Decomposed Mutual Information Estimation for Contrastive Representation Learning

Alessandro Sordoni* [1] **Nouha Dziri*** [2] **Hannes Schulz*** [1] **Geoff Gordon** [1] **Phil Bachman** [1] **Remi Tachet** [1]

Abstract

Recent contrastive representation learning methods rely on estimating mutual information (MI) between multiple views of an underlying context. E.g., we can derive multiple views of a given image by applying data augmentation, or we can split a sequence into views comprising the past and future of some step in the sequence. Contrastive lower bounds on MI are easy to optimize, but have a strong underestimation bias when estimating large amounts of MI. We propose decomposing the full MI estimation problem into a sum of smaller estimation problems by splitting one of the views into progressively more informed subviews and by applying the chain rule on MI between the decomposed views. This expression contains a sum of unconditional and conditional MI terms, each measuring modest chunks of the total MI, which facilitates approximation via contrastive bounds. To maximize the sum, we formulate a contrastive lower bound on the conditional MI which can be approximated efficiently. We refer to our general approach as Decomposed Estimation of Mutual Information (DEMI). We show that DEMI can capture a larger amount of MI than standard non-decomposed contrastive bounds in a synthetic setting, and learns better representations in a vision domain and for dialogue generation.

1. Introduction

The ability to extract actionable information from data in the absence of explicit supervision seems to be a core prerequisite for building systems that can, for instance, learn from few data points or quickly make analogies and transfer to other tasks. Approaches to this problem include generative models (Hinton, 2012; Kingma & Welling, 2014) and self-supervised representation learning approaches, in which the objective is not to maximize likelihood, but to formulate a series of (label-agnostic) tasks that the model needs to solve through its representations (Noroozi & Favaro, 2016; Devlin et al., 2019; Gidaris et al., 2018; Hjelm et al., 2019). Self-supervised learning includes successful models leveraging contrastive learning, which have recently attained comparable performance to their fully-supervised counterparts (Bachman et al., 2019; Chen et al., 2020a).

Recent self-supervised learning methods can be seen as training an encoder f such that it maximizes the mutual information (MI) between representations $f(\cdot)$ of a pair of views x and y of the same input datum, $I(f(x); f(y)) \leq I(x; y)$[1]. For images, different views can be built using random flipping or color jittering (Bachman et al., 2019; Chen et al., 2020a). For sequential data such as conversational text, the views can be past and future utterances in a given dialogue, or a particular word and its surrounding context (Stratos, 2019). Contrastive approaches train representations of pairs of views to be more similar to each other than to representations sampled from a negative sample distribution. The InfoNCE bound on $I(x; y)$ (Oord et al., 2018) has been successful insofar as it enjoys much lower variance than competing approaches (Song & Ermon, 2020a). However, the capacity of the bound is limited by the number of contrastive samples used (McAllester & Stratos, 2020a; Poole et al., 2019) and is therefore likely biased when a large amount of MI needs to be estimated, e.g. between high dimensional objects such as natural images.

The starting point of this paper is to decompose $I(x, y)$ by applying the chain rule on MI to obtain a sum of terms, each containing smaller chunks of the total MI that can be approximated with less bias by contrastive approaches. For example, consider creating a subview x' by removing information from x, e.g. by masking some pixels as depicted in Fig. 1 (*left*). By construction, $I(x', x; y) = I(x'; y) + I(x; y|x') = I(x; y)$. Decomposed Estimation of Mutual Information (DEMI) prescribes learning representations that maximize each term in the sum, by contrastive learning. The condi-

*Equal contribution [1]Microsoft Research [2]University of Alberta. Correspondence to: Alessandro Sordoni <alsordon@microsoft.com>, Nouha Dziri <dziri@cs.ualberta.ca>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]In what follows, we will slightly abuse language and use the expression "maximizing $I(x, y)$" as a shortcut for "maximizing a lower bound on $I(x, y)$ with respect to f".

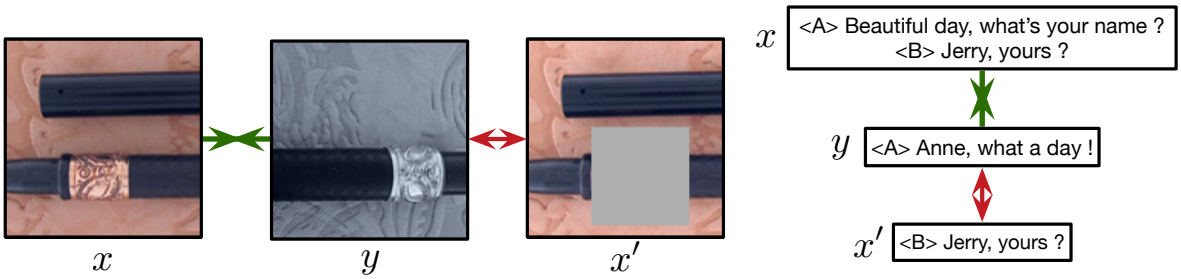

Figure 1: (**left**) Given two augmentations x and y, we create a subview x', which is obtained by occluding some of the pixels in x. We can maximize $I(x;y) \geq I(x';y) + I(x;y|x')$ using a contrastive bound by training x' to be closer to y than to other images from the corpus. Additionally, we train x to be closer to y than to samples from $p(y|x')$, i.e. we can use x' to generate hard negatives y, which corresponds to maximizing conditional MI, and leads the encoder to capture features not explained by x'. (**right**) A fictional dialogue in which x and y represent past and future of the conversation respectively and x' is the "recent past". In this context, the conditional MI term encourages the encoder to capture long-term dependencies that cannot be explained by the most recent utterances.

tional MI term measures the information about y that the model has gained by looking at x given the information already contained in x'. An intuitive explanation of why this term may lead to capturing more of the total MI between views can be found in Fig. 1. For images (*left*), only maximizing $I(x;y)$ could imbue the representations with the overall "shape" of the stick and representations would likely need many negative samples to capture other discriminative features of the image. By maximizing conditional MI, we hope to more directly encourage the model to capture these additional features, e.g. the embossed detailing. In the context of predictive coding on sequential data such as dialogue, by setting x' to be the most recent utterance (Fig. 1, *right*), the encoder is directly encouraged to capture long-term dependencies that cannot be explained by the most recent utterance.

