SELF-SUPERVISED LEARNING FROM A MULTI-VIEW PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised representation learning adopts self-defined signals as supervision and uses the learned representation for downstream tasks, such as contrastive visual representation learning (e.g., SimCLR) for computer vision applications. In this paper, we present to understand self-supervised learning via a multi-view perspective. In particular, we regard the input (e.g., original images) and the self-supervised signals (e.g., augmented images) as two corresponding views of the data. Then, we leverage the multi-view redundant assumption that different views provide redundant information for the downstream tasks (e.g., image augmentation will not affect image's label). Under this assumption, we contribute to 1) provide a better understanding of self-supervised learning by demonstrating the self-supervised learned representation can extract task-relevant information (with a potential loss) and discard task-irrelevant information (with a fixed gap); 2) pave a larger space of composing self-supervised learning objectives such as bridging the gap between prior contrastive/ predictive objectives and designing a new objective to discard task-irrelevant information. We conduct controlled experiments to support our analysis empirically. The experiments are studied in two angles: 1) from visual-only to visual-textual representation learning; and 2) from the scenario that the multi-view assumption holds to the scenario that the assumption is likely to fail.

1 INTRODUCTION

Self-supervised learning (SSL) (Zhang et al., 2016; Devlin et al., 2018; Oord et al., 2018; Tian et al., 2019) learns representations using a proxy objective (i.e., SSL objective) between inputs and self-defined signals. Empirical evidence suggests that the learned representations can generalize well to a wide range of downstream tasks, even when the SSL objective has not utilize any downstream supervision during training. For example, SimCLR (Chen et al., 2020) defines a contrastive loss (i.e., an SSL objective) between images with different augmentations (i.e., one as the input and the other as the self-supervised signal). Then, one can take SimCLR as features extractor and adopt the features to various computer vision applications, spanning image classification, object detection, instance segmentation, and pose estimation (He et al., 2019). Despite showing success in practice, there are only a few work (Arora et al., 2019; Lee et al., 2020; Tosh et al., 2020) providing theoretical insights into SSL. Our work shares a similar goal to explain the success of SSL, from the perspectives of Information Theory (Cover & Thomas, 2012) and multi-view representation[1].

To understand (a subset[2] of) SSL, we start by the following *multi-view assumption*. First, we regard the input and the self-supervised signals as two corresponding views of the data. Using our running example, in SimCLR (Chen et al., 2020), applying image augmentation can be seen as changing the view of an image. Hence, we can regard augmented images (i.e., the input and the self-supervised signal) as images with different views. Second, we adopt a common assumption in multi-view learning: either view alone is (approximately) sufficient for the downstream tasks (see Assumption 1 in prior work (Sridharan & Kakade, 2008)). The assumption suggests that the image augmentations (e.g., changing the style of an image) should not affect the labels of images, or analogously, the self-supervised signal contains most (perhaps not all) of the information that the input has about

[1]The work (Lee et al., 2020; Tosh et al., 2020) are done concurrent and in parallel, and part of their assumptions/ conclusions are similar to ours. We will elaborate the differences more in the related work section.

[2]We discuss the limitations of the multi-view assumption in Section 2.1.

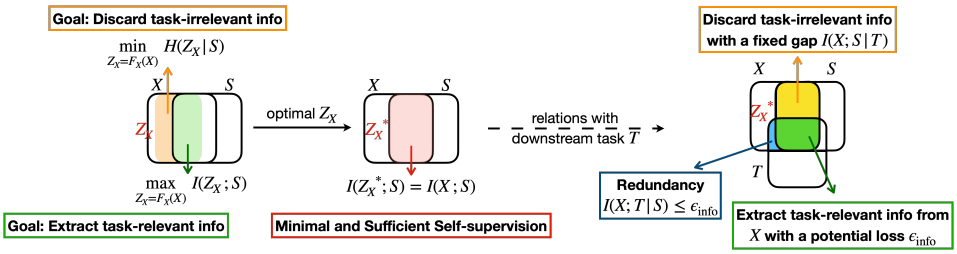

Figure 1: High-level takeaways for our main results using information diagrams. (a) We present to learn minimal and sufficient self-supervision: minimize $H(Z_X|S)$ for discarding task-irrelevant information and maximize $I(Z_X;S)$ for extracting task-relevant information. (b) The resulting learned representation $Z_X{}^*$ contains all task relevant information from the input with a potential loss ϵ_{info} and discards task-irrelevant information with a fixed gap $I(X;S|T)$. (c) Our core assumption: the self-supervised signal is approximately redundant to the input for the task-relevant information.

the downstream tasks. With this assumption, our first contribution is to formally show that the self-supervised learned representations can 1) extract all the task-relevant information (from the input) with a potential loss; and 2) discard all the task-irrelevant information (from the input) with a fixed gap. Then, using classification task as an example, we are able the quantify the smallest generalization error (Bayes error rate) given the discussed task-relevant and task-irrelevant information.

As the second contribution, our analysis 1) connects prior arts for SSL on contrastive (Oord et al., 2018; Bachman et al., 2019; Chen et al., 2020; Tian et al., 2019) and predictive learning (Zhang et al., 2016; Vondrick et al., 2016; Tulyakov et al., 2018; Devlin et al., 2018) approaches; and 2) paves the way to a larger space of composing SSL objectives to extract task-relevant and discard task-irrelevant information simultaneously. For instance, the combination between the contrastive and predictive learning approaches achieves better performance than contrastive- or predictive-alone objective and enjoys less over-fitting problem. We also present a new objective to discard task-irrelevant information. The objective can be easily incorporated with prior self-supervised learning objectives.

