Predicting What You Already Know Helps: Provable Self-Supervised Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Self-supervised representation learning solves auxiliary prediction tasks (known as pretext tasks), that do not require labeled data, to learn semantic representations. These pretext tasks are created solely using the input features, such as predicting a missing image patch, recovering the color channels of an image from context, or predicting missing words in text, yet predicting this *known* information helps in learning representations effective for downstream prediction tasks. This paper posits a mechanism based on approximate conditional independence to formalize how solving certain pretext tasks can learn representations that provably decrease the sample complexity of downstream supervised tasks. Formally, we quantify how the approximate independence between the components of the pretext task (conditional on the label and latent variables) allows us to learn representations that can solve the downstream task with drastically reduced sample complexity by just training a linear layer on top of the learned representation.

1 Introduction

Self-supervised learning revitalizes machine learning models in computer vision, language modeling, and control problems (see reference therein [30, 32, 11, 53, 29]). Training a model with auxiliary tasks based only on input features reduces the extensive costs of data collection and semantic annotations for downstream tasks. It is also known to improve the adversarial robustness of models [24, 7, 8]. Self-supervised learning creates pseudo labels solely based on input features, and solves auxiliary prediction tasks in a supervised manner (pretext tasks). However, the underlying principles of self-supervised learning are mysterious since it is a-priori unclear why predicting what we already know should help. We thus raise the following question:

What conceptual connection between pretext and downstream tasks ensures good representations? What is a good way to quantify this?

As a thought experiment, consider a simple downstream task of classifying desert, forest, and sea images. A meaningful pretext task is to predict the background color of images (known as image colorization [56]). Denote X_1, X_2, Y to be the input image, color channel, and the downstream label respectively. Given knowledge of the label Y, one can possibly predict the background X_2 without knowing much about X_1. In other words, X_2 is approximately independent of X_1 conditional on the label Y. Consider another task of inpainting [40] the front of a building (X_2) from the rest (X_1). While knowing the label "building" (Y) is not sufficient for successful inpainting, adding additional latent variables Z such as architectural style, location, window positions, etc. will ensure that variation in X_2 given Y, Z is small. We can mathematically interpret this as X_1 being approximate conditionally independent of X_2 given Y, Z.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

In the above settings with conditional independence, the only way to solve the pretext task for X_1 is to first implicitly predict Y and then predict X_2 from Y. Even without labeled data, the information of Y is hidden in the prediction for X_2.

Contributions. We propose a mechanism based on approximate conditional independence (ACI) to explain why solving pretext tasks created from known information can learn representations that provably reduce downstream sample complexity. For instance, learned representation will only require $\tilde{\mathcal{O}}(k)$ samples to solve a k-way supervised task under conditional independence (CI). Under ACI (quantified by the norm of a certain partial covariance matrix), we show similar sample complexity improvements.

Related work. There has been a flurry of self-supervised methods lately. One class of methods reconstruct images from corrupted or incomplete versions of it, like denoising auto-encoders [51], image inpainting [40], and split-brain autoencoder [57]. Pretext tasks are also created using visual common sense, including predicting relative position [18, 12, 38], recovering color channels [56], and discriminating images created from distortion [13]. We refer to the above procedures as **reconstruction-based SSL**. Other popular SSL paradigms include contrastive learning and language modeling based methods for text. Our work initiates theoretical understanding for reconstruction-based SSL. Related to our work is the recent theoretical analyses of contrastive learning. [3, 47, 48] that show guarantees for representations from contrastive learning on *linear classification* tasks using different kinds of CI like assumptions. CI and redundancy assumptions on multiple views [31, 2] are used to analyze a canonical-correlation based dimension reduction algorithm. More details are presented in Section A.

