Towards Compositional Interpretability (Extended abstract)

Sean Tull, Robin Lorenz, Stephen Clark
Quantinuum, 17 Beaumont Street, Oxford, UK

Though AI systems based on machine learning (ML) have achieved great practical successes over the past decade, there have been increasing societal concerns over their lack of *interpretability*, which is especially critical for high-stakes problems like in medical, legal or financial contexts. The growing area of *eXplainable AI (XAI)* has aimed to remedy this, but several authors have criticised the dominant approaches of applying 'post-hoc' techniques to black-box models that lack interpretable structure [8, 4].

In our forthcoming[1] article [9] we provide a categorical framework for XAI in which one may both define AI models and analyse their interpretability. We use it to describe a wide range of common AI models in terms of string diagrams (see Fig. 1) and demonstrate the explanatory strengths of those that come with rich interpretable categorical structure, which we call *compositionally interpretable* (CI).

This categorical approach comes with several benefits. It provides a unified diagrammatic perspective on AI models, which makes compositional structure explicit, allows to meaningfully compare different models and serves as an invitation to the categorical perspective for those with an ML background. Amongst many others we discuss linear and rule-based models, neural networks (NN), recurrent NNs and transformers [11], DisCo models [3, 2, 12], conceptual space models [1, 10] and causal models [7, 5, 6]. On the basis of that common ground we then give precise definitions that clarify and disentangle aspects of common intuitions in XAI, in particular, when a model as such is interpretable rather than just affords (approximate) explanations of particular outcomes.

This allows us to characterise the vague notion of 'intrinsically interpretable models' and generalise it beyond the standard linear or rule-based ones. Further examples of CI models include DisCo models in NLP and importantly, causal models. The approach in fact helps clarify the prominent role of causal intuitions in XAI, where for explanatory purposes it is crucial to distinguish between causal structure 'just' induced by a NN-based model and it corresponding to the phenomena that the model is about.

The categorical approach can accommodate classical, probabilistic, as well as quantum semantics of models and indeed is the starting point for defining the interpretability of quantum AI models, which are naturally defined compositionally in terms of circuit diagrams. Finally, the structure of CI models provides interpretability benefits and facilitates new forms of explanations, including that of a *rewrite explanation* for the output of a compositional model.

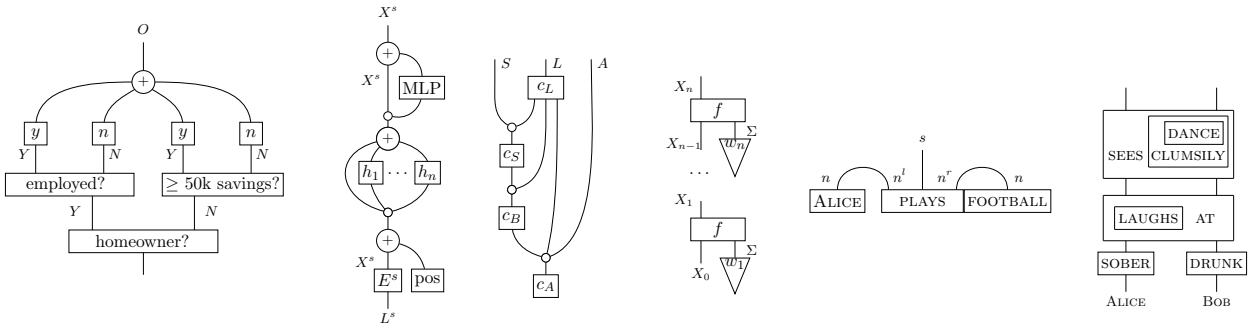

Figure 1: Structure diagrams for decision tree, transformer, causal model, RNN, DisCoCat, DisCoCirc.

Formally, we define a *compositional model* $\mathbb{M}$ as given simply by a *structure* category $\mathbb{S}$, *semantics* category $\mathbf{C}$, and *representation* functor $[\![-]\!]\colon \mathbb{S} \to \mathbf{C}$. Typically both are symmetric monoidal or cd-categories, and $[\![-]\!]$ preserves this structure. More precisely, $\mathbb{S} = \mathbf{Free}(G)$ is a free category of diagrams

[1]The full article is in draft stage, and will be submitted to the arxiv at the end of April 2024.

generated by some fixed signature G, thought of as defining the abstract variables and processes which generate the model. The semantics $\mathbf{C}$ could be taken in a category of neural networks, probabilistic maps, or quantum processes, amongst others. In practice, spelling out a compositional model typically involves drawing a distinguished string diagram, which captures the overall input-output process of the model, and implicitly specifies G.

Next we define an *interpretation* for a model semi-formally, making use of both structure and semantics. This consists of a signature $\mathbb{H}$ of 'human-friendly' terms or concepts, along with a partial structure-preserving map $\mathcal{I}\colon G \to \mathbb{H}$ called the *abstract interpretation*, as well as a family of partial maps from the homsets of $\mathbf{C}$ to $\mathbb{H}$ indexed by types from $\mathbb{S}$, which we call the *concrete interpretation*. Intuitively, an abstract interpretation gives meaning 'labels' to certain variables in a model, for example asserting that a representation space encodes 'brightness', while a concrete interpretation gives meaning to states and processes, for example saying that the state 0 means 'dark' while the state 1 means 'bright'.

