LEARNING ROBUST REPRESENTATIONS VIA MULTI-VIEW INFORMATION BOTTLENECK

Anonymous authors
Paper under double-blind review

ABSTRACT

The information bottleneck method of Tishby et al. (2000) provides an information-theoretic view of representation learning. The original formulation, however, can only be applied in the supervised setting where task-specific labels are available at learning time. We extend this method to the unsupervised setting, by taking advantage of multi-view data, which provides two views of the same underlying entity. A theoretical analysis leads to the definition of a new multi-view model which produces state-of-the-art results on two standard multi-view datasets, Sketchy and MIR-Flickr. We also extend our theory to the single-view setting by taking advantage of standard data augmentation techniques, empirically showing better generalization capabilities when compared to traditional unsupervised approaches.

1 INTRODUCTION

The goal of deep representation learning (LeCun et al., 2015) is to transform a raw observational input, $\mathbf{x}$, into a representation, $\mathbf{z}$, to extract useful information. Significant progress has been made in deep learning via supervised representation learning, where the labels, $\mathbf{y}$, for the downstream task are known while $p(\mathbf{y}|\mathbf{x})$ is learned directly (Sutskever et al., 2012; Hinton et al., 2012). Due to the cost of acquiring large labeled datasets, a recently renewed focus on unsupervised representation learning seeks to generate representations, $\mathbf{z}$, that allow learning of (a priori unknown) target supervised tasks more efficiently, i.e. with fewer labels (Devlin et al., 2018; Radford et al., 2019).

Our work is based on the information bottleneck principle (Tishby et al., 2000) stating that whenever a data representation discards information from the input which is not useful for a given task, it becomes less affected by nuisances, resulting in increased robustness for downstream tasks.

In the supervised setting one can directly apply the information bottleneck method by minimizing the mutual information between $\mathbf{z}$ and $\mathbf{x}$ while simultaneously maximizing the mutual information between $\mathbf{z}$ and $\mathbf{y}$ (Alemi et al., 2017). In the unsupervised setting, discarding only superfluous information is more challenging as without labels one cannot directly identify the relevant information. In this setting recent literature (Devon Hjelm et al., 2019; van den Oord et al., 2018) has instead focused on the InfoMax objective *maximizing* the mutual information between $\mathbf{x}$ and $\mathbf{z}$, $I(\mathbf{x}, \mathbf{z})$, instead of minimizing it.

In this paper, we extend the information bottleneck method to the unsupervised multi-view setting. To do this, we rely on a basic assumption of the multi-view literature – that each view provides the same *task relevant information* (Zhao et al., 2017). Hence, one can improve generalization by discarding from the representation all information which is not shared by both views. We do this through an objective which maximizes the mutual information between the representations of the two views (Multi-View InfoMax) while at the same time reducing the mutual information between each view and its corresponding representation (as with the information bottleneck). The resulting representation contains only the information shared by both views, eliminating the effect of independent factors of variations.

Our contributions are three-fold: (1) We extend the information bottleneck principle to the unsupervised multi-view setting and provide a rigorous theoretical analysis of it's application. (2) We define a new model that empirically leads to state-of-the-art results on two standard multi-view datasets, Sketchy and MIR-Flickr. (3) We further extend our theory to the single-view case, i.e. multiple

views are obtained by standard data augmentation techniques rather than from paired data, and empirically show the robustness of our representations compared to popular unsupervised approaches.

2 PRELIMINARIES AND FRAMEWORK

The challenge of representation learning can be formulated as finding a distribution $p(\mathbf{z}|\mathbf{x})$ that maps data observations $\mathbf{x} \in \mathsf{X}$ into a code space $\mathbf{z} \in \mathsf{Z}$. Whenever the end goal involves predicting a label $\mathbf{y}$, we consider only the $\mathbf{z}$ that are discriminative enough to identify the label. This requirement can be quantified by considering the amount of target information that remains accessible after encoding the data, and is known in literature as sufficiency of $\mathbf{z}$ for $\mathbf{y}$ (Achille & Soatto, 2018):

Definition 1. Sufficiency: A representation $\mathbf{z}$ of $\mathbf{x}$ is sufficient for $\mathbf{y}$ if and only if $I(\mathbf{x}; \mathbf{y}|\mathbf{z}) = 0$

Any model that has access to a sufficient representation $\mathbf{z}$ must be able to predict $\mathbf{y}$ at least as accurately as if it has access to the original data $\mathbf{x}$ instead. In fact, $\mathbf{z}$ is sufficient for $\mathbf{y}$ if and only if the amount of information regarding the task is unchanged by the encoding procedure (see Proposition B.1 in the Appendix):

$$I(\mathbf{x}; \mathbf{y}|\mathbf{z}) = 0 \iff I(\mathbf{x}; \mathbf{y}) = I(\mathbf{y}; \mathbf{z}). \tag{1}$$

Among sufficient representations, the ones that result in better generalization for unlabeled data instances are particularly appealing. When $\mathbf{x}$ has higher information content than $\mathbf{y}$, some of the information in $\mathbf{x}$ must be irrelevant for the prediction task. This can be better understood by subdividing $I(\mathbf{x}; \mathbf{z})$ into three components by using the chain rule of mutual information (see Appendix A):

$$I(\mathbf{x}; \mathbf{z}) = \underbrace{I(\mathbf{x}; \mathbf{z}|\mathbf{y})}_{\text{superfluous information}} + \underbrace{I(\mathbf{x}; \mathbf{y})}_{\text{predictive information}} - \underbrace{I(\mathbf{x}; \mathbf{y}|\mathbf{z})}_{\text{predictive information not in } \mathbf{z}}. \tag{2}$$

Conditional mutual information $I(\mathbf{x}; \mathbf{z}|\mathbf{y})$ represents the information in $\mathbf{z}$ that is not predictive of $\mathbf{y}$, i.e. **superfluous information**. While $I(\mathbf{x}; \mathbf{y})$ is a constant determined by how much label information is accessible from the raw observations; the last term $I(\mathbf{x}; \mathbf{y}|\mathbf{z})$ represents the amount of information regarding $\mathbf{y}$ that is lost by encoding $\mathbf{x}$ into $\mathbf{z}$. Note that this last term is zero whenever $\mathbf{z}$ is sufficient for $\mathbf{x}$. Since the amount of predictive information $I(\mathbf{x}; \mathbf{y})$ is fixed,

Proposition 2.1. *A sufficient representation $\mathbf{z}$ of $\mathbf{x}$ for $\mathbf{y}$ is minimal whenever $I(\mathbf{x}; \mathbf{z}|\mathbf{y})$ is minimal.*

Minimizing the amount of superfluous information can be done directly only in supervised settings. In fact, reducing $I(\mathbf{x}; \mathbf{z})$ without violating the sufficiency constraint necessarily requires making some additional assumptions on the predictive task (see Theorem B.1 in the Appendix). In Section 3 we describe a strategy to safely reduce the information content of a representation even when the label $\mathbf{y}$ is not observed, by exploiting redundant information in the form of an additional view on the data.

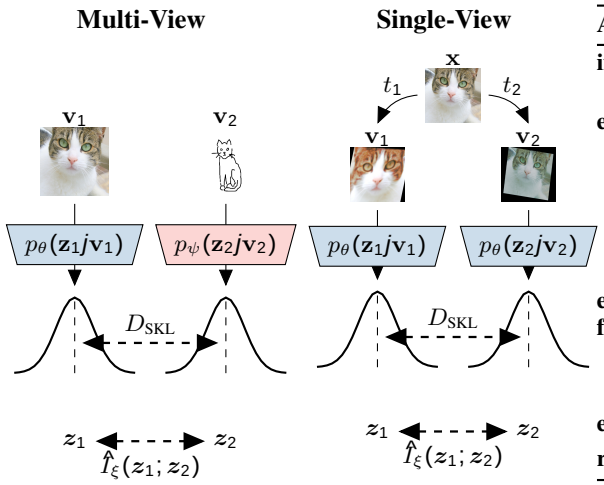
3 MULTI-VIEW INFORMATION BOTTLENECK

Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be two images of the same object from different view-points and $\mathbf{y}$ be its label. Assuming that the object is clearly distinguishable from both $\mathbf{v}_1$ and $\mathbf{v}_2$, any representation $\mathbf{z}$ with all the information that is accessible from both views would also contain the necessary label information. Furthermore, if $\mathbf{z}$ captures only the details that are visible from both pictures, it would reduce the total information content, discarding the view-specific details and reducing the sensitivity of the representation to view-changes. The theory to support this intuition is described in the following where $\mathbf{v}_1$ and $\mathbf{v}_2$ are jointly observed and referred to as data-views.

3.1 SUFFICIENCY AND ROBUSTNESS IN THE MULTI-VIEW SETTING

In this section we extend our analysis of sufficiency and robustness to the multi-view setting.

Intuitively, we can guarantee that $\mathbf{z}$ is sufficient for predicting $\mathbf{y}$ even without knowing $\mathbf{y}$ by ensuring that $\mathbf{z}$ maintains all information which is shared by $\mathbf{v}_1$ and $\mathbf{v}_2$. This intuition relies on a basic assumption of the multi-view environment – that each view provides the same task relevant information. To formalize this we define **redundancy**.



Algorithm 1: $\mathcal{L}_{\text{MIB}}(\theta, \psi; \beta, B)$

if *Multi-View* **then**
  $f(\boldsymbol{v}_1^{(i)}, \boldsymbol{v}_2^{(i)})g_{i=1}^B \quad p(\mathbf{v}_1, \mathbf{v}_2)$;
else
  $f\boldsymbol{x}^{(i)}g_{i=1}^B \quad p(\mathbf{x})$;
  $f(t_1^{(i)}, t_2^{(i)})g_{i=1}^B \quad p^2(\mathbf{t})$;
  for i 1 **to** B **do**
    $\boldsymbol{v}_1^{(i)}, \boldsymbol{v}_2^{(i)} \quad t_1^{(i)}(\boldsymbol{x}^{(i)}), t_2^{(i)}(\boldsymbol{x}^{(i)})$;
  end for
end if
for i 1 **to** B **do**
  $\boldsymbol{z}_1^{(i)}, \boldsymbol{z}_2^{(i)} \quad p_\theta(\mathbf{z}_1|\boldsymbol{v}_1^{(i)}), p_\psi(\mathbf{z}_2|\boldsymbol{v}_2^{(i)})$;
  $L_m^{(i)} \quad D_{SKL}(p_\theta(\mathbf{z}_1|\boldsymbol{v}_1^{(i)})||p_\psi(\mathbf{z}_2|\boldsymbol{v}_2^{(i)}))$;
end for
return $\frac{\beta}{B}\sum_{i=1}^B L_m^{(i)} \quad \hat{I}_\xi(f(\boldsymbol{z}_1^{(i)}, \boldsymbol{z}_2^{(i)})g_{i=1}^B)$

Figure 1: Visualizing our Multi-View Information Bottleneck model for both multi-view and single-view settings. Whenever $p(\mathbf{v}_1)$ and $p(\mathbf{v}_2)$ have the same distribution, the two encoders can share their parameters.

Definition 2. Redundancy: $\mathbf{v}_1$ is redundant with respect to $\mathbf{v}_2$ for $\mathbf{y}$ if and only if $I(\mathbf{y}; \mathbf{v}_1|\mathbf{v}_2) = 0$

Intuitively, a view $\mathbf{v}_1$ is redundant for a task whenever it is irrelevant for the prediction of $\mathbf{y}$ when $\mathbf{v}_2$ is already observed. Whenever $\mathbf{v}_1$ and $\mathbf{v}_2$ are **mutually redundant** ($\mathbf{v}_1$ is redundant with respect to $\mathbf{v}_2$, and vice-versa), we can show the following:

Corollary 1. *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 $\mathbf{v}_2$ ($I(\mathbf{v}_1; \mathbf{v}_2|\mathbf{z}_1) = 0$) then $\mathbf{z}_1$ is as predictive for $\mathbf{y}$ as the joint observation of the two views ($I(\mathbf{v}_1\mathbf{v}_2; \mathbf{y}) = I(\mathbf{y}; \mathbf{z}_1)$).*

In other words, whenever it is possible to assume mutual redundancy, any representation which contains all the information shared by both views (the redundant information) is as useful as both views for predicting the label $\mathbf{y}$.

