

Deep Learning for Computer Vision

# VAEs and Disentanglement

Vineeth N Balasubramanian

Department of Computer Science and Engineering  
Indian Institute of Technology, Hyderabad



# What is Disentanglement?

- Isolating sources of variation in observational data
  - E.g. separating underlying concepts of “**Big Red Apple**”: size (*big*), color (*red*) and shape (*apple*)



NPTEL

# What is Disentanglement?

- Isolating sources of variation in observational data
  - E.g. separating underlying concepts of “**Big Red Apple**”: size (*big*), color (*red*) and shape (*apple*)
- Can we isolate these factors using some representation learning method?



NPTEL

# What is Disentanglement?

- Isolating sources of variation in observational data
  - E.g. separating underlying concepts of “**Big Red Apple**”: size (*big*), color (*red*) and shape (*apple*)
- Can we isolate these factors using some representation learning method?
- Why do we need this?



NPTEL

# What is Disentanglement?



- Isolating sources of variation in observational data
  - E.g. separating underlying concepts of “**Big Red Apple**”: size (*big*), color (*red*) and shape (*apple*)
- Can we isolate these factors using some representation learning method?
- Why do we need this? Useful to generate new images that are not in observed dataset
  - E.g. Generate an image corresponding to “**Small Black Apple**” using a model that was trained on “*Small Black Grapes*” and “*Big Red Apples*”

## Disentanglement: Example

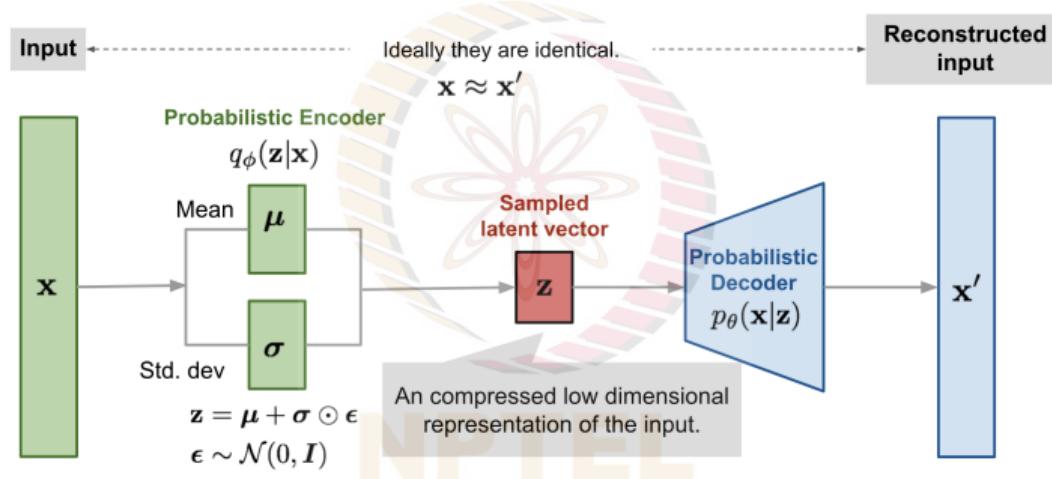


Images generated when latents (dimensions encoding generative factors) corresponding to gender are changed; more control when latents are disentangled

Credit: Chen et al, *Isolating Sources of Disentanglement in Variational Autoencoders*, NeurIPS 2018

# Disentanglement: Why VAEs?

Recall VAEs:



VAEs learn latent variables which can be used to generate data; if these latent variables are disentangled, allows controlled generation of images

Credit: [Lilian Weng](#)

# $\beta$ -VAE<sup>1</sup>

- A variant of VAE which allows disentanglement

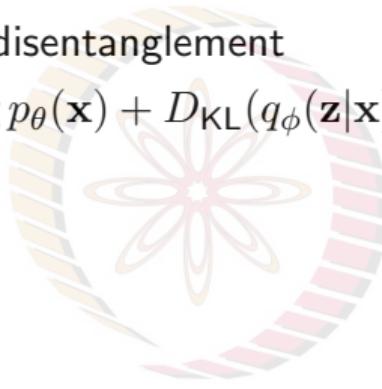


---

<sup>1</sup>Higgins et al, beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, ICLR 2017