One may wonder how DEMI is related to recent approaches maximizing MI between more than two views, amongst them AMDIM (Bachman et al., 2019), CMC (Tian et al., 2019) and SwAV (Caron et al., 2020). Interestingly, these models can be seen as maximizing the sum of MIs between views $I(x,x';y) = I(x';y) + I(x;y)$. E.g., in Bachman et al. (2019), x and x' could be global and local representations of an image, and in Caron et al. (2020), x and x' could be the views resulting from standard cropping and the aggressive multi-crop strategy. This equality is only valid when the views x and x' are statistically independent, which usually does not hold. Instead, DEMI maximizes $I(x,x';y) = I(x';y) + I(x;y|x')$, which always holds. Most importantly, the conditional MI term encourages the encoder to capture more non-redundant information across views.

Our contributions are the following. We show that DEMI can potentially capture more of the total information shared between the original views x and y. We extend existing contrastive MI bounds to conditional MI estimation and present novel computationally tractable approximations. Supplementally, our results offer another perspective on *hard* contrastive examples, i.e., Faghri et al. (2018), given that conditional MI maximization can be achieved by sampling contrastive examples from a partially informed conditional distribution instead of the marginal distribution. We first show in a synthetic setting that DEMI leads to capturing more of the ground-truth MI thus alleviating bias existing in InfoNCE. Finally, we present evidence of the effectiveness of the proposed method in vision and in dialogue generation.

2. Problem Setting

The maximum MI predictive coding framework (McAllester, 2018; Oord et al., 2018; Hjelm et al., 2019) prescribes learning representations of input data such that they maximize MI between inputs and representations. Recent interpretations of this principle create two independently-augmented copies x and y of the same input by applying a set of stochastic transformations twice, and then learn representations of x and y by maximizing the MI of the respective features produced by an encoder $f : \mathcal{X} \to \mathbb{R}^d$ (Bachman et al., 2019; Chen et al., 2020a):

$$\arg\max_f \; I(f(x);f(y)) \leq I(x;y) \tag{1}$$

where the upper bound is due to the data processing inequality. Our starting point to maximize Eq. 1 is the recently proposed InfoNCE lower bound on MI (Oord et al., 2018) which trains $f(x)$ to be closer to $f(y)$ than to the representations of other images drawn from the marginal distribution of the corpus. This can be viewed as a *contrastive* estimation of the MI (Oord et al., 2018) and has been shown to enjoy lower variance than competing approaches (Song & Ermon, 2020a).

2.1. InfoNCE Bound

InfoNCE (Oord et al., 2018) is a lower-bound on $I(x;y)$ obtained by comparing pairs sampled from the joint distribution $x, y_1 \sim p(x,y)$ to pairs x, y_i built using a set of negative examples, $y_{2:K} \sim p(y_{2:K}) = \prod_{k=2}^{K} p(y_k)$, also called *contrastive*, independently sampled from the marginal:

$$I_{NCE}(x,y|\phi,K) = \mathbb{E}\left[\log \frac{e^{\psi(x,y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\psi(x,y_k)}}\right], \quad (2)$$

where the expectation is with respect to $p(x,y_1)p(y_{2:K})$ and ψ is a critic assigning a real valued score to x, y pairs. Usually, ψ is the dot product of the representations after applying an additional transformation g, e.g. an MLP, $\psi(x,y) \triangleq g(f(x))^T g(f(y))$ (Chen et al., 2020a). We provide an exact derivation of this bound in the Appendix[2]. The optimal value of I_{NCE} is reached for a critic proportional to the log-odds between the conditional distribution $p(y|x)$ and the marginal distribution $p(y)$, i.e. the PMI between x and y, $\psi^*(x,y) = \log \frac{p(y|x)}{p(y)} + c(x)$ (Oord et al., 2018; Ma & Collins, 2018; Poole et al., 2019).

InfoNCE has recently been extensively used in self-supervised representation learning given that it enjoys lower variance than some of its competitors such as MINE (Belghazi et al., 2018; Song & Ermon, 2020a). However, the bound is loose if the true mutual information $I(x;y)$ is larger than $\log K$, which is likely when dealing with high-dimensional inputs such as natural images. To overcome this difficulty, recent methods either train with large batch sizes (Chen et al., 2020a) or exploit an external memory of negative samples in order to reduce memory requirements (Chen et al., 2020b; Tian et al., 2020). These methods rely on uniform sampling from the training set in order to form the contrastive sets. Discussion of limits of variational bounds can be found in McAllester & Stratos (2020a).

3. Decomposing Mutual Information

When $\mathcal{X}$ is high-dimensional, the amount of mutual information between x and y will potentially be larger than the amount of MI that I_{NCE} can measure given computational constraints associated with large K and the poor log scaling properties of the bound. We argue that we can ease this estimation problem by creating subviews of x and applying the chain rule on MI to decompose the total MI into a sum of potentially smaller MI terms.

By the data processing inequality, we have: $I(x;y) \geq I(\{x^1,\dots,x^N\};y)$, where $\{x^1,\dots,x^N\}$ are different subviews of x – i.e., views derived from x without adding *any* exogenous information. For example, $\{x^1,\dots,x^N\}$ can represent single utterances in a dialog x, sentences in a document x, or different augmentations of the same image x. Equality is obtained when the set of subviews retains all information about x or if x is in the set.

For ease of exposition and without loss of generality, we consider the case where we have two subviews, x itself and x'. Then, $I(x;y) = I(x,x';y)$ and we can write $I(x,x';y)$ by applying the chain rule for MI:

$$I(x,x';y) = I(x';y) + I(x;y|x'). \quad (3)$$

The conditional MI term can be written as:

$$I(x;y|x') = \mathbb{E}_{p(x,x',y)} \log \frac{p(y|x,x')}{p(y|x')}. \quad (4)$$

This conditional MI is different from the unconditional MI, $I(x;y)$, as it measures the amount of information shared between x and y that cannot be explained by x'.

Lower bounding each term in Eq. 3 with a contrastive bound can potentially lead to a less biased estimator of the total MI. This motivates us to introduce DEMI, a sum of unconditional and conditional lower bounds:

$$I_{DEMI} = I_{NCE}(x';y) + I_{CNCE}(x;y|x') \leq I(x;y), \quad (5)$$

where I_{CNCE} is a placeholder for a lower bound on the conditional MI and will be presented in the next section. Both conditional and unconditional bounds on the MI can capture at most $\log K$ nats of MI. Therefore, DEMI in Eq. 5 potentially allows to capture up to $N \log K$ nats of MI in total, where N is the number of subviews used to describe x. This is strictly larger than $\log K$ in the standard I_{NCE}.

