Representation Learning: A Causal Perspective

Yixin Wang

University of Michigan
yixinw@umich.edu

Representation learning focuses on reducing high-dimensional data into low-dimensional representations that capture essential features. For example, MNIST images, represented as 784-dimensional pixel vectors, can be summarized in fewer dimensions while retaining key information about the digits. Similarly, large text corpora, where each document is represented by a high-dimensional word count vector, can benefit from learning compressed representations. The goal is to map an m-dimensional data point $\boldsymbol{X} \in \mathbb{R}^m$ to a d-dimensional representation $\boldsymbol{Z} \in \mathbb{R}^d$, where $d \ll m$, while retaining the important features of the data.

Standard approaches often involve fitting neural networks or latent variable models such as variational autoencoders (VAEs) to create these low-dimensional representations. In supervised settings, a neural network learns to map high-dimensional data to labels, and the top layer is often taken as the data representation. In unsupervised settings, VAEs or similar methods use latent variables inferred from the data to generate low-dimensional embeddings. These learned representations are expected to perform well in downstream tasks and provide insights into the data's underlying structure.

However, these heuristic approaches frequently encounter problems. For instance, learned representations may capture spurious features that do not generalize well across datasets or tasks, or the representations may be entangled, meaning that a single dimension encodes multiple features, making interpretation difficult. An example would be a neural network trained to classify images of animals that learns to associate certain backgrounds (like grass) with the presence of specific animals (like dogs). Though this feature may correlate with the label, it is not causally related and is unlikely to transfer well to other datasets.

The issue of entangled representations is also significant. When different features (e.g., animal fur and background lighting) are encoded in the same representation dimension, it becomes challenging to interpret the representation or use it for further analysis. Such representations are less informative and harder to manipulate for generating new data.

While concepts like non-spuriousness and disentanglement are natural goals for representation learning, they are often difficult to define formally and even harder to measure and optimize. Without concrete metrics or labeled features, it is challenging to design algorithms that ensure these properties in learned representations.

In this work, we adopt a causal perspective on representation learning, which provides a formal framework to define and enforce these desiderata. Specifically, we use causal notions to define non-spuriousness and efficiency in supervised representation learning and disentanglement in unsupervised learning. These causal relationships between the label and the features captured by the low-dimensional representation $\boldsymbol{Z}$ yield calculable metrics that can be used to assess how well the representation meets these criteria.

In the supervised setting, we focus on non-spuriousness and efficiency. From a causal standpoint, a non-spurious representation captures features that are sufficient causes of the label. This ensures that the representation generalizes well to new datasets. For example, in the case of animal images, capturing the feature "dog-face" is non-spurious because it is sufficient to determine the label "dog," while a feature like "grass" is spurious because it only correlates with the label but does not causally determine it. Efficiency, on the other hand, ensures that the representation does not include redundant features. For instance, knowing both "dog-face" and "four-leg" is unnecessary since "dog-face" alone is sufficient to determine the label. We formalize these concepts using the probabilities of sufficiency (PS) and necessity (PN) from causal inference, providing a solid theoretical basis for assessing representation quality.

In the unsupervised setting, we extend this causal perspective to focus on disentanglement, which requires that different dimensions of the learned representation correspond to distinct, independent features. Disentangled representations allow for greater interpretability and manipulation, as one can vary individual features without affecting others. We formalize disentanglement by ensuring that different features encoded by the representation do not causally affect each other. Although these features may be correlated, they should not have causal connections.

A key challenge in implementing these causal definitions is that not all causal quantities are observable from the data. To address this, we study the observable implications of these causal criteria, a problem known as causal identification. In summary, this work illustrates how a causal perspective can formalize intuitively desirable properties in representation learning.