Visual Causal Feature Learning

Krzysztof Chalupka KJCHALUP@CALTECH.EDU
Pietro Perona PERONA@CALTECH.EDU
Computation and Neural Systems
California Institute of Technology
Pasadena, CA 91125, USA

Frederick Eberhardt FDE@CALTECH.EDU
Humanities and Social Sciences
California Institute of Technology
Pasadena, CA 91125, USA

Abstract

We react to what we see, but what exactly is it that we react to? What are the visual causes of behavior? Can we identify such causes from raw image data? If the visual features are causes, how can we manipulate them? Here we provide a rigorous definition of the *visual cause* of a behavior that is broadly applicable to the visually driven behavior in humans, animals, neurons, robots and other perceiving systems. Our framework generalizes standard accounts of causal learning to settings in which the causal variables need to be constructed from micro-variables (raw image pixels in this case). We prove the Causal Coarsening Theorem, which allows us to gain causal knowledge from observational data with minimal experimental effort. The theorem provides a connection to standard inference techniques in machine learning that identify features of an image that *correlate* with, but may not *cause*, the target behavior. Finally, we propose an active learning scheme to learn a manipulator function that performs optimal manipulations on the image to automatically identify the visual cause of a target behavior. We illustrate our inference and learning algorithms in experiments based on both synthetic and real data. To our knowledge, our account is the first demonstration of true causal feature learning in the literature.

Keywords: Causal Inference, Causal Feature Learning, Bayesian Networks, Neural Networks, Computer Vision

1. Introduction

We present a theoretical framework and inference algorithms for visual causes in images. A visual cause is defined (more formally below) as a function (or *feature*) of raw image pixels that has a *causal effect* on the target behavior of a perceiving system of interest. We present three advances:

1. We provide a definition of the visual cause of a target behavior as a macro-variable that is constructed from the micro-variables (pixels) that make up the image space. The visual cause is distinguished from other macro-variables in that it contains all and only the causal information about the target behavior that is available in the image. We place the visual cause within the standard framework of causal graphical models (Spirtes et al., 2000; Pearl, 2009).

2. We prove the Causal Coarsening Theorem (5), which shows how observational data can be used to learn the visual cause with minimal experimental effort. It connects the present results to standard classification tasks in machine learning.

3. We describe a method to learn the manipulator function, which automatically performs perceptually optimal manipulations on the visual causes.

We illustrate our ideas using synthetic and real-data experiments. Python code that implements our algorithms and reproduces the experimental results will be available online at `http://vision.caltech.edu/~kchalupk/code/visual_causes.tar.gz`.

1.1 Previous Work

Our framework extends the theory of causal graphical models (Spirtes et al., 2000; Pearl, 2009) to a setting in which the input data consists of raw pixel (or other micro-variable) data. In contrast to the standard setting, in which the macro-variables in the statistical dataset already specify the candidate causal relata, the causal variables in our setting have to be constructed from the micro-variables they supervene on, before any causal relations can be established. We emphasize the difference between our method of causal feature *learning* and methods for causal feature *selection* (Guyon et al., 2007; Pellet and Elisseeff, 2008). The latter methods choose the best (under some causal criterion) features from a restricted set of plausible macro-variable candidates. In contrast, our framework efficiently searches the whole space of all the possible macro-variables that can be constructed from an image. Our approach derives its theoretical underpinnings from the theory of computational mechanics of Shalizi and Crutchfield (2001) (see also Shalizi, 2001), but supports a more explicitly causal interpretation than the original description. To our knowledge, this is the first account of true causal feature learning.

1.2 Visual Causal Feature Learning: an Example

Figure 1 presents a paradigmatic case study in visual feature learning, which we will use as a running example. In the model the contents of an image I are caused by binary hidden variables H_1 and H_2 such that if H_1 is on, I contains a vertical bar (v-bar[1]) at a random position, and if H_2 is on, I contains a horizontal bar (h-bar) at a random position. A target behavior $T \in \{0, 1\}$ is caused by H_1 and I, such that $T = 1$ is more likely whenever $H_1 = 1$ and whenever the image contains an h-bar.

We deliberately constructed this example such that the visual cause is very clearly identifiable: manipulating the presence of an h-bar in the image will influence the distribution of T. Thus, we can call the following function $C \colon \mathcal{I} \to \{0, 1\}$ the *causal feature* of I or the *visual cause* of T:

$$C(I) = \begin{cases} 1 & \text{if } I \text{ contains an h-bar} \\ 0 & \text{otherwise.} \end{cases}$$

The presence of a v-bar, on the other hand, is not a causal feature. Manipulating the presence of a v-bar in the image has no effect on the value of H_1 or T. Still, the presence of a v-bar is exactly as strongly correlated with the value of T (via the common cause H_1) as the presence of an h-bar is. We will call the following function $S \colon \mathcal{I} \to \{0, 1\}$ the *spurious correlate* of T in I:

$$S(I) = \begin{cases} 1 & \text{if } I \text{ contains a v-bar} \\ 0 & \text{otherwise.} \end{cases}$$

1. We take a v-bar (h-bar) to consist of a complete column (row) of black pixels.

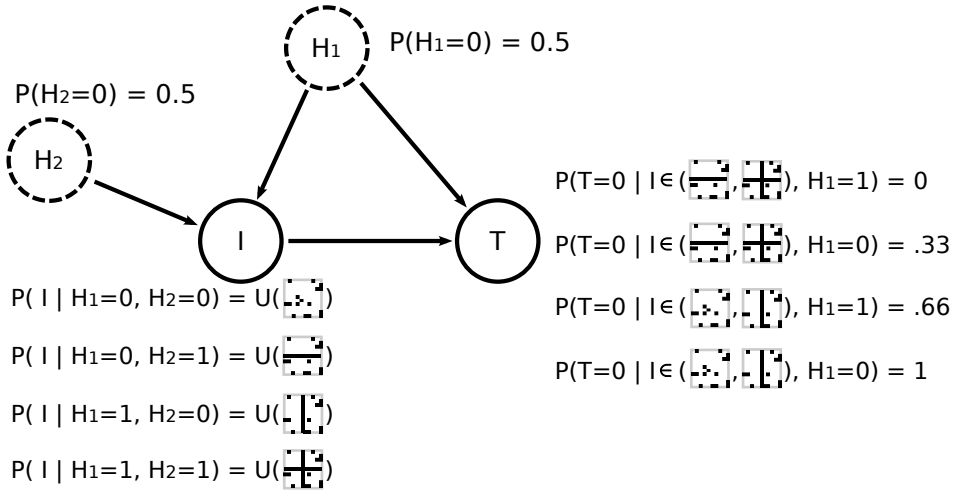


Figure 1: The generative model of our case study. Two binary hidden variables H_1 and H_2 toss unbiased coins. The contents of the image I depend on these variables as follows. If $H_1 = H_2 = 0$, I is chosen uniformly at random from all the images containing no v-bars and no h-bars. If $H_1 = 0, H_2 = 1$, I is chosen uniformly at random from images containing at least one h-bar but no v-bars. If $H_1 = 1$ and $H_2 = 0$, I is chosen uniformly at random from all the images containing at least one v-bar but no h-bars. Finally, if $H_1 = H_2 = 1$, I is chosen from images containing at least one v-bar and at least one h-bar. The distribution of the binary behavior T depends only on the presence of an h-bar in I and the value of H_1. In observational studies, $H_1 = 1$ iff I contains a v-bar. However, a *manipulation* of any specific image $I = i$ that introduces a v-bar (without changing H_1) will in general not change the probability of T occurring. Thus, T does *not* depend causally on the presence of v-bars in I.

Both the presence of h-bars and the presence of v-bars are good predictors of the target variable, but only one of them is a cause. Identifying the visual cause from the image thus requires the ability to distinguish among the correlates of the target variables those that are actually causal, even if the non-causal correlates are (possibly more) strongly correlated with the target.

One thing to note about this example is that the values of S and C stand in a bijective correspondence to the values of H_1 and H_2, respectively. Thus, learning S and C amounts here to learning to read out the two hidden variables from the image. This does not need to be the case in practice. The visual cause and the spurious correlate can be (random) functions of any number of the hidden variables, and can share the same hidden causes.

2. A Theory of Visual Causal Features

In our example the meaning of "The presence of an h-bar is the visual cause of T" is intuitive, as the model is constructed to have an easily describable visual cause. But the example does not provide a general theoretical account of what it takes to be a visual cause in the general case when we do not know

what the causally relevant pixel configurations are. In this section, we provide a general account of how the visual cause is related to the pixel data.

2.1 Visual Causes as Macro-variables

A visual cause is a high-level random variable that is a function (or feature) of the image, which in turn is defined by the random micro-variables that determine the pixel values. The functional relation between the image and the visual cause is, in general, surjective, though in principle it could be bijective. While we are interested in identifying the visual causes of a target behavior, the functional relation between the image pixels and the visual cause should not itself be interpreted as causal. Pixels do not *cause* the features of an image, they *constitute* them, just as the atoms of a table constitute the table (and its features). The difference between the causal and the constitutive relation is that the former requires the possibility of independent manipulation (at least to some extent), whereas by definition one cannot manipulate the visual cause without manipulating the image pixels.

