Semi-Supervised Disentangling of Causal Factors

Sargur N. Srihari
srihari@cedar.buffalo.edu

Topics in Representation Learning

1. Greedy Layer-Wise Unsupervised Pretraining
2. Transfer Learning and Domain Adaptation
3. Semi-supervised Disentangling of Causal Factors
4. Distributed Representation
5. Exponential Gains from depth
6. Providing Clues to Discover Underlying Causes

What makes one representation better than an other?

- Ideal representation is one where features within the representation correspond to the underlying causes of the observed data
  - With separate features or directions in feature space corresponding to different causes
    - So that the representation disentangles the causes from one another
- This motivates approaches in which we seek a good representation for $p(\boldsymbol{x})$
  - Which may also be good for representing $p(\boldsymbol{y}|\boldsymbol{x})$ if $\boldsymbol{y}$ is among the most salient causes of $\boldsymbol{x}$

Contrast with other representations

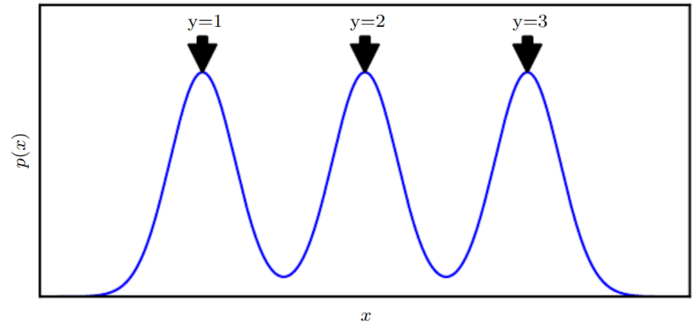
- We are usually concerned with a representation easy to model
  - E.g., independence, sparsity
- Representation that separates causal factors may not be easy to model
- However for many tasks the two coincide
- If a representation $\boldsymbol{h}$ represents many of the underlying causes of the observed $\boldsymbol{x}$, and the outputs $\boldsymbol{y}$ are among the most salient causes, then it is easy to predict $\boldsymbol{y}$ from $\boldsymbol{h}$

How semi-supervised learning can fail

- When is $p(\mathbf{x})$ if of no help to learning $p(\mathbf{y}|\mathbf{x})$?
- Consider where $p(\mathbf{x})$ is uniformly distributed and we want to learn $f(\boldsymbol{x})=\mathrm{E}[\mathbf{y}|\boldsymbol{x}]$
- Clearly observing the training set of $\boldsymbol{x}$ values alone gives us no information about $p(\mathbf{y}|\mathbf{x})$

How semi-supervised can succeed

- Ex: density over $\boldsymbol{x}$ is a mixture over three components, one per value of $\mathbf{y}$
- If components well-separated
  - modeling $p(\mathbf{x})$ reveals where each component is
    - A single labeled example per class enough to learn $p(\mathbf{y}|\mathbf{x})$



- What could tie $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x})$ together?
  - If $\mathbf{y}$ is closely associated with one of the causal factors of $\mathbf{x}$, then $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$ will be strongly tied
    - Unsupervised learning that tries to disentangle the underlying factors of variation is likely to be useful as a semi-supervised learning strategy

Formalizing best possible model

- Assume y is one of the causal factors of $\mathbf{x}$
- Let $\mathbf{h}$ represent all those factors
- The true generative process can be conceived as structured according to this directed model with $\mathbf{h}$ as the parent of $\mathbf{x}$: $p(\mathbf{h,x})=p(\mathbf{x})p(\mathbf{x}|\mathbf{h})$
  - Thus data has marginal probability $p(\boldsymbol{x})=\mathrm{E}_{\mathbf{h}}\, p(\boldsymbol{x}|\boldsymbol{h})$
- Thus we conclude that the best possible model of $\mathbf{x}$ is one that uncovers the above true structure with $\mathbf{h}$ as a latent variable that explains the observed variations in $\boldsymbol{x}$

Ideal representation learning

- It should recover the latent factors
- If y is one of these then it will be easy to predict y from such a representation
- We also see from Bayes rule: $\boxed{p(\mathrm{y}|\mathrm{x}) = \dfrac{p(\mathrm{x}|\mathrm{y})p(\mathrm{y})}{p(\mathrm{x})}}$
- Thus the marginal $p(\mathbf{x})$ is intimately tied to the conditional $p(\mathbf{y}|\mathbf{x})$
  - Knowledge of the structure of the former should help learn the latter
  - Therefore in situations respecting these assumptions, semi-supervised learning should improve performance