By factorizing the mutual information between $\mathbf{v}_1$ and $\mathbf{z}_1$ analogously to Equation 2, we can identify 3 components:

$$I(\mathbf{z}_1; \mathbf{v}_1) = \underbrace{I(\mathbf{v}_1; \mathbf{z}_1|\mathbf{v}_2)}_{\text{superfluous information}} + \underbrace{I(\mathbf{v}_1; \mathbf{v}_2)}_{\text{shared information}} - \underbrace{I(\mathbf{v}_1; \mathbf{v}_2|\mathbf{z}_1)}_{\text{shared information not in } \mathbf{z}_1} .$$

Since $I(\mathbf{v}_1; \mathbf{v}_2)$ is a constant that depends only on the two views and $I(\mathbf{v}_1; \mathbf{v}_2|\mathbf{z}_1)$ must be zero if we want the representation to be sufficient for the label, we conclude that $I(\mathbf{v}_1; \mathbf{z}_1)$ can be reduced by minimizing $I(\mathbf{v}_1; \mathbf{z}_1|\mathbf{v}_2)$. This term intuitively represents the information $\mathbf{z}_1$ contains which is unique to $\mathbf{v}_1$ and not shared by $\mathbf{v}_2$. Since we assumed mutual redundancy between the two views, this information must be irrelevant for the predictive task and, therefore, it can be safely discarded. The proofs and formal assertions for the above statements and Corollary 1 can be found in Appendix B.

The less the two views have in common, the more $I(\mathbf{z}_1; \mathbf{v}_1)$ can be reduced without violating sufficiency for the label, the more robust the resulting representation. At the extreme, $\mathbf{v}_1$ and $\mathbf{v}_2$ share only label information, in which case we can show that $\mathbf{z}_1$ is minimal for $\mathbf{y}$ and our method is identical to the supervised information bottleneck method without needing to access the labels. Conversely, if $\mathbf{v}_1$ and $\mathbf{v}_2$ are identical, then our method degenerates to the InfoMax principle since no information can be safely discarded (see Appendix C).

3.2 THE MULTI-VIEW INFORMATION BOTTLENECK LOSS FUNCTION

Given $\mathbf{v}_1$ and $\mathbf{v}_2$ that satisfy the mutual redundancy condition for a label $\mathbf{y}$, we would like to define an objective function for the representation $\mathbf{z}_1$ of $\mathbf{v}_1$ that discards as much information as possible without losing any label information. In Section 3.1 we showed that we can maintain sufficiency for $\mathbf{y}$ by ensuring that $I(\mathbf{v}_1; \mathbf{v}_2|\mathbf{z}_1) = 0$, and that decreasing $I(\mathbf{z}_1; \mathbf{v}_1|\mathbf{v}_2)$ will increase the robustness of the representation by discarding irrelevant information. So if we combine these two terms using a relaxed Lagrangian objective, then we obtain:

$$\mathcal{L}_1(\theta; \lambda_1) = \underbrace{I_\theta(\mathbf{z}_1; \mathbf{v}_1|\mathbf{v}_2)}_{\text{superfluous information}} + \underbrace{\lambda_1 \, I_\theta(\mathbf{v}_1; \mathbf{v}_2|\mathbf{z}_1)}_{\text{sufficiency of } \mathbf{z}_1 \text{ for predicting } \mathbf{y}}, \tag{3}$$

where θ denotes the dependency on the parameters of the encoder $p_\theta(\mathbf{z}_1|\mathbf{v}_1)$, and λ_1 represents the Lagrangian multiplier introduced by the constrained optimization. Symmetrically, we define a loss $\mathcal{L}_2$ to optimize the parameters ψ of a conditional distribution $p_\psi(\mathbf{z}_2|\mathbf{v}_2)$ that defines a robust sufficient representation $\mathbf{z}_2$ of the second view $\mathbf{v}_2$:

$$\mathcal{L}_2(\psi; \lambda_2) = \underbrace{I_\psi(\mathbf{z}_2; \mathbf{v}_2|\mathbf{v}_1)}_{\text{superfluous information}} + \underbrace{\lambda_2 \, I_\psi(\mathbf{v}_1; \mathbf{v}_2|\mathbf{z}_2)}_{\text{sufficiency of } \mathbf{z}_2 \text{ for predicting } \mathbf{y}}, \tag{4}$$

Although $\mathcal{L}_1$ and $\mathcal{L}_2$ can not be computed directly, by defining $\mathbf{z}_1$ and $\mathbf{z}_2$ on the same domain Z and re-parametrizing the Lagrangian multipliers, their sum can be upper bounded as follows:

$$\mathcal{L}_{MIB}(\theta, \psi; \beta) = - \underbrace{I_{\theta\psi}(\mathbf{z}_1; \mathbf{z}_2)}_{\text{sufficiency of } \mathbf{z}_1 \text{ and } \mathbf{z}_2 \text{ for predicting } \mathbf{y}} + \beta \underbrace{D_{SKL}(p_\theta(\mathbf{z}_1|\mathbf{v}_1)||p_\psi(\mathbf{z}_2|\mathbf{v}_2))}_{\text{superfluous information}}, \tag{5}$$

where D_{SKL} is the symmetrized KL divergence obtained by averaging $D_{\mathsf{KL}}(p_\theta(\mathbf{z}_1|\mathbf{v}_1)||p_\psi(\mathbf{z}_2|\mathbf{v}_2))$ and $D_{\mathsf{KL}}(p_\psi(\mathbf{z}_2|\mathbf{v}_2)||p_\theta(\mathbf{z}_1|\mathbf{v}_1))$, while the coefficient β defines the trade-off between sufficiency and robustness of the representation, which is a hyper-parameter in this work. The resulting Multi-View Infomation Bottleneck (MIB) model (Equation 5) is visualized in Figure 1, while the batch-based computation of the loss function is summarized in Algorithm 1.

The symmetrized KL divergence $D_{SKL}(p_\theta(\mathbf{z}_1|\mathbf{v}_1)||p_\psi(\mathbf{z}_2|\mathbf{v}_2))$ can be computed directly whenever $p_\theta(\mathbf{z}_1|\mathbf{v}_1)$ and $p_\psi(\mathbf{z}_2|\mathbf{v}_2)$ have a known density, while the mutual information between the two representations $I_{\theta\psi}(\mathbf{z}_1; \mathbf{z}_2)$ can be maximized by using any sample-based differentiable mutual information lower bound. Both the Jensen-Shannon I_{JSD} (Devon Hjelm et al., 2019; Poole et al., 2019) and the InfoNCE I_{NCE} (van den Oord et al., 2018) estimators used in this work require introducing an auxiliary parameteric model $C_\xi(\boldsymbol{z}_1, \boldsymbol{z}_2)$, which is jointly optimized during the training procedure. The full derivation for the MIB loss function can be found in Appendix D.

3.3 SELF-SUPERVISION AND INVARIANCE

In this section, we introduce a methodology to build mutually redundant views starting from single observations $\mathbf{x}$ with domain X by exploiting known symmetries of the task.

By picking a class T of functions $t : \mathsf{X} \to \mathsf{W}$ that do not affect label information, it is possible to artificially build views that satisfy mutual redundancy for $\mathbf{y}$ with a procedure similar to data-augmentation. Let $\mathbf{t}_1$ and $\mathbf{t}_2$ be two random variables over T, then $\mathbf{v}_1 := \mathbf{t}_1(\mathbf{x})$ and $\mathbf{v}_2 := \mathbf{t}_2(\mathbf{x})$ must be mutually redundant for $\mathbf{y}$. Since no function in T affects label information ($I(\mathbf{v}_1; \mathbf{y}) = I(\mathbf{v}_2; \mathbf{y}) = I(\mathbf{x}; \mathbf{y})$), a representation $\mathbf{z}_1$ of $\mathbf{v}_1$ that is sufficient to $\mathbf{v}_2$ must contain same amount of predictive information as $\mathbf{x}$. Formal proofs can be found in Appendix B.4.

Whenever the two transformations for the same observations are independent ($I(\mathbf{t}_1; \mathbf{t}_2|\mathbf{x}) = 0$), they introduce uncorrelated variations in the two views. As an example, if T represents a set of small translations, the two views will differ by a small shift. Since this information is not shared, $\mathbf{z}_1$ that contains only common information between $\mathbf{v}_1$ and $\mathbf{v}_2$ will discard fine-grained details regarding the position.

For single-view datasets, we generate the two views $\mathbf{v}_1$ and $\mathbf{v}_2$ by independently sampling two functions from the same function class T with uniform probability. Since the resulting $\mathbf{t}_1$ and $\mathbf{t}_2$

have the same distribution, the two generated views will also have the same marginals. For this reason, the two conditional distributions $p_\theta(\mathbf{z}_1|\mathbf{v}_1)$ and $p_\psi(\mathbf{z}_2|\mathbf{v}_2)$ can share their parameters and only one encoder can be used. Full (or partial) parameter sharing can be also applied in the multi-view settings whenever the two views have the same (or similar) marginal distributions.

4 Related Work

The space of all the possible representations $\mathbf{z}$ of $\mathbf{x}$ for a predictive task $\mathbf{y}$ can be represented as a region in the Information Plane (Tishby et al., 2000). Each representation is characterised by the amount of information regarding the raw observation $I(\mathbf{x};\mathbf{z})$ and the corresponding measure of accessible predictive information $I(\mathbf{y};\mathbf{z})$ (x and y axis respectively on Figure 2). Ideally, a good representation would be maximally informative about the label while retaining a minimal amount of information from the observations (top left corner of the parallelogram). Further details on the Information Plane and the bounds visualized in Figure 2 are described in Appendix E.

Thanks to recent progress in mutual information estimation (Nguyen et al., 2008; Ishmael Belghazi et al., 2018; Poole et al., 2019), the InfoMax principle (Linsker, 1988) has gained attention for unsupervised representation learning (Devon Hjelm et al., 2019; van den Oord et al., 2018). Since the InfoMax objective involves maximizing $I(\mathbf{x};\mathbf{z})$, the resulting representation aims to preserve all the information regarding the raw observations (top right corner in Figure 2). Despite their success, Tschannen et al. (2019) has shown that the effectiveness of the InfoMax models is due to inductive biases introduced by the architecture and estimators rather than the training objective itself, since the InfoMax objective can be trivially maximized by using invertible encoders.

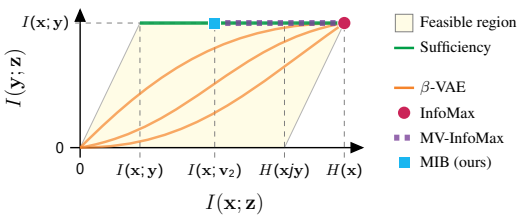


Figure 2: Information Plane determined by $I(\mathbf{x};\mathbf{z})$ (x-axis) and $I(\mathbf{y};\mathbf{z})$ (y-axis). Different objectives are compared based on their target.

On the other hand, Variational Autoencoders (VAEs) (Kingma & Welling, 2014) define a training objective that balances compression and reconstruction error (Alemi et al., 2018) through an hyper-parameter β. Whenever β is close to 0, the VAE objective aims for a lossless representation, approaching the same region of the Information Plane as the one targeted by InfoMax (Barber & Agakov, 2003). When β approaches large values, the representation becomes more compressed, showing increased generalization and disentanglement (Higgins et al., 2017; Burgess et al., 2018), and, as β approaches infinity, $I(\mathbf{z};\mathbf{x})$ goes to zero. During this transition from low to high β, however, there are no guarantees that VAEs will retain label information (Theorem B.1 in the Appendix). The path between the two regimes depends on how well the label information aligns with the inductive bias introduced by encoder (Jimenez Rezende & Mohamed, 2015; Kingma et al., 2016), prior (Tomczak & Welling, 2018) and decoder architectures (Gulrajani et al., 2017; Chen et al., 2017).

