Using Qualitative Techniques with Kolmogorov-Arnold Networks for Explainable AI

Ismael Sanz[a,*,1], **Lledó Museros**[a,1], **Vicente Casales-García**[b,1] **and Luis González-Abril**[b,1]

[a]Universitat Jaume I, Spain
[b]Universidad de Sevilla, Spain
ORCID (Ismael Sanz): https://orcid.org/0000-0001-9670-5627, ORCID (Lledó Museros): https://orcid.org/0000-0001-5521-2666, ORCID (Vicente Casales-García): https://orcid.org/0000-0001-8537-7023, ORCID (Luis González-Abril): https://orcid.org/0000-0002-2532-0946

Abstract. Kolmogorov-Arnold networks (KANs) are neural networks that work by fitting a composition of simple univariate functions. They present several advantages with respect to perceptrons; in particular, they are capable of learning fully symbolic equations, thus generating inherently interpretable models. However, these symbolic representations are not generally easily human-understandable. Through a simple use case, we show how we can use qualitative techniques to find intuitive explanations for KAN-learned models. We show how KANs and qualitative techniques are complementary, and propose future avenues of research.

1 Introduction

Kolmogorov-Arnold Networks (KANs) [7] have recently emerged as a hot topic in the field of neural networks. KANs are based in the Kolmogorov-Arnold representation theorem [5], that states that any continuous multivariate function can be represented as a finite sum of continuous univariate functions and their compositions. Leveraging this theoretical foundation, Kolmogorov-Arnold Networks aim to decompose complex, high-dimensional functions into more manageable univariate components, thereby enhancing both interpretability and computational efficiency.

In the current landscape of Artificial Intelligence techniques, this approach promises several important advantages: firstly, KANs can be more parameter-efficient than an equivalent multi-layer perceptron. KANs are particularly useful in applications where the relationship between input variables is intricate and nonlinear. By breaking down these relationships into simpler, univariate functions, KANs can effectively capture the underlying patterns with fewer parameters, reducing the risk of overfitting. Furthermore, the modular nature of KANs allows for easier adaptation and extension, making them suitable for a wide range of tasks from regression and classification to more complex domains such as time-series prediction [9] and image processing [2].

Secondly, they can produce explainable results, since in principle it's possible to fit functions with a symbolic interpretation. This has motivated a flurry of applications where KANs are used to learn processes which can be modeled as relatively straightforward physics-informed equations. From our point of view this is particularly important, since traditional neural network architectures, while powerful, often operate as black boxes. In response, the field of *Explainable Artificial Intelligence* (XAI) [8] has emerged to study how to open these black boxes, which has important technical, ethical and legal implications. KANs are potentially very useful tools in that regard, as an approach which is both technically robust and interpretable in principle.

However, the fact that KANs are able to generate symbolic functions does not mean that these functions are readily human-interpretable. For that reason, we seek to introduce qualitative reasoning approaches into the KAN framework. Qualitative reasoning focuses on understanding and modeling the behavior of systems without relying solely on quantitative data, providing a complementary perspective that emphasizes the relationships and constraints inherent in the data. By combining KANs with qualitative reasoning, we can develop models that not only perform well but also provide deeper insights into the underlying mechanisms of the phenomena being studied.

The main goal of this paper is to present ideas on how KANs and qualitative techniques can be applied together, with a focus on Explainable AI. We will use a case study in which we use KANs to learn a simple color transformation in images, and we will use a qualitative theory to provide an intuitive explanation of the result, which complements the symbolic formula learned by the KAN. We will then discuss possible avenues of research for further integration of KANs and qualitative approaches.

The paper is structured as follows. Section 2 briefly introduces the motivating case study and how it is solved by using KANs. Section 3 shows how we can use qualitative techniques — in particular, the Qualitative Color Description (QCD) theory — to provide an intuitive explanation of the model learned by the KAN. Finally, Section 4 discusses potential areas of research, and the paper ends with some brief conclusions.

2 KANs: Motivating example

2.1 Brief introduction to KANs

The Kolmogorov-Arnold representation theorem [5] states that if f is a multivariate continuous function on a bounded domain, then it can be written as a finite composition of continuous functions of a single

* Corresponding Author. Email: isanz@uji.es
[1] Equal contribution.

variable using addition. More formally, for a smooth $f : [0, 1]^n \to \mathbb{R}$

$$f(x) = f(x_1, ..., x_n) = \sum_{q=1}^{2n+1} \Phi_q(\sum_{p=1}^{n} \phi_{q,p}(x_p))$$

,

where $\phi_{q,p} : [0, 1] \to \mathbb{R}$ and $\Phi_q : \mathbb{R} \to \mathbb{R}$.

