Contents

1 Prerequisites on Learning Disentanglement

1.1 Learning from Data

Learning from data is commonly understood as the ability of algorithms to improve their performance on a task with experience accumulated from observing data citegoodfellow16dlb. The source of data points x is usually understood as a dataset $X = \{x_i | i \in \{1 \dots n\}\}$ which is a collection of samples from a probability distribution $x_i \sim p(x)$.

1.1.1 Supervised

The term supervised learning denotes the task to learn a mapping from data points x_i to target labels y_i. A supervised algorithm has access to data-label pairs $(y_i, x_i) \sim p(y, x)$, in order to estimate the connection between data points and labels, either in form of a conditional probability $p(y|x)$, or in form of a deterministic function $y = f(x)$. The label y can be either discrete (*e.g.* information about an object class) or continuous (*e.g.* the location of an object part in an image). example for landmarks Recent advances, in particular the effectiveness of neural network models 1.1.3 to on big datasets, have led to progress in supervised learning: For many regression and classification tasks, such as object recognition, image classification or human pose estimation algorithms are performing on a superhuman level; many of these tasks are considered to be solved problemscite here. limitation: labelling data

1.1.2 Unsupervised

why?: usage of unlabeled data -> find structure in data space; transfer learning, multi-task learning Unsupervised learning is the endeavour to learn about structures and patterns in unlabelled data. The learning algorithm then has access to the data distribution $x \sim p(x)$. The task is usually framed as a form of density estimation, to model the entire distribution (cf. sec. 1.2). much harder also unspecified

model-free vs model-based rigid enough to be useful, flexible enough to useful recently data-driven -> flexible

limits of unsupervised learning? how much prior modelling should be employ? -> as much as possible as long as it is good? (link post Inference)

modeling data distribution P(y, x) sampling from distribution possible e.g. outlier detection P(X) has low probability

talk about data compression what does unsupervised even mean? no prior assumptions, no knowledge at all? unspecified.. read on this. notion of truly unsupervised learning actually harmful to progress, ill-defined -> intro

1.1.3 Artificial Neural Networks

Artificial neural networks (NN) are a powerful and flexible tool for function approximation. They are inspired by the connectionist modelling of biological neural networks. In a NN, a function $y = f(x)$ with vector input $x = \{x_i | i = 1 \dots n\}$ and vector output $y = \{y_j | j = 1 \dots m\}$ is modelled by:

$$\begin{aligned} h_j &= a(\sum_i w_{ji} x_i + b_i) \\ y_j &= a'(\sum_i w'_{ji} h_i + b'_i) \end{aligned} \tag{1.1}$$

, with weight matrices w, w', non-linear so-called activation functions a, a' (*e.g.* $a(x) = 0$ for $x < 0$, $a(x) = x$ for $x >= 0$) and bias vectors b, b'. The components h_j are called hidden units or neurons. Neural networks can also comprise multiple hidden layers a la $h_j = a(\sum_i w_{ji} h_i + b_i)$. It can be shown theoretically, that in the limit of infinite hidden units h_j, that NN can approximate any (continuous) function arbitrarily close **?** cite other. In practice, however, networks with more that one layer, referred to as deep neural networks, seem to work better.

For processing image data, one constrains the weight matrices to be only locally connected and to share weights across locations to enforce translation invariance, resulting in *convolutional* neural networks.

feature hierarchy **?** optimization via gradient descent has proven successful (for deep networks called backpropagation)

1.2 Generative Models

> What I cannot create, I do not understand. - R. Feynman

Learning and understanding structure in data by being able to generate the data, is the rationale behind generative modelling. Generative models are mostly applied for unsupervised learning and can be distinguished from discriminative models. While discriminative models are used to model posterior conditionals $p(y|x)$ (*e.g.* for supervised learning (cf. sec. 1.1.1), generative models capture the complete data distribution $p(x)$ in an estimate $\hat{p}(x)$citebishop06ml. Thus, after estimation one can generate samples from this model $\hat{p}$, hence the name generative model. The currently predominant formulations for learning generative models are built on either autoencoding or adversarial formulations:

1.2.1 Autoencoding Models

?? An autoencoding model is learning by reconstructing samples of data, $\hat{x} = f(x)$. To enforce data compression (otherwise the identity function is a trivial solution of autoencoding) the function has an information bottleneck, namely an inferred latent code z of reduced dimension. The autoencoder is then the chain of an encoding function $z = e(x)$ and a decoding function $\hat{x} = d(z) = d(e(x))$.

Whereas the conventional autoencoder consists of deterministic mappings e, d, the **variational autoencoder** models the probability distribution $p(x)$. More specifically, is maximizes a lower bound to the logarithmic likelihood $\log p(x)$ of data x. This so-called variational lower bound $\mathcal{L}$ is given by:

$$\mathcal{L} = \mathbb{E}_{z \sim q(z|x)} \log p(x|z) - \mathrm{KL}(q(z|x)||p(z)) \tag{1.2}$$

Where z introduces latent variables, with a prior distribution $p(z)$. The approximation to the posterior $q(z|x)$ of the latent variables and the posterior of the data given the latent variables $p(x|z)$. If one wants to model the distributions with neural networks, one typically uses Gaussian distributions and lets the networks predict the parameters (mean and variance) based on the image.