We conduct controlled experiments on visual (the first set) and visual-textual (the second set) self-supervised representation learning. The first set of experiments are performed when the multi-view assumption is likely to hold. The goal is to compare different compositions of SSL objectives on extracting task-relevant and discarding task-irrelevant information. The second set of experiments is to challenge the multi-view assumption by having the input and the self-supervised signal lie in very different modalities. Under this cross-modality setting, the task-relevant information may not mostly lie in the shared information between the input and the self-supervised signal. The goal is to examine SSL objectives' generalization, where the multi-view assumption is likely to fail.

2 A Multi-view Information-Theoretical Framework

Notations. For the input, we denote its random variable as X, sample space as $\mathcal{X}$, and outcome as x. We learn a representation (Z_X/ $\mathcal{Z}$/ z_x) from the input through a deterministic mapping F_X: $Z_X = F_X(X)$. For the self-supervised signal, we denote its random variable/ sample space/ outcome as S/ $\mathcal{S}$/ s. Two sample spaces can be different between the input and the self-supervised signal: $\mathcal{X} \neq \mathcal{S}$. The information required for downstream tasks is referred to as "task-relevant information": T/ $\mathcal{T}$/ t. Note that SSL has no access to the task-relevant information. Lastly, we use $I(A;B)$ to represent mutual information, $I(A;B|C)$ to represent conditional mutual information, $H(A)$ to represent the entropy, and $H(A|B)$ to represent conditional entropy for random variables A/B/C. We provide high-level takeaways for our main results in Figure 1. We defer all proofs in Supplementary.

2.1 Multi-view Assumption

In our paper, we regard the input (X) and the self-supervised signals (S) as two views of the data. Here, we provide a table showing different X/S in various SSL frameworks:

Framework	BERT (Devlin et al., 2018)	Look & Listen (Arandjelovic & Zisserman, 2017)	SimCLR (Chen et al., 2020)	Colorization (Zhang et al., 2016)
Inputs (X)	Non-masked Words	Image	Image	Image Lightness
Self-supervised Signals (S)	Masked Words	Audio Stream	Same Image with Augmentation	Image Color

We note that not all SSL frameworks realize the inputs and the self-supervised signals as corresponding views. For instance, Jigsaw puzzles (Noroozi & Favaro, 2016) considers (shuffled) image patches as the input and the positions of the patches as the self-supervised signals. Another example is Learning by Predicting Rotations (Gidaris et al., 2018), which considers an image (rotating with a specific angle) as the input and the rotation angle of the image as the self-supervised signal. We point out

that the frameworks that regard X/S as two corresponding views (Chen et al.; 2020; He et al., 2019) have a much better empirical downstream performance than the frameworks that do not (Noroozi & Favaro, 2016; Gidaris et al., 2018). Our paper hence focuses on the multi-view setting between X/S.

Next, we adopt the common assumption (i.e., *multi-view assumption* (Sridharan & Kakade, 2008; Xu et al., 2013)) in the multi-view learning between the input and the self-supervised signal:

Assumption 1 (Multi-view, restating Assumption 1 in prior work (Sridharan & Kakade, 2008))**.** *The self-supervised signal is approximately redundant to the input for the task-relevant information. In other words, there exist an* $\epsilon_{\text{info}} > 0$ *such that* $I(X;T|S) \leq \epsilon_{\text{info}}$.

Assumption 1 states that, when ϵ_{info} is small, the task-relevant information lies mostly in the shared information between the input and the self-supervised signals. Jusitfication is as follows. For self-supervised visual contrastive learning (Hjelm et al., 2018; Chen et al., 2020), the input and the self-supervised signal are the same image with different augmentations. Using image augmentations can be seen as changing the style of an image while not affecting the content. And we argue that the information required for downstream tasks should only be retained in the content but not the style. Last, we point out the failure cases of the assumption (or have large ϵ_{info}): the input and the self-supervised signal contain very different task-relevant information. For instance, a drastic image augmentation (e.g., adding large noise) may change the content of the image (e.g., the noise completely occludes the objects). Another example is BERT (Devlin et al., 2018), with too much masking, the downstream information may exist differently in the masked (i.e., the self-supervised signals) and the non-masked (i.e., the input) words. Analogously, too much masking makes the non-masked words have no enough context to predict the masked words (Devlin et al., 2018).

2.2 Learning Minimal and Sufficient Representations for Self-supervision

We start by discussing the supervised representation learning. The Information Bottleneck (IB) method (Tishby et al., 2000; Achille & Soatto, 2018) generalizes minimal sufficient statistics to the representations that are minimal (i.e., less complexity) and sufficient (i.e., better fidelity). To learn such representations for downstream supervision, we consider the following objectives:

Definition 1 (Minimal and Sufficient Representations for Downstream Supervision)**.** Let Z_X^{sup} be the sufficient supervised representation and $Z_X^{\text{sup}_{\text{min}}}$ be the minimal and sufficient representation:

$$Z_X^{\text{sup}} = \arg\max_{Z_X} I(Z_X;T) \text{ and } Z_X^{\text{sup}_{\text{min}}} = \arg\min_{Z_X} H(Z_X|T) \text{ s.t. } I(Z_X;T) \text{ is maximized.}$$

To reduce the complexity of the representation Z_X, the prior methods (Tishby et al., 2000; Achille & Soatto, 2018) presented to minimize $I(Z_X;X)$ while ours presents to minimize $H(Z_X|T)$. We provide a justification: minimizing $H(Z_X|T)$ reduces the randomness from T to Z_X, and the randomness is regarded as the incompressibility (Calude, 2013). Hence, minimizing $H(Z_X|T)$ leads to a more compressed representation (discarding redundant information)[3]. Note that we do not constrain T, which can be classification, regression, or clustering.

