Contrastive Multiview Coding

Yonglong Tian [1] **Dilip Krishnan** [2] **Phillip Isola** [1]

Abstract

Humans view the world through many sensory channels, e.g., the long-wavelength light channel, seen by the left eye, or the high-frequency vibrations channel, heard by the right ear. Each view (such as light or sound) is noisy and incomplete, but important factors, such as physics, geometry, and semantics, tend to be shared between all views (e.g., a "dog" can be seen, heard, and felt). We hypothesize that a powerful neural representation is one that models view-invariant factors. Based on this hypothesis, we investigate a contrastive coding scheme, in which a deep representation is learned that aims to maximize mutual information between different views but is otherwise compact. Our approach scales to any number of views, and is view-agnostic. The resulting learned representations perform above the state of the art for downstream tasks such as object classification, compared to formulations based on predictive learning or single view reconstruction, and their performance improves as more views are added.

1. Introduction

A foundational idea in coding theory is to learn compressed representations that nonetheless can be used to reconstruct the raw data. This idea shows up in contemporary representation learning in the form of autoencoders (Salakhutdinov & Hinton, 2009) and generative models (Kingma & Welling, 2013; Goodfellow et al., 2014), which try to represent a data point or distribution as losslessly as possible. Yet lossless representation might not be what we really want; instead we might prefer to keep the "good" information (signal) and throw away the bad (noise).

To an autoencoder, or a maximum likelihood generative model, a bit is a bit. No one bit is better than any other. Our conjecture is that some bits *are* in fact better than others. Some bits code important properties like semantics, physics, and geometry, while others code attributes that we might consider less important, like incidental lighting conditions or thermal noise in a camera's sensor.

We hypothesize that the good bits are the ones that are shared between multiple *views* of the world, for example between multiple sensory modalities like vision, sound, and touch. Under this perspective, "presence of dog" is good information, since dogs can be seen, heard, and felt, but "camera pose" is bad information, since a camera's pose has little or no effect on the acoustic and tactile properties of the imaged scene.

Our goal is therefore to learn representations that capture information shared between multiple sensory views but that are otherwise compact (i.e. throw away the bad information). To do so, we employ contrastive learning, where we learn a feature embedding such that multiple views of the same scene map to nearby points while views of different scenes map to far apart points. In particular, we adapt the recently proposed method of Contrastive Predictive Coding (CPC) (Oord et al., 2018), except we simplify it – removing the recurrent network – and generalize it – show how to apply it to arbitrary collections of views, rather than just to temporal predictions. In reference to CPC, we term our method *Contrastive Multiview Coding* (CMC). The contrastive objective in our formulation, as in CPC, is based on Noise Contrastive Estimation (NCE) (Gutmann & Hyvärinen, 2010). This objective can be understood as attempting to maximize the mutual information between the representations of each view.

The core ideas that we build on: contrastive learning, mutual information maximization, and deep representation learning, are not new and have been explored in the literature on representation and multiview learning (Li et al., 2018; Xu et al., 2013; Arora et al., 2019). Our main contribution is to set up a framework to extend these ideas to any number of views. We show significant benefits to the learned representations, in terms of transfer to tasks such as object recognition.

2. Contrastive Multiview Coding

We consider a collection of M views of the data, denoted as $V_1, \dots, V_M$. For each view V_i, we denote v_i as a random

[1] MIT CSAIL [2] Google Research. Correspondence to: Yonglong Tian <yonglong@mit.edu>.

ICML 2019 workshop on Self-Supervised Learning, Long Beach, California. Copyright 2019 by the author(s).

