AUSTRIA'S LEADING RESEARCH CENTER
FOR DATA-DRIVEN BUSINESS AND BIG DATA ANALYTICS

Does IT Matter?

On Architecture and Modelling Choices in Neural IB-Type Models

Bernhard C. Geiger

Information Theory & Tapas, 25.-27.01.2023

© Know-Center GmbH, www.know-center.at

Acknowledgments

Der Wissenschaftsfonds.

FWF Grant No. J 3765

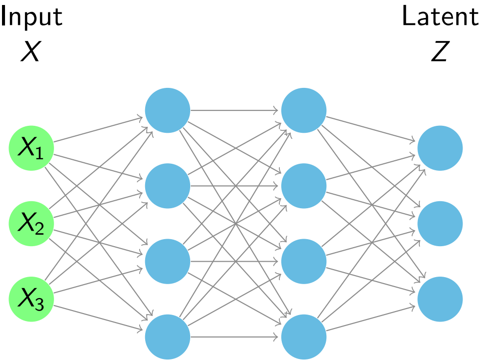
EC H2020 Grant No. 783163

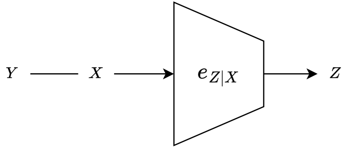
Forschung wirkt.

Joint work with:

R. Ali Amjad (Amazon)
I. S. Fischer (Google)
J. de Freitas (Know-Center)
M. Basirat & P. Roth (TU Graz)
L. Adilova & A. Fischer (Ruhr Univ. Bochum)

Setting: Neural Representation Learning





Information Bottleneck for Representation Learning

IB principle for training DNNs[1]

$$\min_{e_{Z|X} \in \mathcal{E}} I(X;Z) - \beta I(Y;Z)$$

Representation Z should be a *minimal sufficient statistic* for Y:

- sufficiency $\Leftrightarrow$ large $I(Y;Z)$
- minimality $\Leftrightarrow$ small $I(X;Z)$

[1] Tishby and Zaslavsky, "Deep learning and the information bottleneck principle", 2015

Information Bottleneck for Representation Learning

$$\min_{e_{Z|X} \in \mathcal{E}} I(X;Z) - \beta I(Y;Z)$$

- generalization bound for discrete $p_{X,Y}$[2]
- SGD, compression, and generalization behavior[3]
- $I(X;Z)$ for continuous p_X and deterministic $\mathcal{E}$[4]
- setting $Y = f(X)$[5]
- learnability of IB (smallest nontrivial β)[6]
- variational approaches (p_Z and $p_{Y|Z}$ are intractable)

[2] Vera, Piantanida, and Vega, "The Role of the Information Bottleneck in Representation Learning", 2018

[3] Shwartz-Ziv and Tishby, *Opening the Black Box of Deep Neural Networks via Information*, 2017

[4] Amjad and Geiger, "Learning Representations for Neural Network-Based Classification Using the Information Bottleneck Principle", 2020

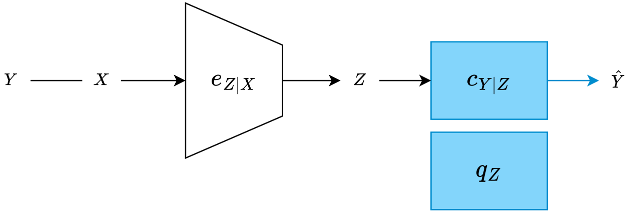
[5] Kolchinsky, Tracey, and Van Kuyk, "Caveats for information bottleneck in deterministic scenarios", 2019

[6] Wu et al., "Learnability for the Information Bottleneck", 2019

Deep Variational Information Bottleneck (VIB)[7]

$$
\begin{aligned}
I(X;Z) &+ \beta H(Y|Z) \\
&\leq \mathbb{E}\left(D\left(e_{Z|X}(\cdot|X) \| q_Z(\cdot)\right)\right) - \beta \mathbb{E}\left(\log c_{Y|Z}(Y|Z)\right)
\end{aligned}
$$

and this upper bound is minimized over $e_{Z|X}$, q_Z, and $c_{Y|Z}$.