The probability distribution over the visual cause is induced by the probability distribution over the pixels in the image and the functional mapping from the image to the visual cause. But since a visual cause stands in a constitutive relation with the image, we cannot without further explanation describe interventions on the visual cause in terms of the standard do-operation (Pearl, 2009). So our goal in this section is to define a macro-variable C, which contains all the causal information available in an image about a given behavior T, and define its manipulation. To make the problem approachable, we introduce two assumptions about the causal relation between the image and the behavior: (i) The target behavior T is subsequent to the image in time, and (ii) the behavior is not represented in the image. These assumptions allow us to avoid the possibility that the target behavior is a cause of features in the image, and thus avoid cycles in the underlying directed causal graph.

2.2 Generative Models: From Micro- to Macro-variables

Let $T \in \{0, 1\}$ represent a visual behavior.[2] Let $\mathcal{I}$ be a discrete space of all the images that can influence the target behavior (in our experiments in Section 4, $\mathcal{I}$ is the space of n-dimensional black-and-white images). We use the following generative model to describe the relation between the images and the target behavior: An image is generated by a finite set of unobserved discrete variables $H_1, \ldots, H_m$ (we write $\mathbf{H}$ for short). The target behavior is then determined by the image and possibly a subset of variables $\mathbf{H}_c \subseteq \mathbf{H}$ that are confounders of the image and the target behavior:

$$P(T, I) = \sum_{\mathbf{H}} P(T \mid I, \mathbf{H}) P(I \mid \mathbf{H}) P(\mathbf{H}) = \sum_{\mathbf{H}} P(T \mid I, \mathbf{H}_c) P(I \mid \mathbf{H}) P(\mathbf{H}). \tag{1}$$

Independent noise that may contribute to the target behavior is marginalized and omitted for the sake of simplicity in the above equation. The noise term incorporates any hidden variables which influence the behavior but stand in no causal relation to the image. Such variables are not directly relevant to the problem. Fig. 2 shows this generative model.

Under this model, we can define an *observational partition* of the space of images $\mathcal{I}$ that groups images into classes that have the same conditional probability $P(T|I)$:

2. An extension of the framework to non-binary, discrete T is easy but complicates the notation significantly

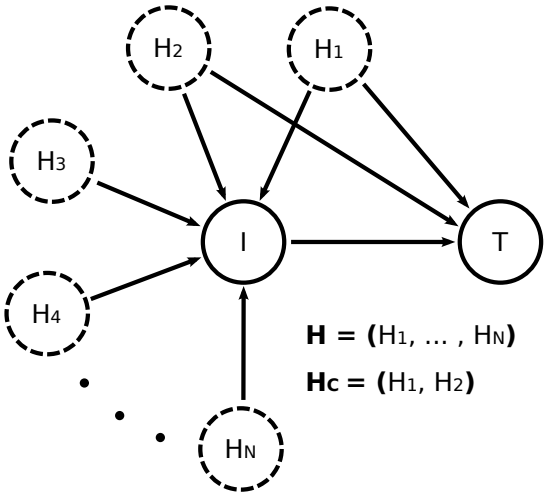


Figure 2: A general model of visual causation. In our model each image I is caused by a number of hidden non-visual variables H_i. The image itself is the only observed cause of a target behavior T. In addition, a (not necessarily proper) subset of the hidden variables can be a cause of the target behavior. These confounders create visual "spurious correlates" of the behavior in I.

Definition 1 (Observational Partition, Observational Class) *The* observational partition $\Pi_o(T, \mathcal{I})$ *of the set $\mathcal{I}$ w.r.t. behavior T is the partition induced by the equivalence relation $\sim$ such that $i \sim j$ if and only if $P(T \mid I = i) = P(T \mid I = j)$. We will denote it as Π_o when the context is clear. A cell of an observational partition is called an* observational class.

In standard classification tasks in machine learning, the observational partition is associated with class labels. In our case, two images that belong to the same cell of the observational partition assign equal *predictive* probability to the target behavior. Thus, knowing the observational class of an image allows us to predict the value of T. However, the predictive probability assigned to an image does not tell us the *causal* effect of the image on T. A familiar example that distinguishes causation from mere prediction is the barometer. A barometer is widely taken to be an excellent predictor of the weather. But changing the barometer needle does not cause an improvement of the weather. It is not a (visual or otherwise) cause of the weather. In contrast, seeing a particular barometer reading may well be a *visual cause* of whether we pack an umbrella.

Our notion of a visual cause depends on the ability to manipulate the image.

Definition 2 (Visual Manipulation) *A* visual manipulation *is the operation $man(I = i)$ that changes (the pixels of) the image to image $i \in \mathcal{I}$, while not affecting any other variables (such as $\mathbf{H}$ or T). That is, the manipulated probability distribution of the generative model in Eq.* (1) *is given by*

$$P(T \mid man(I = i)) = \sum_{\mathbf{H}_c} P(T \mid I = i, \mathbf{H}_c) P(\mathbf{H}_c). \tag{2}$$

The manipulation changes the values of image pixels, but does not change the underlying "world", represented in our model by the H_i that generated the image. Formally, the manipulation is similar to the

do-operator for standard causal models. However, we here reserve the do-operation for interventions on causal *macro*-variables, such as the visual cause of T. We discuss the distinction in more detail below.

We can now define the *causal partition* of the image space (with respect to the target behavior T) as:

Definition 3 (Causal Partition, Causal Class) *The causal partition $\Pi_c(T, \mathcal{I})$ of the set $\mathcal{I}$ w.r.t. behavior T is the partition induced by the equivalence relation $\sim$ defined on $\mathcal{I}$ such that $i \sim j$ if and only if $P(T \mid man(I = i)) = P(T \mid man(I = j))$ for $i, j \in \mathcal{I}$. When the image space and the target behavior are clear from the context, we will indicate the causal partition by Π_c. A cell of a causal partition is called a* causal class.

The underlying idea is that images are considered causally equivalent with respect to T if they have the same causal effect on T. Given the causal partition of the image space, we can now define the visual cause of T:

Definition 4 (Visual Cause) *The* visual cause C *of a target behavior T is a random variable whose value is the causal class of I.*

The visual cause is thus a function over $\mathcal{I}$, whose values correspond to the post-manipulation distributions $C(i) = P(T \mid man(I = i))$. We will write $C(i) = c$ to indicate that the causal class of image $i \in \mathcal{I}$ is c, or in other words, that in image i, the visual cause C takes value c. Knowing C allows us to predict the effects of a visual manipulation $P(T \mid man(I = i))$, as long as we have estimated $P(T \mid man(I = i_k^*))$ for one representative i_k^* of each causal class k.

2.3 The Causal Coarsening Theorem

We are now ready to state our main theorem, which relates the causal and observational partitions for a given $\mathcal{I}$ and T. It turns out that in general the causal partition is a coarsening of the observational partition. That is, the causal partition aligns with the observational partition, but the observational partition subdivides some of the causal classes.

Theorem 5 (Causal Coarsening) *Among all the generative distributions of the form shown in Fig. 2 which induce a given observational partition Π_o, almost all induce a causal partition Π_c that is a coarsening of the Π_o.*

Throughout this article, we use "almost all" to mean "all except for a subset of Lebesgue measure zero". The Causal Coarsening Theorem states that no matter what the observational partition looks like, among all the distributions that could have generated it, the distributions for which the causal partition is not a coarsening of the observational partition are extremely rare. We can thus, in general, safely assume that the corresponding causal partition is aligned with any given observational partition but merges some of the observational classes. Fig. 3 illustrates the relation between the causal and the observational partition implied by the theorem.

We prove the Causal Coarsening Theorem (5) in Appendix A using a technique extending that of Meek (1995). Two points are worth noting here: First, CCT confirms the intuition that the visual causes of a behavior do not contain all the information in the image that could help predict the behavior. Such information, though not itself a cause of the behavior, can be informative about the state of

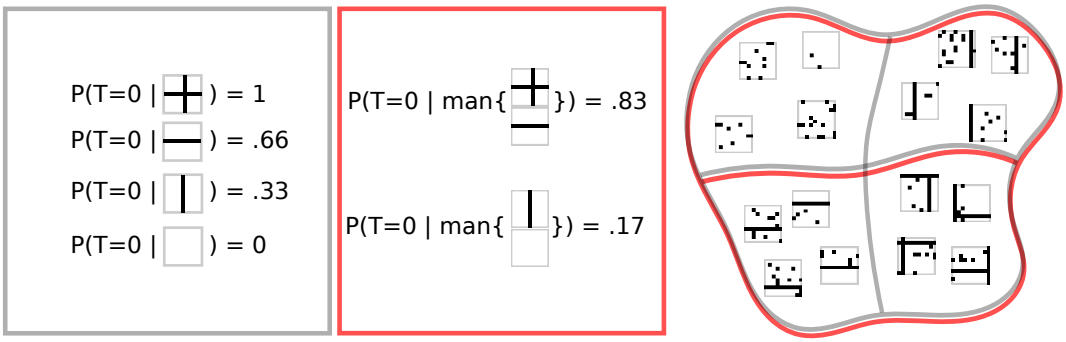


Figure 3: The Causal Coarsening Theorem. The observational probabilities of T (gray frame) induce an observational partition on the space of all the images (right, observational partition in gray). The causal probabilities (red frame) induce a causal partition, indicated on the right in red. The Causal Coarsening Theorem allows us to expect that the observational partition is a coarsening of the causal partition. The observational and causal probabilities correspond to the generative model shown in Fig. 1.

other non-visual causes of the target behavior. Second, CCT allows us to take any classification problem in which the data is divided into observational classes, and assume that the causal labels do not change within each observational class. This will help us develop efficient causal inference algorithms in Section 3.

