Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data

Wei-Ning Hsu, Yu Zhang, and James Glass

Talk by Stefan Wezel

Explainable Machine Learning

January 6, 2021

Overview

- What are disentangled representations (intuition)
- Why disentangled representations
- Formal description of disentangled representations
- SequentialVAE
- Did they achieve disentanglement?
- Other approaches and challenges

What is disentanglement?

Intuition

- encode distinct generating factors in separate subsets of latent space dimensions
- i.e. color as one subspace, translation, as another
- The exact definition is often discussed, we will have a look at a proposed one

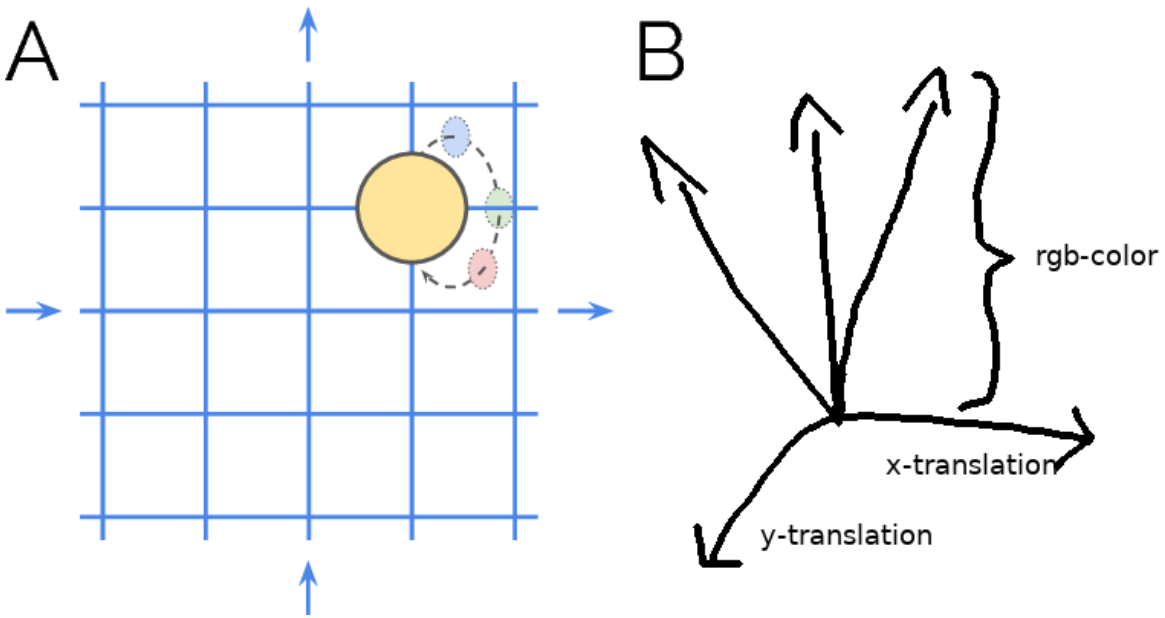

Figure: adslkfjds...Source:Higgins, Irina, et al.

Why learn disentangled representations?

Motivation

- Gives us an exact idea, of what variables were used, to come to a result
  - Fairness in ML (exact)
  - Explainability/Interpretability
  - Overall, a model just becomes more usable if latent variables carry semantic meaning

What are disentangled representations formally?

A field-trip to group theory

- Signal can get shifted or warped
- the set of these transformations make up a symmetry group
- signal's meaning is preserved
- the resulting set of transformed signals are the actions of the symmetry group on the world state

What are disentangled representations formally?

A field-trip to group theory

- This symmetry group can be decomposed into symmetry subgroups
- One affects location
- the other affects frequence

What are disentangled representations formally?

Disentangled Group Action

- Group action $G \times X \mapsto X$
- Group decomposes into direct product $G = G_{shifts} \times G_{warps}$
- Is disentangled with respect to decomposition of G
  - if there is decomposition $X = X_{shifted} \times X_{warped}$
  - and actions $G_{shifts} \times X_{shifted} \mapsto X_{shifted}$
  - and actions $G_{warps} \times X_{warped} \mapsto X_{warped}$

What are disentangled representations formally?

- Let W be the set of world states (all shifts and warps of signal)
- Generative process $b : W \mapsto O$ (voice to audio processing unit)
- Inference process $h : O \mapsto Z$ (observation to latent space)
- $f : W \mapsto Z, f = h \circ b$
- Now, we know, there is a symmetry group acting on W ($G \times W \mapsto W$)
- We want to find corresponding $G \times Z \mapsto Z$ to reflect symmetry structure of W in Z
- More formal: $g \cdot f(w) = f(g \cdot w)$
- This is whats called an equivariant map (famous example: convnet)

What are disentangled representations formally?

Disentangled Representation

- Assume Symmetries of W decompose into direct product $G = G_1 \times ... \times G_n$
- Representation is disentangled if
  - action $\cdot : G \times Z \mapsto Z$
  - equivariant map $f : W \mapsto Z$ between actions on W and Z
  - Decomposition $Z = Z_{shifted} \times Z_{warped}$
  - where $Z_{shifted}$ is only affected by shifts in W (G_{shifts})
  - and Z_{warped} is only affected by warps in W (G_{warps})
- There may be more criteria (preserving group structure, isomorphisms, ...) but for the intuition this is sufficient

Did they achieve disentanglement

...

- With respect to a decomposition into two
- Setting: 10 sentences, 630 speakers
- How can we formulate this in group theory?

How did they do it?

Intuition

- With respect to a decomposition into two
- regularize z2 by sequence dependant prior (lookup table of s-vectors)
- and z1 by sequence independant prior

How did they do it?

Methods

- Sample batch at segment level (instead of sequence level)
- Maximize segment variational lower bound
- (Force z2 to be close to mu2)
- approximation of mu2 is closed form equation (concave function, set derivative to 0)

Challenges

...

- If we really think about it, it is hard for us to define what a disentangled representation should actually be
- Precise biases of what the latent space should be decomposited into can be helpful as well as biases towards the 'form' of these latent subspaces