

# Semi-supervised Learning of Visual Causal Macrovariables

Aruna Jammalamadaka, Lingyi Zhang\*, Joe Comer<sup>†</sup>  
Sasha Strelnikoff, Ryan Mustari, Tsai-Ching Lu, Rajan Bhattacharyya

Intelligent Systems Laboratory, HRL Laboratories, LLC

{ajammalamadaka, lzhang, jfcomer, sstrelnikoff, rmustari, tlu, rbhattach}@hrl.com

## Abstract

Discovery of causally related concepts is one of the key challenges in extracting knowledge from observational data. Lower-dimensional “causal macrovariables” represent concepts which preserve all relevant causal information in high-dimensional systems. Existing causal macrovariable discovery algorithms are limited by assumptions about known and controllable interventions. We propose a variational autoencoder-inspired architecture with regularization terms for semi-supervised causal macrovariable discovery. These terms impose domain knowledge regarding visual causal concepts to differentiate between correlation and causation. Experiments on both synthetic and real-world datasets with known causal dynamics show that our method can discover more concise and precise causal macrovariables than unsupervised methods.

## Introduction

Discovering causal abstractions from observations of complex systems has increasing importance for various high-stakes operational tasks. Decision makers such as intelligence analysts, maintenance, logistics, and planning personnel are increasingly overwhelmed with the high volume of data they need to forage through and evaluate. Deriving recommendations by fusing multiple sources of visual data (e.g., satellite imagery, radar, grids or rasters representing terrain, elevation, or weather) is even more challenging. While some analysts are highly skilled in image processing and machine learning methods, many decision-makers are not, resulting in a need for interpretable and robust visual explanations for analyst recommendations. Motivated by this need, we propose a novel framework to abstract and display causal concepts from visual data, while allowing non-technical domain experts to provide visual annotations on aspects of the data they believe to be causal factors for their decisions.

Causal macrovariables, first introduced in Chalupka, Perona, and Eberhardt (2015), are lower-dimensional representations of high-dimensional “microvariables”  $X$  and  $Y$  which preserve all relevant causal information. They attempt to answer which changes in  $Y$  are caused by which changes in  $X$ ,

\*This author is now at University of Connecticut.

<sup>†</sup>This author is now at iRhythm Technologies, Inc.

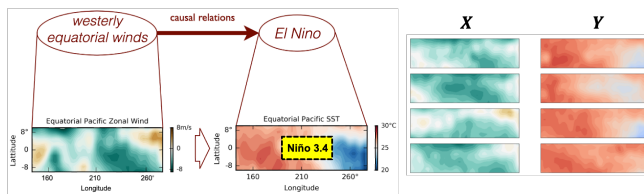


Figure 1: El Niño Dataset, (Chalupka et al., 2016). Left: At the macro-level, westerly equatorial winds are a known cause of El Niño, characterized by deviations in sea surface temperature within the Niño 3.4 region ( $120^{\circ}W$ - $170^{\circ}W$ ,  $5^{\circ}N$ - $5^{\circ}S$ ) (yellow). Figure from Eberhardt (2022). Right: Four examples of microvariables  $X$  and  $Y$ , corresponding to zonal WS and SST maps from a region of the Equatorial Pacific ocean.

given the observed variation in both. Causal macrovariables provide an approximate causal abstraction (Beckers, Eberhardt, and Halpern, 2020) of a cause-effect system  $X \rightarrow Y$ , as opposed to other concept learning paradigms which focus on learning concepts within  $X$  with respect to a categorical classification. Take for example the El Niño phenomenon, which describes a tongue of hot water that extends from the west coast of the Americas into the Pacific Ocean. In this way, the “El Niño” macrovariable supervenes upon the specific sea surface temperatures (SST) which constitute it. As demonstrated in Chalupka et al. (2016), high-dimensional wind speed (WS) ( $X$ ) and SST ( $Y$ ) measurements can be used to discover macrovariables corresponding to westerly equatorial winds causing the El Niño phenomenon (Figure 1). These macrovariables are contextual in that they are unique to the cause-effect system of interest; causal macrovariables capturing aspects of SST which cause California rainfall may have entirely different representations.

As pointed out in Schölkopf et al. (2021), learning causal macrovariables requires some sort of supervision, and which variables can be extracted and their level of abstraction depends on which distribution shifts, explicit interventions, and other supervision signals are available. Unsupervised causal macrovariable discovery algorithms (Chalupka, Perona, and Eberhardt, 2015; Höltgen, 2021) are limited by their need for controllable interventions on the data generating process to distinguish correlation from causation. These interventions are not always possible, e.g., in the climate science example

above.

