

structure, with \mathbf{h} as a latent variable that explains the observed variations in \mathbf{x} . The “ideal” representation learning discussed above should thus recover these latent factors. If \mathbf{y} is one of these (or closely related to one of them), then it will be very easy to learn to predict \mathbf{y} from such a representation. We also see that the conditional distribution of \mathbf{y} given \mathbf{x} is tied by Bayes’ rule to the components in the above equation:

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{y})p(\mathbf{y})}{p(\mathbf{x})}. \quad (15.3)$$

Thus the marginal $p(\mathbf{x})$ is intimately tied to the conditional $p(\mathbf{y} \mid \mathbf{x})$ and knowledge of the structure of the former should be helpful to learn the latter. Therefore, in situations respecting these assumptions, semi-supervised learning should improve performance.

An important research problem regards the fact that most observations are formed by an extremely large number of underlying causes. Suppose $\mathbf{y} = \mathbf{h}_i$, but the unsupervised learner does not know which \mathbf{h}_i . The brute force solution is for an unsupervised learner to learn a representation that captures *all* the reasonably salient generative factors \mathbf{h}_j and disentangles them from each other, thus making it easy to predict \mathbf{y} from \mathbf{h} , regardless of which \mathbf{h}_i is associated with \mathbf{y} .

In practice, the brute force solution is not feasible because it is not possible to capture all or most of the factors of variation that influence an observation. For example, in a visual scene, should the representation always encode all of the smallest objects in the background? It is a well-documented psychological phenomenon that human beings fail to perceive changes in their environment that are not immediately relevant to the task they are performing—see, e.g., [Simons and Levin \(1998\)](#). An important research frontier in semi-supervised learning is determining *what* to encode in each situation. Currently, two of the main strategies for dealing with a large number of underlying causes are to use a supervised learning signal at the same time as the unsupervised learning signal so that the model will choose to capture the most relevant factors of variation, or to use much larger representations if using purely unsupervised learning.

An emerging strategy for unsupervised learning is to modify the definition of which underlying causes are most salient. Historically, autoencoders and generative models have been trained to optimize a fixed criterion, often similar to mean squared error. These fixed criteria determine which causes are considered salient. For example, mean squared error applied to the pixels of an image implicitly specifies that an underlying cause is only salient if it significantly changes the brightness of a large number of pixels. This can be problematic if the task we wish to solve involves interacting with small objects. See figure [15.5](#) for an example