Learning Robust Representations via Multi-View Information Bottleneck

Marco Federici[1], Anjan Dutta[2], Patrick Forré[1], Nate Kushman[3], Zeynep Akata[4]

[1]University of Amsterdam
[2]University of Exeter
[3]Microsoft Research Cambridge
[4]University of Tübingen

May 4, 2020

Table of Contents

Which Representation?

Let $\mathbf{z}$ be a representation of $\mathbf{x}$

Which Representation?

Let $\mathbf{z}$ be a representation of $\mathbf{x}$

- Which $\mathbf{z}$ is useful?

Which Representation?

Let $\mathbf{z}$ be a representation of $\mathbf{x}$

- Which $\mathbf{z}$ is useful?
  - Disentangled

Which Representation?

Let $\mathbf{z}$ be a representation of $\mathbf{x}$

- Which $\mathbf{z}$ is useful?
  - Disentangled
  - Compositional

Which Representation?

Let $\mathbf{z}$ be a representation of $\mathbf{x}$

- Which $\mathbf{z}$ is useful?
  - Disentangled
  - Compositional
  - Abstract

Which Representation?

Let $\mathbf{z}$ be a representation of $\mathbf{x}$

- Which $\mathbf{z}$ is useful?
  - Disentangled
  - Compositional
  - Abstract
  - "Human"

Example

$\mathbf{x}$

Example

x

$\Longrightarrow$

z

"cat" laying on "laptop"

Example

x

$\Longrightarrow$

z

"cat" laying on "laptop"

Task:

Example

x

$\Longrightarrow$

z

"cat" laying on "laptop"

Task:

- Is there a chair?

Example

x **z**

$\Longrightarrow$

"cat" laying on "laptop"

Task:

- Is there a chair?
- How many black pixels are in the picture?

Which Representation?

Which **z** is useful

Which Representation?

Which $\mathbf{z}$ is useful for predicting $\mathbf{y}$?

Which Representation?

Which $\mathbf{z}$ is useful for predicting $\mathbf{y}$?

- $\mathbf{z}$ contains **all** the information regarding $\mathbf{y}$

Which Representation?

Which $\mathbf{z}$ is useful for predicting $\mathbf{y}$?

- $\mathbf{z}$ contains **all** the information regarding $\mathbf{y}$
- $\mathbf{z}$ contains **only** information regarding $\mathbf{y}$

Sufficiency

Sufficiency

Definition

A representation $\mathbf{z}$ of $\mathbf{x}$ is **sufficient** for $\mathbf{y}$ if and only if $I(\mathbf{x};\mathbf{y}) = I(\mathbf{z};\mathbf{y})$

- $\mathbf{z}$ contains **all** the information about $\mathbf{y}$

Minimality

Which representation contains the least information?

Minimality

Which representation contains the least information?

$I(\mathbf{z}; \mathbf{x})$

Minimality

Which representation contains the least information?

$$I(\mathbf{z};\mathbf{x}) = I(\mathbf{z};\mathbf{y}) + I(\mathbf{x};\mathbf{z}|\mathbf{y})$$

Minimality

Which representation contains the least information?

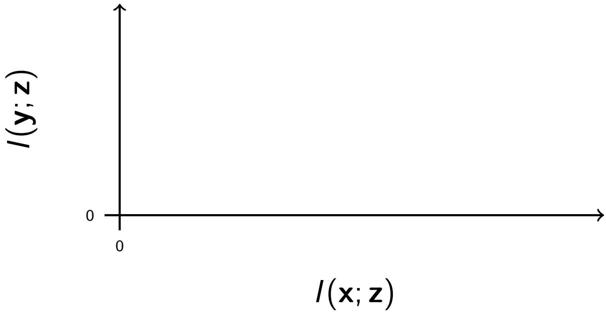
$$I(\mathbf{z};\mathbf{x}) = \underbrace{I(\mathbf{z};\mathbf{y})}_{\substack{\text{predictive}\\\text{information in }\mathbf{z}}} + I(\mathbf{x};\mathbf{z}|\mathbf{y})$$

Minimality

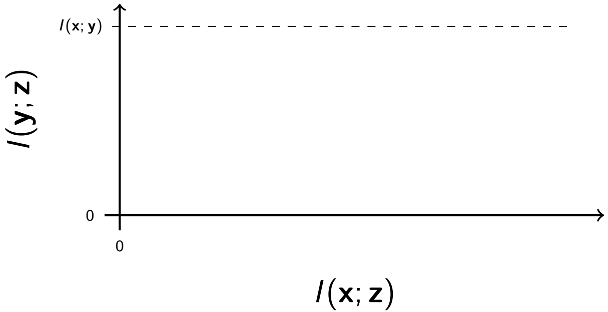
Which representation contains the least information?

