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 8, 2021

Overview

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

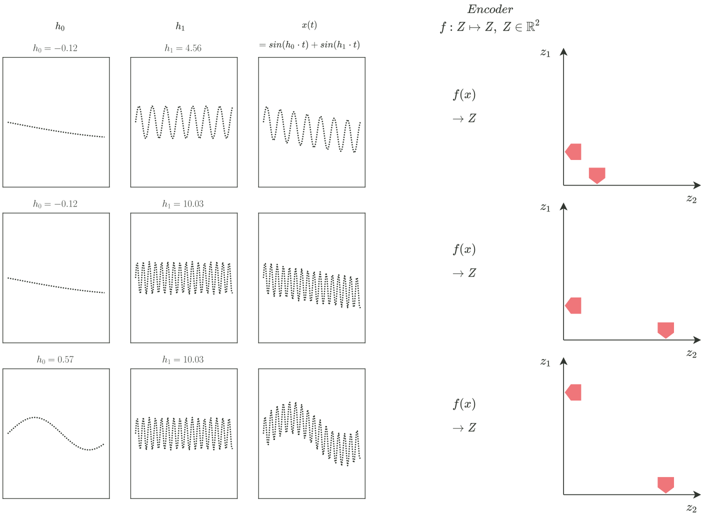
Overview

- Using Sequential VAE (-> Unsupervised representation learning)
- Represent information from different temporal scales in corresponding latent subspaces
- Claim that they achieve disentanglement with respect to sequence (speaker) and segment (content) information
- would mean that those latent variables then can be used separately
  - speaker verification
  - denoising
  - ...

What is disentanglement?

Intuition

- encode distinct generating factors in separate subsets of latent space dimensions



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

Disentangled Representations Formally

A field-trip to group theory: important concepts

- Group
  - tuple of operation and set
  - set is closed under operation, there is identity element, and inverse for every element, associativity
- Symmetry group
  - Group action, that leaves object (defined through set/sets) invariant
- Group action
  - Actions are results of symmetry transformations of set (i.e. set of changed order)
- Direct product
  - $G = G_1 \times ... \times G_n$
  - Group conditions must hold for group and each subgroup

- Disentangled group actions
  - Result of transformations that only change certain aspect of world, but leave others invariant
- Assuming G can be decomposed into direct product symmetry subgroups G_i
- We want mapping $f : W \mapsto Z$
- Symmetry G on W should be preserved in Z, $G \times Z \mapsto Z$
  - $g \cdot f(w) = f(g \cdot w)$ -> equivariant map

Disentangled Representations Formally

A field-trip to group theory: What is disentanglement in terms of group theory?

- Representation is disentangled if
  - equivariant map $f : W \mapsto Z, g \cdot f(w) = f(g \cdot w) \forall g \in G, w \in W$
  - such a map would split Z into independent subspaces, thus satisfying:
    - Decomposition $Z = Z_1 \times ... \times Z_n$
    - where Z_i is only affected by transformations G_i in W
    - and Z_i invariant to all $G_{j \neq i}$ in W
    - Thus each subspace Z_i can be transformed ONLY by the corresponding symmetry of W

- signal $x(t) = sin(h_0 \cdot t) + sin(h_1 \cdot t)$ with $h_0 \sim \mathcal{N}(0,1),\ h_1 \sim \mathcal{N}(5,1)$;
- The set of possible values for h_0, h_1 make up our W
- The group of symmetries acting on this W decompose into $G = G_{h_0} \times G_{h_1}$
- We want to find an equivariant map $f : W \mapsto Z$ with $Z \in \mathbb{Z}^2$
- so that changes of h_0 result ONLY in changes in z_0 and changes of h_1 ONLY in z_2
- Note, that this requires prior knowledge of generating factors in our world

Back to the paper

Did they achieve disentanglement?

- Disentangled with respect to what decomposition?
- Assuming there is a decomposition $G = G_{sequence} \times G_{segment}$
- This should be reflected in $Z = (z_1, z_2)$
- They propose to store sequence information in z_2 and segment information in z_1
- Thus, they need to find an equivariant map $f : W \mapsto Z$, so that z_1 is only affected by actions on $G_{sequence}$ and vice versa

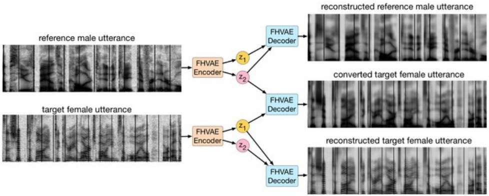
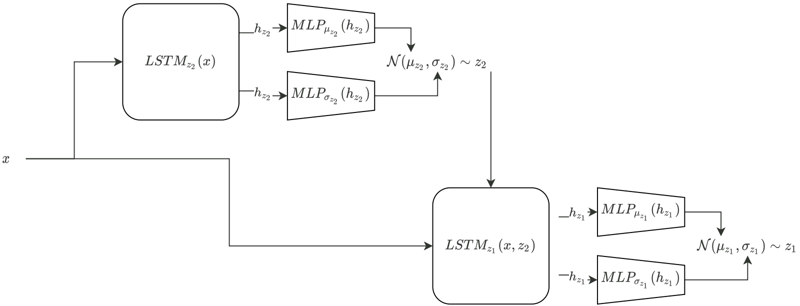
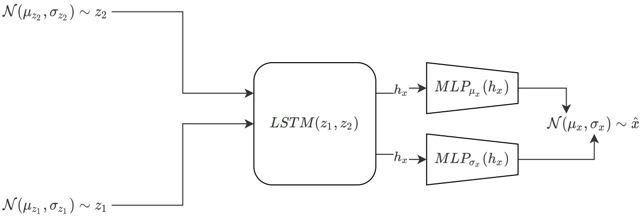