2.4 Visual Causes in a Causal Model consisting of Macro-variables

We can now simplify our generative model by omitting all the information in I unrelated to behavior T. Assume that the observational partition Π_o^T refines the causal partition Π_c^T. Each of the causal classes $c_1, \cdots, c_K$ delineates a region in the image space $\mathcal{I}$ such that all the images belonging to that region induce the same $P(T \mid man(I))$. Each of those regions—say, the k-th one—can be further partitioned into sub-regions $s_1^k, \cdots, s_{M_k}^k$ such that all the images in the m-th sub-region of the k-th causal region induce the same observational probability $P(T \mid I)$. By assumption, the observational partition has a finite number of classes, and we can arbitrarily order the observational classes within each causal class. Once such ordering is fixed, we can assign an integer $m \in \{1, 2, \cdots, M_k\}$ to each image i belonging to the k-th causal class such that i belongs to the m-th observational class among the M_k observational classes contained in c_k. By construction, this integer explains all the variation of the observational class within a given causal class. This suggests the following definition:

Definition 6 (Spurious Correlate) *The* spurious correlate S *is a discrete random variable whose value differentiates between the observational classes contained in any causal class.*

The spurious correlate is a well-defined function on $\mathcal{I}$, whose value ranges between 1 and $\max_k M_k$. Like C, S is a macro-variable constructed from the pixels that make up the image. C and S together contain all and only the visual information in I relevant to T, but only C contains the causal information:

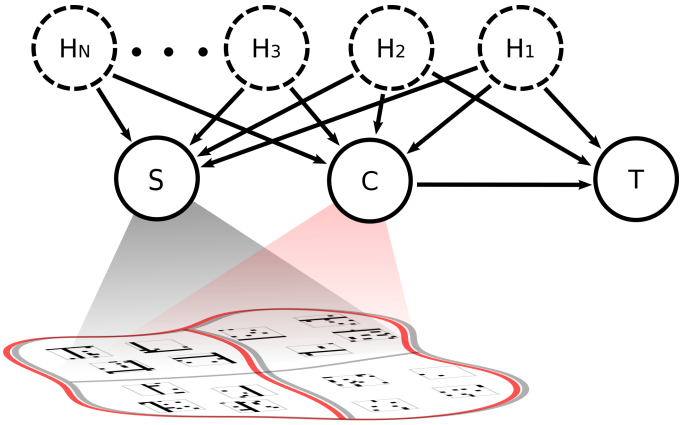


Figure 4: A macro-variable model of visual causation. Using our theory of visual causation we can aggregate the information present in visual micro-variables (image pixels) into the visual cause C, spurious correlate S, and irrelevant noise N. According to Theorem (7), C and S contain all the information about T available in I.

Theorem 7 (Complete Macro-variable Description) *The following two statements hold for C and S as defined above:*

1. $P(T \mid I) = P(T \mid C, S)$.
2. *Any other variable X such that $P(T \mid I) = P(T \mid X)$ has Shannon entropy $H(X) \geq H(C, S)$.*

We prove the theorem in Appendix A. It guarantees that C and S constitute the smallest-entropy macro-variables that encompass all the information about the relationship between T and I. Figure 4 shows the relationship between C, S and T, the image space $\mathcal{I}$ and the observational and causal partitions schematically. C is now a cause of T, S correlates with T due to the unobserved common causes $\mathbf{H}_C$, and any information irrelevant to T is pushed into the independent noise variables (commonly not shown in graphical representations of structural equation models).

The macro-variable model lends itself to the standard treatment of causal graphical models described in Pearl (2009). We can define interventions on the causal variables $\{C, S, T\}$ using the standard *do*-operation. The *do*-operator only sets the value of the intervened variable to the desired value, making it independent of its causes, but it does not (directly) affect the other variables in the system or the relationships between them. This assumption about interventions on variables in causal models is known as *modularity* (Pearl, 2009). However, unlike the standard case where causal variables are separated in location (for example, *smoking* and *lung cancer*), the causal variables in an image may involve the same pixels (for example, C may be the average brightness of the image, whereas S indicates the presence or absence of particular shapes in the image). An intervention on a causal variable using the *do*-operator thus requires that the underlying manipulation of the image respects the state of the other causal variables:

Definition 8 (Causal Intervention on Macro-variables) *Given the set of macro-variables $\{C, S\}$ that take on values $\{c, s\}$ for an image $i \in \mathcal{I}$, an intervention $do(C = c')$ on the macro-variable C is given by*

the manipulation of the image space $man(I = i')$ *such that* $C(i') = c'$ *and* $S(i') = s$*. The intervention* $do(S = s')$ *is defined analogously as the change of the underlying image that keeps the value of* C *constant.*

Definition 8 can easily be extended to a setting in which multiple behaviors $T_1, \cdots, T_M$ are caused by the same I. Each behavior T_m can then have a different cause C_m and spurious correlate S_m, and we define the intervention on C_m as a change in the underlying image that keeps all the other macro-variables constant.

In some cases it can be impossible to manipulate C to a desired value without changing other variables of interest. This is not a problem: The manipulation range of causal variables is often restricted. For example, in medical experiments it is often impossible to increase medicine dosage arbitrarily, as this could cause the death of a patient and thus influence multiple other variables in the system.

3. Causal Feature Learning: Inference Algorithms

Given the theoretical specification of the concepts of interest in the previous section, we can now develop algorithms to learn C, the visual cause of a behavior. In addition, knowledge of C will allow us to specify a *manipulator function*: a function that, given any image, can return a maximally similar image with the desired causal effect.

Definition 9 (Manipulator Function) *Let* C *be the causal variable of* T *and* d *a metric on* $\mathcal{I}$*. The* manipulator function *of* C *is a function* $M_C \colon \mathcal{I} \times \mathcal{C} \to \mathcal{I}$ *such that* $M_C(i, k) = \arg\min_{\hat{i} \in C^{-1}(k)} d(i, \hat{i})$ *for any* $i \in \mathcal{I}, k \in \mathcal{C}$*. In case* $d(i, .)$ *has multiple minima, we group them together into one equivalence class and leave the choice of the representative to the manipulator function.*

The manipulator searches for an image closest to I among all the images with the desired causal effect k. The meaning of "closest" depends on the metric d and is discussed further in Section 3.2 below. Note that the manipulator function can find candidates for the image manipulation underlying the desired causal manipulation $do(C = c)$, but it does not check whether other variables in the system (in particular, the spurious correlate) remain in fact unchanged. Using the closest possible image with desired causal effect is a heuristic approach to fulfilling that requirement.

There are several reasons we might want such a manipulator function:

1. If our goal is to perform causal manipulations on images, the manipulator functions offers an automated solution.

2. A manipulator that uses a given C and produces images with the desired causal effect provides strong evidence that C is indeed the visual cause of the behavior.

3. Using the manipulator function we can enrich our dataset with new datapoints, in hope of achieving better generalization on both the causal and predictive learning tasks. We might know there exist multiple behaviors $T_1, \cdots, T_N$ that depend on I, but we want to manipulate the image to causally influence only one of them (say T_1). If we lack any information about the possible causes of the other behaviors, the best we can do to avoid influencing them is to find the manipulation that changes the image *as little as possible*.

The problem of visual causal feature learning can now be posed as follows: Given an image space $\mathcal{I}$ and a metric d, learn C—the visual cause of T—and the manipulator M_C.

3.1 Causal Effect Prediction

A standard machine learning approach to learning the relation between I and T would be to take an *observational dataset* $\mathcal{D}_{obs} = \{(i_k, P(T \mid i_k))\}_{k=1,\cdots,N}$ and learn a predictor f whose training performance guarantees a low test error (so that $f(i^*) \approx P(T \mid i^*)$ for a test image i^*). In causal feature learning, low test error on observational data is insufficient; it is entirely possible that $\mathcal{D}$ contains spurious information useful in predicting test labels which is nevertheless not causal. That is, the prediction may be highly accurate for observational data, but completely inaccurate for a prediction of the effect of a manipulation of the image. However, we can use the Causal Coarsening Theorem to obtain a causal dataset from the observational data, and then train a predictor on that dataset. Algorithm 1 uses this strategy to learn a function C that, presented with any image $i \in \mathcal{I}$, returns $C(i) \approx P(T \mid man(I = i))$. We use a fixed neural network architecture to learn C, but any differentiable hypothesis class could be susbtituted instead. Differentiability of C is necessary in Section 3.2 in order to learn the manipulator function.