1.2.2 Adversarial Models

?? Generative Adversarial Networks (GAN) consist of two neural networks competing in a zero-sum game. A generator network G is generating images based on a latent code z sampled from a distribution $p(z)$. The discriminator network D is a binary classifier with the task to classify an image as originating from the data distribution p_{data} or from the distribution produced by G. The loss function of G is the negative of the loss of D, such that one can formulate the optimization in a minmax form:

$$\min_D \max_G -\frac{1}{2}\mathbb{E}_{x \sim p_{data}}[\log D(x)] - \frac{1}{2}\mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))] \tag{1.3}$$

The discriminator can also be interpreted as a learned similarity metric to measure the closeness of an image to the data distribution. **?**. The generator is then optimized to make the output indiscriminable from the data distribution.

There are many variants and extensions to this basic principle of learning with an adversarial principle. For example one can learn a discriminator on for a set of image patches cite CoGAN

1.3 Disentangling Representations

?? In supervised learning, a performance measure is naturally given by the metric that is optimized. In the unsupervised setting, judging the performance of a model is less straightforward. For example, when modelling an image domain, one could subjectively rate the quality of the generated image. But even for a qualitative assessment the question arises, how to rate the quality of the latent representation?

> Disentangle as many factors as possible, discarding as little information about the data as is practical. - Y. Bengio, A. Courville and P. Vincent citebengio

Bengio *et al*. define a representation to be useful, if it can be applied to many - in advance unknown - different tasks, while being trained on one particular task. As the downstream tasks can be multifarious, the essential data *information* should be contained. For some tasks only a subset of aspects of the data will be necessary, that is why *disentangled factors* make a representation particularly practical - so goes their reasoning.
The latent representation z learned by generative models captures the essential *information* of the data distribution. That is made sure by requiring to be able to generate the entire distribution from it. Now to the second goal, the *disentangling* of independent generative factors:

1.3.1 Equivariance and Invariance

The definition of factor by change static ... factors should represent elements of real world - change in element -> corresponding change in representational factor - leave other factors representing other elements invariant

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 estimates $\hat{\mathbf{z}}_\mathbf{i}$ 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.

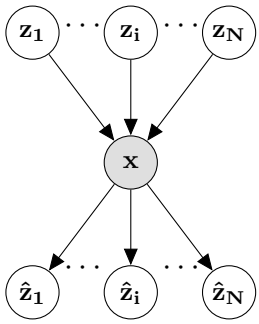

Figure 1.1: Disentangling causal factors means to infer an estimate - *i.e*. a representation - from an image

1.3.2 Shape and Appearance

especially difficult because of domain

1.4 Theoretical Impediments from Causality

?? factors are causal As outlined earlier, the type of knowledge that can be gained by learning from "raw" data is limited. With raw data we mean data x sampled from a $p(x)$. so far fitting curve p(x) to data manifold what is missing to human-level intelligence? (cite lake 2016)

causal learning is a hard problem: instead of only learning statistical measures from data, model also needs to be learned (cite Schoelkopf)

Hypothesis: disentangling factors = estimating causal factors -> needs causal for estimation of causal factors "raw data" insufficient -> need interventional data or model assumptions. we do both: 1. intervene with changes to an image which are assumed to change only one factor. 2. model the causal process of the image generation in the theme of analysis-by-synthesis

what does the causality literature have to say? $\rightarrow$ 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

lacking the tools to accurately estimate causality, researchers shied away from making causal statement. Developing machines with human-like abilities requires discovery and reasoning in terms of causal models. Recently (in the past 30 years), overshadowed by the prominent success of data-driven deep learning, the field of causality has emerged to mathematical rigor.

- ladder of causation: association, intervention, counterfactual - current machine learning mostly on level of association (correlations estimated from "pure" data) -> purely data-driven approach can only go so far humans seem to have innate assumptions on coherence, causality, physics etc. introducing inductive biases

measure: p(x) assume causal model: p(x | a, s) want: p(s) and p(a)

encoding $p(s) = p(s|x)\ p(a) = p(a|x) = p(a|s, x)$

decoding $p(x) = p(x|a, s)p(a)p(s)$

p(x| do(s), do(a))

example: Gaussian only with access to p(x) hopes to find factors p(a, b) = p(a) p(b) (InfoGAN, BetaVAE) what if not full-filled? two-dimensional Gaussian: axis x and y are independent factors. in general any superposition of x and y which is orthogonal, can be found imagine a perfect dimensionality reduction yielding a two-dimensional latent space one can find the axes that correlate most with human understanding of independent factors i.e. pose and appearance. But how can a machine find these axes automatically from raw data? it cant, neither can anyone (including humans) (Pearl). Humans know these factors are independent from observing that they can change independently e.g. from observing someone undressing or changing his pose (i.e. harnessing temporal information, with the assumption of temporal coherence) or by changing the factors themselves e.g. what happens to the image of me if I change my pullover? It can be proven mathematically (Pearl) that interventional data or at least certain (which) causal assumptions about the world are necessary to estimate certain quantities.

1.4.1 Transformation as interventions

we harness intervention p(x| do(a), b)

2 Bibliography