Scalable Partial Explainability in Neural Networks via Flexible Activation Functions (Student Abstract)

Schyler C. Sun,[1 *] **Chen Li,** [1 *] **Zhuangkun Wei,** [3] **Antonios Tsourdos,** [1] **Weisi Guo** [1,2 †]

[1] AI Group, DARTeC, Cranfield University
[2] Alan Turing Institute
[3] School of Engieering, University of Warwick
{Schyler.Sun, C.Li.21,A.Tsourdos, Weisi.Guo}@cranfield.ac.uk, Zhuangkun.Wei@warwick.ac.uk, wguo@turing.ac.uk

Abstract

Current state-of-the-art neural network explanation methods (e.g. Saliency maps, DeepLIFT, LIME, etc.) focus more on the direct relationship between NN outputs and inputs rather than the NN structure and operations itself, hence there still exists uncertainty over the exact role played by neurons. In this paper, we propose a novel neural network structure with Kolmogorov-Arnold Superposition Theorem based topology and Gaussian Processes based flexible activation function to achieve partial explainability of the neuron inner reasoning. The model feasibility is verified in a case study on binary classification of the banknotes.

Introduction

Motivation and Related Work With the increasing popularity of artificial intelligence, there raise several explicit requirements for XAI in different regions, such as EU GDPR (see Recital 71) requires machine learning algorithms to be able to explain their decisions. For neural networks (NN), The most common way to achieving explainability is to evaluate the impact of each input on the output, e.g. Saliency maps, DeepLIFT and LIME can obtain the approximate solution to provide explanation by reverse analysis for instances. However, these methods are more focusing on the direct relation between inputs and output rather than the NN structure and inner operations, hence there still exists uncertainty over the exact role played by neurons. Authors of (Alaa and van der Schaar 2019) propose to demystify blackbox models with symbolic meta-models which leads a pathway to split and explicit the inner operations of NN and inspired us to improve transparency in NN from an activation function (AF) and topology perspective.

Methodology In this paper, we propose to lay our explainability scheme on a fixed topology mode and reveal the role of each neuron by flexible AFs. Kolmogorov–Arnold Superposition Theorem (KST) can offer an approximation to any continuous function in high dimensional space using a finite composition of (a) univariate continuous functions and (b) addition operation (Kolmogorov 1957). Therefore, based on KST, we establish a scalable NN topology as the foundation of our explainability to replace the conventional weight-bias NN so that to embed all NN inner operations in the AFs. In order to achieve the shape flexibility of the AF as required in KST, we propose to model the AFs as noise-contained Gaussian Processes (GP) by fitting the control points, whose coordinates can be tuned by backpropagation in training process so that to to control the posterior GP curve. In this case, we can achieve the following objectives for the AFs: (a) Ensure of the intrinsic autocorrelation within the function for smoothness and explainability and gain both local and global function adjustability owing to the nature of GP; (b) Avoid over-fitting due to its tolerance of the noise (Williams and Rasmussen 2006). At last, we analyze the AFs qualitatively to partially explain the model inner reasoning.

System Model

Neural Networks Topology

According to the Kolmogorov-Arnold Superposition Theorem (KST) (Kolmogorov 1957), for any $D \in \mathbb{N}$, there exist $R \leq 2D$ and continuous functions $\phi_{rd}(\lambda_d) : \mathbb{I} \to \mathbb{R}$ for $d = 1, 2, ..., D$ and $r = 0, 1, ..., R$, such that: for every arbitrary multivariate continuous function $f(\boldsymbol{\lambda}) : \mathbb{I}^D \to \mathbb{R}$, where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, ..., \lambda_d, ..., \lambda_D]^{\mathrm{T}}$ there exist continuous functions $\Phi_r : \mathbb{R} \to \mathbb{R}$ for $r = 0, 1, ..., R$, such that we may define:

$$F(\boldsymbol{\lambda}) = \sum_{r=0}^{R} \Phi_r \left(\sum_{d=1}^{D} \phi_{rd}(\lambda_d) \right) \tag{1}$$

as an approximate realization of function $f(\boldsymbol{\lambda})$; that is, given any $\epsilon > 0$, $|F(\boldsymbol{\lambda}) - f(\boldsymbol{\lambda})| < \epsilon$ for each $\boldsymbol{\lambda} \in \mathbb{I}^D$, which means functions of the form $F(\boldsymbol{\lambda})$ are dense in $C(\mathbb{I}^D)$. Thus, based on KST, we construct our NN topology (examples in Fig.1(a)) and set $\underline{R} = 0, 1, 2, ...$ in (1), the *repetition level*, as the only scaling parameter of topology, which can determine the expressive power of the NN and thus is the key to the tradeoff between model expressive power and explainability (Fig.1(b)) while flexible activation function is applied. Meanwhile, we define each repetition in topology as a *unit*, which enable us to alleviate the complexity of explanation into a fixed mode.

[*] These two authors contributed equally.

[†] Corresponding author.

