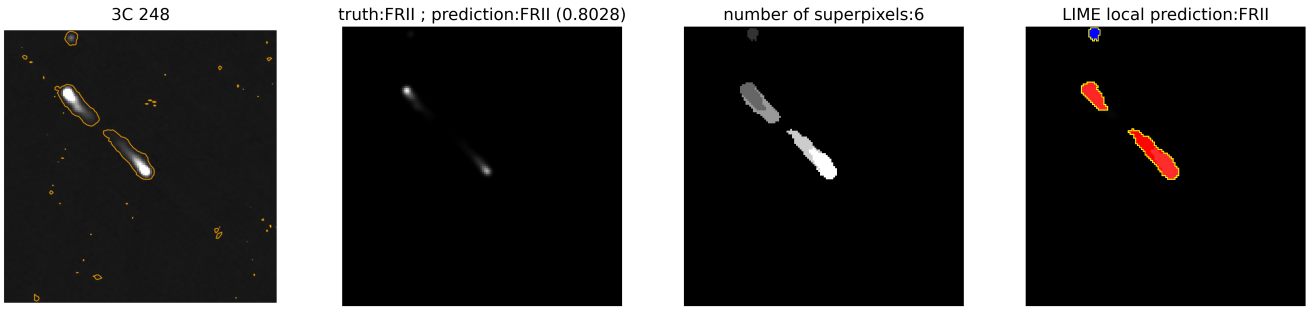

Figure 1. An illustration of sample image model prediction interpretation using **LIME**. From left to right: (1) The first picture shows the contour lines of the original FIRST radio survey image of 3C 248 at 3σ level; (2) the second one shows the same image experinced normalization and was used in the model testing (downloaded from FR-DEEP v2); (3) the third picture illustrates the super-pixels generated by **Felzenszwalb** segmentation method, where the different color of the regions represent different super-pixels; (4) the last one is the interpretation of the picture using LIME. The red regions contribute positively to the FR II class prediction, while the blue one contributes negatively to the same prediction.

the plot, however, another separate source has contributed to the FR II classification negatively. In other words, when there is more than one radio galaxy in an image, the network can no longer claim "The central image source is a FR II radio galaxy". For our network, model user then should visual inspect these images with the aid of generated **LIME** maps.

In terms of early statistical analysis, by visual inspecting the 130 samples in our data testset, we found the network does able to make prediction based on the image central object in most time, especially when an image contains one source only. It generally favors hot spots and those source radio lobes with relatively sharp margins when classifying a source as a FR II object, whereas the situation of FR I source classification is more complicated. **LIME** may also facilitate image mis-classification, though such diagnostics does not always succeed and thus require further investigations. Detailed discussions upon in what aspects can **LIME** help interpret our network predictions will be presented in Tang and Yue et al. (2023, in prep.).

5 Conclusion

We propose the use of **LIME**, a model-agnostic machine learning model interpretation technique to explain deep learning algorithm developed for Fanaroff and Riley radio galaxy morphology classification task. We present a routine of using this technique to explain model prediction behaviour of a trained CNN based classification algorithm. In this work, **LIME** generally segments image into multiple "super-pixels" via **Felzenszwalb** segmentation method, and find those super-pixels in an image that contributed to its model predicted image class. Our early analysis show that for our trained network:

- predict image class mostly based on central source emission regions
- when more than one source presented in the same image, model prediction may be biased.
- FR II source classifications given by the network generally favor hot spots and source radio lobes with sharp margins.

Situations of FR I classification and mis-classification diagnostics are rather complicated, which require further investigation.

Acknowledgements

HT gratefully acknowledges the support from the Shuimu Tsinghua Scholar Program of Tsinghua University; the fellowship of China Postdoctoral Science Foundation 2022M721875; and long lasting support of DoA TAGLAB research group, Tsinghua University and JBCA machine learning group. SY would like to acknowledge the teammate, for their wonderful collaboration and patient support; gratefully thanks Hongsen Yin and Jianxiong Li for the helpful discussions and unreserved supports during the study. SY and ZW are also grateful to the cultivation of Strengthening Foundation Plan conducted by School of Physics and Astronomy, Sun Yat-sen University.

References

[1] Becker, B., Vaccari, M., Prescott, M., and Grobler, T., "CNN architecture comparison for radio galaxy classification", *Monthly Notices of the Royal Astronomical Society*, **503**, 2, May 2021, pp. 1828–1846, doi:10.1093/mnras/stab32510.48550/arXiv.2102.03780.

[2] Bowles, M., Scaife, A. M. M., Porter, F., Tang, H., and Bastien, D. J., "Attention-gating for improved radio galaxy classification", *Monthly Notices of the Royal Astronomical Society*, **501**, 3, February 2021, pp. 4579–4595, doi:10.1093/mnras/staa394610.48550/arXiv.2012.01248.

[3] Fanaroff, B. L. and Riley, J. M., "The morphology of extragalactic radio sources of high and low

luminosity", *Monthly Notices of the Royal Astronomical Society*, **167**, May 1974, pp. 31P–36P, doi:10.1093/mnras/167.1.31P.

[4] Felzenszwalb, P.F., Huttenlocher, D.P. "Efficient Graph-Based Image Segmentation", *International Journal of Computer Vision*, **59**, September 2004, pp. 167–181, https://doi.org/10.1023/B:VISI.0000022288.19776.77

[5] Mohan, D., Scaife, A. M. M., Porter, F., Walmsley, M., and Bowles, M., "Quantifying uncertainty in deep learning approaches to radio galaxy classification", *Monthly Notices of the Royal Astronomical Society*, **511**, 3, April 2022, pp. 3722–3740, doi:10.1093/mnras/stac22310.48550/arXiv.2201.01203.

[6] Norris, R. P., "EMU: Evolutionary Map of the Universe", *Publications of the Astronomical Society of Australia*, **28**, 3, August 2011, pp. 215–248, doi:10.1071/AS1102110.48550/arXiv.1106.3219.

[7] Norris, R. P., "The Evolutionary Map of the Universe pilot survey", *Publications of the Astronomical Society of Australia*, **38**, September 2021. doi:10.1017/pasa.2021.4210.48550/arXiv.2108.00569.

[8] Lao, B., "Artificial intelligence for celestial object census: the latest technology meets the oldest science", *Science Bulletin*, **66**, 21, July 2021, pp. 2145–2147, doi:10.1016/j.scib.2021.07.01510.48550/arXiv.2107.03082.

[9] Tulio Ribeiro, M., Singh, S., and Guestrin, C., ""Why Should I Trust You?": Explaining the Predictions of Any Classifier", *arXiv e-prints*, February 2016, doi:10.48550/arXiv.1602.04938.

[10] Rudnick, L., "Radio Galaxy Classification: #Tags, Not Boxes", *Galaxies*, **9**, 4, October 2021, p. 85, doi:10.3390/galaxies904008510.48550/arXiv.2110.13733.