Algorithm 1: Causal Predictor Training

input : $\mathcal{D}_{obs} = \{(i_1, p_1 = p(T \mid i_1)), \cdots (i_N, p_N = p(T \mid i_N)\}$ – observational data
$\mathcal{P} = \{P_1, \cdots, P_M\}$ – the set of observational classes (so that $\forall_k t_k \in \mathcal{P}$)
`Train`: data $\mapsto$ NN – a neural network training algorithm

output: $C : \mathcal{I} \to [0, 1]$ – the causal variable

1 Pick $\{i_{k_1}, \cdots, i_{k_M}\} \subset \{i_1, \cdots, i_N\}$ s.t. $p_{k_m} = P_m$;
2 Estimate $\hat{C}_m \leftarrow P(T \mid man(I = i_{k_m}))$ for each m;
3 For all k let $\hat{C}(i_k) \leftarrow \hat{C}_m$ if $p_k = P_m$;
4 $\mathcal{D}_{csl} \leftarrow \{(i_1, \hat{C}(i_1)), \cdots, (i_N, \hat{C}(i_N))\}$;
5 $C \leftarrow \texttt{Train}(\mathcal{D}_{csl})$;

In Step 1 the algorithm picks a representative member of each observational class. The Causal Coarsening Theorem tells us that the causal partition coarsens the observational one. That is, in principle (ignoring sampling issues) it is sufficient to estimate $\hat{C}_m = P(T \mid man(I = i_{k_m}))$ for just one image in an observational class m in order to know that $P(T \mid man(I = i)) = \hat{C}_m$ for any other i in the same observational class. The choice of the experimental method of estimating the causal class in Step 2 is left to the user and depends on the behaving agent and the behavior in question. If, for example, T represents whether the spiking rate of a recorded neuron is above a fixed threshold, estimating $P(T \mid man(I = i))$ could consist of recording the neuron's response to i in a laboratory setting multiple times, and then calculating the probability of spiking from the finite sample. The causal dataset created in Step 4 consists of the observational inputs and their causal classes. The causal dataset is acquired through $\mathcal{O}(N)$ experiments, where N is the number of observational classes. The final step of the algorithm trains a neural network that predicts the causal labels on unseen images. The choice of the method of training is again left to the user.

3.2 Causal Feature Manipulation

Once we have learned C we can use the causal neural network to create synthetic examples of images as similar as possible to the originals, but with a different causal label. The meaning of "as similar as

possible" depends on the image metric d (see Definition 9). The choice of d is task-specific and crucial to the quality of the manipulations. For example, let d be the discrete metric,

$$d(i,\hat{\imath}) = \begin{cases} 0 & \text{if } i = \hat{\imath} \\ 1 & \text{otherwise} \end{cases}.$$

Under this metric, a valid (but rather uninteresting) manipulator is

$$M_C(i,k) = \hat{\imath}_k \quad \text{for some} \quad \hat{\imath}_k \in C^{-1}(k),$$

where $\hat{\imath}_k$ is any member of the relevant causal class. A more useful metric could be induced by an L_p norm (so that $d(i,\hat{\imath}) = \|i - \hat{\imath}\|_p$) of which L_1 and L_2 are commonly used in computer vision applications (see Datta et al., 2008, for a discussion on these and other simple image metrics). Another popular choice has been the Earth Mover's Distance (Rubner et al., 2000), which measures the distance between two histogram distributions of image statistics. If we have background knowledge on the type of features that the behaving agent in question considers important in images, we can use a feature function ϕ to map the images into a useful feature space, with $d(i,\hat{\imath}) = \|\phi(i) - \phi(\hat{\imath})\|$. Learned image representations, such as Stacked Autoencoders and Deep Boltzmann Machines, are able to capture natural image statistics useful in discrimination tasks (Bengio et al., 2013), and representing images in such learned feature spaces might improve the causal manipulation results. A similar class of metrics is induced by positive-definite kernels that shift members of $\mathcal{I}$ into an implicit high-dimensional vector space (Schölkopf and Smola, 2002). For a positive-definite kernel k, $d(i,\hat{\imath}) = \sqrt{k(i-\hat{\imath}, i-\hat{\imath})}$ is a metric on $\mathcal{I}$. Image kernels have been used with considerable success in augmenting visual classification algorithms (Harchaoui and Bach, 2007; Grauman and Darrell, 2007; Bosch et al., 2007; Vishwanathan, 2010).

Algorithm 2 proposes one way to learn the manipulator function using a simple manipulation procedure that approximates the requirements of Definition 9 up to local minima. The algorithm (illustrated in Fig. 5) starts off by training a causal neural network in Step 2. If only observational data is available, this can be achieved using Algorithm 1. Next, it randomly chooses a set of images to be manipulated, and their target post-manipulation causal labels. The loop that starts in Step 6 then takes each of those images and searches for the image that, among the images with the same desired causal class, is closest to the original image. Note that the causal class boundaries are defined by the current causal neural net C. Since C is in general a highly nonlinear function and it can be hard to find its inverse sets, we use an approximate solution. The algorithm thus finds the minimum of a weighted sum of $|C(j) - \hat{c}_{iter,k}|$ (the difference of the output image j's label and the desired label $\hat{c}_{iter,k}$) and $d(i_{iter,k}, j)$ (the distance of the output image j from the original image $i_{iter,k}$).

At each iteration, the algorithm performs Q manipulations and causal queries to the agent, which result in new datapoints $(\hat{\imath}_{iter,1}, agentq(\hat{\imath}_{iter,1})), \cdots, (\hat{\imath}_{iter,Q}, agentq(\hat{\imath}_{iter,Q}))$. It's natural to claim that the manipulator performs well if $agentq(\hat{\imath}_{iter,k}) \approx \hat{c}_{iter,k}$ for many k, which means the target causal labels agree with the true causal labels. We thus define the *manipulation error* of the $iter$th iteration $MErr_{iter}$ as

$$MErr_{iter} = \frac{1}{Q}\sum_{k=1}^{Q} |agentq(\hat{\imath}_{iter,k}) - \hat{c}_{iter,k}|. \tag{3}$$

While it is important that our manipulations are accurate, we also want them to be minimal. This suggests considering the *average manipulation distance*

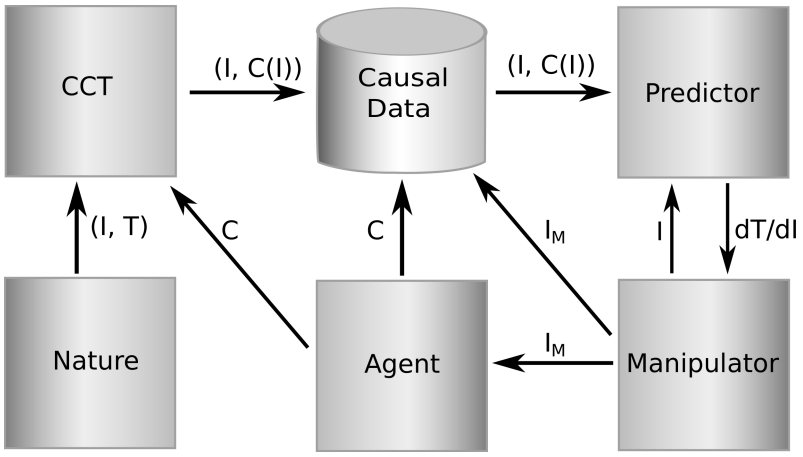


Figure 5: A flowchart of Algorithm 2. Nature creates an observational dataset of images I annotated with behaviors T. Using the Causal Coarsening Theorem and a minimal number of queries to the Agent, we transform the observational data into a causal dataset. The Causal Data is then used to train a Predictor (in our case, a multi-layer perceptron) that predicts the causal label $C(i)$ for any image i. Next, a Manipulator uses the Predictor's gradient to manipulate a chosen set of images. These images serve as experimental causal queries to the Agent. Finally, the manipulated images and their correct causal labels are added to the Causal Dataset, which completes the loop. The manipulation error $MErr$ counts the difference between the current Predictor's estimates and the Agent's responses $C(i_M)$.