## $\beta$ -VAE<sup>1</sup>

- A variant of VAE which allows disentanglement
- Recall **VAE loss**:  $L_{\text{VAE}} = -\log p_{\theta}(\mathbf{x}) + D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x})\|p_{\theta}(\mathbf{z}|\mathbf{x}))$



---

<sup>1</sup>Higgins et al, beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, ICLR 2017

## $\beta$ -VAE<sup>1</sup>

- A variant of VAE which allows disentanglement
- Recall **VAE loss**:  $L_{\text{VAE}} = -\log p_{\theta}(\mathbf{x}) + D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x})\|p_{\theta}(\mathbf{z}|\mathbf{x}))$
- Another way of writing the VAE objective:

$$\max_{\phi, \theta} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} [\mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z})]$$

$$\text{subject to } D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x})\|p_{\theta}(\mathbf{z})) < \delta$$

Maximize probability of generating real data, while keeping distance between real and approximate posterior distributions small (under a small constant  $\delta$ )

---

<sup>1</sup>Higgins et al, beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, ICLR 2017

# $\beta$ -VAE

- VAE maximization objective can then be rewritten as a Lagrangian with a Lagrangian multiplier  $\beta$  under KKT conditions (similar to SVM):

$$\begin{aligned} & \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta(D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x})\|p_{\theta}(\mathbf{z})) - \delta) \\ &= \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x})\|p_{\theta}(\mathbf{z})) + \beta\delta \\ &\geq \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x})\|p_{\theta}(\mathbf{z})) \quad \text{since } \beta, \delta \geq 0 \end{aligned}$$

NPTEL

## $\beta$ -VAE

- VAE maximization objective can then be rewritten as a Lagrangian with a Lagrangian multiplier  $\beta$  under KKT conditions (similar to SVM):

$$\begin{aligned} & \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta(D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z})) - \delta) \\ &= \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z})) + \beta\delta \\ &\geq \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z})) \quad \text{since } \beta, \delta \geq 0 \end{aligned}$$

- $\beta$ -VAE loss hence given by:

$$L_{\text{BETA}}(\phi, \beta) = -\mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) + \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z}))$$

NPTEL

## $\beta$ -VAE

- VAE maximization objective can then be rewritten as a Lagrangian with a Lagrangian multiplier  $\beta$  under KKT conditions (similar to SVM):

$$\begin{aligned} & \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta(D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z})) - \delta) \\ &= \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z})) + \beta\delta \\ &\geq \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z})) \quad \text{since } \beta, \delta \geq 0 \end{aligned}$$

- $\beta$ -VAE loss hence given by:

$$L_{\text{BETA}}(\phi, \beta) = -\mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) + \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z}))$$

- When  $\beta = 1 \rightarrow$  standard VAE

NPTEL

## $\beta$ -VAE

- VAE maximization objective can then be rewritten as a Lagrangian with a Lagrangian multiplier  $\beta$  under KKT conditions (similar to SVM):

$$\begin{aligned} & \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta(D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z})) - \delta) \\ &= \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z})) + \beta\delta \\ &\geq \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z})) \quad \text{since } \beta, \delta \geq 0 \end{aligned}$$

- $\beta$ -VAE loss hence given by:

$$L_{\text{BETA}}(\phi, \beta) = -\mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) + \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z}))$$

- When  $\beta = 1 \rightarrow$  standard VAE
- When  $\beta > 1 \rightarrow$  stronger constraint on latent bottleneck, follow generative process and thus encourage **disentanglement**

## $\beta$ -VAE

- VAE maximization objective can then be rewritten as a Lagrangian with a Lagrangian multiplier  $\beta$  under KKT conditions (similar to SVM):

$$\begin{aligned} & \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta(D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z})) - \delta) \\ &= \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z})) + \beta\delta \\ &\geq \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z})) \quad \text{since } \beta, \delta \geq 0 \end{aligned}$$

- **$\beta$ -VAE** loss hence given by:

$$L_{\text{BETA}}(\phi, \beta) = -\mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) + \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z}))$$

- When  $\beta = 1 \rightarrow$  standard VAE
- When  $\beta > 1 \rightarrow$  stronger constraint on latent bottleneck, follow generative process and thus encourage **disentanglement**
- Could limit representation capacity of  $\mathbf{z}$ , creating a trade-off between reconstruction quality and extent of disentanglement

Credit: [Lilian Weng](#)

# $\beta$ -TCVAE<sup>2</sup>

- **Disadvantage of  $\beta$ -VAE:** Trade-off between disentanglement and reconstruction capability. How can we get both?

