Contents

1 Intro

Computer vision is the endeavour to algorithmically discern patterns in images, which reflect structures and processes in the physical world. For an understanding of the world, that is captured in an image, an algorithm needs to model the underlying causal elements, that contribute to generating the image. For example, when shown an image ..

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 managable 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 -> 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 -> 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.

How 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.

1.1 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

Part I

Appendix

A Lists

A.1 List of Figures

A.2 List of Tables

B Bibliography

Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John V Guttag. Synthesizing images of humans in unseen poses. *arXiv preprint arXiv:1804.07739*, 2018.

Irving Biederman. Recognition-by-components: A theory of human image understanding. *Psychol. Rev.*, 1987.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

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

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

Rodrigo de Bem, Arnab Ghosh, Thalaiyasingam Ajanthan, Ondrej Miksik, N Siddharth, and Philip H S Torr. Dgpose: Disentangled semi-supervised deep generative models for human body analysis. *arXiv preprint arXiv:1804.06364*, 2018.

Emily L Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *NIPS*, 2017.

Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. *CVPR*, 2018.

Pedro F Felzenszwalb, Ross B Girshick, David A McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010.

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.

Catalin Ionescu, Fuxin Li, and Cristian Sminchisescu. Latent structured models for human pose estimation. In *ICCV*, 2011.

Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Conditional image generation for learning the structure of visual objects. *NIPS*, 2018.

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

Michael Lam, Behrooz Mahasseni, and Sinisa Todorovic. Fine-grained recognition as hsnet search for informative image parts. In *CVPR*, 2017.

Karel Lenc and Andrea Vedaldi. Learning covariant feature detectors. In *ECCV Workshops*, 2016.

Zejian Li, Yongchuan Tang, and Yongxing He. Unsupervised disentangled representation learning with analogical relations. In *IJCAI*, 2018.

Jongin Lim, Youngjoon Yoo, Byeongho Heo, and Jin Young Choi. Pose transforming network: Learning to disentangle human posture in variational auto-encoded latent space. *Pattern Recognit. Lett.*, 2018.

Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017a.

Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. *CVPR*, 2017b.

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

Grégoire Mesnil, Antoine Bordes, Jason Weston, Gal Chechik, and Yoshua Bengio. Learning semantic representations of objects and their parts. *Mach Learn*, 2013.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *ECCV*, 2016.

Tu Dinh Nguyen, Truyen Tran, Dinh Q Phung, and Svetha Venkatesh. Learning parts-based representations with nonnegative restricted boltzmann machine. In *ACML*, 2013.

David Novotny, Diane Larlus, and Andrea Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *CVPR*, 2017.

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

Marco Pedersoli, Radu Timofte, Tinne Tuytelaars, and Luc J Van Gool. Using a deformation field model for localizing faces and facial points under weak supervision. In *CVPR*, 2014.

Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015.

Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *TPAMI*, 2017.

David A Ross and Richard S Zemel. Learning parts-based representations of data. *JMLR*, 2006.

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.

Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. *CVPR*, 2018.

Saurabh Singh, Abhinav Gupta, and Alexei A Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *NIPS*, 2017a.

James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017b.

Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

Yue Wu and Qiang Ji. Robust facial landmark detection under significant head poses and occlusion. *CVPR*, 2015.

Xianglei Xing, Ruiqi Gao, Tian Han, Song-Chun Zhu, and Ying Nian Wu. Deformable generator network: Unsupervised disentanglement of appearance and geometry. *arXiv preprint arXiv:1806.06298*, 2018.

Wei Yang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016.

Xiang Yu, Feng Zhou, and Manmohan Chandraker. Deep deformation network for object landmark localization. In *ECCV*, 2016.

Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *TPAMI*, 2016.

Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015.

Erklärung:

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den (Datum) . . . .