# Visual Causal Feature Learning

**Krzysztof Chalupka**
Computation and Neural Systems
California Institute of Technology
Pasadena, CA, USA

**Pietro Perona**
Electrical Engineering
California Institute of Technology
Pasadena, CA, USA

**Frederick Eberhardt**
Humanities and Social Sciences
California Institute of Technology
Pasadena, CA, USA

## Abstract

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. 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.

## 1 INTRODUCTION

Visual perception is an important trigger of human and animal behavior. The visual cause of a behavior can be easy to define, say, when a traffic light turns green, or quite subtle: apparently it is the increased symmetry of features that leads people to judge faces more attractive than others (Grammer and Thornhill, 1994). Significant scientific and economic effort is focused on visual causes in advertising, entertainment, communication, design, medicine, robotics and the study of human and animal cognition. Visual causes profoundly influence our daily activity, yet our understanding of what constitutes a visual cause lacks a theoretical basis. In practice, it is well-known that images are composed of millions of variables (the pixels) but it is functions of the pixels (often called 'features') that have meaning, rather than the pixels themselves.

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:

- 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 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), thereby contributing to an account of how to construct causal variables.
- We prove the Causal Coarsening Theorem (CCT), 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.
- 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, as well as reproduces some of the experimental results, is available online at `http://vision.caltech.edu/~kchalupk/code.html`.

We chose to develop the theory within the context of *visual* causes as this setting makes the definitions most intuitive and is itself of significant practical interest. However, the framework and results can be equally well applied to extract causal information from any aggregate of micro-variables on which manipulations are possible. Examples include auditory, olfactory and other sensory stimuli; high-dimensional neural recordings; market data in finance; consumer data in marketing. There, causal feature learning is both of theoretical ("What is the cause?") and practical ("Can we automatically manipulate it?") importance.

### 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 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 (Shalizi and Crutchfield, 2001; Shalizi, 2001), but supports a more explicitly causal interpretation by incorporating the possibility of confounding and interventions. We take the distinction between interventional and observational distributions to be one of the key features of a causal analysis. Since we allow for unmeasured common causes of the features in the image and the target behavior, we have to distinguish between the plain conditional probability distribution of the target behavior ($T$) given the (observed) image ($I$) and the distribution of the target behavior given that the observed image was manipulated (i.e. $P(T|I)$ vs. $P(T|do(I))$). Hoel et al. (2013), who develop a similar model to investigate the relationship between causal micro- and macro-variables, avoid this distinction by assuming that all their data was generated from what in our setting would be the manipulated distribution $P(T|do(I))$. The extant literature on causal learning from image or video data does not generally consider the aggregation from pixel variables into causal macro-variables, but instead starts from annotated or pre-defined features of the image (see e.g. Fire and Zhu (2013a,b)).

### 1.2 CAUSAL FEATURE LEARNING: AN EXAMPLE

Fig. 1 presents a paradigmatic case study in visual causal feature learning, which we will use as a running example. The contents of an image $I$ are caused by external, non-visual 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.

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

We deliberately constructed this example such that the visual cause is 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 $H_1$ or $T$. Still, the presence of a v-bar is 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 $I$:

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

Both the presence of h-bars and the presence of v-bars are good individual (and even better joint) 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.

While the values of $S$ and $C$ in our example stand in a bijective correspondence to the values of $H_1$ and $H_2$, respectively, this is only to keep the illustration simple. In general, the visual cause and the spurious correlate can be probabilistic functions of any number of (not necessarily independent) hidden variables, and can share the same hidden causes.

