Disentanglement
Basic Architectures, Ideas, and Metrics

Chi-Ching Hsu

05.04.2022

Seminar in Deep Neural Networks (FS 2022)

Disentanglement

Assumption: data is generated from independent generating factors

Generating factors z

Smile: 0.50
Skin tone: 0.50
Gender: 0.01
Glasses: 0.02
Hair color: 0.88

Latent variables c

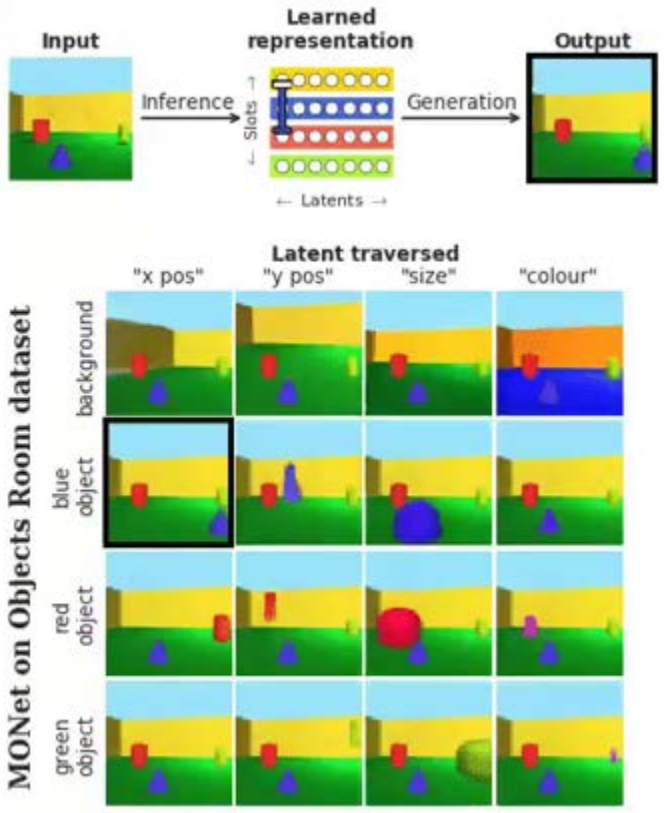
Smile: 0.50
Skin tone: 0.50
Gender: 0.01
Glasses: 0.02
Hair color: 0.88

Usually unknown

Better unsupervised

https://en.wikipedia.org/wiki/Mona_Lisa

Disentanglement



https://twitter.com/pfau/status/1275821516219985930

Motivations

- Downstream tasks (next talk)

encoder

Latent variables c

Smile: 0.50
Skin tone: 0.50
Gender: 0.01
Glasses: 0.02
Hair color: 0.88

decoder

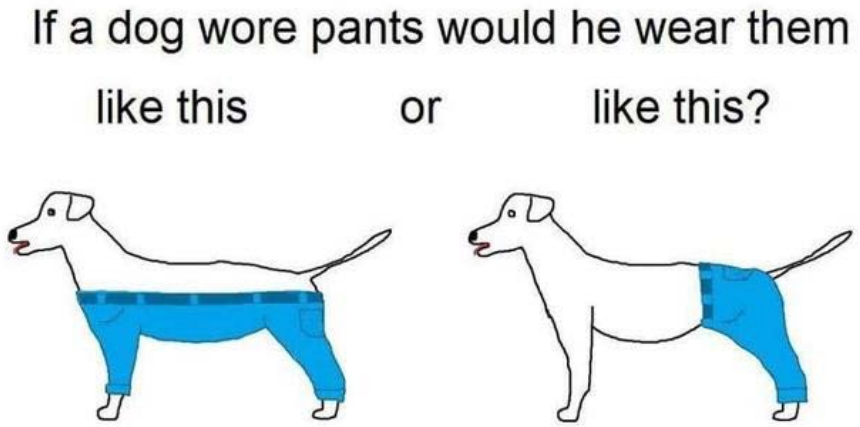
Find generating factors

Generate new data

https://en.wikipedia.org/wiki/Mona_Lisa

Challenges

- Quantify the quality of disentanglement is difficult
- Evaluation
  - Visual inspection (only qualitatively)
  - Knowing ground-truth generating factors (most of the case not possible)