A related field of research is representation learning, where the goal is to learn factors of variation within a single high-dimensional variable  $X$  (e.g., (Chen et al., 2018; Truble et al., 2021; Locatello et al., 2019)), or with respect to a scalar value  $y$  (e.g., image classification in (Mitrovic et al., 2020)). These methods often utilize self-supervision in the form of strong assumptions about the data generating process (e.g., contrastive learning or concept independence). Our method refrains from making such strong assumptions and addresses the case where  $Y$  may also be high-dimensional.

In this paper, we choose to impose human supervision in the form of domain-expert-provided regions of interest (ROI) to enhance the practical use of causal macrovariable discovery methods. To our knowledge, we are the first to enable an end-user to provide this supervision in the visual domain, which (although not explored in this paper) paves the way for passive collection of supervision using eye-tracking devices. Our specific contributions are

- Extensions of (Holtgen, 2021) which provide more stable and semantically meaningful results by leveraging nonlinear spatial relationships in image data.
- A modified loss function which incorporates regularization terms to impose domain expert supervision in the form of cause or effect ROIs.
- Quantitative evaluation of discovered causal macrovariables which more concisely and precisely capture known causal dynamics on both synthetic and real-world datasets.

The remainder of this paper is structured as follows. *Problem Setup and Related Work* provides a mathematical definition of causal macrovariables and summarizes related work in deriving them from two high-dimensional datasets. *Supervised Causal Autoencoder* outlines our novel “SCAE” architecture and its advantages. *Experiments* describes the synthetic and real-world datasets, metrics, and results. Finally, *Discussion* concludes with a summary of the paper and potential avenues for future work.

## Problem Setup and Related Work

We assume a dataset of paired observations  $\mathbf{X}, \mathbf{Y} = \{(X_i, Y_i)\}_{i=1}^N$  of two high-dimensional, continuous random microvariables  $X$  and  $Y$  where we believe that there are aspects of  $X$  which cause aspects of  $Y$ . The dimension of  $X$  and  $Y$  need not be the same, however, let us assume for ease of notation that they are both dimension  $M$ . Our goal is to find functions  $f$  and  $g$  yielding  $K$ -dimensional macrovariables  $\bar{X} = f(X)$  and  $\bar{Y} = g(Y)$  which ideally capture all and only the causal information between  $X$  and  $Y$ , where  $K \ll M$ . As opposed to the prior work described below, our method also assumes the availability of annotated ROIs  $\alpha(X_i)$  and/or  $\alpha(Y_i)$ , which correspond to visual causes or effects respectively, for a subset of observations  $i \in S$ .

Chalupka, Perona, and Eberhardt (2015) attempt to solve this problem using a method called Causal Feature Learning (CFL). The method finds equivalence classes of the conditional distribution  $P(Y|X)$  by grouping together  $X_i$ ’s that

are observed to cause similar  $Y_i$ ’s and  $Y_i$ ’s which are observed to be caused by similar  $X_i$ ’s. This grouping produces an observation-based (*correlational*) partition of the space of  $X$  and  $Y$  such that  $f$  and  $g$  classify observed microvariables into scalar macrovariable categories. They also prove the *Causal Coarsening Theorem*, which dictates that if one can perform interventions the true *causal* macrovariables resulting from equivalence classes of  $P(Y|do(X = x))$  can be found by further merging (or coarsening) the correlational macrovariables. This idea is renamed “refinements” in Mitrovic et al. (2020). The drawbacks of this method are that it does not allow for continuous-valued macrovariables (which provide a richer representation), does not provide a data-driven method for determining the number of causal macrovariables, and does not lend itself easily to ROI-based supervision.

Holtgen (2021) proposes a new characterization of causal macrovariables as information bottlenecks (Tishby, Pereira, and Bialek, 2000) between  $X$  and  $Y$ , using a novel neural network architecture called the Causal Autoencoder (CAE). Inspired by the variational autoencoder (VAE), the approach consists of training two connected stochastic neural networks ( $net_X$  and  $net_Y$ ) simultaneously. Each network encoder provides a dimension  $M$  to dimension  $K$  compression of the input ( $X$  or  $Y$ ) with respect to the output ( $Y$  or  $X$ ) such that  $\bar{X} = (x_1, \dots, x_K)$  and  $\bar{Y} = (y_1, \dots, y_K)$ . Because this method overcomes the drawbacks of CFL, we explain it in more detail below and provide our extensions in *Supervised Causal Autoencoder*.

We compare our results to CFL and CAE due to the lack of supervised methods for causal macrovariable discovery. Since CFL and CAE are entirely unsupervised, SCAE does not claim to “outperform” them, rather, our comparisons provide empirical evidence that the proposed method can effectively incorporate image-based annotations from domain experts that are familiar with visual causes and effects present in the data.

