Causal Abstractions of Neural Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Structural analysis methods (e.g., probing and feature attribution) are increasingly important tools for neural network analysis. We propose a new structural analysis method grounded in a formal theory of *causal abstraction* that provides rich characterizations of model-internal representations and their roles in input/output behavior. In this method, neural representations are aligned with variables in interpretable causal models, and then *interchange interventions* are used to experimentally verify that the neural representations have the causal properties of their aligned variables. We apply this method in a case study to analyze neural models trained on Multiply Quantified Natural Language Inference (MQNLI) corpus, a highly complex NLI dataset that was constructed with a tree-structured natural logic causal model. We discover that a BERT-based model with state-of-the-art performance successfully realizes the approximate causal structure of the natural logic causal model, whereas a simpler baseline model fails to show any such structure, demonstrating that neural representations encode the compositional structure of MQNLI examples.

1 Introduction

Explainability and interpretability have long been central issues for neural networks, and they have taken on renewed importance as such models are now ubiquitous in research and technology. Recent structural evaluation methods seek to reveal the internal structure of these "black box" models. Structural methods include probes, attributions (feature importance methods), and interventions (manipulations of model-internal states). These methods can complement standard behavioral techniques (e.g., performance on gold evaluation sets), and they can yield insights into how and why models make the predictions they do. However, these tools have their limitations, and it has often been assumed that more ambitious and systematic causal analysis of such models is beyond reach.

Although there is a sense in which neural networks are "black boxes", they have the virtue of being completely closed and controlled systems. This means that standard empirical challenges of causal inference due to lack of observability simply do not arise. The challenge is rather to identify high-level causal regularities that *abstract away* from irrelevant (but arbitrarily observable and manipulable) low-level details. Our contribution in this paper is to show that this challenge can be met. Drawing on recent innovations in the formal theory of causal abstraction [5, 27, 1, 2], we offer a methodology for meaningful causal explanations of neural network behavior.

Our methodology consists of three stages. (1) Formulate a hypothesis by defining a causal model that might explain network behavior. Candidate causal models can be naturally adapted from theoretical and empirical modeling work in linguistics and cognitive sciences. (2) Search for an alignment between neural representations in the network and variables in the high-level causal model. (3) Verify experimentally that the neural representations have the same causal properties as their aligned high-level variables using the *interchange intervention* method of [12].

As a case study, we apply this methodology to LSTM-based and BERT-based natural language inference (NLI) models trained on the logically complex Multiply Quantified NLI (MQNLI) dataset

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

of [11]. This challenging dataset was constructed with a tree-structured natural logic causal model [21, 35, 16]. Our BERT-based model has the structure of a standard NLI classifier, and yet it is able to perform well on MQNLI (88%), a result Geiger et al. achieved only with highly customized task-specific models. By contrast, our LSTM-based model is much less successful (46%).

The obvious scientific question in this case study is what drives the success of the BERT-based model on this challenging task. To answer this we employ our methodology. (1) We formulate hypotheses by defining simplified variants of the natural logic causal model. (2) We search over potential alignments between neural representations in BERT and variables in our high-level causal models. (3) We perform interchange interventions on the BERT model for each alignment. We find that our BERT model partially realizes the causal structure of the natural logic causal model; crucially, the LSTM model does not. High-level causal explanation for system behavior is often considered a gold standard for interpretability, one that may be thought quixotic for complex neural models [19]. The point of our case study is to show that, using the methodology we propose, this high standard can be achieved.

We conclude by comparing our methodology to the probing and the attribution method of integrated gradients [33]. We argue that correlational methods such as probing are unable to provide a causal characterization of models. While attribution methods do measure causal properties, our method of aligning neural representations with variables in a causal model provides a richer causal characterization. (We make these arguments formally in Appendix A.)

2 Related Work

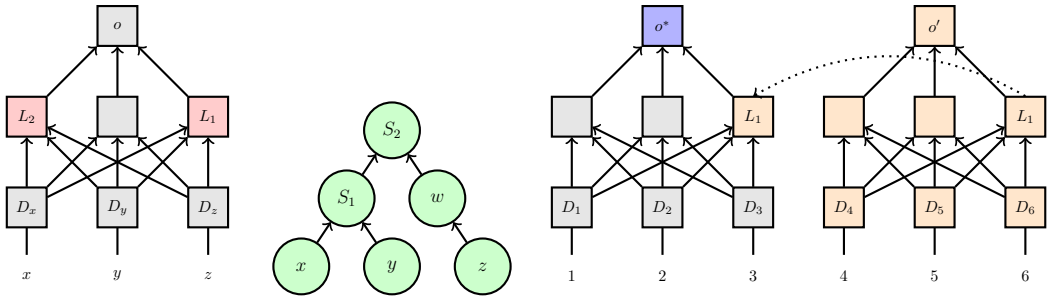
Probes Probes are supervised models trained on the internal representations of networks with the goal of determining what those internal representations encode [15, 24, 34, 7]. Probes are fundamentally unable to directly measure causal properties of neural representations. We argue this formally in Appendix A.1 by presenting a simple network in which a probe finds information that, by construction, plays no causal role in the network. [25], [9], and [12] have argued that probes are poor indirect evidence of causal properties as well.

