What Makes for Good Views for Contrastive Learning?

By: Yonglong Tian etc.
On: NeurIPS 2020

Presented by: Xiaobing Chen
Feb. 16, 2022

Outlines

1. **Contrastive Learning & Mutual Information**
   - Contrastive Learning Framework
   - Relationship between contrastive loss and mutual information

2. InfoMin Principle

- Impact of positive pairs to downstream task
- InfoMin priciple

3. Learning Views for Contrastive Learning

- Unsupervised view learning
- Semi-supervised view learning

4. Conclusions and Discussion

1 Contrastive Learning & Mutual Information

Contrastive Learning

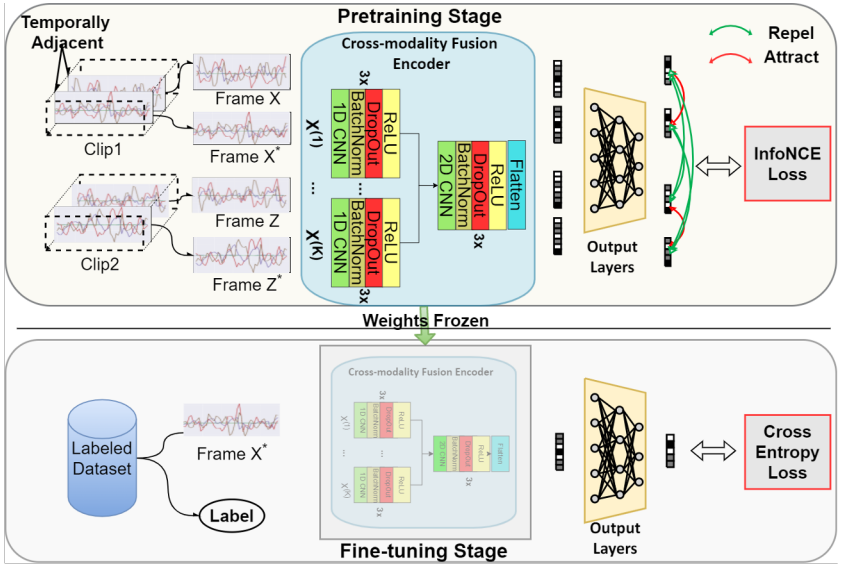


Fig. 1. An example of contrastive learning framework in human activity recognition.

- **Contrastive Learning (CL)**: learn representations by a "contrastive" loss which pushes apart dissimilar data pairs while pulling together similar pairs [1].
- **Relationship with unsupervised learning**: CL is a **subset** of unsupervised learning.
- **Two stages**: 1. pretrain encoders in **pretext tasks** 2. apply frozen encoder to downstream tasks
- **Positive pairs selection**: create multiple views of each datapoint, such as randomly augmenting an image twice

InfoNCE & Mutual Information (MI)

- **Commonly used Objective function** – InfoNCE [2]:

Suppose $p(\mathbf{v_1})$ and $p(\mathbf{v_2})$ are the data distribution of two views and $\mathbf{v}_{1,\mathrm{i}} \sim p(\mathbf{v_1})$ is the target sample, then its positive sample is $\mathbf{v}_{2,\mathrm{i}} \sim p(\mathbf{v_2} \mid \mathbf{v}_{1,\mathrm{i}})$ and negative sample is $\mathbf{v}_{2,\mathrm{i}} \sim p(\mathbf{v_2})$. InfoNCE loss is defined by:

$$\mathcal{L}_{\mathrm{NCE}} = -\mathbb{E}\left[\log \frac{e^{h(\mathbf{v}_{1,\mathrm{i}}, \mathbf{v}_{2,\mathrm{i}})}}{\sum_{j=1}^{K} e^{h(\mathbf{v}_{1,\mathrm{i}}, \mathbf{v}_{2,\mathrm{j}})}}\right]$$

- **InfoNCE & Mutual Information (MI)**: Minimizing InfoNCE loss equivalently maximizes a lower bound of mutual information of two views, i.e.,

$$I(\mathbf{v_1}; \mathbf{v_2}) \geq \log(K) - \mathcal{L}_{\mathrm{NCE}} = I_{\mathrm{NCE}}(\mathbf{v_1}; \mathbf{v_2}).$$

- **Sufficient Encoder Assumption**:

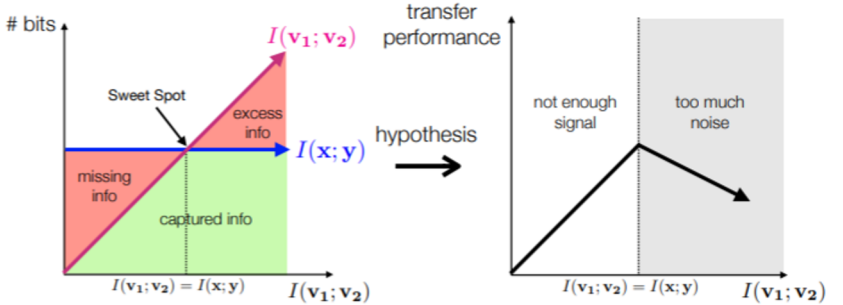
Given sufficient encoder $f(\cdot)$, representations can be denoted as $\mathbf{z_1} = f(\mathbf{v_1})$ and $\mathbf{z_2} = f(\mathbf{v_2})$. Then, the mutual information of $\mathbf{z_1}$ and $\mathbf{z_2}$ satisfies

$$I(\mathbf{z_1}; \mathbf{z_2}) = I(\mathbf{v_1}; \mathbf{v_2})$$

Optimal positive pairs will maximize MI.

