ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

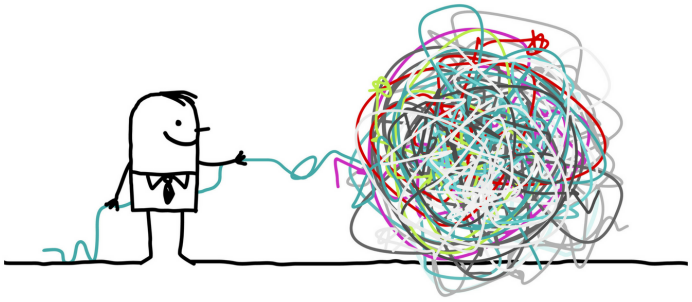
Distributed Computing

Prof. R. Wattenhofer

Advanced Topics in Deep Learning - Disentangled Representations

Deep neural networks have been very successful at automatic extraction of meaningful features from data. Manual feature engineering is often not necessary anymore, and we can instead focus on designing the architecture of the neural networks. However, due to the complexity of neural networks, the extracted features are themselves highly complex and can often not be interpreted by humans. Instead, the neural network is treated as a black box and we have to trust external evaluation metrics such as train and test error.

Often it would be beneficial to understand what kinds of hidden representations, or features, the model has actually learned. Several methods exist that are particularly suited for learning meaningful hidden representations. The model we will be mainly looking at in this thesis is the Variational Autoencoder (VAE), a deep generative model. VAEs have been shown to be able to disentangle simple data generating factors from a highly complex input space. For example, when being fed with images of faces, a VAE might automatically learn to encode the direction of the lighting in a single hidden variable. After training, we can vary each hidden variable and observe the effect on the output, which let's us analyze, albeit manually, what kinds of features the model has learned. Several VAE extensions such as β-VAE focus mainly on the disentanglement aspect and also introduce disentanglement metrics, making evaluation easier.



So far, VAEs (and other generative models) have been most succesful in the image domain, where most architectures are based on Convolutional Neural Networks (CNNs), which by themselves are known to be powerful feature extractors. In sequential and/or discrete domains, such as music and language processing, there has been limited success. In this thesis we want to investigate the causes for this. In particular, we want to compare different neural network architectures and their disentanglement capabilities when using them as building blocks for VAEs. Depending on the results we will go more towards improving existing methods for learning disentangled features and having a close look at current disentanglement metrics, or try to learn disentangled representations for sequence tasks.

If this sounds interesting to you, do not hesitate to contact us.

Requirements: Knowledge in Deep Learning, or solid background in Machine Learning. Implementation experience is an advantage. You should be able to read and understand the first 12 chapters of the "Deep Learning Book" by Goodfellow et al. (available for free online from MIT press). If you are interested in the topic but new to deep learning we expect

you to complete an introductory deep learning course before applying for the thesis, such as Andrew Ng's coursera course (use the free trial!)[1] or this Udacity course[2].

Interesting Papers:

- Deep Convolutional Inverse Graphics Network (one of the first papers on learning deep disentangled representations) `http://papers.nips.cc/paper/5851-deep-convolutional-inverse-graphics-network.pdf`
- Autoencoding Variational Bayes (original VAE) `https://arxiv.org/pdf/1312.6114.pdf`
- β-VAE (Disentanglement metric 1) `https://openreview.net/pdf?id=Sy2fzU9gl`
- Understanding disentangling in β-VAE `https://arxiv.org/pdf/1804.03599.pdf`
- Disentangling by factorising (Factor-VAE, disentanglement metric 2) `https://arxiv.org/pdf/1802.05983.pdf`
- Isolating Sources of Disentanglement in Variational Autoencoders (Disentanglement metric 3) `https://arxiv.org/pdf/1802.04942.pdf`
- Independently Controllable Factors `https://arxiv.org/pdf/1708.01289.pdf`
- InfoGAN `http://papers.nips.cc/paper/6399-infogan-interpretable-representation-learning-by-information-maximizing-generative-adversarial-nets.pdf`

Interested? Please contact us for more details!

Contacts

- Gino Brunner: `brunnegi@ethz.ch`, ETZ G63
- Oliver Richter: `richtero@ethz.ch`, ETZ G63

[1] `https://www.coursera.org/specializations/deep-learning`
[2] `https://classroom.udacity.com/courses/ud730`