Attribution Methods Attribution methods aim to quantify the degree to which a network representation contributes to the output prediction of the model, for a specific example or set of examples [38, 31, 29, 3, 33]. [6] develop an attribution method that explicitly treats neural models as structured causal models and computes the individual causal effect of a feature to determine its attribution. While attribution methods certainly measure causal properties, assigning each neural representation a single scalar value is a limited characterization of that representation's role in network behavior. We address this issue directly in Appendix A.3 by presenting a simple network in which the integrated gradients method [33] cannot determine between two neural representations that, by construction, have distinct causal roles in the network.

Interventions Intervention methods actively manipulate the value of neural representations during the network evolution and measure the activity of causally downstream neurons. Several analysis methods use interventions on input representations to provide causal characterizations [10, 23], and some also construct interpretable models from these characterizations [20, 26]. Model internal intervention have been used for task-specific analyses [30, 13, 12], but, to our knowledge, only [9], [36], and this paper develop a general methodological approach founded on such interventions.

Causal Abstraction Theories of causal abstraction define when a high-level model faithfully abstracts a low-level model in a way that preserves causal relationships [5, 27, 1, 2]. The notion of abstraction that we employ is here is a relatively simple one called *constructive abstraction* [1]. Informally, a high-level model is a constructive abstraction of a low-level model if there is a way to partition the variables in the low-level model where each high-level variable can be assigned to a low-level partition cell, such that there is a systematic correspondence between interventions on the low-level partition cells and interventions on the high-level variables.

There are two properties of constructive abstraction that make it ideal for neural network analysis. First, the information content of partitions of low-level variables is determined by the high-level variables that they correspond to. For neural networks, the partitions of low-level variables are sets of neurons, and our method supports reasoning at the level of vector representations (sets



(a) Network N_+ with L_1 highlighted as the hypothesized location encoding $x + y$ and L_2 as the location encoding $w = z$.

(b) The causal model C_+ that first computes $S_1 = x + y$ and $w = z$, before computing the final output $S_2 = w + S_1$.

(c) Network-level interchange intervention. The network processes two different input sequences. The representation created at L_1 for input sequence $(1, 2, 3)$ is replaced by the corresponding representation created for input sequence $(4, 5, 6)$.

Figure 1: Our motivating example. The interchange intervention method depicted schematically at right can be used to test the hypothesis that the model C_+ is a causal abstraction of the network N_+.

of neurons). Second, the causal dependencies between partitions of low-level variables are not necessarily preserved as causal dependencies between the high-level variables corresponding to these partitions. For example, the low-level model might be a fully connected neural network, whereas the high-level model might have much sparser connections. For neural network analysis, this means we can find causal abstractions that have far simpler causal structures than the underlying neural networks. We provide an example in the next section.

3 Causal Abstractions of Neural Networks

We now describe our methodology in more detail, illustrating the relevant concepts with an example of a neural network performing basic arithmetic. Specifically, suppose that we have a neural network N_+ that takes in three vector representations D_x, D_y, D_z representing the integers x, y, and z, and outputs the sum of the three inputs: $N_+(D_x, D_y, D_z) = x + y + z$. We seek an informative causal explanation of this network's behavior.

Formulating a Hypothesis A human performing this task might follow an algorithm in which they add together the first two numbers and then add that sum to the third number. We can hypothesize that the behavior of N_+ is explained by this algorithm. Specifically, the network combines D_x and D_y to create an internal representation at some location L_1 encoding $a = x + y$; it encodes z at some location L_2; and L_1 and L_2 are composed to encode $a + z$ at the location of the output representation. This hypothesis is given schematically in Figure 1a.

Following our methodology, we first define the causal model C_+ in Figure 1b. Our informal hypothesis that a neural network's behavior is explained by a simple algorithm can then be restated more formally: C_+ is a constructive abstraction of the neural network N_+.

Alignment Search Now that we have hypothesized that the causal model C_+ is a causal abstraction of the network N_+, the next step is to align the neural representations in N_+ with the variables in C_+. The input embeddings D_x, D_y, and D_z must be aligned with the input variables x, y, and z and the output neuron o must be aligned with the output variable S_2. That leaves the intermediate variables S_1 and w to be aligned with neural representations at some undetermined locations L_1 and L_2. If this were an actual experiment (see below), we would perform an *alignment search* to consider many possible values for L_1 and L_2. Each alignment is a hypothesis about where the network N_+ stores and uses the values of S_1 and w.

Interchange Interventions Finally, for each alignment, we would experimentally determine whether the neural representations at L_1 and L_2 have the same causal properties as S_1 and w. The basic experimental technique we use is an *interchange intervention*, in which a neural representation created during prediction on an input is interchanged with the representation created for a second

input [12]. We now show informally that this method can be used to prove that the causal model C_+ is a constructive abstraction of the neural network N_+ (see Appendix H for details).

