Causal Scrubbing and Abstractions

Dylan X, Joyee C

April 2023

1 Introduction

In this document, we will connect our understanding of causal scrubbing to abstractions and see how to define and find what we want in the latter. We first define *sample-wise* causal scrubbing as follows:

Definition 1 (Sample-wise causal scrubbing)**.** *Given any $h = (G, I, c)$, after running the scrubbing algorithm on G on some input, we want the expected difference between scrubbed and normal output of G for any input x (rather than the difference in expectations over $x \sim D$) to be close to zero.*

We further define *exact* sample-wise causal scrubbing to require that the scrubbed and normal output be exactly the same for any input x.

Using these definitions, we then define our model of abstractions based on G and I. For any node $n_I \in I$, let $n_G = c(n_I)$ be the corresponding node in G. We then impose some topological ordering on G and I such that, for instance, the activations of certain nodes are computed from the activations of their "parent" nodes, like layers in neural networks. Let $\{p_I\}$ and $\{p_G\}$ be the set of parents of n_I and n_G respectively; note that c restricted to $\{p_I\}$ is a subset of $\{p_G\}$ since c is a graph homomorphism. We then make the following assumption:

Lemma 1.1. *For any inputs from our dataset $x, x' \in D$, let $n_G(x)$ and $n_I(x)$ be the activation of node n_G and n_I on input x respectively. Then if $n_G(x) = n_G(x')$, then $n_I(x) = n_I(x')$.*

In some sense, n_I is less "detailed", or loses more information than n_G, because it does not distinguish between inputs that n_G does not distinguish on. Let $V(n_G) = \{n_G(x)\}$ and $V(n_I = \{n_I(x)\}$ be the set of all values that n_G and n_I could have respectively. We can thus say $g : V(n_G) \to V(n_I)$ maps $n_G(x) \mapsto n_I(x)$ for all such x, x' (and is well defined). (I guess the rest of the map is ambiguous?) Similarly, we say g maps every $p_G = c(p_I) \in \{p_G\}$ to its corresponding $p_I \in \{p_I\}$. Finally, let f and f' calculate the activation of a node in G and I respectively given the activations of its parents.

We then define what it means for I to be a proper abstraction using a consistency condition:

Definition 2. *For any node $n_I \in I$, if this diagram commutes:*

$$\begin{tikzcd}
\{p_G\} \arrow[r, "f"] \arrow[d, "g"'] & n_G \arrow[d, "g"] \\
\{p_I\} \arrow[r, "f'"] & n_I
\end{tikzcd}$$

then I is a proper abstraction.

We also need some way to compare the outputs of G and I, as they may not be the same type/inherently comparable. Let O_G and O_I be the set of all possible outputs of G and I (e.g. the activation of their respective "last" nodes) and let $o : O_I \to O_G$ "convert" o_I into o_G (i.e. $o_I \sim o_G$ iff $o(o_I) = o_G$). Then we say I is an o-abstraction if g on o_I simply equals o. I is a proper abstraction if it is an o-abstraction on the trivial function $o : o_I(x) \mapsto o_G(x)$.

We thus show:

Theorem 1.2. *Any hypothesis $h = (G, I, c)$ that is a proper abstraction is accepted by exact sample-wise causal scrubbing.*

Proof. For any $p_I \in \{p_I\}$, suppose we have some $x, x' \in D$ such that $p_I(x) = p_I(x')$. It follows that the commun

2 Construction in the Other Direction

We just showed, in some sense, that a consistent abstraction leads to a good interpretation by exact causal scrubbing standards. For the other direction, we first present a notion of a "less specific" hypothesis (technically less than or equally specific):

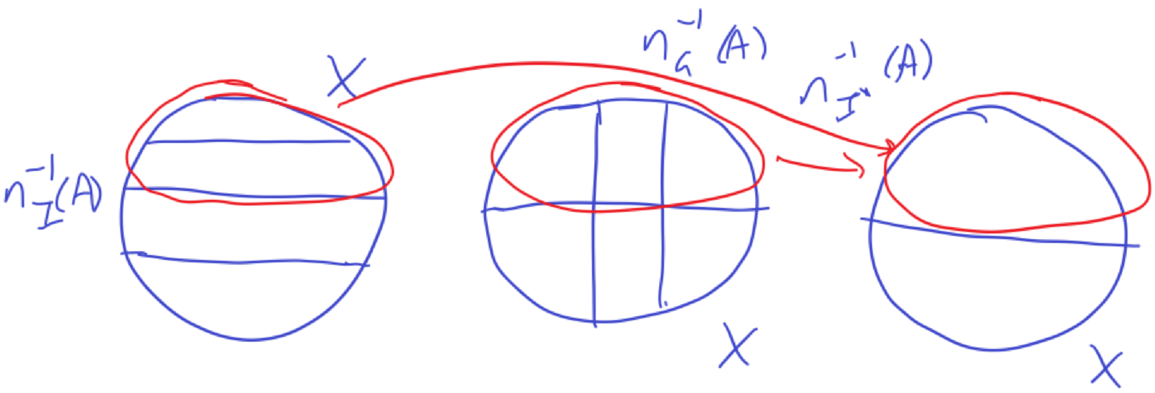
Definition 3. *Given $h = (G, I, c)$, let I' be an alternative interpretation of G. Let $d : I' \to I$ be an injective graph homomorphism. Then I' is less specific than I, denoted $I \leq I'$ for graphs, if and only if we can find the value of every node $n_{I'}$ from the value of $d(n_{I'}) := n_I$.*

