

# Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data

Wei-Ning Hsu, Yu Zhang, and James Glass  
Talk by Stefan Wezel

Seminar ML4S

February 15, 2021

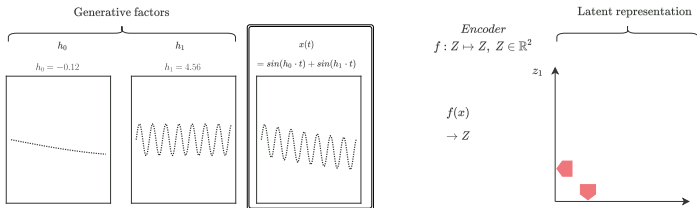
- Introduction
- What are Disentangled Representations? (intuition)
- Why Disentangled Representations
- Formal Description of Disentangled Representations
- Disentanglement in the Context of the Paper
- The Factorized Hierarchical VAE Model
- Results

- Propose Sequential Factorized Hierarchical VAE (FHVAE)
- Focus on speech data
  - Sequence level (Speaker, ...) attributes
  - Segment level (content, noise, ...) attributes
- Reflect different temporal scales in latent space

# What is disentanglement?

## Intuition

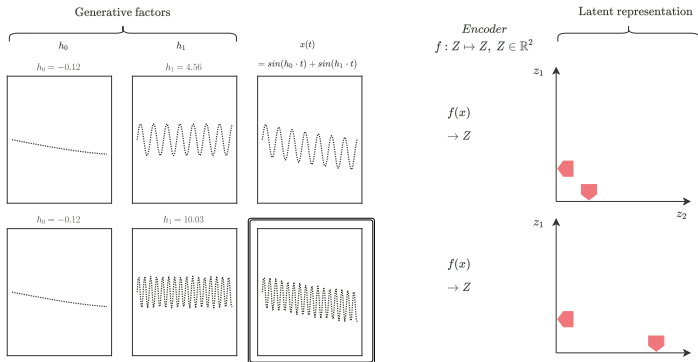
- Encode distinct generating factors in separate subsets of latent space dimensions



# What is disentanglement?

## Intuition

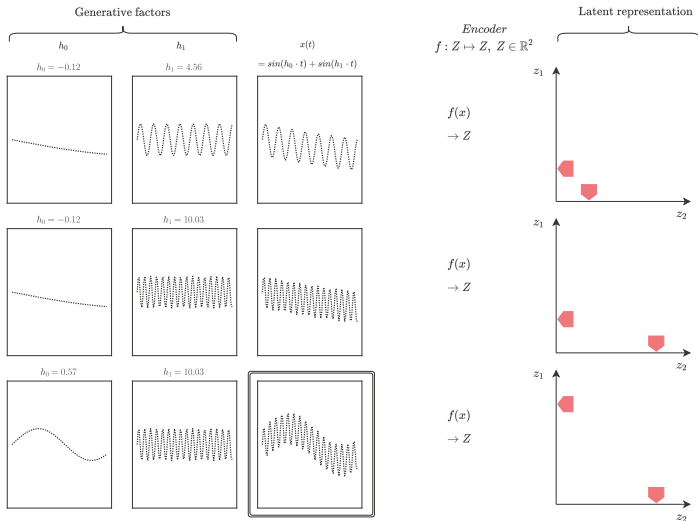
- Encode distinct generating factors in separate subsets of latent space dimensions



# What is disentanglement?

## Intuition

- 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
  - Operation and non-empty set  $G = (\circ, G)$
  - Set closed under operation, identity element, inverse elements, associativity
- Symmetry group
  - Set of transformations that leave object  $X$  invariant
  - Operation is composition of transformations
- Group action
  - Results of symmetry transformations on object  $X$
  - I.e. set of changed order
- Direct product
  - $G = G_1 \times \dots \times G_n$