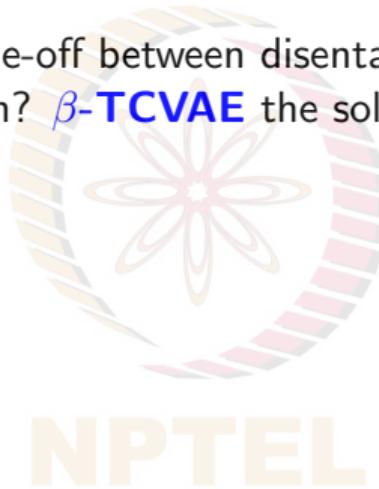


---

<sup>2</sup>Chen et al, Isolating Sources of Disentanglement in Variational Autoencoders, NeurIPS 2018

# $\beta$ -TCVAE<sup>2</sup>

- **Disadvantage of  $\beta$ -VAE:** Trade-off between disentanglement and reconstruction capability. How can we get both?  $\beta$ -TCVAE the solution



---

<sup>2</sup>Chen et al, Isolating Sources of Disentanglement in Variational Autoencoders, NeurIPS 2018

## $\beta$ -TCVAE<sup>2</sup>

- **Disadvantage of  $\beta$ -VAE:** Trade-off between disentanglement and reconstruction capability. How can we get both?  $\beta$ -TCVAE the solution
- KL-divergence term can be decomposed as:

$$D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x})\|p_{\theta}(\mathbf{z})) = \underbrace{I_q(\mathbf{z}, \mathbf{n})}_{\text{index-code mutual information (MI)}} + \underbrace{D_{\text{KL}}(q_{\phi}(\mathbf{z})\|p_{\theta}(\mathbf{z}))}_{\text{marginal KL to prior}}$$

NPTEL

<sup>2</sup>Chen et al, Isolating Sources of Disentanglement in Variational Autoencoders, NeurIPS 2018

## $\beta$ -TCVAE<sup>2</sup>

- **Disadvantage of  $\beta$ -VAE:** Trade-off between disentanglement and reconstruction capability. How can we get both?  $\beta$ -TCVAE the solution
- KL-divergence term can be decomposed as:

$$D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x})\|p_{\theta}(\mathbf{z})) = \underbrace{I_q(\mathbf{z}, \mathbf{n})}_{\text{index-code mutual information (MI)}} + \underbrace{D_{\text{KL}}(q_{\phi}(\mathbf{z})\|p_{\theta}(\mathbf{z}))}_{\text{marginal KL to prior}}$$

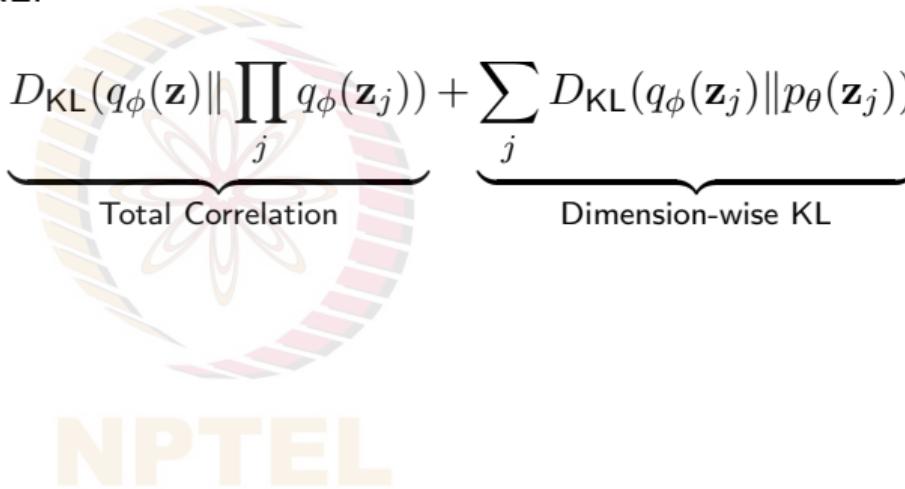
- **Marginal KL to prior** more important to learn disentangled representations; reducing MI might be causing poor reconstruction. What to do?