Then, we present SSL objectives to learn sufficient (and minimal) representations for self-supervision:

Definition 2 (Minimal and Sufficient Representations for Self-supervision)**.** Let Z_X^{ssl} be the sufficient self-supervised representation and $Z_X^{\text{ssl}_{\text{min}}}$ be the minimal and sufficient representation:

$$Z_X^{\text{ssl}} = \arg\max_{Z_X} I(Z_X;S) \text{ and } Z_X^{\text{ssl}_{\text{min}}} = \arg\min_{Z_X} H(Z_X|S) \text{ s.t. } I(Z_X;S) \text{ is maximized.}$$

Definition 2 defines our self-supervised representation learning strategy. Now, we are ready to associate the supervised and self-supervised learned representations:

Theorem 1 (Task-relevant information with a potential loss ϵ_{info})**.** *The supervised learned representations (i.e.,* $I(Z_X^{\text{sup}};T)$ *and* $I(Z_X^{\text{sup}_{\text{min}}};T)$*) contain all the task-relevant information in the input (i.e.,* $I(X;T)$*). The self-supervised learned representations (i.e.,* $I(Z_X^{\text{ssl}};T)$ *and* $I(Z_X^{\text{ssl}_{\text{min}}};T)$*) contain all the task-relevant information in the input with a potential loss* ϵ_{info}*. Formally,*

$$I(X;T) = I(Z_X^{\text{sup}};T) = I(Z_X^{\text{sup}_{\text{min}}};T) \geq I(Z_X^{\text{ssl}};T) \geq I(Z_X^{\text{ssl}_{\text{min}}};T) \geq I(X;T) - \epsilon_{\text{info}}.$$

[3] We do not claim $H(Z_X|T)$ minimization is better than $I(Z_X;X)$ minimization for reducing the complexity in the representations Z_X. In Supplementary, we will show that $H(Z_X|T)$ minimization and $I(Z_X;X)$ minimization are interchangeable under our framework's setting.

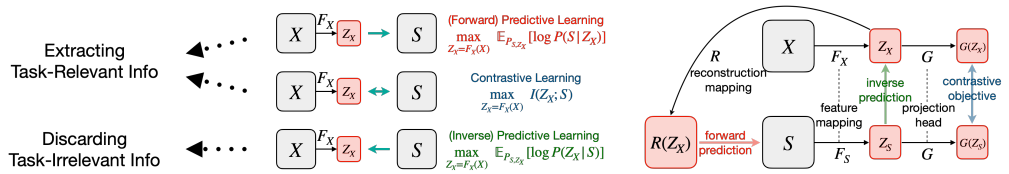

Figure 2: Remarks on contrastive and predictive learning objectives for self-supervised learning. Between the representation Z_X and the self-supervised signal S, *contrastive objective* performs mutual information maximization and *predictive objectives* perform log conditional likelihood maximization. We show that the SSL objectives aim at extracting task-relevant and discarding task-irrelevant information. Last, we summarize the computational blocks for practical deployments for these objectives.

When ϵ_{info} is small, Theorem 1 indicates that the self-supervised learned representations can extract as much task-relevant information as the supervised one. While when ϵ_{info} is not small, the self-supervised learned representations may not always lead to good downstream performance. This result has also been observed in prior work (Tschannen et al., 2019) and InfoMin (Tian et al., 2020), which claim the representations with maximal mutual information may not have the best performance.

Theorem 2 (Task-irrelevant information with a fixed compression gap $I(X;S|T)$)**.** *The sufficient self-supervised representation (i.e., $I(Z_X^{\text{ssl}};T)$) contains more task-irrelevant information in the input than then the sufficient and minimal self-supervised representation (i.e., $I(Z_X^{\text{ssl}_{\min}};T)$). The later contains an amount of the information, $I(X;S|T)$, that cannot be discarded from the input. Formally,*

$$I(Z_X^{\text{ssl}};X|T) = I(X;S|T) + I(Z_X^{\text{ssl}};X|S,T) \geq I(Z_X^{\text{ssl}_{\min}};X|T) = I(X;S|T) \geq I(Z_X^{\text{sup}_{\min}};X|T) = 0.$$

Theorem 2 indicates that a compression gap (i.e., $I(X;S|T)$) exists when we discard the task-irrelevant information from the input. To be specific, $I(X;S|T)$ is the amount of the shared information between the input and the self-supervised signal excluding the task-relevant information. Hence, $I(X;S|T)$ would be large if the downstream tasks requires only a portion of the shared information.

2.3 Connections with Contrastive and Predictive Learning Objectives

Theorem 1 and 2 state that our self-supervised learning strategies (i.e., $\min H(Z_X|S)$ and $\max I(Z_X;S)$ defined in Definition 2) can extract task-relevant and discard task-irrelevant information. A research question now emerges:

"*What are the practical aspects of the presented self-supervised learning strategies?*"

To answer this question, we present 1) the connections with prior SSL objectives, especially for contrastive (Oord et al., 2018; Bachman et al., 2019; Chen et al., 2020; Tian et al., 2019; Hjelm et al., 2018; He et al., 2019) and predictive (Zhang et al., 2016; Pathak et al., 2016; Vondrick et al., 2016; Tulyakov et al., 2018; Peters et al., 2018; Devlin et al., 2018) learning objectives, showing that these objectives are extracting task-relevant information; and 2) a new *inverse predictive learning* objective to discard task-irrelevant information. We illustrate important remarks in Figure 2.