**Causal Autoencoder** The CAE architecture consists of two encoder-decoder networks,  $net_X$  and  $net_Y$ , tied together by a linear function between their information bottlenecks. In the following, we describe  $net_X$ , but  $net_Y$  has the same structure with input and output inverted. In a standard  $\beta$ -VAE framework (Higgins et al., 2017), the information bottleneck is meant to trade-off compression and reconstruction of the input  $X$ . In the CAE, the information bottleneck is instead computed between  $\bar{X}$  (containing causes) and  $\bar{Y}$  (containing effects). Assuming that causal macrovariables  $\bar{X}$  and  $\bar{Y}$  contain all causal connections between  $X$  and  $Y$ , they must also contain all mutual information shared by  $X$  and  $Y$ , i.e.,  $\mathcal{I}(\bar{X}; \bar{Y}) = \mathcal{I}(X; Y) = \mathcal{I}(\bar{Y}; \bar{X}) = \mathcal{I}(Y; X)$ . Therefore, the optimal assignment of  $\bar{X}$  minimizes the information bottleneck functional  $\mathcal{L} = \mathcal{I}(\bar{X}; X) - \beta \cdot \mathcal{I}(\bar{X}; \bar{Y})$  where  $\beta$  governs the trade-off between compression and prediction.

The goal of  $net_X$  is to learn three functions: the encoder  $f(X) = \bar{X}$ , the bottleneck-to-bottleneck relationship between macrovariables  $\hat{\bar{Y}} = h(\bar{X})$ , and the decoder  $\hat{Y} = j(\bar{X})$ . The corresponding loss function for  $net_X$  is,

$$\text{loss}_{net_X} = d_1(Y, \hat{Y}) + \beta D_{KL}(\mathcal{N}(0, 1) | q(\bar{X} | X)) + \gamma d_2(\bar{Y}, \hat{\bar{Y}}) \quad (1)$$

where  $d_1$  and  $d_2$  are appropriate distance metrics (in our case mean squared error),  $q(\bar{X} | X)$  denotes the activation distribution of bottleneck neurons over input samples, and  $D_{KL}$  denotes the Kullback-Leibler (KL) divergence.  $\gamma$  and  $\beta$  are hyperparameters used to weight different terms of the loss function. Similarly,  $net_Y$  performs the same compression, reconstruction, and estimate of relation to  $\bar{X}$ , but with respect to  $Y$ . The total loss is then  $\text{loss}_{net_X} + \text{loss}_{net_Y}$ .

## Supervised Causal Autoencoder

Due to the symmetry of the CAE (and mutual information in general), the causal direction between  $X$  and  $Y$  need not be known *a priori*. However, instead of post-processing macrovariables using transfer entropy, additive noise models, or other causal scoring methods (Guyon, Statnikov, and Batu, 2019) to determine causal direction, we restrict our application to the case where aspects of  $X$  cause aspects of  $Y$ , since we assume the causal direction would be known in most real-world decision-making scenarios. Given this restriction, we no longer require our macro-level functional causal model  $h$  to be constrained by a linear relationship. We relax this constraint by adding one extra neuron between each  $x_k$  and  $y_k$ , but keep the injective mapping between macrovariable dimensions in order to maintain strong and sparse correlations. We also convert the encoders and decoders to convolutional neural networks (CNNs) to better capture spatial image features. These changes enhance the stability of discovered causal macrovariables, as detailed in *Results*.

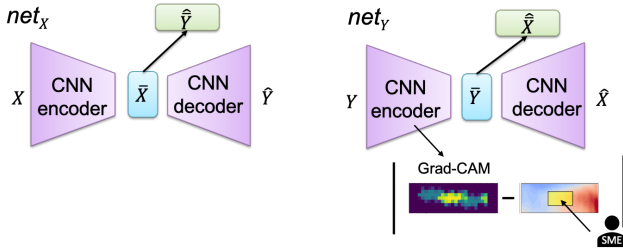


Figure 2: Top: For input  $X$ ,  $net_X$  learns a lower-dimensional embedding  $\bar{X}$  (causal macrovariable), from which it predicts  $\hat{Y}$  as well as the bottleneck layer  $\bar{Y}$  of  $net_Y$ . Similarly,  $net_Y$  (right) predicts  $\hat{X}$  and  $\bar{X}$  from  $Y$ . Bottom: A regularizing loss term is computed between annotated visual causal concepts and heatmaps produced by Grad-CAM.

