Contrastive Multiview Coding

Authors: Yonglong Tian, Dilip Krishnan, Phillip Isola
Presented by: Hattie Zhou, Akshay Singh Rana

Motivation

Motivation

- We want to learn a representation that keeps the “good” information and throws away the “noise”

Motivation

- We want to learn a representation that keeps the “good” information and throws away the “noise”
- What is “good” information? What is “noise”?

Motivation - Multiview Learning

- We adopt the view that good bits are things that are shared across different views of the same scene

Motivation - Multiview Learning

- We adopt the view that good bits are things that are shared across different views of the same scene
- In contrast, noise is view-specific information that is not shared and not useful for differentiating between scenes

Motivation - Multiview Learning

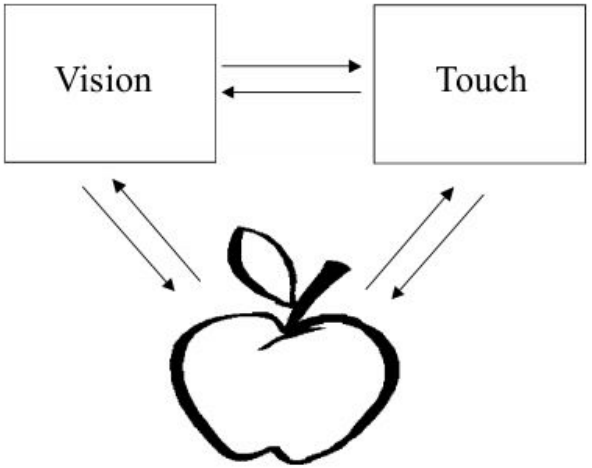


Figure 1. Illustration of the time-locked mappings of two sensory systems to the events in the world and to each other. Because visual and haptic systems actively collect information — by moving hands, by moving eyes — the arrows connecting these systems to each other also can serve as teaching signals for each other.

Smith & Gasser 2005. The Development of Embodied Cognition: Six Lessons from Babies

Motivation - Contrastive Learning

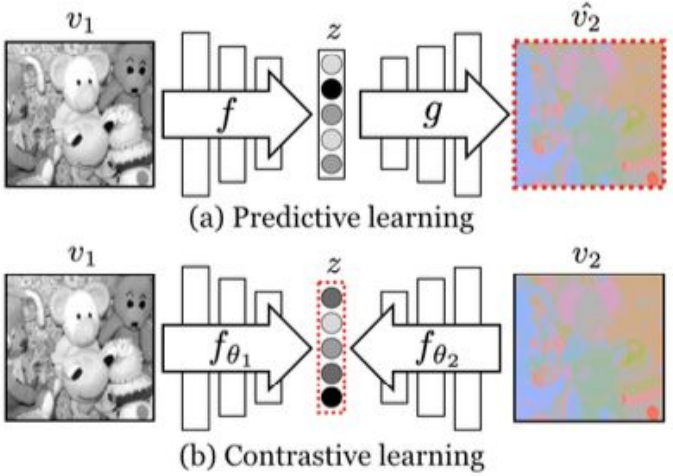
- We want to find features that are useful in distinguishing between different scenes, but are shared among similar scenes

Motivation - Contrastive Learning

- We want to find features that are useful in distinguishing between different scenes, but are shared among similar scenes
- Contrastive learning learns a representation that are close for positive pairings and far apart for negative pairings

Motivation - Contrastive Learning

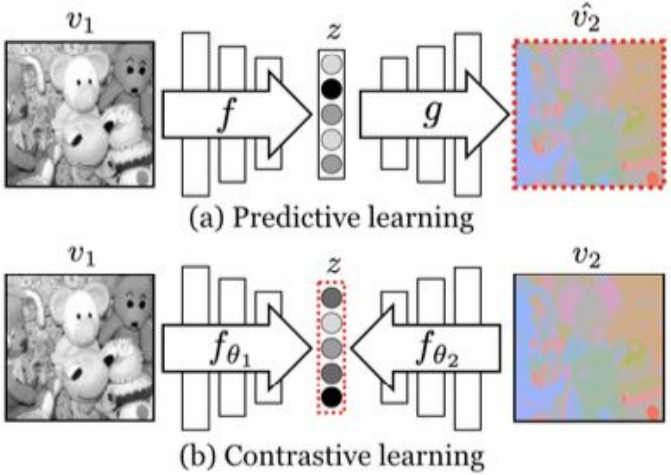
- We want to find features that are useful in distinguishing between different scenes, but are shared among similar scenes
- Contrastive learning learns a representation that are close for positive pairings and far apart for negative pairings



Method Details

Building on top of CPC

- This paper builds on top of Contrastive Predictive Coding, with two main differences:
  1. Removes the autoregressive component
  2. Generalizes CPC to multiple views



Loss Function

- InfoNCE objective:

$$\mathcal{L}_{contrast}^{V_1,V_2} = - \mathop{\mathbb{E}}_{\{v_1^1, v_2^1, \ldots, v_2^{k+1}\}} \left[\log \frac{h_\theta(\{v_1^1, v_2^1\})}{\sum_{j=1}^{k+1} h_\theta(\{v_1^1, v_2^j\})} \right] \quad (2)$$

- Cross entropy loss, essentially the probability that v_1^1 is the most similar to v_2^1 among all the examples

$$\log\left(\frac{e^{\text{similarity}(\,\,)}}{e^{\text{similarity}(\,\,)} + e^{\text{similarity}(\,\,)} + e^{\text{similarity}(\,\,)}} \right)$$

$$h_\theta(\{v_1, v_2\}) = \exp\left(\frac{f_{\theta_1}(v_1) \cdot f_{\theta_2}(v_2)}{\|f_{\theta_1}(v_1)\| \cdot \|f_{\theta_2}(v_2)\|} \cdot \frac{1}{\tau} \right)$$

Amit Chaudhary. The Illustrated SimCLR Framework.

Loss Function

- InfoNCE objective:

$$\mathcal{L}_{contrast}^{V_1,V_2} = - \mathop{\mathbb{E}}_{\{v_1^1, v_2^1, \ldots, v_2^{k+1}\}} \left[\log \frac{h_\theta(\{v_1^1, v_2^1\})}{\sum_{j=1}^{k+1} h_\theta(\{v_1^1, v_2^j\})} \right] \quad (2)$$

- Cross entropy loss, essentially the probability that v_1^1 is the most similar to v_2^1 among all the examples

$$\log\left(\frac{e^{\text{similarity}(\,\cdot\,,\,\cdot\,)}}{e^{\text{similarity}(\,\cdot\,,\,\cdot\,)} + e^{\text{similarity}(\,\cdot\,,\,\cdot\,)} + e^{\text{similarity}(\,\cdot\,,\,\cdot\,)}} \right)$$

$$h_\theta(\{v_1, v_2\}) = \exp\left(\frac{f_{\theta_1}(v_1) \cdot f_{\theta_2}(v_2)}{\|f_{\theta_1}(v_1)\| \cdot \|f_{\theta_2}(v_2)\|} \cdot \frac{1}{\tau} \right)$$

Amit Chaudhary. The Illustrated SimCLR Framework.

Loss Function

A Simple Framework for Contrastive Learning of Visual Representations

Name	Negative loss function	Gradient w.r.t. $\boldsymbol{u}$
NT-Xent	$\boldsymbol{u}^T\boldsymbol{v}^+/\tau - \log \sum_{\boldsymbol{v}\in\{\boldsymbol{v}^+,\boldsymbol{v}^-\}} \exp(\boldsymbol{u}^T\boldsymbol{v}/\tau)$	$(1 - \frac{\exp(\boldsymbol{u}^T\boldsymbol{v}^+/\tau)}{Z(\boldsymbol{u})})/\tau\boldsymbol{v}^+ - \sum_{\boldsymbol{v}^-} \frac{\exp(\boldsymbol{u}^T\boldsymbol{v}^-/\tau)}{Z(\boldsymbol{u})}/\tau\boldsymbol{v}^-$
NT-Logistic	$\log \sigma(\boldsymbol{u}^T\boldsymbol{v}^+/\tau) + \log \sigma(-\boldsymbol{u}^T\boldsymbol{v}^-/\tau)$	$(\sigma(-\boldsymbol{u}^T\boldsymbol{v}^+/\tau))/\tau\boldsymbol{v}^+ - \sigma(\boldsymbol{u}^T\boldsymbol{v}^-/\tau)/\tau\boldsymbol{v}^-$
Margin Triplet	$-\max(\boldsymbol{u}^T\boldsymbol{v}^- - \boldsymbol{u}^T\boldsymbol{v}^+ + m, 0)$	$\boldsymbol{v}^+ - \boldsymbol{v}^-$ if $\boldsymbol{u}^T\boldsymbol{v}^+ - \boldsymbol{u}^T\boldsymbol{v}^- < m$ else $\mathbf{0}$

Table 2. Negative loss functions and their gradients. All input vectors, i.e. $\boldsymbol{u}, \boldsymbol{v}^+, \boldsymbol{v}^-$, are ℓ_2 normalized. NT-Xent is an abbreviation for "Normalized Temperature-scaled Cross Entropy". Different loss functions impose different weightings of positive and negative examples.

Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations

Connection to Mutual Information

- The optimal critic $h^*_\theta(\{v_1, v_2\})$ should capture the probability that the positive pair comes from the data distribution $p(v_1, v_2)$, while all the other pairs come from the noise distribution $p(v_1)p(v_2)$
- This turns out to be proportional to the ratio between the joint and product of the marginal distributions

$$h^*_\theta(\{v_1, v_2\}) \quad \propto \quad \frac{p(z_1, z_2)}{p(z_1)p(z_2)}$$

Connection to Mutual Information

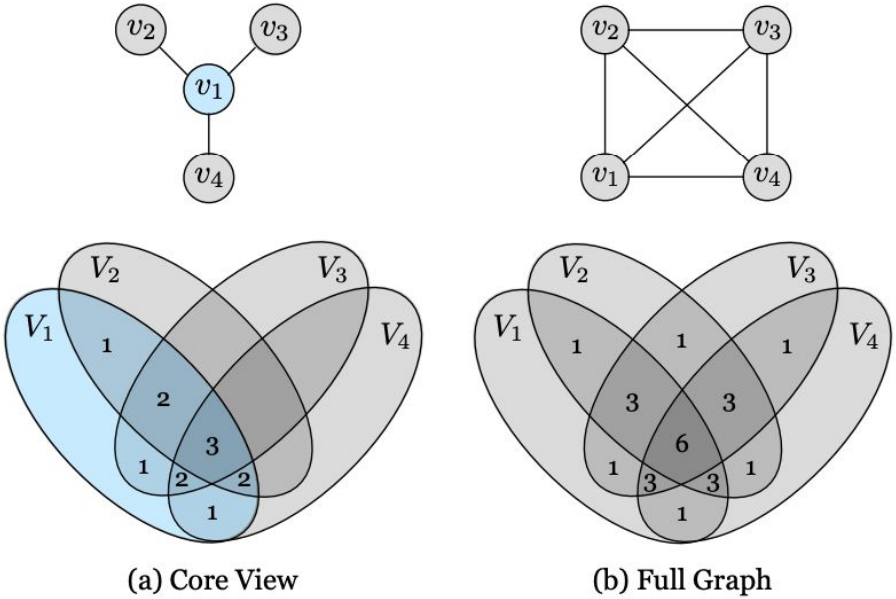
- We see that the loss is thus a quantity related to MI

$$\mathcal{L}_{contrast}^{V_1,V_2} = - \mathop{\mathbb{E}}_{\{v_1^1, v_2^1, \ldots, v_2^{k+1}\}} \left[\log \frac{h_\theta(\{v_1^1, v_2^1\})}{\sum_{j=1}^{k+1} h_\theta(\{v_1^1, v_2^j\})} \right] \quad (2)$$

- We can show that the loss contributes to a lower bound for MI

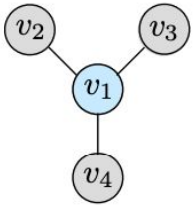
$$I(z_i; z_j) \geq \log(k) - \mathcal{L}_{contrast} \quad (6)$$

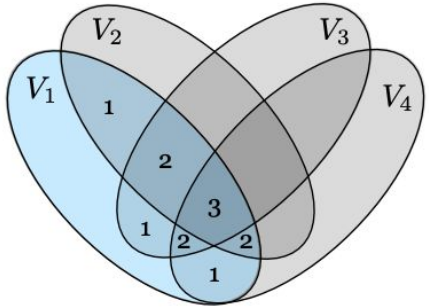
Extension to Multiview



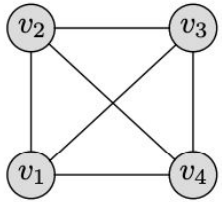
Extension to Multiview

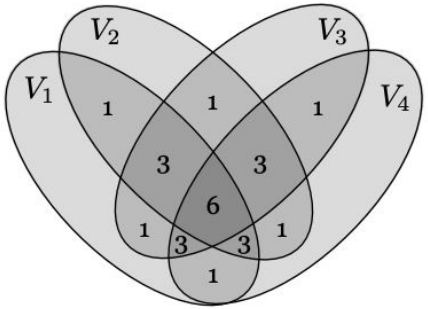
$$\mathcal{L}_C = \sum_{j=2}^{M} \mathcal{L}(V_1, V_j)$$





(a) Core View

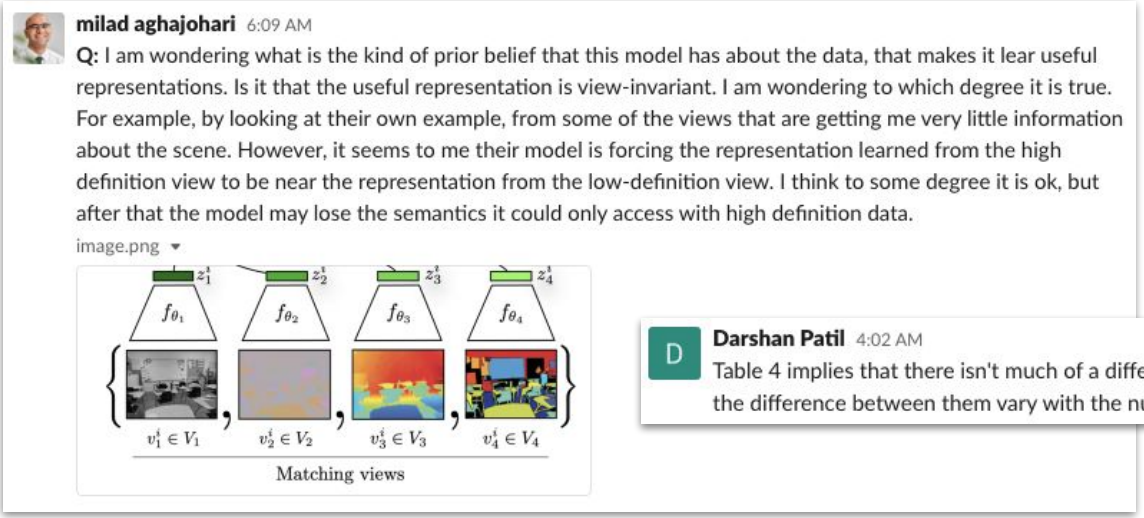




(b) Full Graph

$$\mathcal{L}_F = \sum_{1 \le i < j \le M} \mathcal{L}(V_i, V_j)$$

Extension to Multiview



milad aghajohari 6:09 AM

Q: I am wondering what is the kind of prior belief that this model has about the data, that makes it lear useful representations. Is it that the useful representation is view-invariant. I am wondering to which degree it is true. For example, by looking at their own example, from some of the views that are getting me very little information about the scene. However, it seems to me their model is forcing the representation learned from the high definition view to be near the representation from the low-definition view. I think to some degree it is ok, but after that the model may lose the semantics it could only access with high definition data.

image.png

Darshan Patil 4:02 AM

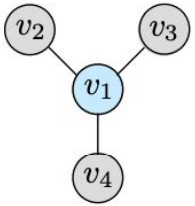
Table 4 implies that there isn't much of a difference between core view and full graph when using 4 views. Does the difference between them vary with the number/type of views?