Contrastive Learning (is extracting task-relevant information). We define the contrastive learning objective as maximizing the dependency/contrastiveness between the learned representation Z_X and the self-supervised signal S, which suggests maximizing the the mutual information $I(Z_X;S)$. Given Theorem 1, we know that maximizing $I(Z_X;S)$ results in Z_X containing (approximately) all the information required for the downstream tasks from the input X. To deploy the contrastive learning objective, we suggest contrastive predictive coding (CPC) (Oord et al., 2018)[4], which is a mutual information lower bound with low variance (Poole et al., 2019; Song & Ermon, 2019):

$$\text{L}_{CL} := \max_{\substack{Z_S = F_S(S),\\ Z_X = F_X(X), G}} \mathbb{E}_{(z_{s1},z_{x1}),\cdots,(z_{sn},z_{xn})\sim P^n(Z_S,Z_X)} \left[\frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{\langle G(z_{xi}),G(z_{si})\rangle}}{\frac{1}{n}\sum_{j=1}^{n} e^{\langle G(z_{xi}),G(z_{sj})\rangle}} \right], \quad (1)$$

where $F_S : \mathcal{S} \to \mathcal{Z}$ is a deterministic mapping and G is a project head that projects a representation in $\mathcal{Z}$ into a lower-dimensional vector. If the input and self-supervised signals share the same

[4]Other contrastive learning objectives can be other mutual information lower bounds such as DV-bound or NWJ-bound (Belghazi et al., 2018) or its JS-divergence (Poole et al., 2019; Hjelm et al., 2018) variants. Among different objectives, Tschannen et al. (2019) have suggested that the objectives with large variance (e.g., DV-/NWJ-bound (Belghazi et al., 2018)) may lead to worsen performance compared to the low variance counterparts (e.g., CPC (Oord et al., 2018) and JS-div. (Poole et al., 2019)).

sample space, i.e., $\mathcal{X} = \mathcal{S}$, we can impose $F_X = F_S$ (e.g., self-supervised visual representation learning (Chen et al., 2020)). The projection head, G, can be an identity, a linear, or a non-linear mapping. Last, we note that modeling equation 1 often requires a large batch size (e.g., large n in equation 1) to ensure a good downstream performance (He et al., 2019; Chen et al., 2020).

Forward Predictive Learning (is extracting task-relevant information). We define the forward predictive learning as encouraging the learned representation Z_X to reconstruct the self-supervised signal S, which suggests maximizing the log conditional likelihood $\mathbb{E}_{P_{S,Z_X}}[\log P(S|Z_X)]$. By the chain rule, $I(Z_X;S) = H(S) - H(S|Z_X)$, where $H(S)$ is irrelevant to Z_X. Hence, maximizing $I(Z_X;S)$ is equivalent to maximizing $-H(S|Z_X) = \mathbb{E}_{P_{S,Z_X}}[\log P(S|Z_X)]$, which is the predictive learning objective. Together with Theorem 1, if z_x can perfectly reconstruct s for any $(s, z_x) \sim P_{S,Z_X}$, then Z_X contains (approximately) all the information required for the downstream tasks from the input X. A common approach to avoid intractability in computing $\mathbb{E}_{P_{S,Z_X}}[\log P(S|Z_X)]$ is assuming a variational distribution $Q_\phi(S|Z_X)$ with ϕ representing the parameters in $Q_\phi(\cdot|\cdot)$. Specifically, we present to maximize $\mathbb{E}_{P_{S,Z_X}}[\log Q_\phi(S|Z_X)]$, which is a lower bound of $\mathbb{E}_{P_{S,Z_X}}[\log P(S|Z_X)]$[5]. $Q_\phi(\cdot|\cdot)$ can be any distribution such as Gaussian or Laplacian and ϕ can be a linear model, a kernel method, or a neural network. Note that the choice of the reconstruction type of loss depends on the distribution type of $Q_\phi(\cdot|\cdot)$, not a fixed one. If we let $Q_\phi(S|Z_X)$ be Gaussian $\mathcal{N}\Big(S|R(Z_X), \sigma\mathbf{I}\Big)$ with $\sigma\mathbf{I}$ as a diagonal matrix[6], the objective becomes:

$$\mathrm{L}_{FP} := \max_{Z_X = F_X(X), R} \mathbb{E}_{s, z_x \sim P_{S,Z_X}} \Big[- \|s - R(z_x)\|_2^2 \Big], \tag{2}$$

where $R : \mathcal{Z} \to \mathcal{S}$ is a deterministic mapping to reconstruct S from Z and we ignore the constants derived from the Gaussian distribution. Last, in most real-world applications, the self-supervised signal S has a much higher dimension than the representation Z_X. Hence, modeling a conditional generative model $Q_\phi(S|Z_X)$ will be challenging. For example, considering S as $224 \times 224 \times 3$ image and Z_X as $64-$dimensional vector.