We first intervene on the causal model. Consider two inputs $\mathbf{a}, \mathbf{a}' \in (\mathbb{N}_9)^3$ where $\mathbb{N}_9$ is the set of integers 0–9. Let $\mathbf{a} = (x, y, z)$ and $\mathbf{a}' = (x', y', z')$. Define

$$C_+^{S_1 \leftarrow \mathbf{a}'}(\mathbf{a}) = x' + y' + z \tag{1}$$

to be the output provided by C_+ when S_1, the variable representing the intermediate sum, is intervened on and set to the value $x' + y'$. Thus, for example, if $C_+(1, 2, 3) = 6$, and $\mathbf{a}' = (4, 5, 6)$, then $C_+^{S_1 \leftarrow \mathbf{a}'}(1, 2, 3) = 4 + 5 + 3 = 12$.

Next, we intervene on the neural network N_+. Let $\mathbf{D}$ be an embedding space that provides unique representations for $\mathbb{N}_9$, and consider two inputs $D = (D_x, D_y, D_z)$ and $D' = (D_{x'}, D_{y'}, D_{z'})$, where all D_i and $D_{i'}$ are drawn from $\mathbf{D}$. In parallel with (1), define

$$N_+^{L_1 \leftarrow D'}(D) \tag{2}$$

to be the output provided by N_+ processing the input D when the representation at location L_1 is replaced with the representation at location L_1 created when N_+ is processing the input D'. This process is depicted in Figure 1c.

With these two definitions, we can define what it means to test the hypothesis that N_+ computes $x + y$ at position L_1. Where $D_{\mathbf{a}}$ is an embedding for $\mathbf{a}$ and $D_{\mathbf{a}'}$ is an embedding for $\mathbf{a}'$, we test:

$$C_+^{S_1 \leftarrow \mathbf{a}'}(\mathbf{a}) = N_+^{L_1 \leftarrow D_{\mathbf{a}'}}(D_{\mathbf{a}}) \tag{3}$$

If this equality holds for all $\mathbf{a}$ and $\mathbf{a}'$, then we can conclude that, for every intervention on S_1, there is an equivalent intervention on L_1. If we can establish a corresponding claim for w and L_2, then we have shown that C_+ is a constructive abstraction of N_+, since the inputs' relationships are established by our embedding and there are no other interventions on C_+ to test.

Analysis Suppose that all of our intervention experiments verify our hypothesis that C_+ is a constructive abstraction of N_+ with variables S_1 and w aligned to neural representations at L_1 and L_2. How exactly does this explain network behavior?

Proving the constructive abstraction relation holds with this particular alignment tells us crucial information about the neural representations at L_1 and L_2, specifically (1) what information is encoded in the representations and (2) the causal role these representations play in network behavior.

Neural representations encode the values of the high-level variables they are aligned with. The location L_1 encodes the value $S_1 = x + y$ and the location L_2 encodes the value w. This is similar to what probing achieves. However, our method is crucially different from probing. In probing, information content is established through purely correlational properties, meaning a neural representation with no causal role in network behavior can be successfully probed. In causal abstraction analysis, information content is established through purely causal properties, ensuring that the neural representation is actually implicated in model behavior.

Neural representations play a parallel causal role to their aligned high-level variables. At the location L_1, D_x and D_y are composed to form a neural representation with content $x+y$ that is then composed with L_2 to create an output. The fact that S_1 doesn't depend on z tells us that while L_1 depends on D_z and representations at L_1 may even correlate with z, the information about z is not causally represented at L_1. At the location L_2, the value of z is simply repeated and then composed with L_1 to create a final output.

Compare this characterization of the causal roles of neural representations with that provided by attribution methods, which would assign L_1 and L_2 a single scalar value based on their contribution to the final output. This gives us no information about what representations are being composed together at L_1 and L_2, nor what representations are composed from representations at L_1 and L_2. (See Appendix E for a formal argument)

Our method assigns causally impactful information content, but also identifies the abstract causal structure along which representations are composed. It encompasses and improves on correlational and attribution methods.

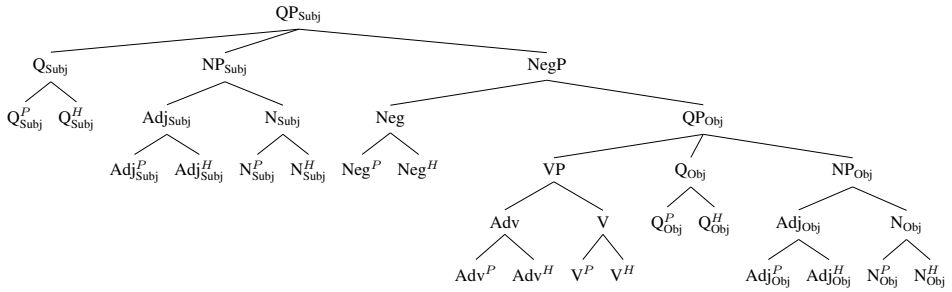


Figure 2: The causal structure of the high-level natural logic causal model C_{NatLog} that performs inference on MQNLI. The superscripts P and H stand for 'premise' and 'hypothesis' and the subscripts 'Obj' and 'Subj' stand for 'subject' and 'object'. The node labels are used to explain the experimental results in Section 5 and Section 5.2.

