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 10, 2021

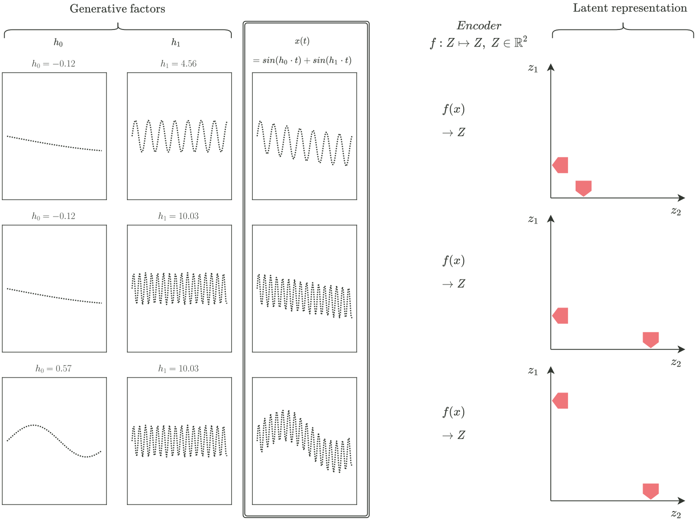
Overview

- Introduction
- What are disentangled representations (intuition)
- Why disentangled representations
- Formal description of disentangled representations
- Disentanglement in the context of the paper
- Factorized Hierarchical VAE

Overview

- Propose Sequential Factorized Hierarchical VAE (FHVAE)
- Learn factorized latent space
- Focus on speech data
  - Sequence level (Speaker, ...) representation
  - Segment level (content, noise, ...) representation
- Exploit different temporal scales of speech sequence data

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



Why learn disentangled representations?

Motivation

- Explainability/Interpretability
- Fairness
- Scientific modeling
- Speaker verification
- Denoising

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
  - Set of transformations that leave object W (i.e. another set) invariant
  - Operation is composition of transformations
- Group action
  - Actions are results of symmetry transformations on object (i.e. set of changed order (permuation))
- Direct product
  - $G = G_1 \times ... \times G_n$

- Disentangled group actions
  - Result of certain subset G_i of transformations that only change certain subset W_i of object, but leave others invariant
- $\rightarrow$ If we observe disentangled group actions in the world, we want to model those
- Then we can assume G can be decomposed into direct product symmetry subgroups G_i
- To model those, we want to find mapping $f : W \mapsto Z$
- Symmetry G on W should be preserved in Z
  - $g \cdot f(w) = f(g \cdot w) \rightarrow$ 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 $f : W \mapsto Z$
- 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 G_i on W (or on Z)

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 $Z = (z_1, z_2)$
- z_1 segment, z_2 sequence
- Find an equivariant map $f : W \mapsto Z$, so that z_2 is only affected by actions of $G_{sequence}$ and vice versa

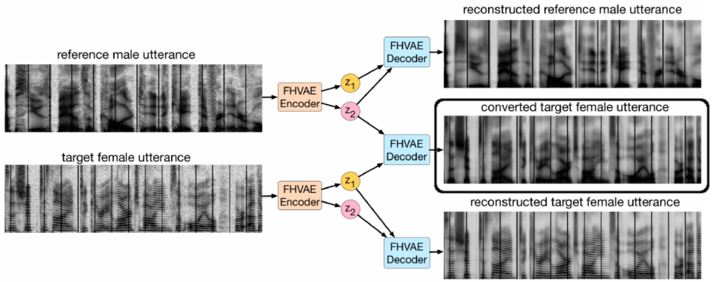
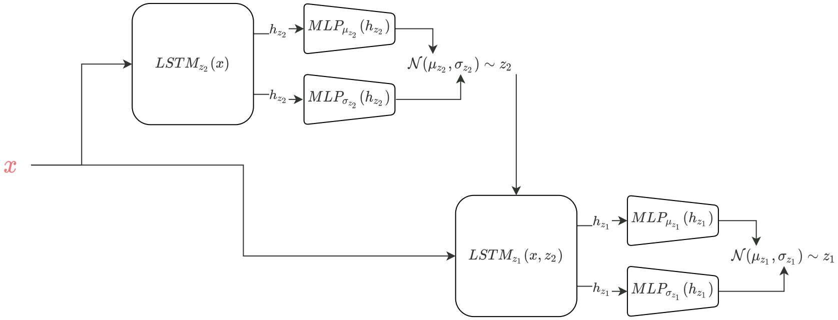


