

1 Introduction

Computer vision is the endeavour to algorithmically understand patterns in images. Structures and processes in the physical world interact in complex ways to generate an image. The image is then a mirror, in which these elements of the world are reflected and patterned. To recognize the patterns in an image, means to take this mirror as a window to understand and observe reality, *i.e.* to observe and measure the causal elements that contributed to the image generation. Formally this can be posed as an inference problem: a number of latent variables z_i interacted in certain ways to cause the image x . An inference algorithm aims at recovering the states of these variables from an image. The recovered state can be seen as an estimate or representation of the true latent variable. Typically, objects appear in an intricate interaction of many factors of variation. For

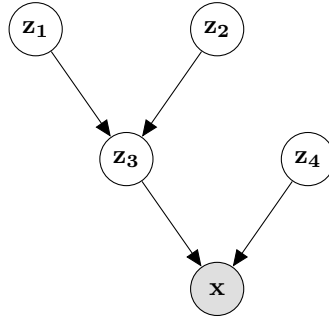


Figure 1.1: Causal elements z_i contributing to generation of image x

example, given the object class of people, variation can be in visual appearance such as the persons clothing or skin color or variation in geometric shape determined by a persons pose or body physique. In order to gain a conceptual understanding of the world, disentangling the underlying factors of variation is a crucial step, as has been argued in numerous works, Desjardins et al. [2012], Bengio et al. [2013], Chen et al. [2016], Higgins et al. [2017], Eastwood and Williams [2018]. For articulated object classes the most prominent factors are geometric shape and visual appearance. Disentangling these factors is a difficult problem due to the intricate interplay of shape and appearance under articulation. The complexity enters, as a variation in shape is a change of the images domain rather than a change of its values Shu et al. [2018]. Consider a person raising his arm: the color and texture of his pullover sleeve intrinsically does not change, but appears at a different location in the image. An efficient model for shape should cover all possible states of the object and preserve the local linkage to its intrinsic appearance.

Finding an abstract image representation, can be framed as a problem of understanding, such that an image representation contains the states of the factors that led to the image including objects and the physical states of objects. A factorized representation

should then represent each causal element and its state individually, in a disentangled manner: A change in the real causal element should correspond to an equivalent change in the abstract representational factor, while leaving the other factors, that represent other causes, unchanged. On the one hand there are pragmatic reasons to aim at extracting disentangled factors from images: to successfully transfer a representation between different tasks, typically only a few factors are relevant ?. Efficient transfer and multi-task learning should account for this. On the other hand, learning to capture external mechanisms in appropriate internal representations, can be seen as a goal in its own. It enables machines to reason about the world Pearl [2018]. In addition, once disentangled, a factor can be manipulated individually to make a targeted change. Thought experiments like *"imagine, how ridiculous you would look, if you wore that hot pants"* are manageable tasks for human imagination, but are out of the league for currently used generative image models Goodfellow et al. [2014], Kingma and Welling [2013], that typically rely on uninterpretable vector spaces with entangled dimensions. In the sense of generative modelling, disentangling factors could as well lead the way from a science of images to a science of imagination Mahadevan [2018].

But how to learn a disentangled representation from scratch, *i.e.* from pure data? As we will find out, disentangling causal factors from raw image data, without any side information is impossible theoretically and can only work based on statistical assumptions. Lets consider an example to illustrate this point: Statistical residues \rightarrow will depends on statistical nature of data (e.g. Gaussian with two dimensions $P(s, a)$ current machine learning: association, probability distribution modeling.

How do humans disentangle? 1. association: "conditioning" 2. Apart from pure association \rightarrow access to video information e.g. video information: how do objects behave across time?

3. Learning by interacting: knowing change by changing. second rung on causal ladder (Pearl): intervention. (\cdot , acting) What happens if I do? $P(s, do(a))$ Others: counterfactual (imagining), association. In humans e.g. egomotion cues: how does image on retina change if I move.

How disentangle? change factor \rightarrow image change equivariantly, leave others invariant \rightarrow equivariance, invariance

change can be mimicked artificially Intelligent pattern recognition algorithms, fuelled by sensory data as learning material alone, may ultimately drive the way to a full-blown artificial intelligence, reasoning about the world on its own. - That is the reasoning behind data-driven and assumptionless machine learning approaches that have conquered several research communities. A theoretical objection to driving-only-with-data comes from the causal literature: For an understanding of the world, an algorithm needs to model causal processes, that cause an image to be generated.

1.1 Contributions I

Hypothesis: learning shape requires abstracting away appearance \rightarrow hence disentangling
Hypothesis ii): learning disentanglement from pure data is fundamentally constrained.

need to take causal literature into account -> disentangling causal factors will need assumptions on causal model and/or interventional/interactional data (instead of raw data).

- validate and evaluate method developed by Lorenz *et al.* 2018 for disentangling
- overview over state-of-the-art disentangling, analysis of future directions
- explain method in context to these
- evaluate unsupervised shape learning:
 - human faces, bodies (CelebA, Human3.6M)
 - animal faces, bodies (cats, dogs, birds)
 - composite objects (dancing pair)
- make own video dataset
 - for disentangling human pose and appearance (heidelbergpose)
 - for articulated animal video (dogs)
 - for composite object (pair dancing salsa)
- ablation study (reconstruction, equivariance loss, transformations)
- qualitative comparison to non-disentangling composite shape learning (Zhang) may be a petty detail or a simple hack/trick (reconstruction on other image) but makes all the difference in terms of causal information
- evaluating disentanglement
 - reID
 - pose estimation

result: soa in shape learning, (first) unsupervised disentangling of articulated shape and appearance

2 Bibliography

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *TPAMI*, 2013.

Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *NIPS*, 2016.

Guillaume Desjardins, Aaron Courville, and Yoshua Bengio. Disentangling factors of variation via generative entangling. *arXiv preprint arXiv:1210.5474*, 2012.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. *ICLR*, 2018.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2013.

Sridhar Mahadevan. Imagination machines: A new challenge for artificial intelligence. In *AAAI*, 2018.

Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. In *WSDM*, 2018.

Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Güler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, 2018.