2 Preliminary

We use lower case symbols (x) to denote scalar quantities, bold lower case symbols ($\boldsymbol{x}$) for vector values, capital letters (X) for random variables, and capital and bold letters $\boldsymbol{X}$ for matrices. P_X denotes the probability law of random variable X.We use standard $\mathcal{O}$ notation to hide universal factors and $\tilde{\mathcal{O}}$ to hide log factors. $\|\cdot\|$ stands for ℓ_2-norm for vectors or Frobenius norm for matrices.

Linear conditional expectation. $\mathbb{E}^L[Y|X]$ denotes the best linear predictor of Y given X, while $\mathbb{E}[Y|X] \equiv \min_f \mathbb{E}[\|Y - f(X)\|^2]$ is the best predictor of Y given X.

(Partial) covariance matrix. For random variables X, Y, we denote $\boldsymbol{\Sigma}_{XY}$ to be covariance matrix of X and Y. For simplicity in most cases, we assume $\mathbb{E}[X] = 0$ and $\mathbb{E}[Y] = 0$; thus we do not distinguish $\mathbb{E}[XY]$ and $\boldsymbol{\Sigma}_{XY}$. The partial covariance matrix between X and Y given Z is: $\boldsymbol{\Sigma}_{XY|Z} := \boldsymbol{\Sigma}_{XY} - \boldsymbol{\Sigma}_{XZ}\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\Sigma}_{ZY}$.

Sub-gaussian random vectors. A random vector $X \in \mathbb{R}^d$ is ρ^2-sub-gaussian if for every fixed unit vector $\boldsymbol{v} \in \mathbb{R}^d$, the variable $\boldsymbol{v}^\top X$ is ρ^2-sub-gaussian, i.e., $\mathbb{E}[e^{s\cdot \boldsymbol{v}^\top (X - \mathbb{E}[X])}] \le e^{s^2\rho^2/2}$ ($\forall s \in \mathbb{R}$).

2.1 Setup and methodology

We denote by X_1 the input variable, X_2 the target random variable for the pretext task, and Y the label for the downstream task, with $X_1 \in \mathcal{X}_1 \subset \mathbb{R}^{d_1}, X_2 \in \mathcal{X}_2 \subset \mathbb{R}^{d_2}$ and $Y \in \mathcal{Y} \subset \mathbb{R}^k$. If $\mathcal{Y}$ is finite with $|\mathcal{Y}| = k$, we assume $\mathcal{Y} \subset \mathbb{R}^k$ is the one-hot encoding of the labels. $P_{X_1X_2Y}$ denotes the joint distribution over $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}$. P_{X_1Y}, P_{X_1} denote the corresponding marginal distributions. Our proposed self-supervised learning procedure is as follows:

Step 1 (pretext task): Learn representation $\psi(\boldsymbol{x}_1)$ through $\psi := \arg\min_{g\in\mathcal{H}} \mathbb{E}\,\|X_2 - g(X_1)\|_F^2$, where $\mathcal{H}$ can be different for different settings that we will specify and discuss later.

Step 2 (downstream task): Perform linear regression on Y with $\psi(X_1)$, i.e. $f(\boldsymbol{x}_1) := (\boldsymbol{W}^*)^\top \psi(\boldsymbol{x}_1)$, where $\boldsymbol{W}^* \leftarrow \arg\min_{\boldsymbol{W}} \mathbb{E}_{X_1,Y}[\|Y - \boldsymbol{W}^\top\psi(X_1)\|^2]$. Namely we learn $f(\cdot) = \mathbb{E}^L[Y|\psi(\cdot)]$.

Performance of the learned representation on the downstream task depends on the following quantities.
Approximation error. We measure this for a learned representation ψ by learning a linear function on top of it for the downstream task. Denote $e_{\text{apx}}(\psi) = \min_{\boldsymbol{W}} \mathbb{E}[\|f^*(X_1) - \boldsymbol{W}\psi(X_1)\|^2]$, where $f^*(\boldsymbol{x}_1) = \mathbb{E}[Y|X_1 = \boldsymbol{x}_1]$ is the optimal predictor for the task. This gives a measure of how well ψ can do with when given infinite samples for the task.