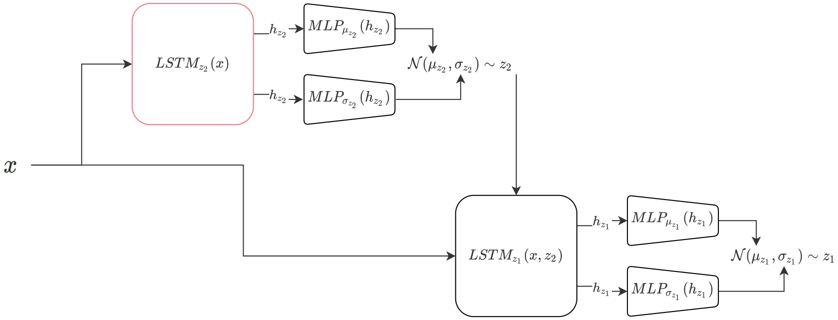
Image source: Hsu et al., 2017

Encoder$_\theta$



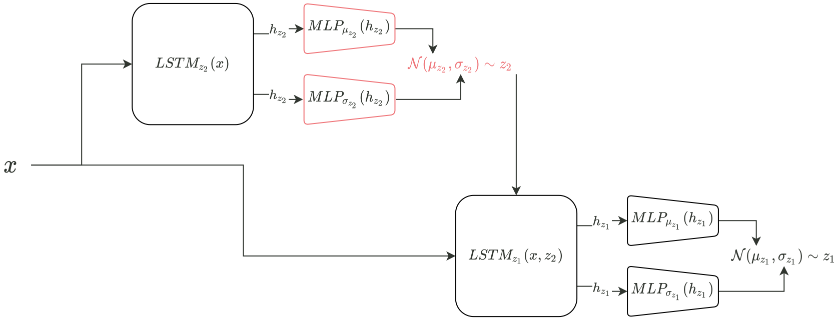
FHVAE

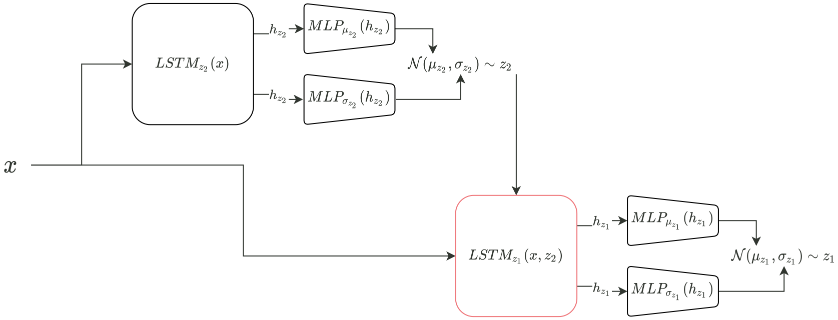
Encoder$_\theta$



FHVAE

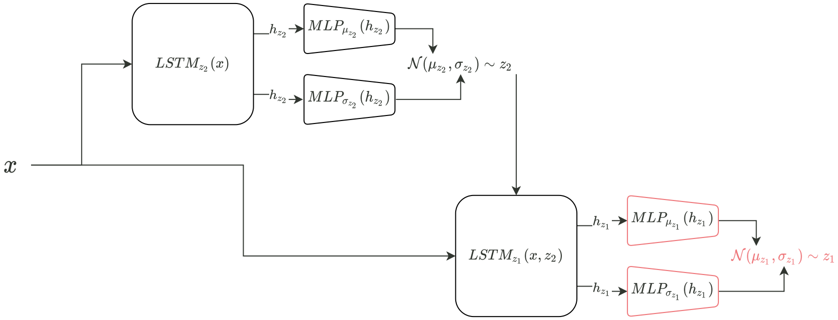
Encoder$_\theta$

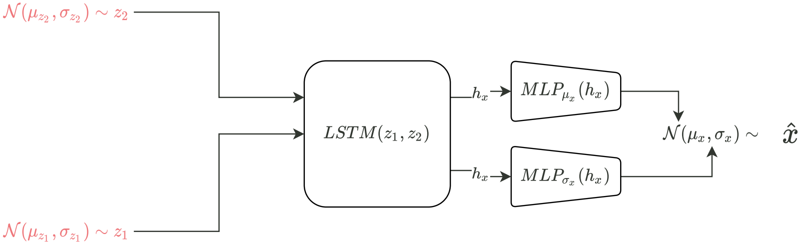



$LSTM_{z_2}(x)$
h_{z_2}
h_{z_2}
$MLP_{\mu_{z_2}}(h_{z_2})$
$MLP_{\sigma_{z_2}}(h_{z_2})$
$\mathcal{N}(\mu_{z_2}, \sigma_{z_2}) \sim z_2$
x
$LSTM_{z_1}(x, z_2)$
h_{z_1}
h_{z_1}
$MLP_{\mu_{z_1}}(h_{z_1})$
$MLP_{\sigma_{z_1}}(h_{z_1})$
$\mathcal{N}(\mu_{z_1}, \sigma_{z_1}) \sim z_1$