SCAE builds upon the CAE by incorporating a regularization term which forces the encoders to pay attention to provided ROIs to derive visual cause and effect concepts within the  $X$  and  $Y$  images, respectively. The regularization provides a third trade-off in the loss function, beyond the standard compression and reconstruction for  $\beta$ -VAEs. As stated in Shen et al. (2022), this form of supervision has the

advantage of allowing for semi-supervision, as opposed to latent conditional models which require supervision to apply to every sample. The supervision term is implemented using a Grad-CAM heatmap (Selvaraju et al., 2017) to identify the regions that each  $x_k$  or  $y_k$  is focusing on within each ( $X_i$  or  $Y_i$ ) image. Other explainable artificial intelligence methods that produce heatmaps (e.g., SHAP features of Lundberg and Lee (2017)) could also be used. We compute the  $L1$ -norm between this heatmap and a binary image mask of the annotated ROI as follows,

$$\phi(Y_i) = |A_k(Y_i) - \alpha(Y_i)|_1, \quad (2)$$

where  $A_k(Y_i)$  is the Grad-CAM heatmap for  $Y_i$  with respect to the supervised concept  $y_k$ , and  $\alpha(Y_i)$  is an image mask annotating the corresponding visual concept in that same image. The loss function for  $net_Y$  becomes,

$$\text{loss}_{net_Y} = d_1(X, \hat{X}) + \beta D_{KL}(\mathcal{N}(0, 1) | q(\bar{Y} | Y)) + \lambda \sum_{i \in S} \phi(Y_i) \quad (3)$$

where  $\lambda$  is the weight of the supervised loss term. Due to the symmetry of the architecture, we can use this term to impose supervision on either  $net_X$  or  $net_Y$ , or both. The idea is similar to the Grad-CAM-based regularization applied in Pillai and Pirsivash (2021), where it is used for self-supervision via data augmentation.

To train the SCAE-CNN architecture, we use multiple phases. In Phase 1, we train the network for  $n$  epochs with only the  $d_1$  and  $\phi$  loss terms (setting  $\beta$  and  $\gamma$  to zero). In Phase 2, we set  $\beta$  to its optimal value for the next  $n$  epochs, while keeping  $\gamma$  at zero. Finally, in Phase 3 we set  $\gamma$  to its optimal value (as in the CAE) and train the network for those remaining terms. This allows the network to first pick up the supervision information, while minimizing training inconsistencies. As mentioned in Höltgen (2021), causal macrovariables cannot be unique because there are infinitely many transformations of derived macrovariables for which causal relationships hold; however, training with supervision in Phase 1 allows us to guide the network to find the macrovariable representation that best captures the ROIs specified by the domain expert. The network is less likely to get stuck in a local minimum when we start with strong guidance on the proper gradient direction, and this also enhances the stability of the training process.

## Experiments

We evaluate our method with two experimental datasets: the synthetic dataset (with causal ground-truth) introduced in Höltgen (2021), and the climate dataset introduced in Chalupka et al. (2016). In both datasets, we empirically show that given enough supervision, our supervised framework is able to perform the aforementioned causal coarsening.

**Datasets and Metrics** As is often necessary for evaluation of novel causal discovery methods, we simulate a synthetic dataset for which the causal ground truth is known. In

this synthetic dataset,  $X$  and  $Y$  are random variables in  $\mathbb{R}^{64}$  which can be visualized as  $8 \times 8$  pixel gray-scale images. The generative model for the data has two-dimensional macrovariables:  $\bar{X} = (x_1, x_2)$  and  $\bar{Y} = (y_1, y_2)$  which correspond to pixel value averages in the top/bottom halves of  $X$  and the right/left halves of  $Y$ , respectively. Scalar values per sample are generated as

$$x_1 := c_1 + n_1^X, \quad y_1 := c_1^3 + n_1^Y, \quad y_2 := \tanh(x_2) + n_2^Y$$

with unobserved confounder  $c_1$  and  $x_2 \sim U([-1, 1])$  and  $n_1^X, n_1^Y, n_2^Y \sim U([-0.2, 0.2])$  all uniformly distributed and mutually independent. Pixel values within each half of each image are generated by distributing the appropriate scalar value shift (plus noise) across all but one pixel, then shifting the last pixel by the remaining amount. As shown in Figure 3,  $x_1$  and  $y_1$  have a common cause  $c_1$  (i.e., they are correlated but  $x_1$  does not directly cause  $y_1$ ), whereas  $x_2$  is a direct cause of  $y_2$ . Supervision is imposed by highlighting the left half of  $Y$  images corresponding to  $y_2$  (yellow, Figure 3) indicating that this is an effect of interest.

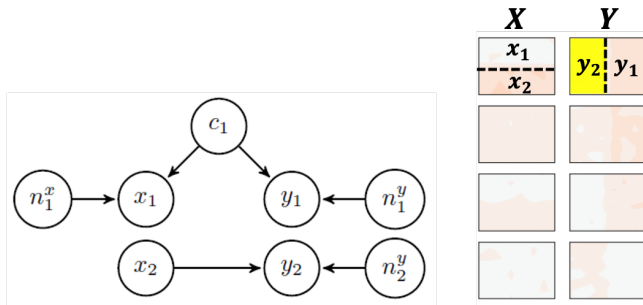


Figure 3: Synthetic dataset, adapted from Höltingen (2021). Left: Generative model. Right: Four examples each of  $X$  and  $Y$ , which are  $8 \times 8$  gray-scale images. The causal effect of interest corresponds to  $y_2$  (yellow).