Estimation error. We measure sample complexity of ψ on the downstream task and assume access to n_2 i.i.d. samples $(\boldsymbol{x}_1^{(1)}, \boldsymbol{y}^{(1)}), \cdots, (\boldsymbol{x}_1^{(n_2)}, \boldsymbol{y}^{(n_2)})$ drawn from P_{X_1Y}. We express the n_2 samples collectively as $\boldsymbol{X}_1^{\text{down}} \in \mathbb{R}^{n_2 \times d_1}$, $\boldsymbol{Y} \in \mathbb{R}^{n_2 \times k}$ and overload notation to say $\psi(\boldsymbol{X}_1^{\text{down}}) = [\psi(\boldsymbol{x}_1^{(1)}) | \psi(\boldsymbol{x}_1^{(2)}) \cdots | \psi(\boldsymbol{x}_1^{(n_2)})]^\top \in \mathbb{R}^{n_2 \times d_2}$. We perform linear regression on the learned representation ψ and are interested in the excess risk that measures generalization.

$$\hat{\boldsymbol{W}} \leftarrow \arg\min_{\boldsymbol{W}} \frac{1}{2n_2} \|\boldsymbol{Y} - \psi(\boldsymbol{X}_1^{\text{down}})\boldsymbol{W}\|_F^2; \;\; \text{ER}_\psi(\hat{\boldsymbol{W}}) \coloneqq \frac{1}{2}\,\mathbb{E}\,\|f^*(X_1) - \hat{\boldsymbol{W}}^\top \psi(X_1)\|_2^2$$

3 Guaranteed recovery with approximate conditional independence

In this section, we first focus on the case when input X_1 and pretext target X_2 are conditionally independent (CI) given the downstream label Y. While this is a strong assumption that is rarely satisfied in practice, it helps us understand the role of CI with clean results and builds up to our main results with ACI with latent variables in Section E. We show how CI helps under two settings: (a) when the function class used for ψ is universal, (b) when ψ is restricted to be a linear function of given features. For now we assume access to a large amount of unlabeled data so as to learn the optimal ψ^* perfectly and this will be relaxed later in Section E. The general recipe for the results is as follows:

1. Find a closed-form expression for the optimal solution ψ^* for the pretext task.
2. Use conditional independence to argue that $e_{\text{apx}}(\psi^*)$ is small.
3. Exploit the low rank structure of ψ^* to show small estimation error on downstream tasks.

Data assumption. Suppose $Y = f^*(X_1) + N$, where $f^* = \mathbb{E}[Y|X_1]$ and hence $\mathbb{E}[N] = 0$. We assume N is σ^2-subgaussian. For simplicity, we assume non-degeneracy: $\boldsymbol{\Sigma}_{X_iX_i}, \boldsymbol{\Sigma}_{YY}$ are full rank.

Assumption 3.1. *Let $X_1 \in \mathbb{R}^{d_1}, X_2 \in \mathbb{R}^{d_2}$ be random variables from some unknown distribution. Let label $Y \in \mathcal{Y}$ be a discrete random variable with $k = |\mathcal{Y}| < d_2$. We assume conditional independence: $X_1 \perp X_2 | Y$.*

Here Y can be interpreted as the multi-class labels where k is the number of classes. For regression problems, one can think about Y as the discretized values of continuous labels. We do not specify the dimension for Y since Y could be arbitrarily encoded but the results only depend on k and the variance of Y (conditional on the input X_1).

Universal function class. Suppose we learn the optimal ψ^* among all measurable functions The optimal function ψ^* in this case is naturally given by conditional expectation: $\psi^*(\boldsymbol{x}_1) = \mathbb{E}[X_2|X_1 = \boldsymbol{x}_1]$. We now show that CI implies that ψ^* is good for downstream tasks, which is not apriori clear.