## 2 A THEORY OF VISUAL CAUSAL FEATURES

In our example the identification of the visual cause with the presence of an h-bar is intuitively obvious, as the model is constructed to have an easily describable visual cause. But the example does not provide a 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

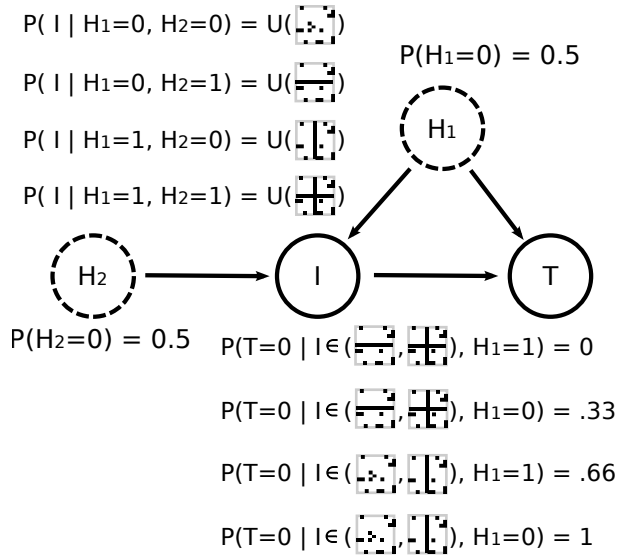


Figure 1: Our case study generative model. Two binary hidden (non-visual) variables $H_1$ and $H_2$ toss unbiased coins. The content of the image $I$ depends on these variables as follows. If $H_1 = H_2 = 0$, $I$ is chosen uniformly at random from all the images containing no v-bars and no h-bars. If $H_1 = 0$ and $H_2 = 1$, $I$ is chosen uniformly at random from all 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 $I$.

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). Our goal will be 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 (natural) assumptions about the causal relation between the image and the behavior: (i) The value of the target behavior $T$ is determined subsequently to the image in time, and (ii) the variable $T$ is in no way represented in the image. These assumptions exclude the possibility that $T$ is a cause of features in the image or that $T$ can be seen as causing itself.

## 2.2 GENERATIVE MODELS: FROM MICRO- TO MACRO-VARIABLES

Let $T \in \{0, 1\}$ represent a target 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:

$$\begin{aligned} P(T, I) &= \sum_{\mathbf{H}} P(T \mid I, \mathbf{H}) P(I \mid \mathbf{H}) P(\mathbf{H}) \\ &= \sum_{\mathbf{H}} P(T \mid I, \mathbf{H}_c) P(I \mid \mathbf{H}) P(\mathbf{H}). \end{aligned} \quad (1)$$

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 $P(T \mid I)$:

**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* $\Pi_o$ *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

[2] An extension of the framework to non-binary, discrete $T$ is easy but complicates the notation significantly. An extension to the continuous case is beyond the scope of this article.

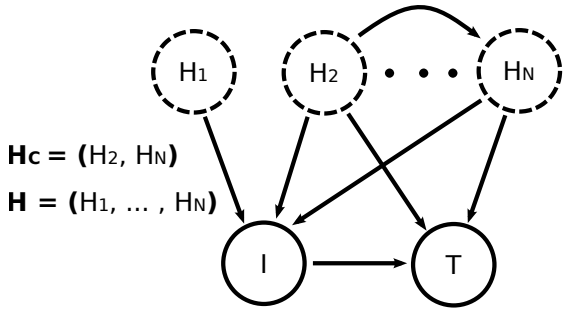


Figure 2: A general model of visual causation. In our model each image $I$ is caused by a number of hidden non-visual variables $H_i$, which need not be independent. The image itself is the only observed cause of a target behavior $T$. In addition, a (not necessarily proper) subset of the hidden variables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

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 $T$. For example, 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 $P(T \mid man(I = i)) = \sum_{\mathbf{H}_c} P(T \mid I = i, \mathbf{H}_c)P(\mathbf{H}_c)$.*

The manipulation changes the values of the 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 $\Pi_c$. 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 stands in a bijective relation to 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

Our main theorem 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 may subdivide 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 $\Pi_o$, almost all induce a causal partition $\Pi_c$ that is a coarsening of the $\Pi_o$.*