Concurrent work applies the InfoMax principle in Multi-View settings (Ji et al., 2019; Hénaff et al., 2019; Tian et al., 2019; Bachman et al., 2019), aiming to maximize mutual information between the representation $\mathbf{z}$ of a first data-view $\mathbf{x}$ and a second one $\mathbf{v}_2$. The target representation for the Multi-View InfoMax (MV-InfoMax) models should contain at least the amount of information in $\mathbf{x}$ that is predictive for $\mathbf{v}_2$, targeting the region $I(\mathbf{z};\mathbf{x}) \geq I(\mathbf{x};\mathbf{v}_2)$ on the Information Plane. Whenever $\mathbf{x}$ is redundant with respect to $\mathbf{v}_2$ for $\mathbf{y}$, the representation must be also sufficient for $\mathbf{y}$ (Corollary 1). Since $\mathbf{z}$ has no incentive in discarding any information regarding $\mathbf{x}$, a representation that is optimal according to the InfoMax principle is also optimal for MV-InfoMax. Our model with $\beta = 0$ (Equation 5) belong to this family of objectives since the minimality term is discarded.

In contrast to all of the above, our work is the first to explicitly identify and discard superfluous information from the representation in the unsupervised multi-view setting. The idea of discarding only irrelevant information was introduced in Tishby et al. (2000) and identified as one of the possible reasons behind the generalization capabilities of deep neural networks by Tishby & Zaslavsky (2015) and Achille & Soatto (2018), but so far has been utilized only in supervised settings (Alemi et al., 2017). Conversely, β-VAE models remove information indiscriminately without explicitly

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

Method	mAP@all		Prec@200	
SaN (Yu et al., 2017)	0.208		0.292	
GN Triplet (Sangkloy et al., 2016)	0.529		0.716	
Siamese CNN (Qi et al., 2016)	0.481		0.612	
Siamese-AlexNet	0.518		0.690	
Triplet-AlexNet	0.573		0.761	
DSH (Liu et al., 2017)	0.711		**0.866**	
GDH (Zhang et al., 2018)	0.810		-	
MV-InfoMax[2]	0.008		0.008	
MIB	**0.856**	**0.005**	0.848	0.005
MIB (64-bits)	0.851	0.004	0.834	0.003

Table 1: Examples of the two views and class label from the Sketchy dataset (on the left) and comparison between MIB and other popular models in literature on the sketch-based image retrieval task (on the right). denotes models that use a 64-bits binary representation.

identifying which information is superfluous, and the InfoMax and Multiview-InfoMax methods do not explicitly try to remove superfluous information at all. In fact, among the representations that are optimal according to Multi-View InfoMax (purple dotted line in Figure 2), the MIB objective results in the the representation with the least superfluous information, i.e. the most robust.

5 Experiments

In this section we demonstrate the effectiveness of our model against state-of-the-art baselines in both the Multi-View and Single-View setting. In the Single-View setting, we also estimate the coordinates on the Information Plane for each of the baseline methods as well as our method to validate the theory in Sec. 3.

5.1 Multi-View Tasks

We compare MIB on the sketch-based image retrieval (Sangkloy et al., 2016) and Flickr multiclass image classification tasks with domain specific and prior Multi-View learning methods.

Sketchy The Sketchy dataset (Sangkloy et al., 2016) consists of 12,500 images and 75,471 hand-drawn sketches of objects from 125 classes. As in Liu et al. (2017), we also include another 60,502 images from the ImageNet (Deng et al., 2009) from the same classes, which results in total 73,002 natural object images. As per the experimental protocol of Zhang et al. (2018), a total of 6,250 sketches (50 sketches per category) are randomly selected and removed from the training set for testing purpose, which leaves 69,221 sketches for training the model. The sketch-based image retrieval task is a ranking of 73,002 natural images according to the unseen test (query) sketch. Retrieval is done for our model by generating representations for the query sketch as well as all natural images, and ranking the image by the euclidean distance of their representation from the sketch representation. The baselines use various domain specific ranking methodologies. Model performance is computed based on the class of the ranked pictures corresponding to the query. The training set consists of pairs of image $\mathbf{v}_1$ and sketch $\mathbf{v}_2$ randomly selected from the same class, to ensure that both views contain the equivalent label information (mutual redundancy).

Following recent prior works (Zhang et al., 2018; Dutta & Akata, 2019), we use features extracted from images and sketches by a VGG (Simonyan & Zisserman, 2014) architecture trained for classification on the TU-Berlin dataset (Eitz et al., 2012). The resulting flattened 4096-dimensional feature vectors are fed to our image and sketch encoders to produce a 64-dimensional representation. Both encoders consist of neural networks with hidden layers of 2048 and 1024 units respectively. Size of the representation and regularization strength β are tuned on a validation sub-split. We evaluate MIB on five different train/test splits[1] and report mean and standard deviation in Table 5.1. Further details on our training procedure and architecture are in Appendix F.

Table 5.1 shows that the our model achieves strong performance for both mean average precision (mAP@all) and precision at 200 (Prec@200), suggesting that the representation is able to capture

[1]Processed dataset and splits will be publicly released on paper acceptance

Method	mAP	
Original Inputs^{y}	0.48	
CCA^{y}	0.529	
Contrastive^{y} (Hermann & Blunsom, 2014)	0.565	
DCCA^{y} (Andrew et al., 2013)	0.573	
MVAE-var^{y} (Ngiam et al., 2011)	0.595	
VCCA^{y} (Wang et al., 2016)	0.605	
VCCA-private^{y}	0.615	
bi-VCCA-private^{y}	0.626	
MV-InfoMax	**0.751**	**0.02**
MIB (β = 10 ^{3})	**0.749**	**0.02**
MIB (β = 1)	0.688	0.02

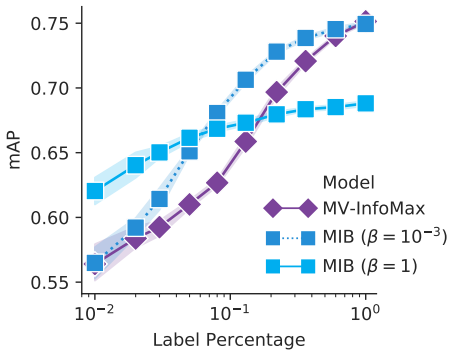


Figure 3: Left: mean average precision (mAP) of the classifer trained on different Multi-View representations for the MIR-Flickr task. Right: comparing the performance for different values of β and percentages of given labels. Each model uses encoders of comparable size, producing a 1024d representation. ^{y} results from Wang et al. (2016).

the common class information between the paired pictures and sketches. The effectiveness of MIB on the retrieval task can be mostly imputed to the regularization introduced with the symmetrized KL divergence between the two encoded views. Other than discarding view-private information, this term actively aligns the representations of $\mathbf{v}_1$ and $\mathbf{v}_2$, making the MIB model especially suitable for retrieval tasks

MIR-Flickr The MIR-Flickr dataset (Huiskes & Lew, 2008) consists of 1M images annotated with 800K distinct user tags. Each image is represented by a vector of 3,857 hand-crafted image features ($\mathbf{v}_1$), while the 2,000 most frequent tags are used to produce a 2000-dimensional multi-hot encoding ($\mathbf{v}_2$) for each picture. The dataset is divided into labeled and unlabeled sets that respectively contain 975K and 25K images, where the labeled set also contains 38 distinct topic classes together with the user tags.

Training images with less than two tags are removed, which reduces the total number of training samples to 749,647 pairs (Sohn et al., 2014; Wang et al., 2016). The labeled set contains 5 different splits of train, validation and test sets of size 10K/5K/10K respectively. Following a standard procedure in literature (Srivastava & Salakhutdinov, 2014; Wang et al., 2016), we train our model on the unlabeled pairs of images and tags. Then a multi-label logistic classifier is trained from the representation of 10K labeled train images to the corresponding macro-categories. The quality of the representation is assessed based on the performance of the trained logistic classifier on the labeled test set. Each encoder consists of a multi-layer perceptron of 4 hidden layers with ReLU activations learning two 1024-dimensional representations $\mathbf{z}_1$ and $\mathbf{z}_2$ for images $\mathbf{v}_1$ and tags $\mathbf{v}_2$ respectively. Examples of the two views, labels, and further details on the training procedure are in Appendix F.

Our MIB model is compared with other popular Multi-View learning models in Figure 3 for $\beta = 0$ (Multi-View InfoMax), $\beta = 1$ and $\beta = 10^{3}$ (where $\beta = 10^{3}$ was the result of tuning β on our validation set). Although the tuned MIB performs similarly to Multi-View InfoMax with a large number of labels, it outperforms it when fewer labels are available. Furthermore, by choosing a larger β the accuracy of our model drastically increases in scarce label regimes (see right side of Figure 3).

A possible reason for the effectiveness of MIB against some of the other baselines may be our ability to use mutual information estimators that do not require reconstruction. Both Multi-View VAE (MVAE) and Deep Variational CCA (VCCA) rely on a reconstruction term to capture cross-modal information, which can introduce bias that decreases performance.

[2]These results are included only for completeness, as the Multi-View InfoMax objective does not produce consistent representations for the two views so there is no straight-forward way to use it for ranking.

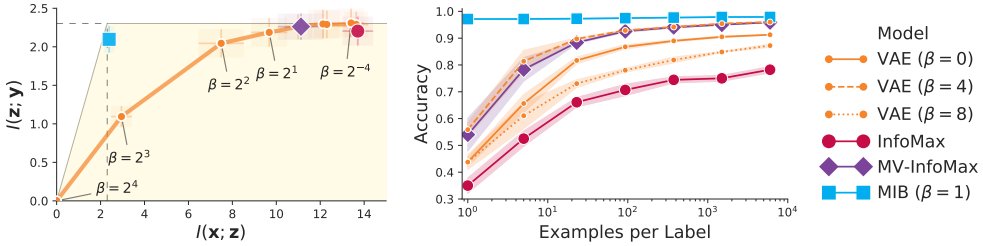


Figure 4: Comparing the representations obtained with different objectives on MNIST dataset. The empirical estimation of the coordinates on the Information Plane (in nats on the left) is followed by the respective classification accuracy for different percentages of given labels. Representations that discard more observational information tend to perform better in scarce label regimes.

5.2 Self-supervised Single-View Task

In this section, we compare the performance of different unsupervised learning models by measuring their data efficiency and empirically estimating the coordinates of their representation on the Information Plane. Since accurate estimation of mutual information is extremely expensive (McAllester & Stratos, 2018), we focus on relatively small experiments that aim to uncover the difference between popular approaches for representation learning.

The dataset is generated from MNIST by creating the two views, $\mathbf{v}_1$ and $\mathbf{v}_2$, via the application of data augmentation consisting of small affine transformations and independently pixel corruption to each image. These are kept small enough to ensure that label information is not effected. Each pair of views is generated from the same underlying image, so no label information is used in this process. Further details on the data augmentation are in Appendix F. To evaluate we model, we train the representation model on the unlabeled multi-view dataset just described, and then fix the representation model. A logistic regression model is trained using these representations along with some set of labels for the training set, and we report the accuracy of this model on a disjoint test set as is standard for the multi-view literature(Tschannen et al., 2019). We estimate $I(\mathbf{x};\mathbf{z})$ and $I(\mathbf{y};\mathbf{z})$ using mutual information estimation networks trained from scratch on the final representations using batches of joint samples $\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{z}^{(i)})\}_{i=1}^{B} \sim p(\mathbf{x},\mathbf{y})p_\theta(\mathbf{z}|\mathbf{x})$. All models are trained using the same encoder architecture consisting of 2 layers of 1024 hidden units with ReLU activations, resulting in 64-dimensional representations. The data augmentation procedure was also applied for single-view architectures and models were trained for 1 million iterations with batch size $B = 64$.