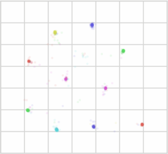


[7] Alemi et al., "Deep Variational Information Bottleneck", 2017

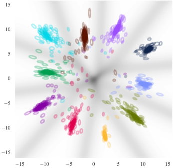
Deep Variational Information Bottleneck

This (and similar) approaches yield[8,9]

- simple latent representation
- improved generalization
- adversarial robustness



taken from [8]



taken from [9]

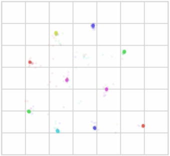
[8] Kolchinsky, Tracey, and Wolpert, "Nonlinear Information Bottleneck", 2019

[9] Alemi et al., "Deep Variational Information Bottleneck", 2017

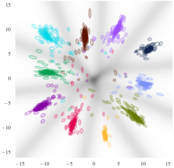
Deep Variational Information Bottleneck

This (and similar) approaches yield[8,9]

- simple latent representation
- improved generalization
- adversarial robustness



taken from [8]



taken from [9]

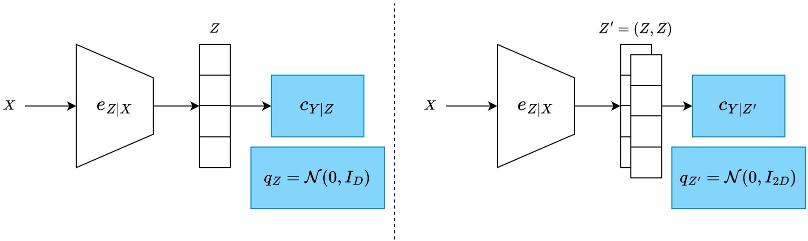
But how much is due to IT?

[8] Kolchinsky, Tracey, and Wolpert, "Nonlinear Information Bottleneck", 2019

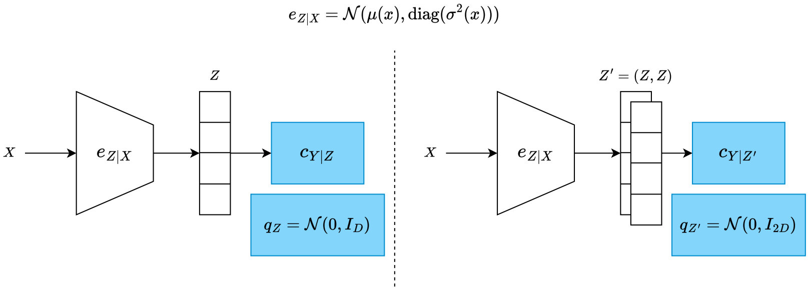
[9] Alemi et al., "Deep Variational Information Bottleneck", 2017

Effect of Latent Dimension

$$e_{Z|X} = \mathcal{N}(\mu(x), \mathrm{diag}(\sigma^2(x)))$$



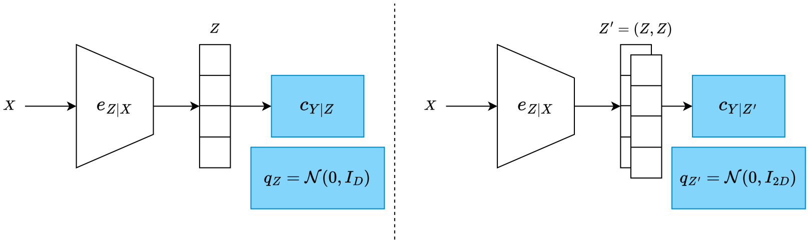
Effect of Latent Dimension



$$I(X;Z) = I(X;Z')$$

Effect of Latent Dimension

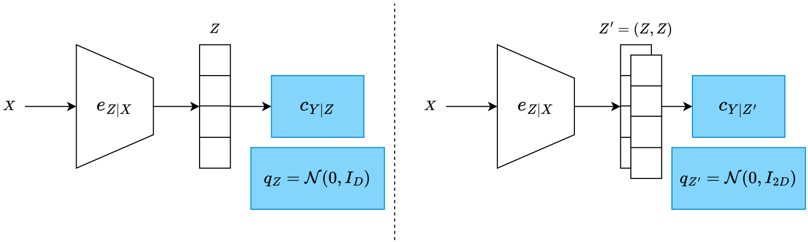
$$e_{Z|X} = \mathcal{N}(\mu(x), \mathrm{diag}(\sigma^2(x)))$$



$$\mathbb{E}\left(D\left(e_{Z'|X}(\cdot|X)\|q_{Z'}(\cdot)\right)\right) = 2 \cdot \mathbb{E}\left(D\left(e_{Z|X}(\cdot|X)\|q_Z(\cdot)\right)\right)$$