Several of these compositions can be combined as layers, thus creating a *Kolmogorov-Arnold Network* (KAN) of arbitrary depths and widths. In each edge of a KAN, there is a univariate function that is fitted to the training data. Thus, a suitable family of basis functions must be selected; the original implementation uses B-splines, but other options are certainly possible. In the next section we show how KANs work with an example.

2.2 *Motivating example: Reconstructing watermarked images with a KAN*

To illustrate the use of KANs, we'll use a very simple digital image processing example. Consider the process of *watermarking* images using a mask. For example, Figure 1 shows a sample photograph, and Figure 2 a watermarked version using a simple mask. Our task is to learn the transformation between the masked pixels in the original photograph and the corresponding ones on the watermarked image.

Figure 1. Original image

This transformation is better defined in a color model such as HSV (Hue-Saturation-Value), which uses human-understandable concepts rather than uninterpretable RGB values. Thus, after transforming the image to HSV, we learn three separate KANs, where each takes as input three separate values (H, S, V) and outputs the transformed Hue, Satutation and Value respectively. We try to keep the KANs as simple as possible, using the smallest network that provides a good result. Other than that, we do not perform any kind of hyperparameter optimization on the KANs. The implementation is done in Python using the PyKAN package.[2]

The learned KAN for Hue is shown in Figure 3. It's a three-layer KAN with three inputs and one output, and the fitted component functions are displayed.

PyKAN uses B-splines as basis functions. After the initial splines are fitted, it's possible to fit a well-known symbolic function that is approximated by these splines. For instance, the bottom-left function

[2] https://github.com/KindXiaoming/pykan

Figure 2. Watermarked image

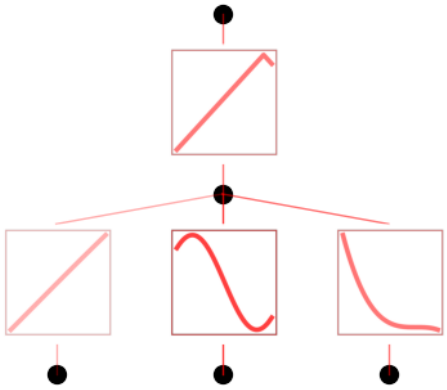

Figure 3. Trained KAN for Hue, showing the learned B-spline functions

can be interpreted as a polynomial, and the function to the left as a trigonometric function such as sine or cosine. The best-fitting function, if any, is selected from a set of well-known candidate functions. After this step, the KAN is retrained with the new component functions. This allows the creation of a fully symbolic representation for the result. In this case, the resulting function is

$$0.9 - 0.17\left|-0.07\left(0.77 - x_V\right)^3 - 0.04\sin\left(4.63x_S + 0.81\right) - 23.09\tanh\left(0.25x_H - 0.05\right) + 4.15\right|$$

where x_H, h_S and h_V represent the input hue, saturation and value respectively. Note that, while symbolic, this can hardly be considered to be a human-friendly explanation of the model, even though it is arguably better in that respect than just having the weights of a neural network. To be fair, in this case it should be certainly possible to achieve a simpler expression by using more training data and a bit of hyperparameter tuning, but our point is that the fact that KANs can produce a symbolic expression does not automatically mean that the result is readily understandable by humans. Incidentally, the expression obtained for the Value KAN is slightly simpler, while the one for the Saturation is far more complex.

Figures 4 and 5 show the learned KANs for Saturation and Value respectively. The learning metrics are reasonable: from about 60000 pixels in the training set, we achieve test RMSE values of beteeen 0.01 and 0.05, which are good enough for this simple case. The KANs train in a few minutes on a Macbook without special GPU support.

