STAT 8101: Applied Causality (Spring'22) Week 8: Mar 14, 2022 - Mar 20, 2022

On Disentangled Representations Learned from Correlated Data

Han Lin hl3199@columbia.edu MS in Computer Science

This paper provides comprehensive empirical study about learning representations from correlated data. Section 2 and 3 contain good illustrations of several aspects about why we need & how to resolve the problem of correlated data:

- Why such correlation might be problematic: The authors first start from ELBO loss, and mention that the prior for latent variables could only be factorized when they are independent. But in reality, they could be correlated due to direct causal relationship or from unobserved confounders. Therefore, if we still use the assumption that these priors are independent, then we could generate combinations of correlated factors by putting mass outside the training distribution.
- Why not encode strongly correlated factors into a single latent variable: The authors provides three reasons, It's nice to connect disentangled representations with robustness to distribution shifts.
- How to resolve these latent correlations: The author provides two methods, one needs few labels, and the other is a weakly supervised method by making some modification on beta-VAE. This weak supervised method seems interesting, and I'm planning to read this Ada-GVAE in more detail later.
- The pairwise entanglement metric proposed in the appendix B seems simple but effective.

Some other thinking: I'm currently interested in resolving causal confusions in imitation learning (Haan 2019), where the goal is to learn a good policy from the tuples of observational images and corresponding expert actions. It seems that VAE methods could be also effective there to decompose observational images into several latent variables. To resolve the spurious causal relationships, the author provides a procedure to mask out some portion of latent variables, to see if the resulting graph is still able to predict the correct action. But it seems that this is just a method similar to random dropout and does not really learn the causal relationships. The experiments in this Trauble's paper assumes that we already know which observational variables we have (azimuth, object size, etc), and then use VAE to learn latent variables, which is not the case for Haan's paper, where we do not know explicitly what are the observational variables. So I'm wondering how could we design some methods that could identify the causal relationship and do an effective filter of spurious correlations?