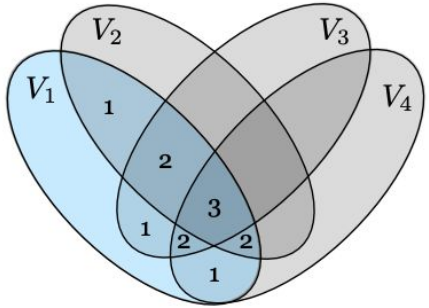
Pierre-André Brousseau 11:06 AM

Apologies for this late question. In the full-graph and core-view setup, you can end up quite with a large amount of contrastive losses all of them equally weighted. Has there been work (e.g Meta Learning) on allowing these losses some flexibility as some pairs of views obviously have better mutual information than others? In this sense, you would only need a full-graph as it could recover the best core-view.

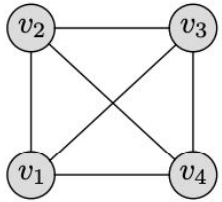
Extension to Multiview

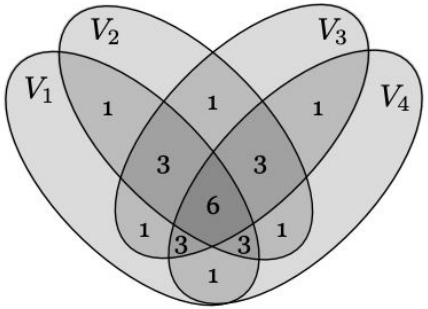
$$\mathcal{L}_C = \sum_{j=2}^{M} \mathcal{L}(V_1, V_j)$$





(a) Core View





(b) Full Graph

$$\mathcal{L}_F = \sum_{1 \le i < j \le M} \mathcal{L}(V_i, V_j)$$

Implementation

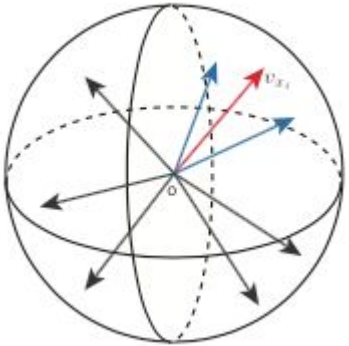
- Having more negative samples lead to better representation, so perhaps we would like to use all available negative examples
- But it also makes the softmax calculation prohibitively expensive

Implementation

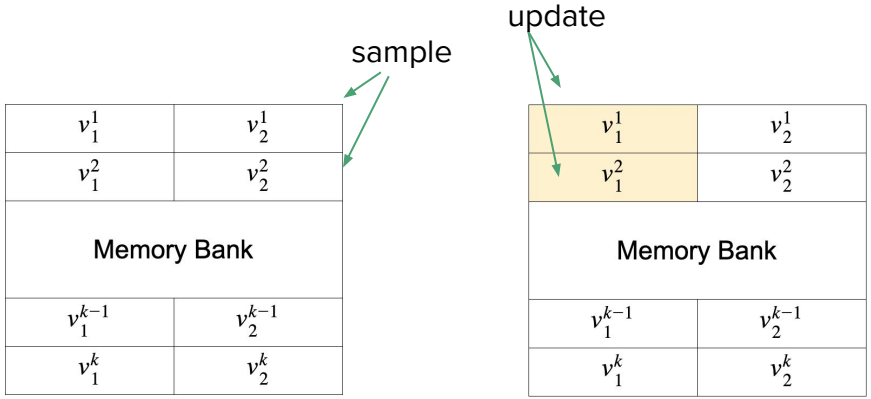
- Having more negative samples lead to better representation, so perhaps we would like to use all available negative examples
- But it also makes the softmax calculation prohibitively expensive
- In the paper, they sample m negative examples and do (m+1)-way softmax classification, and make use of a memory bank

Implementation

- To reduce computation, they make use of a **memory bank**



Randomly initialize embeddings for each example



Store the representations in a memory bank

Sample the negative examples from the memory bank, and update embedding for batch examples

Choosing Good Views

View-Invariant Representations

- We can cast many of the SSL methods as multi-view learning

View-Invariant Representations

- We can cast many of the SSL methods as multi-view learning
- CPC learns from two views: past vs future
- Deep Infomax: global vs local features

View-Invariant Representations

- We can cast many of the SSL methods as multi-view learning
- CPC learns from two views: past vs future
- Deep Infomax: global vs local features
- Exemplar-CNN: augmentations of the same image
- Colorization: grayscale vs colour
- Learning using Videos: object at different timestamps

What makes good views?

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola. What Makes for Good Views for Contrastive Learning?

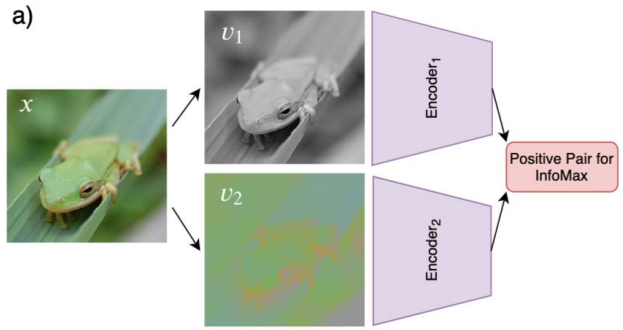
What makes good views?

- Intuitively, we want the views to share in information that is useful for predicting the downstream task
- but not share too much nuisance information, which may hinder generalization and sample complexity

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola. What Makes for Good Views for Contrastive Learning?

What makes good views?

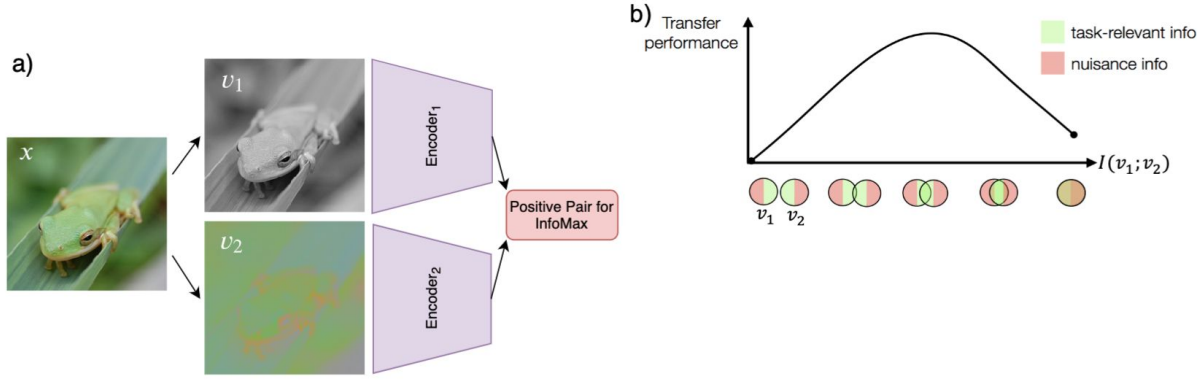
- Intuitively, we want the views to share in information that is useful for predicting the downstream task
- but not share too much nuisance information, which may hinder generalization and sample complexity



Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola. What Makes for Good Views for Contrastive Learning?

What makes good views?

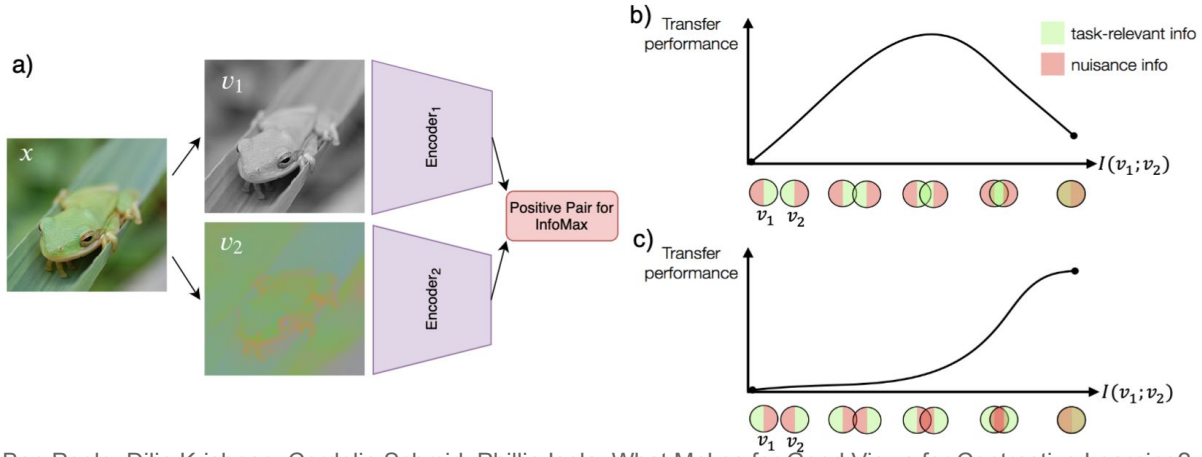
- Intuitively, we want the views to share in information that is useful for predicting the downstream task
- but not share too much nuisance information, which may hinder generalization and sample complexity



Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola. What Makes for Good Views for Contrastive Learning?