4. Contrastive Conditional MI Estimation

One of the difficulties in computing DEMI is estimating the conditional MI. In this section, we provide bounds and approximations of this quantity. First, we show that we can readily extend InfoNCE:

Proposition 1 (Conditional InfoNCE). *I_{CNCE} is a lower-bound on $I(x;y|x')$ and verifies the properties below:*

$$I_{CNCE}(x;y|x',\phi,K) = \mathbb{E}\left[\log \frac{e^{\phi(x',x,y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\phi(x',x,y_k)}}\right], \quad (6)$$

1. $I_{CNCE} \leq I(x;y|x')$.
2. $\phi^* = \arg\sup_\phi I_{CNCE} = \log \frac{p(y|x',x)}{p(y|x')} + c(x,x')$.
3. $\lim_{K\to\infty} I_{CNCE}(x;y|x',\phi^*,K) = I(x;y|x')$.

The expectation is taken with respect to $p(x,x',y_1)p(y_{2:K}|x')$ and the expression is upper bounded by $\log K$. The proof can be found in Sec. A.2 and

[2]The derivation in Oord et al. (2018) presented an approximation and therefore was not properly a bound. An alternative, exact derivation of the bound can be found in (Poole et al., 2019).

follows closely the derivation of the InfoNCE bound by applying a result from (Barber & Agakov, 2003). A related derivation of this bound was also presented in Foster et al. (2020) for optimal experiment design.

Eq. 6 shows that a lower bound on the conditional MI can be obtained by sampling contrastive sets from the proposal distribution $p(y|x')$ (instead of from the marginal $p(y)$ as in Eq. 2). Indeed, since we want to estimate the MI conditioned on x', we should allow our contrastive distribution to condition on x'. Note that ϕ is now a function of three variables. One of the biggest hurdles in computing Eq. 6 is the access to many samples from $p(y|x')$, which is unknown and usually challenging to obtain. In order to overcome this, we propose various solutions next.

4.1. Variational Approximation

It is possible to obtain a bound on the conditional MI by approximating the unknown conditional distribution $p(y|x')$ with a variational distribution $q_\xi(y|x')$, leading to the following proposition:

Proposition 2 (Variational I_{CNCE}). *For any variational approximation $q_\xi(y|x')$ in lieu of $p(y|x')$, with $p(\cdot|x') \ll q_\xi(\cdot|x')$ for any x', we have:*

$$I_{VAR}(x, y|x', \phi, \xi, K) = \tag{7}$$
$$\mathbb{E}\left[\log \frac{e^{\phi(x', x, y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\phi(x', x, y_k)}}\right] - \mathbb{E}\Big[KL\left(p(y|x') \,\|\, q_\xi\right)\Big],$$

1. $I_{VAR} \leq I(x; y|x')$.
2. *If* $q_\xi(y|x') = p(y|x')$, $I_{VAR} = I_{CNCE}$.
3. $\lim_{K\to\infty} \sup_\phi I_{VAR}(x; y|x', \phi, \xi, K) = I(x; y|x')$.

where the first expectation is taken with respect to $p(x, x', y_1)q_\xi(y_{2:K}|x')$ and the second with respect to $p(x')$. See Sec. A.3 for the proof. Note that this bound side-steps the problem of requiring access to an arbitrary number of negative samples from the unknown $p(y|x')$ by i.i.d. sampling from the known and tractable $q_\xi(y|x')$. For example, q_ξ can be a conditional flow-based image generation model (Kingma & Dhariwal, 2018) or a transformer language model for text (Zhang et al., 2020). We prove that as the number of examples goes to ∞, optimizing the bound w.r.t. ϕ converges to the true conditional MI. Interestingly, this holds true for any q_ξ, though the choice of q_ξ will most likely impact the convergence rate of the estimator.

Eq. 7 is superficially similar to the ELBO (Evidence Lower BOund) objective used to train VAEs (Kingma & Welling, 2014), where q_ξ plays the role of the approximate posterior (although the KL direction in the ELBO is inverted). This parallel suggests that, assuming the variational family contains p, the optimal solution w.r.t. ξ may not verify $p(y|x') = q_\xi(y|x')$ for all values of K and ϕ, i.e. there could be solutions for which some of the KL divergence is traded for additional nats on the contrastive cost. However, we see trivially that if we ignore the dependency of the first expectation term on q_ξ (i.e. we "detach" the gradient of the expectation w.r.t ξ) and only optimize ξ to minimize the KL term, then it is guaranteed that $p(y|x') = q_\xi(y|x')$, for any K and ϕ. Thus, by the second property in Proposition 2, optimizing $I_{VAR}(\phi, \xi^*, K)$ w.r.t. ϕ will correspond to optimizing I_{CNCE}.

In practice, the latter observation significantly simplifies the estimation problem as one can minimize a Monte-Carlo approximation of the KL divergence w.r.t ξ by standard supervised learning: we can efficiently approximate the KL by taking samples from $p(y|x')$. Those can be directly obtained by using the joint samples from $p(x, y)$ included in the training set and computing x' from x.[3] However, maximizing I_{VAR} can still be challenging as it requires estimating a distribution over potentially high-dimensional inputs and efficiently sampling a large number of negative examples from it. In the next section, we provide an importance sampling approximation of I_{CNCE} that bypasses this issue.

4.2. Importance Sampling Approximation

The optimal critic for I_{NCE} is $\psi^*(x', y) = \log \frac{p(y|x')}{p(y)} + c(x')$, for any c. Assuming access to $\psi^*(x', y)$, it is possible to use importance sampling to produce approximate expectations from $p(y|x')$. This is achieved by first sampling $\tilde{y}_{1:M} \sim p(y)$ and then resampling $K \leq M$ ($K > 0$) examples i.i.d. from the normalized importance distribution $w_k = \frac{\exp \psi^*(x', \tilde{y}_k)}{\sum_{m=1}^{M} \exp \psi^*(x', \tilde{y}_m)}$. This process is also called "sampling importance resampling" (SIR) and we can write the corresponding distribution as $p_{SIR}(y_k) = w_k \delta(y_k \in \tilde{y}_{1:M}) p(\tilde{y}_{1:M})$. As $M/K \to \infty$, it is guaranteed to produce samples from $p(y|x')$ (Rubin, 1987).