Throughout this article, we use "almost all" to mean "all except for a subset of Lebesgue measure zero". Fig. 3 illustrates the relation between the causal and the observational partition implied by the theorem. We prove the CCT in Supplementary Material A using a technique that extends that of Meek (1995): We show that (1) restricting the space of all the possible $P(T, H, I)$ to only the distributions compatible with a fixed observational partition puts a linear constraint on the distribution space; (2) requiring that the CCT be false puts a non-trivial polynomial constraint on this subspace, and finally, (3) it follows that the theorem holds for almost all distributions that agree with the given observational partition. The proof strategy indicates a close connection between the CCT and the faithfulness assumption (Spirtes et al., 2000). We note that the measure-zero subset where $\Pi_C$ does not coarsen $\Pi_O$ can indeed be non-empty. We provide such counter-examples in Supplementary Material B.

Two points are worth noting here: First, the CCT is interesting inasmuch as the visual causes of a behavior do not contain all the information in the image that predict the behavior. Such information, though not itself a cause of

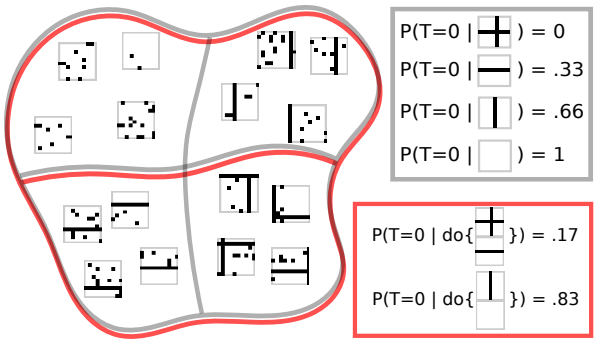


Figure 3: The Causal Coarsening Theorem. The observational probabilities of $T$ given $I$ (gray frame) induce an observational partition on the space of all the images (left, observational partition in gray). The causal probabilities (red frame) induce a causal partition, indicated on the left in red.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

the behavior, can be informative about the state of other non-visual causes of the target behavior. Second, the 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 $\Pi_o^T$ refines the causal partition $\Pi_c^T$. 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 an 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.*

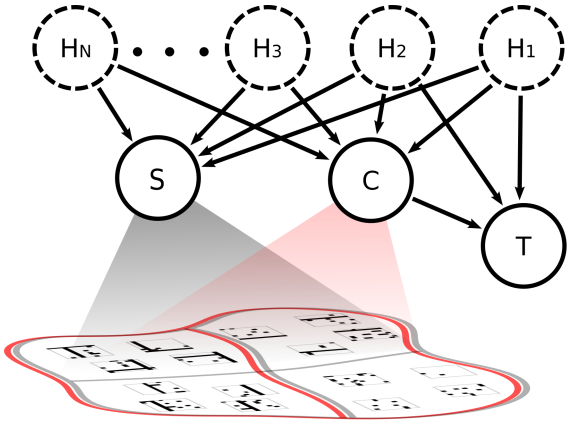


Figure 4: A macro-variable model of visual causation. Using our theory of visual causation we can aggregate the information present in visual micro-variables (image pixels) into the visual cause $C$ and spurious correlate $S$. According to Theorem 7, $C$ and $S$ contain all the information about $T$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

The spurious correlate is a well-defined function on $\mathcal{I}$, whose value ranges between 1 and $\max_k M_k$. Like $C$, the spurious correlate $S$ 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:

**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 Supplementary Material C. It guarantees that $C$ and $S$ constitute the smallest-entropy macro-variables that encompass all the information about the relationship between $T$ and $I$. Fig. 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).[3]

The macro-variable model lends itself to the standard treatment of causal graphical models described in Pearl

[3]We note that $C$ may retain predictive information about $T$ that is not causal, i.e. it is not the case that all spurious correlations can be accounted for in $S$. See Supplementary Material D for an example.

