1 Introduction

Computer vision is the scientific endeavour to algorithmically understand patterns in images. Structures and objects in the physical world interact in complex ways to generate an image. The image then acts as a mirror, in which these elements of the world are reflected and leave patterns. To recognize patterns in an image, means in essence to observe the reality lurking behind this mirror, *i.e*. to measure the causal elements that contributed to the image generation.

1.1 Why Disentangle Causal Factors?

Figure 1.1: *"Imagine, how ridiculous you would look if you wore that hot pants"* - Thought experiments are a targeted manipulation of a disentangled representation.

So far, we framed disentangling generative factors as similar to a scientific measurement process. An interaction of physical elements in the world is captured in an image, which can be treated as a scientific measurement of reality. Discerning patterns, and disentangling sources of patterns from each other, will then be defined as *understanding* the world. What can be gained from such an *understanding*?

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 [1]. 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 step to automate visual reasoning itself. Once disentangled, a factor can be manipulated individually to make a targeted change in an image. Thereby, not only humans may change images at will, but also machines may reason about the world [2], 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"* (cf. Fig. 1.1) are manageable tasks for the human imagination, but are out of the league for currently used generative image models [3, 4], that typically rely on non-interpretable vector spaces with entangled dimensions. Building imagination machines has been proposed as a goal for artificial intelligence research recently [5]. To imagine, is to manipulate of an internal model to generate internal images. In this sense, in the context of generative modelling, disentangling factors could as well lead the way from a science of images to a science of imagination.

1.2 How not to Disentangle.

Figure 1.2: The image captions are generated by a deep neural network (Neuraltalk2) [6]. Yet, common sense understanding of psychological and physical entities in terms of causal relationships and narratives is absent [7]. Instead, the neural network seems to capture mere associations.

Can machines tell a story? Carefully observe your own mind, when viewing the images shown in Fig. 1.2: observe how the human mind immediately interprets and jumps to conclusions, tries to tell itself a story that explains an image, whereas the machine (in this case, NeuralTalk2 [6]), is comically descriptive in contrast. The missing *common sense* may be due to a missing causal reasoning, due to a missing disentangled causal representation of the world. But how to learn a disentangled representation from scratch, *i.e.* from raw image data? - As we will find out (in Sec. **??**), disentangling causal factors from raw image data, without any side information is impossible theoretically, but can only work with model assumptions or interventional data.

Let us consider an example to visualize the fundamental problem: 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 and appearance)? Let us suppose the distribution of variation follows a two-dimensional Gaussian distribution, one dimension for pose, one for appearance. The learning algorithm has access to randomly sampled images from this distribution. An optimal data compression algorithm will be able to fit a function from the images to the two-dimensional subspace, which explains (by assumption in this example) the variation in the dataset. But are the two

dimensions, that the algorithms finds disentangled? No. In fact, any linear combination of pose and appearance axes and its orthogonal complement are equally valid to span the subspace of underlying variation. Just from observing a two-dimensional Gaussian, no meaning will be attached to the axes. In practice, this problem is often circumvented by first fitting a generative model to the image dataset and *afterwards* interpolating in the latent space to determine (by human judgment) the axes of interest (here the pose or appearance axis). The meaning of pose and appearance as independent factors comes from the fact, that it is easily possible in the real world to change one factor without the other. A person that walks (hence changes shape) without loosing clothes on the way (not changing appearance) is a trivial example for that.

In summary, on the basis of dataset statistics alone one cannot disentangle causal factors, if the information about how to select the axes, *i.e*. which factors to separate, is not contained in the raw data. Fitting a model to the data distribution, does in general not give insight into how the data was generated.

1.3 How can Humans Disentangle?

The dichotomy between humans and machines is constructed, of course, because on a fundamental level humans are machines. But in this context, the distinction between humans and machines shall refer to the current gap between human and machine learning performance, in terms of inferring generative factors and reasoning (again, cf. Fig. 1.2). So, what advantageous characteristics does the human mind have, that are lacking in data-driven machine learning algorithms?

Priors. Whether acquired or inherited, certain inductive priors seem to guide the human learning in its early phases [7]. Archetypal knowledge of psychology [8], a universal grammar for language [9] and causal intuitions on everyday physics [10] are some of the cognitive priors, that could explain the intuitive psychology, the rapid language acquisition and the remarkable causal inference from limited amount of data.

Data. Not only quantity, but also quality of data. Machine learning on images is commonly posed as the task to learn from randomly sampled images from a data set. But humans do not perceive the world in arbitrary samples. To humans, the world appears in a temporal sequence, which reveals how generative factors change and persevere across time. Instead of focusing on datasets with static images, sampled at random so that the images may have nothing to do with each other, algorithms should use video datasets and harness the rich temporal information.

Another key difference is, that humans interact with their environment. That means, humans know how factors change, not only by observing them changing, but also by changing them. Anyone, who has watched a human infant play, can affirm that the learning mind is obsessed with interaction and change. The inevitable destruction around a young human is no accident, but a result of curious learning. Whether by performing consciously controlled experiments or by subconscious cues [11]: Interaction is crucial for a learning mind.

