1 Introduction

Computer vision is the scientific endeavour to algorithmically understand patterns in images. Structures and processes in the physical world interact in complex ways to generate an image, the image acts as a mirror, in which these elements of the world are reflected and leave patterns. To recognize these patterns in an image, means to use this mirror as a window to observe the reality lurking behind it, *i.e.* to measure the causal elements that contributed to the image generation. Formally this can be posed as an inference problem: a number of latent variables $\mathbf{z_1} \dots \mathbf{z_N}$ has interacted in certain ways to cause the existence of the observed image $\mathbf{x}$. An inference algorithm aims at recovering these latent variables from the observation, *i.e.* the image. These recoveries can be seen as an estimate $\hat{\mathbf{z}}_\mathbf{i}$ for - or a representation of - the true latent variables $\mathbf{z_i}$. A graphical model of the process is shown in figure 1.1. A disentangled representation should then represent each causal element and its state independently: A change in the real causal element $\mathbf{z_i}$ should correspond to an equivalent change in the abstract representational factor $\hat{\mathbf{z}}_\mathbf{i}$, while leaving the other factors $\hat{\mathbf{z}}_\mathbf{j}, j \neq i$, that represent other causes, unchanged.

Typically, objects appear in an intricated interaction of many factors of variation. multiple levels of interaction For 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. For articulated object classes the most prominent factors are geometric shape and visual appearance. Disentangling these factors is a difficult problem due to the intricated 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.

1.1 Why disentangle causal factors?

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. Once disentangled, a factor can be manipulated individually to make a targeted change. This enables machines to reason about the world Pearl [2018], by simulating changes to factors internally in their model of the world. Thought experiments like *"imagine, how ridiculous you would look, if you wore that hot pants"* are managable tasks for human imagination, but are out of

Figure 1.1: (a) Causal elements $\mathbf{z_i}$ contributing to generation of image $\mathbf{x}$, (b) Shape $\mathbf{s}$ and appearance $\mathbf{a}$ influencing the image of an object.

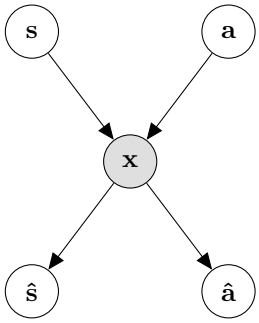

Figure 1.2: Disentangling causal factors: *e.g.* inferring an estimate of object shape $\hat{\mathbf{s}}$ and appearance $\hat{\mathbf{a}}$ from an image

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].

1.2 How Not to Disentangle?

But how to learn a disentangled representation from scratch, *i.e.* from raw image 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 abstract example to illustrate this point: Given an image dataset of human persons that has strong variation in the pose and in the appearance of the persons, how to find these two underlying axes of variation (pose, appearance)? Lets suppose the distribution of varation follows a two-dimensional Gaussian distribution, one dimension for pose, one for appearance. The learning algorithm has access to random samples from this distribution. An intelligent data compression algorithm (such as a variational autoencoder) will be able to fit a function from the images to the two dimensional subspace which explains (by assumption in this example) all of the varition in the dataset. statistical results

Statistical residues -> will depends on statistical nature of data (e.g. Gaussian with two dimensions P(s, a) current machine learning: association, probability distribution

modeling.

1.3 How do humans disentangle?

First it should be noted that this question is not

and we have to speculate about the nature of...

1. association: "conditioning" 2. Apart from pure association → 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. (, 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. *Interaction is crucial.* → barometer example: How to find out the causal connection between a barometer and the weather. There cannot be an abstract intelligence, which finds out about the world purely by observation. The intelligence has to interact with the world, it has to be in the world. before this becomes too philosophical infer causation from correlation RCT

1.4 How to disentangle?

change factor → image change equivariantly, leave others invariant → 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.

1.5 Contributions I

Hypothesis: learning shape requires abstracting away appearance -> 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

<<<<< HEAD

Part I

Appendix

A Bibliography

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.

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.
=======

B Bibliography

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.

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.
»»»> intro