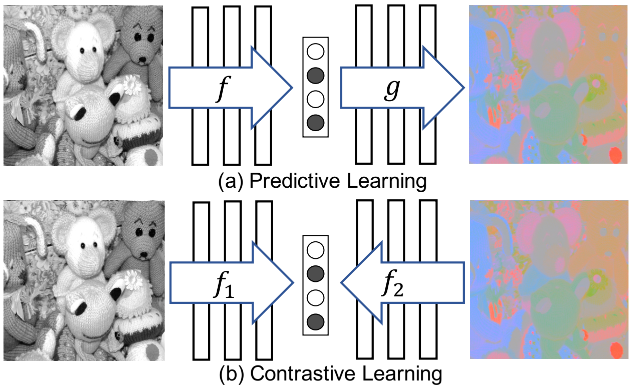

Figure 1. Predictive Learning vs Contrastive Learning. The predictive learning framework (**Top**) learns representations by predicting one view from another view. While in the contrastive learning umbrella, representations are learnt by contrasting paired and unpaired views in latent space.

variable representing samples following $v_i \sim \mathcal{P}(V_i)$.

2.1. Preliminary: Predictive Learning

Let V_1 and V_2 represent two views of a dataset. For instance, V_1 might be the luminance of a particular image and V_2 the chrominance. We define the *predictive learning* setup as a deep nonlinear transformation from v_1 to v_2 through latent variables z, as shown in Fig. 1. Formally, $z = f(v_1)$ and $\hat{v_2} = g(z)$, where f and g represent the encoder and decoder respectively and $\hat{v_2}$ is the prediction of v_2 given v_1. The parameters of the encoder and decoder models are then trained using an objective function that tries to bring $\hat{v_2}$ "close to" v_2. Simple examples of such an objective include the $\mathcal{L}_1$ or $\mathcal{L}_2$ loss functions. The predictive approach has been extensively used in representation learning, for example, colorization (Zhang et al., 2016; 2017) and predicting sound from vision (Owens et al., 2016).

2.2. Contrasting with Two Views

The idea behind contrastive learning is learning by discriminating, or comparing between samples from different distributions. The contrastive loss (Hadsell et al., 2006) sets a margin to embed semantically similar samples close to each other and dissimilar samples far apart. Recently, the InfoNCE loss (Oord et al., 2018) trained with softmax yields a score function in favor of temporally congruent samples.

Given a dataset of V_1 and V_2 that consists of a collection of samples $\{v_1^i, v_2^i\}_{i=1}^N$, we consider contrasting congruent and incongruent pairs. Formally, we refer to samples from the joint distribution as positives, i.e., $x \sim p(v_1, v_2)$ or $x = \{v_1^i, v_2^i\}$, and samples from the product of marginals as negatives, i.e., $y \sim p(v_1)p(v_2)$ or $y = \{v_1^i, v_2^j\}$.

We learn a score function $h_\theta(\cdot)$ favoring positive samples x but disfavoring negative samples y. This function is trained to correctly select a single positive sample x out of a set $S = \{x, y_1, y_2, ..., y_k\}$ which contains k other negatives. The objective we minimize is:

$$\mathcal{L}_{contrast} = -\mathop{\mathbb{E}}_{S}\left[\log \frac{h_\theta(x)}{h_\theta(x) + \sum_{i=1}^{k} h_\theta(y_i)}\right] \tag{1}$$

We implement this score function $h_\theta(\cdot)$ as a neural network. To extract compact latent representations of v_1 and v_2, we employ two encoders $f_{\theta_1}(\cdot)$ and $f_{\theta_2}(\cdot)$ with parameters θ_1 and θ_2 respectively. The latent represention are extracted as $z_1 = f_{\theta_1}(v_1)$, $z_2 = f_{\theta_2}(v_2)$. On top of these features, the score is computed as the exponential of a bivariate function of z_1 and z_2, which here is a bilinear function parameterized by W_{12}. Then $\theta = \{\theta_1, \theta_2, W_{12}\}$ and $h_\theta(\cdot)$ is:

$$h_\theta(\{v_1, v_2\}) = e^{f_{\theta_1}(v_1)^T W_{12} f_{\theta_2}(v_2)} \tag{2}$$

The contrastive learning paradigm maximizes the mutual information between the variable z_1 and z_2. A proof is given by (Oord et al., 2018), demonstrating that:

$$I(z_1; z_2) \geq \log(k) - \mathcal{L}_{contrast} \tag{3}$$

We now put recent related work in the above framework. CPC (Oord et al., 2018) considers 2-view models; our method coincides with CPC in the specific case of two views and sequential data. Deep Infomax (DIM) (Hjelm et al., 2019) maximizes the mutual information between the input and output of a neural net f, i.e. it maximizes $I(x, f(x))$. Both these methods learn a representation from two views but these are two views of the same underlying sensory modality. We apply CMC to the case where each view is of a different physical signal. This approach may be advantageous as many nuisance factors, such as sensor noise, are shared within but not across modalities.