Thus, with these three KANs, we are able to fully specify the trans-

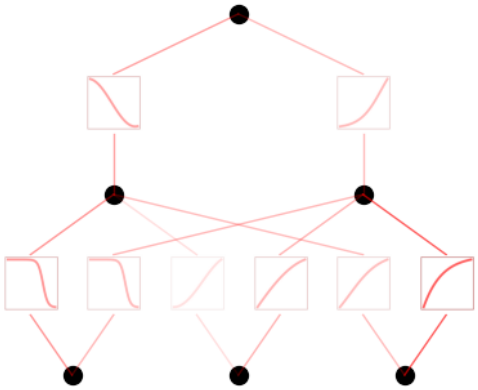

Figure 4. Trained KAN for Saturation

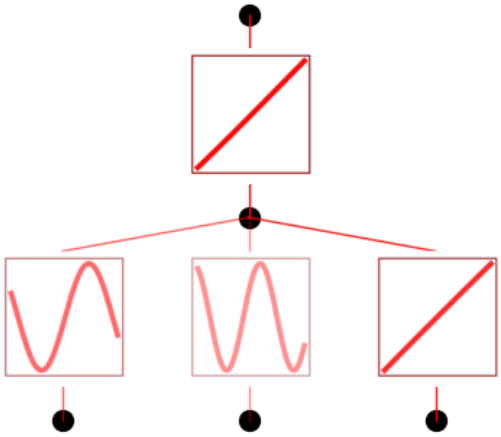

Figure 5. Trained KAN for Value

formation from the color of a pixel in the original image to its corresponding one in the watermarked version. However, it's not clear at all that this result is explainable to humans. In the following section we show how we can use a qualitative color model to better understand this result.

3 Improving interpretability with a qualitative model

3.1 The QCD model

The QCD model [3] defines a reference system in the HSL color space (a variant of HSV) for qualitative color description, which is built according to Figure 6 and defined as:

$QC_{RS} = \{uH, uS, uL, QC_{NAME1..5}, QC_{INT1..5}\}$

where uH is the unit of Hue; uS is the unit of Saturation; uL is the unit of Lightness; $QC_{NAME1..5}$ refers to the color names; and $QC_{INT1..5}$ refers to the intervals of HSL coordinates associated with each color. The chosen QC_{NAME} and QC_{INT} are:

$$QC_{NAME1} = \{black, darkgrey, grey, lightgrey, white\}$$
$$QC_{INT1} = \{[0,20), [20,30), [30,50), [50,75), [75,100) \in uL \mid \forall\, uH \wedge uS \in [0,20]\}$$
$$QC_{NAME2} = \{red, orange, yellow, green, turquoise, blue, purple, pink\}$$

Figure 6. QCD color model

$$QC_{INT2} = \{(335,360] \wedge [0,20], (20,50], (50,80], (80,160], (160,200], (200,260], (260,300], (300,335] \in uH \mid uS \in (50,100] \wedge uL \in (40,55]\}$$
$$QC_{NAME3} = \{pale\text{-}red, pale\text{-}orange, pale\text{-}yellow, \ldots, pale\text{-}blue, pale\text{-}purple, pale\text{-}pink\}$$
$$QC_{INT3} = \{\forall\, QC_{INT2} \mid uS \in (20,50] \wedge uL \in (40,55]\}$$
$$QC_{NAME4} = \{light\text{-}red, light\text{-}orange, light\text{-}yellow, \ldots, light\text{-}blue, light\text{-}purple, light\text{-}pink\}$$
$$QC_{INT4} = \{\forall\, QC_{INT2} \mid uS \in (50,100] \wedge uL \in (55,100]\}$$
$$QC_{NAME5} = \{dark\text{-}red, dark\text{-}orange, dark\text{-}yellow, \ldots, dark\text{-}blue, dark\text{-}purple, dark\text{-}pink\}$$
$$QC_{INT5} = \{\forall\, QC_{INT2} \mid uS \in (50,100] \wedge uL \in (20,40]\}$$

In summary, the QCD defines two set of basic color labels (one monochromatic, one chromatic), which can be combined with "adjectives" (dark, light, pale) that capture meaningful variations in saturation and lightness in an intuitive way.

3.2 QCD interpretation of the KAN model

We can use the QCD to provide a qualitative interpretation of the transformation learned by the KANs. For each pixel in the original image which is covered by the watermarking mask, we compute the corresponding color as transformed by the learned KAN model. Then, we find the QCD label of both colors, thus obtaining a qualitative version of the mapping from the original to the watermarked colors. The results are shown in Table 1. The label on left side of the arrow represents the color in the original image, and the label on the right side represents the corresponding color on the watermarked image. Note that the mapping is not always one to one; in some cases,

some qualitative colors on the original image map to different qualitative labels in the watermarked image.