$$I(\mathbf{z};\mathbf{x}) = \underbrace{I(\mathbf{z};\mathbf{y})}_{\substack{\text{predictive} \\ \text{information in } \mathbf{z}}} + \underbrace{I(\mathbf{x};\mathbf{z}|\mathbf{y})}_{\substack{\text{superfluous} \\ \text{information}}}$$

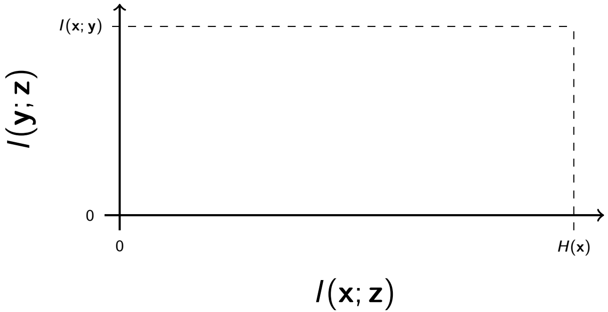
Information Plane



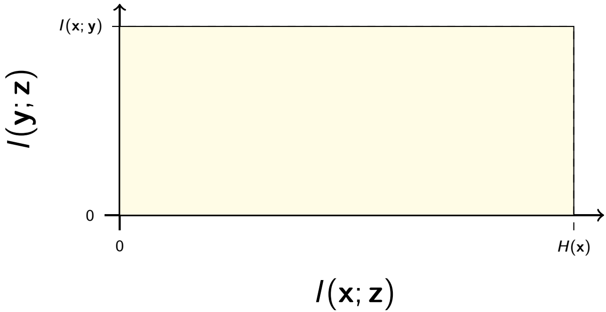
Information Plane



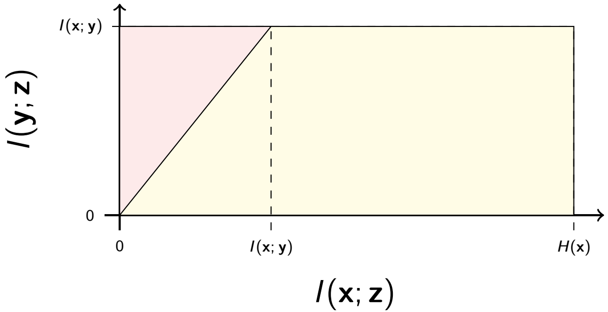
Information Plane



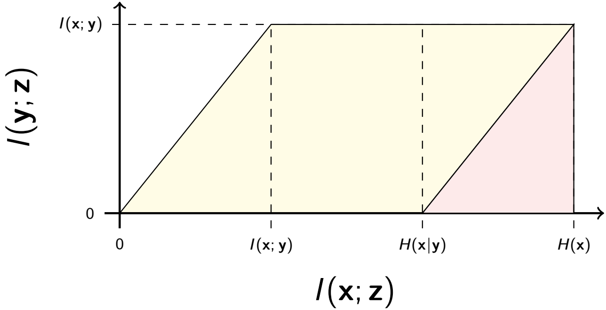
Information Plane



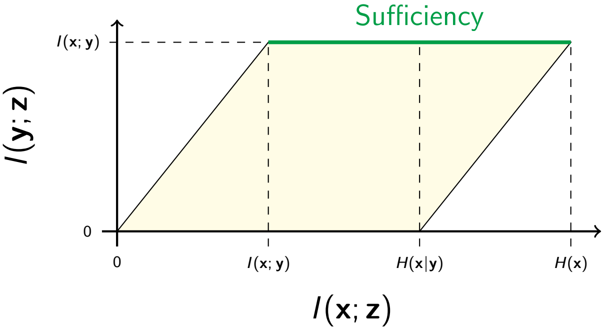
Information Plane



Information Plane