(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 (see the *modularity assumption* in Pearl (2009)). However, unlike the standard case where causal variables are separated in location (e.g. *smoking* and *lung cancer*), the causal variables in an image may involve the same pixels: $C$ may be the average brightness of the image, whereas $S$ may indicate 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 $man(I = i')$ 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.*

In some cases it can be impossible to manipulate $C$ to a desired value without changing $S$. We do not take this to be a problem special to our case. In fact, in the standard macro-variable setting of causal analysis we would expect interventions to be much more restricted by physical constraints than we are with our interventions in the image space.

# 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 $M_C \colon \mathcal{I} \times \mathcal{C} \rightarrow \mathcal{I}$ 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 the desired causal effect is a heuristic approach to fulfilling that requirement.

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

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

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 (recall the barometer example). However, we can use the CCT 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.

In Step 1 the algorithm picks a representative member of each observational class. The CCT 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,

**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 $\forall k, p_k \in \mathcal{P}, 1 \leq k \leq N$)
`Train` – a neural net training algorithm

**output**: $C \colon \mathcal{I} \to [0, 1]$ – 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 $C \leftarrow \texttt{Train}(\mathcal{D}_{csl})$;

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. In our experiments, we use a metric induced by an $L_2$ norm. Alternatives include other $L_p$-induced metrics, distances in implicit feature spaces induced by image kernels (Harchaoui and Bach, 2007; Grauman and Darrell, 2007; Bosch et al., 2007; Vishwanathan, 2010) and distances in learned representation spaces (Bengio et al., 2013).

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, inspired by the active learning techniques of uncertainty sampling (Lewis and Gale, 1994) and density weighing (Settles and Craven, 2008), 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

**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` – a neural net training algorithm
nIters – number of experiment iterations
Q – number of queries per iteration
$\alpha$ – manipulation tuning parameter
$\texttt{A} \colon \mathcal{I} \to \mathcal{C}$ – an oracle for $P(T \mid do(I))$

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

1 **for** $l \leftarrow 1$ **to** nIters **do**
2 $\quad C \leftarrow \texttt{Train}(\mathcal{D}_{csl})$;
3 $\quad$ Choose manipulation starting points $\{i_{l,1}, \cdots, i_{l,\mathsf{Q}}\}$ at random from $\mathcal{D}_{csl}$;
4 $\quad$ Choose manipulation targets $\{\hat{c}_{l,1}, \cdots, \hat{c}_{l,\mathsf{Q}}\}$ such that $\hat{c}_{l,k} \neq c_{l,k}$;
5 $\quad$ **for** $k \leftarrow 1$ **to** Q **do**
6 $\quad\quad \hat{\imath}_{l,k} \leftarrow \operatorname{argmin}_{j \in \mathcal{I}} (1-\alpha)|C(j) - \hat{c}_{l,k}| + \alpha\, d(j, i_{l,k})$;
7 $\quad$ **end**
8 $\quad \mathcal{D}_{csl} \leftarrow \mathcal{D}_{csl} \cup \{(\hat{\imath}_{l,1}, \texttt{A}(\hat{\imath}_{l,1})), \cdots, (\hat{\imath}_{l,\mathsf{Q}}, \texttt{A}(\hat{\imath}_{l,\mathsf{Q}}))\}$;
9 **end**

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 $|C(j) - \hat{c}_{l,k}|$ (the difference of the output image $j$'s label and the desired label $\hat{c}_{l,k}$) and $d(i_{l,k}, j)$ (the distance of the output image $j$ from the original image $i_{l,k}$).

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

$$MErr_l = \frac{1}{Q} \sum_{k=1}^{Q} |A(\hat{\imath}_{l,k}) - \hat{c}_{l,k}|. \tag{2}$$

While it is important that our manipulations are accurate, we also want them to be minimal. Another measure of interest is thus the *average manipulation distance*

$$MDist_l = \frac{1}{Q} \sum_{k=1}^{Q} d(I_{l,k}, \hat{\imath}_{l,k}). \tag{3}$$

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.

## 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 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 (see Fig. 5). Both restrictions increase the clarity of presentation. We use Algorithms 1 and 2 (with minor modifications imposed by the binary nature of the images) to learn the visual cause of behavior $T$.