Figure 4 summarizes the results. The empirical measurements of mutual information reported on the Information Plane are consistent with the theoretical analysis reported in Sec. 4: models that retain less information about the data while maintaining the maximal amount of predictive information, result in better classification performance at low-label regimes, confirming the hypothesis that discarding irrelevant information yields robustness and more data-efficient representations. Notably, the MIB model retains almost exclusively label information, hardly decreasing the classification performance when only one label is used for each data point.

6 Conclusions and Future Work

In this work, we introduce a novel method that relies on multiple data-views to produce robust representation for downstream task. Our model is compared empirically against other approaches in the literature of sketch-based image retrieval, Multi-View and unsupervised learning. The strong performances obtained in the different areas show that Multi-View Information Bottleneck can be practically applied to various tasks for which the paired observations are either available or are artificially produced.

Future work will address optimization strategy, flexibility of the encoding distribution and the role of different choices for data augmentation. We believe that this direction of exploration would be able to bridge the Information Bottleneck principle and invariant neural networks, which are able to exploit known symmetries and structure of the data to remove superfluous information.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of Invariance and Disentanglement in Deep Representations. *JMLR*, 2018.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep Variational Information Bottleneck. In *ICLR*, 2017.

Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a Broken ELBO. In *ICML*, 2018.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, 2013.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning Representations by Maximizing Mutual Information Across Views. *arXiv*, 2019.

David Barber and Felix Agakov. The im algorithm: A variational approach to information maximization. In *NIPS*, 2003.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv*, 2018.

Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational Lossy Autoencoder. In *ICLR*, 2017.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2018.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019.

Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM TOG*, 2012.

Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *TPAMI*, 2013.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. PixelVAE: A Latent Variable Model for Natural Images. In *ICLR*, 2017.

Olivier J. Hénaff, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-Efficient Image Recognition with Contrastive Predictive Coding. *arXiv*, 2019.

Karl Moritz Hermann and Phil Blunsom. Multilingual Distributed Representations without Word Alignment. In *ICLR*, 2014.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *SPM*, 2012.

Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *ICMIR*, pp. 39–43, 2008.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. MINE: Mutual Information Neural Estimation. In *ICML*, 2018.

Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *ICCV*, 2019.

Danilo Jimenez Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. In *ICML*, 2015.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *NIPS*, 2016.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015.

R. Linsker. Self-organization in a perceptual network. *Computer*, 1988.

Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep Sketch Hashing: Fast Free-hand Sketch-Based Image Retrieval. In *CVPR*, 2017.

David McAllester and Karl Stratos. Formal Limitations on the Measurement of Mutual Information. *arXiv*, 2018.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, 2011.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. In *NIPS*, 2008.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On Variational Bounds of Mutual Information. In *ICML*, 2019.

Y. Qi, Y. Song, H. Zhang, and J. Liu. Sketch-based image retrieval via siamese convolutional neural network. In *ICIP*, 2016.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 2016.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, 2014.

Kihyuk Sohn, Wenling Shang, and Honglak Lee. Improved multimodal deep learning with variation of information. In *NIPS*, 2014.

Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. *JMLR*, 2014.

Wanhua Su, Yan Yuan, and Mu Zhu. A relationship between the average precision and the area under the roc curve. In *ICTIR*, 2015.

Ilya Sutskever, Geoffrey E Hinton, and A Krizhevsky. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Multiview Coding. *arXiv*, 2019.

Naftali Tishby and Noga Zaslavsky. Deep Learning and the Information Bottleneck Principle. In *ITW*, 2015.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv*, 2000.

Jakub M. Tomczak and Max Welling. VAE with a VampPrior. In *AISTATS*, 2018.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On Mutual Information Maximization for Representation Learning. *arXiv*, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv*, 2018.

Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep Variational Canonical Correlation Analysis. *arXiv*, 2016.

Qian Yu, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Sketch-a-Net that Beats Humans. *IJCV*, 2017.

Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In *ECCV*, 2018.

Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *IF*, 2017.

A PROPERTIES OF MUTUAL INFORMATION AND ENTROPY

In this section we enumerate some of the properties of mutual information that are used to prove the theorems reported in this work. For any random variables $\mathbf{w}$, $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$:

(P_1) Positivity:

$$I(\mathbf{x};\mathbf{y}) \geq 0, I(\mathbf{x};\mathbf{y}|\mathbf{z}) \geq 0$$

(P_2) Chain rule:

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

(P_3) Chain rule (Multivariate Mutual Information):

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

(P_4) Positivity of discrete entropy:
For discrete $\mathbf{x}$

$$H(\mathbf{x}) \geq 0, H(\mathbf{x}|\mathbf{y}) \geq 0$$

(P_5) Entropy and Mutual Information

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

B THEOREMS AND PROOFS

B.1 ON SUFFICIENCY

Proposition B.1. *Let $\mathbf{x}$ and $\mathbf{y}$ be random variables with joint distribution $p(\mathbf{x},\mathbf{y})$. Let $\mathbf{z}$ be a representation of $\mathbf{x}$, then $\mathbf{z}$ is sufficient for $\mathbf{y}$ if and only if $I(\mathbf{x};\mathbf{y}) = I(\mathbf{y};\mathbf{z})$*

Hypothesis:

(H_1) $I(\mathbf{y};\mathbf{z}|\mathbf{x}) = 0$

Thesis:

(T_1) $I(\mathbf{x};\mathbf{y}|\mathbf{z}) = 0 \iff I(\mathbf{x};\mathbf{y}) = I(\mathbf{y};\mathbf{z})$

Proof.

$$\begin{aligned} I(\mathbf{x};\mathbf{y}|\mathbf{z}) &\overset{(P_3)}{=} I(\mathbf{x};\mathbf{y}) - I(\mathbf{x};\mathbf{y};\mathbf{z}) \overset{(P_3)}{=} I(\mathbf{x};\mathbf{y}) - I(\mathbf{y};\mathbf{z}) - I(\mathbf{y};\mathbf{z}|\mathbf{x}) \\ &\overset{(H_1)}{=} I(\mathbf{x};\mathbf{y}) - I(\mathbf{y};\mathbf{z}) \end{aligned}$$

Since both $I(\mathbf{x};\mathbf{y})$ and $I(\mathbf{y};\mathbf{z})$ are non-negative (P_1), $I(\mathbf{x};\mathbf{y}|\mathbf{z}) = 0 \iff I(\mathbf{y};\mathbf{z}) = I(\mathbf{x};\mathbf{y})$ □

B.2 NO FREE GENERALIZATION

Theorem B.1. *Let $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{y}$ be random variables with joint distribution $p(\mathbf{x},\mathbf{y},\mathbf{z})$. Let $\mathbf{z}^0$ be a representation of $\mathbf{x}$ that satisfies $I(\mathbf{x};\mathbf{z}) > I(\mathbf{x};\mathbf{z}^0)$, then it is always possible to find a label $\mathbf{y}$ for which $\mathbf{z}^0$ is not predictive for $\mathbf{y}$ while $\mathbf{z}$ is.*

Hypothesis:

(H_1) $I(\mathbf{y};\mathbf{z}^0|\mathbf{x}) = 0$

Thesis:

$(T_1)\ \ I(\mathbf{x};\mathbf{z}^0) < I(\mathbf{x};\mathbf{z}) \implies \exists \mathbf{y}. I(\mathbf{y};\mathbf{z}) > I(\mathbf{y};\mathbf{z}^0) = 0$

Proof. By construction.

1. We first factorize $\mathbf{x}$ as a function of two independent random variables (Proposition 2.1 Achille & Soatto (2018)) by picking $\mathbf{y}$ such that:

   $(C_1)\ \ I(\mathbf{y};\mathbf{z}^0) = 0$
   $(C_2)\ \ \mathbf{x} = f(\mathbf{z}^0, \mathbf{y})$

   for some deterministic function f. Note that such $\mathbf{y}$ always exists.

2. Since $\mathbf{x}$ is a function of $\mathbf{y}$ and $\mathbf{z}^0$:

   $(C_4)\ \ I(\mathbf{x};\mathbf{z}|\mathbf{y}\mathbf{z}^0) = 0$

Considering $I(\mathbf{y};\mathbf{z})$:

$$
\begin{aligned}
I(\mathbf{y};\mathbf{z}) &\overset{(P_3)}{=} I(\mathbf{y};\mathbf{z}|\mathbf{x}) + I(\mathbf{x};\mathbf{y};\mathbf{z}) \\
&\overset{(P_1)}{\geq} I(\mathbf{x};\mathbf{y};\mathbf{z}) \\
&\overset{(P_3)}{=} I(\mathbf{x};\mathbf{z}) - I(\mathbf{x};\mathbf{z}|\mathbf{y}) \\
&\overset{(P_3)}{=} I(\mathbf{x};\mathbf{z}) - I(\mathbf{x};\mathbf{z}|\mathbf{y}\mathbf{z}^0) - I(\mathbf{x};\mathbf{z};\mathbf{z}^0|\mathbf{y}) \\
&\overset{(C_2)}{=} I(\mathbf{x};\mathbf{z}) - I(\mathbf{x};\mathbf{z};\mathbf{z}^0|\mathbf{y}) \\
&\overset{(P_3)}{=} I(\mathbf{x};\mathbf{z}) - I(\mathbf{x};\mathbf{z}^0|\mathbf{y}) + I(\mathbf{x};\mathbf{z}^0|\mathbf{y}\mathbf{z}) \\
&\overset{(P_1)}{\geq} I(\mathbf{x};\mathbf{z}) - I(\mathbf{x};\mathbf{z}^0|\mathbf{y}) \\
&\overset{(P_3)}{=} I(\mathbf{x};\mathbf{z}) - I(\mathbf{x};\mathbf{z}^0) + I(\mathbf{x};\mathbf{y};\mathbf{z}^0) \\
&\overset{(P_3)}{=} I(\mathbf{x};\mathbf{z}) - I(\mathbf{x};\mathbf{z}^0) + I(\mathbf{y};\mathbf{z}^0) - I(\mathbf{y};\mathbf{z}^0|\mathbf{x}) \\
&\overset{(P_1)}{\geq} I(\mathbf{x};\mathbf{z}) - I(\mathbf{x};\mathbf{z}^0) - I(\mathbf{y};\mathbf{z}^0|\mathbf{x}) \\
&\overset{(H_1)}{=} I(\mathbf{x};\mathbf{z}) - I(\mathbf{x};\mathbf{z}^0)
\end{aligned}
$$

Whenever $I(\mathbf{x};\mathbf{z}) > I(\mathbf{x};\mathbf{z}^0)$, $I(\mathbf{y};\mathbf{z})$ must be strictly positive, while $I(\mathbf{y};\mathbf{z}^0) = 0$ by construction. Therefore such $\mathbf{y}$ exists. □

Corollary B.1.1. *Let $\mathbf{z}^0$ be a representation of $\mathbf{x}$ that discards observational information. There is always a label $\mathbf{y}$ for which a $\mathbf{z}^0$ is not predictive, while the original observations are.*
Hypothesis:

(H_1) *$\mathbf{x}$ is discrete (empirical distribution)*

$(H_2)\ \ I(\mathbf{z}^0;\mathbf{x}) < H(\mathbf{x})$

Thesis:

$(T_1)\ \ \exists \mathbf{y}. I(\mathbf{y};\mathbf{x}) > I(\mathbf{y};\mathbf{z}^0) = 0$

Proof. By construction using Theorem B.1.