The objective corresponding to this process is:

$$I_{SIR}(x, y|x', \phi, K) = \tag{8}$$
$$\mathbb{E}_{p(x', x, y_1) p_{SIR}(y_{2:K})}\left[\log \frac{e^{\phi(x', x, y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\phi(x', x, y_k)}}\right]$$

Note the dependence of p_{SIR} on w_k and hence ψ^*. SIR is known to increase the variance of the estimator (Skare et al., 2003) and is wasteful given that only a smaller set of $K < M$ examples are actually used for MI estimation.

To provide a cheap approximation of the SIR estimator, we split the denominator of Eq. 8 into a positive term in-

[3] The ability to perform that computation is usually a key assumption in self-supervised learning approaches.

volving y_1 and a sum of contributions coming from negative examples $y_{2:K}$, and we rewrite the latter as an average $(K-1)\sum_{k=2}^{K}\frac{1}{K-1}e^{\phi(x',x,y_k)}$. Now, we can use the normalized importance weights w_k to estimate that term under the resampling distribution. Formally, we have the following approximation:

Proposition 3 (Importance Sampled I_{CNCE}). *Assuming* $\psi^* = \arg\sup_\psi I_{NCE}(x', y)$ *and* $w_k = \frac{\exp\psi^*(x',y_k)}{\sum_{k=2}^{M}\exp\psi^*(x',y_m)}$, *we have the following two properties, where:*

$$I_{IS}(x, y|x', \phi, K) = \mathbb{E}\left[\log \frac{e^{\phi(x',x,y_1)}}{\frac{1}{K}(e^{\phi(x',x,y_1)} + (K-1)\sum_{k=2}^{K} w_k e^{\phi(x',x,y_k)})}\right], \tag{9}$$

1. $\lim_{K\to\infty}\sup_\phi I_{IS}(x; y|x', \phi, K) = I(x; y|x')$,
2. $\lim_{K\to\infty}\arg\sup_\phi I_{IS} = \log\frac{p(y|x',x)}{p(y|x')} + c(x, x')$.

where the expectation is with respect to $p(x', x, y_1)p(y_{2:K})$. The proof can be found in Sec. A.4. I_{IS} skips the resampling step by up-weighting the negative contribution to the normalization term of examples that have large probability under the resampling distribution, i.e. that have large w_k. As detailed in the appendix, this approximation is cheap to compute given that the negative samples are sampled from the marginal distribution $p(y)$ and we avoid the need for the resampling step. We hypothesize that I_{IS} has less variance than I_{SIR} as it does not require the additional resampling step. The proposition shows that as the number of negative examples goes to infinity, the proposed approximation converges to the true value of the conditional MI, and, in the limit of $K \to \infty$, optimizing I_{IS} w.r.t. ϕ converges to the conditional MI and the optimal ϕ converges to the optimal I_{CNCE} solution.

4.3. Boosted Critic Approximation

Proposition 3 shows that the optimal critic ϕ^* estimates the desired log ratio only in the limit of $K \to \infty$. Hereafter, we generalize the results presented in Ma & Collins (2018) and show that we can accurately estimate the conditional log-ratio with the following proposition.

Proposition 4 (Boosted Critic Estimation). *Assuming* $\psi^* = \arg\sup_\psi I_{NCE}(x', y)$, *the following holds, with:*

$$I_{BO}(x, y|x', \phi, K) = \mathbb{E}\left[\log \frac{e^{\psi^*(x',y_1)+\phi(x',x,y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\psi^*(x',y_k)+\phi(x',x,y_k)}}\right], \tag{10}$$

1. $I_{BO} \leq I(x, x'; y)$,
2. $\phi^* = \arg\sup_\phi I_{BO} = \log\frac{p(y|x',x)}{p(y|x')} + c(x, x')$.

where the expectation is with respect to $p(x, x', y_1)p(y_{2:K})$. The proof is straightforward and is in Sec. A.5.

We refer to Eq. 10 as *boosted critic estimation* due to the fact that optimizing ϕ captures residual information not expressed in ψ^*. Perhaps surprisingly, I_{BO} provides an almost embarrassingly simple way of estimating the desired log-ratio *for any* K. It corresponds to estimating an InfoNCE like bound, where negative samples come from the easily-sampled marginal $p(y)$ and the critic is shifted by the optimal critic for $I_{NCE}(x', y)$. However, this comes at the cost of not having a valid approximation of the conditional MI. Indeed, by 1., I_{BO} is a lower-bound on the total MI, not on the conditional MI. As we show in the next section, we can get an estimate of the conditional MI by using I_{BO} to estimate the conditional critic in an accurate manner and I_{IS} to evaluate the conditional MI.

5. Experiments

The goal of our experiments is two-fold: (1) to test whether DEMI leads to a better estimator of the total MI, and whether our proposed conditional MI approximations are accurate; (2) to test whether DEMI helps in estimating better representations for natural data. We verify (1) in a synthetic experiment where we control the total amount of MI between Gaussian covariates. Then, we verify (2) on a self-supervised image representation learning domain and explore an additional application to natural language generation in a sequential setting: conversational dialogue.

5.1. Synthetic Data

We extend Poole et al. (2019)'s two variable setup to three variables. We posit that $\{x, x', y\}$ are three Gaussian covariates, $x, x', y \sim \mathcal{N}(0, \Sigma)$ and we choose Σ such that we can control the total mutual information $I(x, x'; y)$, $I \in \{5, 10, 15, 20\}$ (see Appendix for pseudo-code and details of the setup). We aim to estimate the total MI $I(x, x'; y)$ and compare the performance of our approximators in doing so. We limit this investigation to contrastive estimators although other estimators and non lower-bounds exist (e.g. DoE (McAllester & Stratos, 2020b)). For more details see App. A.6.