The “El Niño Dataset” is the real-world climate science dataset from *Introduction*. This dataset assumes that macrovariables within zonal WS maps ( $X$ ) cause macrovariables within SST maps ( $Y$ ). The goal of causal feature learning in this domain is to discover the El Niño phenomenon from 35 years of climate data from a region in the Equatorial Pacific ocean. The dataset enables us to compare our results with the experimental results of both CFL and CAE.

Although we cannot obtain ground-truth causal macrovariables for this dataset, Chalupka et al. (2016) introduce a precision metric to measure the quality of the resulting macrovariables based on known qualities of El Niño. The National Oceanic and Atmospheric Administration (NOAA) defines El Niño as a positive three-month running mean SST anomaly of more than  $0.5^\circ C$  in the Niño 3.4 region. Therefore, in CFL precision is computed per macrovariable category as the percentage of  $Y$  observations with a spatial average within the Niño 3.4 region that is  $> 0.5$  warmer than the dataset mean. In SCAE, a macrovariable  $\bar{Y}$  is a continuous-valued  $K$ -dimensional vector which can represent a mixture of multiple “effect” concepts. We consider an image  $Y_i$  to “contain”

an effect concept  $y_k$  if it produces an activation value that lies within the top 25th percentile of  $y_k$ ’s activation values (i.e.,  $\{Y_i \mid g_k(Y_i) > P_{75}(g_k(\mathbf{Y}))\}$ ). Precision is then computed per concept over the set of images containing that concept.

Precision is computed similarly for the synthetic dataset; for each  $y_k$  we calculate the percentage of  $Y_i$ s containing that  $y_k$  which have a spatial average within the left half of the image that is greater than the average of the total dataset.

Qualitatively, climate scientists have noted that a larger-than-average westerly wind component in the west-equatorial region is a feature associated with the causes of El Niño (Di Liberto, 2014). This suggests that there should be a causal concept representing WS maps with this characteristic mapped to the effect concept representing El Niño. In this dataset, we provide the Niño 3.4 region as an image mask to guide our construction of causal macrovariables (highlighted in yellow, Figure 1).

**Implementation details** For the synthetic dataset, we use a 2-layer CNN ([input, output] channels: [1,8], [8,16]) and a 3x3 kernel. For the El Niño dataset we use a 3-layer CNN ([input, output] channels: [1,32], [32,64], [64,64]) and a 3x3 kernel. As with most neural architectures, the number of layers has a strong impact on performance; too many layers will overfit to the training data and too few will underfit.

Following the CAE paper, we conducted a grid-search for  $\gamma, \beta$  ( $\gamma = 1, 0.1, 0.01, \beta = 1, 0.1, 0.01$ ) in order to find the best result for both our architecture and the CAE. The total number of bottleneck neurons was set to the same as in CAE: 4 for the synthetic data and 16 for the El Niño data (this parameter was not searched). In the synthetic dataset, we trained 1000 epochs (100, 100, 800 for Phases 1,2,3) with hyperparameters  $\gamma = 0.1, \beta = 0.01$ , and  $\lambda = 10^{-6}$ . For the El Niño dataset, we train a total of 2000 epochs (200, 200, 1600 for Phases 1,2,3) using hyperparameters  $\gamma = 1, \beta = 0.01$ , and  $\lambda = 10^{-6}$ . As in the CAE, we consider bottleneck neurons as “informative” (representing a causal concept  $x_k$  or  $y_k$ ) if the standard deviation of their activation distribution across samples is  $> 0.95$ . We additionally filter out bottleneck neurons which are only informative in *either*  $net_X$  or  $net_Y$ , since these represent information which is needed to reconstruct  $\hat{X}$  or  $\hat{Y}$ , but is not relevant to the causal connection between the two datasets. This is akin to the idea of preserving the “content” and discarding the “style” in the representation learning literature. Because we are asking the model to reconstruct  $Y$  from information in  $X$  (and vice versa), it is unsurprising that the optimal  $\beta$  values for both datasets are significantly less than 1, implying that we are favoring reconstruction over constraining the capacity of the bottleneck in order to achieve a low overall loss.