1. Set $\mathbf{z} = \mathbf{x}$:

   $(C_1)\ \ I(\mathbf{x};\mathbf{z}) \overset{(P_5)}{=} H(\mathbf{x}) - H(\mathbf{x}|\mathbf{z}) \overset{(H_1)}{=} H(\mathbf{x})$

2. $I(\mathbf{z}^{\emptyset};\mathbf{x}) < H(\mathbf{x}) \overset{(C_1)}{\implies} I(\mathbf{z}^{\emptyset};\mathbf{x}) < I(\mathbf{x};\mathbf{z})$

Since the hypothesis are met, we conclude that there exist $\mathbf{y}$ such that $I(\mathbf{y};\mathbf{x}) > I(\mathbf{y};\mathbf{z}^{\emptyset}) = 0$ □

B.3 MULTI-VIEW

B.3.1 MULTI-VIEW REDUNDANCY AND SUFFICIENCY

Proposition B.2. *Let $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{y}$ be random variables with joint distribution $p(\mathbf{v}_1, \mathbf{v}_2, \mathbf{y})$. Let $\mathbf{z}_1$ be a representation of $\mathbf{v}_1$, then:*

$$I(\mathbf{v}_1;\mathbf{y}|\mathbf{z}_1) \leq I(\mathbf{v}_1;\mathbf{v}_2|\mathbf{z}_1) + I(\mathbf{v}_1;\mathbf{y}|\mathbf{v}_2)$$

Hypothesis:

(H1) $I(\mathbf{y};\mathbf{z}_1|\mathbf{v}_2\mathbf{v}_1) = 0$

Thesis:

(T_1) $I(\mathbf{v}_1;\mathbf{y}|\mathbf{z}_1) \leq I(\mathbf{v}_1;\mathbf{v}_2|\mathbf{z}_1) + I(\mathbf{v}_1;\mathbf{y}|\mathbf{v}_2)$

Proof. Since $\mathbf{z}_1$ is a representation of $\mathbf{v}_1$:

(C_1) $I(\mathbf{y};\mathbf{z}_1|\mathbf{v}_2\mathbf{v}_1) = 0$

Therefore:

$$\begin{aligned}
I(\mathbf{v}_1;\mathbf{y}|\mathbf{z}_1) &\overset{(P_3)}{=} I(\mathbf{v}_1;\mathbf{y}|\mathbf{z}_1\mathbf{v}_2) + I(\mathbf{v}_1;\mathbf{v}_2;\mathbf{y}|\mathbf{z}_1) \\
&\overset{(P_3)}{=} I(\mathbf{v}_1;\mathbf{y}|\mathbf{v}_2) - I(\mathbf{v}_1;\mathbf{y};\mathbf{z}_1|\mathbf{v}_2) + I(\mathbf{v}_1;\mathbf{v}_2;\mathbf{y}|\mathbf{z}_1) \\
&\overset{(P_3)}{=} I(\mathbf{v}_1;\mathbf{y}|\mathbf{v}_2) - I(\mathbf{y};\mathbf{z}_1|\mathbf{v}_2) + I(\mathbf{y};\mathbf{z}_1|\mathbf{v}_2\mathbf{v}_1) + I(\mathbf{v}_1;\mathbf{v}_2;\mathbf{y}|\mathbf{z}_1) \\
&\overset{(P_1)}{\leq} I(\mathbf{v}_1;\mathbf{y}|\mathbf{v}_2) + I(\mathbf{y};\mathbf{z}_1|\mathbf{v}_2\mathbf{v}_1) + I(\mathbf{v}_1;\mathbf{v}_2;\mathbf{y}|\mathbf{z}_1) \\
&\overset{(H_1)}{=} I(\mathbf{v}_1;\mathbf{y}|\mathbf{v}_2) + I(\mathbf{v}_1;\mathbf{v}_2;\mathbf{y}|\mathbf{z}_1) \\
&\overset{(P_3)}{=} I(\mathbf{v}_1;\mathbf{y}|\mathbf{v}_2) + I(\mathbf{v}_1;\mathbf{v}_2|\mathbf{z}_1) - I(\mathbf{v}_1;\mathbf{v}_2|\mathbf{z}_1\mathbf{y}) \\
&\overset{(P_1)}{\leq} I(\mathbf{v}_1;\mathbf{y}|\mathbf{v}_2) + I(\mathbf{v}_1;\mathbf{v}_2|\mathbf{z}_1)
\end{aligned}$$

□

Proposition B.3. *Let $\mathbf{v}_1$ be a redundant view with respect to $\mathbf{v}_2$ for $\mathbf{y}$. Any representation $\mathbf{z}_1$ of $\mathbf{v}_1$ that is sufficient for $\mathbf{v}_2$ is also sufficient for $\mathbf{y}$.*

Hypothesis:

(H1) $I(\mathbf{y};\mathbf{z}_1|\mathbf{v}_2\mathbf{v}_1) = 0$

(H_2) $I(\mathbf{y};\mathbf{v}_1|\mathbf{v}_2) = 0$

Thesis:

(T_1) $I(\mathbf{v}_1;\mathbf{v}_2|\mathbf{z}_1) = 0 \implies I(\mathbf{v}_1;\mathbf{y}|\mathbf{z}_1) = 0$

Proof. Using the results from Theorem B.2:

$$I(\mathbf{v}_1;\mathbf{y}|\mathbf{z}_1) \overset{(Th_{B.2})}{\leq} I(\mathbf{v}_1;\mathbf{y}|\mathbf{v}_2) + I(\mathbf{v}_1;\mathbf{v}_2|\mathbf{z}_1) \overset{(H_2)}{=} I(\mathbf{v}_1;\mathbf{v}_2|\mathbf{z}_1)$$

Therefore $I(\mathbf{v}_1;\mathbf{v}_2|\mathbf{z}_1) = 0 \implies I(\mathbf{v}_1;\mathbf{y}|\mathbf{z}_1) = 0$ □

Theorem B.2. *Let* $\mathbf{v}_1$, $\mathbf{v}_2$ *and* $\mathbf{y}$ *be random variables with distribution* $p(\mathbf{v}_1, \mathbf{v}_2, \mathbf{y})$. *Let* $\mathbf{z}$ *be a representation of* $\mathbf{v}_1$, *then*

$$I(\mathbf{y}; \mathbf{z}_1) \geq I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2) - I(\mathbf{v}_1; \mathbf{v}_2|\mathbf{z}_1) - I(\mathbf{v}_1; \mathbf{y}|\mathbf{v}_2) - I(\mathbf{v}_2; \mathbf{y}|\mathbf{v}_1)$$

Hypothesis:

(H_1) $I(\mathbf{y}; \mathbf{z}_1|\mathbf{v}_1\mathbf{v}_2) = 0$

Thesis:

(T_1) $I(\mathbf{y}; \mathbf{z}_1) \geq I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2) - I(\mathbf{v}_1; \mathbf{v}_2|\mathbf{z}_1) - I(\mathbf{v}_1; \mathbf{y}|\mathbf{v}_2) - I(\mathbf{v}_2; \mathbf{y}|\mathbf{v}_1)$

Proof.

$$\begin{aligned}
I(\mathbf{y}; \mathbf{z}_1) &\overset{(P_3)}{=} I(\mathbf{y}; \mathbf{z}_1|\mathbf{v}_1\mathbf{v}_2) + I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2; \mathbf{z}_1) \\
&\overset{(H_1)}{=} I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2; \mathbf{z}_1) \\
&\overset{(P_3)}{=} I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2) - I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2|\mathbf{z}_1) \\
&\overset{(P_2)}{=} I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2) - I(\mathbf{y}; \mathbf{v}_1|\mathbf{z}_1) - I(\mathbf{y}; \mathbf{v}_2|\mathbf{z}_1\mathbf{v}_1) \\
&\overset{(P_3)}{=} I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2) - I(\mathbf{y}; \mathbf{v}_1|\mathbf{z}_1) - I(\mathbf{y}; \mathbf{v}_2|\mathbf{v}_1) + I(\mathbf{y}; \mathbf{v}_2; \mathbf{z}_1|\mathbf{v}_1) \\
&\overset{(P_3)}{=} I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2) - I(\mathbf{y}; \mathbf{v}_1|\mathbf{z}_1) - I(\mathbf{y}; \mathbf{v}_2|\mathbf{v}_1) + I(\mathbf{y}; \mathbf{z}_1|\mathbf{v}_1) - I(\mathbf{y}; \mathbf{z}_1|\mathbf{v}_1\mathbf{v}_2) \\
&\overset{(H_1)}{=} I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2) - I(\mathbf{y}; \mathbf{v}_1|\mathbf{z}_1) - I(\mathbf{y}; \mathbf{v}_2|\mathbf{v}_1) + I(\mathbf{y}; \mathbf{z}_1|\mathbf{v}_1) \\
&\overset{(P_1)}{\geq} I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2) - I(\mathbf{y}; \mathbf{v}_1|\mathbf{z}_1) - I(\mathbf{y}; \mathbf{v}_2|\mathbf{v}_1) \\
&\overset{(Prop_{B.2})}{\geq} I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2) - I(\mathbf{v}_1; \mathbf{y}|\mathbf{v}_2) - I(\mathbf{v}_1; \mathbf{v}_2|\mathbf{z}_1) - I(\mathbf{y}; \mathbf{v}_2|\mathbf{v}_1)
\end{aligned}$$

□

Corollary B.2.1. *Let* $\mathbf{v}_1$ *and* $\mathbf{v}_2$ *be mutually redundant views for* $\mathbf{y}$. *Let* $\mathbf{z}_1$ *be a representation of* $\mathbf{v}_1$ *that is sufficient for* $\mathbf{v}_2$. *Then:*

$$I(\mathbf{y}; \mathbf{z}_1) = I(\mathbf{v}_1\mathbf{v}_2; \mathbf{y})$$

Hypothesis:

(H_1) $I(\mathbf{y}; \mathbf{z}_1|\mathbf{v}_1\mathbf{v}_2) = 0$

(H_2) $I(\mathbf{y}; \mathbf{v}_1|\mathbf{v}_2) + I(\mathbf{y}; \mathbf{v}_2|\mathbf{v}_1) = 0$

(H_3) $I(\mathbf{v}_2; \mathbf{v}_1|\mathbf{z}) = 0$

Thesis:

(T_1) $I(\mathbf{y}; \mathbf{z}_1) = I(\mathbf{v}_1\mathbf{v}_2; \mathbf{y})$

Proof. Using Theorem B.2

$$\begin{aligned}
I(\mathbf{y}; \mathbf{z}_1) &\overset{(Th_{B.2})}{\geq} I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2) - I(\mathbf{v}_1; \mathbf{y}|\mathbf{v}_2) - I(\mathbf{v}_1; \mathbf{v}_2|\mathbf{z}_1) - I(\mathbf{y}; \mathbf{v}_2|\mathbf{v}_1) \\
&\overset{(H_2)}{=} I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2) - I(\mathbf{v}_1; \mathbf{v}_2|\mathbf{z}_1) \\
&\overset{(H_3)}{=} I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2)
\end{aligned}$$

Since $I(\mathbf{y}; \mathbf{z}_1) \leq I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2)$ is a consequence of the data processing inequality, we conclude that $I(\mathbf{y}; \mathbf{z}_1) = I(\mathbf{y}; \mathbf{v}_1\mathbf{v}_2)$ □

B.4 SUFFICIENCY AND AUGMENTATION

Let $\mathbf{x}$ and $\mathbf{y}$ be random variables with domain X and Y respectively. Let T be a class of functions $t : \mathsf{X} \to \mathsf{W}$ and let $\mathbf{t}_1$ and $\mathbf{t}_2$ be a random variables over T that depends only on $\mathbf{x}$. For the theorems and corollaries discussed in this section, we are going to consider the independence assumption that can be derived from the graphical model $\mathcal{G}$ reported in Figure 5.

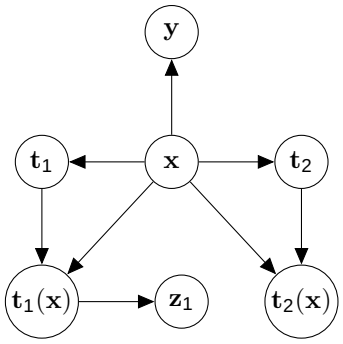


Figure 5: Visualization of the graphical model $\mathcal{G}$ that relates the observations $\mathbf{x}$, label $\mathbf{y}$, functions used for augmentation $\mathbf{t}_1$, $\mathbf{t}_2$ and the representation $\mathbf{z}_1$.