Models. Humans are able to imagine. That presupposes an internal model of the world,

to which specific changes of representational factors can be applied. In machine learning, fitting neural network models as functions to approximate datasets has seen tremendous progress recently, to the point, that it is considered a solved problem. This progress is mainly due to the effectiveness of neural networks to fit high-dimensional functions. But a probabilistic fit to a dataset, however complex and rich, is not a causal model. Even if one were to obtain a probabilistic model over all images the world (one could start with *e.g.* ImageNet [12]), this would tell very little about the real-world (causal) relationships between objects.

What can we learn from these differences? An algorithm to understand the world: should contain useful *prior* assumptions to efficiently use *data* that contains the necessary causal relationships and interactions, to learn a useful *model* of the world.

1.4 Problem Formulation

In this thesis, we focus our efforts on image datasets which feature a single object class. Typically, objects appear in an intricate interaction of many causal factors, that can vary. For example, in an image dataset of people, the persons can have a different clothing, skin color, body physique and posture. The image generation process itself may further add factors like illumination, viewpoint or contrast. The manifold factors of possible variation in objects can be classed into to prominent categories: the objects geometric versus its visual properties. The object's geometry subsumes factors like pose, viewpoint and physique to all of which we collectively refer to as *shape*. The object's visual properties are the complement of shape. They encompass qualities like color, texture and shading and are referred to as *appearance*. Disentangling shape from appearance is a difficult problem, due to their intricate interplay, especially under heavy object articulation. Consider a person raising an arm: the color and texture of the pullover sleeve intrinsically does not change, but appears at a different location in the image. The complexity for modelling this interaction enters, as a variation in shape is a change of the domain of appearance rather than a change of its values [13]. Even simple shape models [14] offer difficult optimization problems. An efficient model for shape should cover all possible states of the object and preserve the local linkage to its intrinsic appearance. For this we partition both shape and appearance.

1.5 Contributions

This thesis provides evidence for two hypotheses:

- *Hypothesis* i)*: Unsupervised learning of object shape benefits from abstracting away the complement of shape, namely the object appearance. Explaining away the appearance factor can be achieved by a disentangled generative modelling of both factors.*

- *Hypothesis* ii)*: Learning unsupervised disentanglement without any assumptions is fundamentally impossible. In accordance with the literature on causal learning [2], disentangling causal factors requires model assumptions and/or interventional data - instead of observational (raw) data.*

To address these hypotheses, we *explain*, *validate* and *evaluate* a method for unsupervised shape learning: *Unsupervised Part-wise Disentanglement of Shape and Appearance* developed by Lorenz *et al*. 2018.

To *explain*, after theoretical prerequisites (Chapter **??**) we give an overview over state-of-the-art unsupervised disentangling literature and situate the proposed method in relation to the literature (Chapter **??**). In particular, we carve out the necessary aspects of an approach for disentangling causal factors and analyze the current state of research in order to indicate future directions. We subsequently disclose our method (Chapter **??**).

To *validate*, we show that the proposed method outperforms the state-of-the-art for unsupervised learning of object shape on miscellaneous datasets, featuring human and animal faces and bodies (Chapter **??**). We also contribute several self-made video datasets for disentangling human pose from appearance, for articulated animal motion and for articulated composite objects. We highlight the specific challenges of these datasets and elucidate how the proposed method tackles them.

To *evaluate*, we perform ablation studies on critical components of the method. In addition, we compare to a part-wise shape learning method which does make the goal of disentangling explicit. To show that the disentanglement is indeed achieved, we evaluate the disentanglement performance against a shape-supervised state-of-the-art disentanglement method and perform favorably (Chapter **??**).

In short, our results are a big improvement upon the state-of-the-art in unsupervised object shape learning. This confirms the first hypothesis. To complement the learned shape in a generative process, object appearance is disentangled from shape. The achieved disentanglement with our causal assumptions, and the not-achieved disentanglement when dropping these assumptions, confirms the second hypothesis.

2 Bibliography

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *TPAMI*, 2013. 1

[2] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. In *WSDM*, 2018. 1, 5

[3] 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. 2

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2013. 2

[5] Sridhar Mahadevan. Imagination machines: A new challenge for artificial intelligence. In *AAAI*, 2018. 2

[6] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[7] Josh Tenenbaum. Building machines that learn and think like people. In *AAMAS*, 2018. 2, 3

[8] Carl G Jung. Collected works of cg jung: The archetypes and the collective unconscious (vol. ix), 1968. 3

[9] Noam Chomsky et al. *New horizons in the study of language and mind*. Cambridge University Press, 2000. 3

[10] Ernő Téglás, Edward Vul, Vittorio Girotto, Michel Gonzalez, Joshua B Tenenbaum, and Luca L Bonatti. Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, 2011. 3

[11] Matthew B Wall and Andrew T Smith. The representation of egomotion in the human brain. *Current Biology*, 2008. 3

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *ICCV*, 2015. 4

[13] 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. 4

[14] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *ECCV*, 1998. 4