In Figure 2 (top), we compare the estimate of the MI obtained by InfoNCE and DEMI, which maximizes I_{DEMI} (Eq. 5). To be comparable with InfoNCE in terms of total number of negative examples used, DEMI uses half as many negative examples for computing each term in the sum $(K/2)$. For all amounts of true MI, and especially for larger amounts, DEMI can capture more nats than InfoNCE with an order of magnitude less examples. We also report the upper-bounds on InfoNCE ($\log K$) and DEMI ($2\log K/2$).

Maximizing I_{DEMI} assumes access to negative samples from

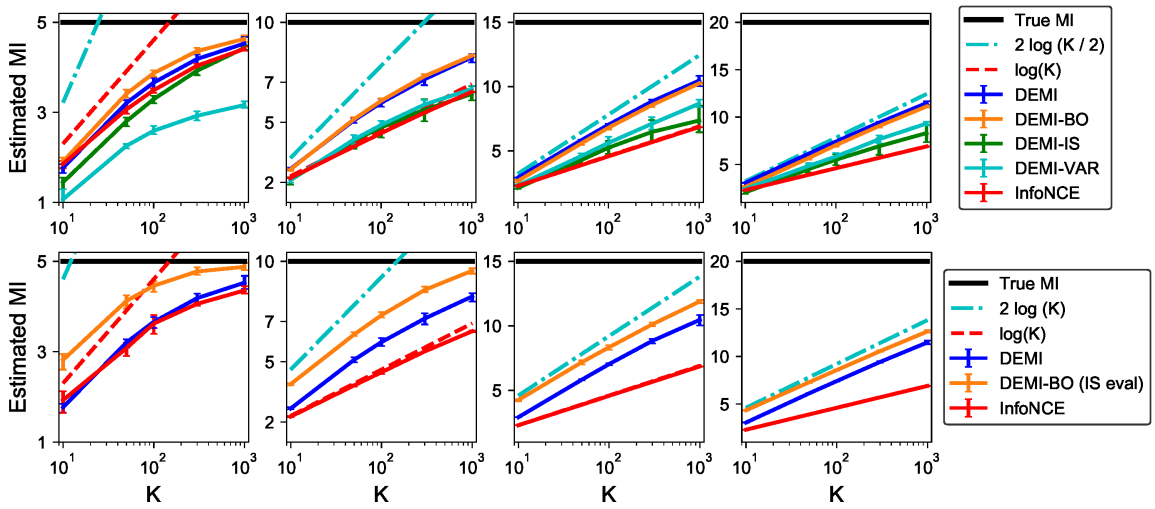

Figure 2: Estimation of $I(x, x'; y)$ for three Gaussian covariates x, x', y as function of the number of negative samples K. (**top**) DEMI maximizes I_{DEMI} with $K/2$ examples for unconditional and conditional bounds (K total) and assume access to the ground-truth $p(y|x)$. DEMI-IS learns the conditional critic using I_{IS}, DEMI-BO using I_{BO}, DEMI-VAR using I_{VAR}. We plot the total MI estimated by I_{DEMI} when learning the conditional critics using our approximations. We see that (1) DEMI captures more MI than InfoNCE for the same number of K and (2) I_{BO} accurately estimates the conditional critic without access to samples from $p(y|x')$ while I_{IS} suffers from significant variance. (**bottom**) We assess whether we can form a good estimator of the total MI *without* access to $p(y|x')$ neither at training nor at evaluation time. Here, DEMI-BO trains the conditional critic by I_{BO} and evaluates the total MI by $I_{NCE} + I_{IS}$.

$p(y|x')$, which is an unrealistic assumption in practice. To verify the effectiveness of our approximations, we train the conditional critics using I_{BO} (DEMI-BO), I_{IS} (DEMI-IS) and I_{VAR} (DEMI-VAR) and we evaluate the total MI using I_{DEMI} (we assume access to $p(y|x')$ only at evaluation time). This allows us to verify whether it is possible to reliably estimate the conditional critic in the absence of negative samples from $p(y|x')$. It is interesting to note how the critic learnt by I_{IS} suffers high variance and does not lead to a good estimate of the total MI when evaluated with I_{CNCE}. DEMI-VAR still outperforms InfoNCE for higher values of total MI, but seems to suffer in the case of small MIs. For this experiment, we update q_ξ at the same rate as ϕ. Improvements could be obtained by updating q_ξ more frequently, similarly to the asynchronous updates successfully used in the GAN literature (Mescheder et al., 2018). I_{BO} accurately estimates the critic.

In Figure 2 (bottom), we show that it is possible to obtain an estimate of the total MI without access to $p(y|x')$ neither at training nor evaluation time. We first learn the conditional critic using I_{BO} and compute $I_{NCE} + I_{IS}$ by using the estimated critic. Figure 2 (bottom) reports the results. For this experiment, we share the same set of K negative examples for both conditional and unconditional MI and therefore we report the upper bound $2 \log K$.

5.2. Vision

5.2.1. ImageNet

Setup We study self-supervised learning of image representations using 224×224 images from ImageNet (Deng et al., 2009). The evaluation is performed by fitting a linear classifier to the task labels using the pre-trained representations only, that is, we fix the weights of the pre-trained image encoder f. We build upon InfoMin (Tian et al., 2020). All hyperparameters for training and evaluation are the same as in Tian et al. (2020). All models use a momentum-contrastive memory buffer of $K = 65536$ examples (Chen et al., 2020b). All models use a Resnet50 backbone and are trained for 200 epochs. We report transfer learning performance by freezing the encoder on STL-10, CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), Stanford Cars (Krause et al., 2013), Caltech-UCSD Birds (CUB) (Welinder et al., 2010) and Oxford 102 Flowers (Nilsback & Zisserman, 2008).

Views Each input image is independently augmented into two views x and y using a stochastically applied transformation following Tian et al. (2020). This uses random resized crop, color jittering, gaussian blur, rand augment, color dropping, and jigsaw as augmentations. We experiment two ways of creating the subview x' of x: `cut`, which applies cutout to x, and `crop` which is inspired by Caron et al. (2020) and consists in cropping the image aggressively and resizing the resulting crops to 96×96. To do so, we use the `RandomResizedCrop` from the

Table 1: Accuracy for self-supervised learning on Imagenet-100 (IN100) and on full Imagenet (IN1K), measured by linear evaluation. $x \leftrightarrow y$ denotes standard contrastive matching between views. In DEMI, we use the same base InfoMin architecture but augments the loss function with conditional MI maximization across views. InfoMin (multi) considers x' just as an additional view and therefore discards conditional MI maximization. All models use a standard Resnet-50 and are trained for 200 epochs. The right part of the table reports transfer learning performance of our model trained on IN1K.