# Disentangled Representations Formally

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

- Disentangled group actions
  - Result of subset of symmetries  $G_i$  that only change subset  $X_i$  of object, but leave other  $X_{j \neq i}$  invariant
- $\rightarrow$  If we observe disentangled group actions in the world, we want to model those
- We can assume  $G$  can be decomposed into direct product of symmetry subgroups  $G_i$

# Disentangled Representations Formally

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

- We want to find symmetry preserving mapping  $f : X \mapsto Z$

$$\begin{array}{ccc} X & \xrightarrow{G} & X \\ f \downarrow & & f \downarrow \\ Z & \xrightarrow{G} & Z \end{array}$$

- $\rightarrow$  Equivariant map  $g \cdot f(x) = f(g \cdot x)$
- Result is disentangled representation
  - Decomposition  $Z = Z_1 \times \dots \times Z_n$
  - $Z_i$  only affected by symmetry  $G_i$  on  $X$
  - $Z_i$  invariant to all  $G_{j \neq i}$

# Back to the paper

Did they achieve disentanglement?

- Disentangled with respect to what decomposition?
- Assume decomposition  $G = G_{sequence} \times G_{segment}$
- Reflect decomposition in  $Z = (z_1, z_2)$
- $z_1$ : segment,  $z_2$ : sequence
- Find equivariant map  $f : W \mapsto Z$ , so that  $z_2$  is only affected by symmetries  $G_{sequence}$  and vice versa

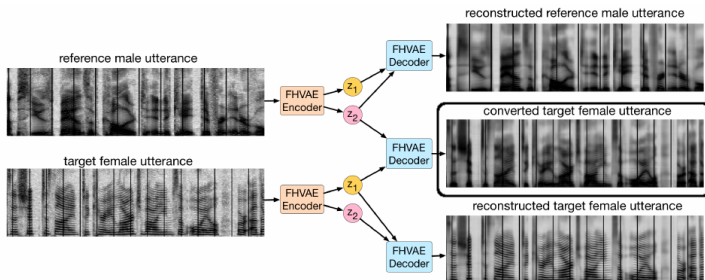
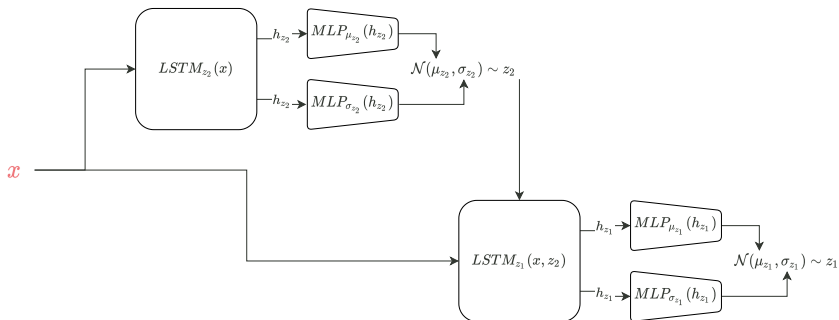
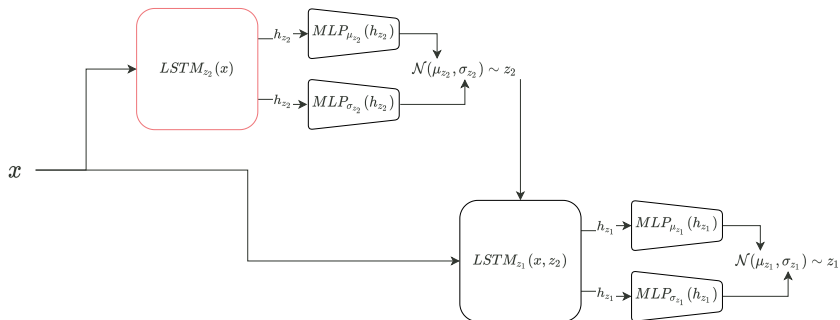
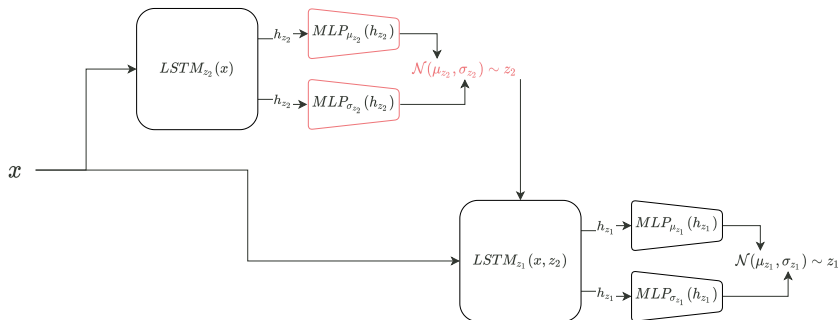
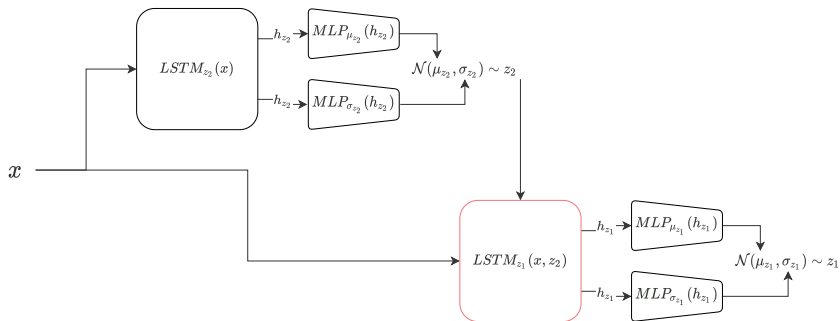