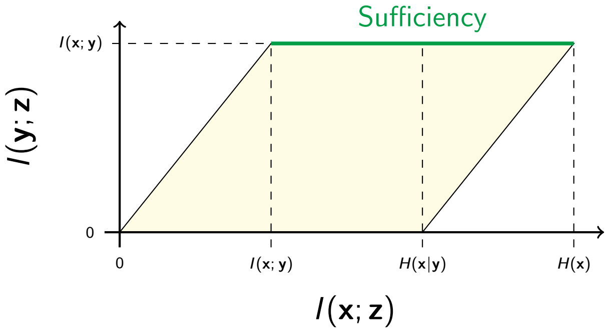


Information Plane

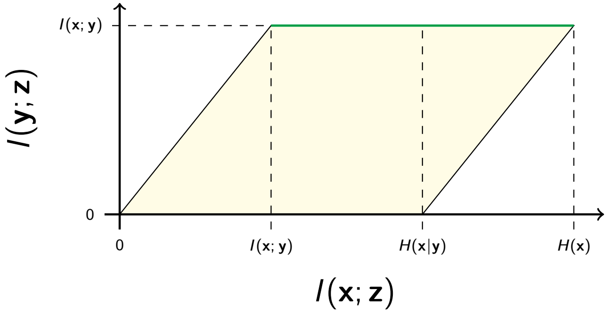


- $I(\mathbf{y};\mathbf{z}) = I(\mathbf{x};\mathbf{y})$

Information Plane

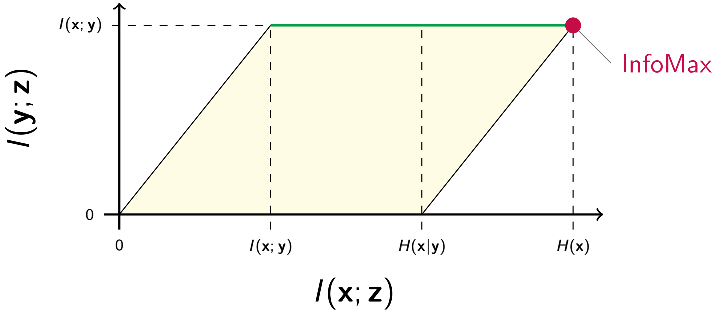


Information Plane



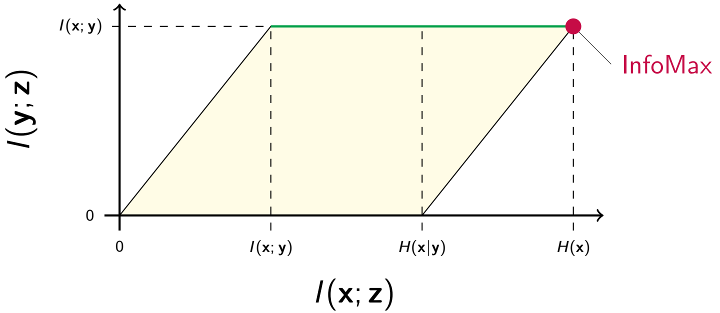
- InfoMax?

Information Plane



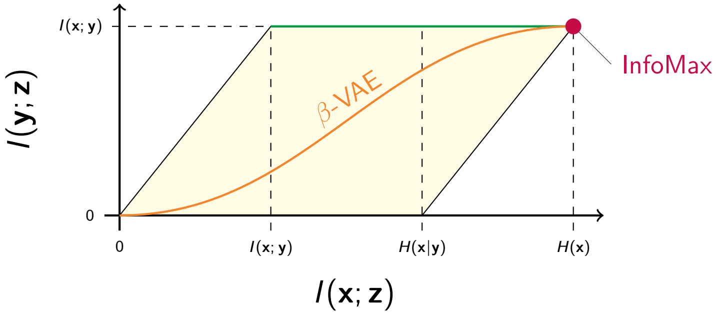
- InfoMax?

Information Plane



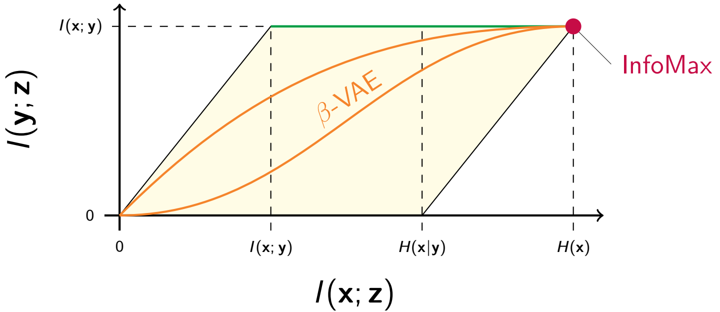
- β-VAE?

Information Plane



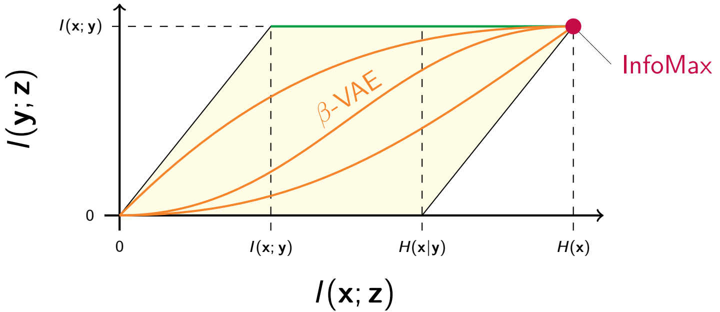
- β-VAE?

Information Plane



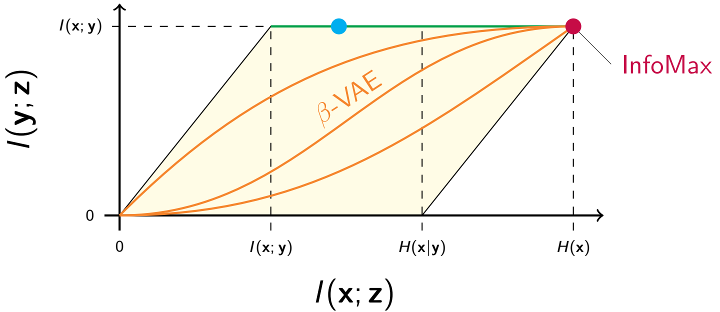
- β-VAE?