ε every ε baker ε ε ε eats ε no ε bread
contradiction
ε no angry baker ε ε ε eats ε no ε bread

ε every silly professor ε ε ε sells not every ε book
neutral
ε every silly professor ε ε ε sells not every ε chair

not every sad baker ε ε fairly admits not every odd idea
entailment
ε some ε baker does not ε admits ε no ε idea

(a) MQNLI examples. The ε token serves as padding (but still attended to by the model) and ensures a perfect alignment between both premises and hypotheses and across all examples. It is semantically an identity element.

Model	Train	Dev	Test
CBoW	88.04	54.18	53.99
TreeNN	67.01	54.01	53.73
CompTreeNN	99.65	80.17	80.21
BiLSTM	99.42	46.41	46.32
BERT	99.99	**88.25**	**88.50**

(b) MQNLI results. The first three models are from Geiger et al. 11, where the CompTreeNN is a task-specific model not suitable for general NLI and functions as an idealized upperbound. Our results show that BERT-based models can surpass this without such alignments.

Table 1: MQNLI examples (left) and MQNLI results (right).

4 The Natural Language Inference Task and Models

Multiply Quantified NLI Dataset The Multiply Quantified NLI (MQNLI) dataset of [11] contains templatically generated English-language NLI examples that involve very complex interactions between quantifiers, negation, and modifiers. We provide a few examples in Table 1a; the emptystring symbol ε ensures perfect alignments at the token level both between premises and hypotheses and across all examples.

The MQNLI examples are labeled using an algorithmic implementation of the natural logic of [22] over tree structures, and the dataset distribution includes a method for creating train/dev/test splits that vary in their difficulty. In the hardest setting, the train set is provably the minimal set of examples required to ensure that the dev and test sets can be perfectly solved by a simple symbolic model; in the easier settings, the train set redundantly encodes necessary information, which might allow a model to perform perfectly in assessment by memorization despite not having found a truly general solution. For a fuller review of the dataset, see Appendix B.

MQNLI is a fitting benchmark given our goals for a few reasons. First, we can focus on the hardest splits that can be generated, which will stress-test our NLI architectures in a standard behavioral way. Second, the MQNLI labeling algorithm is itself a causal model of the data-generating process. Figure 2 summarizes this model in tree form and it is presented in full detail in [11]. This allows us to rigorously assess whether a neural network has learned to implement variants of this causal model. The complexity of the MQNLI examples creates many opportunities to do this in linguistically interesting ways.

Models We evaluated two models on MQNLI: a randomly initialized multilayered Bidirectional LSTM (BiLSTM; [28]) and a BERT-based classifier model in which the English `bert-base`

parameters [8] are fine-tuned on the MQNLI train set. Output predictions are computed using the final representation above the [CLS] token. Models are trained to predict the relation of every pair of aligned phrases in Figure 2. Additional model and training details are given in Appendix C.

Results Table 1b summarizes the results of our BERT and BiLSTM models on the hardest fair generalization task [11] creates with MQNLI. We find that our BiLSTM model is not able to learn this task, and that our BERT model is able to achieve high accuracy. The only models in [11] able to achieve above 50% accuracy were task specific tree-structured models with the structure of the tree in Figure 2. Thus, our BERT-based model is the first general-purpose model able to achieve good performance on this hard generalization task.

A natural hypothesis is that the BERT-based model achieves this high performance *because* it has in effect induced some approximation to the tree-like structure of the data-generating process in its own internal layers. With the present methodology we are actually in a position to test this hypothesis.

5 A Case Study in Structural Neural Network Analysis

5.1 Causal Abstractions of Neural NLI models

Formulating Our Hypotheses We proceed just as we did for the simple motivating example in Section 3, except that we are now seeking to assess the extent to which the natural logic algebra in Figure 2 is a causal abstraction of the trained neural models in the above section.

The hallmark of Figure 2 is that it defines an alignment between premise and hypothesis at both lexical and phrasal levels. This permits us to run interchange interventions in a naturally compositional way. For a given non-leaf node N in Figure 2, let C^N_{NatLog} be a submodel of C_{NatLog} that computes the relation between the aligned phrases under N and uses them to compute the final output relation between premise and hypothesis . For example, let $C^{\text{NP}_{\text{Obj}}}_{NatLog}$ be the submodel of C_{NatLog} that computes the relation between the two aligned object phrases and then uses that relation in computing the final output relation between premise and hypothesis. We would like to ask whether our trained neural models also compute this relation between object noun phrases and use it to make a final prediction. We can pose this same question for other nodes which correspond to a pair of aligned subphrases.