2.3. More than Two Views

In this work, we present more general formulations of Eq. 1 which can handle any number of views. Such formulations include "core view" and "full graph" paradigms, which offer different tradeoffs between efficiency and effectiveness. These formulations are visualized in Fig. 2.

Suppose we have a collection of M views $V_1, \ldots, V_M$. The "core view" formulation sets apart one view that we want to optimize over, say V_1, and builds pair-wise representations between V_1 and each other view $V_j, j > 1$, by optimizing the sum of a set of pair-wise objectives:

$$\mathcal{L}_C = \sum_{j=2}^{M} \mathcal{L}_{contrast}(V_1, V_j) \tag{4}$$

A second formulation is the "full graph" where we consider all pairs $(i, j), i \neq j$, and build $\binom{n}{2}$ relationships in all. By

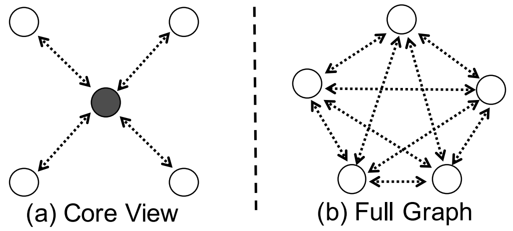

Figure 2. We generalize our CMC to M views following two different formulations: (1) "core view" specifies one view, and all other views are contrasted against that view; (2) "full graph" contrasts every pair and representations for all views are jointly learned.

involving all pairs, the objective function that we optimize is:

$$\mathcal{L}_F = \sum_{1 \leq i < j \leq M} \mathcal{L}_{contrast}(V_i, V_j) \tag{5}$$

2.4. Implementation

Better representations using $\mathcal{L}_{contrast}$ in Eq. 1 are learnt by using many negative samples. To avoid very large batch sizes, we consider two practical implementations introduced by previous methods (Wu et al., 2018; Hjelm et al., 2019). (1) **Memory-based.** We maintain a dynamic memory bank to store latent features for each data sample. Therefore, during training we can efficiently retrieve large number of negative pairs from the memory bank without recomputing the features. This allows us to approximate Eq. 1 with large amount of negatives by using NCE (Gutmann & Hyvärinen, 2010). (2) **Patch-based.** Instead of contrasting features from the last layer, patch-based method contrasts feature from the last layer with features from previous layers. For instance, we use features from the last layer of $f_{\theta_1}(\cdot)$ to contrast with feature points from feature maps produced by the first several conv layers of $f_{\theta_2}(\cdot)$. This is equivalent to contrast between global patch from one view with local patches from the other view.

3. Experiments

We first evaluate our CMC framework on image representation learning benchmarks. Then we extends it to more than two views and analyze it's effectiveness.

3.1. CMC on Images

Given a dataset of RGB images, we convert them to the *Lab* image color space, and split each image into *L* and *ab* channels, as originally proposed in SplitBrain autoencoders (Zhang et al., 2017). Each split represents a view of the orginal image and is passed through a seprate encoder. As in SplitBrain, we design these two encoders by evenly splitting a given deep network into sub-networks. By concatenating representations layer-wise from these two encoders, we achieve the final representation of an input image. The quality of such a representation is evaluated by freezing the weights of encoder and training linear or non-linear classifiers on top of each layer.