Information Plane



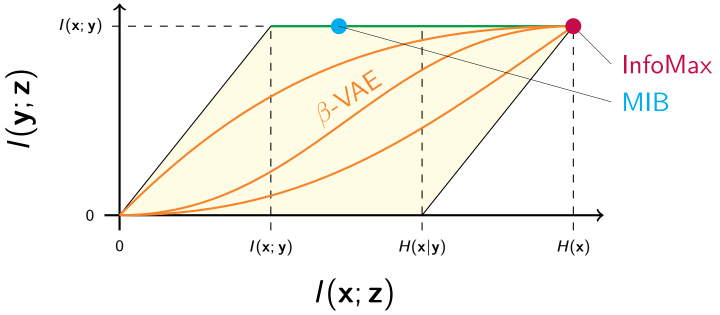
- Can we do better without label supervision?

Information Plane



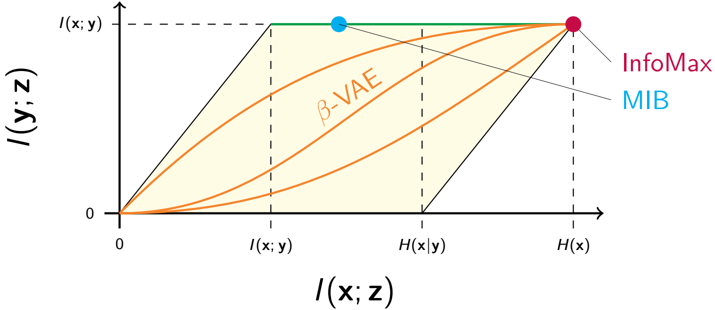
- Can we do better without label supervision?

Information Plane



- Can we do better without label supervision?
  Yes!

Information Plane



- Can we do better without label supervision?
  Yes! By exploiting redundant information and properties of the task.

Multi-View Learning

$\mathbf{v}_1$

Multi-View Learning

$\mathbf{v}_1$

$\mathbf{v}_2$

Multi-View Learning

$\mathbf{v}_1$

$\mathbf{v}_2$

Common

Not Common

Multi-View Learning

$\mathbf{v}_1$

$\mathbf{v}_2$

Common

- Pointy Ears
- Paws
- Fur

Not Common

Multi-View Learning

$\mathbf{v}_1$

$\mathbf{v}_2$

Common

- Pointy Ears
- Paws
- Fur

Not Common

- Laptop
- Table
- Sofa

Redundancy

Redundancy

Definition

A view $\mathbf{v}_1$ is **redundant** with respect to $\mathbf{v}_2$ for $\mathbf{y}$ if and only if $I(\mathbf{v}_1; \mathbf{y} | \mathbf{v}_2) = 0$

Redundancy

Definition

A view $\mathbf{v}_1$ is **redundant** with respect to $\mathbf{v}_2$ for $\mathbf{y}$ if and only if $I(\mathbf{v}_1; \mathbf{y}|\mathbf{v}_2) = 0$

- $I(\mathbf{v}_1; \mathbf{y}|\mathbf{v}_2) = 0 \iff p(\mathbf{y}|\mathbf{v}_2) = p(\mathbf{y}|\mathbf{v}_1, \mathbf{v}_2)$

Redundancy

Definition

A view $\mathbf{v}_1$ is **redundant** with respect to $\mathbf{v}_2$ for $\mathbf{y}$ if and only if $I(\mathbf{v}_1; \mathbf{y}|\mathbf{v}_2) = 0$

- $I(\mathbf{v}_1; \mathbf{y}|\mathbf{v}_2) = 0 \iff p(\mathbf{y}|\mathbf{v}_2) = p(\mathbf{y}|\mathbf{v}_1, \mathbf{v}_2)$

Example

Redundancy

Definition

A view $\mathbf{v}_1$ is **redundant** with respect to $\mathbf{v}_2$ for $\mathbf{y}$ if and only if $I(\mathbf{v}_1; \mathbf{y} | \mathbf{v}_2) = 0$

- $I(\mathbf{v}_1; \mathbf{y} | \mathbf{v}_2) = 0 \iff p(\mathbf{y}|\mathbf{v}_2) = p(\mathbf{y}|\mathbf{v}_1, \mathbf{v}_2)$

Example

- $\mathbf{v}_1$: abstract of a paper

Redundancy

Definition

A view $\mathbf{v}_1$ is **redundant** with respect to $\mathbf{v}_2$ for $\mathbf{y}$ if and only if $I(\mathbf{v}_1; \mathbf{y}|\mathbf{v}_2) = 0$