Image source: Hsu et al., 2017



x
$LSTM_{z_2}(x)$
h_{z_2}
h_{z_2}
$MLP_{\mu_{z_2}}(h_{z_2})$
$MLP_{\sigma_{z_2}}(h_{z_2})$
$\mathcal{N}(\mu_{z_2}, \sigma_{z_2}) \sim z_2$
$LSTM_{z_1}(x, z_2)$
h_{z_1}
h_{z_1}
$MLP_{\mu_{z_1}}(h_{z_1})$
$MLP_{\sigma_{z_1}}(h_{z_1})$
$\mathcal{N}(\mu_{z_1}, \sigma_{z_1}) \sim z_1$



$\mathcal{N}(\mu_{z_2}, \sigma_{z_2}) \sim z_2$
$\mathcal{N}(\mu_{z_1}, \sigma_{z_1}) \sim z_1$
$LSTM(z_1, z_2)$
h_x
h_x
$MLP_{\mu_x}(h_x)$
$MLP_{\sigma_x}(h_x)$
$\mathcal{N}(\mu_x, \sigma_x) \sim \hat{x}$

$$\mathcal{L}(\theta, \phi, X) = \sum_{n=1}^{N} \mathcal{L}(\theta, \phi; x^{(n)} | \tilde{\mu_2}) + log\ p_\theta + const.$$

$$\mathcal{L}(\theta, \phi; x^{(n)} | \tilde{\mu_2}) = \mathbb{E}_{q_\phi(z_1^{(n)}, z_2^{(n)} | x^{(n)})} \left[log\ p_\theta(x^{(n)} | z_1^{(n)}, z_2^{(n)}) \right]$$

$$-\mathbb{E}_{q_\phi(z_2^{(n)} | x^{(n)})} \left[D_{KL}(q_\phi(z_1^{(n)} | x^{(n)}, z_2^{(n)}) || \underbrace{p_\theta(z_1^{(n)})}_{\text{sequ. ind.}}) \right]$$

$$-D_{KL}(q_\phi(z_2^{(n)} | x^{(n)}) || \underbrace{p_\theta(z_2^{(n)} | \tilde{\mu_2})}_{\text{seq. dep. prior}})$$

FHVAE

Regularization

- What is the sequence dependent prior μ_2?
  - imagine a word vector (->s-vector)
  - $g(y) = \mu_2$ can be viewed as a differentiable lookup table (embedding in tf, pytorch)
- regularize z2 by sequence dependent prior
- and z1 by sequence independent prior (i.e. standard Gaussian)
- μ_2 can be found in closed form

FHVAE

Self-supervised discriminative objective

- Train z_2 to predict sequence index (of sequence in lookup table)

Features	Dimension	α	Raw	LDA (12 dim)	LDA (24 dim)
i-vector	48	-	10.12%	6.25%	5.95%
	100	-	9.52%	6.10%	5.50%
	200	-	9.82%	6.54%	6.10%
$\boldsymbol{\mu}_2$	16	0	5.06%	4.02%	-
	16	10^{-1}	4.91%	4.61%	-
	16	10^{0}	3.87%	3.86%	-
	16	10^{1}	**2.38%**	**2.08%**	-
	32	10^{1}	**2.38%**	**2.08%**	**1.34%**
$\boldsymbol{\mu}_1$	16	10^{0}	22.77%	15.62%	-
	16	10^{1}	27.68%	22.17%	-
	32	10^{1}	22.47%	16.82%	17.26%

Figure: Comparison of speaker verification equal error rate (EER) on TIMIT dataset (lower is better)

Disentangled Representations Formally

A field-trip to group theory: Disentangle our example formally

- Signal can get shifted or warped
- the set of these transformations make up a symmetry group
- This can be decomposed into shifts and warps/subsets of original set (all shifted $\times$ all warped)
- Either content is preserved, or speaker is preserved
- the resulting set of transformed signals are the actions of the symmetry group on the world state

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

- 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 $G_{warps} \times X_{warped} \mapsto X_{warped}$

- 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 symmetry transformations G of W decompose into direct product $G = G_1 \times ... \times G_n$
- Representation is disentangled if
  - equivariant map $f : W \mapsto Z, g \cdot f(w) = f(g \cdot w) \forall g \in G, w \in W$
  - such a map would split Z into independent subspaces, thus satisfying:
    - 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})
    - Thus each subspace can be transformed by the corresponding symmetry (like shift or warp independently)
- 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 terms?

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