Model	Views	IN100	IN1K	STL10	C10	C100	CARS	CUB	FLOWERS
SimCLR (Chen et al., 2020a)	$x \leftrightarrow y$	-	66.6	-	90.6	71.6	50.3	-	91.2
MocoV2 (Chen et al., 2020b)	$x \leftrightarrow y$	-	67.5	-	-	-	-	-	-
InfoMin (Tian et al., 2020)	$x \leftrightarrow y$	74.9	70.1	96.2	92.0	73.2	48.1	41.7	93.2
InfoMin (multi)	$x, x' \leftrightarrow y$	77.2	70.2	95.9	92.6	74.5	49.2	42.1	94.7
DEMI	$x, x' \leftrightarrow y$	**78.6**	**70.8**	**96.4**	**92.8**	**75.0**	**51.8**	**43.6**	**95.0**

`torchvision.transforms` module with $\mathtt{s} = (0.05, 0.14)$.

Models Our baseline, InfoMin, maximizes $I_{NCE}(x, y)$. We also report an enhanced baseline InfoMin (multi), which maximizes $I_{NCE}(x, y) + I_{NCE}(x', y)$ and aims to verify whether additional gains can be obtained by estimating conditional MI rather than just using x' as an additional view. We recur to I_{BO} to estimate the conditional critic[4]. DEMI maximizes four terms: $I_{NCE}(x'; y) + I_{BO}(x; y|x') + I_{NCE}(x; y) + I_{BO}(x'; y|x)$. This corresponds to maximizing both decompositions of the joint $I(x, x'; y)$. Differently from MI estimation, we found to be important for representation learning to maximize both decompositions, which include $I_{NCE}(x; y)$ in the objective. The computation of the conditional MI terms can be efficiently done by reusing the logits of the two unconditional MI (Listing 1).

Results Table 1 reports the average accuracy of linear evaluations obtained by 3 pretraining seeds. DEMI obtains 3.7% improvement (78.6±0.2) compared to the baseline InfoMin for Imagenet100 (IN100) and 0.7% (70.8±0.1) for full Imagenet (IN1K). Although not reported, the `crop` strategy performs better than the `cut` strategy (which obtains 70.5±0.1 on average IN1K). One hypothesis is that cutout introduces image patches that do not follow the pixel statistics in the corpus. InfoMin (multi) ablates conditional MI maximization and shows that introducing the additional view is helpful in low-data setting such as IN100, but can only slightly improve performance in IN1K. It is interesting to note that DEMI improves transfer learning performance the most in the fine-grained classification benchmarks CARS and CUB, where it is particularly important to capture detailed information about the input image (Yang et al., 2018). This serves as indication that the representations learnt by DEMI can extract more information about each input.

5.2.2. CIFAR-10

We also experiment on CIFAR-10 building upon SimCLR (Chen et al., 2020b), which uses a standard ResNet-50 architecture by replacing the first 7×7 Conv of stride 2 with 3×3 Conv of stride 1 and also remove the max pooling operation. In order to generate the views, we use Inception crop (flip and resize to 32×32) and color distortion. We train with learning rate 0.5, batch-size 800, momentum coefficient of 0.9 and cosine annealing schedule. Our energy function is the cosine similarity between representations scaled by a temperature of 0.5 (Chen et al., 2020b). We obtain a top-1 accuracy of 94.7% using a linear classifier compared to 94.0% reported in Chen et al. (2020b) and 95.1% for a supervised baseline with same architecture.

[4] Although not reported explicitly, we found that I_{IS} leads very similar performance with a slightly higher variance across seeds.

```
def compute_demi(x, xp, y, f, f_ema, g, g_bo):
  f_x, f_xp, k_y = f(x), f(xp), g(f_ema(y))
  # NCE heads
  q_x, q_xp = g(f_x), g(f_xp)
  # conditional NCE heads
  q_bo_x, q_bo_xp = g_bo(f_x), g_bo(f_xp)
  # compute NCE critics
  s_x_y = dot(q_x, cat(k_y, memory))
  s_xp_y = dot(q_xp, cat(k_y, memory))
  # compute conditional NCE critics
  s_bo_xp_y = dot(q_bo_xp, cat(k_y, memory))
  s_bo_x_y = dot(q_bo_x, cat(k_y, memory))
  # compute NCE bounds
  nce_x_y = -log_softmax(s_x_y)[0]
  nce_xp_y = -log_softmax(s_xp_y)[0]
  # compute BO estimator
  bo_x_xp = -log_softmax(
    s_xp_y.detach() + s_bo_x_y)[0]
  bo_xp_x = -log_softmax(
    s_x_y.detach() + s_bo_xp_y)[0]
  return (nce_x_y + nce_xp_y + bo_x_xp + bo_xp_x)
```

Listing 1: PyTorch-style pseudo-code for DEMI in InfoMin. We use I_{BO} to estimate the critic for conditional MI.

5.3. Dialogue

Setup We experiment with language modeling task on the Wizard of Wikipedia (WoW) dataset (Dinan et al., 2019). We evaluate our models using automated metrics and human evaluation. For automated metrics, we report perplexity (ppl), BLEU (Papineni et al., 2002). We report a comprehensive set of metrics in the Appendix (Sec B). We build upon GPT2 (Radford et al., 2019), and fine-tune it by language modeling (LM) on the dialogue corpus. In addition to the LM loss, we maximize MI between representations of the

Table 2: Perplexity, BLEU and side-by-side human evaluation on WoW (Dinan et al., 2019). H- columns indicate whether DEMI was preferred (✓) or not (✗), or neither (=) at $\alpha = 0.01$.

Model	ppl	BLEU	H-rel	H-hum	H-int
GPT2	19.21	0.78	✓	✓	✓
TransferTransfo	19.32	0.75	✓	✓	✓
GPT2-MMI	19.30	0.65	✓	✓	✓
InfoNCE	18.85	0.80	=	✓	✓
DEMI	**18.70**	**0.82**	=	=	=
Human	–	–	✗	✗	✗

past and future utterances in each dialogue, i.e. the predictive coding framework (Elias, 1955; McAllester & Stratos, 2020a). We consider past and future in a dialogue as views of the same conversation. Given L utterances $(x_1, \dots, x_L)$, we set $y = (x_{k+1}, \dots, x_L)$, $x = (x_1, \dots, x_k)$ and $x' = x_k$, where $(.)$ denotes concatenation and k is randomly chosen between $2 < k < L$. The goal is therefore to imbue representations with information about the future that cannot be solely explained by the most recent utterance x'. The representations of past and future are the state corresponding to the last token in the last layer in GPT2.