2 InfoMin Principle

Three Regimes of Information Captured



- **MI and downstream task**: the amount and type of information shared between v1 and v2 (i.e., I(v1; v2)) determines the performance on downstream tasks.
- Consider the MI I(x; y) in downstream task, there are three regimes of performance regarding to the information captured in I(v1; v2).

1. *Missing information*: When $I(\mathbf{v_1};\mathbf{v_2}) < I(\mathbf{x};\mathbf{y})$, there is information about the task-relevant variable that is discarded by the view, degrading performance.
2. *Sweet spot*: When $I(\mathbf{v_1};\mathbf{y}) = I(\mathbf{v_2};\mathbf{y}) = I(\mathbf{v_1};\mathbf{v_2}) = I(\mathbf{x};\mathbf{y})$, the only information shared between $\mathbf{v_1}$ and $\mathbf{v_2}$ is task-relevant, and there is no irrelevant noise.
3. *Excess noise*: As we increase the amount of information shared in the views beyond $I(\mathbf{x};\mathbf{y})$, we begin to include additional information that is irrelevant for the downstream task. This can lead to worse generalization on the downstream task

MI and Performance

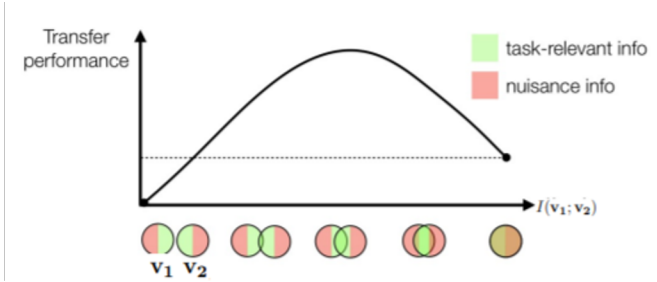


Fig. 2. Reverse-U shaped curve of MI and transfer performance.

- Appropriate task-relevant info shared in MI gives the best downstream performance.

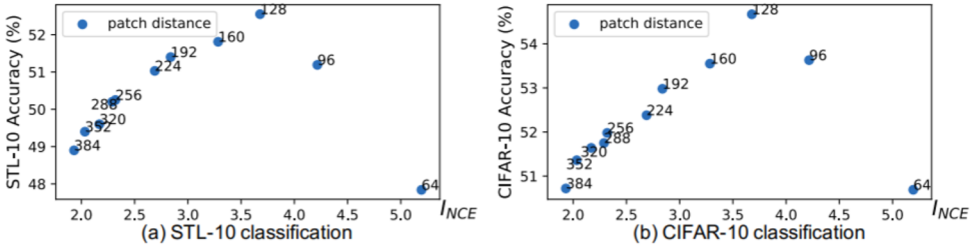


Fig. 3. Empirical demonstration of the relationship between MI and transfer performance.

- Reducing I(v1; v2) with spatial distance:
  - Randomly crop two patches from the same image
  - Change the distance between the centers of the two patches
  - Pretrain the encoder in the contrastive learning
  - Freeze the encoder and train a linear classifier in the downstream task

InfoMin Principle

- **InfoMin Principle for Contrastive Learning**: optimal positive pairs are supposed to share minimal information **NECESSARY** to the downstream task.

Proposition 3.1. *Suppose f_1 and f_2 are minimal sufficient encoders. Given a downstream task $\mathcal{T}$ with label $\mathbf{y}$, the optimal views created from the data $\mathbf{x}$ are $(\mathbf{v_1}^*, \mathbf{v_2}^*) = \arg\min_{\mathbf{v_1},\mathbf{v_2}} I(\mathbf{v_1};\mathbf{v_2})$, subject to $I(\mathbf{v_1};\mathbf{y}) = I(\mathbf{v_2};\mathbf{y}) = I(\mathbf{x};\mathbf{y})$. Given $\mathbf{v_1}^*, \mathbf{v_2}^*$, the representation $\mathbf{z_1^*}$ (or $\mathbf{z_2^*}$) learned by contrastive learning is optimal for $\mathcal{T}$ (Def 3), thanks to the minimality and sufficiency of f_1 and f_2.*

Impact of Shared Information

- Experimental dataset: **colorful moving-MNIST,** where each frame has three factors of variation: the class of the digit, the position of the digit, and the class of background image (randomly chosen from 10 classes)