Inverse Predictive Learning (is discarding task-irrelevant information). We define the inverse predictive learning as encouraging the self-supervised signal S to reconstruct the learned representation Z_X, which suggests maximizing the log conditional likelihood $\mathbb{E}_{P_{S,Z_X}}[\log P(Z_X|S)]$. Given Theorem 2 together with $-H(Z_X|S) = \mathbb{E}_{P_{S,Z_X}}[\log P(Z_X|S)]$, we know if s can perfectly reconstruct z_x for any $(s, z_x) \sim P_{S,Z_X}$ under the constraint that $I(Z_X;S)$ is maximized, then Z_X discards the task-irrelevant information, excluding $I(X;S|T)$. Similar to the forward predictive learning, we use $\mathbb{E}_{P_{S,Z_X}}[\log Q_\phi(Z_X|S)]$ as a lower bound of $\mathbb{E}_{P_{S,Z_X}}[\log P(Z_X|S)]$. In our deployment, we take the advantage of the design in equation 1 and let $Q_\phi(Z_X|S)$ be Gaussian $\mathcal{N}\Big(Z_X|F_S(S), \sigma\mathbf{I}\Big)$:

$$\mathrm{L}_{IP} := \max_{Z_S = F_S(S), Z_X = F_X(X)} \mathbb{E}_{z_s, z_x \sim P_{Z_S,Z_X}} \Big[- \|z_x - z_s\|_2^2 \Big]. \tag{3}$$

Note that optimizing equation 3 alone results in a degenerated solution, e.g., learning Z_X and Z_S to be the same constant.

Composing SSL Objectives (to extract task-relevant and discard task-irrelevant information simultaneously). We have connected the SSL strategy presented in Definition 2 to contrastive and predictive learning objectives. Bringing their practical aspects together (equation 1, equation 2, and equation 3), we can pave the way to a larger space of composing SSL objectives:

$$\mathrm{L}_{SSL} = \lambda_{CL}\mathrm{L}_{CL} + \lambda_{FP}\mathrm{L}_{FP} + \lambda_{IP}\mathrm{L}_{IP}, \tag{4}$$

where λ_{CL}, λ_{FP}, and λ_{IP} are hyper-parameters.

[5] $\mathbb{E}_{P_{S,Z_X}}[\log P(S|Z_X)] = \max_{Q_\phi} \mathbb{E}_{P_{S,Z_X}}[\log Q_\phi(S|Z_X)] + D_{\mathrm{KL}}\Big(P(S|Z_X) \,\|\, Q_\phi(S|Z_X)\Big) \geq \max_{Q_\phi} \mathbb{E}_{P_{S,Z_X}}[\log Q_\phi(S|Z_X)]$.

[6] The assumption of identity covariance in the Gaussian is only a particular parameterization of the distribution $Q(\cdot|\cdot)$. Other examples are MocoGAN (Tulyakov et al., 2018), which assumes Q is Laplacian (i.e., ℓ_1 reconstruction loss) and ϕ is a deconvolutional network (Long et al., 2015). Transformer-XL (Dai et al., 2019) assumes Q is a categorical distribution (i.e., cross entropy loss) and ϕ is a Transformer network (Vaswani et al., 2017). Although Gaussian with diagonal covariance is not the best assumption, it is perhaps the simplest one.

2.4 THEORETICAL ANALYSIS - BAYES ERROR RATE FOR DOWNSTREAM CLASSIFICATION

So far, we see the practical aspects of our designed self-supervised learning strategies. Now, we provide an theoretical analysis on the representations' generalization error when T is a categorical variable . We use Bayes error rate as an example, which stands for the irreducible error (smallest generalization error (Feder & Merhav, 1994)) when learning an arbitrary classifier from the representation to infer the labels. In specific, let P_e be the Bayes error rate of arbitrary learned representations Z_X and $\hat{T}$ as the estimation for T from our classifier, $P_e := \mathbb{E}_{z_x \sim P_{Z_X}}[1 - \max_{t \in T} P(\hat{T} = t | z_x)]$.

To begin with, we present a general form of sample complexity with mutual information ($I(Z_X; S)$) estimation using empirical samples from distribution $P_{Z_X,S}$. Let $P^{(n)}_{Z_X,S}$ denote the (uniformly sampled) empirical distribution of $P_{Z_X,S}$ and $\hat{I}^{(n)}_{\theta}(Z_X; S) := \mathbb{E}_{P^{(n)}_{Z_X;S}}[\hat{f}_\theta(z_x, s)]$ with $\hat{f}_\theta$ being the estimated log density ratio (i.e., $\log p(s|z_x)/p(s)$).

Proposition 1 (Mutual Information Neural Estimation, restating Theorem 1 by Tsai et al. (2020))**.** *Let $0 < \delta < 1$. There exists $d \in \mathbb{N}$ and a family of neural networks $\mathcal{F} := \{\hat{f}_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ where Θ is compact, so that $\exists \theta^* \in \Theta$, with probability at least $1 - \delta$ over the draw of $\{z_{xi}, s_i\}_{i=1}^n \sim P^{\otimes n}_{Z_X,S}$,* $\left|\widehat{I}^{(n)}_{\theta^*}(Z_X; S) - I(Z_X; S)\right| \leq O\left(\sqrt{\frac{d + \log(1/\delta)}{n}}\right)$.