https://twitter.com/pfau/status/1275821516219985930

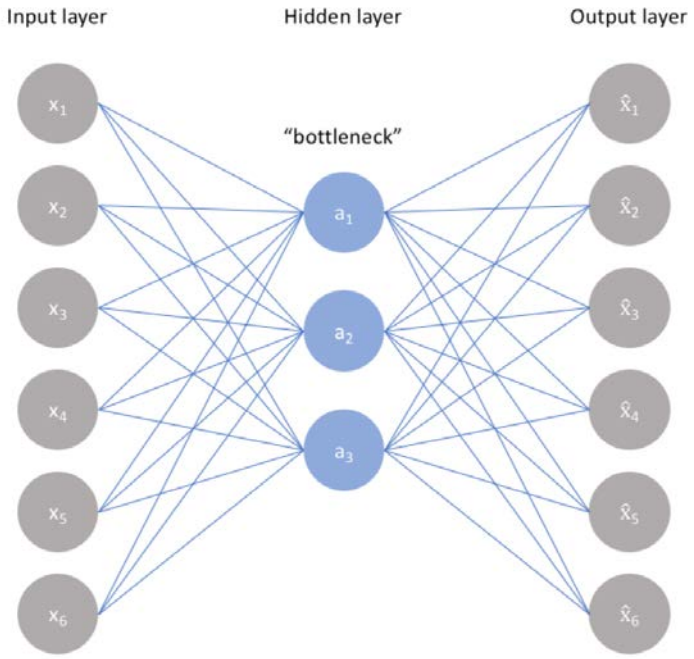
Variational Autoencoder (VAE)

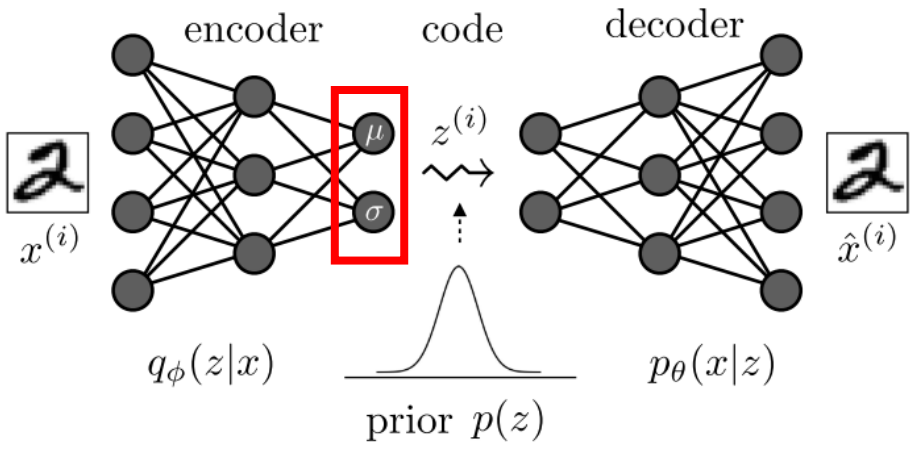
Variational Autoencoder (VAE)

- *Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).*
- *Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).*

Variational Autoencoder (VAE)

deterministic vector → probabilistic distribution

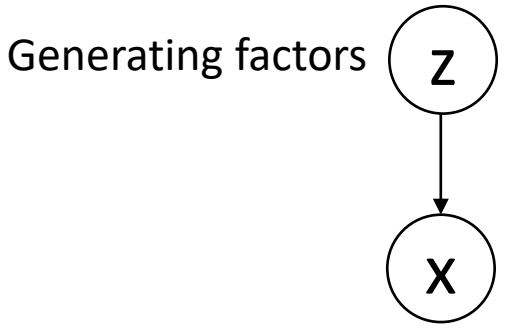




(a) Variational Autoencoder (VAE) framework.