Figure 5 (top) shows the progress of the training process. The first step (not shown in the figure) uses the CCT 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. We then follow Algorithm 2 to train the manipulator iteratively. Fig. 5 (bottom) 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%. Each column shows example manipulations of a particular kind. 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. Fig. 5 (bottom, a) shows that after training on the causally-coarsened observational dataset, the manipulator fails about 40% of the time. In Fig. 5 (b), after twenty manipulator learning iterations, only six manipulations out of a hundred are unsuccessful. Furthermore, the causally irrelevant image pixels are also much better preserved than at iteration 1. 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.

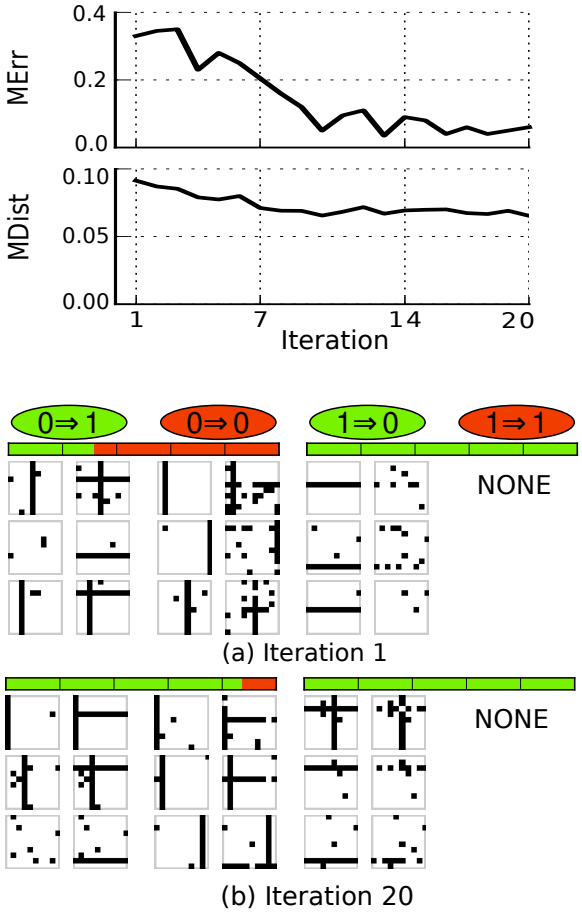


Figure 5: Manipulator learning for GRATING. **Top.** The plots show the progress of our manipulator function learning algorithm over twenty iterations of experiments for the GRATING problem. The manipulation error decreases quickly with progressing iterations, whereas the manipulation distance stays close to constant. **Bottom.** Original and manipulated GRATING images. See text for the details.

### 4.2 THE MNIST ON MTURK EXPERIMENT

In this experiment we start with the MNIST dataset of handwritten digits. In our terminology, this – as well as any standard vision dataset – is already causal data: the labels are assigned in an experimental setting, not "in nature".

Consider the following binary human behavior: $T = 1$ if a human observer answers affirmatively to the question

