Disentangling via dynamics

Benjamin Peters*
Zuckerman Institute
Columbia University
New York, NY 10027
benjamin.peters@columbia.edu

Abstract

1 Introduction

1.1 Disentangling via a hierarchy of constrained dynamics

The aim of this project is to learn disentangled object representations in an unsupervised way from dynamic input by predicting the next input frame in a sequence of images. Hierarchical predictive networks are inspired by the hierarchical architecture of the brain and by the idea of analysis by synthesis of unsupervised learning (e.g., Helmholtz).

Recently, several attempts have been made to build deep, end-to-end trainable hierarchical predictive models. E.g., Lotter16 developed the PredNet architecture [...]. One observation in these models is that they do not necessarily make use of all of the hierarchal levels for prediction (). One reason for this might be that the latent space of the lowest level $z_t^{(1)}$ of the hierarchy together with a complex non-linear transition function f that maps $z_t^{(1)}$ to $z_{t+1}^{(1)}$ is already expressive enough for predicting in image space. As a consequence of a complex transition function, the latent space of $z^{(1)}$ can be highly entangled. The aim of this project is to learn meaningful, disentangled, hierarchical object represenations.

The implementational details are vague:

Consider a dynamic scene in which an object moves across the visual field. A VAE encodes frames x_t ($t \in [1 \dots T]$) via $q(z_t|x_t)$ and tries to predict x_t from its current latent state representation $p(x_t|z_t)$. In order to learn dynamics of this object/scene in a meaningful latent space, prediction across time occurs in the latent space, such that $\hat{q}(z_{t+1}|z_t) = \mathcal{N}(f(z_t), \Sigma(z_t))$ and we want to have a 'good' prediction in latent space (e.g., reducing KL-div between $\hat{q}(z_{t+1}|z_t)$ and $q(z_{t+1}|x_{t+1})$, or simply euclidean loss on predicting the first moment/mean of q).

If z is a disentangled latent representation it would contain variables that encode the object's appearance as well as dynamic variables (position, speed). In this case, the function f would become an identity mapping for the appearance variables and a simple linear transformation for the set of dynamics variables. The idea here is: If we apply an appropriate capacity constraints on f (e.g., an L1 norm on $|f(z) - I|$ and $|\Sigma(z)|$, or low entropy on $\hat{q}(z_{t+1}|z_t)$) this will facilitate learning of disentangled representations in z. Moreover, this would provide a straight forward interpretation of disentanglement in the context of dynamics: representations are disentangled in latent space if their transitional dynamics are simple.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Preprint. Work in progress.

Not all aspects of the data might be well represented by constrained dynamics in latent space at a single hierarchical level. The general idea therefore is to learn hierarchical representations that integrate and predict over different time-scales (and possibly also receive differently complex transformations of the input, e.g., Ladder VAEs).

1.2 Related work

Hsieh18 predict video sequences (e.g., moving MNIST) by decomposing the latent representation into a time-invariant content vector and a time-dependent pose vector. Kosiorek18 similarly decompose the latent representation into a 'what' and 'where' representations. The proposed approach here would instead of having no dynamics on some variables vs complex non-linear dynamics on other variables, place a general constraint on the dynamics.

1.3 challenges

- z wants to encode/predict every aspect of x. But not all details of x might be relevant for a 'meaningful' representation, some aspects of the transition of x_t -> x_{t+1} might be unpredictable or irrelevant for a 'meaningful' representation in latent space but may still drive the reconstruction loss to a large extent .
  - constrain capacity on z (-> beta-VAE)?
  - Variational lossy AE Chen2016: `http://arxiv.org/abs/1611.02731`
- the reconstruction loss might become a problem if we have stochastic dynamics (e.g., motion of the object has a stochastic component).
  - Henaff2017: Prediction under uncertainty with error encoding[2]
  - Denton2018: Stochastic video generation with a learned prior[3]
- dark-room problem: optimizing prediction loss in latent space over both parameters that determine prediction in latent space as well as the encoding into latent space might result in degenerate solutions. (-> predictive contrastive loss,)

[2] `https://arxiv.org/pdf/1711.04994.pdf`
[3] `https://arxiv.org/pdf/1802.07687.pdf`