- $I(\mathbf{v}_1; \mathbf{y}|\mathbf{v}_2) = 0 \iff p(\mathbf{y}|\mathbf{v}_2) = p(\mathbf{y}|\mathbf{v}_1, \mathbf{v}_2)$

Example

- $\mathbf{v}_1$: abstract of a paper
- $\mathbf{v}_2$: full text of the same paper

Redundancy

Definition

A view $\mathbf{v}_1$ is **redundant** with respect to $\mathbf{v}_2$ for $\mathbf{y}$ if and only if $I(\mathbf{v}_1; \mathbf{y}|\mathbf{v}_2) = 0$

- $I(\mathbf{v}_1; \mathbf{y}|\mathbf{v}_2) = 0 \iff p(\mathbf{y}|\mathbf{v}_2) = p(\mathbf{y}|\mathbf{v}_1, \mathbf{v}_2)$

Example

- $\mathbf{v}_1$: abstract of a paper
- $\mathbf{v}_2$: full text of the same paper
- $\mathbf{y}$: score assigned by reviewer 3

Mutual Redundancy

Mutual Redundancy

Definition

Two views $\mathbf{v}_1$ and $\mathbf{v}_2$ are **mutually redundant** for $\mathbf{y}$ if and only if $\mathbf{v}_1$ is redundant with respect to $\mathbf{v}_2$ for $\mathbf{y}$ and vice-versa

Mutual Redundancy

Definition

Two views $\mathbf{v}_1$ and $\mathbf{v}_2$ are **mutually redundant** for $\mathbf{y}$ if and only if $\mathbf{v}_1$ is redundant with respect to $\mathbf{v}_2$ for $\mathbf{y}$ and vice-versa

Example

- $\mathbf{v}_1$: sentence in French

Mutual Redundancy

Definition

Two views $\mathbf{v}_1$ and $\mathbf{v}_2$ are **mutually redundant** for $\mathbf{y}$ if and only if $\mathbf{v}_1$ is redundant with respect to $\mathbf{v}_2$ for $\mathbf{y}$ and vice-versa

Example

- $\mathbf{v}_1$: sentence in French
- $\mathbf{v}_2$: same sentence in English

Mutual Redundancy

Definition

Two views $\mathbf{v}_1$ and $\mathbf{v}_2$ are **mutually redundant** for $\mathbf{y}$ if and only if $\mathbf{v}_1$ is redundant with respect to $\mathbf{v}_2$ for $\mathbf{y}$ and vice-versa

Example

- $\mathbf{v}_1$: sentence in French
- $\mathbf{v}_2$: same sentence in English
- $\mathbf{y}$: semantics

Mutual Redundancy and Sufficiency

Mutual Redundancy and Sufficiency

Theorem

Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be two mutually redundant views for a target $\mathbf{y}$ and let $\mathbf{z}_1$ be a representation of $\mathbf{v}_1$. If $\mathbf{z}_1$ is sufficient for $\mathbf{v}_2$ ($I(\mathbf{v}_1; \mathbf{v}_2) = I(\mathbf{z}_1; \mathbf{v}_2)$) then $\mathbf{z}_1$ is as predictive for $\mathbf{y}$ as the joint observation of the two views ($I(\mathbf{v}_1\mathbf{v}_2; \mathbf{y}) = I(\mathbf{y}; \mathbf{z}_1)$).

Multi-View Minimality

Multi-View Minimality

$I(\mathbf{z}_1; \mathbf{v}_1)$

Multi-View Minimality

$$I(\mathbf{z}_1;\mathbf{v}_1) = \underbrace{I(\mathbf{z}_1;\mathbf{v}_2)}_{\substack{\text{predictive}\\ \text{information for } \mathbf{v}_2}} + \underbrace{I(\mathbf{v}_1;\mathbf{z}_1|\mathbf{v}_2)}_{\substack{\text{superfluous}\\ \text{information for } \mathbf{v}_2}}$$

Multi-View Minimality

$$I(\mathbf{z}_1;\mathbf{v}_1) = \underbrace{I(\mathbf{z}_1;\mathbf{v}_2)}_{\substack{\text{predictive}\\ \text{information for } \mathbf{v}_2}} + \underbrace{I(\mathbf{v}_1;\mathbf{z}_1|\mathbf{v}_2)}_{\substack{\text{superfluous}\\ \text{information for } \mathbf{v}_2}}$$

Loss function for a parametric encoder $p_\theta(\mathbf{z}_1|\mathbf{v}_1)$:

Multi-View Minimality

$$I(\mathbf{z}_1;\mathbf{v}_1) = \underbrace{I(\mathbf{z}_1;\mathbf{v}_2)}_{\substack{\text{predictive}\\ \text{information for } \mathbf{v}_2}} + \underbrace{I(\mathbf{v}_1;\mathbf{z}_1|\mathbf{v}_2)}_{\substack{\text{superfluous}\\ \text{information for } \mathbf{v}_2}}$$

