FINDING ORIENTATION IN MODEL INTERPRETABILITY

How should you interpret your deep learning model?

A map to navigate among the main techniques

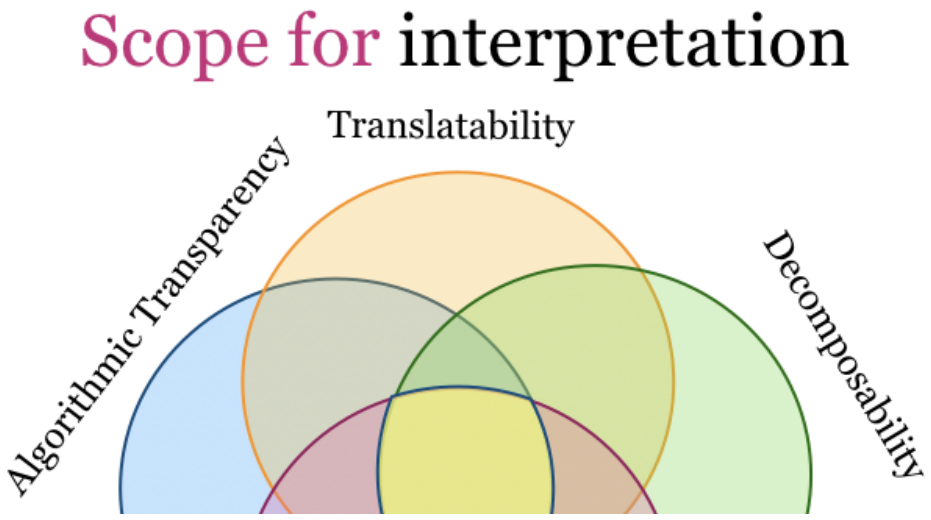
Mara Graziani Follow
Apr 24 · 4 min read ★

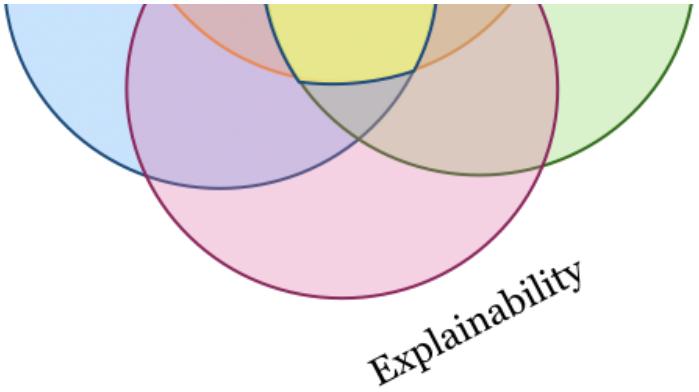
Machine learning interpretability has been a fast growing research field for the last ten years, at least. If you would like to introduce some interpretability to your deep learning model, it could be a bit tricky at first to find orientation among all the techniques that have been developed. This post proposes a map to navigate the sea of tools according to the purpose of the interpretation. Some of the main approaches in the literature are grouped in four different areas, defined as scopes for interpretability. Each area is linked to a grouping of techniques in the literature of deep learning interpretability that can address the specific requirements for that scope. In this way it is possible to navigate from one interpretability scope to a specific set of tools.

Note that this post only covers the main ideas and branches in interpretability research and does not offer exhaustive coverage of all the techniques! You can find at the bottom of this post the links to the papers and techniques that are covered by our analysis.

Understanding the scope of interpretability

The first thing to address is understanding what do you need the most in your interpretability analysis. What does your need for interpretability look like? Are you delivering the interpretation to end-users of your deep model? Or is interpretability needed for developmental purposes, and maybe spot some unwanted behavior? These kind of questions can guide you towards the method that suits the most to your needs.





Clarifying automated procedures (transparency), "translating" decisions to users (translatability), understanding components individually (decomposability) and giving contrasting explanations (explainability) are four main scopes for seeking model interpretation.