Effect of Latent Dimension

$$e_{Z|X} = \mathcal{N}(\mu(x), \mathrm{diag}(\sigma^2(x)))$$



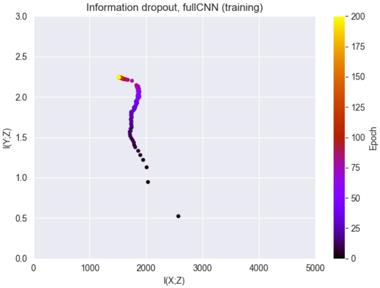
Hyperparameter β must be chosen jointly with latent dimension.

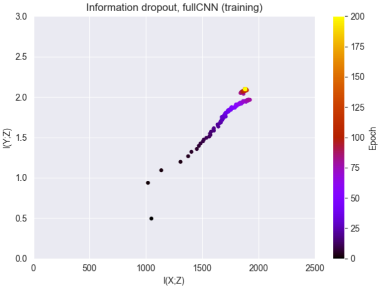
Effect of Latent Dimension (cont'd)

In [the context of the β-VAE] it makes sense to normalise β by latent $\mathbf{z}$ size [...] in order to compare its different values across different latent layer sizes [...] We found that larger latent $\mathbf{z}$ layer sizes require higher constraint pressures (higher β values) [...].[10]

[10] Higgins et al., "β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework", 2017

Effect of Latent Dimension (cont'd)





Fully convolutional NN with only 25% of the filters (right) shows initially (!) lower estimates of the variational bound[11]

[11] Adilova, Geiger, and Fischer, *Information Plane Analysis for Dropout Neural Networks*, 2022

Effect of Variational Marginal

$$I(X;Z) = \min_{q_Z} \mathbb{E}\left(D\left(e_{Z|X}(\cdot|X) \| q_Z(\cdot)\right)\right)$$

Effect of Variational Marginal

Selecting a family $\mathcal{Q}$ (Gaussian, etc.):

$$I(X;Z) \leq \min_{q_Z \in \mathcal{Q}} \mathbb{E}\left(D\left(e_{Z|X}(\cdot|X) \| q_Z(\cdot)\right)\right)$$

Effect of Variational Marginal

Selecting a *factorized* family, i.e., $q_Z = \prod q_{Z_i}$:

$$I(X;Z) \leq \min_{\{q_{Z_i}\}} \mathbb{E}\left(D\left(e_{Z|X}(\cdot|X) \| \prod q_{Z_i}(\cdot)\right)\right)$$

Effect of Variational Marginal

Selecting a *factorized* family, i.e., $q_Z = \prod q_{Z_i}$:

$$I(X;Z) = \min_{\{q_{Z_i}\}} \mathbb{E}\left(D\left(e_{Z|X}(\cdot|X) \| \prod q_{Z_i}(\cdot)\right)\right) - D\left(p_Z \| \prod p_{Z_i}\right)$$

[12] Achille and Soatto, "Information Dropout: Learning Optimal Representations Through Noisy Computation", 2018

Effect of Variational Marginal

Selecting a *factorized* family, i.e., $q_Z = \prod q_{Z_i}$:

$$I(X;Z) = \min_{\{q_{Z_i}\}} \mathbb{E}\left(D\left(e_{Z|X}(\cdot|X) \| \prod q_{Z_i}(\cdot)\right)\right) - D\left(p_Z \| \prod p_{Z_i}\right)$$

Minimizing the variational bound on $I(X;Z)$ *simultaneously* minimizes total correlation of Z (disentanglement)[12]

[12] Achille and Soatto, "Information Dropout: Learning Optimal Representations Through Noisy Computation", 2018

Information Dropout[13]