Loss function for a parametric encoder $p_\theta(\mathbf{z}_1|\mathbf{v}_1)$:

$$\mathcal{L}(\theta;\lambda) = \underbrace{I_\theta(\mathbf{v}_1;\mathbf{z}_1|\mathbf{v}_2)}_{\text{minimality of } \mathbf{z}_1 \text{ for } \mathbf{v}_2} - \lambda \underbrace{I_\theta(\mathbf{v}_2;\mathbf{z}_1)}_{\text{sufficiency of } \mathbf{z}_1 \text{ for } \mathbf{v}_2}$$

MIB Loss function

$$\mathcal{L}_1(\theta;\lambda_1) = I_\theta(\mathbf{v}_1;\mathbf{z}_1|\mathbf{v}_2) - \lambda_1 \; I_\theta(\mathbf{v}_2;\mathbf{z}_1)$$
$$\mathcal{L}_2(\psi;\lambda_2) = I_\psi(\mathbf{v}_2;\mathbf{z}_2|\mathbf{v}_1) - \lambda_2 \; I_\theta(\mathbf{v}_1;\mathbf{z}_2)$$

MIB Loss function

$$\mathcal{L}_1(\theta;\lambda_1) = I_\theta(\mathbf{v}_1;\mathbf{z}_1|\mathbf{v}_2) - \lambda_1\ I_\theta(\mathbf{v}_2;\mathbf{z}_1)$$
$$\mathcal{L}_2(\psi;\lambda_2) = I_\psi(\mathbf{v}_2;\mathbf{z}_2|\mathbf{v}_1) - \lambda_2\ I_\theta(\mathbf{v}_1;\mathbf{z}_2)$$

$$\frac{1}{2}\mathcal{L}_1(\theta;\lambda_1) + \frac{1}{2}\mathcal{L}_2(\psi;\lambda_2) \leq$$

MIB Loss function

$$\mathcal{L}_1(\theta;\lambda_1) = I_\theta(\mathbf{v}_1;\mathbf{z}_1|\mathbf{v}_2) - \lambda_1\ I_\theta(\mathbf{v}_2;\mathbf{z}_1)$$
$$\mathcal{L}_2(\psi;\lambda_2) = I_\psi(\mathbf{v}_2;\mathbf{z}_2|\mathbf{v}_1) - \lambda_2\ I_\theta(\mathbf{v}_1;\mathbf{z}_2)$$

$$\frac{1}{2}\mathcal{L}_1(\theta;\lambda_1) + \frac{1}{2}\mathcal{L}_2(\psi;\lambda_2) \le$$
$$- I_{\theta\psi}(\mathbf{z}_1;\mathbf{z}_2) + \beta\ D_{SKL}(p_\theta(\mathbf{z}_1|\mathbf{v}_1)||p_\psi(\mathbf{z}_2|\mathbf{v}_2))$$

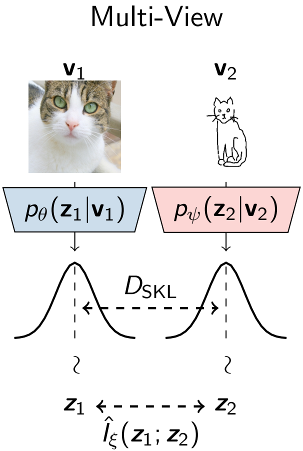
MIB Loss function

$$\mathcal{L}_1(\theta;\lambda_1) = I_\theta(\mathbf{v}_1;\mathbf{z}_1|\mathbf{v}_2) - \lambda_1\ I_\theta(\mathbf{v}_2;\mathbf{z}_1)$$
$$\mathcal{L}_2(\psi;\lambda_2) = I_\psi(\mathbf{v}_2;\mathbf{z}_2|\mathbf{v}_1) - \lambda_2\ I_\theta(\mathbf{v}_1;\mathbf{z}_2)$$

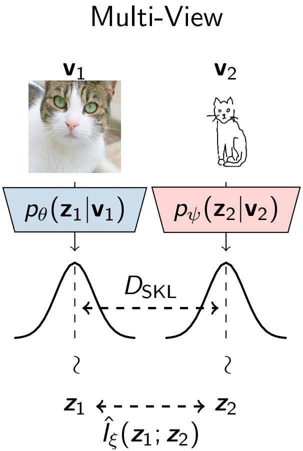
$$\mathcal{L}_{\text{MIB}}(\theta,\psi;\beta) :=\ - I_{\theta\psi}(\mathbf{z}_1;\mathbf{z}_2) + \beta\ D_{SKL}(p_\theta(\mathbf{z}_1|\mathbf{v}_1)||p_\psi(\mathbf{z}_2|\mathbf{v}_2))$$