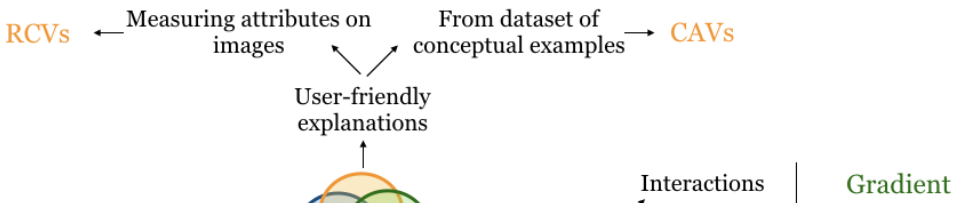
Depending on the interpretability scope, there could be different ways to analyse the same task. This diagram collects four main scopes for interpretation*. Each of them can lead to a different interpretability analysis.

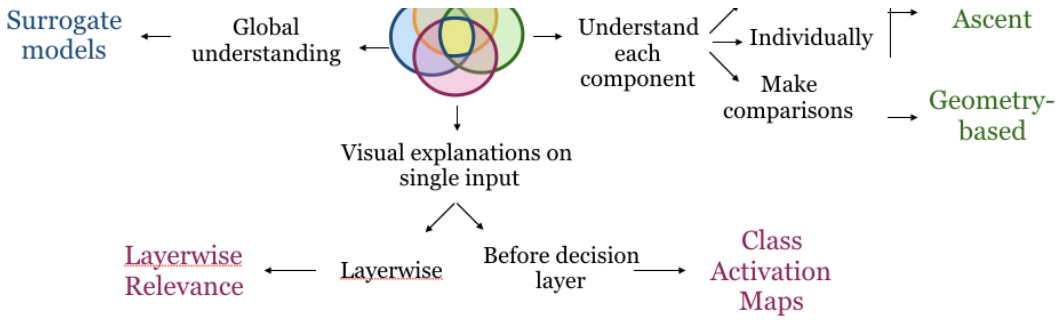
In some cases, you may be looking for something called **"algorithmic transparency"**, namely a clear and understandable procedure that brings the model from an input to an output. In other cases, you may be interested in **"translating"** what is happening in the algorithm to a specific user. Interpretability of machine learning models for medical diagnostics, for example, needs to be translated to different users, like the patient, the clinician or the technician using the diagnostic machine. Each of them has a different technical knowledge and understands the interpretability analysis only when given at the correct level of technicality.

Decomposability focuses on understanding in depth the role of each component in our tool. This is great for debugging purposes, for example for identifying unwanted behavior in our model. **Explanations** are particularly useful to clarify why a certain decision was taken by the network. In particular, contrastive explanations (in the form of why a certain decision was taken rather than another) seem to be intuitive and valuable to non-experts**.

These four areas of "scopes for interpretability" are partially overlapping and partially complementary. If you are interested in having a global overview of your model's behavior, then your desired target is probably somewhere at the intersection of all of them. Alternatively, you can navigate towards one specific area rather than another depending on the main scope for your need of interpretation.

Start by understanding what the interpretation is needed for.





A map to choose the deep learning interpretability technique most appropriate to our needs.

If interpretability is needed to understand each component in the deep learning model, like the activations of a neuron, a layer or a channel gradient-ascent based techniques are a good starting point. If you are rather interested in comparisons across networks, then you may look into geometric approaches such as Singular Vector Canonical Correlation Analysis.

Deep Learning Explained in 7 Steps - Updated | Data Driven Investor

Self-driving cars, Alexa, medical imaging - gadgets are getting super smart around us with the help of deep learning...

www.datadriveninvestor.com

For a global understanding of the network's decision function, you may start from surrogate models. Finally, if you are looking for translatability, namely for generating user-centric explanations, you could use concept attribution techniques such as Concept Activation Vectors and Regression Concept Vectors.

> *Thank you for reading this post on model interpretability! We hope you found this useful and interesting. If so, please let us or the author know!*
>
> *Note that this post is not exhaustive of all techniques and methods. Share in the comments what do you think about it, and if you would recommend some other technique!*

Finally, if you would like to know more, listen to our latest talk about Machine Learning Interpretability.

Footnotes:

* Some of these were analyzed by Lipton in "The Mythos of Model Interpretability"

* These questions were found from a sociological perspective more intuitive and more valuable, even to non-experts in computer science. Also other types of explanations are included in explainability, however, such as associative, interventionist and counterfactuals explanations. [1] Explanation in AI:Insights from the Social Sciences, Tim Miller (https://arxiv.org/pdf/1706.07269.pdf)

Machine Learning Model Interpretability Deep Learning Artificial Intelligence Medical Imaging