Disentangled Representation with Autoencoders for efficient DRL

M. Falzari

March 2, 2022

Quick recap

Dimensionality Reduction

reduce the number of dimensions/features while maintaing the most important information

Quick recap

Dimensionality Reduction

reduce the number of dimensions/features while maintaing the most important information

Sparse Autoencoders

- main goal is to learn $f(g(x)) = x$
- g is the encoder and f is the decoder
- $g(x) = z$ where $z \in R^n$ and $x \in R^m$ and $n \ll m$

Quick recap

Dimensionality Reduction

reduce the number of dimensions/features while maintaing the most important information

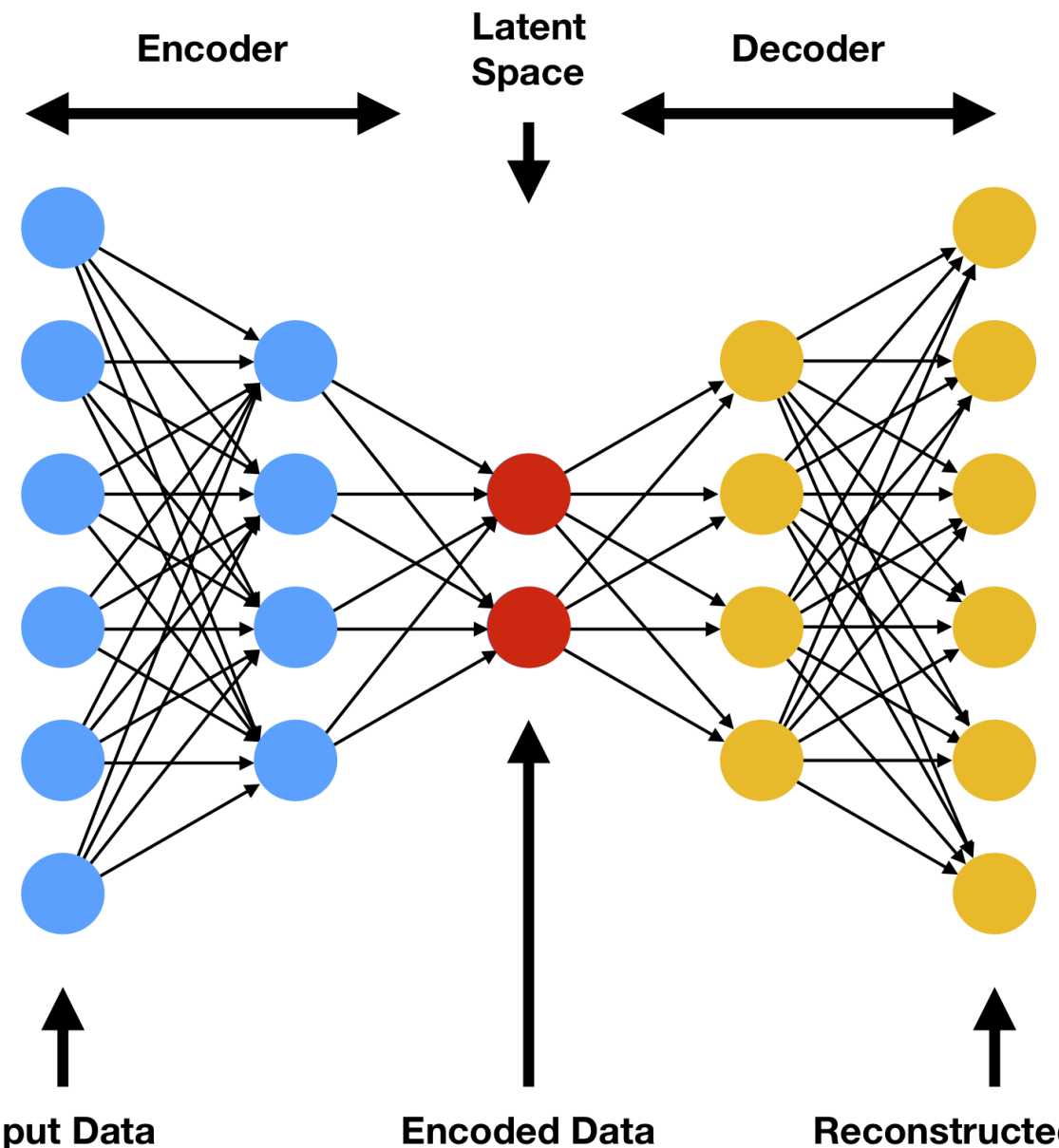
Sparse Autoencoders

- main goal is to learn $f(g(x)) = x$
- g is the encoder and f is the decoder
- $g(x) = z$ where $z \in R^n$ and $x \in R^m$ and $n \ll m$