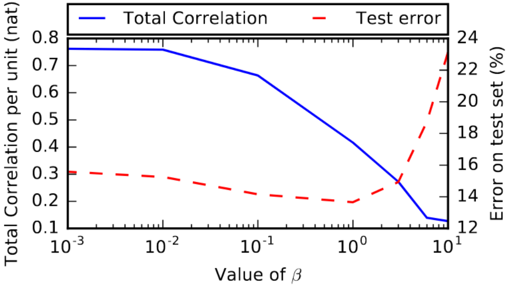


Fig. 5: Plot of the test error and total correlation for different values of β of the final layer of the All-CNN-32 network with Softplus activations trained on CIFAR-10 with 25% of the filters. Increasing β the test error decreases (we prevent

taken from [13]

[13] Achille and Soatto, "Information Dropout: Learning Optimal Representations Through Noisy Computation", 2018

Effect of Equivalent Information-Theoretic Functionals

Since $Y - X - Z$, we have

$$I(X;Z) = I(X,Y;Z) = I(X;Z|Y) + I(Y;Z).$$

Effect of Equivalent Information-Theoretic Functionals

Since $Y - X - Z$, we have

$$I(X;Z) = I(X,Y;Z) = I(X;Z|Y) + I(Y;Z).$$

Thus,

$$\begin{aligned} &I(X;Z) + \beta H(Y|Z) \\ &\qquad \leq \mathbb{E}\left(D\left(e_{Z|X}(\cdot|X) \| q_Z(\cdot)\right)\right) - \beta \mathbb{E}\left(\log c_{Y|Z}(Y|Z)\right) \end{aligned}$$

Effect of Equivalent Information-Theoretic Functionals

Since $Y - X - Z$, we have

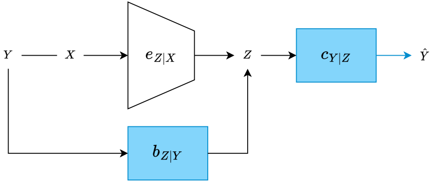
$$I(X;Z) = I(X,Y;Z) = I(X;Z|Y) + I(Y;Z).$$

Thus,

$$\begin{aligned} &I(X;Z|Y) + (\beta-1)H(Y|Z) \\ &\quad \leq \mathbb{E}\left(D\left(e_{Z|X}(\cdot|X)\|b_{Z|Y}(\cdot|Y)\right)\right) - (\beta-1)\mathbb{E}\left(\log c_{Y|Z}(Y|Z)\right) \end{aligned}$$

Conditional Entropy Bottleneck (CEB)[15]

$$I(X;Z|Y) + (\beta - 1)H(Y|Z)$$
$$\leq \mathbb{E}\left(D\left(e_{Z|X}(\cdot|X) \| b_{Z|Y}(\cdot|Y)\right)\right) - (\beta - 1)\mathbb{E}\left(\log c_{Y|Z}(Y|Z)\right)$$



- better accuracy and adversarial robustness than VIB[14]
- ...which purportedly is due to CEB yielding a tighter bound on the information bottleneck functional

[14] Fischer and Alemi, "CEB Improves Model Robustness", 2020

[15] Fischer, "The Conditional Entropy Bottleneck", 2020

Conditional Entropy Bottleneck (cont'd)

Theorem 1. *If VCEB is constrained to a consistent classifier-backward encoder pair, and if* $\mathcal{Q} \supseteq \{q_Z \colon q_Z(z) = \sum_y b_{Z|Y}(z|y) p_Y(y),\ b_{Z|Y} \in \mathcal{B}\}$, *then*

$$\min_{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, q_Z\in\mathcal{Q}} \mathcal{L}_{\text{VIB}} \leq \min_{\substack{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, b_{Z|Y}\in\mathcal{B} \\ (c_{\hat{Y}|Z}, b_{Z|Y}) \text{ consistent}}} \mathcal{L}_{\text{VCEB}}. \tag{13a}$$

If VIB and VCEB are constrained to a consistent classifier–marginal and classifier-backward encoder pair, respectively, and if $\mathcal{B} \supseteq \{b_{Z|Y} \colon b_{Z|Y}(z|y) = c_{\hat{Y}|Z}(y|z) q_Z(z) / p_Y(y),\ q_Z \in \mathcal{Q}, c_{\hat{Y}|Z} \in \mathcal{C}\}$, *then*

$$\min_{\substack{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, q_Z\in\mathcal{Q} \\ (c_{\hat{Y}|Z}, q_Z) \text{ consistent}}} \mathcal{L}_{\text{VIB}} \geq \min_{\substack{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, b_{Z|Y}\in\mathcal{B} \\ (c_{\hat{Y}|Z}, b_{Z|Y}) \text{ consistent}}} \mathcal{L}_{\text{VCEB}}. \tag{13b}$$

A fortiori, (13b) continues to hold if VCEB is not constrained to a consistent classifier-backward encoder pair.