<sup>2</sup>Chen et al, Isolating Sources of Disentanglement in Variational Autoencoders, NeurIPS 2018

# $\beta$ -TCVAE<sup>3</sup>

- Further decompose marginal KL:

$$D_{\text{KL}}(q_{\phi}(\mathbf{z}) \| p_{\theta}(\mathbf{z})) = \underbrace{D_{\text{KL}}(q_{\phi}(\mathbf{z}) \| \prod_j q_{\phi}(\mathbf{z}_j))}_{\text{Total Correlation}} + \underbrace{\sum_j D_{\text{KL}}(q_{\phi}(\mathbf{z}_j) \| p_{\theta}(\mathbf{z}_j))}_{\text{Dimension-wise KL}}$$



<sup>3</sup>Chen et al, Isolating Sources of Disentanglement in Variational Autoencoders, NeurIPS 2018

## $\beta$ -TCVAE<sup>3</sup>

- Further decompose marginal KL:

$$D_{\text{KL}}(q_{\phi}(\mathbf{z}) \| p_{\theta}(\mathbf{z})) = \underbrace{D_{\text{KL}}(q_{\phi}(\mathbf{z}) \| \prod_j q_{\phi}(\mathbf{z}_j))}_{\text{Total Correlation}} + \underbrace{\sum_j D_{\text{KL}}(q_{\phi}(\mathbf{z}_j) \| p_{\theta}(\mathbf{z}_j))}_{\text{Dimension-wise KL}}$$

- **Total Correlation** important for learning disentangled representation

NPTEL

<sup>3</sup>Chen et al, Isolating Sources of Disentanglement in Variational Autoencoders, NeurIPS 2018

## $\beta$ -TCVAE<sup>3</sup>

- Further decompose marginal KL:

$$D_{\text{KL}}(q_{\phi}(\mathbf{z}) \| p_{\theta}(\mathbf{z})) = \underbrace{D_{\text{KL}}(q_{\phi}(\mathbf{z}) \| \prod_j q_{\phi}(\mathbf{z}_j))}_{\text{Total Correlation}} + \underbrace{\sum_j D_{\text{KL}}(q_{\phi}(\mathbf{z}_j) \| p_{\theta}(\mathbf{z}_j))}_{\text{Dimension-wise KL}}$$

- Total Correlation** important for learning disentangled representation
- Hence, final  $\beta$ -TCVAE loss:

$$-\mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) + I_q(\mathbf{z}, \mathbf{n}) + \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}) \| \prod_j q_{\phi}(\mathbf{z}_j)) + \sum_j D_{\text{KL}}(q_{\phi}(\mathbf{z}_j) \| p_{\theta}(\mathbf{z}_j))$$

<sup>3</sup>Chen et al, Isolating Sources of Disentanglement in Variational Autoencoders, NeurIPS 2018

## $\beta$ -TCVAE<sup>3</sup>

- Further decompose marginal KL:

$$D_{\text{KL}}(q_{\phi}(\mathbf{z}) \| p_{\theta}(\mathbf{z})) = \underbrace{D_{\text{KL}}(q_{\phi}(\mathbf{z}) \| \prod_j q_{\phi}(\mathbf{z}_j))}_{\text{Total Correlation}} + \underbrace{\sum_j D_{\text{KL}}(q_{\phi}(\mathbf{z}_j) \| p_{\theta}(\mathbf{z}_j))}_{\text{Dimension-wise KL}}$$

- **Total Correlation** important for learning disentangled representation
- Hence, final  $\beta$ -TCVAE loss:

$$-\mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) + I_q(\mathbf{z}, \mathbf{n}) + \beta D_{\text{KL}}(q_{\phi}(\mathbf{z}) \| \prod_j q_{\phi}(\mathbf{z}_j)) + \sum_j D_{\text{KL}}(q_{\phi}(\mathbf{z}_j) \| p_{\theta}(\mathbf{z}_j))$$

- Weight  $\beta > 1$  to disentangle without affecting reconstruction

<sup>3</sup>Chen et al, Isolating Sources of Disentanglement in Variational Autoencoders, NeurIPS 2018

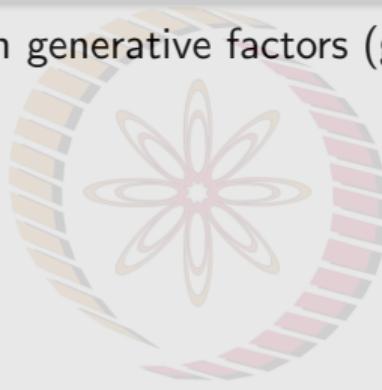
# Disentangled Representation Learning: How to evaluate?