Algorithm 2: Manipulator Function Learning

input : $d\colon \mathcal{I} \times \mathcal{I} \to \mathbb{R}_+$ – a metric on the image space
$\mathcal{D}_{csl} = \{(i_1, c_1), \cdots (i_N, c_N)\}$ – causal data
$\mathcal{C} = \{C_1, \cdots, C_M\}$ – the set of causal classes (so that $\forall_i c_i \in \mathcal{C}$)
`Train`: data $\mapsto$ NN – a neural network training algorithm
nIters – number of manipulation experiment iterations
Q – number of manipulated queries per iteration
alpha – manipulation tuning parameter
`agentq`: $\mathcal{I} \to \mathcal{C}$ – an oracle that obtains $P(T \mid do(I))$

output: $M_C\colon \mathcal{I} \times \mathcal{C} \to \mathcal{I}$ – the manipulator function

for iter $\leftarrow 1$ **to** nIters **do**
  $C \leftarrow \texttt{Train}(\mathcal{D}_{csl})$;
  Choose manipulation starting points $\{i_{\text{iter},1}, \cdots, i_{\text{iter,Q}}\}$ at random from $\mathcal{D}_{csl}$;
  Choose manipulation targets $\{\hat{c}_{\text{iter},1}, \cdots, \hat{c}_{\text{iter,Q}}\}$ such that $\hat{c}_{\text{iter},k} \neq c_{\text{iter},k}$;
  for $k \leftarrow 1$ **to** Q **do**
    $\hat{i}_{\text{iter},k} \leftarrow \underset{j \in \mathcal{I}}{\operatorname{argmin}} \; (1 - \text{alpha})|C(j) - \hat{c}_{\text{iter},k}| + \text{alpha}\; d(j, i_{\text{iter},k})$;
  end
  $\mathcal{D}_{csl} \leftarrow \mathcal{D}_{csl} \cup \{(\hat{i}_{\text{iter},1}, \texttt{agentq}(\hat{i}_{\text{iter},1})), \cdots, (\hat{i}_{\text{iter,Q}}, \texttt{agentq}(\hat{i}_{\text{iter,Q}}))\}$;
end

$$MDist_{iter} = \frac{1}{Q} \sum_{k=1}^{Q} d(I_{iter,k}, \hat{i}_{iter,k}). \tag{4}$$

A natural variant of Algorithm 2 is to set $nIters$ to a large integer and break the loop when one or both of these performance criteria reaches a desired value.

The algorithm is inspired by an active learning technique called uncertainty sampling (Lewis and Gale, 1994) with a variation on density weighing (Settles and Craven, 2008). Uncertainty sampling—and in particular least-confidence sampling, which we use—is a general active learning technique in which a dataset is updated with datapoints whose predicted labels have largest uncertainty under the current classifier. These are precisely the points that lie on the decision boundaries. Density weighing suggests that points from more densely populated input regions should be queried more often.

Settles (2010) argued that different kinds of uncertainty sampling can be variously effective for different tasks. In particular, for multi-class classification some techniques might be more efficient than least-confidence sampling for improving the classifier's performance. In our case however, least-confidence sampling is very natural as it leads to training a classifier that works well for the manipulation procedure described below.

4. Experiments

In order to illustrate the concepts presented in this article we perform two causal feature learning experiments. The first experiment, called GRATING, uses observational and causal data generated by the

model from Section 1.2. The GRATING experiment confirms that our system can learn the ground truth cause and ignore the spurious correlates of a behavior. The second experiment, MNIST, uses images of hand-written digits (LeCun et al., 1998) to exemplify the use of the manipulator function on slightly more realistic data: in this example, we transform an image into a maximally similar image with another class label.

We deliberately chose problems that are simple from the computer vision point of view. Our goal is to develop the theory of visual causal feature learning and show that it has feasible algorithmic solutions; we are at this point not engineering advanced computer vision systems.

4.1 The GRATING Experiment

In this experiment we generate data using the model of Fig. 1, with two minor differences: H_1 and H_2 only induce one v-bar or h-bar in the image and we restrict our observational dataset to images with only about 3% of the pixels filled with random noise. Both restrictions increase the clarity of presentation. Example images are shown in most figures throughout the article, including Fig. 7 which contains the experiment's results.

We use Algorithms 1 and 2 (with minor modifications imposed by the binary nature of the images) to learn the visual cause of behavior T. A Python code package that reproduces the experiments and contains the details of the implementation is available at `http://vision.caltech.edu/~kchalupk/code/visual_causes.tar.gz`. Fig. 6 shows the progress of the training process. The first step (not shown in the figure) is to use the Causal Coarsening Theorem to learn the causal labels on the observational data. We then train a simple neural network (a fully connected network with one hidden layer of 100 units) on this data. The same network is used on Iteration 1 to create new manipulated exemplars. After about ten iterations of the manipulator learning loop (Algorithm 2, Steps 2– 8), the manipulation error (Eq. 3) reaches about 10%, and follows a decreasing trend throughout iteration 20. Fig. 7 illustrates the difference between the manipulator on Iteration 1 (which fails almost 40% of the time) and Iteration 20, where the error is about 6%. The fully-trained manipulator correctly learned to manipulate the presence of the h-bar to cause changes in T, and ignores the v-bar that is strongly correlated with the behavior but does not cause it.

4.2 The MNIST Experiment

In this experiment we start off with the MNIST dataset of handwritten digits. In our terminology, this is already a causal dataset: the labels are assigned in an experimental setting, not "in nature".[3] In fact, most vision datasets that consist of images and class labels can be considered causal datasets: the labeling process assumes there are no non-visual confounders that influence the labeler's decision as to which class an image belongs to.

Consider the following binary human behavior: $T = 1$ if a human observer answers affirmatively to the question "Does this image contain the digit '7'?", $T = 0$ if the observer judges that the image does not contain the digit '7'. For simplicity we will assume that for any image either $P(T = 1 \mid man(I)) = 0$ or $P(T = 1 \mid man(I)) = 1$, although this is not precisely true: some images in Fig. 9 (explained below) are rather ambiguous. Our task is to learn the manipulator function that will take any image and modify

3. To be perfectly correct, MNIST is quite special in comparison with other computer vision datasets in that the digits were drawn manually by high-school students and government employees as requested by the dataset creators; this is different from the usual paradigm of presenting an image to be labeled to a worker.

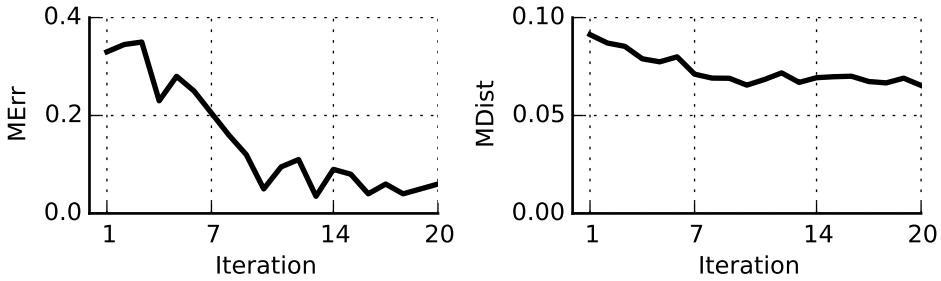


Figure 6: Manipulator learning for GRATING. The plots show the progress of our manipulator function learning algorithm over ten iterations of experiments for the GRATING problem. The manipulation error decreases quickly with progressing iterations, whereas the manipulation distance stays close to constant.

it minimally such that it will become a passable '7' if it was not before, or will stop resembling a '7' if it did originally.

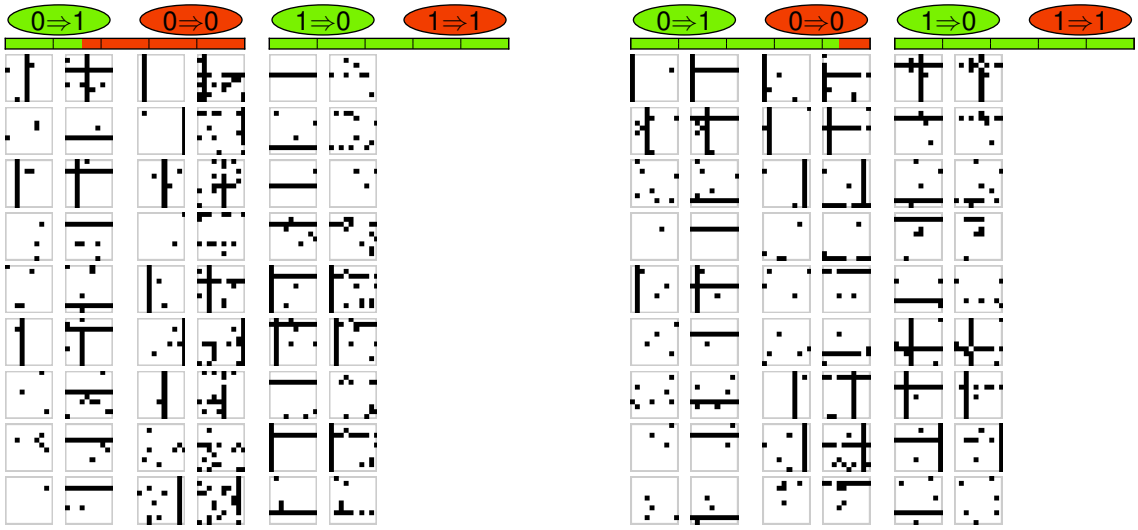
Apart from being an exercise in learning a manipulator function, this example investigates an interesting property of neural networks noticed by Szegedy et al. (2014). They noticed that for a given neural network that achieves very low test error on MNIST, a '9' (for example) can be easily modifed to be classified as '7' but be perceptually indistinguishable (to a human) from the original '9'. Our manipulator learning procedure should make sure that the visual causes of the neural network's classification align with what causes a human to classify an image as '7'.