**Results: Synthetic Dataset** The relaxation of linear constraints on the  $\bar{X}$  to  $\bar{Y}$  mapping, and addition of CNN encoders and decoders contribute to the stability of the results across random initializations. To quantify this, we ran 10 random network initializations without supervision on the synthetic dataset and computed the mean and standard deviation of the number of informative concepts  $K$  (Table 1). Since

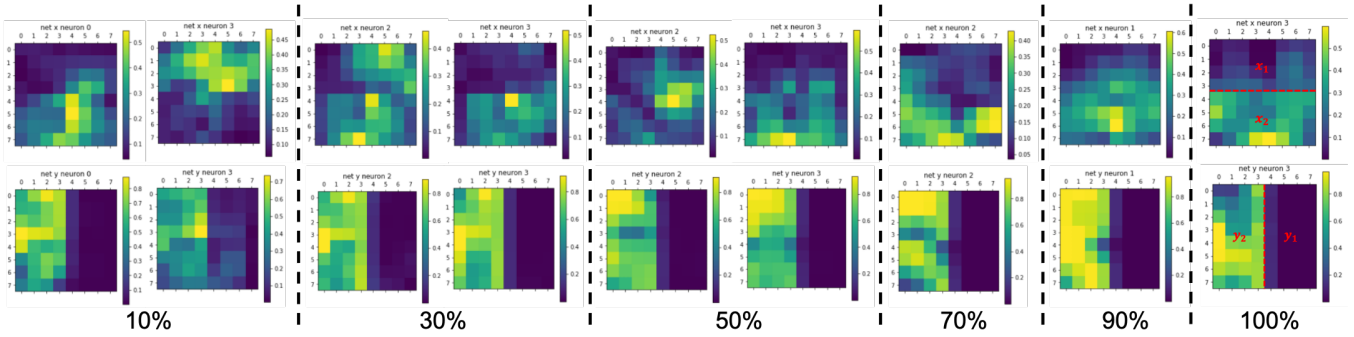


Figure 4: Synthetic dataset Grad-CAM results from semi-supervision for  $net_x$  (top row) and  $net_y$  (bottom row). Each of the dashed vertical lines indicate a different proportion of image masks provided to the system, according to the percentage below. Greater than 50% supervision is needed in order to collapse the two correlated concepts to the single ground truth causal relationship ( $x_2 \rightarrow y_2$ ).

we are not imposing supervision, we expect the system to discover two correlation-based concepts ( $K = 2$ ). Although the results in the table are not statistically significant, we find that the mean is closest to 2 and has the least variability when using the CNN and nonlinear adjustments.

	NN	CNN
Linear	1.67 +/- 0.47	1.7 +/- 0.45
Nonlinear	1.67 +/- 0.47	1.9 +/- 0.3

Table 1: Mean +/- standard deviation of  $K$  in an unsupervised setting across 10 random initializations.

Subject matter expertise, such as highlighted image regions, may be expensive or otherwise difficult to obtain. Figure 4 shows the Grad-CAM result from varying the percentage of annotated image masks per batch. The visual concepts represented by  $\bar{X}$  and  $\bar{Y}$  seem to vary smoothly, but  $K$  drops from two to one (the ground truth value) when 50% of images (5000 samples) are supervised. Using annotations on 100% of  $Y$  images, SCAE correctly detects one causal macrovariable relationship  $x_2 \rightarrow y_2$  in the synthetic dataset. Although we have only supervised  $net_y$  to focus on  $y_2$ , we observe that  $net_x$  has also been driven to focus on  $x_2$  due to the inter-connectivity of the overall loss function. Figure 5 further demonstrates further that the concepts corresponding to neuron 3 of both bottlenecks, correctly capture  $x_2$  and  $y_2$  from the ground truth model. Concepts  $x_1$  and  $y_1$  are still captured by neuron 0 of  $net_x$  and neuron 2 of  $net_y$  respectively, however, their functional relationship (correlation) has been broken.

Visualizations of discovered concepts for the synthetic dataset are omitted to save space, since they look very similar to the Grad-CAM heatmaps shown in Figure 4. Precision metrics show that causal concepts are clearly differentiated by whether the spatial average of values within the  $y_2$  region fall above or below the dataset average. CAE discovers macrovariables which capture both the correlative relationship between  $x_1$  and  $y_1$  (precision= 0.5, since these are random with respect to  $y_2$ ) and the causal relationship between  $x_2$  and  $y_2$

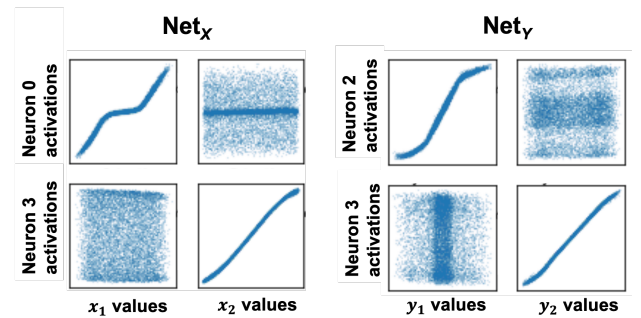


Figure 5: Scatter plots of  $net_x$  and  $net_y$  informative bottleneck neuron activation values versus the ground-truth variable values of the synthetic dataset. This indicates that neuron 3 of  $net_x$  and  $net_y$  does indeed capture  $x_2$  and  $y_2$ , respectively.