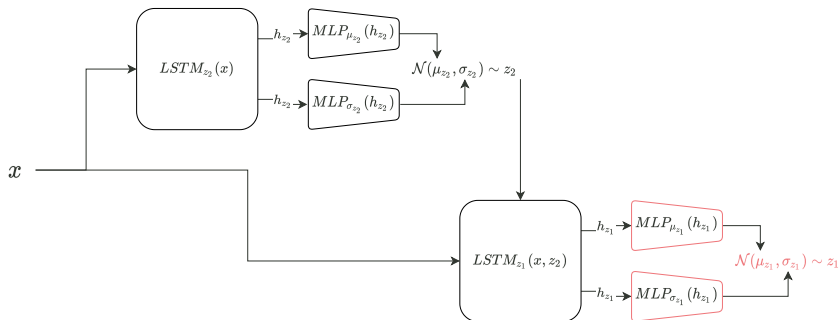
Image source: Hsu et al., 2017



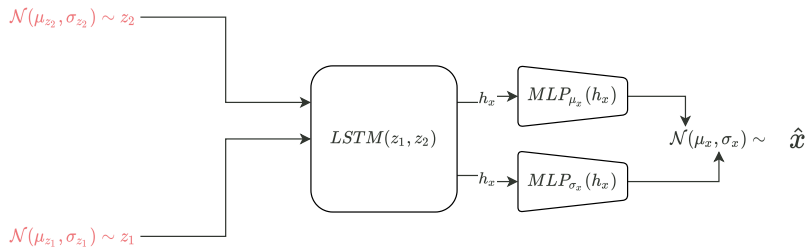


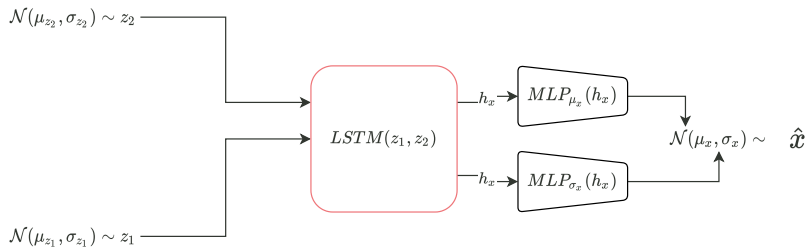


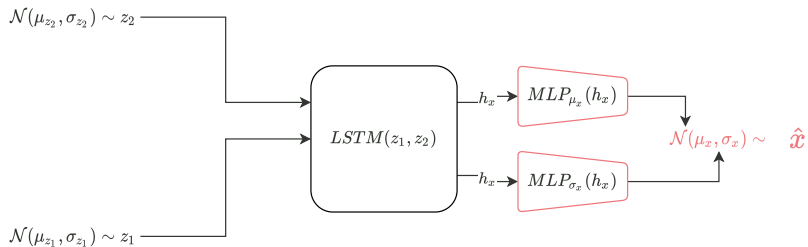












$$\mathcal{L}(\theta, \phi, X) = \sum_{n=1}^N \underbrace{\mathcal{L}(\theta, \phi; x^{(n)} | \mu_2) + \log p_{\theta}(\mu_2) + \text{const.}}_{\text{var. lower bound}} + \underbrace{\alpha \cdot \log(i | z_2^{(i,n)})}_{\text{discrim.obj.}}$$

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

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

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

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

- What is sequence dependent prior  $\mu_2$  (s-vector)?
  - Imagine a word vector
  - S-vector for every sequence
  - Similar sequence attributes  $\rightarrow$  s-vectors close in euclidean space
  - $g(\text{sequence id}) = \mu_2$  as (differentiable) lookup table
    - $\rightarrow$  Embedding in pytorch, tensorflow

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

$$\log p(\text{sequence } id^{(i)} | z_2^{(i,n)}) = \log \frac{p(\mu_{z_2}^{(i,n)} | \mu_2^{(i)})}{\sum_{j=1}^{num \text{ seqs}} p(\mu_{z_2}^{(i,n)} | \mu_2^{(j)})}$$

- Encourage  $z_2^{(i)}$  to be close to corresponding  $\mu_2^{(i)}$
- and far from all other sequence's s-vectors  $\mu_2^{(j \neq i)}$

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

- Joint objective to encourage factorization
- $\alpha$  hyperparameter to weigh discriminative objective



- Task: Speaker verification
  - Allows quantitative analysis of performance
  - Assess quality of disentanglement