Other techniques? and why Autoencoders?

visit `https://172.104.159.41/thesis/summary.html` section Line of Thoughts

Sparse Autoencoders



Representation Learning

Overview

- Representation Learning is highly related to Dimensionality reduction.
- Important to notice a good dimensionality reduction does not necessarily learn a good representation and the opposite is also true
- Autoencoders are a "neutral" technique that gives a lot of flexibility during training and enables to have more control on different tradeoffs that are intrinsic of the problem (i.e. dimensionality reduction vs representation learning)

Why create low dimensional latent space for DRL?

DARLA (DisentAngled Representation Learning Agent)

- It shows that all DRL techniques (implicitly) maps the high dimensional state-space to a low dimensional state-space and then maps this low dimensional state-space to the action space.
- Therefore, we want to remove this concern from the DRL. We also want to do this because we do not want the representation to be biased by the DRL objective (which, in a nutshell, maximizes the rewards)

MDPs Generalization

What is it?

- it is the concept that exists only one natural world and we can sample MDPs from it.
- Each MDP will have the same action space. (If this does not apply, it close to impossible to have transfer)
- Different State spaces but some structural similarity (i.e. isomorphisms)
- so if we want to generalize we need to be able to have the same representation for different MDPs
- In order word, generalization in this context means be able to find the common state space between MDPs (this must exists since we assumed that we are sampling these MDPs from one single "natural world")

Disentangled Representation?

Definition

There is not yet a definition on which the community agrees upon. The best and the most formal attempt was done in Towards a definition of disentangled representations. (exploiting concept of physics and group theory)

Disentangled Representation?

Definition

There is not yet a definition on which the community agrees upon. The best and the most formal attempt was done in Towards a definition of disentangled representations. (exploiting concept of physics and group theory)

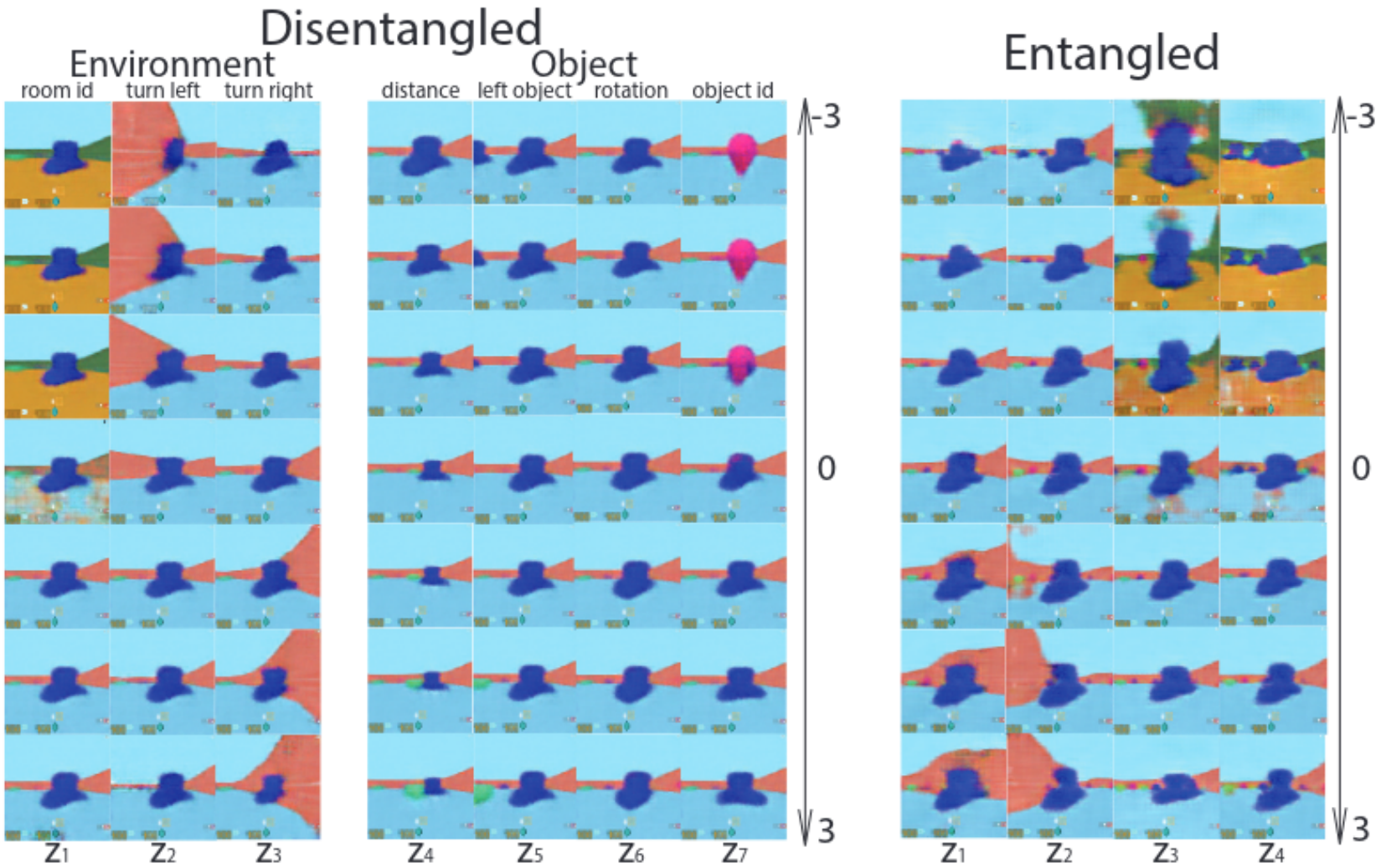
In nutshell (quoting directly)