Models We evaluate our introduced models against different baselines. GPT2 is a basic small pre-trained model fine-tuned on the dialogue corpus. TransferTransfo (Wolf et al., 2019) augments the standard next-word prediction loss in GPT2 with the next-sentence prediction loss similar to Devlin et al. (2019). GPT2-MMI follows MMI-bidi (Li et al., 2016); we generate 50 responses from GPT2 and then rank them based on a trained backward model $p_{GPT2}(x|y)$. For the InfoNCE baseline, we only maximize the unconditional MI between x and y and sample negative futures from the marginal distribution $p(y)$. DEMI maximizes conditional MI by recurring to I_{VAR} and using GPT2 itself as the variational approximation. GPT2 is a generative model therefore we can simply sample a set of negative futures from $p_{GPT2}(y|x')$, that is, by restricting the amount of contextual information GPT2 is allowed to consider. To speed up training, the negative sampling of future candidates is done offline. We also tried I_{BO} in this setting and obtained similar results.

Results Table 2 shows results on the validation set obtained by 3 pretraining seeds. For the test set results and sample dialogue exchanges, please refer to the Appendix. The automated metrics indicate that DEMI representations result in higher quality responses. We also perform human evaluation on randomly sampled 1000 WoW dialogue contexts. We present the annotators with a pair of candidate responses consisting of InfoNCE, DEMI and baseline responses. They were asked to compare the pairs regarding interestingness, relevance and humanness, using a 3-point Likert scale (Zhang et al., 2020). In Table 2, we see that overall responses generated by DEMI were strongly preferred to other models but not to the gold response. Bootstrap confidence intervals and p-values (t-test, following Zhang et al., 2020) indicate significant improvements at α=0.01.

6. Related Works

Representation learning based on MI maximization has been applied in various domains such as images (Grill et al., 2020; Caron et al., 2020), words (Mikolov et al., 2013; Stratos, 2019), graphs (Velickovic et al., 2019), RL (Mazoure et al., 2020) and videos (Jabri et al., 2020), exploiting noise-contrastive estimation (NCE) (Gutmann & Hyvärinen, 2012), InfoNCE (Oord et al., 2018) and variational objectives (MINE) (Hjelm et al., 2019). InfoNCE have gained recent interest w.r.t. variational approaches due to its lower variance (Song & Ermon, 2020a) and superior performance in downstream tasks. InfoNCE however can underestimate large amounts of true MI given that it is capped at $\log K$. Poole et al. (2019) propose to trade-off between variance and bias by interpolating variational and contrastive bounds. Song & Ermon (2020b) propose a modification to InfoNCE for reducing bias where the critic needs to jointly identify multiple positive samples at the same time. Our proposal to scaffold the total MI estimation into a sequence of smaller estimation problems shares similarities with the recent telescopic estimation of density ratio (Rhodes et al., 2020) which is based on variational approximations. Instead, we build upon InfoNCE, propose new results on contrastive conditional MI estimation and apply it to self-supervised representation learning. Other MINE-based approaches of conditional MI estimation can be found in the recent (Mondal et al., 2020). Our contrastive bound in Eq. 5 is reminiscent of conditional noise-contrastive estimation (Ceylan & Gutmann, 2018), which generalizes NCE for data-conditional noise distributions (Gutmann & Hyvärinen, 2012): our result is an interpretation in terms of conditional MI.

7. Conclusion

We decompose the original cross-view MI into a sum of conditional and unconditional MI terms (DEMI). We provide several contrastive approximations to the conditional MI and verify their effectiveness in various domains. Incorporating more than two terms in the decomposition is straightforward and could be investigated in the future. Recent work questioned whether MI maximization itself is at the core of the recent success in representation learning (Rainforth et al., 2018; Tschannen et al., 2020). These showed that capturing a larger amount of mutual information between views may not correlate to better downstream performance. Other desirable properties of the representation space may play an important role (Wang & Isola, 2020). Although

we acknowledge these results, we posit that devising more effective ways to maximize MI will still prove useful in representation learning, especially if paired with architectural inductive biases or explicit regularization methods.

Acknowledgements

We would like to acknowledge Jiaming Song and Mike Wu for the insightful discussions and the anonymous reviewers for their helpful comments.

References

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. In Proc. Conf. on Neural Information Processing Systems (NeurIPS), pp. 15509–15519, 2019.

Barber, D. and Agakov, F. The im algorithm: A variational approach to information maximization. In Proc. Conf. on Neural Information Processing Systems (NIPS), pp. 201–208, 2003.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Hjelm, R. D., and Courville, A. C. Mutual information neural estimation. In Proc. Int. Conf. on Machine Learning (ICML), pp. 530–539, 2018.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In Proc. Conf. on Neural Information Processing Systems (NeurIPS), 2020.

Ceylan, C. and Gutmann, M. U. Conditional noise-contrastive estimation of unnormalised models. In Proc. Int. Conf. on Machine Learning (ICML), pp. 725–733, 2018.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. In Proc. Int. Conf. on Machine Learning (ICML), pp. 1597–1607, 2020a.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020b.

Cremer, C., Morris, Q., and Duvenaud, D. Reinterpreting importance-weighted autoencoders. Proc. Int. Conf. on Learning Representations (ICLR), 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. Conf. Assoc. for Computational Linguistics (ACL), pp. 4171–4186, 2019.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. Wizard of wikipedia: Knowledge-powered conversational agents. In Proc. Int. Conf. on Learning Representations (ICLR), 2019.

Elias, P. Predictive coding–i. IRE Transactions on Information Theory, 1(1):16–24, 1955.

Faghri, F., Fleet, D. J., Kiros, J. R., and Fidler, S. VSE++: improving visual-semantic embeddings with hard negatives. In Proc. British Machine Vision Conference (BMVC), pp. 12, 2018.