What makes good views?

- Intuitively, we want the views to share in information that is useful for predicting the downstream task
- but not share too much nuisance information, which may hinder generalization and sample complexity



Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola. What Makes for Good Views for Contrastive Learning?

What makes good views?

- Suppose we have two views of X, V1 and V2, and a downstream task label Y
- We want what is shared between V1 and V2 to be only the component of X that is useful in predicting Y

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola. What Makes for Good Views for Contrastive Learning?

What makes good views?

- Suppose we have two views of X, V1 and V2, and a downstream task label Y
- We want what is shared between V1 and V2 to be only the component of X that is useful in predicting Y
- This means the optimal view choices are

$$(\mathbf{v_1}^*, \mathbf{v_2}^*) = \min_{\mathbf{v_1},\mathbf{v_2}} I(\mathbf{v_1}; \mathbf{v_2})$$

- Subject to I(v1; y) = I(v2; y) = I(x,y)
- Example: treat images from the same class as congruent pairs and different classes as incongruent pairs

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola. What Makes for Good Views for Contrastive Learning?

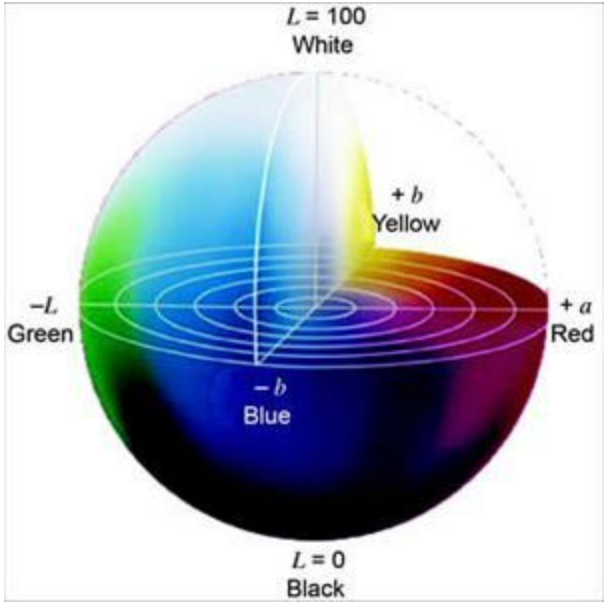
Experiments

Experiments

ImageNet

- Views:
  Luminance: L
  Chrominance: ab

- Negative Sample:
  (L, ab from random image)



Experiments

ImageNet

- Views:
  Luminance: L
  Chrominance: ab

- Negative Sample:
  (L, ab from random image)

Setting	ResNet-50 x0.5	ResNet-50 x1	ResNet-50 x2
$\{L, ab\}$	57.5 / 80.3	64.0 / 85.5	68.3 / 88.2
$\{Y, DbDr\}$	58.4 / 81.2	64.8 / 86.1	69.0 / 88.9
$\{Y, DbDr\}$ + RA	**60.0 / 82.3**	**66.2 / 87.0**	**70.6 / 89.7**

Experiments

ImageNet

- Comparison with semi-supervised methods

Method	ImageNet Classification Accuracy				
	conv1	conv2	conv3	conv4	conv5
ImageNet-Labels	19.3	36.3	44.2	48.3	50.5
Random	11.6	17.1	16.9	16.3	14.1
Data-Init [42]	17.5	23.0	24.5	23.2	20.6
Context [17]	16.2	23.3	30.2	31.7	29.6
Colorization [84]	13.1	24.8	31.0	32.6	31.8
Jigsaw [54]	19.2	30.1	34.7	33.9	28.3
BiGAN [18]	17.7	24.5	31.0	29.9	28.0
SplitBrain† [85]	17.7	29.3	35.4	35.2	32.8
Counting [55]	18.0	30.6	34.3	32.5	25.7
Inst-Dis [79]	16.8	26.5	31.8	34.1	35.6
RotNet [22]	18.8	31.7	38.7	38.2	36.5
DeepCluster [11]	12.9	29.2	38.2	39.8	36.1
AET [83]	**19.3**	32.8	**40.6**	39.7	37.7
CMC($\{Y, DbDr\}$)	18.3	**33.7**	38.3	**40.5**	**42.8**

Experiments

ImageNet

- Relation with number of negative samples

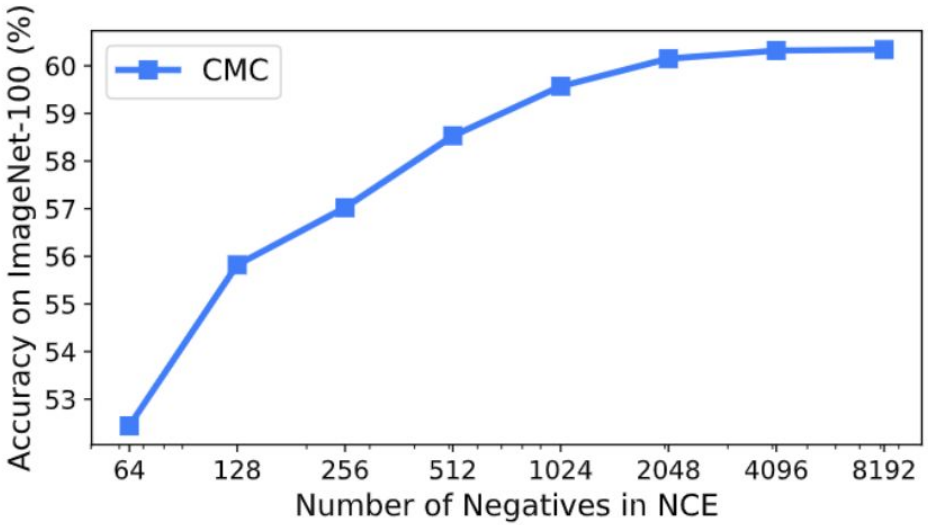
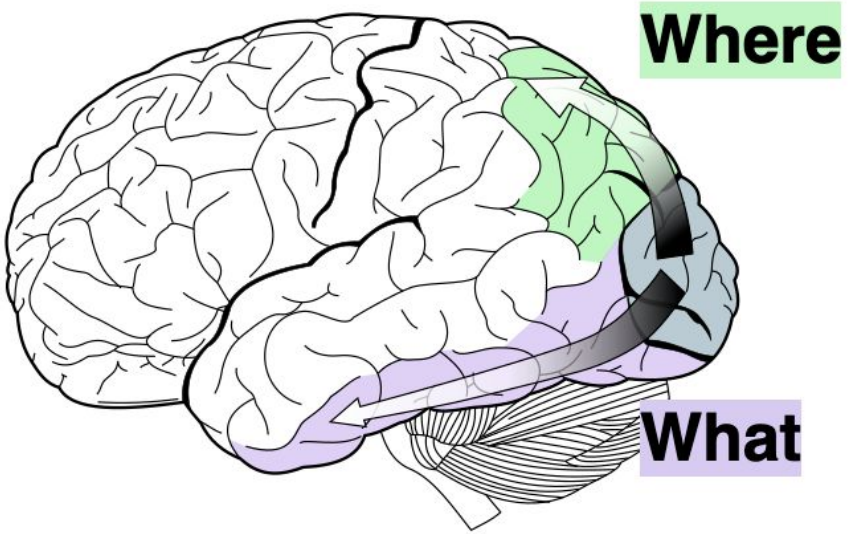


Figure 6: We plot the number of negative examples m in NCE-based contrastive loss against the accuracy for 100 randomly chosen classes of Imagenet 100. It is seen that the accuracy steadily increases with m.