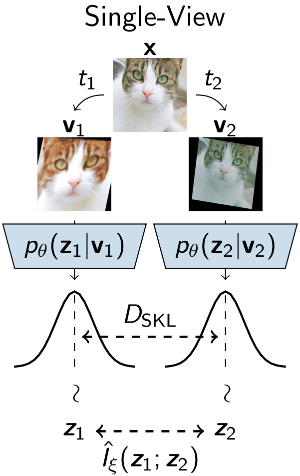
Model

Model


Multi-View
$\mathbf{v}_1$
$\mathbf{v}_2$
$p_\theta(\mathbf{z}_1|\mathbf{v}_1)$
$p_\psi(\mathbf{z}_2|\mathbf{v}_2)$
D_{SKL}
z_1
z_2
$\hat{I}_\xi(z_1; z_2)$

Model


Multi-View
$\mathbf{v}_1$
$\mathbf{v}_2$
$p_\theta(\mathbf{z}_1|\mathbf{v}_1)$
$p_\psi(\mathbf{z}_2|\mathbf{v}_2)$
D_{SKL}
z_1
z_2
$\hat{I}_\xi(z_1; z_2)$


Single-View
$\mathbf{x}$
t_1
t_2
$\mathbf{v}_1$
$\mathbf{v}_2$
$p_\theta(\mathbf{z}_1|\mathbf{v}_1)$
$p_\theta(\mathbf{z}_2|\mathbf{v}_2)$
D_{SKL}
z_1
z_2
$\hat{I}_\xi(z_1; z_2)$

Training

Algorithm 1: Sampling

if *Multi-View* **then**

  $\{(\boldsymbol{v}_1^{(i)}, \boldsymbol{v}_2^{(i)})\}_{i=1}^{B} \sim p(\mathbf{v}_1, \mathbf{v}_2)$;

else

  $\{\boldsymbol{x}^{(i)}\}_{i=1}^{B} \sim p(\mathbf{x})$;

  $\{(t_1^{(i)}, t_2^{(i)})\}_{i=1}^{B} \sim p^2(\mathbf{t})$;

  for $i \leftarrow 1$ **to** B **do**

    $\boldsymbol{v}_1^{(i)} \leftarrow t_1^{(i)}(\boldsymbol{x}^{(i)})$;

    $\boldsymbol{v}_2^{(i)} \leftarrow t_2^{(i)}(\boldsymbol{x}^{(i)})$;

  end for

end if

Training

Algorithm 2: $\mathcal{L}_{\text{MIB}}(\theta, \psi; \beta, B)$

for $i \leftarrow 1$ **to** B **do**

$\quad \boldsymbol{z}_1^{(i)} \sim p_\theta(\mathbf{z}_1 | \boldsymbol{v}_1^{(i)})$;

$\quad \boldsymbol{z}_2^{(i)} \sim p_\psi(\mathbf{z}_2 | \boldsymbol{v}_2^{(i)})$;

$\quad \mathcal{L}_m^{(i)} \leftarrow D_{SKL}(p_\theta(\mathbf{z}_1 | \boldsymbol{v}_1^{(i)}) || p_\psi(\mathbf{z}_2 | \boldsymbol{v}_2^{(i)}))$;

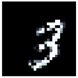
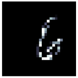
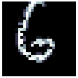
end for

$\mathcal{L}_s \leftarrow -\hat{I}_\xi(\{(\boldsymbol{z}_1^{(i)}, \boldsymbol{z}_2^{(i)})\}_{i=1}^B)$;

return $\mathcal{L}_s + \frac{\beta}{B} \sum_{i=1}^B \mathcal{L}_M^{(i)}$

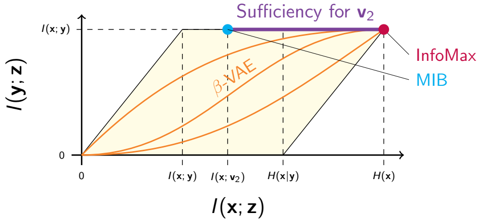
MNIST: Task

MNIST: Task

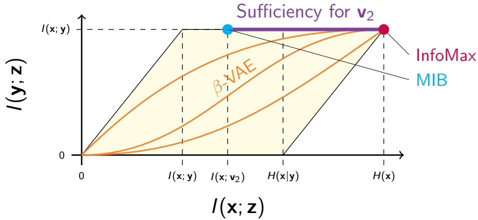
$\mathbf{v}_1 \in [0, 1]^{28\times28}$	$\mathbf{v}_2 \in [0, 1]^{28\times28}$	$\mathbf{y} \in [10]$
		"3"
		"6"