Tschannen, Michael, Olivier Bachem, and Mario Lucic. "Recent advances in autoencoder-based representation learning." *arXiv preprint arXiv:1812.05069* (2018).
https://www.jeremyjordan.me/autoencoders/

Variational Autoencoder (VAE)



$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \quad (1)$$

Computational expensive or even intractable

$$p(x) = \int_z p(x|z)p(z)dz \quad (2)$$

$$minD_{KL}(q(z|x)||p(z|x)) \quad (3)$$

min(3) by max(4)

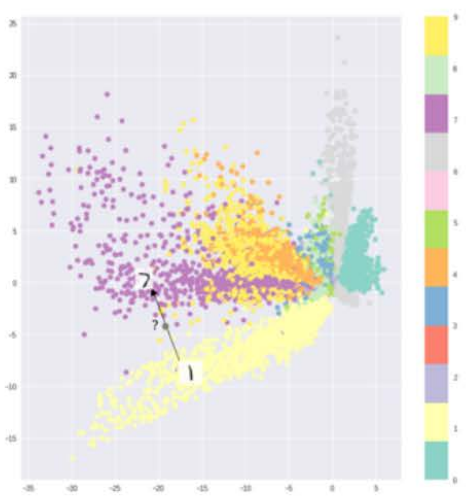
$$ELBO = \underbrace{E_{z\sim q(z|x)}[logp(x|z)] - D_{KL}(q(z|x)||p(z))}_{\text{The evidence lower bound}} \quad (4)$$

Variational Autoencoder (VAE)

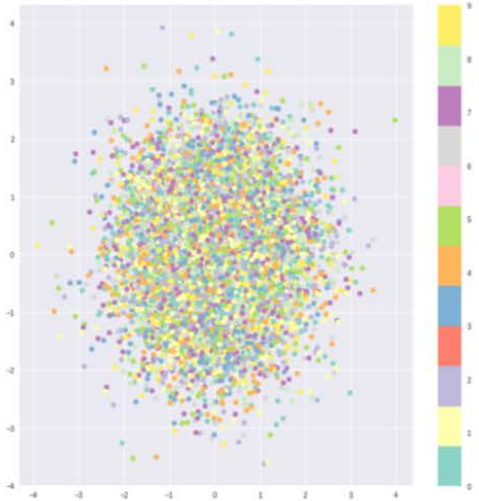
$$ELBO = E_{z\sim q(z|x)}[logp(x|z)] - D_{KL}(q(z|x)||p(z))$$

1 2

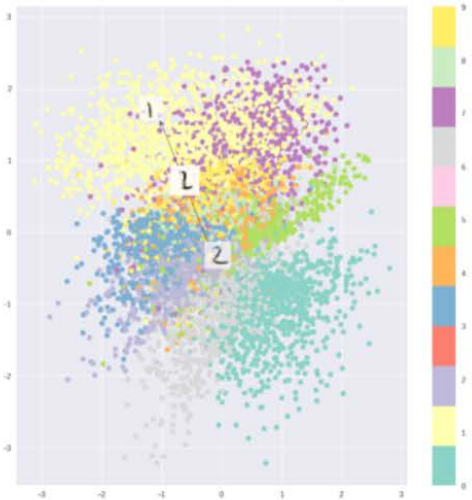
1 Only reconstruction loss



2 Only KL divergence



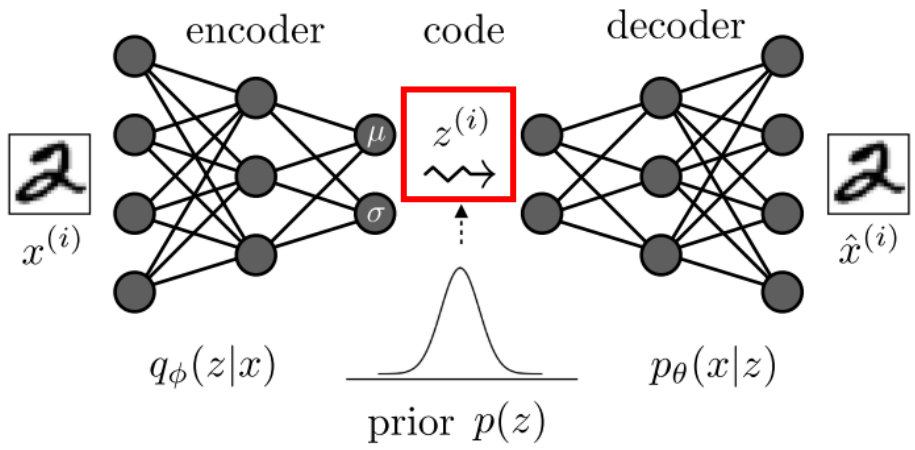
1 + 2 Combination



https://www.jeremyjordan.me/variational-autoencoders/

Variational Autoencoder (VAE)