...a fair comparison (network architectures) shows that there cannot be an ordering.[16]

[16] Geiger and Fischer, "A Comparison of Variational Bounds for the Information Bottleneck Functional", 2020

Conditional Entropy Bottleneck (cont'd)

Theorem 1. *If VCEB is constrained to a consistent classifier-backward encoder pair, and if* $\mathcal{Q} \supseteq \{q_Z \colon q_Z(z) = \sum_y b_{Z|Y}(z|y) p_Y(y),\ b_{Z|Y} \in \mathcal{B}\}$, *then*

$$\min_{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, q_Z\in\mathcal{Q}} \mathcal{L}_{\mathrm{VIB}} \leq \min_{\substack{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, b_{Z|Y}\in\mathcal{B} \\ (c_{\hat{Y}|Z}, b_{Z|Y})\ \text{consistent}}} \mathcal{L}_{\mathrm{VCEB}}. \tag{13a}$$

If VIB and VCEB are constrained to a consistent classifier–marginal and classifier-backward encoder pair, respectively, and if $\mathcal{B} \supseteq \{b_{Z|Y} \colon b_{Z|Y}(z|y) = c_{\hat{Y}|Z}(y|z) q_Z(z) / p_Y(y),\ q_Z \in \mathcal{Q}, c_{\hat{Y}|Z} \in \mathcal{C}\}$, *then*

$$\min_{\substack{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, q_Z\in\mathcal{Q} \\ (c_{\hat{Y}|Z}, q_Z)\ \text{consistent}}} \mathcal{L}_{\mathrm{VIB}} \geq \min_{\substack{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, b_{Z|Y}\in\mathcal{B} \\ (c_{\hat{Y}|Z}, b_{Z|Y})\ \text{consistent}}} \mathcal{L}_{\mathrm{VCEB}}. \tag{13b}$$

A fortiori, (13b) continues to hold if VCEB is not constrained to a consistent classifier-backward encoder pair.

...a fair comparison (network architectures) shows that there cannot be an ordering.[16] Then why is CEB better than VIB?

[16] Geiger and Fischer, "A Comparison of Variational Bounds for the Information Bottleneck Functional", 2020

Conditional Entropy Bottleneck (cont'd)

Selecting a *factorized* family, i.e., $b_{Z|Y} = \prod b_{Z_i|Y}$:

$$I(X;Z|Y) = \min_{\{b_{Z_i|Y}\}} \mathbb{E}\left(D\left(e_{Z|X}(\cdot|X) \| \prod b_{Z_i|Y}(\cdot)\right)\right) - \mathbb{E}\left(D\left(p_{Z|Y} \| \prod p_{Z_i|Y}\right)\right)$$

[17] Amjad and Geiger, *Class-Conditional Compression and Disentanglement: Bridging the Gap between Neural Networks and Naive Bayes Classifiers*, 2019

Conditional Entropy Bottleneck (cont'd)

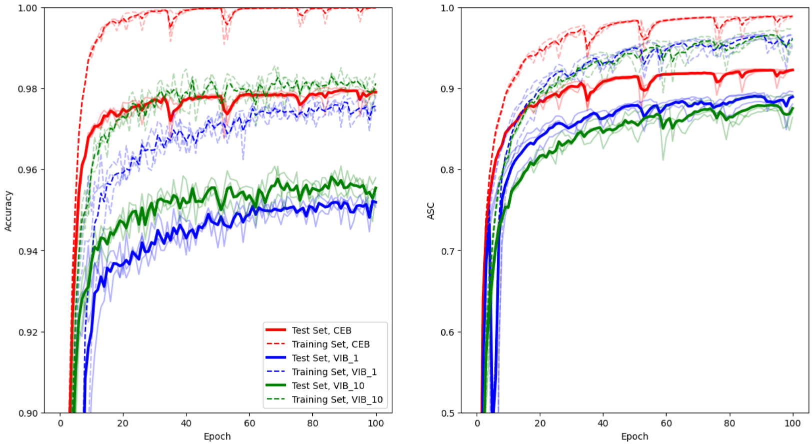
Selecting a *factorized* family, i.e., $b_{Z|Y} = \prod b_{Z_i|Y}$:

$$I(X;Z|Y) = \min_{\{b_{Z_i|Y}\}} \mathbb{E}\left(D\left(e_{Z|X}(\cdot|X) \| \prod b_{Z_i|Y}(\cdot)\right)\right) - \mathbb{E}\left(D\left(p_{Z|Y} \| \prod p_{Z_i|Y}\right)\right)$$

Minimizing the variational bound on $I(X;Z|Y)$ *simultaneously* minimizes conditional total correlation of Z (conditional disentanglement)[17]

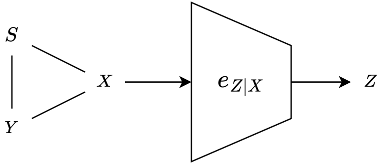
[17] Amjad and Geiger, *Class-Conditional Compression and Disentanglement: Bridging the Gap between Neural Networks and Naive Bayes Classifiers*, 2019

Conditional Entropy Bottleneck (cont'd)



16-dimensional latent space, $\beta = 5$

Invariant Representation Learning



$$\min_{e_{Z|X}} I(S;Z) - \alpha I(X;Z) - \beta I(Y;Z)$$

- CPFSI[18]
- privacy funnel[19]

$$\min_{e_{Z|X}} I(S;Z) + \alpha I(X;Z) - \beta I(Y;Z)$$

- fair bottleneck[19]
- CLUB[20]
- IBSI[21]

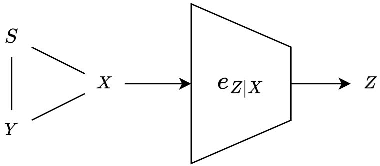
[18] Freitas and Geiger, *FUNCK: Information Funnels and Bottlenecks for Invariant Representation Learning*, 2022

[19] Rodríguez-Gálvez, Thobaben, and Skoglund, "A Variational Approach to Privacy and Fairness", 2021

[20] Razeghi et al., *Bottlenecks CLUB: Unifying Information-Theoretic Trade-offs Among Complexity, Leakage, and Utility*, 2022

[21] Moyer et al., "Invariant Representations without Adversarial Training", 2018

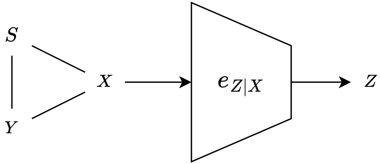
Effect of Equivalent Variational Terms



$$(S, Y) - X - Z$$

$$\min_{e_{Z|X}} I(S;Z) + \alpha I(X;Z) - \beta I(Y;Z) \qquad \min_{e_{Z|X}} I(S;Z) - \alpha I(X;Z) - \beta I(Y;Z)$$