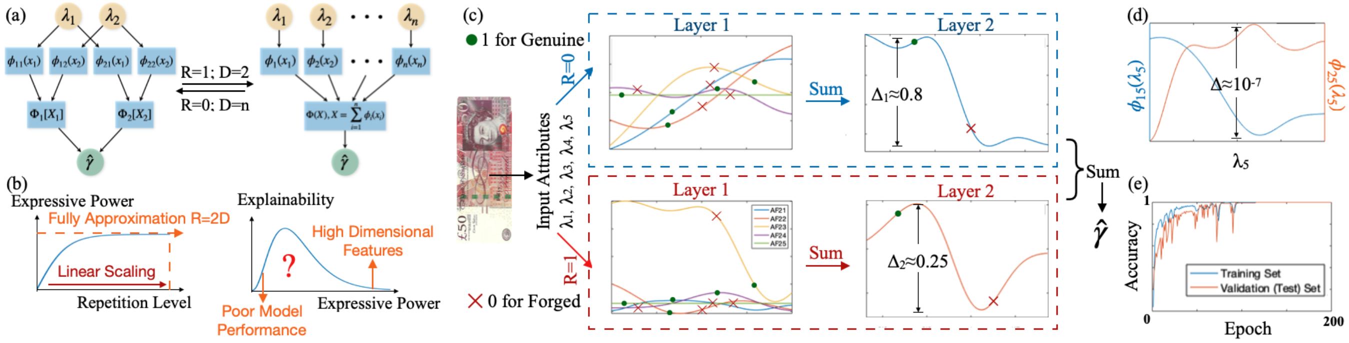

Figure 1: (a) The proposed KST-based topology; (b) Tradeoff between explainability and model performance; (c) The well-trained neurons in our case study. (d) Model explainability on noise attribute; (d) Model performance.

Activation Function

Initialization From the continuous AF domain, finite number of control points are taken randomly as $\boldsymbol{x} = [x_1, x_2, ..., x_n]^{\mathrm{T}}$, with $\boldsymbol{y} = \phi(\boldsymbol{x}) = [y_1, y_2, ..., y_n]^{\mathrm{T}}$ while each activation function is assumed to follow a latent GP plus noise ϵ with these control points. Then, the initial posterior GP mean function can be obtained after maximizing the log marginal likelihood as (Williams and Rasmussen 2006):

$$\overline{\phi(x_*)} = \boldsymbol{k}^{\mathrm{T}}(\boldsymbol{x}, x_*)(\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 \boldsymbol{I_n})^{-1} \boldsymbol{y} \tag{2}$$

where σ^2 is the variance of the additive zero-mean Gaussian noise ϵ and $k(x_i, x_j)$ is the kernel function which can quantify the covariance between every two points. In our experiments, rational quadratic (RQ) kernel is the default choice due to its robustness.

Backpropagation With the symbolic expression of AFs, the model loss can be represented by the coordinates of the control points. Let $L(\boldsymbol{\Theta})$ denote the loss for a batch of instances, where $\boldsymbol{\Theta} = \{\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{y}^{\mathrm{T}}\}$. In each backpropagation epoch, we can perform an update on control points' coordinate with a learning rate η as:

$$\begin{bmatrix} x_i^{t+1} \\ y_i^{t+1} \end{bmatrix} \leftarrow \begin{bmatrix} x_i^t \\ y_i^t \end{bmatrix} - \eta \begin{bmatrix} \frac{\partial L(\boldsymbol{\Theta})}{\partial x_i} \\ \frac{\partial L(\boldsymbol{\Theta})}{\partial y_i} \end{bmatrix} \tag{3}$$

Case Study and Result Discussion

System Setup In this section, we perform a case study on a four-attributes binary classification dataset for banknote authentication [1] with attribute information: (λ_1) variance of wavelet transformed image; (λ_2) skewness of wavelet transformed image; (λ_3) kurtosis of wavelet transformed image; (λ_4) entropy of image. Each attribute is standardized into $[-1, 1]$ interval for clearer illustrating. Furthermore, we add an attribute with artificial noise (λ_5), which uniformly distribute over $[-1, 1]$, in order to evaluate the model's ability to deal with noise. The proposed NN is trained to give the score of the banknotes' image ("1" for genuine and "0" for forged). In our experiment, we set 6 control points in each neuron and $R = 1$ while RQ kernel is used in the GP.

Results Discussion Fig.1(e) shows the classification accuracy achieve 100% for both training and test set after around 120 epochs, which is higher than the classical SVM (99.2%) on this dataset. Fig.1(c) visualizes the well-trained AFs in our NN model. At layer 2, the AF1 and AF2 give different value ranges with $\Delta_1 \approx 0.8 > \Delta_2 \approx 0.25$ which indicates that the unit 1 ($R = 0$) has a more decisive impact than unit 2 ($R = 1$) on the model result. Besides, the green points and the red crosses give two representative examples on how genuine and forged banknotes features are operated in each neuron which gives an explanation for the outputs with inner NN reasoning. Fig.1(d) demonstrates that our model has the robustness to noise attribute λ_5 – AF15 and AF25 give extremely little contribution to the model result other than overfit the data. By analyzing the trained model reversely, scientists can anticipate potential risks in advance about how banknote would be forged to pass the detector under this case. Meanwhile, for AI users, the one-dimensional visual functions flow offer the transparency and partial explainability on how model result come from each attribute input, which can enhance the trust to the model.

Conclusion

In this paper, we propose a novel neural network structure with Kolmogorov-Arnold Superposition Theorem based topology and Gaussian Processes based flexible activation function to achieve partial explainability of the neuron inner reasoning. The model feasibility is verified in a case study on binary classification of the banknotes.

References

Alaa, A. M.; and van der Schaar, M. 2019. Demystifying Black-box Models with Symbolic Metamodels. In *Advances in Neural Information Processing Systems*, 11301–11311.

Kolmogorov, A. N. 1957. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. In *Doklady Akademii Nauk*, volume 114, 953–956.

Williams, C. K.; and Rasmussen, C. E. 2006. *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA.

[1] Data Source: https://archive.ics.uci.edu/ml/datasets/banknote+authentication