One finds that while ML models can formally be cast as compositional models by drawing a string diagram expressing their architecture, most (such as transformers) are opaque boxes for which only their inputs and outputs tend to be interpretable, but not the intermediate variables and processes. Indeed, as a result most XAI methods focus on probing and partially characterising the input-output behaviour.

A stronger interpretability however obtains for models whose *internal* structure is also interpretable, and moreover is interesting as compositional structure, i.e. neither a black box, nor a fully connected NN. We refer to such models as compositionally interpretable. This notion is best understood as a graded one, ranging from classical examples of rule-based models and the sequential-only composition of recurrent neural networks, to more richly compositional models such as DisCo models in NLP [3, 12] and interpretable causal models [6].

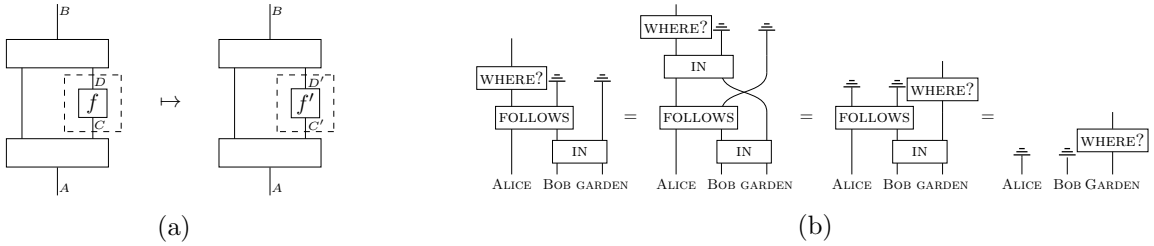

Figure 2: (a) Illustration of diagram surgery; (b) rewrite explanation, making use of signalling arguments.

To make the benefits of such manifestly interpretable CI models explicit, we outline several explanation techniques afforded by such models. Firstly, *signalling arguments* can be used to provide hard constraints on which inputs can affect a given output. These arguments require non-trivial compositional structure, and so are typically not possible for models given simply by NNs.

Secondly, the ability to perform *diagram surgery* allows one to ask 'what if' questions about a process described by an interpretable diagram. This notion includes interventions on causal models [5, 6] and as an explanatory technique generalises the *counterfactual explanations (CFEs)* commonly studied in XAI, which inspect the output of a model after modifying its input, to allow intervention on arbitrary internal processes in a model.

Finally, we propose the notion of a rewrite explanation as an XAI technique applicable to CI models. Given some output, a rewrite explanation consists of knowledge about local aspects of a model, given by (approximate) equations between interpreted diagrams, along with a rewriting proof that these lead to the given output. This kind of local reasoning generalises the 'traceability' aspect of intrinsically interpretable models, providing CI models with the ability to give explanations for, and guarantees on, their behaviour.

Ultimately, the aim of this work is to show how the compositional approach to AI, developed in recent years within the applied category theory community, can provide a useful perspective on questions of interpretability, both in clarifying models and their interpretation and in defining new kinds of models that are manifestly interpretable and compositional. In future work, we hope to explore the extent to which such meaningful compositional structure can be learned from data, explore the role of quantum models more deeply, and further sharpen our proposed explainability tools for CI models.

References

[1] Joe Bolt, Bob Coecke, Fabrizio Genovese, Martha Lewis, Dan Marsden, and Robin Piedeleu. Interacting conceptual spaces i: Grammatical composition of concepts. *Conceptual spaces: Elaborations and applications*, pages 151–181, 2019.

[2] Bob Coecke. The mathematics of text structure. *Joachim Lambek: The Interplay of Mathematics, Logic, and Linguistics*, pages 181–217, 2021.

[3] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen J Clark. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36(1):345–384, 2010.

[4] Timo Freiesleben and Gunnar König. Dear xai community, we need to talk! fundamental misconceptions in current xai research. *arXiv preprint arXiv:2306.04292*, 2023.

[5] Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. Causal inference by string diagram surgery. In *Foundations of Software Science and Computation Structures: 22nd International Conference, FOSSACS 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6–11, 2019, Proceedings 22*, pages 313–329. Springer, 2019.

[6] Robin Lorenz and Sean Tull. Causal models in string diagrams. *arXiv preprint arXiv:2304.07638*, 2023.

[7] Judea Pearl. *Causality*. Cambridge university press, 2009.

[8] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019.

[9] Sean Tull, Robin Lorenz, and Stephen Clark. Forthcoming: 'Towards Compositional Interpretability'. 2024.

[10] Sean Tull, Razin A Shaikh, Sara Sabrina Zemljic, and Stephen Clark. From conceptual spaces to quantum concepts: formalising and learning structured conceptual models. *Quantum Machine Intelligence*, 2024.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[12] Vincent Wang-Mascianica, Jonathon Liu, and Bob Coecke. Distilling text into circuits. *arXiv preprint arXiv:2301.10595*, 2023.