Experiments

Videos

- Compared with human visual cortex:
  Ventral Stream: Object Recognition
  Dorsal Stream: Motion



The dorsal stream (green) and ventral stream (purple) are shown. They originate from a common source in the visual cortex

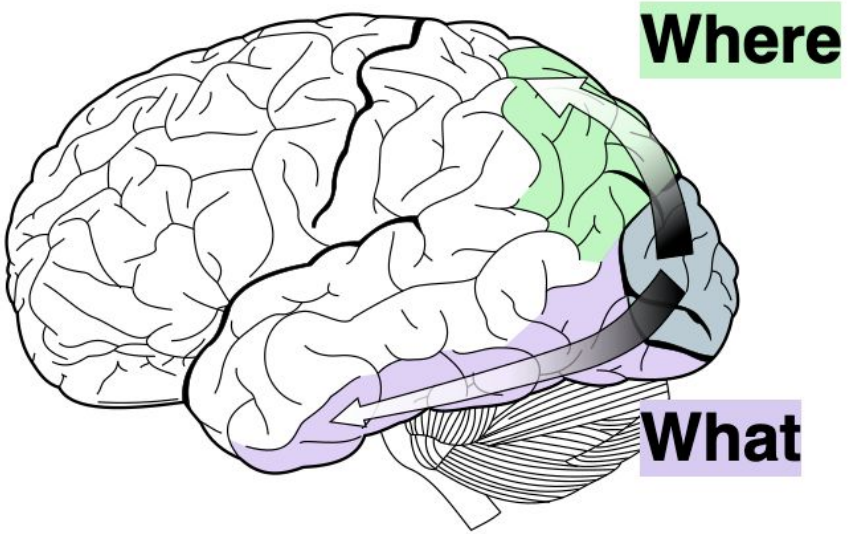
Experiments

Videos

- Compared with human visual cortex:
  Ventral Stream: Object Recognition
  Dorsal Stream: Motion

- Views:
  Image at time t: i_t
  Ventral Stream: Neighbouring Frame (i_t, i_{t+k})
  Dorsal Stream: Optical Flow (i_t, f_t)

- Negative Samples:
  A random frame from a random video



The dorsal stream (green) and ventral stream (purple) are shown. They originate from a common source in the visual cortex

Experiments

Videos

Method	# of Views	UCF-101	HMDB-51
Random	-	48.2	19.5
ImageNet	-	67.7	28.0
VGAN* [77]	2	52.1	-
LT-Motion* [46]	2	53.0	-
TempCoh [52]	1	45.4	15.9
Shuffle and Learn [50]	1	50.2	18.1
Geometry [21]	2	55.1	23.3
OPN [44]	1	56.3	22.1
ST Order [10]	1	58.6	25.0
Cross and Learn [66]	2	58.7	**27.2**
CMC (V)	2	55.3	-
CMC (D)	2	57.1	-
CMC (V+D)	3	**59.1**	26.7

Most methods either use single RGB view or additional optical flow view, while VGAN explores sound as the second view.

Experiments

Videos

Increasing the number of views of the data from 2 to 3 (using both streams instead of one) provides a boost for UCF-101.

Extending CMC to multiple views

Method	# of Views	UCF-101	HMDB-51
Random	-	48.2	19.5
ImageNet	-	67.7	28.0
VGAN* [77]	2	52.1	-
LT-Motion* [46]	2	53.0	-
TempCoh [52]	1	45.4	15.9
Shuffle and Learn [50]	1	50.2	18.1
Geometry [21]	2	55.1	23.3
OPN [44]	1	56.3	22.1
ST Order [10]	1	58.6	25.0
Cross and Learn [66]	2	58.7	**27.2**
CMC (V)	2	55.3	-
CMC (D)	2	57.1	-
CMC (V+D)	3	**59.1**	26.7

Most methods either use single RGB view or additional optical flow view, while VGAN explores sound as the second view.

Experiments

Videos

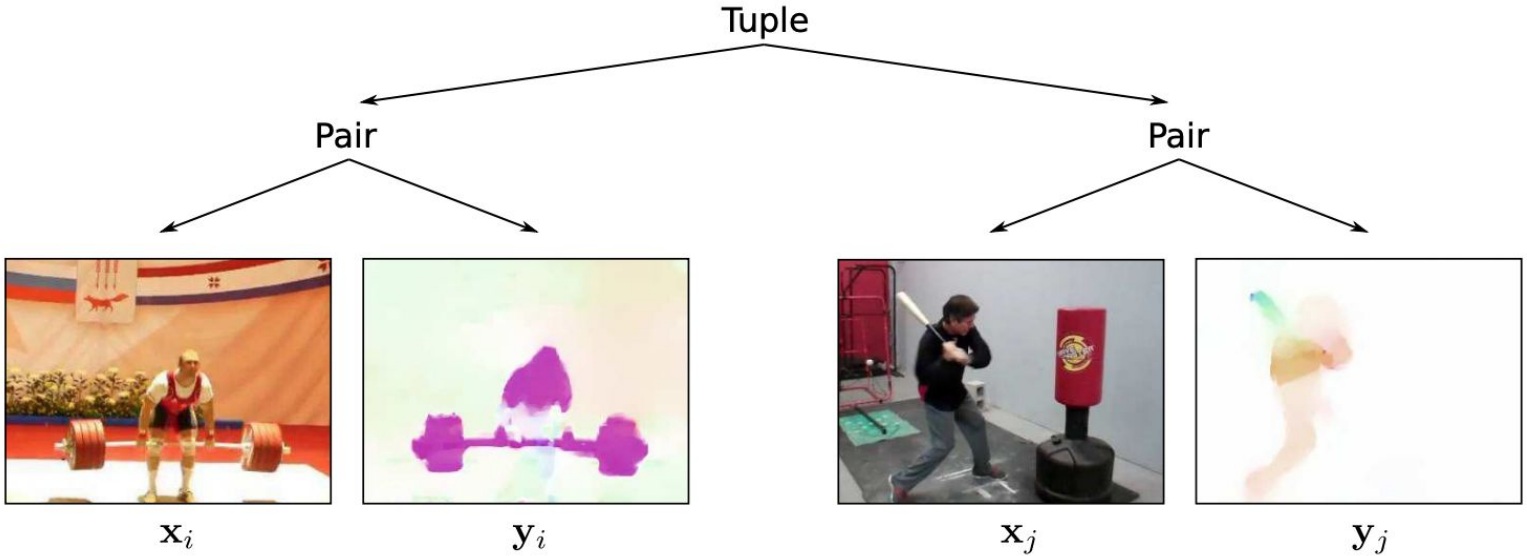
Cross and Learn

Method	# of Views	UCF-101	HMDB-51
Random	-	48.2	19.5
ImageNet	-	67.7	28.0
VGAN* [77]	2	52.1	-
LT-Motion* [46]	2	53.0	-
TempCoh [52]	1	45.4	15.9
Shuffle and Learn [50]	1	50.2	18.1
Geometry [21]	2	55.1	23.3
OPN [44]	1	56.3	22.1
ST Order [10]	1	58.6	25.0
Cross and Learn [66]	2	58.7	**27.2**
CMC (V)	2	55.3	-
CMC (D)	2	57.1	-
CMC (V+D)	3	**59.1**	26.7

Most methods either use single RGB view or additional optical flow view, while VGAN explores sound as the second view.