# Disentangled Representation Learning: How to evaluate?

## Mutual Information Gap (MIG)

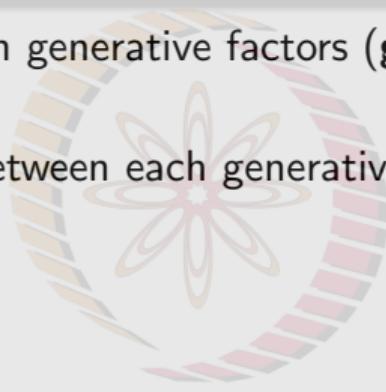
- Use mutual information between generative factors ( $g$ ) and latent dimensions ( $z$ ) in some way; how?



# Disentangled Representation Learning: How to evaluate?

## Mutual Information Gap (MIG)

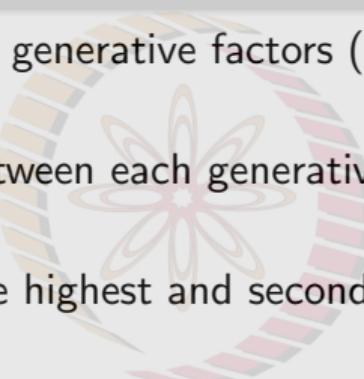
- Use mutual information between generative factors ( $\mathbf{g}$ ) and latent dimensions ( $\mathbf{z}$ ) in some way; how?
- Compute mutual information between each generative factor ( $\mathbf{g}_i$ ) and each latent dimension ( $\mathbf{z}_i$ )



# Disentangled Representation Learning: How to evaluate?

## Mutual Information Gap (MIG)

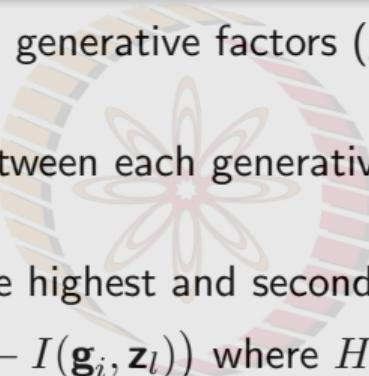
- Use mutual information between generative factors ( $\mathbf{g}$ ) and latent dimensions ( $\mathbf{z}$ ) in some way; how?
- Compute mutual information between each generative factor ( $\mathbf{g}_i$ ) and each latent dimension ( $\mathbf{z}_i$ )
- For each  $\mathbf{g}_i$ , take  $\mathbf{z}_j, \mathbf{z}_l$  that have highest and second highest mutual information with  $\mathbf{g}_i$



# Disentangled Representation Learning: How to evaluate?

## Mutual Information Gap (MIG)

- Use mutual information between generative factors ( $\mathbf{g}$ ) and latent dimensions ( $\mathbf{z}$ ) in some way; how?
- Compute mutual information between each generative factors ( $\mathbf{g}_i$ ) and each latent dimension ( $\mathbf{z}_i$ )
- For each  $\mathbf{g}_i$ , take  $\mathbf{z}_j, \mathbf{z}_l$  that have highest and second highest mutual information with  $\mathbf{g}_i$
- $\text{MIG} = \frac{1}{K} \sum_{i=1}^K \frac{1}{H(\mathbf{g}_i)} (I(\mathbf{g}_i, \mathbf{z}_j) - I(\mathbf{g}_i, \mathbf{z}_l))$  where  $H(\mathbf{g}_i)$  is entropy of  $\mathbf{g}_i$  and  $0 \leq I(\mathbf{g}_i, \mathbf{z}_j) \leq H(\mathbf{g}_i)$