Intuitively, a hypothesis being less specific than another hypothesis is one way in which a hypothesis can be "better" than another. We then show that:

Theorem 2.1. *For any (G, I, c) accepted by causal scrubbing, we can construct a less specific interpretation $I' \geq I$ that is both accepted by causal scrubbing and is a proper abstraction.*

Proof. Let X be the set of all possible inputs (e.g. a vector with inputs between $[0, 1)$) and A be the set of all possible activations. (Note that $V(n_G)$ is the range of $n_G(x)$ on A, and similarly for $V(n_I)$.) Then we can partition X into subsets $n^{-1}(A) \subset X$ depending on the output of $n^{-1}(a) = \{x|n(x) = a\}$ for some activation $a \in V(n)$.

We then define I' and bijective $d : I' \to I$ such that $d(n_{I'}) = n_I$ for any $n_{I'} \in I'$ and $n_{I'}^{-1}(A)$ partitions X into the minimum disjoint blocks such that all $n_G^{-1}(a_1)$ and $n_I^{-1}(a_2)$ are subsets of some $n_{I'}^{-1}(a)$ and do not intersect any other block. When this is true, we say that $n_G^{-1}(A) \geq n_{I'}^{-1}(A)$, $n_I^{-1}(A) \geq n_{I'}^{-1}(A)$, and $n_{I'}^{-1}(A)$ is the "join" of $n_G^{-1}(A)$ and $n_I^{-1}(A)$. In other words, $n_G^{-1}(a_1) \cap n_I^{-1}(a_2) \neq \emptyset$ if and only if $n_G^{-1}(a_1) \subseteq n_{I'}^{-1}(a)$ and $n_I^{-1}(a_2) \subseteq n_{I'}^{-1}(a)$ for some $a \in A$. An example is shown below:



Since this partition covers all of X, $n_{I'}(x)$ is well defined. Also, note that the function $n_{I'}(x) = a$ for all x simply groups X into one partition, which contains all $n_G^{-1}(a_1)$ and $n_I^{-1}(a_2)$ by definition. Hence for any n_G and n_I, there exists a partition with disjoint block(s) containing all intersecting $n_G^{-1}(a_1)$ and $n_I^{-1}(a_2)$, meaning such a partition with the smallest possible disjoint blocks exists. Thus this construction of I' exists for any G and I. I' is also less specific than I, as $n_{I'}(x) = d^{-1}(n_I(x))$ for any x.

We now show that I' is also a proper abstraction; by Theorem 1.2, I' will then also be accepted by causal scrubbing. By construction, for any $n_G, n_{I'}$, $n_G^{-1}(a_1)$ is always contained by some $n_{I'}^{-1}(a)$. This is because there must exist some $n_I^{-1}(a_2)$ that intersects with $n_G^{-1}(a_1)$, as otherwise n_I^{-1} would not be a complete partition of X, which implies $n_G^{-1}(a_1) \subseteq n_{I'}^{-1}(a)$ and $n_I^{-1}(a_2) \subseteq n_{I'}^{-1}(a)$ for some $a \in A$. We then let well-defined $g : n_G \to n_{I'}$ send $a_1 = n_G(x) \mapsto a$ for all x.

$$\begin{tikzcd}
 & \{p_G\} \arrow[r] \arrow[dd] & n_G \arrow[dd] \\
X \arrow[ur, dashed] \arrow[dr, dashed] \arrow[urr, bend left] \arrow[drr, bend right] & & \\
 & \{p_I\} \arrow[r] & n_I
\end{tikzcd}$$

We now only need to show that there exists an $f' : \{p_I\} \to n_I$. If f' exists, then because the maps $\{p_G\} \to \{p_I\}$ and $n_G \to n_I$ already commute on X, and $f : \{p_G\} \to n_G$ is already defined, then the consistency condition is satisfied. Let $\{p_G^{-1}\}$ partition X based on the ordered tuple $(p_G(x))$ for some ordering of the parents of n_G, and define $\{p_I^{-1}\}$ similarly. Then $n_G^{-1}(A) \geq \{p_G^{-1}\}(A)$ and $n_I^{-1}(A) \geq \{p_I^{-1}\}(A)$ because $n^{-1}(A)$ must be less granular/detailed then the partitions of X of its parents for any node n in G or I, as otherwise n could not be calculated from its parents.

Now consider any subset $n_{I'}^{-1}(a) \subseteq X$ in the partition of X by $n_{I'}^{-1}$. By definition, this subset contains all given elements

3 Remaining problems

Both of the above proofs analyze on the scale of every node, which inductively results in an entire graph satisfying what we want. Unfortunately, this method may not generalize to "inexact" causal scrubbing or approximate abstractions, which of course are far more applicable. For instance, small errors in node-level abstraction (i.e. $g(f(\{p_G\}) - f(g(\{p_G\})$ may propagate to large errors in the final output of G. Similarly, large errors for each node may "cancel out" to a negligible error in the final output, causing a bad abstraction to be accepted by causal scrubbing. Both of these possibilities should be examined further.