Figure 8 shows progress of the manipulator training. As in Fig. 6, we see that the manipulation error decreases slowly but steadily. It takes 20 iterations of presenting 200 causal queries to a human to achieve a manipulation error of about 10%. Fig. 9 shows that on the first iteration our neural network exhibits behavior described by Szegedy et al. (2014). For example, in the second "double-column" of the figure we can see an image of a '3' that when manipulated still looks like the same '3' to the human observer; but the manipulated image is considered a '7' by the network. At the end of the training, however, such cases do not occur. There are still some failed manipulations, but they look more like extremely noisy attempts at creating '7's than the original digits.

In this experiment, we again used a simple neural network with only one hidden layer of 100 neurons. This is arguably not a state-of-the-art learning machine that can achieve great MNIST classification performance: In our experiments, the test error was about 4% consistently throughout the experiments. An experiment that uses a modern convolutional neural net should produce significantly more eye-pleasing results and belongs to our future work directions.

5. Discussion

Modern causal discovery algorithms presuppose that the set of causal variables is well-defined and meaningful. What exactly this presupposition entails is unclear, but there are clear counter-examples: x and $2x$ cannot be two distinct causal variables. There are also well understood problems when causal variables



(a) Iteration 1 (b) Iteration 10

Figure 7: Original and manipulated GRATING images. Each column shows example manipulations of a particular kind (nine cases per column, unless only fewer were available at the given iteration). Columns with green labels indicate successful manipulations of which there are two kinds: switching the causal variable on ($0 \Rightarrow 1$, "adding the h-bar"), or switching it off ($1 \Rightarrow 0$, "removing the h-bar"). Red-labeled columns show cases in which the manipulator failed to influence the cause: That is, each red column shows an original image and its manipulated version which the manipulator believes should cause a change in T, but which does not induce such change. The red/green horizontal bars show the percentage of success/error for each manipulation direction. **(a)** After training on the causally-coarsened observational dataset, the manipulator functions fails about 40% of the time. **(b)** After twenty manipulator learning iterations, only six manipulations out of a hundred are unsuccessful. The causally irrelevant image pixels are also much better preserved than at iteration 1.

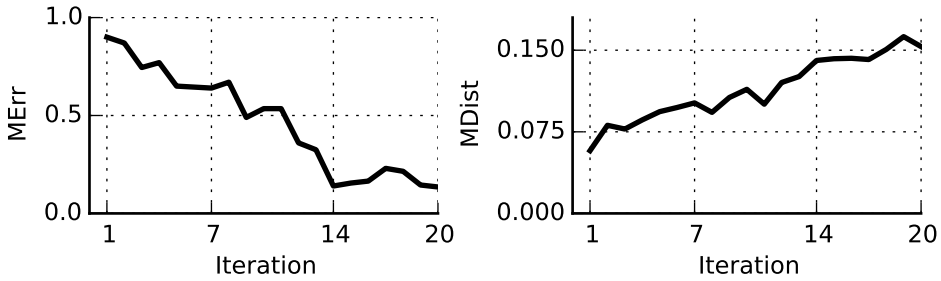
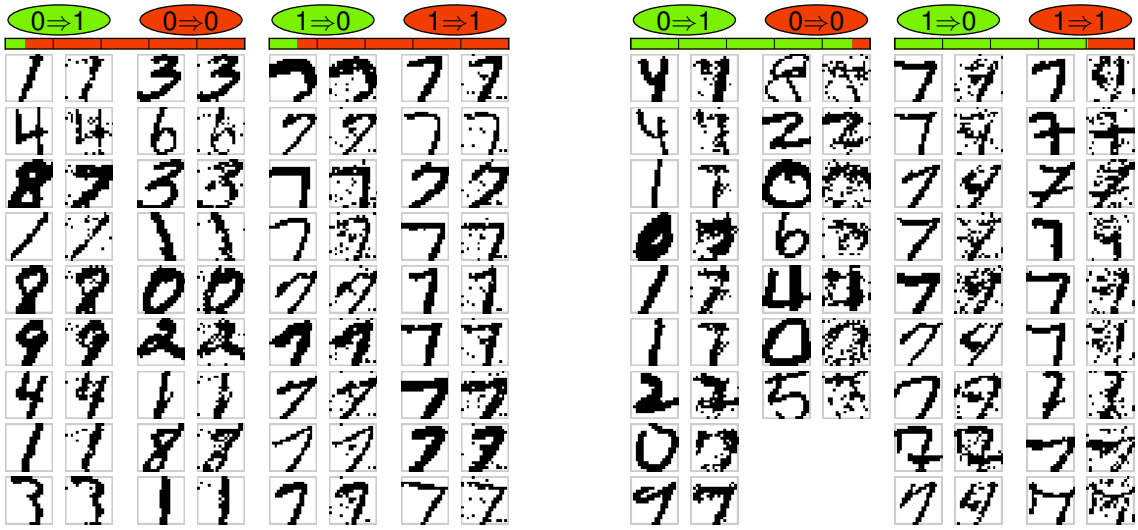


Figure 8: Manipulator Learning for MNIST. The plots show the progress of our learning algorithm for the manipulator function over twenty iterations of experiments for the MNIST problem. The manipulation error decreases steadily throughout training. In contrast to the GRATING experiment, here the manipulation distance grows with training. This is because significant portions of an image need to change to achieve the desired causal effect (for example, manipulating an '8' to look like a '7' requires removing large portions of the image).

are aggregates of other variables (Chu et al., 2003; Spirtes and Scheines, 2004). We provide an account of how causal macro-variables can supervene on micro-variables that they are constituted of.

Our learning method provides a link between causal discovery methods and neural network models that have recently enjoyed tremendous success in the field of computer vision (LeCun et al., 1998; Krizhevsky et al., 2012; Russakovsky et al., 2014; Simonyan and Zisserman, 2014a,b; Vinyals et al., 2014; Karpathy and Fei-Fei, 2014). One interpretation of our results is that we show how to learn *causal* as opposed to *observational* neural networks. Much progress has been made recently in image classification (Krizhevsky et al., 2012), in object detection and classification (Dollar et al., 2012) and in fine-grained classification (Branson et al., 2014; Zhang et al., 2014). However, detection and classification performance suffers when the test images obey statistics that are different from the training set's (Torralba and Efros, 2011). Furthermore, experiments show that it is easy to produce images that a human would classify in one class and a classification algorithm would classify in another, as shown by Szegedy et al. (2014). Current methods discover correlations between pixels and categories and may be somewhat blind to the aspects of an image that cause a human to classify it correctly. We have proposed an approach that discovers causal relationships between pixels and classes, and thus may help machines to be more robust to previously unseen image variations and better aware of image content.

In the framework of causal graphical models, which effectively provides an axiomatization of causality, the causal variables are simply treated as primitves. No theory is provided of how they are obtained or what constraints they must satisfy. This article is to our knowledge the first attempt to clarify how (at least in the case of images) one may construct a set of well-defined causal macro-variables that can function as basic relata in a causal graphical model. This step strikes us as essential if causal methodology is to be successful in areas where we do not have clearly delineated candidate causes or where causes supervene on micro-variables, such as in climate science and neuroscience, economics and—in our specific case—computer vision.



(a) Iteration 1 (b) Iteration 20

Figure 9: Original and manipulated MNIST images. We encourage the reader to zoom the figure out (if viewing on the computer screen) or look at it from a distance (if viewing on paper) to emulate the experimental setup in which a human judged the presence of a '7' in images of side length of about 1° of the visual field. In this task, switching the causal variable on means "making a 7 ", and switching it off means "removing the 7 " – all the time with smallest possible change of the image w.r.t. the L1 norm. **(a)** After training on the MNIST dataset, the manipulator function fails about 90% of the time. **(b)** At the end of training, the manipulation error falls to only 13%. The manipulator function now knows to add a bar at the top of a '1', to remove part of a '0', to change the strokes of a '4' etc, in order to make a '7'; and to complete the loop to make a '7' into a '9', or to simply add significant amounts of noise to the image to make the '7' hard to read.

Acknowledgments

KC's and PP's work was supported by the ONR MURI grant N00014-10-1-0933. FE would like to thank Cosma Shalizi for the introduction to computational mechanics and many relevant results this paper builds on. KC would like to thank the creators of the Theano (Bergstra et al., 2010; Bastien et al., 2012) and Pylearn2 (Goodfellow et al., 2013) Python packages that enabled him to implement the neural networks used in this article easily and efficiently.

Appendix A. Theorem Proofs

This appendix contains the proofs of the theorems stated in the main text of the article. All the variables are considered discrete and T is presumed binary, although extension of the proofs to any discrete T is easy (and complicates the notation significantly).

Before we prove the Causal Coarsening Theorem, we prove its less general version in order to split the rather complex proof of CCT into two parts.