NPTEL

# Disentangled Representation Learning: How to evaluate?

## Mutual Information Gap (MIG)

- Use mutual information between generative factors ( $\mathbf{g}$ ) and latent dimensions ( $\mathbf{z}$ ) in some way; how?
- Compute mutual information between each generative factors ( $\mathbf{g}_i$ ) and each latent dimension ( $\mathbf{z}_i$ )
- For each  $\mathbf{g}_i$ , take  $\mathbf{z}_j, \mathbf{z}_l$  that have highest and second highest mutual information with  $\mathbf{g}_i$
- $\text{MIG} = \frac{1}{K} \sum_{i=1}^K \frac{1}{H(\mathbf{g}_i)} (I(\mathbf{g}_i, \mathbf{z}_j) - I(\mathbf{g}_i, \mathbf{z}_l))$  where  $H(\mathbf{g}_i)$  is entropy of  $\mathbf{g}_i$  and  $0 \leq I(\mathbf{g}_i, \mathbf{z}_j) \leq H(\mathbf{g}_i)$
- Averaging by  $K$  and normalizing by  $H(\mathbf{g}_i)$  provides values between 0 and 1

# Disentangled Representation Learning: How to evaluate?

## Mutual Information Gap (MIG)

- Use mutual information between generative factors ( $\mathbf{g}$ ) and latent dimensions ( $\mathbf{z}$ ) in some way; how?
- Compute mutual information between each generative factor ( $\mathbf{g}_i$ ) and each latent dimension ( $\mathbf{z}_i$ )
- For each  $\mathbf{g}_i$ , take  $\mathbf{z}_j, \mathbf{z}_l$  that have highest and second highest mutual information with  $\mathbf{g}_i$
- $\text{MIG} = \frac{1}{K} \sum_{i=1}^K \frac{1}{H(\mathbf{g}_i)} (I(\mathbf{g}_i, \mathbf{z}_j) - I(\mathbf{g}_i, \mathbf{z}_l))$  where  $H(\mathbf{g}_i)$  is entropy of  $\mathbf{g}_i$  and  $0 \leq I(\mathbf{g}_i, \mathbf{z}_j) \leq H(\mathbf{g}_i)$
- Averaging by  $K$  and normalizing by  $H(\mathbf{g}_i)$  provides values between 0 and 1
- $\text{MIG} \rightarrow 0$  : bad disentanglement,  $\text{MIG} \rightarrow 1$  : good disentanglement

# Disentangled Representation Learning: How to evaluate?

## Mutual Information Gap (MIG)

- Use mutual information between generative factors ( $\mathbf{g}$ ) and latent dimensions ( $\mathbf{z}$ ) in some way; how?
- Compute mutual information between each generative factor ( $\mathbf{g}_i$ ) and each latent dimension ( $\mathbf{z}_i$ )
- For each  $\mathbf{g}_i$ , take  $\mathbf{z}_j, \mathbf{z}_l$  that have highest and second highest mutual information with  $\mathbf{g}_i$
- $\text{MIG} = \frac{1}{K} \sum_{i=1}^K \frac{1}{H(\mathbf{g}_i)} (I(\mathbf{g}_i, \mathbf{z}_j) - I(\mathbf{g}_i, \mathbf{z}_l))$  where  $H(\mathbf{g}_i)$  is entropy of  $\mathbf{g}_i$  and  $0 \leq I(\mathbf{g}_i, \mathbf{z}_j) \leq H(\mathbf{g}_i)$
- Averaging by  $K$  and normalizing by  $H(\mathbf{g}_i)$  provides values between 0 and 1
- $\text{MIG} \rightarrow 0$  : bad disentanglement,  $\text{MIG} \rightarrow 1$  : good disentanglement
- Why not simply use MI? Why MI gap?