Method	classifier	conv5	fc7	Strided Crop
AE	MLP	62.19	55.78	-
BiGAN		71.53	67.18	-
SplitBrain[†]		72.35	63.15	-
DIM	MLP	72.57	70.00	76.97
CPC		-	-	77.81
CMC[†](Patch)	Linear	76.65	79.25	82.58
CMC[†](Patch)	MLP	80.14	80.11	**83.43**
CMC[†](Memory)	Linear	80.69	84.73	-
CMC[†](Memory)	MLP	**83.03**	**85.06**	-
Supervised		68.70		

Table 1. Classification accuracy on STL-10. For all methods we compare against, we include the numbers that are reported in the DIM (Hjelm et al., 2019) paper, except for SplitBrain, which is our reimplementation. Methods marked with † only have half the number of parameters because of splitting.

3.1.1. STL-10

Setup. We adopt the same data augmentation strategy and network architecture as those in DIM (Hjelm et al., 2019). For a fair comparison with DIM, the patch based contrastive loss is employed during unsupervised pre-training. With the weights of the pre-trained encoder frozen, a two-layer fully connected network (MLP) or a linear classifier is trained on top of different layers to perform 10-way classification. We also investigated the strided crop strategy proposed in CPC (Oord et al., 2018). At last, we evaluate the memory-based implementation for comparison.

The family of contrastive learning methods, such as DIM, CPC, and CMC, achieve higher classification accuracy than other methods such as SplitBrain that use predictive learning; or BiGAN that use adversarial learning. CMC significantly outperforms DIM and CPC. We hypothesize that this outperformance results from the modeling of cross-view mutual information, where view-specific noisy details are discarded. Finally, we notice that the predictive learning methods suffer from a big drop in performance when the encoding layer is switched from conv5 to fc7. On the other hand, the contrastive learning approaches are much more stable across layers. From a practical perspective, this is a significant advantage as the selection of specific layers should ideally not change downstream performance by too much.

3.1.2. ImageNet

Setup. To compare with other methods, we adopt standard AlexNet and split it into two encoders. Because of splitting, each layer only connects to half of the neurons in the previous layer, and therefore the number of parameters in our model halves. Two variants of CMC are considered:

	ImageNet Classification Accuracy			
Method	conv2	conv3	conv4	conv5
ImageNet-Labels	36.3	44.2	48.3	50.5
Random	17.1	16.9	16.3	14.1
(Krähenbühl et al., 2015)	23.0	24.5	23.2	20.6
(Doersch et al., 2015)	23.3	30.2	31.7	29.6
(Zhang et al., 2016)	24.8	31.0	32.6	31.8
(Noroozi & Favaro, 2016)	30.1	34.7	33.9	28.3
(Donahue et al., 2017)	24.5	31.0	29.9	28.0
(Zhang et al., 2017)†	29.3	35.4	35.2	32.8
(Noroozi et al., 2017)	30.6	34.3	32.5	25.7
(Wu et al., 2018)	26.5	31.8	34.1	35.6
(Gidaris et al., 2018)	31.7	**38.7**	38.2	36.5
(Caron et al., 2018)	29.2	38.2	39.8	36.1
CMC†(Patch)	30.8	34.2	37.5	38.1
CMC†(Memory)	**33.5**	38.1	**40.4**	**42.6**

Table 2. Top-1 classification accuracy on 1000 classes of ImageNet (Deng et al., 2009). We compare our CMC method with other approaches by training 1000-way linear classifiers on top of the feature maps of each layer, as proposed by (Zhang et al., 2016). Methods marked with † only have half the number of parameters compared to others, because of splitting.

patch-based and memory-based.

ImageNet classification task. Following (Zhang et al., 2016), we evaluate task generalization of the learned representation by training 1000-way *linear* classifiers on top of different layers. Table 2 shows the results of comparing the two variants of CMC against other models, both predictive and contrastive. The CMC Memory variant is the best among all these methods; futhermore the CMC methods tend to perform better at higher convolutional layers, similar to the other contrastive model Inst-Dis (Wu et al., 2018). The memory-based CMC model consistently performs better than the patch-based model due to the use of many more negative examples.