Auxiliary Theorem *Among all the generative models of the form discussed in Fig. 2, the subset of distributions $P(T, H, I)$ for which the observational partition does not refine the causal partition is measure zero.*

Proof Our proof is inspired by a proof used by Meek (1995) to prove that almost all distributions compatible with a given causal graph are faithful. The proof strategy is thus first to express the proposition that for a given distribution, the observational partition does not refine the causal partition as a polynomial equation on the space of all distributions compatible with the model. We then show that this polynomial equation is not trivial, i.e. there is at least one distribution that is not its root. By a simple algebraic lemma, this will prove the theorem. We extend Meek's proof technique in our usage of Fubini's Theorem for the Lebesgue integral. It allows us to "split" the polynomial constraint into multiple different constraints along several of the distribution parameters. This allows for additional flexibility in creating useful assumptions (in our proof, the assumption that the datapoints have well-defined causal classes, but the observational class can still vary freely).

Assume that T is binary and $H = (H_1, \cdots, H_M)$, I are discrete variables (say $|H_i| = K_i, |I| = N$, though N can be very large. We will use the notation $K \triangleq K_1 \times \cdots \times K_i$ for simplicity later on). The discreteness assumption is not crucial, but will simplify the reasoning. We can factorize the joint as $P(T, H, I) = P(T \mid H, I)P(I \mid H)P(H)$. $P(T \mid H, I)$ can be parametrized by $|H_1| \times \cdots \times |H_M| \times |I| = K \times N$ parameters, $P(I \mid H)$ by $(N-1) \times K$ parameters, and $P(H)$ by another K parameters, all of which are independent. Call the parameters, respectively,

$$\alpha_{h,i} \triangleq P(T = 0 \mid H = h, I = i) \tag{5}$$

$$\beta_{i,h} \triangleq P(I = i \mid H = h) \tag{6}$$

$$\gamma_h \triangleq P(H = h) \tag{7}$$

We will denote parameter vectors as

$$\alpha = (\alpha_{h_1,i_1}, \cdots, \alpha_{h_K,i_N}) \in \mathbb{R}^{h_K \times i_N} \tag{8}$$

$$\beta = (\beta_{i_1,h_1}, \cdots, \beta_{i_{N-1},h_K}) \in \mathbb{R}^{i_{N-1} \times h_K} \tag{9}$$

$$\gamma = (\gamma_{h_1}, \cdots, \gamma_{h_K}) \in \mathbb{R}^{h_K}, \tag{10}$$

where the indices are arranged in lexicographical order. This creates a one-to-one correspondence of each possible joint distribution $P(T, H, I)$ with a point $(\alpha, \beta, \gamma) \in P[\alpha, \beta, \gamma] \subset \mathbb{R}^{(h_K)^3 \times i_N \times i_{N-1}}$, where $P[\alpha, \beta, \gamma]$ is the $(h_K)^3 \times i_N \times i_{N-1}$-dimensional simplex of multinomial distributions.

To proceed with the proof, we first pick any point in the $P(T \mid H, I) \times P(H)$ space: that is, we fix the values of α and γ. The only free parameters are now $\beta_{i,h}$ for all values of i, h; varying these values

creates a subset of the space of all the distributions which we will call

$$P[\alpha, \gamma] = \{(\alpha, \beta, \gamma) \mid \beta \in [0,1]^{i_{N-1} \times h_K}\}. \tag{11}$$

$P[\alpha, \gamma]$ is a subset of $P[\alpha, \beta, \gamma]$ isometric to the $[0,1]^{i_{N-1} \times h_K}$-dimensional simplex of multinomials. We will use the term $P[\alpha, \gamma]$ to refer both the subset of $P[\alpha, \beta, \gamma]$ and the lower-dimensional simplex it's isometric to, remembering that the latter comes equipped with the Lebesgue measure on $\mathbb{R}^{i_{N-1} \times h_K}$.

Now we are ready to show that the subset of $P[\alpha, \gamma]$ which does not satisfy the Causal Coarsening constraint is of measure zero with respect to the Lebesgue measure. To see this, first note that since α and γ are fixed, each image i has a well-defined causal class $C(i) = \sum_h \alpha_{h,i}\gamma_h$. The Causal Coarsening constraint says "For every pair of images i, j such that $P(T \mid i) = P(T \mid j)$ it holds that $C(i) = C(j)$." The subset of $P[\alpha, \gamma]$ of all distributions that don't satisfy the constraint consists of the $P(T, H, I)$ for which for some i, j it holds that $P(T = 0 \mid i) = P(T = 0 \mid j)$ and $C(i) \neq C(j)$. Take any pair i, j for which $C(i) \neq C(j)$ (if such a pair does not exist, then the Causal Coarsening constraint holds for all the distributions in $P[\alpha, \gamma]$). We can write

$$P(T = 0 \mid i) = \sum_h P(T = 0 \mid h, i)P(h \mid i) \tag{12}$$

$$= \frac{1}{P(i)} \sum_h P(T = 0 \mid h, i)P(i \mid h)P(h). \tag{13}$$

Since the same equation applies to $P(T = 0 \mid j)$, the constraint $P(T \mid i) = P(T \mid j)$ can be rewritten

$$\frac{1}{P(i)} \sum_h P(T = 0 \mid h, i)P(i \mid h)P(h) = \frac{1}{P(j)} \sum_h P(T = 0 \mid h, j)P(j \mid h)P(h), \tag{14}$$

$$P(j) \sum_h P(T = 0 \mid h, i)P(i \mid h)P(h) - P(i) \sum_h P(T = 0 \mid h, j)P(j \mid h)P(h) = 0, \tag{15}$$

Which we can rewrite in terms of the independent parameters (after defining $\alpha_{0,h,i} = \alpha_{h,i}$ and $\alpha_{1,h,i} = 1 - \alpha_{h,i}$) and further simplify as

$$\left(\sum_{t \in \{0,1\}} \sum_h \alpha_{t,h,j}\gamma_h\beta_{j,h}\right) \sum_h \alpha_{0,h,i}\gamma_h\beta_{i,h} - \left(\sum_{t \in \{0,1\}} \sum_h \alpha_{t,h,i}\gamma_h\beta_{i,h}\right) \sum_h \alpha_{0,h,j}\gamma_h\beta_{j,h} = 0, \tag{16}$$

$$\left(\sum_h \alpha_{1,h,j}\gamma_h\beta_{j,h}\right) \sum_h \alpha_{0,h,i}\gamma_h\beta_{i,h} - \left(\sum_h \alpha_{1,h,i}\gamma_h\beta_{i,h}\right) \sum_h \alpha_{0,h,j}\gamma_h\beta_{j,h} = 0, \tag{17}$$

$$\left(\sum_h (1 - \alpha_{h,j})\gamma_h\beta_{j,h}\right) \sum_h \alpha_{h,i}\gamma_h\beta_{i,h} - \left(\sum_h (1 - \alpha_{h,i})\gamma_h\beta_{i,h}\right) \sum_h \alpha_{h,j}\gamma_h\beta_{j,h} = 0, \tag{18}$$

$$\left(\sum_h \gamma_h\beta_{j,h}\right) \sum_h \alpha_{h,i}\gamma_h\beta_{i,h} - \left(\sum_h \gamma_h\beta_{i,h}\right) \sum_h \alpha_{h,j}\gamma_h\beta_{j,h} = 0, \tag{19}$$

which is a polynomial constraint on $P[\alpha, \gamma]$ (note that to keep the notation manageable, we have omitted the dependent term $1 - \sum_h \gamma_h$ from the equations.) By a simple algebraic lemma (proven by Okamoto, 1973), if the above constraint is not trivial (that is, if there exists β for which the constraint does not hold), the subset of $P[\alpha, \gamma]$ on which it holds is measure zero.

To see that Eq. (19) does not always hold, note that if for *any* h^* we set $\beta_{i,h^*} = 1$ (and thus $\beta_{i,h} = 0$ for any $h \neq h^*$) and $\beta_{j,h^*} = 1$, the equation reduces to $(\gamma_{h^*})^2(\alpha_{h_i,i} - \alpha_{h_j,h}) = 0$. Thus if Eq. (19) was trivially true, we would have $\alpha_{h,i} = \alpha_{h,j}$ or $\gamma_h = 0$ for all h. However, this implies $C(i) = C(j)$, which contradicts our assumption.

We have now shown that the subset of $P[\alpha, \gamma]$ which consists of distributions for which $P(T \mid i) = P(T \mid j)$ (even though $C(i) \neq C(j)$) is Lebesgue measure zero. Since there are only finitely many pairs of images i, j for which $C(i) \neq C(j)$, the subset of $P[\alpha, \gamma]$ of distributions which violate the Causal Coarsening constraint is also Lebesgue measure zero. The remainder of the proof is a direct application of Fubini's theorem.