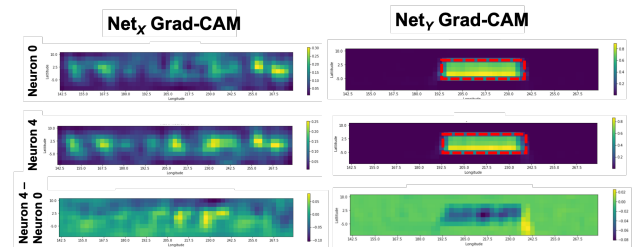


Figure 6: Grad-CAM heatmap for informative bottleneck neurons 0 and 4 for El Niño data (Left:  $net_x$ , Right:  $net_y$ )

(precision= 1.0). SCAE coarsens these such that only the causal relationship between  $x_2$  and  $y_2$  (precision= 1.0) is preserved.

**Results: El Niño Dataset** Using Niño 3.4 annotations on 100% of  $Y$  images, SCAE detects two causal concepts for the El Niño dataset. We do not know the correct  $K$  in this case, but attempt to judge the quality of our results based on the characteristics described in *Datasets and Metrics* as

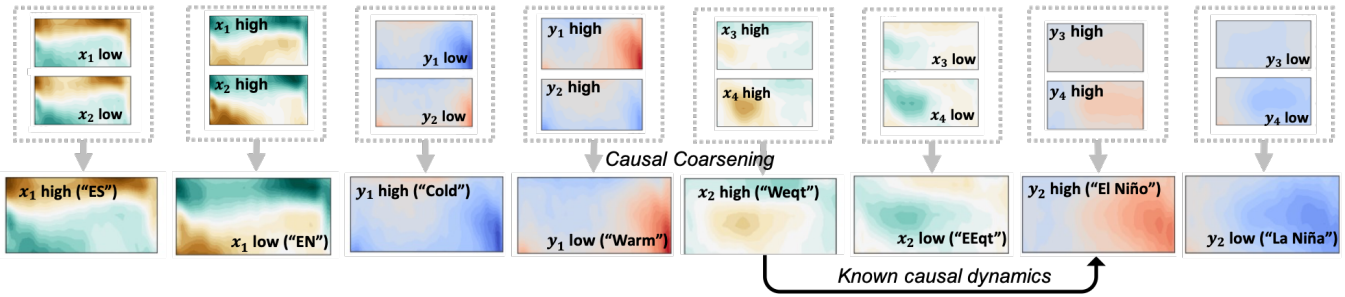


Figure 7: Visualization of  $K = 4$  concepts discovered by CAE (Top) and  $K = 2$  concepts discovered by SCAE (Bottom) for the El Niño dataset. Gray boxes indicate coarsening of correlational concepts into causal ones. SCAE results are also labeled using CFL discovered categories for WS and SST: Easterly Equatorial (EEqt), Westerly Equatorial (WEqt), Easterly North of Equator (EN) and Easterly South of Equator (ES), Cold, El Niño, La Niña, and Warm.

compared to the four macrovariables discovered by CAE and CFL. Fig. 6 shows the final Grad-CAM heatmap for neuron 0 ( $x_1, y_1$ ) and neuron 4 ( $x_2, y_2$ ). Macrovariables from  $net_X$  seemingly capture deviations in wind speeds across the image frame. By taking the difference in Grad-CAM heatmaps (bottom row), we can see that Neuron 4 focuses more on westerly equatorial winds. This agrees with the results in Figure 7, i.e., the more extreme “WEqt” and “El Niño” concepts correspond to  $x_2$  and  $y_2$ . Discovered concepts from  $net_Y$  both strongly focus on temperature deviations within the Niño 3.4 region, with negligible differences.

We display discovered visual causal concepts by taking the top and bottom 25th percentile of their activation values, followed by the mean of the images that produced them, and then subtracting off the mean image for the whole dataset. Figure 7 shows these mean-subtracted average images with respect to each concept for CAE (Top) and SCAE (Bottom). Inset text in quotes provides closest mapping to the semantic labels of macrovariables discovered by CFL in Chalupka et al. (2016), where the number of macrovariables  $|\bar{X}| = |\bar{Y}| = 4$  must be specified *a priori*. According to the precision metric, SCAE’s  $y_2$  is able to detect El Niño with 100% precision, as opposed to the maximum 85% precision from CAE’s  $y_4$ . SCAE does this while preserving the known causal relationship between Westerly Equatorial Winds (“WEqt”) and the occurrence of the El Niño phenomenon. However, the more remarkable result is that SCAE is able to achieve this improved performance using high and low activation values for only 2 cause-effect concepts, as opposed to the four visually redundant macrovariables from CFL or CAE. We attribute this to the use of continuous-valued macrovariables and supervision for causal coarsening, respectively.