  - Use s-vector  $\mu_2$  to predict speaker
- Compare i-vector baseline
  - i-vector used in SOTA speaker verification approaches
  - Low dimensional subspace of GMM universal background model
  - Contains speaker information (content-independent)

- Unsupervised speaker verification (Raw column)
- Metric: equal error rate (lower is better)
- $\mu_1$  based on  $z_1$  as sanity check

Features	Dimension	$\alpha$	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%
$\mu_2$	16	0	5.06%	4.02%	-
	16	$10^{-1}$	4.91%	4.61%	-
	16	$10^0$	3.87%	3.86%	-
	16	$10^1$	<b>2.38%</b>	<b>2.08%</b>	-
	32	$10^1$	<b>2.38%</b>	<b>2.08%</b>	<b>1.34%</b>
$\mu_1$	16	$10^0$	22.77%	15.62%	-
	16	$10^1$	27.68%	22.17%	-
	32	$10^1$	22.47%	16.82%	17.26%

- Evidence towards disentangling with respect to sequence-segment decomposition
  - Other decompositions may prove more challenging
- Good performance on speaker verification and denoising task
- I had trouble disentangling simple examples
- Questions for you:
  - Is learning disentangled representations worth the effort?
  - Have there been situations where you wished for an interpretable latent space?
  - Do you know any successful models where equivariant maps are used?

## 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).
- Hsu, W.N. and Glass, J., 2018. Scalable factorized hierarchical variational autoencoder training. arXiv preprint arXiv:1804.03201.
- Higgins, I., Amos, D., Pfau, D., Raeaniere, 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.

- 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 decomposed into can be helpful as well as biases towards the 'form' of these latent subspaces