Lemma 3.1 (Approximation error)**.** *Suppose random variables X_1, X_2, Y satisfy Assumption 3.1, and matrix $\boldsymbol{A} \in \mathbb{R}^{\mathcal{Y} \times d_2}$ with $\boldsymbol{A}_{y,:} \coloneqq \mathbb{E}[X_2|Y = \boldsymbol{y}]$ is of rank $k = |\mathcal{Y}|$. Then $e_{apx}(\psi^*) = 0$.*

This tells us that although f^* could be nonlinear in $\boldsymbol{x}_1$, it is guaranteed to be linear in $\psi^*(\boldsymbol{x}_1)$. Note that Y does not have to be linear in X_2. We provide this simple example for better understanding:

Example 3.1. *Let $Y \in \{-1, 1\}$ be binary labels, and X_1, X_2 be $2-$mixture Gaussian random variables with $X_1 \sim \mathcal{N}(Y\boldsymbol{\mu}_1, \mathbf{I}), X_2 \sim \mathcal{N}(Y\boldsymbol{\mu}_2, \mathbf{I})$. In this example, $X_1 \perp X_2 | Y$. Although $f^* = \mathbb{E}[Y|X_2]$ is not linear, $\mathbb{E}[Y|\psi]$ is linear: $\psi(\boldsymbol{x}_1) = P(Y = 1|X_1 = \boldsymbol{x}_1)\boldsymbol{\mu}_2 - P(Y = -1|X_1 = \boldsymbol{x}_1)\boldsymbol{\mu}_2$ and $f^*(\boldsymbol{x}_1) = P(Y = 1|X_1 = \boldsymbol{x}_1) - P(Y = -1|X_1 = \boldsymbol{x}_1) \equiv \boldsymbol{\mu}_2^T \psi(\boldsymbol{x}_1)/\|\boldsymbol{\mu}_2\|^2$.*

Given that ψ^* is good for downstream, we now care about the sample complexity. We will need to assume that the representation has some nice concentration properties. We make an assumption about the whitened data $\psi^*(X_1)$ to ignore scaling factors.

Assumption 3.2. *We assume the whitened feature variable $U \coloneqq \boldsymbol{\Sigma}_\psi^{-1/2}\psi(X_1)$ is a ρ^2-subgaussian random variable, where $\boldsymbol{\Sigma}_\psi = \mathbb{E}[\psi(X_1)\psi(X_1)^\top]$.*

We note that all bounded random variables satisfy sub-gaussian property.

Theorem 3.2 (General conditional independence)**.** *Fix a failure probability $\delta \in (0, 1)$, under the same assumption as Lemma 3.1 and Assumption 3.2 for ψ^*, if additionally $n \gg \rho^4(k + \log(1/\delta))$,*

then the excess risk of the learned predictor $\boldsymbol{x}_1 \to \hat{\boldsymbol{W}}^\top \psi^*(\boldsymbol{x}_1)$ *on the downstream task satsifies:*

$$\mathrm{ER}_{\psi^*}[\hat{\boldsymbol{W}}] \leq \mathcal{O}\left(\frac{k+\log(k/\delta)}{n_2}\sigma^2\right).$$

Function class induced by feature maps. Given feature map $\phi_1 : \mathcal{X}_1 \to \mathbb{R}^{D_1}$, we consider the function class $\mathcal{H}_1 = \{\psi : \mathcal{X}_1 \to \mathbb{R}^{d_2} | \exists \boldsymbol{B} \in \mathbb{R}^{d_2 \times D_1}, \psi(\boldsymbol{x}_1) = \boldsymbol{B}\phi_1(\boldsymbol{x}_1)\}$.