## Discussion

The existence of causal macrovariables in any microvariable system is not guaranteed. However, we have shown that in cases where macro-level descriptions of micro-level dynamics are assumed to exist, it is possible to utilize some amount of human supervision in order to enhance the representation of discovered macrovariables such that they more concisely represent the known causal dynamics of the system. This work attempts to push the state-of-the-art in disentanglement

and causal representation learning by considering a high-dimensional  $Y$  where conventional class labels are undefined. This limits the ability to produce data augmentations which can be used for self-supervised learning, because we do not know (without human supervision) what types of augmentations preserve relevant effects in the  $Y$  images.

Our method of supervision is limited in its ability to specify causal factors which cannot be represented visually, as well as the need to completely annotate any image in  $S$ , i.e., the user cannot annotate only one of multiple causal factors present in the image. In addition, if there are multiple causal factors within one image, they must be annotated as separate image masks if we seek to disentangle them. Despite these limitations, we show that in a synthetically generated dataset (where the causal ground-truth is known), our methods are able to distinguish between correlation and direct causation, a differentiation which is otherwise unobtainable from purely observational data. In addition, we are able to capture the macrovariables corresponding to El Niño and La Niña with higher precision and more concisely than existing methods.

Further investigation into the amount of supervision required to obtain accurate causal macrovariables in various settings is surely needed. Although annotating thousands of input or output images in a dataset may seem tedious, this number should stay relatively constant as the size of the dataset increases, assuming stationarity of causal concepts. Our trained architecture can be used for automated prediction of effect macrovariables (e.g., El Niño) from any new input  $X_i$  (e.g., WS map), which saves time and effort in applications where subject matter experts currently analyze high-dimensional datasets manually. Towards this goal of enhancing practicality and operationalizing the system, we are also pursuing the use of eye-tracking data from human subject experiments as means of passive supervision.

## Acknowledgement

This material is based upon work supported by the Office of Naval Research (ONR) under Contract No. N0001419C2024. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research (ONR).

## References

- Beckers, S.; Eberhardt, F.; and Halpern, J. Y. 2020. Approximate causal abstractions. In *Uncertainty in Artificial Intelligence*, 606–615. PMLR.
- Chalupka, K.; Bischoff, T.; Perona, P.; and Eberhardt, F. 2016. Unsupervised discovery of el nino using causal feature learning on microlevel climate data. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 72–81.
- Chalupka, K.; Perona, P.; and Eberhardt, F. 2015. Visual causal feature learning. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 181–190. AUAI Press.
- Chen, R. T.; Li, X.; Grosse, R. B.; and Duvenaud, D. K. 2018. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems* 31.
- Di Liberto, T. 2014. The walker circulation: Ensos atmospheric buddy. *NOAA Climate.gov*.
- Eberhardt, F. 2022. Causal emergence: When distortions in a map obscure the territory. <https://simons.berkeley.edu/talks/causal-emergence-when-distortions-map-obscure-territory>. Simons Institute for the Theory of Computing - Causality Program, 2022-02-14, Accessed: 2022-06-04.
- Guyon, I.; Statnikov, A.; and Batu, B. B. 2019. *Cause effect Pairs in machine learning*. Springer.
- Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.
- Höltgen, B. 2021. Encoding causal macrovariables. In *Workshop on Causal Inference and Machine Learning: Why now? (WHY)*, NeurIPS.
- Locatello, F.; Tschannen, M.; Bauer, S.; Rätsch, G.; Schölkopf, B.; and Bachem, O. 2019. Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258*.
- Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30.
- Mitrovic, J.; McWilliams, B.; Walker, J.; Buesing, L.; and Blundell, C. 2020. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*.
- Pillai, V., and Pirsiavash, H. 2021. Explainable models with consistent interpretations. *UMBC Student Collection*.
- Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Towards causal representation learning 2021. *arXiv preprint arXiv:2102.11107*.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.
- Shen, X.; Liu, F.; Dong, H.; Lian, Q.; Chen, Z.; and Zhang, T. 2022. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research* 23:1–55.
- Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.
- Träuble, F.; Creager, E.; Kilbertus, N.; Locatello, F.; Dittadi, A.; Goyal, A.; Schölkopf, B.; and Bauer, S. 2021. On disentangled representations learned from correlated data. In *International Conference on Machine Learning*, 10401–10412. PMLR.