FHVAE

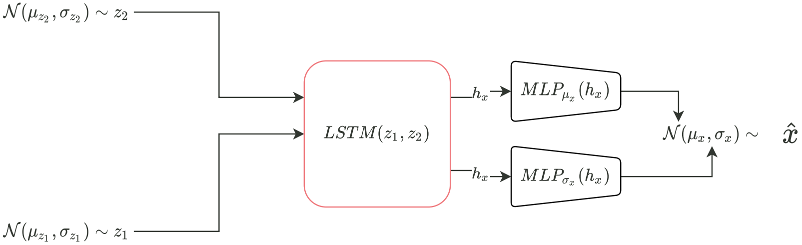
Encoder$_\theta$




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

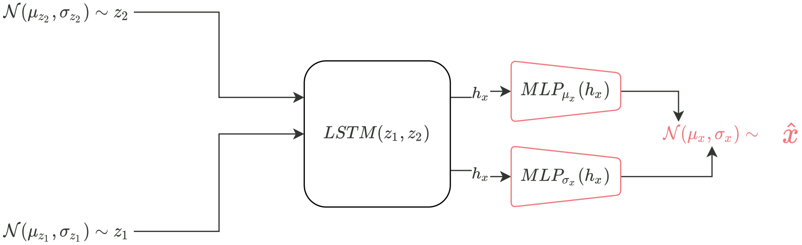
FHVAE

Decoder$_\phi$



FHVAE

Decoder$_\phi$



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

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

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

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

and $log\ p_\theta(\tilde{\mu_2}) = log\ p(z_2^{(i,n)} | \tilde{\mu_2}^{(i)}) - log(\sum_{j=1}^{M} p(z_2^{(i,n)} | \mu_2^{(j)}))$

FHVAE

S-vector μ_2

- What is the sequence dependent prior μ_2 (s-vector)?
  - imagine a word vector
  - s-vector for every sequence
  - Ideally, similarities in sequences should be reflected in s-vectors close in euclidian space
  - $g(sequence\ id) = \mu_2$ can be viewed as a differentiable lookup table (embedding in tf, pytorch)
  - For test (where there is no seq.id., it can be found in closed form solution)

Objective - Sequence variational lower bound

$$\mathcal{L}(\theta, \phi, X) = \underbrace{log\ p(x|z_1, z_2)}_{\text{reconstruction}}$$
$$- \underbrace{D_{KL}(\mathcal{N}(\mu_{z_1}, \sigma_{z_1})||\mathcal{N}(0, 1))}_{\text{regularize } z_1 \text{ through global prior}}$$
$$- \underbrace{D_{KL}(\mathcal{N}(\mu_{z_2}, \sigma_{z_2})||\mathcal{N}(\tilde{\mu_2}, 0.5))}_{\text{regularize } z_2 \text{ through seq. dep. prior}}$$
$$+ \underbrace{log\ p(\tilde{\mu_2}) \cdot \frac{1}{seq.\ length}}_{\text{prob. of } \tilde{\mu_2} \text{ under standard Gaussian prior}}$$

$$log\ p(sequence\ id|z_2) = \underbrace{CrossEntropy(\frac{-(\mu_{z_2} - \mu_2)^2}{\sigma_{z_2}^2},\ sequence\ id)}_{\text{Try to predict sequence id with } z_2}$$

FHVAE

Objective - Discriminative segment variational lower bound

$$\mathcal{L}^{dis}(\theta, \phi; x) = \mathcal{L}(\theta, \phi, X) + \alpha \cdot log\ p(sequence\ id|z_2)$$

- Combined objective
  - encourage factorization
- discriminative objective can be adjusted through α hyperparameter
  - encourage μ_{z_2} to become more meaningful

- Task: Speaker verification
  - Allows quantitative analysis of performance and of quality of disentanglement
  - use FHVAE's s-vector μ_2 to predict speaker
- Compare i-vector baseline
  - i-vector is used in SOTA speaker verification approaches
  - low dim subspace of GMM universal background model
  - subspace of speaker (content-independent) information

- Unsupervised speaker verification (Raw column)
- Equal error rates (lower is better)
- use μ_1 as sanity check

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%

- Some evidence towards disentangling with respect to sequence-segment decomposition
  - other decompositions may prove more challenging
  - speaker gender, speaker age, language as more fine grained decompositions
- I had trouble disentangling simple examples
- Good performance on speaker verification and denoising task

References

- Hsu, W.N., Zhang, Y. and Glass, J., 2017. Unsupervised learning of disentangled and interpretable representations from sequential data. In Advances in neural information processing systems (pp. 1878-1889).
- Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D. and Lerchner, A., 2018. Towards a definition of disentangled representations. arXiv preprint arXiv:1812.02230.
- Scott, W.R., 2012. Group theory. Courier Corporation.
- Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

- 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

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

- 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?

- 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?

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

- 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