Cross and Learn

Modalities: RGB and Optical Flow



Cross and Learn

Cross Modal

Vertical Distance decreases

Diversity

Horizontal Distance increases

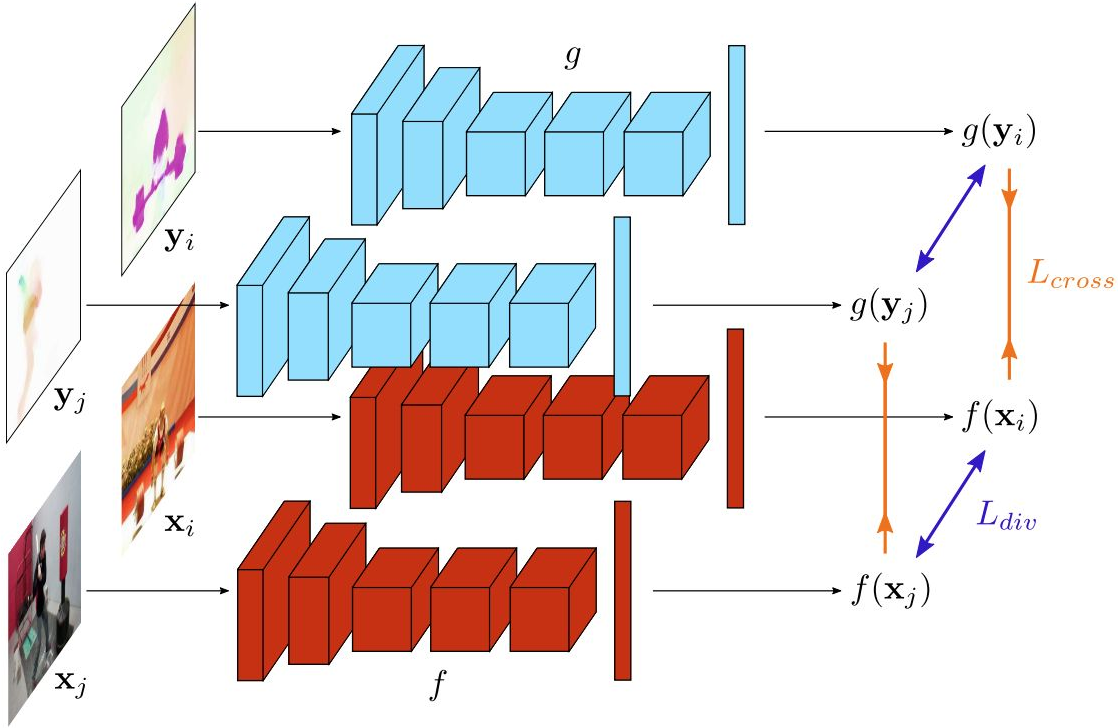


Fig. 2. Visualization of our model. Networks of same color share weights and are denoted by f (red, dark) and g (blue, bright) respectively. f and g are trained so that the cross-modal distance (vertical) decreases while the distance between different pairs (horizontal) increases

Cross and Learn

$$L_{cross}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{x}_j, \mathbf{y}_j) = \frac{1}{2}\left[\mathrm{d}(f(\mathbf{x}_i), g(\mathbf{y}_i)) + \mathrm{d}(f(\mathbf{x}_j), g(\mathbf{y}_j))\right]$$

$$L_{div}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{x}_j, \mathbf{y}_j) = -\frac{1}{2}\left[\mathrm{d}(f(\mathbf{x}_i), f(\mathbf{x}_j)) + \mathrm{d}(g(\mathbf{y}_i), g(\mathbf{y}_j))\right]$$

$$L(\mathbf{x}_i, \mathbf{y}_i, \mathbf{x}_j, \mathbf{y}_j) = L_{cross}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{x}_j, \mathbf{y}_j) + L_{div}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{x}_j, \mathbf{y}_j).$$

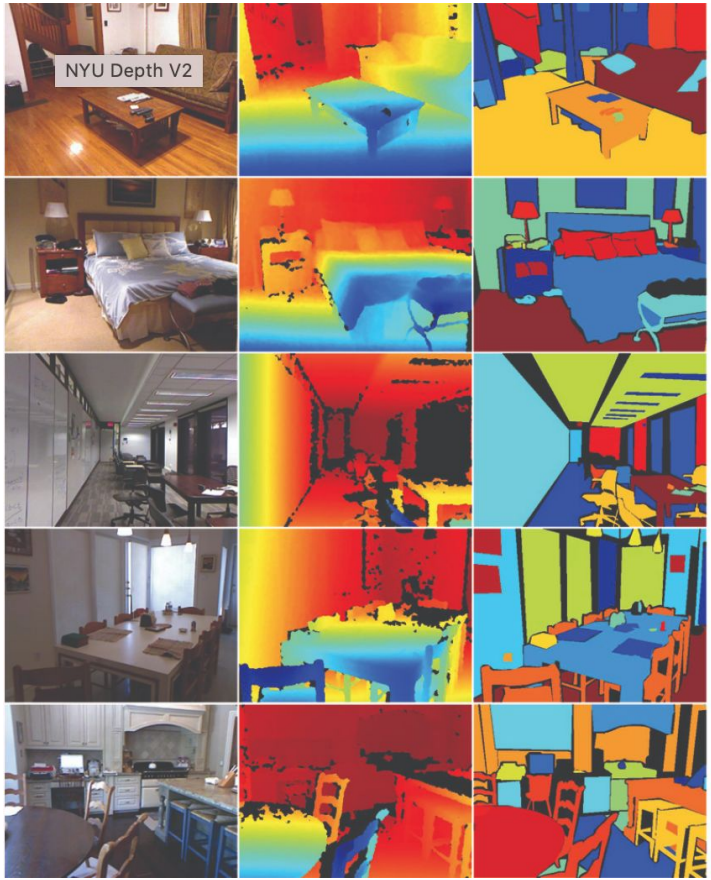
Experiments

NYU Depth V2

1449 densily labeled pairs

Views:

Luminance (L)
Chrominance (ab)
Depth
Surface Normal



Samples of the RGB Image, the raw depth image, and the class labels

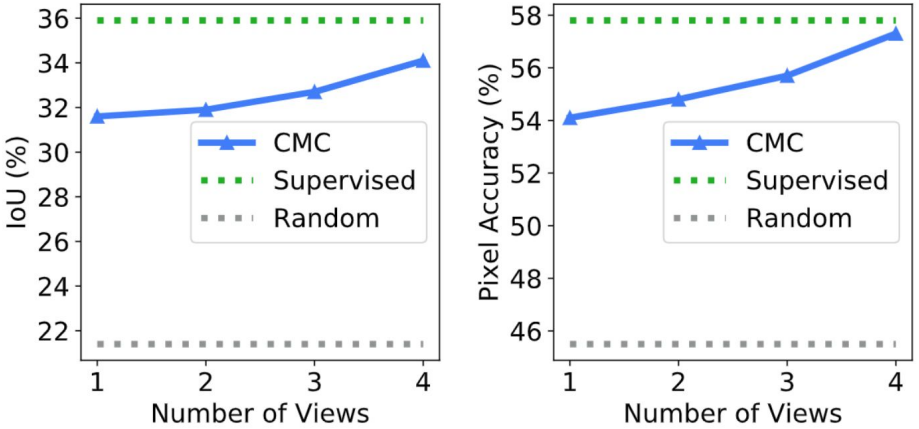
Experiments

NYU Depth V2

1449 densily labeled pairs

Views:

Luminance (L)
Chrominance (ab)
Depth
Surface Normal



Intersection over Union (IoU) (left) andPixel Accuracy (right) for the NYU Depth-V2 dataset, as CMC is trained with increasingly more views from 1 to 4. The views are (inorder of inclusion): L, ab, depth and surface normals.

	Pixel Accuracy (%)	mIoU (%)
Random	45.5	21.4
CMC (core-view)	57.1	34.1
CMC (full-graph)	57.0	34.4
Supervised	**57.8**	**35.9**

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.