- Powerful and widely used



(a) Variational Autoencoder (VAE) framework.

Latent variables c

c^1 Smile: 0.50
c^2 Skin tone: 0.50
c^3 Gender: 0.01
c^4 Glasses: 0.02
c^5 Hair color: 0.88

Tschannen, Michael, Olivier Bachem, and Mario Lucic. "Recent advances in autoencoder-based representation learning." *arXiv preprint arXiv:1812.05069* (2018).

β-VAE

- ICLR 2017

Encourage disentanglement

$$ELBO = E_{z\sim q(Z|X)}[logp(x|z)] - \beta D_{KL}(q(z|x)||p(z))$$

$\beta > 1 \rightarrow$ more disentangled

Higgins, Irina, et al. "beta-vae: Learning basic visual concepts with a constrained variational framework." (2016).

β-TCVAE

- NeurIPS 2018, TC: total correlation

$$ELBO = E_{z\sim q(z|x)}[logp(x|z)] - \beta D_{KL}(q(z|x)||p(z))$$

$$\mathbb{E}_{p(n)}\Big[\mathrm{KL}\big(q(z|n)||p(z)\big)\Big] = \underbrace{\mathrm{KL}\left(q(z,n)||q(z)p(n)\right)}_{\text{(i) Index-Code MI}} + \underbrace{\mathrm{KL}\big(q(z)||\prod_j q(z_j)\big)}_{\text{(ii) Total Correlation}} + \underbrace{\sum_j \mathrm{KL}\left(q(z_j)||p(z_j)\right)}_{\text{(iii) Dimension-wise KL}}$$

$$\mathcal{L}_{\beta-\mathrm{TC}} := \mathbb{E}_{q(z|n)p(n)}[\log p(n|z)] - \alpha I_q(z;n) - \beta\,\mathrm{KL}\big(q(z)||\prod_j q(z_j)\big) - \gamma\sum_j \mathrm{KL}\left(q(z_j)||p(z_j)\right)$$

$\alpha = \gamma = 1$, change only β → more disentangled

Chen, Ricky TQ, et al. "Isolating sources of disentanglement in variational autoencoders." Advances in neural information processing systems 31 (2018).

Mechanisms encourage disentanglement

Why VAE disentangled?

- The design of the VAE does not suggest any disentanglement
- Not fully clear

- *Rolinek, Michal, Dominik Zietlow, and Georg Martius. "Variational autoencoders pursue PCA directions (by accident)." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.*
- *Burgess, Christopher P., et al. "Understanding disentangling in beta-VAE." arXiv preprint arXiv:1804.03599 (2018).*

Mechanisms encourage disentanglement

Regularization of the encoding distribution

- Reweighting the ELBO: β-VAE, β-TCVAE, …

$$ELBO = E_{z\sim q(z|x)}[logp(x|z)] - \beta D_{KL}(q(z|x)||p(z))$$

$$\mathcal{L}_{\beta-\mathrm{TC}} := \mathbb{E}_{q(z|n)p(n)}[\log p(n|z)] - \alpha I_q(z;n) - \beta\ \mathrm{KL}\big(q(z)||\prod_j q(z_j)\big) - \gamma \sum_j \mathrm{KL}\left(q(z_j)||p(z_j)\right)$$

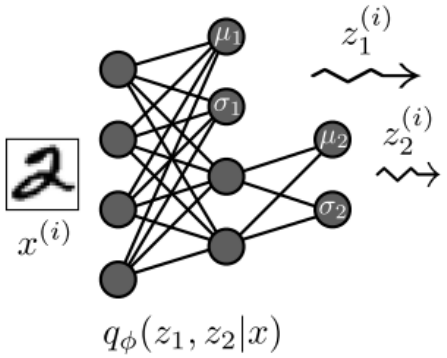
Tschannen, Michael, Olivier Bachem, and Mario Lucic. "Recent advances in autoencoder-based representation learning." *arXiv preprint arXiv:1812.05069* (2018).