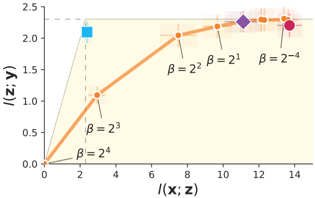
- **Task**: Create a representation $\mathbf{z}_1$ of $\mathbf{v}_1$ to predict $\mathbf{y}$

MNIST: Information Plane

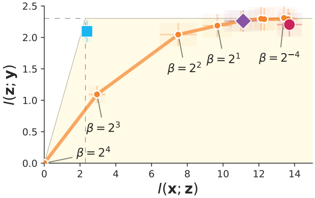


MNIST: Information Plane

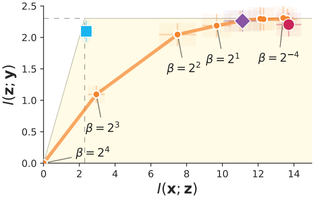


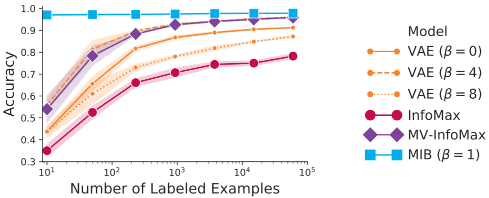


MNIST Results



MNIST Results





MNIST Representation

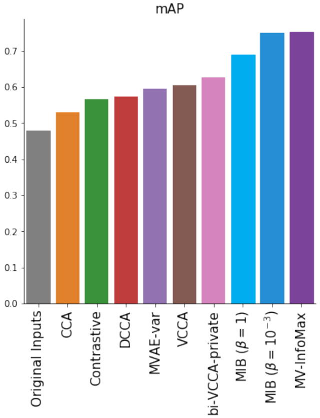
MIR Flick: Task

MIR Flick: Task

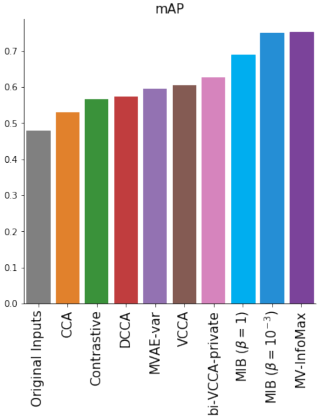
$\mathbf{v}_1 \in \mathbb{R}^{3857}$	$\mathbf{v}_2 \in \{0,1\}^{2000}$	$\mathbf{y} \in \{0,1\}^{38}$
	"watermelon", "hilarious", "chihuahua", "dog"	"animals", "dog", "food"
	"colors", "cores", "centro", "comercial", "building"	"clouds", "sky", "structures"

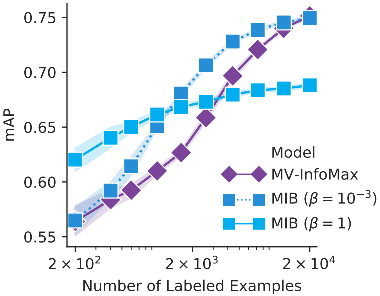
- **Task**: Learn a representation $\mathbf{z}_1$ which is useful to predict $\mathbf{y}$

MIR Flickr: Results



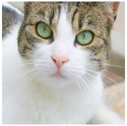
MIR Flickr: Results





Sketchy: Task

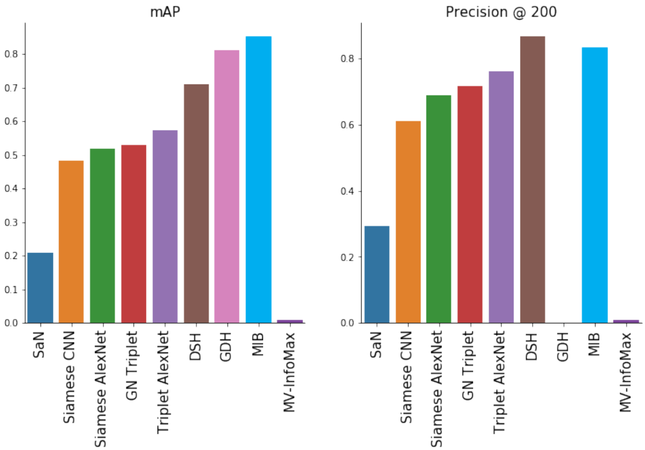
Sketchy: Task

$\mathbf{v}_1 \in \mathbb{R}^{4096}$	$\mathbf{v}_2 \in \mathbb{R}^{4096}$	$\mathbf{y} \in [125]$
		"cat"
		"apple"

- **Task**: Retrieve images of the same class as the query sketch

Sketchy: Results

Sketchy: Results



Discussion and Future Work

Future direction of exploration include:

- Extending MIB to more than 2 views

Discussion and Future Work

Future direction of exploration include:

- Extending MIB to more than 2 views
- Connecting mutual redundancy with invariant neural networks

Thank you for your attention