# Disentangled Representation Learning: How to evaluate?

## Mutual Information Gap (MIG)

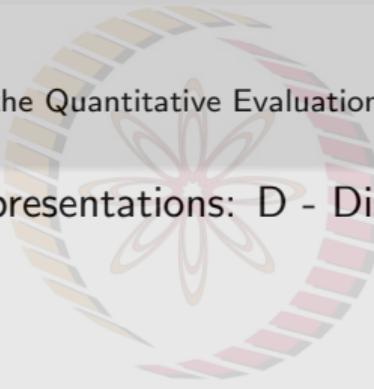
- Use mutual information between generative factors ( $\mathbf{g}$ ) and latent dimensions ( $\mathbf{z}$ ) in some way; how?
- Compute mutual information between each generative factor ( $\mathbf{g}_i$ ) and each latent dimension ( $\mathbf{z}_i$ )
- For each  $\mathbf{g}_i$ , take  $\mathbf{z}_j, \mathbf{z}_l$  that have highest and second highest mutual information with  $\mathbf{g}_i$
- $\text{MIG} = \frac{1}{K} \sum_{i=1}^K \frac{1}{H(\mathbf{g}_i)} (I(\mathbf{g}_i, \mathbf{z}_j) - I(\mathbf{g}_i, \mathbf{z}_l))$  where  $H(\mathbf{g}_i)$  is entropy of  $\mathbf{g}_i$  and  $0 \leq I(\mathbf{g}_i, \mathbf{z}_j) \leq H(\mathbf{g}_i)$
- Averaging by  $K$  and normalizing by  $H(\mathbf{g}_i)$  provides values between 0 and 1
- $\text{MIG} \rightarrow 0$  : bad disentanglement,  $\text{MIG} \rightarrow 1$  : good disentanglement
- Why not simply use MI? Why MI gap? **Homework!** (*Hint: Read metric section in Chen et al, Isolating Sources of Disentanglement in Variational Autoencoders, NeurIPS 2018*)

# Disentangled Representation Learning: How to evaluate?

## DCI Metric<sup>a</sup>

<sup>a</sup>Eastwood and Williams, A Framework for the Quantitative Evaluation of Disentangled Representations, ICLR 2018

- Considers three properties of representations: D - Disentanglement, C - Completeness, I - Informativeness



# Disentangled Representation Learning: How to evaluate?

## DCI Metric<sup>a</sup>

<sup>a</sup>Eastwood and Williams, A Framework for the Quantitative Evaluation of Disentangled Representations, ICLR 2018

- Considers three properties of representations: D - Disentanglement, C - Completeness, I - Informativeness
- Train a model (e.g.  $\beta$ -VAE) to get latent representations



# Disentangled Representation Learning: How to evaluate?

## DCI Metric<sup>a</sup>

<sup>a</sup>Eastwood and Williams, A Framework for the Quantitative Evaluation of Disentangled Representations, ICLR 2018

- Considers three properties of representations: D - Disentanglement, C - Completeness, I - Informativeness
- Train a model (e.g.  $\beta$ -VAE) to get latent representations
- Get latent representation of each image in a dataset



# Disentangled Representation Learning: How to evaluate?

## DCI Metric<sup>a</sup>

<sup>a</sup>Eastwood and Williams, A Framework for the Quantitative Evaluation of Disentangled Representations, ICLR 2018

- Considers three properties of representations: D - Disentanglement, C - Completeness, I - Informativeness
- Train a model (e.g.  $\beta$ -VAE) to get latent representations
- Get latent representation of each image in a dataset
- Train  $k$  linear regressors (one for each  $\mathbf{g}_i$ ),  $f_1 \dots f_k$ , to predict  $\mathbf{g}_i$  given  $\mathbf{z}$

# Disentangled Representation Learning: How to evaluate?

## DCI Metric<sup>a</sup>

<sup>a</sup>Eastwood and Williams, A Framework for the Quantitative Evaluation of Disentangled Representations, ICLR 2018

- Considers three properties of representations: D - Disentanglement, C - Completeness, I - Informativeness
- Train a model (e.g.  $\beta$ -VAE) to get latent representations
- Get latent representation of each image in a dataset
- Train  $k$  linear regressors (one for each  $\mathbf{g}_i$ ),  $f_1 \dots f_k$ , to predict  $\mathbf{g}_i$  given  $\mathbf{z}$
- From the regressors, we get  $W_{ij}$  (how much  $\mathbf{z}_i$  is important to predict  $\mathbf{g}_j$ )

# Disentangled Representation Learning: How to evaluate?

## DCI Metric<sup>a</sup>

<sup>a</sup>Eastwood and Williams, A Framework for the Quantitative Evaluation of Disentangled Representations, ICLR 2018

- Considers three properties of representations: D - Disentanglement, C - Completeness, I - Informativeness
- Train a model (e.g.  $\beta$ -VAE) to get latent representations
- Get latent representation of each image in a dataset
- Train  $k$  linear regressors (one for each  $\mathbf{g}_i$ ),  $f_1 \dots f_k$ , to predict  $\mathbf{g}_i$  given  $\mathbf{z}$
- From the regressors, we get  $W_{ij}$  (how much  $\mathbf{z}_i$  is important to predict  $\mathbf{g}_j$ )
- Create a **relative importance matrix**  $R$  such that  $R_{ij} = |W_{ij}|$