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1798-1828.

Chen, Ricky TQ, et al. "Isolating sources of disentanglement in variational autoencoders." Advances in neural information processing systems 31 (2018).

Mechanisms encourage disentanglement

- Choice of the encoding and decoding distribution or model family
  - Hierarchical encoder



- Choice of a flexible prior distribution $p(z)$ of the representation

Tschannen, Michael, Olivier Bachem, and Mario Lucic. "Recent advances in autoencoder-based representation learning." *arXiv preprint arXiv:1812.05069* (2018).

Evaluation Metrics

Unsupervised Disentanglement Ranking (UDR)

- Unsupervised
- Assumptions

Disentangled representations are similar
Entangled representations are different

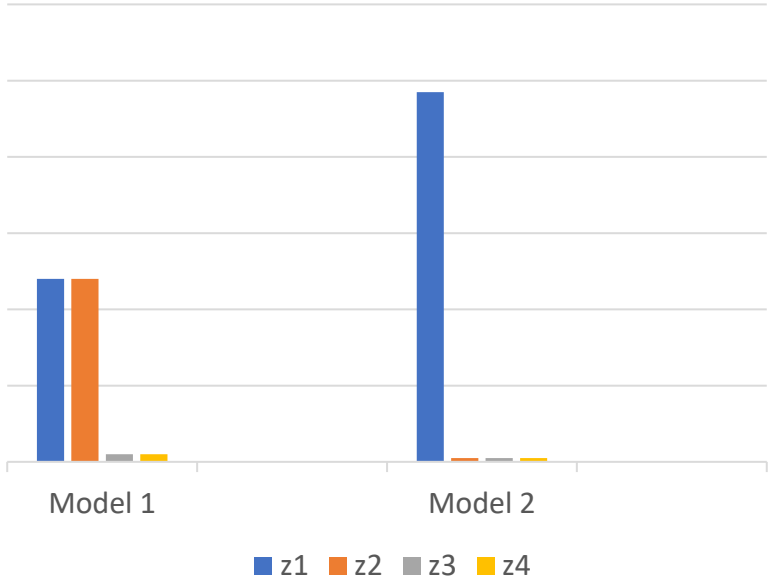
- Train many models and compare latent representation pairwise with similarity matrix

Duan, Sunny, et al. "Unsupervised model selection for variational disentangled representation learning." arXiv preprint arXiv:1905.12614 (2019).

Mutual Information Gap (MIG)

- Supervised
- A generating factor should only be captured by one latent variable

Gap small
Not disentangled

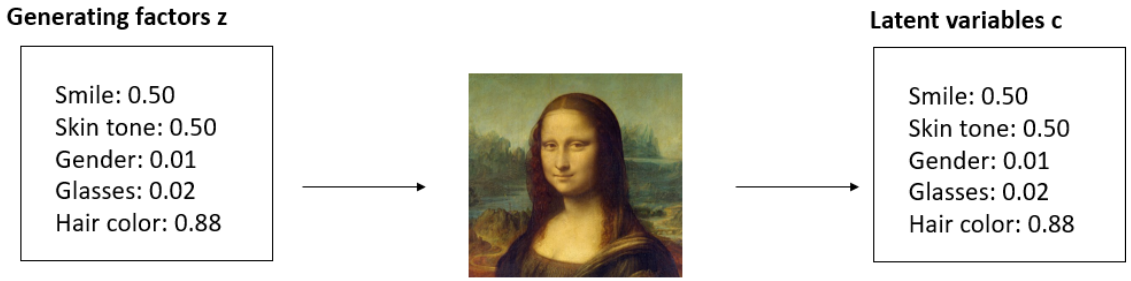


Gap large
Disentangled

Chen, Ricky TQ, et al. "Isolating sources of disentanglement in variational autoencoders." Advances in neural information processing systems 31 (2018).