Claim 3.3 (Closed form solution)**.** *The optimal function in* $\mathcal{H}$ *is* $\psi^*(\boldsymbol{x}_1) = \boldsymbol{\Sigma}_{X_2\phi_1}\boldsymbol{\Sigma}_{\phi_1\phi_1}^{-1}\boldsymbol{x}_1$*, where* $\boldsymbol{\Sigma}_{X_2\phi_1} := \boldsymbol{\Sigma}_{X_2\phi_1(X_1)}$ *and* $\boldsymbol{\Sigma}_{\phi_1\phi_1} := \boldsymbol{\Sigma}_{\phi_1(X_1)\phi_1(X_1)}$.

We again show the benefit of CI, this time only comparing the performance of ψ^* to the original features ϕ_1. Since ψ^* is linear in ϕ_1, it cannot have smaller approximation error than ϕ_1. However CI will ensure that ψ^* has the same approximation error as ϕ_1 and enjoys much better sample complexity.

Lemma 3.4 (Approximation error)**.** *If Assumption 3.1 is satisfied, and if the matrix* $\boldsymbol{A} \in \mathbb{R}^{\mathcal{Y}\times d_2}$ *with* $\boldsymbol{A}_{y,:} := \mathbb{E}[X_2|Y = \boldsymbol{y}]$ *is of rank* $k = |\mathcal{Y}|$*. Then* $e_{apx}(\psi^*) = e_{apx}(\phi_1)$.

We additionally need an assumption on the residual $a(\boldsymbol{x}_1) := \mathbb{E}[Y|X_1 = \boldsymbol{x}_1] - \mathbb{E}^L[Y|\phi_1(\boldsymbol{x}_1)]$.

Assumption 3.3. *(Bounded approx. error; Condition 3 in [26])) We have almost surely*

$$\|\boldsymbol{\Sigma}_{\phi_1\phi_1}^{-1/2}\phi_1(X_1)a(X_1)^\top\|_F \leq b_0\sqrt{k}$$

Theorem 3.5. *(CI with approximation error) Fix a failure probability* $\delta \in (0,1)$*, under the same assumption as Lemma 3.4, Assumption 3.2 for* ψ^* *and Assumption 3.3, if* $n_2 \gg \rho^4(k+\log(1/\delta))$*, then the excess risk of the learned predictor* $\boldsymbol{x}_1 \to \hat{\boldsymbol{W}}^\top\psi^*(\boldsymbol{x}_1)$ *on the downstream task satisfies:*

$$\mathrm{ER}_{\psi^*}[\hat{\boldsymbol{W}}] \leq e_{apx}(\phi_1) + \mathcal{O}\left(\frac{k+\log(k/\delta)}{n_2}\sigma^2\right).$$

Theorem 3.5 is also true with Assumption B.3 instead of exact CI, if we replace k by km. Therefore with SSL, the requirement of labels is reduced from complexity for $\mathcal{H}$ to $\mathcal{O}(k)$ (or $\mathcal{O}(km)$).

Remark 3.1. *We note that since* $X_1 \perp X_2|Y$ *ensures* $X_1 \perp h(X_2)|Y$ *for any deterministic function* h*, we could replace* X_2 *by* $h(X_2)$ *and all results hold. Therefore we could replace* X_2 *with* $h(X_2)$ *in our algorithm especially when* $d_2 < km$*.*

3.1 Beyond conditional independence

Informally we present the excess risk bound with further relaxed assumptions: 1) For pretext task we learn $\tilde{\psi}$ with finite samples and achieve: $\mathbb{E}\,\|\tilde{\psi}(X_1) - \psi^*(X_1)\|_F^2 \leq \epsilon_{\text{pre}}^2$; 2) Given Y and latent variable Z, X_1 and X_2 have small dependence that is captured by ϵ_{CI}. (A formal definition of ϵ_{CI} is given in Definition E.2.) Under this setting, if we learn a linear model trained on $\tilde{\psi}(\boldsymbol{X}_1^{\text{down}})$:

$$\hat{\boldsymbol{W}} \leftarrow \arg\min_{\boldsymbol{W}} \frac{1}{2n_2}\|\boldsymbol{Y} - \tilde{\psi}(\boldsymbol{X}_1^{\text{down}})\boldsymbol{W}\|_F^2,\ \mathrm{ER}_{\tilde{\psi}}(\hat{\boldsymbol{W}}) := \mathbb{E}_{X_1}\,\|f_{\mathcal{H}}^*(X_1) - \hat{\boldsymbol{W}}^\top\tilde{\psi}(X_1)\|_2^2.$$