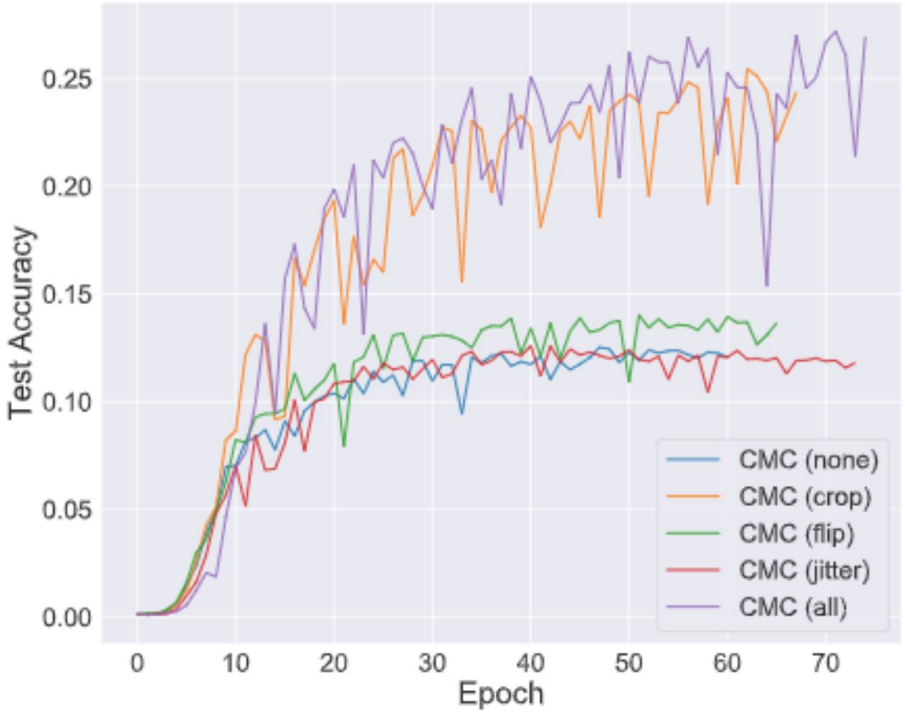
On Mutual Information in Contrastive Learning for Visual Representations

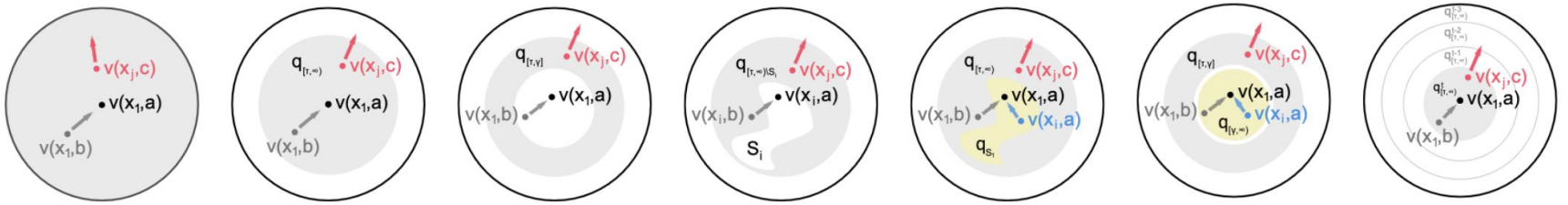
Mike Wu[■], **Chengxu Zhuang**[★], **Milan Mossé**[■♦], **Daniel Yamins**[■★], **Noah Goodman**[■★]
Department of Computer Science (■), Philosophy (♦), and Psychology (★)
Stanford University
{wumike, chengxuz, mmosse19, yamins, ngoodman}@stanford.edu

- Augmentation is the crux of learning a good representation. Without it, the contrastive learning would not enjoy the success it has found in practice.

- Derives a new lower bound on mutual information that supports sampling negative examples from a restricted distribution

Augmentation is the crux of learning a good representation.





(a) IR (b) BALL (c) RING (d) CAVE (e) LA (f) RING$^{\text{s-neigh}}$ (g) BALL$^{\text{anneal}}$

Derives a new lower bound on mutual information that supports sampling negative examples from a restricted distribution

Model	Top1	Top5	Model	Top1	Top5
IR	43.2	67.0	CMC	48.2	72.4
BALL	45.3	68.6	BALL$^{\text{Lab}}$	48.9	73.1
CAVE	46.6	70.4	CAVE$^{\text{Lab}}$	49.2	73.3
LA	48.0	72.4	–	–	–
BALL$_{\text{anneal}}$	47.3	71.1	BALL$^{\text{Lab+anneal}}$	49.7	73.6
CAVE$_{\text{anneal}}$	**48.4**	**72.5**	CAVE$^{\text{Lab+anneal}}$	**50.5**	**74.0**

(a) ImageNet: Classification Accuracy

Experiments

Mutual Information between Views

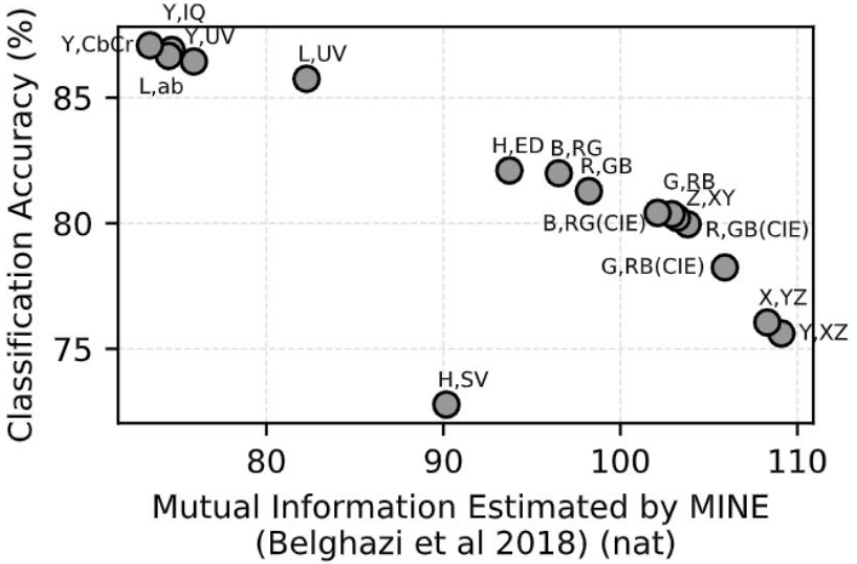
Good collection of views is one that shares some information but not too much.

Experiments

Mutual Information between Views

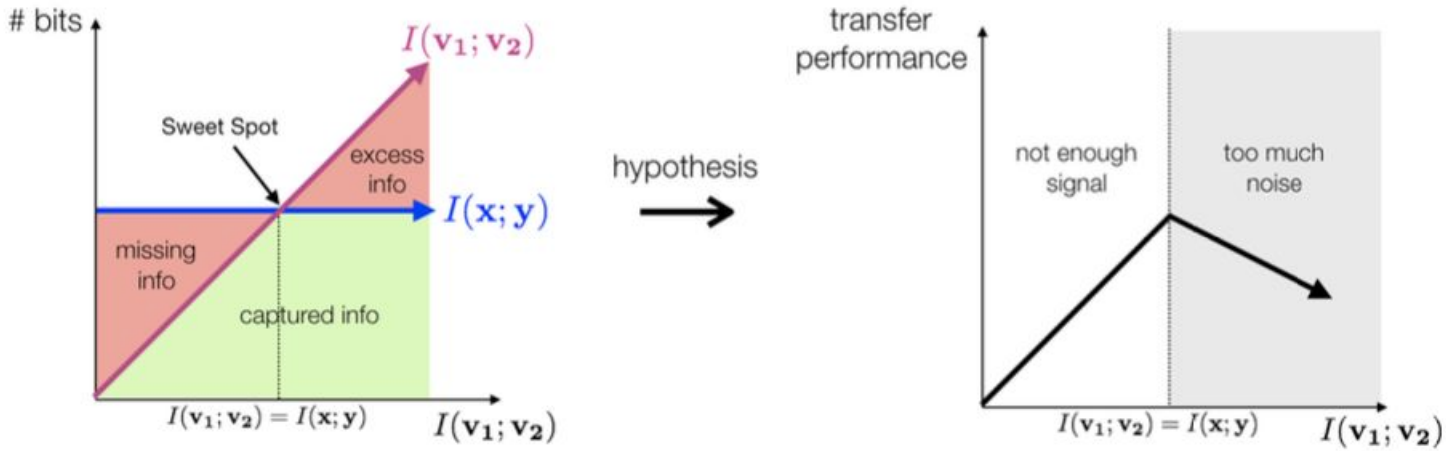
Good collection of views is one that shares some information but not too much.

- Splitting each color-space into two views like (L, ab), (R, gb)
- Uses the MINE estimator to estimate the mutual information between the views
- Trained a linear classifier on the learned representations on the STL-10 dataset



Classification accuracy against estimated Mutual Information between channels of different color spaces

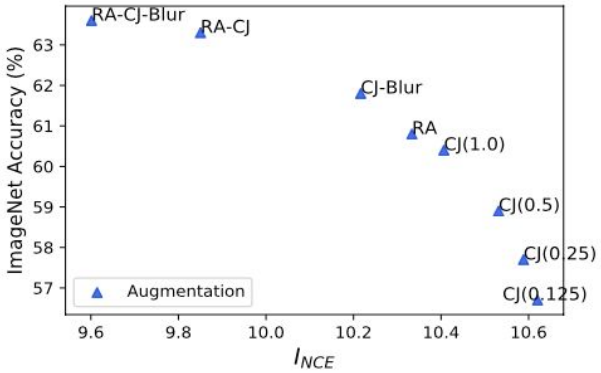
What makes good views?



Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola. What Makes for Good Views for Contrastive Learning?

Relationship to Data Augmentation

- As we increase the intensity of data augmentation and decrease mutual information, we observe a steady increase in downstream accuracy



(a) I_{NCE} $v.s$ Accuracy

PyTorch-style data augmentation

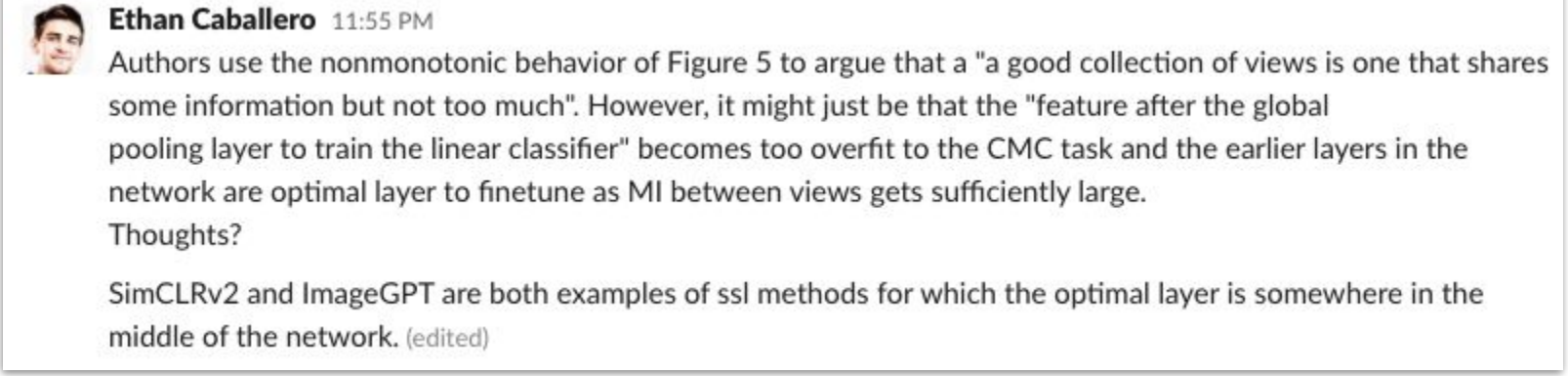
```
RandomResizedCrop(scale=(0.2, 1.0))
RandomHorizontalFlip()
# CJ(x): random color jitter with x
cj = ColorJitter([0.8,0.8,0.8,0.4]*x)
RandomApply([cj], p=0.8)
# Blur: random blurring
blur = Blur(sigma=(0.1,2.0))
RandomApply([blur], p=0.5)
# RA: RandAugment
rnd_augment()
RandomGrayscale(p=0.2),
```

(b) Data Augmentation

Figure 10: (a) data augmentation as InfoMin on ImageNet with linear projection head; (b) illustration of step-by-step data augmentation used in InfoMin.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola. What Makes for Good Views for Contrastive Learning?

What makes good views?



Ethan Caballero 11:55 PM

Authors use the nonmonotonic behavior of Figure 5 to argue that a "a good collection of views is one that shares some information but not too much". However, it might just be that the "feature after the global pooling layer to train the linear classifier" becomes too overfit to the CMC task and the earlier layers in the network are optimal layer to finetune as MI between views gets sufficiently large.
Thoughts?

SimCLRv2 and ImageGPT are both examples of ssl methods for which the optimal layer is somewhere in the middle of the network. (edited)

Learning Effective Views

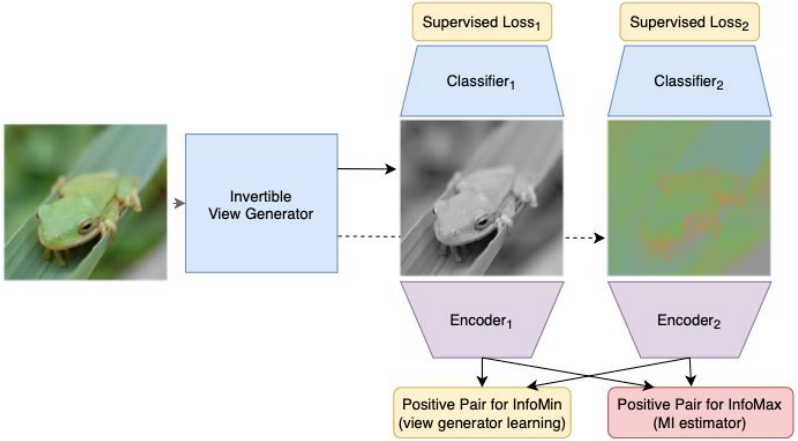
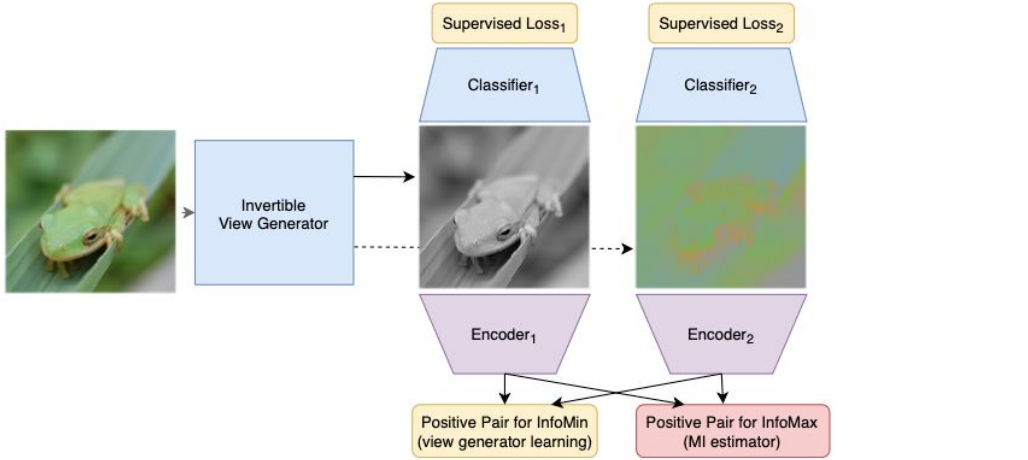


Figure 7: Schematic of contrastive representation learning with a learned view generator. An input image is split into two views using an invertible view generator. To learn the view generator, we optimize the losses in yellow: minimizing information between views while ensuring we can classify the object from each view. The encoders used to estimate mutual information are always trained to maximize the InfoNCE lower bound. After learning the view generator, we reset the weights of the encoders, and train with a fixed view generator without the additional supervised classification losses.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola. What Makes for Good Views for Contrastive Learning?

Learning Effective Views



$$\min_{g,c_1,c_2} \max_{f_1,f_2} \underbrace{I_{NCE}^{f_1,f_2}(g(X)_1; g(X)_{2:3})}_{\text{unsupervised: reduce } I(v_1;v_2)} + \underbrace{\mathcal{L}_{ce}(c_1(g(X)_1), y) + \mathcal{L}_{ce}(c_2(g(X)_{2:3}), y)}_{\text{supervised: keep } I(v_1;y) \text{ and } I(v_2;y)}$$

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola. What Makes for Good Views for Contrastive Learning?

Learning Effective Views

Table 2: Comparison of different view generators by measuring STL-10 classification accuracy: *supervised*, *unsupervised*, and *semi-supervised*

Method	RGB	YDbDr
unsupervised	82.4 ± 3.2	84.3 ± 0.5
supervised	79.9 ± 1.5	78.5 ± 2.3
semi-supervised	$\mathbf{86.0 \pm 0.6}$	$\mathbf{87.0 \pm 0.3}$
raw input	81.5 ± 0.2	86.6 ± 0.2

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola. What Makes for Good Views for Contrastive Learning?