Effect of Equivalent Variational Terms

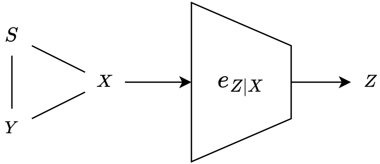


$$(S, Y) - X - Z$$

$$\min_{e_{Z|X}} (1+\alpha) I(X;Z) - I(X;Z|S) - \beta I(Y;Z)$$

$$\min_{e_{Z|X}} (1-\alpha) I(X;Z) - I(X;Z|S) - \beta I(Y;Z)$$

Effect of Equivalent Variational Terms



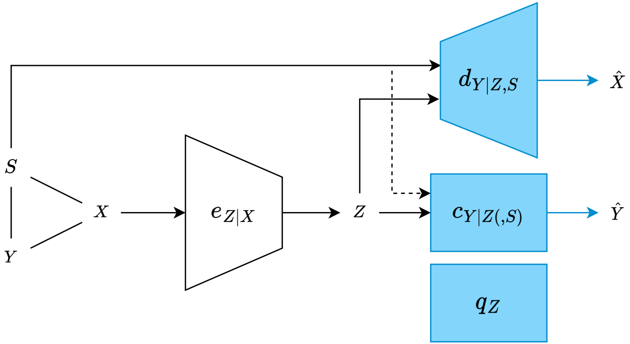
$$(S, Y) - X - Z$$

$$\min_{e_{Z|X}} (1+\alpha) I(X;Z) - I(X;Z|S) - \beta I(Y;Z)$$

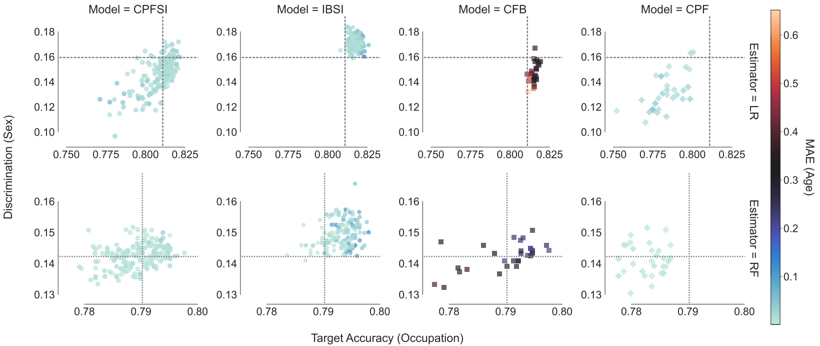
$$\min_{e_{Z|X}} (1-\alpha) I(X;Z) - I(X;Z|S) - \beta I(Y;Z)$$

The mutual information term for reconstruction is always maximized!

Invariant Represenation Learning (cont'd)



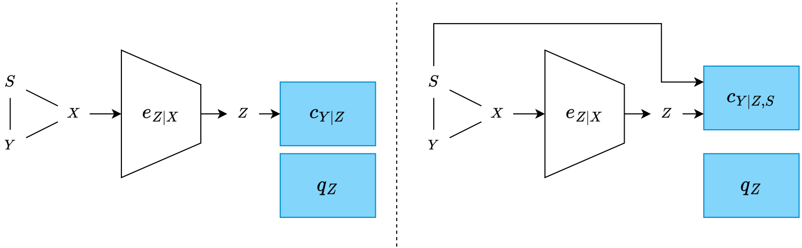
Invariant Represenation Learning (cont'd)



Representation learning (32-dimensinal) on the Dutch dataset, different trade-off parameters[22]

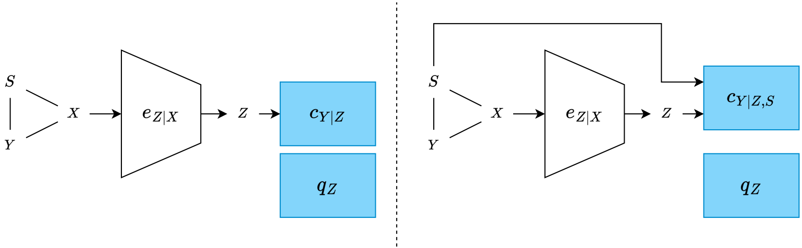
[22]Freitas and Geiger, *FUNCK: Information Funnels and Bottlenecks for Invariant Representation Learning*, 2022

To Condition or Not To Condition?



$$\min_{e_{Z|X}\in\mathcal{E}} I(S;Z),\ \mathcal{E} \text{ s.t. } H(Y|Z)\leq\varepsilon \qquad \min_{e_{Z|X}\in\mathcal{E}'} I(S;Z),\ \mathcal{E}' \text{ s.t. } H(Y|Z,S)\leq\varepsilon$$

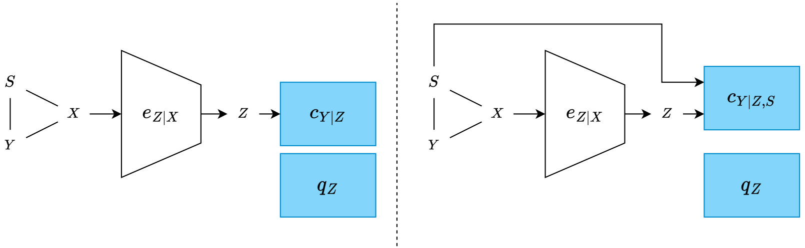
To Condition or Not To Condition?