This proposition shows that there exists a neural network θ^*, with high probability, $\widehat{I}^{(n)}_{\theta^*}(Z_X; S)$ can approximate $I(Z_X; S)$ with n samples at rate $O(1/\sqrt{n})$. Under this network θ^* and the same parameters d and δ, we are ready to present our main results on the Bayes error rate. Formally, let $|T|$ be T's cardinalitiy and $\mathrm{Th}(x) = \min\{\max\{x, 0\}, 1 - 1/|T|\}$ as a thresholding function:

Theorem 3 (Bayes Error Rates for Arbitrary Learned Representations)**.** *For an arbitrary learned representations Z_X, $P_e = \mathrm{Th}(\bar{P}_e)$ with*

$$\bar{P}_e \leq 1 - \exp\Big(- \Big(H(T) + I(X; S|T) + I(Z; X|S, T) - \hat{I}^{(n)}_{\theta^*}(Z_X; S) + O\big(\sqrt{\frac{d + \log(1/\delta)}{n}}\big)\Big)\Big).$$

Given arbitrary learned representations (Z_X), Theorem 3 suggests the corresponding Bayes error rate (P_e) is small when: 1) the estimated mutual information $\big(\widehat{I}^{(n)}_{\theta^*}(Z_X; S)\big)$ is large; 2) a larger number of samples n are used for estimating the mutual information; and 3) the task-irrelevant information $\big($the compression gap $I(X; S|T)$ and the superfluous information $I(Z; X|S, T)$, defined in Theorem 2$\big)$ is small. The first and the second results supports the claim that maximizing $I(Z_X; S)$ may learn the representations that are beneficial to downstream tasks. The third result implies the learned representations may perform better on the downstream task when the compression gap is small. Additionally, $Z^{\mathrm{ssl}_{\min}}$ is preferable than Z^{ssl} since $I(Z^{\mathrm{ssl}_{\min}}; X|S, T) = 0$ and $I(Z^{\mathrm{ssl}}; X|S, T) \geq 0$.

Theorem 4 (Bayes Error Rates for Self-supervised Learned Representations)**.** *Let $P^{\mathrm{sup}}_e / P^{\mathrm{ssl}}_e / P^{\mathrm{ssl}_{\min}}_e$ be the Bayes error rate of the supervised or the self-supervised learned representations $Z^{\mathrm{sup}}_X / Z^{\mathrm{ssl}}_X / Z^{\mathrm{ssl}_{\min}}_X$. Then, $P^{\mathrm{ssl}}_e = \mathrm{Th}(\bar{P}^{\mathrm{ssl}}_e)$ and $P^{\mathrm{ssl}_{\min}}_e = \mathrm{Th}(\bar{P}^{\mathrm{ssl}_{\min}}_e)$ with*

$$-\frac{\log(1 - P^{\mathrm{sup}}_e) + \log 2}{\log(|T|)} \leq \{\bar{P}^{\mathrm{ssl}}_e, \bar{P}^{\mathrm{ssl}_{\min}}_e\} \leq 1 - \exp\Big(- (\log 2 + P^{\mathrm{sup}}_e \cdot \log|T| + \epsilon_{\mathrm{info}})\Big).$$

Given our self-supervised learned representations (Z^{ssl}_X and $Z^{\mathrm{ssl}_{\min}}_X$), Theorem 4 suggests a smaller upper bound of P^{ssl}_e (or $P^{\mathrm{ssl}_{\min}}_e$) when the redundancy between the input and the self-supervised signal (ϵ_{info}, defined in Assumption 1) is small. This result implies the self-supervised learned representations may perform better on the downstream task when the multi-view redundancy is small.

3 CONTROLLED EXPERIMENTS

This section aims at providing empirical supports for Theorems 1/ 2 and comparing different SSL objectives. We design two sets of experiments: The first one is when the input and self-supervised signals lie in the same modality (visual) and are likely to satisfy the multi-view redundancy assumption (Assumption 1). The second one is when the input and self-supervised signals lie in very different modalities (visual and textual), thus challenging the SSL objective's generalization ability.

(a) Omniglot (Composing SSL Objectives)

Objective	Trained for	Test Accuracy
L_{CL}	500 epochs	$85.59 \pm 0.05\%$
$\mathrm{L}_{CL} + \mathrm{L}_{IP}$	500 epochs	$85.90 \pm 0.09\%$
L_{FP}	20000 epochs	$84.36 \pm 0.09\%$
$\mathrm{L}_{FP} + \mathrm{L}_{IP}$	20000 epochs	$84.51 \pm 0.08\%$
$\mathrm{L}_{CL} + 0.005\mathrm{L}_{FP}$	800 epochs	$85.77 \pm 0.17\%$
$\mathrm{L}_{CL} + 0.005\mathrm{L}_{FP} + \mathrm{L}_{IP}$	800 epochs	$85.92 \pm 0.08\%$

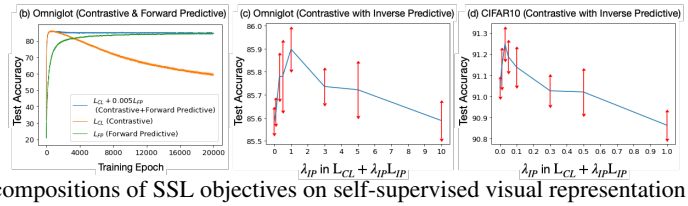

Figure 3: Comparisons for different compositions of SSL objectives on self-supervised visual representation training. We report mean and its standard error from 5 random trials.

Experiment I - Visual Representation Learning. We use Omniglot dataset (Lake et al., 2015) [7] in this experiment. The training set contains images from 964 characters, and the test set contains 659 characters. There are no characters overlap between the training and test set. Each character contains twenty examples drawn from twenty different people. We regard image as input (X) and generate self-supervised signal (S) by first sampling an image from the same character as the input image and then applying translation/ rotation to it. Furthermore, we represent task-relevant information (T) by the labels of the image. Under this self-supervised signal construction, the exclusive information in X or S are drawing styles (i.e., by different people) and image augmentations, and only their shared information contribute to T. To formally show the later, if T representing the label for X/S, then $P(T|X)$ and $P(T|S)$ are Dirac. Hence, $T \perp\!\!\!\perp S|X$ and $T \perp\!\!\!\perp X|S$, suggesting Assumption 1 holds.