Proposition B.4. *Whenever* $I(\mathbf{t}_1(\mathbf{x}); \mathbf{y}) = I(\mathbf{t}_2(\mathbf{x}); \mathbf{y}) = I(\mathbf{x}; \mathbf{y})$ *the two views* $\mathbf{t}_1(\mathbf{x})$ *and* $\mathbf{t}_2(\mathbf{x})$ *must be mutually redundant for* $\mathbf{y}$.

Hypothesis:

(H_1) *Independence relations determined by* $\mathcal{G}$

Thesis:

$$(T_1)\ \ I(\mathbf{t}_1(\mathbf{x}); \mathbf{y}) = I(\mathbf{t}_2(\mathbf{x}); \mathbf{y}) = I(\mathbf{x}; \mathbf{y}) \implies I(\mathbf{t}_1(\mathbf{x}); \mathbf{y} | \mathbf{t}_2(\mathbf{x})) + I(\mathbf{t}_2(\mathbf{x}); \mathbf{y} | \mathbf{t}_1(\mathbf{x})) = 0$$

Proof.

1. Considering $\mathcal{G}$ we have:

   $(C_1)\ \ I(\mathbf{t}_1(\mathbf{x}); \mathbf{y} | \mathbf{x}\mathbf{t}_2(\mathbf{x})) = 0$

   $(C_2)\ \ I(\mathbf{y}; \mathbf{t}_2(\mathbf{x}) | \mathbf{x}) = 0$

2. Since $\mathbf{t}_2(\mathbf{x})$ is uniquely determined by $\mathbf{x}$ and $\mathbf{t}_2$:

   $(C_3)\ \ I(\mathbf{t}_2(\mathbf{x}); \mathbf{y} | \mathbf{x}\mathbf{t}_2) = 0$

3. Consider $I(\mathbf{t}_1(\mathbf{x}); \mathbf{y} | \mathbf{t}_2(\mathbf{x}))$

$$
\begin{aligned}
I(\mathbf{t}_1(\mathbf{x}); \mathbf{y} | \mathbf{t}_2(\mathbf{x})) &\overset{(P_3)}{=} I(\mathbf{t}_1(\mathbf{x}); \mathbf{y} | \mathbf{x}\mathbf{t}_2(\mathbf{x})) + I(\mathbf{t}_1(\mathbf{x}); \mathbf{y}; \mathbf{x} | \mathbf{t}_2(\mathbf{x})) \\
&\overset{(C_1)}{=} I(\mathbf{t}_1(\mathbf{x}); \mathbf{y}; \mathbf{x} | \mathbf{t}_2(\mathbf{x})) \\
&\overset{(P_3)}{=} I(\mathbf{y}; \mathbf{x} | \mathbf{t}_2(\mathbf{x})) - I(\mathbf{y}; \mathbf{x} | \mathbf{t}_1(\mathbf{x})\mathbf{t}_2(\mathbf{x})) \\
&\overset{(P_1)}{\leq} I(\mathbf{y}; \mathbf{x} | \mathbf{t}_2(\mathbf{x})) \\
&\overset{(P_3)}{=} I(\mathbf{y}; \mathbf{x}) - I(\mathbf{y}; \mathbf{x}; \mathbf{t}_2(\mathbf{x})) \\
&\overset{(P_3)}{=} I(\mathbf{y}; \mathbf{x}) - I(\mathbf{y}; \mathbf{t}_2(\mathbf{x})) + I(\mathbf{y}; \mathbf{t}_2(\mathbf{x}) | \mathbf{x}) \\
&\overset{(P_3)}{=} I(\mathbf{y}; \mathbf{x}) - I(\mathbf{y}; \mathbf{t}_2(\mathbf{x})) + I(\mathbf{y}; \mathbf{t}_2(\mathbf{x}) | \mathbf{t}_2\mathbf{x}) + I(\mathbf{y}; \mathbf{t}_2(\mathbf{x}); \mathbf{t}_2 | \mathbf{x}) \\
&\overset{(C_3)}{=} I(\mathbf{y}; \mathbf{x}) - I(\mathbf{y}; \mathbf{t}_2(\mathbf{x})) + I(\mathbf{y}; \mathbf{t}_2(\mathbf{x}); \mathbf{t}_2 | \mathbf{x}) \\
&\overset{(P_3)}{=} I(\mathbf{y}; \mathbf{x}) - I(\mathbf{y}; \mathbf{t}_2(\mathbf{x})) + I(\mathbf{y}; \mathbf{t}_2(\mathbf{x}) | \mathbf{x}) - I(\mathbf{y}; \mathbf{t}_2(\mathbf{x}) | \mathbf{t}_2\mathbf{x}) \\
&\overset{(P_1)}{\geq} I(\mathbf{y}; \mathbf{x}) - I(\mathbf{y}; \mathbf{t}_2(\mathbf{x})) + I(\mathbf{y}; \mathbf{t}_2(\mathbf{x}) | \mathbf{x}) \\
&\overset{(C_2)}{\geq} I(\mathbf{y}; \mathbf{x}) - I(\mathbf{y}; \mathbf{t}_2(\mathbf{x}))
\end{aligned}
$$

Therefore $I(\mathbf{y}; \mathbf{x}) = I(\mathbf{y}; \mathbf{t}_2(\mathbf{x})) \implies I(\mathbf{t}_1(\mathbf{x}); \mathbf{y} | \mathbf{t}_2(\mathbf{x})) = 0$

The proof for $I(\mathbf{y}; \mathbf{x}) = I(\mathbf{y}; \mathbf{t}_1(\mathbf{x})) \implies I(\mathbf{t}_2(\mathbf{x}); \mathbf{y} | \mathbf{t}_1(\mathbf{x})) = 0$ is symmetric, therefore we conclude $I(\mathbf{t}_1(\mathbf{x}); \mathbf{y}) = I(\mathbf{t}_2(\mathbf{x}); \mathbf{y}) = I(\mathbf{x}; \mathbf{y}) \implies I(\mathbf{t}_1(\mathbf{x}); \mathbf{y} | \mathbf{t}_2(\mathbf{x})) + I(\mathbf{t}_2(\mathbf{x}); \mathbf{y} | \mathbf{t}_1(\mathbf{x})) = 0$ □

Theorem B.3. *Let $I(\mathbf{t}_1(\mathbf{x}); \mathbf{y}) = I(\mathbf{t}_2(\mathbf{x}); \mathbf{y}) = I(\mathbf{x}; \mathbf{y})$. Let $\mathbf{z}_1$ be a representation of $\mathbf{t}_1(\mathbf{x})$. If $\mathbf{z}_1$ is sufficient for $\mathbf{t}_2(\mathbf{x})$ then $I(\mathbf{x}; \mathbf{y}) = I(\mathbf{y}; \mathbf{z}_1)$.*

Hypothesis:

(H_1) *Independence relations determined by $\mathcal{G}$*

(H_2) $I(\mathbf{t}_1(\mathbf{x}); \mathbf{y}) = I(\mathbf{t}_2(\mathbf{x}); \mathbf{y}) = I(\mathbf{x}; \mathbf{y})$

Thesis:

(T_1) $I(\mathbf{t}_1(\mathbf{x}); \mathbf{t}_2(\mathbf{x}) | \mathbf{z}_1) = 0 \implies I(\mathbf{x}; \mathbf{y}) = I(\mathbf{y}; \mathbf{z}_1)$

Proof. Since $\mathbf{t}_1(\mathbf{x})$ is redundant for $\mathbf{t}_2(\mathbf{x})$ (Proposition B.4) any representation $\mathbf{z}_1$ of $\mathbf{t}_1(\mathbf{x})$ that is sufficient for $\mathbf{t}_2(\mathbf{x})$ must also be sufficient for $\mathbf{y}$ (Theorem B.2). Using Proposition B.1 we have $I(\mathbf{y}; \mathbf{z}_1) = I(\mathbf{y}; \mathbf{t}_1(\mathbf{x}))$. Since $I(\mathbf{y}; \mathbf{t}_1(\mathbf{x})) = I(\mathbf{y}; \mathbf{x})$ by hypothesis, we conclude $I(\mathbf{x}; \mathbf{y}) = I(\mathbf{y}; \mathbf{z}_1)$ □

C EQUIVALENCES OF DIFFERENT OBJECTIVES

Different objectives in literature can be seen as a special case of the Multi-View Information Bottleneck principle. In this section we show that the supervised version of Information Bottleneck is equivalent to the corresponding Multi-View version whenever the two redundant views have only label information in common. A second subsection show equivalence between InfoMax and Multi-View Information Bottleneck whenever the two views are identical.

C.1 MULTI-VIEW INFORMATION BOTTLENECK AND SUPERVISED INFORMATION BOTTLENECK

Whenever the two mutually redundant views $\mathbf{v_1}$ and $\mathbf{v_2}$ have only label information in common (or when one of the two views is the label itself) the Multi-View Information Bottleneck objective is equivalent to the respective supervised version. This can be shown by proving that $I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{v_2}) = I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{y})$, i.e. a representation $\mathbf{z_1}$ of $\mathbf{v_1}$ that is sufficient and minimal for $\mathbf{v_2}$ is also sufficient and minimal for $\mathbf{y}$.

Proposition C.1. *Let $\mathbf{v_1}$ and $\mathbf{v_2}$ be mutually redundant views for a label $\mathbf{y}$ that share only label information. Then a sufficient representation $\mathbf{z_1}$ of $\mathbf{v_1}$ for $\mathbf{v_2}$ that is minimal for $\mathbf{v_2}$ is also a minimal representation for $\mathbf{y}$.*

Hypothesis:

(H_1) $I(\mathbf{v_1}; \mathbf{y}|\mathbf{v_2}) + I(\mathbf{v_2}; \mathbf{y}|\mathbf{v_1}) = 0$

(H_2) $I(\mathbf{v_1}; \mathbf{v_2}) = I(\mathbf{v_1}; \mathbf{v}_y)$

(H_3) $I(\mathbf{v_1}; \mathbf{v_2}|\mathbf{z_1}) = 0$

Thesis:

(T_1) $I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{v_2}) = I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{y})$

Proof.

1. Consider $I(\mathbf{v_1}; \mathbf{z})$:

$$\begin{aligned} I(\mathbf{v_1}; \mathbf{z_1}) &\overset{(P_3)}{=} I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{v_2}) + I(\mathbf{v_1}; \mathbf{v_2}; \mathbf{z_1}) \\ &\overset{(P_3)}{=} I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{v_2}) + I(\mathbf{v_1}; \mathbf{v_2}) - I(\mathbf{v_1}; \mathbf{v_2}|\mathbf{z_1}) \\ &\overset{(H_3)}{=} I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{v_2}) + I(\mathbf{v_1}; \mathbf{v_2}) \\ &\overset{(H_1)}{=} I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{v_2}) + I(\mathbf{v_1}; \mathbf{y}) \end{aligned}$$

2. Using Corollary 1, from (H_2) and (H_3) follows $I(\mathbf{v_1}; \mathbf{y}|\mathbf{z_1}) = 0$

3. $I(\mathbf{v_1}; \mathbf{z})$ can be alternatively expressed as:

$$\begin{aligned} I(\mathbf{v_1}; \mathbf{z_1}) &\overset{(P_3)}{=} I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{y}) + I(\mathbf{v_1}; \mathbf{y}; \mathbf{z_1}) \\ &\overset{(P_3)}{=} I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{y}) + I(\mathbf{v_1}; \mathbf{y}) - I(\mathbf{v_1}; \mathbf{y}|\mathbf{z_1}) \\ &\overset{(Cor_1)}{=} I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{y}) + I(\mathbf{v_1}; \mathbf{y}) \end{aligned}$$

Equating 1 and 3, we conclude $I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{v_2}) = I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{y})$. □

C.2 MULTI-VIEW INFORMATION BOTTLENECK AND INFOMAX

Whenever $\mathbf{v_1} = \mathbf{v_2}$, a representation $\mathbf{z_1}$ of $\mathbf{v_1}$ that is sufficient for $\mathbf{v_1}$ must contain all the original information. Furthermore since $I(\mathbf{v_1}; \mathbf{z_1}|\mathbf{v_1}) = 0$ for every representation, no superfluous information can be identified and removed. As a consequence, a minimal sufficient representation $\mathbf{z_1}$ of $\mathbf{v_1}$ for $\mathbf{v_1}$ is any representation for which mutual information is maximal, hence InfoMax.