For each α, γ, call the (measure zero) subset of $P[\alpha, \gamma]$ that violates the Causal Coarsening constraint $z[\alpha, \gamma]$. Let $Z = \cup_{\alpha,\gamma} z[\alpha, \gamma] \subset P[\alpha, \beta, \gamma]$ be the set of all the joint distributions which violate the Causal Coarsening constraint. We want to prove that $\mu(Z) = 0$, where μ is the Lebesgue measure. To show this, we will use the indicator function

$$\hat{z}(\alpha, \beta, \gamma) = \begin{cases} 1 & \text{if } \beta \in z[\alpha, \gamma], \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

By the basic properties of positive measures we have

$$\mu(Z) = \int_{P[\alpha,\beta,\gamma]} \hat{z} \, d\mu. \tag{21}$$

It is a standard application of Fubini's Theorem for the Lebesgue integral to show that the integral in question equals zero. For simplicity of notation, let

$$\mathcal{A} = \mathbb{R}^{hK \times iN} \tag{22}$$

$$\mathcal{B} = \mathbb{R}^{iN \times hK} \tag{23}$$

$$\mathcal{G} = \mathbb{R}^{hK}. \tag{24}$$

We have

$$\int_{P[\alpha,\beta,\gamma]} \hat{z} \, d\mu = \int_{\mathcal{A}\times\mathcal{B}\times\mathcal{G}} \hat{z}(\alpha, \beta, \gamma) \, d(\alpha, \beta, \gamma) \tag{25}$$

$$= \int_{\mathcal{A}\times\mathcal{G}} \int_{\mathcal{B}} \hat{z}(\alpha, \beta, \gamma) \, d(\beta) \, d(\alpha, \gamma) \tag{26}$$

$$= \int_{\mathcal{A}\times\mathcal{G}} \mu(z[\alpha, \gamma]) \, d(\alpha, \gamma) \tag{27}$$

$$= \int_{\mathcal{A}\times\mathcal{G}} 0 \, d(\alpha, \gamma) \tag{28}$$

$$= 0. \tag{29}$$

Equation (27) follows as $\hat{z}$ restricted to $P[\alpha, \gamma]$ is the indicator function of $z[\alpha, \gamma]$.

This completes the proof that Z, the set of joint distributions over T, H and I that violate the Causal Coarsening constraint, is measure zero. ■

We are now ready to prove the main theorem.

Theorem (Causal Coarsening Theorem) *Among all the generative models of the form discussed in Fig. 2 that have distributions $P(T, \mathbf{H}, I)$, which induce some given observational partition Π_o,* almost all *induce a causal partition Π_c that is a coarsening of Π_o.*

Proof Any variables that appear in this proof without definition are defined in the proof of the Auxiliary Theorem. We take the same α, β, γ parametrization of distributions. Fixing an observational partition means fixing a set of observational constraints (OCs)

$$P(T \mid i_1^1) = \cdots = P(T \mid i_{N_1}^1), \tag{30}$$

$$\vdots \tag{31}$$

$$P(T \mid i_1^K) = \cdots = P(T \mid i_{N_K}^K), \tag{32}$$

where $1 \leq K \leq N$ is the number of observational classes. Since $P(T, H, I) = P(H \mid T, I)P(T \mid I)P(I)$, $P(T \mid i)$ is an independent parameter in the unrestricted $P(T, H, I)$, and the OCs reduce the number of independent parameters of the joint by $\sum_{k=1}^{K}(N_k - 1)$. We want to express this parameter-space reduction in terms of the α, β and γ parameterization and then apply the proof of the Auxiliary Theorem. To do this, for each observational class k, choose a representative image $\hat{\imath}^k$ such that $P(T \mid i_l^k) = P(T \mid \hat{\imath}^k) \forall_{l \in 1 \cdots N_k}$. Then for each $i_l^k \neq \hat{\imath}^k$ it holds that

$$P(T, i_l^k) = P(T \mid \hat{\imath}^k) P(i_l^k) \tag{33}$$

or

$$\sum_h P(T, h, i_l^k) = P(T \mid \hat{\imath}^k) \sum_h P(h, i_l^k). \tag{34}$$

Picking an arbitrary h_0, we can separate the left-hand side as

$$P(T, h_0, i_l^k) = P(T \mid \hat{\imath}^k) \sum_h P(h, i_l^k) - \sum_{h \neq h_0} P(T, h, i_l^k). \tag{35}$$

Finally, this equation can be rewritten in terms of α, β and γ as

$$\alpha_{h_0,i} \beta_{i,h_0} \gamma_{h_0} = P(T \mid \hat{\imath}^k) \sum_h \beta_{h,i_l^k} \gamma_h - \sum_{h \neq h_0} \alpha_{h,i_l^k} \beta_{i_l^k} \gamma_h, \tag{36}$$

or

$$\alpha_{h_0,i} = \frac{1}{\beta_{i,h_0} \gamma_{h_0}} \left(P(T \mid \hat{\imath}^k) \sum_h \beta_{h,i_l^k} \gamma_h - \sum_{h \neq h_0} \alpha_{h,i_l^k} \beta_{i_l^k} \gamma_h \right). \tag{37}$$

for any $i_l^k \neq \hat{\imath}^k$. There are precisely $\sum_{k=1}^{K}(N_k - 1)$ such equations, altogether equivalent to the observational constraints. Thus we can express any $P(T, H, I)$ distribution that is consistent with a given

observational partition in terms of the full range of β and γ parameters, and a restricted number of independent α parameters. The rest of the proof now follows exactly like the proof of the Auxiliary Theorem and shows that within this restricted parameter space, the parameters for which the (fixed) observational partition is not a refinement of the causal partition is measure zero. ■

Theorem (Complete Macro-variable Description) *The following two statements hold for C and S as defined above:*

1. $P(T \mid I) = P(T \mid C, S)$.
2. *Any other variable X such that $P(T \mid I) = P(T \mid X)$ has Shannon entropy $H(X) \geq H(C, S)$.*

Proof The first part follows by construction of S. For the second part, note that by the CCT there is a bijective correspondence between the pairs of values (c, s) and the observational probabilities $P(T \mid I)$. Call this correspondence f, that is $f(c, s) = P(T \mid c, s)$ and $f^{-1}(p) = (c, s \text{ s.t. } P(T|c, s) = p)$. Further, define g as the function on X, with $g\colon x \mapsto P(T \mid x)$. But since $P(T \mid X) = P(T \mid I)$, we have $(c, s) = f^{-1}(g(x))$. That is, the value of C and S is a function of the value of X, and thus the entropy of C and S is smaller than the entropy of X. ■

References

F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I.J. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, and Y. Bengio. Theano: new features and speed improvements. In *NIPS 2012 deep learning workshop*, 2012.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: a CPU and GPU math compiler in Python. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010.

A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *6th ACM International Conference on Image and Video Retrieval*, pages 401–408, 2007.

S. Branson, G. Van Horn, and C. Wah. The Ignorant Led by the Blind: A Hybrid Human–Machine Vision System for Fine-Grained Categorization. *International Journal of Computer Vision*, 108(1-2): 3–29, 2014.

T. Chu, C. Glymour, R. Scheines, and P. Spirtes. A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, 19(9):1147–1152, 2003.

R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2), 2008.

P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012.

I.J. Goodfellow, D. Warde-Farley, P. Lamblin, V. Dumoulin, M. Mirza, R. Pascanu, J. Bergstra, F. Bastien, and Y. Bengio. Pylearn2: a machine learning research library. *arXiv preprint arXiv:1308.4214*, 2013.

K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8:725–260, 2007.

I. Guyon, A. Elisseeff, and C. Aliferis. Causal feature selection. In *Computational Methods of Feature Selection Data Mining and Knowledge Discovery Series*, pages 63–85. Chapman and Hall/CRC, 2007.

Z. Harchaoui and F. Bach. Image classification with segmentation graph kernels. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. Technical report, Stanford University, 2014. URL `http://cs.stanford.edu/people/karpathy/deepimagesent/`.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *ACM SIGIR Seventeenth Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

C. Meek. Strong completeness and faithfulness in Bayesian networks. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 411–418, 1995.

M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1(4):763–765, 1973.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

J. P. Pellet and A. Elisseeff. Using Markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9:1295–1342, 2008.

Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014. URL `http://arxiv.org/abs/1409.0575`.

B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

B. Settles. Active learning literature survey. Technical report, University of Wisconsin, Madison, 2010.

B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Conference on Empirical Methods in Natural Langauge Processing*, pages 1070–1079, 2008.

C. R. Shalizi. *Causal architecture, complexity and self-organization in the time series and cellular automata*. PhD thesis, University of Wisconsin at Madison, 2001.

C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3-4):817–879, 2001.

K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Z Ghahramani, M Welling, C Cortes, ND Lawrence, and KQ Weinberger, editors, *Advances in Neural Information Processing Systems 27*, 2014a.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014b. URL http://arxiv.org/abs/1409.1556.

P. Spirtes and R. Scheines. Causal inference of ambiguous manipulations. *Philosophy of Science*, 71(5): 833–845, 2004.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. Massachusetts Institute of Technology, 2nd ed. edition, 2000.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.

A. Torralba and A.A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A Neural Image Caption Generator. *arXiv preprint arXiv:1411.4555*, 2014. URL http://arxiv.org/abs/1411.4555.

S. V. N. Vishwanathan. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based R-CNNs for fine-grained category detection. In *ECCV 2014*, pages 834–849, 2014. URL http://link.springer.com/chapter/10.1007/978-3-319-10590-1_54.