We train the feature mapping $F_X(\cdot)$ with SSL objectives (see eq. equation 4), set $F_S(\cdot) = F_X(\cdot)$, let $R(\cdot)$ to be symmetrical to $F_X(\cdot)$, and have $G(\cdot)$ to be an identity mapping. We use $1 - NN$ classification to evaluate the downstream performance of the representations. On the test set, we fix the mapping and randomly select 5 examples per character as the labeled examples. Then, we classify the rest of the examples using the 1-nearest neighbor classifier based on feature (i.e., $Z_X = F_X(X)$) cosine similarity. The random performance on this task stands at $\frac{1}{659} \approx 0.15\%$. One may refer to Supplementary for more details.

▷ *Results & Discussions.* In Figure 3, we evaluate the generalization ability on the test set for different SSL objectives. Figure 3 (a)/(b) suggest that, comparing to L_{FP}, L_{CL} 1) reaches better test accuracy; 2) requires shorter training epochs to reach the best performance; and 3) suffers from overfitting with long-epoch training. Combining both of them ($\mathrm{L}_{CL} + 0.005\mathrm{L}_{FP}$) brings their advantages together. We also find that adding L_{IP} in the objective can boost model performance. According to Theorem 2, the improved performance suggests a more compressed representation results in better performance for the downstream tasks. Nonetheless, in Figure 3 (c), we find the performance is sensitive to the hyper-parameter λ_{IP} for combining L_{IP}. In specific, the performance first improves and later deteriorates as λ_{IP} increases. Last, we show that combining L_{CL} and L_{IP} together can also improve performance in SOTA SSL framework. In Figure 3 (d), we provide experiment with SimCLR (Chen et al., 2020) on CIFAR10 (Krizhevsky et al., 2009) , where $\lambda_{IP} = 0$ refers to the exact same setup as in SimCLR (which considers only L_{CL}).

Experiment II - Visual-Textual Representation Learning. We provide experiments using MS COCO dataset (Lin et al., 2014) that contains 328k multi-labeled images with 2.5 million labeled instances from 91 objects. Each image has 5 annotated captions describing the relationships between objects in the scenes. We regard image as input (X) and its textual descriptions as self-supervised signal (S). Since vision and text are two very different modalities, the multi-view redundancy may not be satisfied, which means ϵ_{info} may be large in Assumption 1.

We adopt L_{CL} (+$\lambda_{IP}\mathrm{L}_{\mathrm{IP}}$) as our SSL objective. We use ResNet18 (He et al., 2016) image encoder for $F_X(\cdot)$ (trained from scratch or fine-tuned on ImageNet (Deng et al., 2009) pre-trained weights), BERT-uncased (Devlin et al., 2018) text encoder for $F_S(\cdot)$ (trained from scratch or BookCorpus (Zhu et al., 2015)/Wikipedia pre-trained weights), and a linear layer for $G(\cdot)$. After performing self-supervised visual-textual representation learning, we consider the downstream multi-label classification over 91 categories. We evaluate learned visual representation (Z_X) using *downstream linear evaluation protocol* (Oord et al., 2018; Hénaff et al., 2019; Tian et al., 2019; Hjelm et al., 2018; Bachman et al., 2019; Tschannen et al., 2019). Specifically, a linear classifier is trained from the self-supervised learned (fixed) representation to the labels on the training set. Commonly used metrics for multi-label classification are reported on MS COCO validation set: Micro ROC-AUC, Hamming Loss, and Subset Accuracy. One may refer to Supplementary for more details on these metrics.

[7] More complex datasets such as CIFAR10 (Krizhevsky et al., 2009) or ImageNet (Deng et al., 2009), to achieve similar performance, require a much larger training scale from contrastive to forward predictive objective. For example, on ImageNet, MoCo (He et al., 2019) uses 8 GPUs for its contrastive objective and ImageGPT (Chen et al.) uses 2048 TPUs for its forward predictive objective. We choose the Omniglot to ensure fair comparisons among different self-supervised learning objectives under reasonable computation constraint.

(a) MS COCO (Using L_{CL} as SSL objective)

Setting	Micro ROC-AUC (↑)	Hamming Loss (↓)	Subset Acc. (↑)
Cross-modality Self-supervised Learning			
Raw BERT + Raw ResNet	0.5963 ± 0.0034	0.0363 ± 0.0003	0.0166 ± 0.0017
Pre-trained BERT + Raw ResNet	0.5915 ± 0.0035	0.0359 ± 0.0002	0.0163 ± 0.0011
Raw BERT + Pre-trained ResNet	0.7049 ± 0.0040	0.0265 ± 0.0003	0.2081 ± 0.0063
Pre-trained BERT + Pre-trained ResNet	0.7065 ± 0.0026	0.0269 ± 0.0004	0.2123 ± 0.0040
Non Self-supervised Learning			
Only Pre-trained ResNet	0.6761 ± 0.0045	0.0277 ± 0.0003	0.1719 ± 0.0015

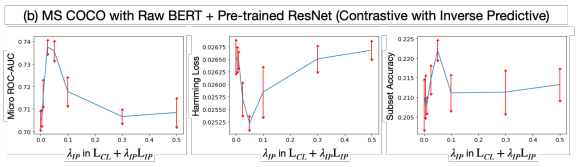

Figure 4: Comparisons for different settings on self-supervised visual-textual representation training. We report metrics on MS COCO validation set with mean and standard deviation from 5 random trials. *Micro ROC-AUC / Subset Accuracy* are the higher the better and *Hamming Loss* is the lower the better.