D LOSS COMPUTATION

Starting from Equation 3, we consider the sum of the losses $\mathcal{L}_1(\theta;\lambda_1)$ and $\mathcal{L}_2(\psi;\lambda_2)$ that aim to create the minimal sufficient representations $\mathbf{z_1}$ and $\mathbf{z_2}$ respectively:

$$\mathcal{L}_{1+2}(\theta,\psi;\lambda_1,\lambda_2) = (I_\theta(\mathbf{v_1};\mathbf{z_1}|\mathbf{v_2}) + I_\psi(\mathbf{v_2};\mathbf{z_2}|\mathbf{v_1})) + (\lambda_1 I_\theta(\mathbf{v_1};\mathbf{v_2}|\mathbf{z_1}) + \lambda_2 I_\psi(\mathbf{v_1};\mathbf{v_2}|\mathbf{z_1})) \tag{6}$$

Considering $\mathbf{z_1}$ and $\mathbf{z_2}$ on the same domain Z, $I_\theta(\mathbf{v_1};\mathbf{z_1}|\mathbf{v_2})$ can be expressed as:

$$\begin{aligned}
I_\theta(\mathbf{v_1};\mathbf{z_1}|\mathbf{v_2}) &= \mathsf{E}_{\boldsymbol{v}_1,\boldsymbol{v}_2 \;\; p(\mathbf{v_1},\mathbf{v_2})}\mathsf{E}_{\boldsymbol{z} \;\; p_\theta(\mathbf{z_1} j \mathbf{v_1})}\left[\log \frac{p_\theta(\mathbf{z_1}=\boldsymbol{z}|\mathbf{v_1}=\boldsymbol{v_1})}{p_\theta(\mathbf{z_1}=\boldsymbol{z}|\mathbf{v_2}=\boldsymbol{v_2})}\right] \\
&= \mathsf{E}_{\boldsymbol{v}_1,\boldsymbol{v}_2 \;\; p(\mathbf{v_1},\mathbf{v_2})}\mathsf{E}_{\boldsymbol{z} \;\; p_\theta(\mathbf{z_1} j \mathbf{v_1})}\left[\log \frac{p_\theta(\mathbf{z_1}=\boldsymbol{z}|\mathbf{v_1}=\boldsymbol{v_1})}{p_\psi(\mathbf{z_2}=\boldsymbol{z}|\mathbf{v_2}=\boldsymbol{v_2})}\frac{p_\psi(\mathbf{z_2}=\boldsymbol{z}|\mathbf{v_2}=\boldsymbol{v_2})}{p_\theta(\mathbf{z_1}=\boldsymbol{z}|\mathbf{v_2}=\boldsymbol{v_2})}\right] \\
&= D_{\mathsf{KL}}(p_\theta(\mathbf{z_1}|\mathbf{v_1})||p_\psi(\mathbf{z_2}|\mathbf{v_2})) - D_{\mathsf{KL}}(p_\theta(\mathbf{z_2}|\mathbf{v_1})||p_\psi(\mathbf{z_2}|\mathbf{v_2})) \\
&\leq D_{\mathsf{KL}}(p_\theta(\mathbf{z_1}|\mathbf{v_1})||p_\psi(\mathbf{z_2}|\mathbf{v_2}))
\end{aligned}$$

Note that the bound is tight whenever $p_\psi(\mathbf{z_2}|\mathbf{v_2})$ coincides with $p_\theta(\mathbf{z_1}|\mathbf{v_2})$. This happens whenever $\mathbf{z_1}$ and $\mathbf{z_2}$ produce a consistent encoding. Analogously $I_\psi(\mathbf{v_2};\mathbf{z_2}|\mathbf{v_1})$ is upper bounded by $D_{\mathsf{KL}}(p_\psi(\mathbf{z_2}|\mathbf{v_2})||p_\theta(\mathbf{z_1}|\mathbf{v_1}))$.

$I_\theta(\mathbf{v_1};\mathbf{v_2}|\mathbf{z_1})$ can be rephrased as:

$$\begin{aligned}
I_\theta(\mathbf{v_1};\mathbf{v_2}|\mathbf{z_1}) &= I(\mathbf{v_1};\mathbf{v_2}) - I_\theta(\mathbf{z_1};\mathbf{v_2}) \\
&\overset{(P_2)}{=} I(\mathbf{v_1};\mathbf{v_2}) - I_\theta(\mathbf{z_1};\mathbf{z_2}\mathbf{v_2}) - I_\theta(\mathbf{z_1};\mathbf{z_2}|\mathbf{v_2}) \\
&= \;\; I(\mathbf{v_1};\mathbf{v_2}) - I_\theta(\mathbf{z_1};\mathbf{z_2}\mathbf{v_2}) \\
&= I(\mathbf{v_1};\mathbf{v_2}) - I_\theta(\mathbf{z_1};\mathbf{z_2}) - I_{\theta\psi}(\mathbf{z_1};\mathbf{v_2}|\mathbf{z_2}) \\
&\leq I(\mathbf{v_1};\mathbf{v_2}) - I_{\theta\psi}(\mathbf{z_1};\mathbf{z_2})
\end{aligned}$$

Where $\overset{(P_2)}{=}$ follows from $\mathbf{z_2}$ representation of $\mathbf{v_2}$. The bound reported in this equation is tight whenever $\mathbf{z_2}$ is sufficient for $\mathbf{z_1}$. This happens whenever $\mathbf{z_2}$ contains all the information regarding $\mathbf{z_1}$ (and therefore $\mathbf{v_1}$). Once again, the same bound can symmetrically be used to define $I_\theta(\mathbf{v_1};\mathbf{v_2}|\mathbf{z_2}) \leq I(\mathbf{v_1};\mathbf{v_2}) - I_{\theta\psi}(\mathbf{z_1};\mathbf{z_2})$.

Since $I(\mathbf{v_1};\mathbf{v_2})$ is constant in θ and ψ, the loss function in Equation 6 can be upper-bounded with;

$$\mathcal{L}_{1+2}(\theta,\psi;\lambda_1,\lambda_2) \leq 2D_{SKL}(p_\theta(\mathbf{z_1}|\mathbf{v_1})||p_\psi(\mathbf{z_2}|\mathbf{v_2})) - (\lambda_1+\lambda_2)I_{\theta\psi}(\mathbf{z_1};\mathbf{z_2}) \tag{7}$$

Where:

$$D_{SKL}(p_\theta(\mathbf{z_1}|\mathbf{v_1})||p_\psi(\mathbf{z_2}|\mathbf{v_2})) := \frac{1}{2}D_{\mathsf{KL}}(p_\theta(\mathbf{z_1}|\mathbf{v_1})||p_\psi(\mathbf{z_2}|\mathbf{v_2})) + \frac{1}{2}D_{\mathsf{KL}}(p_\psi(\mathbf{z_2}|\mathbf{v_2})||p_\theta(\mathbf{z_1}|\mathbf{v_1}))$$

Lastly, multiplying both terms with $\beta := \frac{2}{\lambda_1+\lambda_2}$ and re-parametrizing the objective, we obtain:

$$\mathcal{L}_{MIB}(\theta,\psi;\beta) = -I_{\theta\psi}(\mathbf{z_1};\mathbf{z_2}) + \beta\, D_{SKL}(p_\theta(\mathbf{z_1}|\mathbf{v_1})||p_\psi(\mathbf{z_2}|\mathbf{v_2})) \tag{8}$$

E INFORMATION PLANE

Every representation $\mathbf{z}$ of $\mathbf{x}$ must satisfy the following constraints:

- $0 \leq I(\mathbf{y};\mathbf{z}) \leq I(\mathbf{x};\mathbf{y})$: The amount of label information ranges from 0 to the total predictive information accessible from the raw observations $I(\mathbf{x};\mathbf{y})$.
- $I(\mathbf{y};\mathbf{z}) \leq I(\mathbf{x};\mathbf{z}) \leq I(\mathbf{y};\mathbf{z}) + H(\mathbf{x}|\mathbf{y})$: The representation must contain more information about the observations than about the label. When $\mathbf{x}$ is discrete, the amount of discarded label information $I(\mathbf{x};\mathbf{y}) - I(\mathbf{y};\mathbf{z})$ must be smaller than the amount of discarded observational information $H(\mathbf{x}) - I(\mathbf{x};\mathbf{z})$, which implies $I(\mathbf{x};\mathbf{z}) \leq I(\mathbf{y};\mathbf{z}) + H(\mathbf{x}|\mathbf{y})$.

Proof. Since $\mathbf{z}$ is a representation of $\mathbf{x}$:

(C_1) $I(\mathbf{y};\mathbf{z}|\mathbf{x}) = 0$

Considering the four bounds separately:

1. $I(\mathbf{y};\mathbf{z}) \geq 0$: Follows from P_1
2. $I(\mathbf{x};\mathbf{z}) \geq I(\mathbf{y};\mathbf{z})$: Follows from (T_1) in Lemma B.1
3. $I(\mathbf{y};\mathbf{z}) \leq I(\mathbf{y};\mathbf{x})$: Data processing inequality

$$\begin{aligned} I(\mathbf{y};\mathbf{z}) &\overset{(P_3)}{=} I(\mathbf{y};\mathbf{z}|\mathbf{x}) + I(\mathbf{y};\mathbf{z};\mathbf{x}) \\ &\overset{(C_1)}{=} I(\mathbf{y};\mathbf{z};\mathbf{x}) \\ &\overset{(P_3)}{=} I(\mathbf{x};\mathbf{y}) - I(\mathbf{x};\mathbf{y}|\mathbf{z}) \\ &\overset{(P_1)}{\leq} I(\mathbf{x};\mathbf{y}) \end{aligned}$$

4. $I(\mathbf{x};\mathbf{z}) \leq I(\mathbf{y};\mathbf{z}) + H(\mathbf{x}|\mathbf{y})$:

$$\begin{aligned} I(\mathbf{x};\mathbf{z}) &\overset{(P_3)}{=} I(\mathbf{x};\mathbf{z}|\mathbf{y}) + I(\mathbf{x};\mathbf{y};\mathbf{z}) \\ &\overset{(P_3)}{=} I(\mathbf{x};\mathbf{z}|\mathbf{y}) + I(\mathbf{y};\mathbf{z}) - I(\mathbf{y};\mathbf{z}|\mathbf{x}) \\ &\overset{(C_1)}{=} I(\mathbf{x};\mathbf{z}|\mathbf{y}) + I(\mathbf{y};\mathbf{z}) \\ &\overset{(H_2+P_4)}{\leq} I(\mathbf{x};\mathbf{z}|\mathbf{y}) + H(\mathbf{x}|\mathbf{y}\mathbf{z}) + I(\mathbf{y};\mathbf{z}) \\ &\overset{(P_5)}{=} H(\mathbf{x}|\mathbf{y}) + I(\mathbf{y};\mathbf{z}) \end{aligned}$$

Note that (H_2) is needed only to prove bound 4. For continuous $\mathbf{x}$ bounds 1, 2 and 3 still hold. □

F EXPERIMENTAL PROCEDURE AND DETAILS

F.1 MODELING

The two stochastic encoders $p_\theta(\mathbf{z}_1|\mathbf{v}_1)$ and $p_\psi(\mathbf{z}_2|\mathbf{v}_2)$ are modeled by Normal distributions parametrized with neural networks $(\boldsymbol{\mu}_\theta, \boldsymbol{\sigma}^2_\theta)$ and $(\boldsymbol{\mu}_\psi, \boldsymbol{\sigma}^2_\psi)$ respectively:

$$\begin{aligned} p_\theta(\mathbf{z}_1|\mathbf{v}_1) &:= \mathcal{N}\left(\mathbf{z}_1|\boldsymbol{\mu}_\theta(\mathbf{v}_1), \boldsymbol{\sigma}^2_\theta(\mathbf{v}_1)\right) \\ p_\psi(\mathbf{z}_2|\mathbf{v}_2) &:= \mathcal{N}\left(\mathbf{z}_2|\boldsymbol{\mu}_\psi(\mathbf{v}_2), \boldsymbol{\sigma}^2_\psi(\mathbf{v}_2)\right) \end{aligned}$$

Since the density of the two encoders can be evaluated, the symmetrized KL-divergence in equation 4 can be directly computed. On the other hand, $I_{\theta\psi}(\mathbf{z}_1;\mathbf{z}_2)$ requires the use of a mutual information estimator.