Table 1. Mapping of QCD colors under the watermarking transformation

$black \mapsto grey$
$dark\_green \mapsto light\_grey$
$dark\_grey \mapsto grey \vert light\_grey$
$dark\_yellow \mapsto light\_grey$
$grey \mapsto light\_grey$
$light\_blue \mapsto white$
$light\_green \mapsto white$
$light\_grey \mapsto white \vert light\_grey$
$light\_red \mapsto light\_red \vert white \vert light\_grey$
$light\_yellow \mapsto white \vert light\_grey$
$pale\_blue \mapsto light\_grey$
$pale\_green \mapsto light\_grey$
$pale\_red \mapsto light\_grey$
$pale\_yellow \mapsto light\_grey$
$white \mapsto white$

Note how, in this case, some patterns are obvious:

- First of all, by using a qualitative representation the color labels are immediately understandable. For example, RGB coordinates $(1, 0.14, 0.19)$ or HSV coordinates $(357\,\mathrm{deg}, 0.85, 1)$, and all perceptually similar sections of the color space are just referred to by the natural language label "red".
- By examining the table, we can see that the transformation corresponds to a "lightening" of the colors, transforming $dark$ colors to their $light$ versions, and converting $light$ colors to $white$ in some cases. Thus, the interpretation becomes immediately obvious.
- Also note that some colors are transformed to labels in the gray scale. This corresponds to a well-known feature of the human vision system, modeled by the QCD, in which very muted colors are perceived as gray (i.e. their chromatic hue is lost), even though if we examine the quantitative coordinates of such colors the hue is unchanged.
- Finally, note how these descriptions correspond to the way in which a human would describe the difference between the watermarked image and the original one. The watermarked areas are normally thought of as "whitened", "lightened" or "muted" with respect to the original version.

These reasons illustrate why qualitative representations are an excellent fit for model explainability in general. And, in this particular case, they provide a natural complement to the symbolic expressions learned by the KAN.

4 Discussion: enhancing the interpretability of KANs with qualitative techniques

In the previous section we have shown how a qualitative representation can be used to provide an intuitive explanation of the result of a model. This is called a post-hoc explanation, and it can certainly be applied to models other than a KAN. However, KANs have some specific properties that make them especially interesting to be used in combination with qualitative techniques. Here we provide two promising examples:

First of all, remember that KANs depend on a suitable family of basis functions to be fitted. In the base implementation we have used in this paper, these functions are B-splines. However, in principle many other function basis are possible; some that have been recently tried are e.g. radial basis functions [6] and wavelets [1]. While using qualitative functions directly is not possible since they are not differentiable, it is indeed possible to use some basis functions that are more readily interpretable in a qualitative way, such as fuzzy basis functions [4].

Another aspect in which qualitative approaches are potentially useful is as *constraints*. It's possible to incorporate constraints to guide the training of the a KAN; this has been used, for example, to incorporate physical knowledge into the system. Of course, there is a long tradition in the field of qualitative reasoning of defining qualitative theories to be used in this way, and in principle it should be possible to incorporate domain knowledge into KAN training using qualitative reasoning techniques.

We consider that these aspects are promising avenues of research that combine the strengths of KAN and qualitative reasoning techniques.

5 Conclusion

In this paper we have introduced Kolmogorov-Arnold networks, and how they can be used to obtain symbolic approximations of functions. After applying them to a simple case study, we have seen how these symbolic formulas can be hard to interpret. As a solution, we have applied the QCD qualitative color theory to find an intuitive explanation of the result. Finally, we have introduced several topics for further research: finding basis functions which are suitable for generating qualitative interpretations, and the incorporation of qualitative constraints into the training process. We consider that KANs and qualitative approaches are complementary approaches, and that these directions of research may provide useful results.

Acknowledgements

This research has been partially funded by the Spanish Ministry of Science under grants PID2021-123152OB-C22 and PDC2021-121097-I00 both funded by the MCIN/AEI/10.13039/501100011033.

References

[1] Z. Bozorgasl and H. Chen. Wav-KAN: Wavelet Kolmogorov-Arnold networks, 2024.
[2] M. Cheon. Kolmogorov-arnold network for satellite image classification in remote sensing, 2024.
[3] Z. Falomir, L. Museros, and L. Gonzalez-Abril. A model for colour naming and comparing based on conceptual neighbourhood. an application for comparing art compositions. *Knowledge-Based Systems*, 81:1–21, 2015.
[4] H. M. Kim and J. Mendel. Fuzzy basis functions: comparisons with other basis functions. *IEEE Transactions on Fuzzy Systems*, 3(2):158–168, 1995. doi: 10.1109/91.388171.
[5] A. Kolmogorov. On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. *Proceedings of the USSR Academy of Sciences*, pages 179––182, 1956. English translation: Amer. Math. Soc. Transl., 17 (1961), pp. 369–373.
[6] Z. Li. Kolmogorov-Arnold networks are radial basis function networks, 2024.
[7] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark. KAN: Kolmogorov-Arnold Networks, 2024.
[8] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, and S. Stumpf. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024.
[9] C. J. Vaca-Rubio, L. Blanco, R. Pereira, and M. Caus. Kolmogorov-arnold networks (kans) for time series analysis, 2024.