$$\min_{e_{Z|X}\in\mathcal{E}} I(S;Z),\ \mathcal{E} \text{ s.t. } H(Y|Z)\leq\varepsilon \qquad \min_{e_{Z|X}\in\mathcal{E}'} I(S;Z),\ \mathcal{E}' \text{ s.t. } H(Y|Z,S)\leq\varepsilon$$

$$\mathcal{E}\subseteq\mathcal{E}'$$

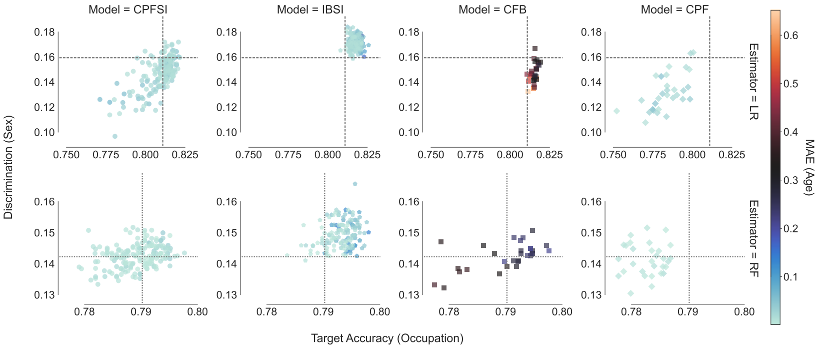
To Condition or Not To Condition?



$$\min_{e_{Z|X}\in\mathcal{E}} I(S;Z),\ \mathcal{E} \text{ s.t. } H(Y|Z)\le\varepsilon \qquad \min_{e_{Z|X}\in\mathcal{E}'} I(S;Z),\ \mathcal{E}' \text{ s.t. } H(Y|Z,S)\le\varepsilon$$

$$H(Y|Z,S)\le H(Y|Z)\le H(Y|Z,S)+H(S)$$

(No?) Effect of Conditioning



Why does CFB perform so well?

Conclusions

- Information-theoretic objectives are *just one* of many interdependent ingredients

Conclusions

- Information-theoretic objectives are *just one* of many interdependent ingredients
  - architecture choices (latent dimension size, etc.)

Conclusions

- Information-theoretic objectives are *just one* of many interdependent ingredients
  - architecture choices (latent dimension size, etc.)
  - choice of variational approach/bound

Conclusions

- Information-theoretic objectives are *just one* of many interdependent ingredients
  - architecture choices (latent dimension size, etc.)
  - choice of variational approach/bound
  - modelling choices (factorized distributions, conditioning, etc.)

Conclusions

- Information-theoretic objectives are *just one* of many interdependent ingredients
  - architecture choices (latent dimension size, etc.)
  - choice of variational approach/bound
  - modelling choices (factorized distributions, conditioning, etc.)
  - choice of the optimization method

Conclusions

- Information-theoretic objectives are *just one* of many interdependent ingredients
  - architecture choices (latent dimension size, etc.)
  - choice of variational approach/bound
  - modelling choices (factorized distributions, conditioning, etc.)
  - choice of the optimization method
- that can reinforce or even negate the chosen objective.

Conclusions

- Information-theoretic objectives are *just one* of many interdependent ingredients
  - architecture choices (latent dimension size, etc.)
  - choice of variational approach/bound
  - modelling choices (factorized distributions, conditioning, etc.)
  - choice of the optimization method
- that can reinforce or even negate the chosen objective.
- To what extent can the operational goals (compression?, invariance, etc.) be captured by IT cost functions?

Conclusions

- Information-theoretic objectives are *just one* of many interdependent ingredients
  - architecture choices (latent dimension size, etc.)
  - choice of variational approach/bound
  - modelling choices (factorized distributions, conditioning, etc.)
  - choice of the optimization method
- that can reinforce or even negate the chosen objective.
- To what extent can the operational goals (compression?, invariance, etc.) be captured by IT cost functions?

Thanks!