3.2. Extending CMC to More Views

We further extend our CMC learning framework to multiview scenarios. We experiment on the NYU-Depth-V2 (Nathan Silberman & Fergus, 2012) dataset. We focus more on understanding the behavior and effectiveness of CMC rather than competing with current state-of-the-arts. The views we consider are: luminance (L channel), chrominance (ab channel), depth, surface normal, and semantic labels.

Setup. To extract features from each view, we use a neural network with 5 convolutional layers, and 2 fully connected layer. As the size of the dataset is small, we adopt the patch-based contrastive objective to increase the number of negative pairs. Patches with a size of 128 × 128 are randomly cropped from the original images for contrastive learning.

To measure the quality of the learned representation, we consider the task of predicting semantic labels from the representation of L. We follow the *core view paradigm* and use L are the core view, thus learning a set of representations on L by contrasting different views with L. A UNet style architecture (Ronneberger et al., 2015) is utilized to perform the segmentation task. Contrastive training is performed on the above architecture that is equivalent of the UNet's encoder. After contrastive training is completed, we initialize the encoder weights of the UNet from the L encoder (which are equivalent architectures) and keep them frozen. Only the decoder is trained during this finetuning stage.

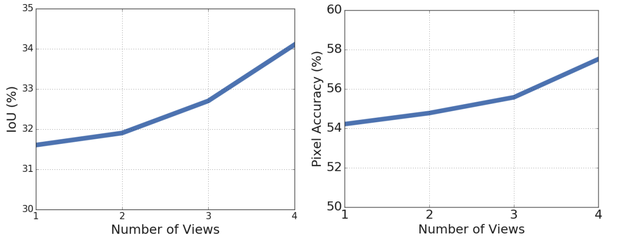

Figure 3. We show the Intersection over Union (IoU) (top) and Pixel Accuracy (bottom) for the NYU-Depth-V2 dataset, as CMC is trained with increasingly more views from 1 to 4. As more views are added, both these metrics steadily increase. The views are (in order of inclusion): L, ab, depth and surface normals.

	Pixel Accuracy (%)	mIoU (%)
Random	45.5	21.4
CMC	57.1	34.1
Supervised	**57.8**	**35.9**

Table 3. Results on the task of predicting semantic labels from L channel representation which is learnt using the patch-based contrastive loss and all 4 views. We compare CMC with *Random* and *Supervised* baselines, which serve as lower and upper bounds respectively.

Since we use the patch-based contrastive loss, in the 1 view setting case, CMC coincides with DIM (Hjelm et al., 2019). The 2-4 view cases contrast L with ab, and then sequentially add depth and surface normals, but in all cases, the patch based loss is used because the amount of data is small. The semantic labeling results are measured by mean IoU over all classes and pixel accuracy are shown in Fig. 3. We see that the performance steadily improves as new views are added. We have tested different orders of adding the views, and they all follow a similar pattern.

We also compare CMC with two baselines. First, we randomly initialize and freeze the encoder, and we call this the *Random* baseline; it serves as a lower bound since the representation is just a random projection. Rather than freezing the randomly initialized encoder, we could train it jointly with the decoder. This end-to-end *Supervised* baseline serves as an upper bound. The results are presented in Table 3, which shows CMC produces high quality feature maps even though it's unaware of the downstream task.

References

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, 2018.

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

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1422–1430, 2015.

Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial feature learning. In *International Conference on Learning Representations*, 2017.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. IEEE, 2006.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Krähenbühl, P., Doersch, C., Donahue, J., and Darrell, T. Data-dependent initializations of convolutional neural networks. *arXiv preprint arXiv:1511.06856*, 2015.

Li, Y., Yang, M., and Zhang, Z. M. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

Nathan Silberman, Derek Hoiem, P. K. and Fergus, R. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

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

Noroozi, M., Pirsiavash, H., and Favaro, P. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5898–5906, 2017.

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

Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., and Freeman, W. T. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2405–2413, 2016.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Salakhutdinov, R. and Hinton, G. Deep boltzmann machines. In *Artificial intelligence and statistics*, pp. 448–455, 2009.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

Xu, C., Tao, D., and Xu, C. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

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

Zhang, R., Isola, P., and Efros, A. A. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1058–1067, 2017.