Alignment Search For each N, we search for an alignment between a neural representations in N_{NLI} and the variable N in C^N_{NatLog}. In principle, any location in the network could be the right one for any causal model. Testing every hypothesis in this space would be intractable. Thus, for each C^N_{NatLog}, we consider a restricted class of locations based on the value of N:

- Q_{Subj}, Adj_{Subj}, N_{Subj}, Neg, Adv, V, Q_{Obj}, Adj_{Obj}, N_{Obj}: hidden representations above the two descendant leaf tokens.
- NP_{Subj}, VP, and NP_{Obj}: same but above the four descendant leaf tokens.
- QP_{Obj}: hidden representations above Q^P_{Obj} and Q^H_{Obj}.
- NegP: same but above Neg^P and Neg^H.
- All nodes (for BERT): same but above [CLS] and [SEP].

Interchange Interventions As in Section 3, we first focus on our high-level causal model, in this case Figure 2. Consider a non-leaf node N from Figure 2 and two input token sequences e and e' from MQNLI. Define

$$C^{N \leftarrow e'}_{NatLog}(e) \tag{4}$$

to be the output provided by the causal model C^N_{NatLog} when processing input e where the relation between the aligned subphrases under the node N is changed to the relation between those subphrases in e'. For example, simplifying for the sake of exposition, suppose e is (*every happy person*, *some happy baker*), which has output label entailment, and suppose e' is (*some happy baker*, *no* ϵ *baker*), which has output label contradiction. We wish to intervene on the noun phrase, so $N = \text{NP}$. In e, the noun phrase relation is reverse entailment; in e', it is entailment. Thus, $C^{\text{NP} \leftarrow e'}_{NatLog}(e)$ changes the object noun phrase relation in e to entailment while holding everything else about e constant. This results in the output label **neutral**.

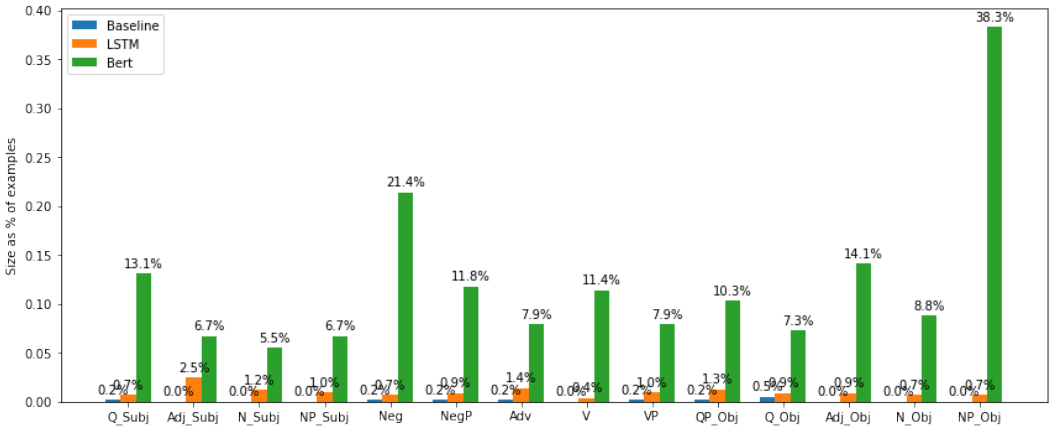


Figure 3: The largest subset of examples on which C^N_{NatLog} is an abstraction of an LSTM and BERT model both trained on MQNLI, and a baseline BERT model with no fine-tuning at all, where N is a non-leaf node of the tree in Figure 2. We record the size of such subsets as a percentage of the total 1000 examples. On this subset, we know that the neural models compute a representation of the relation between the aligned subphrases under N and use this information to make a final prediction.

Next, we consider interventions in a neural model N_{NLI}. Define

$$N^{L \leftarrow e'}_{NLI}(e) \tag{5}$$

to be the output provided by N_{NLI} processing the input e when the representation at location L is replaced with the representation at location L created when N_{NLI} is processing e'. This is exactly the process depicted in Figure 1a, except now the networks are the complex trained networks of Section 4.

Our hypothesis linking Figure 2 with a model N_{NLI} takes the same form as (3). The causal model C^N_{NatLog} is a constructive abstraction of N_{NLI} when, for some representation location L, it is the case that, for all MQNLI examples e and e', we have

$$C^{N \leftarrow e'}_{NatLog}(e) = N^{L \leftarrow e'}_{NLI}(e) \tag{6}$$

This asserts a correspondence between interventions on the representations at L in network N_{NLI} and interventions on the variable N in the causal model C^N_{NatLog}. If it holds, then N_{NLI} computes the relation between the aligned phrases under the node N and uses this information to compute the relation between the premise and hypothesis.