Brute force for large no of causes

- Most observations are formed by an extremely large no of causes
- Suppose $\mathbf{y}=\mathrm{h}_i$, but the unsupervised learner does not know which h_i
- The brute-force solution is for an unsupervised learner to learn a representation that captures *all* the reasonably salient generative factors h_j
  - and disentagles them from each other thus making it easy to predict $\mathbf{y}$ from $\mathbf{h}$ regardless of which h_i is associated with $\mathbf{y}$

Brute force is infeasible

- It is not possible to capture all or most of the factors of variation that influence the observation
- Ex: should the representation always encode all the smallest objects in the background?
- Research frontier in semi-supervised learning: *What* to encode in each situation

Two ways to deal with many causes

- Two main strategies to deal with a large no of underlying causes:

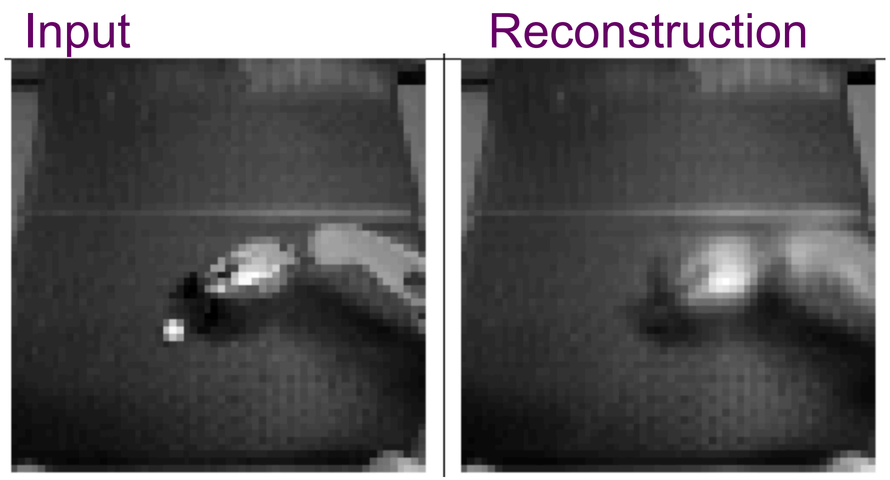
1. Use a supervised learning signal at the same time as the unsupervised learning signal so that the model will choose to capture the most relevant factors of variation
2. Use much larger representations if using purely unsupervised learning

Modifying definition of saliency

- Emerging strategy for unsupervised learning is to modify the definition of which underlying causes are most salient
- Autoencoders and generative models usually optimize a fixed criterion, say MSE
- These fixed criteria determine which causes are considered salient
  - Ex: MSE applied to pixels implies that an underlying cause is salient only if it significantly changes the brightness of a large no of pixels
    - Problematic if task involves interacting with small objects
      - Example next

Failure of salience detection

- Autoencoder trained with MSE for a robotics task fails to reconstruct a ping pong ball



  - The autoencoder has limited capacity and training with MSE did not identify ball as salient enough
  - Same robot succeeds with larger objects
    - Such as baseballs which are more salient according to MSE

Other definitions of salience

- If a group of pixels follows a highly recognizable pattern even if that pattern does not involve extreme brightness or darkness then that pattern could be considered salient
- One way to implement such a definition of salience is called generative adversarial networks (GANs)

GANs to detect saliency

- A generative model is trained to fool a feedforward classifier
- The feedforward classifier attempts to recognize all samples from the generative model as being fake and all samples from the training set as being real
- Any structured pattern that the feedforward network can recognize is highly salient.
- The networks learn how to determine what is salient

Models generating human heads

- Models trained to generate human heads neglect to generate the ears when trained with MSE
- But generate ears when trained with GANs
- Because the ears are not especially bright or dark compared to surrounding skin
- But their highly recognizable shape and and consistent position means the feedforward network can easily learn to detect them

Predictive generative network

- Importance of learning which features are salient

Ground Truth MSE Adversarial