Here $f_{\mathcal{H}}^*$ is the best function in the function class $\mathcal{H}$. Then we get:

$$\mathrm{ER}_{\tilde{\psi}}(\hat{\boldsymbol{W}}) \leq \tilde{O}\left(\frac{d_2}{n_2} + \epsilon_{\text{CI}}^2 + \epsilon_{\text{pre}}^2\right).$$

Here d_2 could be improved to k (or km with latent variables $Z \in \mathcal{Z}, |\mathcal{Z}| = m$) with principle component regression. We defer the formal statements and proofs to Appendix E. Finally, we also show empirical validations of our main results in Appendix I.

4 Conclusion

In this work we theoretically quantify how an approximate conditional independence assumption that connects pretext and downstream task data distributions can give sample complexity benefits of self-supervised learning on downstream tasks. Our theoretical findings are also supported by experiments on simulated data and also on real CV and NLP tasks. We would like to note that approximate CI is only a sufficient condition for a useful pretext task. We leave it for future work to investigate other mechanisms by which pretext tasks help with downstream tasks.

References

[1] Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

[2] Rie Kubota Ando and Tong Zhang. Two-view feature generation model for semi-supervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 25–32, 2007.

[3] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[4] Charles R Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

[5] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

[6] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[7] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11190–11201, 2019.

[8] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. *arXiv preprint arXiv:2003.12862*, 2020.

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

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[11] 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.

[12] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[13] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015.

[14] Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

[15] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017.

[16] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

[17] Kenji Fukumizu, Francis R Bach, Michael I Jordan, et al. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009.

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

[19] Shiry Ginosar, Kate Rakelly, Sarah Sachs, Brian Yin, and Alexei A Efros. A century of portraits: A visual historical record of american high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–7, 2015.

[20] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[21] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[22] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

[23] 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*, pages 770–778, 2016.

[24] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pages 15637–15648, 2019.

[25] 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.

[26] Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1, 2012.

[27] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

[28] Tzee-Ming Huang. Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, 38(4):2047–2091, 2010.

[29] Eric Jang, Coline Devin, Vincent Vanhoucke, and Sergey Levine. Grasp2vec: Learning object representations from self-supervised grasping. *arXiv preprint arXiv:1811.06964*, 2018.

[30] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[31] Sham M Kakade and Dean P Foster. Multi-view regression via canonical correlation analysis. In *International Conference on Computational Learning Theory*, pages 82–96. Springer, 2007.

[32] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1920–1929, 2019.

[33] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *Proceedings of the International Conference on Learning Representations*, 2018.

[34] Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[35] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

[36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 2013.

[37] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

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

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

[40] 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*, pages 2536–2544, 2016.

[41] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[42] Michael Reed. *Methods of modern mathematical physics: Functional analysis*. Elsevier, 2012.

[43] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[44] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[45] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013.

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

[47] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive estimation reveals topic posterior information to linear models. *arXiv preprint arXiv:2003.02234*, 2020.

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

[49] 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.

[50] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[51] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

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

[53] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[54] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018.

[55] Han Yang, Xiao Yan, Xinyan Dai, and James Cheng. Self-enhanced gnn: Improving graph neural networks using model outputs. *arXiv preprint arXiv:2002.07518*, 2020.

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

[57] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

[58] Zaiwei Zhang, Zhenxiao Liang, Lemeng Wu, Xiaowei Zhou, and Qixing Huang. Path-invariant map networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11084–11094, 2019.