1 Prerequisites on Learning Disentanglement

1.1 Theoretical Impediments from Causality

Generative factors represent causal elements. Learning a disentangled representation of generative factors is then understood as causal inference. In accordance with the causal literature, we can make statements about the type of knowledge, that can be gained by the type of data provided. It turns out that from "raw" image data - raw data meaning images x sampled from $p(x)$, without further assumptions, it is impossible learn a disentangled representation z. We start with a primer for causal learning (sec. 1.1.1), outline which inductive biases are needed for disentanglement (sec. 1.1.2) and assess how one can instantiate such biases for disentangling the factors of shape and appearance in images (sec. 1.1.3, sec. 1.1.4)).

1.1.1 Causal Learning

Learning to infer causality is harder than statistical learning. We outline the basic problem for the case of two variables x_1, x_2: statistical learning aims at estimating probabilistic properties such as $p(x_1, x_2)$ or $p(x_2|x_1)$ from data. A well-known theme is that statistical correlation does not imply causation. Less well-known is Reichenbachs principle [1, 2], that states: if two random variables are statistically dependent, then there exists a third variable that influences both or a direct causal link between them (Fig. 1.1). In addition to estimating the probability distribution, also the causal structure has to be inferred [1].

We start with an intuitive example problem: How to learn the causal connection between a barometer and the weather? If the barometer is working well, there exists a clear correlation between the weather condition and the needle position. Given a dataset showing both barometer and corresponding weather condition, a capable machine learning algorithm will be able to capture this correlation. However, it will fail to understand the causal direction, since this is not possible from the data. Imagine how a human would go about solving this problem: Having a mechanistic model of the world he could reason about the precise causal mechanism relating weather to humidity to needle position. For example a model of: weather influencing air pressure influencing barometer needle position. What if one has no prior knowledge? A solution of childlevel simplicity is, to force the needle to move with a finger. Without the power of voodoo magic, the weather will not change. Hence causality has to go other way or via a third latent variable influencing both *i.e*. air pressure. To conclude, the strength of association (correlation) can be estimated with observational data alone, this can answer the question: how likely will it

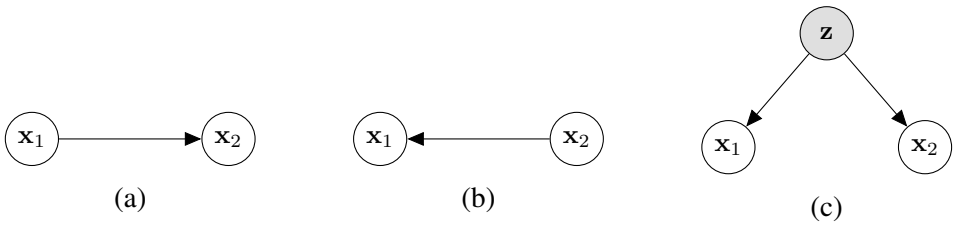

Figure 1.1: Correlation implies causation - if x_1 and x_2 correlate, a) x_1 may cause x_2, b) x_1 may be caused by x_2 or c) both are contingent on a latent cause z

rain, if the barometer needle sinks? But not: how would the weather change if I force the barometer needle to sink?

Pearl [3] distinguishes between three types of questions, that can be answered by different types of knowledge:

1. Association. What if I see . . . ?
2. Intervention. What if I do . . . ?
3. Counterfactual. What if I had done . . . ?

The levels of this *ladder of causation* [3] are separate not only conceptually, but in the type of data or assumptions that have to be made in order to access them. In particular, by unsupervised learning from observational data only the first level is accessible. The second level requires interactional data or model assumptions, while the third is inaccessible without an explicit model. The answers to these hypothetical questions (counterfactuals) lie by definition not in the data (facts).

1.1.2 Disentangling requires Interventions or Model Assumptions

The results from the study of causal inference also entail that "purely" unsupervised disentangling, *i.e.* estimating $\hat{\mathbf{z}}_\mathbf{i}$ from samples $x \sim p(x)$, is impossible. A rigerous proof for this can be found in [4]. Current machine learning operates mostly on the level of association, estimating (complex) correlations from raw data. As we have seen, this purely data-driven approach can only go so far. In contrast, humans seem to have the ability to interact with their environment and have innate assumptions on coherence, causality, physics etc., which introduce inductive priors. To bring *i)* interventions and *ii)* model assumptions to our problem of disentangling shape and appearance, we *i)* apply changes to an image, which are assumed to change only one factor and *ii)* model the causal process of the image generation in the theme of analysis-by-synthesis.

1.1.3 Image Transformation as Intervention

1.1.4 Analysis-by-Synthesis as Model Assumption

2 Bibliography

[1] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017. 1

[2] H. Reichenbach. *The Direction of Time*. University of California Press, 1956. 1

[3] Judea Pearl and Dana Mackenzie. *The Book of Why*. Hachette Book Group, 2018. 2

[4] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. *arXiv*, 2018. 2