Disentanglement, Completeness, Informativeness (DCI)

- Assumption: it is possible to recover latent variable c from the data

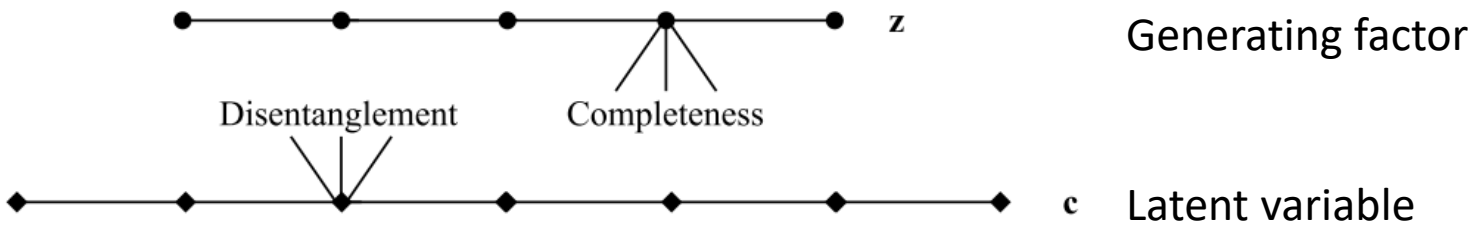


- Disentanglement (for c): one latent variable only learn one generating factor
- Completeness (for z): one generating factor only be captured by one latent variable

Eastwood, Cian, and Christopher KI Williams. "A framework for the quantitative evaluation of disentangled representations." *International Conference on Learning Representations*. 2018.

Disentanglement, Completeness, Informativeness (DCI)

- Disentanglement (for c): one latent variable only learn one generating factor
- Completeness (for z): one generating factor only be captured by one latent variable (similar to MIG)



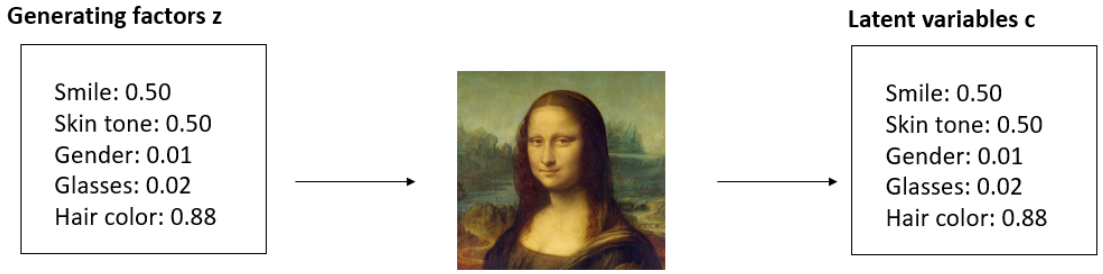
- Informativeness: the amount of information that a representation captures about the underlying factors of variation

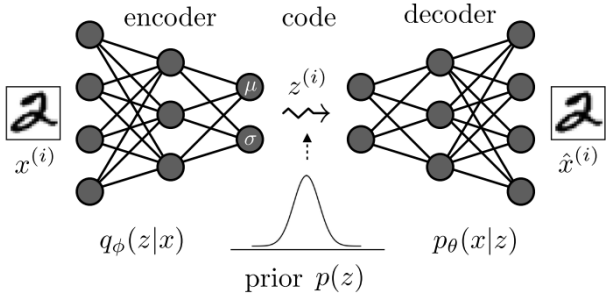
$$\hat{z}_j = f_j(c)$$
$$E(z_j, \hat{z}_j)$$

Eastwood, Cian, and Christopher KI Williams. "A framework for the quantitative evaluation of disentangled representations." *International Conference on Learning Representations*. 2018.

Summary

- Disentanglement
- VAE and disentanglement
- Why VAE disentangled?
- Mechanisms encourage disentanglement
- Evaluation metrics: UDR, MIG, DCI
  - Still difficult to evaluate





(a) Variational Autoencoder (VAE) framework.

Questions?