Intuitevely, we define a vector representation as disentangled, if it can be decomposed into a number of subspaces,each one of which is compatible with, and can be transformed independently by a unique symmetry transformation

Still not clear?

Let's say we want to present an environment that has only a solid and this solid has a colour and a position. Ideally, we want to have a 3D vector where one dimension represents the shape one the position and one the colour. This, in a superficial point of view, is a disentangled representation.

Examples of Entangled vs Disentangled representations



But why Disentangled representations?

Also here the literature is not clear. There are a lot of papers which shows that having such a representation has the following benefits on DRL

- increase the sample-efficiency
- decrease the sensitivity to nuisance variables (i.e. variables that are not too important for the decision process)
- Better performance in terms of generalization

But why Disentangled representations?

Also here the literature is not clear. There are a lot of papers which shows that having such a representation has the following benefits on DRL

- increase the sample-efficiency
- decrease the sensitivity to nuisance variables (i.e. variables that are not too important for the decision process)
- Better performance in terms of generalization

Formally though, there is no theory of why is the case, a good starting point is the paper Are Disentangled Representations Helpful for Abstract Visual Reasoning? Here they show experimentally (once again) that having such a representation results in the aforementioned properties

Interesting point (2022 survey)

What we will do in the thesis?

We want to see whether these experimental gains also translates to harder and more complex environment

What we will do in the thesis?

We want to see whether these experimental gains also translates to harder and more complex environment

if that will be the case we also want to address whether we can generalize in a zero-shot transfer situation

What we will do in the thesis?

We want to see whether these experimental gains also translates to harder and more complex environment

if that will be the case we also want to address whether we can generalize in a zero-shot transfer situation

The architectures we want to test are:

- Sparse Autoencoders (already implemented)
- Variational Autoencoders (already implemented)
- β Variational Autoencoders
- Mutual information Variational Autoencoders
- Adversarial Variational Bayes (already implemented)

Ideas for future research

- test other Disentanglement AE/GAN architectures (e.g. FactorVAE,CasualVAE,DreamingVAE).
- explicitly focus on transfer (maybe with fine-tuning instead of zero-shot)
- test different DRL algorithm to see how this impact the performance
- test different method of training the AE (for example on-line, see active perceptions frameworks and/or active learning currently in development by Microsoft research closely followed by Yoshua Bengio)
- In general, the idea of representation learning + DRL seems to be a really interesting and not fully explored path. (see The Consciusness Prior by Yoshua Bengio)

Reference

For more reference and in-depth explanation of the research process see `https://172.104.159.41/thesis/summary.html` which is constantly update with every single step we are taking and the motivation/explanation of why we are taking such steps