# Disentangled Representation Learning: How to evaluate?

## DCI Metric: Disentanglement

- Degree to which a representation disentangles underlying factors of variation
- Disentanglement score of  $i^{th}$  latent:  $D_i = (1 - H(P_i))$  where  $H$  is entropy and  $P_{ij} = \frac{R_{ij}}{\sum_k R_{ik}}$ , importance of  $\mathbf{z}_i$  to predict  $\mathbf{g}_j$
- Total disentanglement score:  $D = \sum_i \rho_i D_i$  where  $\rho_i = \frac{\sum_j R_{ij}}{\sum_{ij} R_{ij}}$ , relative latent importance used to normalize the score

NPTEL

# Disentangled Representation Learning: How to evaluate?

## DCI Metric: Disentanglement

- Degree to which a representation disentangles underlying factors of variation
- Disentanglement score of  $i^{th}$  latent:  $D_i = (1 - H(P_i))$  where  $H$  is entropy and  $P_{ij} = \frac{R_{ij}}{\sum_k R_{ik}}$ , importance of  $\mathbf{z}_i$  to predict  $\mathbf{g}_j$
- Total disentanglement score:  $D = \sum_i \rho_i D_i$  where  $\rho_i = \frac{\sum_j R_{ij}}{\sum_{ij} R_{ij}}$ , relative latent importance used to normalize the score

## DCI Metric: Completeness

- Degree to which each underlying generative factor is captured by a single latent variable
- For each generative factor  $\mathbf{g}_j$ ,  $C_j = (1 - H(P_j))$  where the distribution  $P_j$  is as above
- If a single latent variable contributes to  $\mathbf{g}_j$ 's prediction, score is 1 (**complete**); if all latent variables equally contribute to  $\mathbf{g}_j$ 's prediction, score is 0 (**maximally overcomplete**)

# Disentangled Representation Learning: How to evaluate?

## DCI Metric: Informativeness

- Amount of useful information a representation captures about underlying factors
- Useful for natural tasks which require knowledge of important attributes of data; e.g. for classification task, representation should capture information about object of interest

NPTEL

# Disentangled Representation Learning: How to evaluate?

## DCI Metric: Informativeness

- Amount of useful information a representation captures about underlying factors
- Useful for natural tasks which require knowledge of important attributes of data; e.g. for classification task, representation should capture information about object of interest
- **Informativeness** of  $\mathbf{z}$  about  $\mathbf{g}_j$  quantified by prediction error  $E(\mathbf{g}_j, \hat{\mathbf{g}}_j)$  where  $\hat{\mathbf{g}}_j = f_j(\mathbf{z})$

NPTEL

# Disentangled Representation Learning: How to evaluate?

## DCI Metric: Informativeness

- Amount of useful information a representation captures about underlying factors
- Useful for natural tasks which require knowledge of important attributes of data; e.g. for classification task, representation should capture information about object of interest
- **Informativeness** of  $\mathbf{z}$  about  $\mathbf{g}_j$  quantified by prediction error  $E(\mathbf{g}_j, \hat{\mathbf{g}}_j)$  where  $\hat{\mathbf{g}}_j = f_j(\mathbf{z})$
- Note that  $I$  value depends on capacity of model  $f_i$  also

# Homework

## Readings

- [Lilian Weng, From Autoencoders to Beta-VAE](#)
- [Prashnna Gyawali, Disentanglement with VAEs: A Review](#)
- (Optional) Papers on respective slides



## Questions

- Why is MI Gap and not MI used as a metric for disentanglement?

NPTFI

# References

- 
-  Ricky TQ Chen et al. "Isolating sources of disentanglement in variational autoencoders". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2610–2620.
  -  Cian Eastwood and Christopher KI Williams. "A framework for the quantitative evaluation of disentangled representations". In: *International Conference on Learning Representations*. 2018.
  -  Irina Higgins et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *ICLR*. 2017.
  -  Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2013. arXiv: [1312.6114 \[stat.ML\]](https://arxiv.org/abs/1312.6114).