Foster, A., Jankowiak, M., O'Meara, M., Teh, Y. W., and Rainforth, T. A unified stochastic gradient approach to designing bayesian-optimal experiments. In Proc. Int. Conf. on Artificial Intelligence and Statistics, pp. 2959–2969, 2020.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. In Proc. Int. Conf. on Learning Representations (ICLR), 2018.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent - A new approach to self-supervised learning. In Proc. Conf. on Neural Information Processing Systems (NeurIPS), 2020.

Gutmann, M. U. and Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. Journal of Machine Learning Research, 13:307–361, 2012.

Hinton, G. E. A practical guide to training restricted boltzmann machines. In Neural networks: Tricks of the trade, pp. 599–619. Springer, 2012.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In Proc. Int. Conf. on Learning Representations (ICLR), 2019.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. In Proc. Int. Conf. on Learning Representations (ICLR), 2020.

Jabri, A., Owens, A., and Efros, A. A. Space-time correspondence as a contrastive random walk. In Proc. Conf. on Neural Information Processing Systems (NeurIPS), 2020.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In Proc. Conf. on Neural Information Processing Systems (NeurIPS), pp. 10236–10245, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In Proc. Int. Conf. on Learning Representations (ICLR), 2014.

Krause, J., Deng, J., Stark, M., and Fei-Fei, L. Collecting a large-scale dataset of fine-grained cars. Technical report, 2013.

Krizhevsky, A. et al. Learning multiple layers of features from tiny images. Technical report, 2009.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. A diversity-promoting objective function for neural conversation models. In Proc. Conf. Assoc. for Computational Linguistics (ACL), pp. 110–119, 2016.

Li, M., Roller, S., Kulikov, I., Welleck, S., Boureau, Y.-L., Cho, K., and Weston, J. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In Proc. Conf. Assoc. for Computational Linguistics (ACL), pp. 4715–4728, 2020.

Ma, Z. and Collins, M. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 3698–3707, 2018.

Mazoure, B., Tachet des Combes, R., Doan, T., Bachman, P., and Hjelm, R. D. Deep reinforcement and infomax learning. In Proc. Conf. on Neural Information Processing Systems (NeurIPS), 2020.

McAllester, D. Information theoretic co-training. arXiv preprint arXiv:1802.07572, 2018.

McAllester, D. and Stratos, K. Formal limitations on the measurement of mutual information. In Int. Conf. on Artificial Intelligence and Statistics (AISTATS), pp. 875–884, 2020a.

McAllester, D. and Stratos, K. Formal limitations on the measurement of mutual information. In Chiappa, S. and Calandra, R. (eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of Proceedings of Machine Learning Research, pp. 875–884. PMLR, 26–28 Aug 2020b.

Mescheder, L. M., Geiger, A., and Nowozin, S. Which training methods for gans do actually converge? In Proc. Int. Conf. on Machine Learning (ICML), pp. 3478–3487, 2018.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

Mondal, A. K., Bhattacharjee, A., Mukherjee, S., Asnani, H., Kannan, S., and P., P. A. C-MI-GAN : Estimation of conditional mutual information using minmax formulation. In Proc. Conf. on Uncertainty in Artificial Intelligence (UAI), pp. 849–858, 2020.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. ICVGIP '08, pp. 722–729, USA, 2008. IEEE Computer Society.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In Proc. European Conf. on Computer Vision, pp. 69–84. Springer, 2016.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In Proc. Conf. Assoc. for Computational Linguistics (ACL), pp. 311–318, 2002.

Poole, B., Ozair, S., van den Oord, A., Alemi, A., and Tucker, G. On variational bounds of mutual information. In Proc. Int. Conf. on Machine Learning (ICML), pp. 5171–5180, 2019.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. OpenAI Blog, 2019.

Rainforth, T., Kosiorek, A. R., Le, T. A., Maddison, C. J., Igl, M., Wood, F., and Teh, Y. W. Tighter variational bounds are not necessarily better. In Proc. Int. Conf. on Machine Learning (ICML), pp. 4274–4282, 2018.

Rhodes, B., Xu, K., and Gutmann, M. U. Telescoping density-ratio estimation. arXiv preprint arXiv:2006.12204, 2020.

Rubin, D. B. The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. Journal of the American Statistical Association, 82(398):543–546, 1987.

Skare, Ø., Bølviken, E., and Holden, L. Improved sampling-importance resampling and reduced bias importance sampling. Scandinavian Journal of Statistics, 30(4):719–737, 2003.

Song, J. and Ermon, S. Understanding the limitations of variational mutual information estimators. In Proc. Int. Conf. on Learning Representations (ICLR), 2020a.

Song, J. and Ermon, S. Multi-label contrastive predictive coding. In Proc. Conf. on Neural Information Processing Systems (NeurIPS), 2020b.

Stratos, K. Mutual information maximization for simple and accurate part-of-speech induction. In Proc. Conf. Assoc. for Computational Linguistics (ACL), pp. 1095–1104, 2019.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. arXiv preprint arXiv:1906.05849, 2019.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning. arXiv preprint arXiv:2005.10243, 2020.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On mutual information maximization for representation learning. In Proc. Int. Conf. on Learning Representations (ICLR), 2020.

Velickovic, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. Deep graph infomax. In Proc. Int. Conf. on Learning Representations (ICLR), 2019.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Proc. Int. Conf. on Machine Learning (ICML), pp. 9929–9939, 2020.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-ucsd birds 200. 2010.

Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., and Weston, J. Neural text generation with unlikelihood training. In Proc. Int. Conf. on Learning Representations (ICLR), 2020.

Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. Transfertransfo: A transfer learning approach for neural network based conversational agents. In Proc. Conf. on Neural Information Processing Systems (NeurIPS) CAI Workshop, 2019.

Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., and Wang, L. Learning to navigate for fine-grained classification. In Proc. of the European Conf. on Computer Vision (ECCV), pp. 420–435, 2018.

Zhang, Y., Galley, M., Gao, J., Gan, Z., Li, X., Brockett, C., and Dolan, B. Generating informative and diverse conversational responses via adversarial information maximization. In Proc. Conf. on Neural Information Processing Systems (NeurIPS), pp. 1815–1825, 2018.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. DIALOGPT : Large-scale generative pre-training for conversational response generation. In Proc. Conf. Assoc. for Computational Linguistics (ACL), pp. 270–278, 2020.