To facilitate the optimization, the hyper-parameter β is slowly increased during training, starting from a small value $\approx 10^{-4}$ to its final value with an exponential schedule. This is because the mutual information estimator is trained together with the other architectures and, since it starts from a random initialization, it requires an initial warm-up. Starting with bigger β results in the encoder collapsing into a fixed representation. The update policy for the hyper-parameter during training has not shown strong influence on the representation, as long as the mutual information estimator network has reached full capacity.

All the experiments have been performed using the Adam optimizer with a learning rate of 10^{-4} for both encoders and the estimation network. Higher learning rate can result in instabilities in the training procedure. The results reported in the main text relied on the Jensen-Shannon mutual information estimator (Devon Hjelm et al., 2019) since the InfoNCE counterpart (van den Oord et al., 2018) generally resulted in worse performances that could be explained by the effect of the factorization of the critic network (Poole et al., 2019).

F.2 SKETCHY EXPERIMENTS

- **Input:** The two views for the sketch-based classification task consist of 4096 dimensional sketch and image features extracted from two distinct VGG-16 network models which were pre-trained on images and sketches from the TU-Berlin dataset Eitz et al. (2012) for end-to-end classification. The feature extractors are frozen during the training procedure of for the two representations. Each training iteration used batches of size $B = 128$.
- **Encoder and Critic architectures:** Both sketch and image encoders consist of multi-layer perceptrons of 2 hidden ReLU units of size 2,048 and 1,024 respectively with an output of size 2x64 that parametrizes mean and variance for the two Gaussian posteriors. The critic architecture also consists of a multi layer perceptron of 2 hidden ReLU units of size 512.
- β **update policy:** The initial value of β is set to 10^{-4}. Starting from the 10,000$^{\text{th}}$ training iteration, the value of β is exponentially increased up to 1.0 during the following 250,000 training iterations. The value of β is then kept fixed to one until the end of the training procedure (500,000 iterations).
- **Evaluation:** All natural images are used as both training sets and retrieval galleries. The 64 dimensional real outputs of sketch and image representation are compared using Euclidean distance. For having a fair comparison other methods that rely on binary hashing (Liu et al., 2017; Zhang et al., 2018), we used Hamming distance on a binarized representation (obtained by applying iterative quantization Gong et al. (2013) on our real valued representation). We report the mean average precision (mAP@all) and precision at top-rank 200 (Prec@200) Su et al. (2015) on both the real and binary representation to evaluate our method and compare it with prior works.

F.3 MIR-FLICKR EXPERIMENTS

Figure 6: Examples of pictures $\mathbf{v}_1$, tags $\mathbf{v}_2$ and category labels $\mathbf{y}$ for the MIR-Flickr dataset (Srivastava & Salakhutdinov, 2014).

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

- **Input:** Whitening is applied to the handcrafted image features. Batches of size $B = 128$ are used for each update step.
- **Encoders and Critic architectures:** The two encoders consists of a multi layer perceptron of 4 hidden ReLU units of size 1,024, which exactly resemble the architecture used in Wang et al. (2016). Both representations $\mathbf{z}_1$ and $\mathbf{z}_2$ have a size of 1,024, therefore the two architecture output a total of 2x1,024 parameters that define mean and variance of the respective factorized Gaussian posterior. Similarly to the Sketchy experiments, the critic is consists of a multi-layer perceptron of 2 hidden ReLU units of size 512.
- β **update policy:** The initial value of β is set to 10^{-8}. Starting from 150000$^{\text{th}}$ iteration, β is set to exponentially increase up to 1.0 (and 10^{-3}) during the following 150,000 iterations.
- **Evaluation:** Once the models are trained on the *unlabeled* set, the representation of the 25,000 *labeled* images is computed. The resulting vectors are used for training and evaluating a multi-label logistic regression classifier on the respective splits. The optimal parameters (such as β) for our model are chosen based on the performance on the validation set. In Table 3, we report the aggregated mean of the 5 test splits as the final value mean average precision value.

F.4 MNIST EXPERIMENTS

- **Input:** The two views $\mathbf{v}_1$ and $\mathbf{v}_2$ for the MNIST dataset are generated by applying small translation ([0-10]%), rotation ([-15,15] degrees), scale ([90,110]%), shear ([-15,15] degrees) and pixel corruption (20%). Batches of size $B = 64$ samples are used during training.

- **Encoders, Decoders and Critic architectures:** All the encoders used for the MNIST experiments consist of neural networks with two hidden layers of 1,024 units and ReLU activations, producing a 2x64-dimensional parameter vector that is used to parameterize mean and variance for the Gaussian posteriors. The decoders used for the VAE experiments also consist of the networks of the same size. Similarly, the critic architecture used for mutual information estimation consists of two hidden layers of 1,204 units each and ReLU activations.

- β **update policy:** The initial value of β is set to 10^{3}, which is increased with an exponential schedule starting from the 50,000$^{\text{th}}$ until 1the 50,000$^{\text{th}}$ iteration. The value of β is then kept constant until the 1,000,000$^{\text{th}}$ iteration. The same annealing policy is used to trained the different β-VAEs reported in this work.

- **Evaluation:** The trained representation are evaluated following the well-known protocol described in Tschannen et al. (2019); Tian et al. (2019); Bachman et al. (2019); van den Oord et al. (2018). Each logistic regression is trained 5 different balanced splits of the training set for different percentages of training examples, ranging from 1 example per label to the whole training set. The accuracy reported in this work has been computed on the disjoint test set. Mean and standard deviation are computed according to the 5 different subsets used for training the logistic regression. Mean and variance for the mutual information estimation reported on the Information Plane (Figure 4) are computed by training two estimation networks from scratch on the final representation of the non-augmented train set. The two estimation architectures consist of 2 hidden layers of 2048 and 1024 units each, and have been trained with batches of size $B = 256$ for a total of approximately 25,000 iterations. The Jensen-Shannon mutual information lower bound is maximized during training, while the numerical estimation are computed using an energy-based bound (Poole et al., 2019; Devon Hjelm et al., 2019). The final values for $I(\mathbf{x}; \mathbf{z})$ and $I(\mathbf{y}; \mathbf{z})$ are computed by averaging the mutual information estimation on the whole dataset. In order to reduce the variance of the estimator, the lowest and highest 5% are removed before averaging. This practical detail makes the estimation more consistent and less susceptible to numerical instabilities.

G ABLATION STUDIES

G.1 DIFFERENT RANGES OF DATA AUGMENTATION

Figure 7 visualizes the effect of different ranges of corruption probabily as data augmentation strategy to produce the two views $\mathbf{v}_1$ and $\mathbf{v}_2$. The MV-InfoMax Model does not seem to get any advantage from the use increasing amount of corruption, and it representation remains approximately in the same region of the information plane. On the other hand, the models trained with the MIB objective are able to take advantage of the augmentation to remove irrelevant data information and the representation transitions from the top right corner of the Information Plane (no-augmentation) to the top-left. When the amount of corruption approaches 100%, the mutual redundancy assumption is clearly violated, and the performances of MIB deteriorate. In the initial part of the transitions between the two regimes (which corresponds to extremely low probability of corruption) the MIB models drops some label information that is quickly re-gained when pixel corruption becomes more frequent. We hypothesize that this behavior is due to a problem with the optimization procedure, since the corruption are extremely unlikely, the Monte-Carlo estimation for the symmetrized Kullback-Leibler divergence is more biased. Using more examples of views produced from the same data-point within the same batch could mitigate this issue.

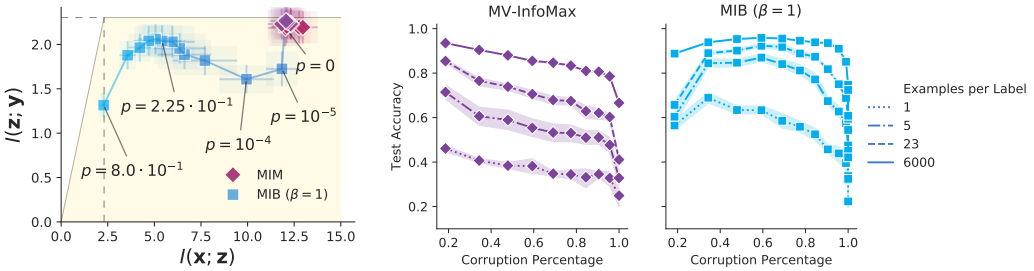


Figure 7: Visualization of the coordinates on the Information Plane (plot on the left) and prediction accuracy (center and right) for the MV-InfoMax and MIB objectives with different amount of training labels and corruption percentage used for data-augmentation.

G.2 Effect of β

The hyper-parameter β (Equation 5) determines the trade-off between sufficiency and minimality of the representation for the second data view. When β is zero, the training objective of MIB is equivalent to the Multi-View InfoMax target, since the representation has no incentive to discard any information. When $0 < \beta \leq 1$ the sufficiency constrain is enforced, while the superfluous information is gradually removed from the representation. Values of $\beta > 1$ can result in representations that violate the sufficiency constraint, since the minimization of $I(\mathbf{x}; \mathbf{z}|\mathbf{v}_2)$ is prioritized. The trade-off resulting from the choice of different β is visualized in Figure 8 and compared against β-VAE. Note that in each point of the pareto-front the MIB model results in a better trade-off between $I(\mathbf{x}; \mathbf{z})$ and $I(\mathbf{y}; \mathbf{z})$ when compared to β-VAE. The effectiveness of the Multi-View Information Bottleneck model is also justified by the corresponding values of predictive accuracy.

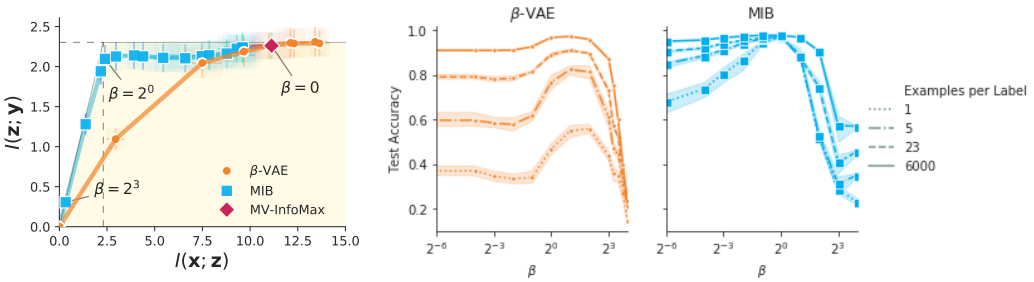


Figure 8: Visualization of the coordinates on the Information Plane (plot on the left) and prediction accuracy (center and right) for the β-VAE, Multi-View InfoMax and Multi-View Information Bottleneck objectives with different amount of training labels and different values of the respective hyperparameter β.