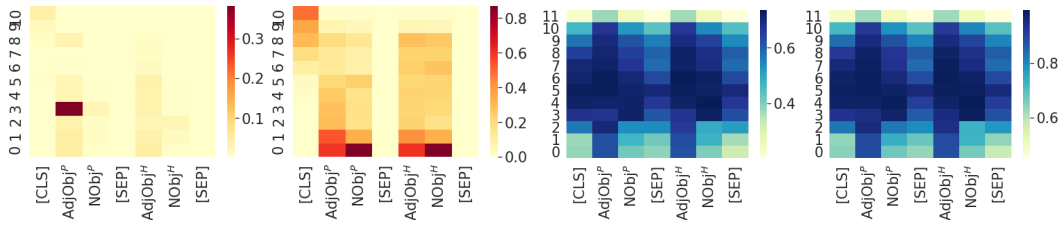
We call a pair of examples (e, e') *successful* if it satisfies equation (6), i.e., interventions in both the target causal model and neural model produce equal results. In addition, to isolate the causal impact of our interventions, we specifically focus on pairs (e, e') for which performing the intervention produces a different output value than without the intervention:

$$C^{N \leftarrow e'}_{NatLog}(e) \neq C^N_{NatLog}(e) \tag{7}$$

We call a pair (e, e') *impactful* if it satisfies (7).

Quantifying Partial Success Equation (6) universally quantifies over all examples. We do not expect this kind of perfect correspondence to emerge in practice for real problems: neural network training is often approximate and variable in nature, and even our best model does not achieve *perfect* performance. However, we can still ask how widely (6) holds for a given model. To do this, we seek to find the *largest subset* of MQNLI on which C^N_{NatLog} is an abstraction of our neural models, for each non-leaf node N in C_{NatLog}.

More specifically, considering each example in MQNLI as a vertex in a graph, we add an undirected edge between two examples e_i, e_j if and only if both the ordered pairs (e_i, e_j) and (e_j, e_i) satisfy (6). In other words, C^N_{NatLog} is an abstraction of a neural model on a subset of examples S of MQNLI if and only if all examples in S form a *clique*.



(a) Clique size. (b) Interchange success. (c) Probe selectivity. (d) Probe Accuracy.

Figure 4: Interchange intervention and probing results for the NP_{Obj} position. Vertical axes denote layers of BERT and horizontal axes denote the token position of hidden representations. The intervention success rates reported here are calculated based on intervention experiments with a change in the output label. Clique sizes are reported as % of 1000 examples. We did not conduct interchange experiments on the layer 11 of BERT because all representations on this layer except the one above [CLS] causally affected the result.

The number of intervention experiments we need to run scales quadratically with the number of inputs we consider, so we sample 1000 MQNLI examples, producing a total of $1000 \times 1000 = 1\text{M}$ ordered pairs. We only consider examples for which that the neural network outputs a correct label. For each node N and each of its corresponding neural network locations L, we perform interventions on all of these pairs.

Results and Analysis For each target causal model node N and neural network representation location L, we construct a graph as described above with 1000 examples as vertices and add an edge between two examples e_i and e_j if and only if *both* (e_i, e_j) and (e_j, e_i) are successful. We then find the largest clique in this graph with at least one impactful edge and record its size.

Figure 3 shows, for each causal model node N, the maximum size of cliques found among all neural locations. With this stricter *impactful* criterion (as opposed to simply using intervention success), our results show that, for almost all nodes N our target causal model, C^N_{NatLog} is indeed a causal abstraction of BERT, on a significant number of examples in our dataset. These subsets are much smaller for the BiLSTM model.

We will now explain what these successful results tell us about (1) the information encdoded by representations, and (2) the causal role these representations play in network behavior.

This analysis is similar to the analysis of our hypothetical addition example in Section 3, except for two crucial differences. First, for each variable N, we are hypothesizing that the causal model C^N_{NatLog} is an abstraction of N_{NLI}, where as in the addition example there was only one model. To account for this difference, we will be taking $N = \text{NP}_{\text{Obj}}$ as a paradigm case, with the results an analysis for other nodes in Appendix D. Second, we only achieved partial experimental success, whereas in the addition example we assumed complete success. Crucially, this means that the following analysis will only be valid on subsets of the input space on which the abstraction relation holds between N_{NLI} and $C^{\text{NP}_{\text{Obj}}}_{NatLog}$.

We visualize the results of our intervention experiments for the node NP_{Obj} in Figure 4. Observe that the alignment with the largest subset of inputs aligns the NP_{Obj} variable in $C^{\text{NP}_{\text{Obj}}}_{NatLog}$ with the neural representation on the fourth layer of BERT above the $\text{Adj}^P_{\text{Obj}}$ token. Because neural representations encode the value of their aligned variables and play a parallel causal role to their high-level variables, we know that, on this subset of input examples, at the fourth neural representation above the $\text{Adj}^P_{\text{Obj}}$ token, the four input embeddings for the object nouns and adjectives in the premise and hypothesis are composed to form a neural representation with information content of the relation between the object noun phrases in the premise and hypothesis. Then this representation is composed with the other input-embeddings to create an output representing the relation between the premise and hypothesis. We can perform a similar analysis for every other high-level variable, using the heatmaps from Appendix D.