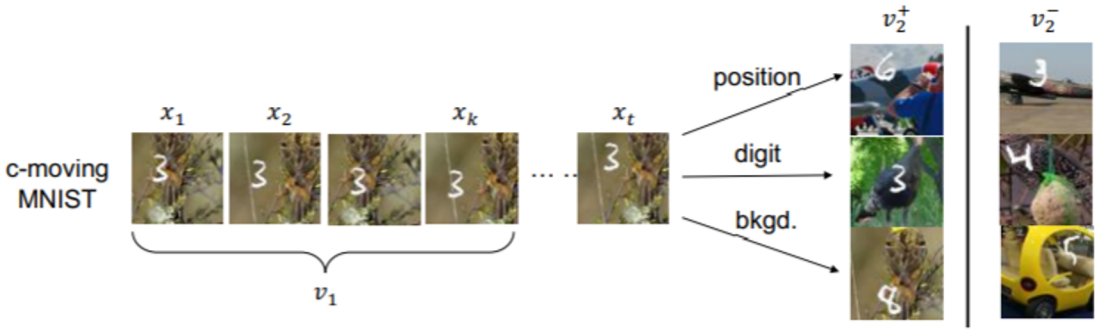


Figure 4: Illustration of the Colorful-Moving-MNIST dataset. In this example, the first view $\mathbf{v_1}$ is a sequence of frames containing the moving digit, e.g., $\mathbf{v_1} = x_{1:k}$. The matched second view $\mathbf{v_2^+}$ share some factor with x_t that $\mathbf{v_1}$ can predict, while the unmatched view $\mathbf{v_2^-}$ does not share factor with x_t.

- Positive pairs are constructed in **single factor shared** and **multi-factor shared** scenario.

Impact of Shared Information

	$I(\mathbf{v_1}; \mathbf{v_2})$	digit cls. error rate (%)	background cls. error rate (%)	digit loc. error pixels
Single Factor	*digit*	**16.8**	88.6	13.6
	bkgd	88.6	**51.7**	16.1
	pos	57.9	87.6	**3.95**
Multiple Factors	*bkgd, digit, pos*	88.8	56.3	16.2
	bkgd, digit	88.2	53.9	16.3
	bkgd, pos	88.8	53.8	15.9
	digit, pos	**14.5**	88.9	13.7
Supervised		3.4	45.3	0.93

Table 1. Performance on three different tasks regarding various positive pairs in the pretraining.

- The performance is significantly affected by what is shared between v1 and v2. If the downstream task is relevant to one factor, I(v1; v2) should include that factor rather than others.
- One factor can overwhelm another; for instance, whenever background is shared, the latent representation leaves out information for discriminating or localizing digits.

3

Learning Views for Contrastive Learning

Unsupervised View Learning

- **Goal**: to train a generator g(.) which produces optimal views based on InfoMin principle.
- **Objective function**: Given an input image X, the splitting over channels is represented as {X1, X2:3}. The generator transform the positive pair into other views and encoder f1 and f2 to encode each view. The objective function is given by

$$\min_{g} \max_{f_1, f_2} I_{\text{NCE}}^{f_1, f_2}\left(g(X)_1; g(X)_{2:3}\right)$$

- Encoder f1 and f2 aim to maximize the MI while the generator try to minimize it.
- After the joint training with unlabeled data, two encoder and a generator are transferred to the downstream task.

Semi-supervised View Learning

- **Goal**: to introduce the label information to the unsupervised view learning since optimal views are task-dependent.
- **Objective function**: Given some labeled data, the objective function can be revised by introducing the cross-entropy loss,

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

where c1 and c2 are two classifiers.

- Semi-supervised view learning: labeled batch and unlabeled batch are iteratively used. The second term, cross-entropy loss, is only included when labeled data are used.

Experiments

- STL-10 dataset: include 10 classes; 96x96 pixels color images; 5k labeled; 100k unlabeled.
- Experimental settings: each image with two color spaces, RGB and YDbDr

Method (# of Images)	RGB	YDbDr
unsupervised (100k)	82.4 ± 3.2	84.3 ± 0.5
supervised (5k)	79.9 ± 1.5	78.5 ± 2.3
semi-supervised (105k)	$\mathbf{86.0 \pm 0.6}$	$\mathbf{87.0 \pm 0.3}$
raw views	81.5 ± 0.2	86.6 ± 0.2

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

- The semi-supervised view generator significantly outperforms the supervised one, validating the importance of reducing I(v1; v2).

4 Conclusions & Discussions

Conclusions & Discussions

- The mutual information of two views and the performance on downstream task follows a **reverse-U** shape relationship.
- Goodness of positive pairs is **task-relevant**.
- **InfoMin principle**: except task-relevant information, positive pairs should share as less info as possible.

Reference

[1]. Tian, Yonglong, et al. "What makes for good views for contrastive learning?." Advances in Neural Information Processing Systems 33 (2020): 6827-6839.

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

Thanks