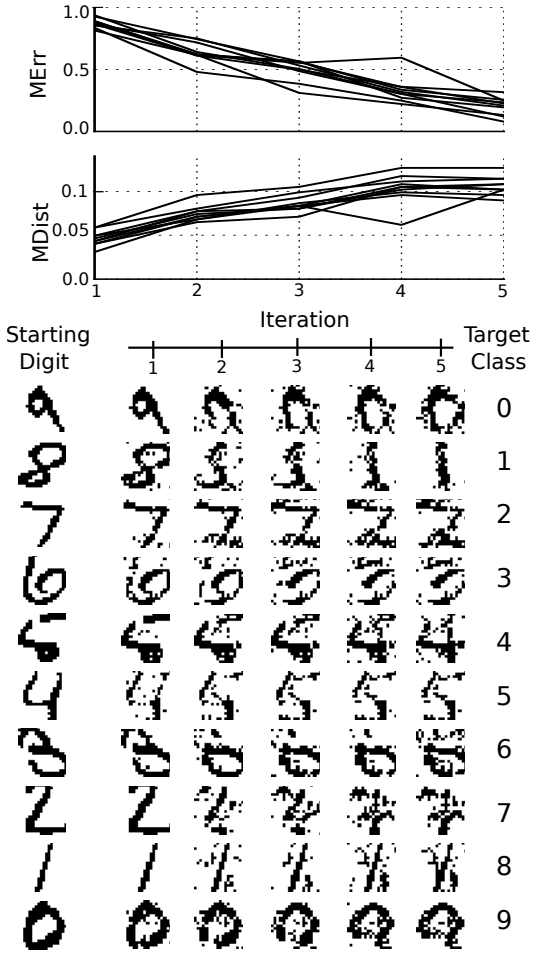


Figure 6: Manipulator Learning for MNIST ON MTURK. **Top.** In contrast to the GRATING experiment, here the manipulation distance grows as the manipulation error decreases. This is because a successful manipulator needs to change significant parts of each image (such as continuous strokes). **Bottom.** Visualization of manipulator training on randomly selected (not cherry-picked) MNIST digits. See text for the details.

"Does this image contain the digit '7'?", while $T = 0$ 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 $P(T = 1 \mid man(I)) = 1$. Our task is to learn the manipulator function that will take any image and modify it minimally such that it will become a '7' if it was not before, or will stop resembling a '7' if it did originally.

We conduct the manipulator training separately for all the ten MNIST digits using human annotators on Amazon Mechanical Turk. The exact training procedure is described in Supplementary Material E. Fig. 6 (top) shows training progress. As in Fig. 5, the manipulation error decreases with training. Fig. 6 (bottom) visualizes the manipulator training progress. In the first row we see a randomly chosen MNIST "9" being manipulated to resemble a "0", pushed through successive "0-vs-all" manipulators trained at iterations 0, 1, ..., 5 (iteration 1 shows what the neural net takes to be the closest manipulation to change the "9" to a "0" purely on the basis of the non-manipulated data). Further rows perform similar experiments for the other digits. The plots show how successive manipulators progressively remove the original digits' features and add target class features to the image.

# 5 DISCUSSION

We provide a link between causal reasoning and neural network models that have recently enjoyed tremendous success in the fields of machine learning and computer vision (LeCun et al., 1998; Russakovsky et al., 2014). Despite very encouraging results in image classification (Krizhevsky et al., 2012), object detection (Dollar et al., 2012) and fine-grained classification (Branson et al., 2014; Zhang et al., 2014), some researchers have found that visual neural networks can be easily fooled using adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2014). The learning procedure for our manipulator function could be viewed as an attempt to train a classifier that is robust against such examples. The procedure uses causal reasoning to improve on the boundaries of a standard, correlational classifier (Fig. 5 and 6 show the improvement). However, the ultimate purpose of a causal manipulator network is to extract truly causal features from data and automatically perform causal manipulations based on those features.

A second contribution concerns the field of causal discovery. 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 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.

This article is an attempt to clarify how one may construct a set of well-defined causal macro-variables that 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—vision.

**Acknowledgements**

KC's work was funded by the Qualcomm Innovation Fellowship 2014. 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 pointers to many relevant results this paper builds on.

## References


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

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.

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.

A. S. Fire and S. C. Zhu. Using causal induction in humans to learn and infer causality from video. *The Annual Meeting of the Cognitive Science Society (CogSci)*, 2013a.

A. S. Fire and S. C. Zhu. Learning Perceptual Causality from Video. *AAAI Workshop: Learning Rich Representations from Low-Level Sensors*, 2013b.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*, 2014.

K. Grammer and R. Thornhill. Human (Homo sapiens) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of Comparative Psychology*, 108(3):233–242, 1994.

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.

E. P. Hoel, L. Albantakis, and G. Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49): 19790–19795, 2013.

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.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

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

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.

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.

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.

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.