We additionally would like to acknowledge that focusing on MQNLI could reinforce normative notions of how English is used and how reasoning in language works that could be exclusionary.

5.2 Comparison with Other Structural Analysis Methods

Probes Probes are fundamentally unable to measure causal properties of representations, instead determining the information content of representations using a modified measure of the correlation between neural representations and symbolic variables. We probed neural representation locations for the relation between aligned subexpressions on a subset of 12,800 randomly selected MQNLI examples. For a pair of aligned subexpressions below a node N in Figure 2, we probe the columns above the same set of restricted class of tokens as described in Section 5.1.

To evaluate these probes, we report accuracy as well as *selectivity* as defined by [14]: probe accuracy minus control accuracy, where *control accuracy* is the train set accuracy of a probe with the same architecture but trained on a control task to factor out probe success that can be attributed to the probe model itself. Our control task is to learn a random mapping from node types to semantic relations; see Appendix F for full details on how this task was constructed.

Figure 4 summarizes our probing results for $N = \text{NP}_{\text{Obj}}$, along with corresponding interchange intervention results for comparison. Probes tell us that information about the relation between the aligned noun phrases is encoded in nearly all of the locations we considered, and using the selectivity metric does not result in any qualitative change. In contrast, our intervention heatmaps indicate only a small number of locations store this information in a causally relevant way. Clearly, our intervention experiments are far more discriminating than probes. Appendix D provides examples involving other variables along with the intervention experiments, where the general trend of interchange interventions being more discriminate holds.

Integrated Gradients Attribution methods that estimate feature importance can measure causal properties of neural representations (see Appendix A.2), but a single feature importance method is an impoverished characterization of a the representations role in network behavior. Whereas our interchange interventions gave us high-level information about how a neural representation is composed and what it is composed into, attribution methods simply tell us "how much" a representation contributes to the network output on a give input. Moreover, intervention interchanges provide a rich, high-level characterization of causal structure on a space of inputs.

We use integrated gradients on our models to verify the intuitive hypothesis that if a premise and hypothesis differ by a single token, then the neural representations above that token should be more causally responsible for the network output than other representations. For example, given premise 'Every sleepy cat meows' and hypothesis 'Some hungry cat meows' the attributive modifier position is different and the rest are matched. The neural representations above the adjective tokens *sleepy* and *hungry* should be more important for the network output than others, because if those adjectives were the same, the example label would change from **neutral** to **entailment**.

We summarize the results of our integrated gradient experiments in Appendix E, where we confirm our intuitive hypothesis. The results are consistent with the results of our interchange interventions, as locations that had successful interchange interventions were locations where large attribution was assigned, and locations that did not have success were above locations where small attribution was assigned. However, our intervention experiments characterized the information content of neural representations and the high-level graph structure along which they are composed.

6 Conclusion

We have introduced a methodology for deriving interpretable causal explanations of neural network behaviors, grounded in a formal theory of causal abstraction. The methodology involves first *formulating a hypothesis* in the form of a high-level, interpretable causal model, then *searching for an alignment* between the neural network and the causal model, and finally *verifying experimentally* that the neural representations encode the same causal properties and information content as the corresponding components of the high-level causal model. As a case study demonstrating the feasibility of the approach, we analyzed neural models trained on the formidable MQNLI dataset. Guided by the intuition that success on this challenging task may call for a way of recapitulating the causal structure of the natural logic model that generates the MQNLI data, we were able to verify the hypothesis that a state-of-the-art BERT-based model partially realizes this structure, whereas baseline models that do not perform as well fail to do so. This suggestive case study demonstrates that our theoretically grounded methodology can work in practice.

References

[1] S. Beckers and J. Y. Halpern. Abstracting causal models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2678–2685, Jul. 2019. doi: 10.1609/aaai.v33i01.33012678. URL https://ojs.aaai.org/index.php/AAAI/article/view/4117.

[2] S. Beckers, F. Eberhardt, and J. Y. Halpern. Approximate causal abstractions. In R. P. Adams and V. Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 606–615, Tel Aviv, Israel, 22–25 Jul 2020. PMLR. URL http://proceedings.mlr.press/v115/beckers20a.html.

[3] A. Binder, G. Montavon, S. Bach, K. Müller, and W. Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. *CoRR*, abs/1604.00825, 2016. URL http://arxiv.org/abs/1604.00825.

[4] S. Bongers, P. Forré, J. Peters, B. Schölkopf, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *arXiv.org preprint*, arXiv:1611.06221v4 [stat.ME], Oct. 2020. URL https://arxiv.org/abs/1611.06221v4.

[5] K. Chalupka, F. Eberhardt, and P. Perona. Multi-level cause-effect systems. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 361–369, Cadiz, Spain, 09–11 May 2016. PMLR. URL http://proceedings.mlr.press/v51/chalupka16.html.