▷ *Results & Discussions.* First, Figure 4 (a) suggests that the SSL strategy can still work when the input and self-supervised signals lie in different modalities. For example, pre-trained ResNet with BERT (either raw or the pre-trained one) outperforms pre-trained ResNet alone. We also see that the self-supervised learned representations benefit more if the ResNet is pre-trained but not the BERT. This result is in accord with the fact that object recognition requires more understanding in vision, and hence the pre-trained ResNet is preferrable than the pre-trained BERT. Next, Figure 4 (b) suggests that the self-supervised learned representations can be further improved by combining L_{CL} and L_{IP}: $\mathrm{L}_{CL} + \lambda_{IP}\mathrm{L}_{IP}$, suggesting L_{IP} may be a useful objective to discard task-irrelevant information. As a summary, we show that, even under the scenario when the Assumption 1 is likely to fail (i.e., the information required for downstream tasks differs between the input and the self-supervised signal), we can still learn useful representations via self-supervised learning strategies.

4 Related Work

Prior work by Arora et al. (2019) and the recent concurrent work (Lee et al., 2020; Tosh et al., 2020) are landmarks for theoretically understanding the success of SSL. In particular, Arora et al. (2019); Lee et al. (2020) showed a decreased sample complexity for downstream supervised tasks when adopting contrastive learning objectives (Arora et al., 2019) or predicting the known information in the data (Lee et al., 2020). Tosh et al. (2020) showed that the linear functions of the learned representations are nearly optimal on downstream prediction tasks. By viewing the input and the self-supervised signal as two corresponding views of the data, we discuss the differences among these works and ours. On the one hand, the work by Arora et al. (2019); Lee et al. (2020) assume strong independence between the views conditioning on the downstream tasks , i.e., $I(X; S|T) \approx 0$. On the other hand, the work by Tosh et al. (2020) and ours assume strong independence between the downstream task and one view conditioning on the other view, i.e., $I(T; X|S) \approx 0$. Prior work (Balcan et al., 2005; Du et al., 2010) have compared these two assumptions and pointed out the former one ($I(X; S|T) \approx 0$) is too strong and not likely to hold in practice. We note that all these related work and ours have shown the self-supervised learned representations can extract most of the task-relevant information. Our work additionally presents to discard task-irrelevant information and quantifies the amount of information that cannot be discarded.

Our method also resembles the InfoMax principle (Linsker, 1988; Hjelm et al., 2018) and the Multi-view Information Bottleneck method (Federici et al., 2020). The InfoMax principle aims at preserving the information of itself, while ours aims at extracting the information in the self-supervised signal. On the other hand, to reduce the amount of the redundant information that is not required for downstream tasks, the Multi-view Information Bottleneck method proposed to minimize the conditional mutual information $I(Z_X; X|S)$, while ours propose to minimize the conditional entropy $H(Z_X|S)$. We argue that minimizing the conditional entropy is simpler as directly minimizing the conditional mutual information requires solving a min-max optimization problem (Mukherjee et al., 2020).

On the empirical side, we explain why contrastive (Oord et al., 2018; Bachman et al., 2019; Chen et al., 2020) and predictive learning (Zhang et al., 2016; Pathak et al., 2016; Vondrick et al., 2016; Chen et al.) approaches can unsupervised extract task-relevant information. Different from these work, we present an objective to discard task-irrelevant information and show its combination with existing contrastive or predictive objectives benefits the performance.

5 Conclusion

This work studies both theoretical and empirical perspectives on self-supervised learning. We show that the self-supervised learned representations could extract task-relevant information (with a potential loss) and discard task-irrelevant information (with a fixed gap), along with their practical deployments such as contrastive and predictive learning objectives. In the future, we plan to connect our framework and recent SSL methods that cannot be easily fit into our analysis: e.g., BYOL (Grill et al., 2020), SWAV (Caron et al., 2020), and Unifromality-Alignment (Wang & Isola, 2020).

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pp. 15509–15519, 2019.

Maria-Florina Balcan, Avrim Blum, and Ke Yang. Co-training and expansion: Towards bridging theory and practice. In *Advances in neural information processing systems*, pp. 89–96, 2005.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Cristian S Calude. *Information and randomness: an algorithmic perspective*. Springer Science & Business Media, 2013.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Mark Chen, Alec Radford, Rewon Child, Jeff Wu, and Heewoo Jun. Generative pretraining from pixels.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jun Du, Charles X Ling, and Zhi-Hua Zhou. When does cotraining work in real data? *IEEE Transactions on Knowledge and Data Engineering*, 23(5):788–799, 2010.

Meir Feder and Neri Merhav. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40(1):259–266, 1994.

M Federici, A Dutta, P Forré, N Kushmann, and Z Akata. Learning robust representations via multi-view information bottleneck. International Conference on Learning Representation, 2020.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Olivier J Hénaff, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

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. *arXiv preprint arXiv:1808.06670*, 2018.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *arXiv preprint arXiv:2008.01064*, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. Ccmi: Classifier based conditional mutual information estimation. In *Uncertainty in Artificial Intelligence*, pp. 1083–1093. PMLR, 2020.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pp. 69–84. Springer, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*, 2019.

Karthik Sridharan and Sham M Kakade. An information theoretic framework for multi-view learning. 2008.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. *arXiv preprint arXiv:2008.10150*, 2020.

Yao-Hung Hubert Tsai, Han Zhao, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Neural methods for point-wise dependency estimation. *arXiv preprint arXiv:2006.05553*, 2020.

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1526–1535, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances in neural information processing systems*, pp. 613–621, 2016.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020.

Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.