[6] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian. Neural network attributions: A causal perspective. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 981–990, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/chattopadhyay19a.html.

[7] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://www.aclweb.org/anthology/W19-4828.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

[9] Y. Elazar, S. Ravfogel, A. Jacovi, and Y. Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. In *Proceedings of the 2020 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Nov. 2020. doi: 10.18653/v1/W18-5426.

[10] A. Feder, N. Oved, U. Shalit, and R. Reichart. CausaLM: Causal Model Explanation Through Counterfactual Language Models. *Computational Linguistics*, pages 1–54, 05 2021. ISSN 0891-2017. doi: 10.1162/coli_a_00404. URL https://doi.org/10.1162/coli_a_00404.

[11] A. Geiger, I. Cases, L. Karttunen, and C. Potts. Posing fair generalization tasks for natural language inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4475–4485, Stroudsburg, PA, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1456. URL https://www.aclweb.org/anthology/D19-1456.

[12] A. Geiger, K. Richardson, and C. Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.16. URL https://www.aclweb.org/anthology/2020.blackboxnlp-1.16.

[13] M. Giulianelli, J. Harding, F. Mohnert, D. Hupkes, and W. Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5426. URL https://www.aclweb.org/anthology/W18-5426.

[14] J. Hewitt and P. Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://www.aclweb.org/anthology/D19-1275.

[15] D. Hupkes, S. Bouwmeester, and R. Fernández. Analysing the potential of seq-to-seq models for incremental interpretation in task-oriented dialogue. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 165–174, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5419. URL https://www.aclweb.org/anthology/W18-5419.

[16] T. F. Icard and L. S. Moss. Recent progress on monotonicity. *Linguistic Issues in Language Technology*, 9(7):1–31, January 2013.

[17] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[18] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un)reliability of saliency methods. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019.

[19] T. P. Lillicrap and K. P. Kording. What does it mean to understand a neural network?, 2019.

[20] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[21] B. MacCartney and C. D. Manning. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 193–200, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1654536.1654575.

[22] B. MacCartney and C. D. Manning. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands, Jan. 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W09-3714.

[23] C. Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

[24] M. Peters, M. Neumann, L. Zettlemoyer, and W.-t. Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1179. URL https://www.aclweb.org/anthology/D18-1179.

[25] A. Ravichander, Y. Belinkov, and E. Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance?, 2020.

[26] M. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-3020. URL `https://www.aclweb.org/anthology/N16-3020`.

[27] P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*. Association for Uncertainty in Artificial Intelligence (AUAI), Aug. 2017. URL `http://auai.org/uai2017/proceedings/papers/11.pdf`. *equal contribution.

[28] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[29] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Not just a black box: Learning important features through propagating activation differences. *CoRR*, abs/1605.01713, 2016. URL `http://arxiv.org/abs/1605.01713`.

[30] P. Soulos, R. T. McCoy, T. Linzen, and P. Smolensky. Discovering the compositional structure of vector representations with role learning networks. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 238–254, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.23. URL `https://www.aclweb.org/anthology/2020.blackboxnlp-1.23`.

[31] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, 12 2014.

[32] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.

[33] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/sundararajan17a.html`.

[34] I. Tenney, D. Das, and E. Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL `https://www.aclweb.org/anthology/P19-1452`.

[35] J. van Benthem. A brief history of natural logic. In M. Chakraborty, B. Löwe, M. Nath Mitra, and S. Sarukki, editors, *Logic, Navya-Nyaya and Applications: Homage to Bimal Matilal*, 2008.

[36] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020.

[37] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[38] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Our paper presents a general-purpose methodology for neural network analysis, with a motivating example involving arithmetic and a case-study involving natural language inference.
   (b) Did you describe the limitations of your work? [Yes] We are upfront throughout our analysis that we are making many idealizing assumptions for the sake of computational tractability.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] At the end of our analysis, we highlight that focusing on MQNLI could reinforce normative notions of how English is used and how reasoning in language works that could be exclusionary.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The theoretical results in our appendix are properly situated in the literature on causal abstraction, and our assumptions are clear.
   (b) Did you include complete proofs of all theoretical results? [Yes] While we do not prove any novel theorems in this paper, we did prove that our hypothetical arithmetic experiment and our actual NLI experiments prove the causal abstraction relations we claim hold in fact do hold.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Our supplemental material contains the necessary code to reproduce our experiments. If published, we will add a github repo.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] This information is all available in the appendix
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] No, we did not compute error bars for our tables. This would have been too computationally expensive.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We provide a rough estimate the compute and type of resources used in the appendix section on training details.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite the creators of the MQNLI dataset
   (b) Did you mention the license of the assets? [No] There are no licenses to mention
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Our code for this paper is included in supplementall material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The only dataset we used is publically available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] This is not applicable to this paper.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] We did not use crowdsourcing
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] We did not use crowdsourcing
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] We did not use crowdsourcing