南京航空航天大學

Nanjing University of Aeronautics and Astronautics

ParNeC 模式识别与神经计算研究组

PAttern Recognition and NEural Computing

Contrastive Learning with Stronger Augmentations

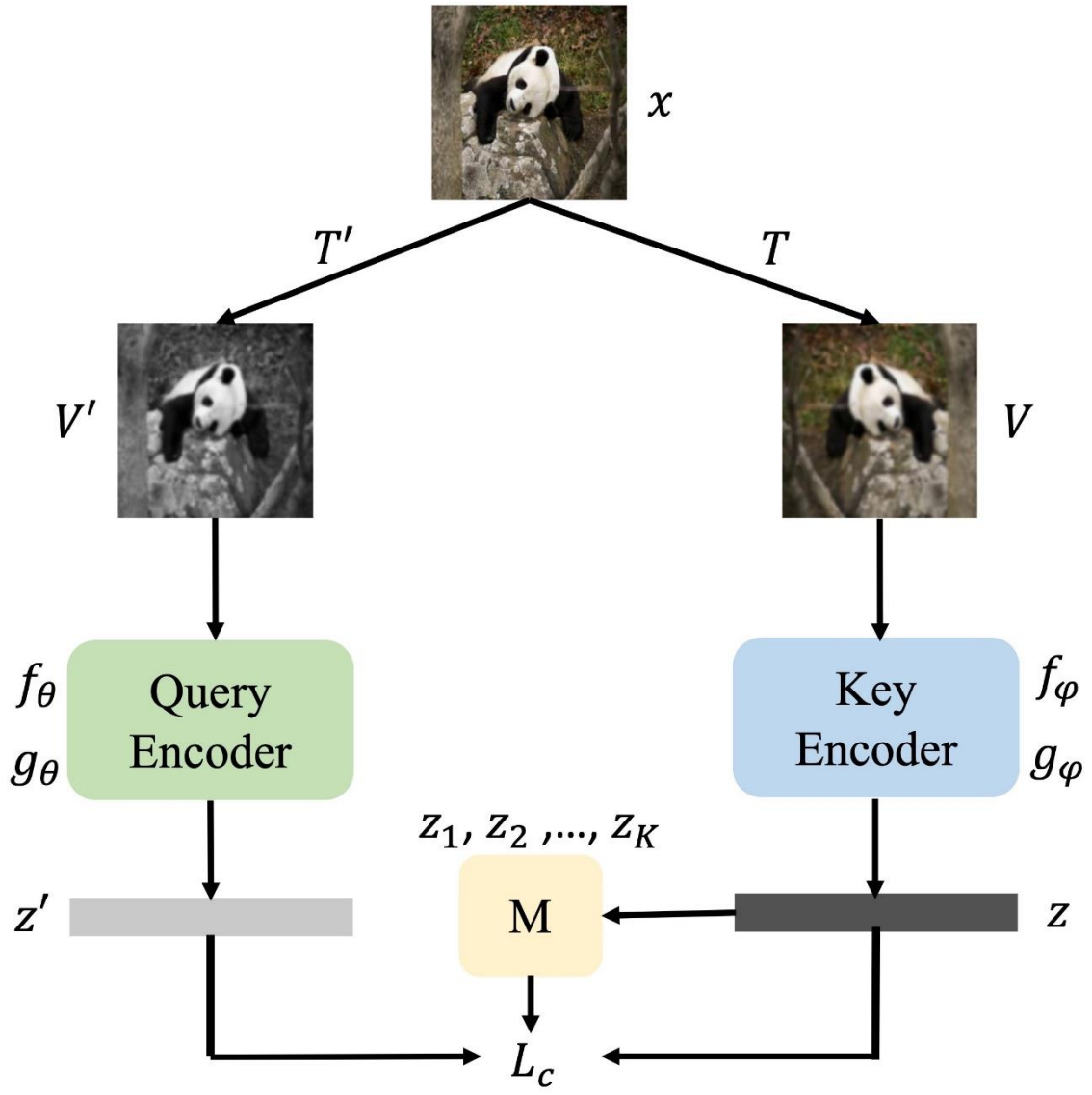
Xiao Wang and Guo-Jun Qi, *Senior Member, IEEE*

arXiv 2021

Contrastive Learning

Contrastive Instance Learning Framework

- Contrastive Loss
  - $\mathcal{L}_C = \mathbb{E}_{i \in B}[-\log \frac{\mathcal{Q}(i, i+)}{\mathcal{Q}(i, i+) + \sum_{k=1}^{K} \mathcal{Q}(i, k)}]$
- Similarities of Positive Pairs
  - $\mathcal{Q}(i, i+) = \exp\left(sim(z_i', z_i)/\tau\right)$
- Similarities of Negative Pairs
  - $\mathcal{Q}(i, k) = \exp\left(sim(z_i', z_k)/\tau\right)$
- Cosine similarity
  - $sim(z_i', z_k) = \frac{{z_i'}^T z_k}{\|z_i'\| \cdot \|z_k\|}$



Motivation

Lack of enough studies on the potential of positive pairs

- For positive pairs, the data augmentation is carefully designed.

(a) Original

(b) Crop and resize

(c) Crop and resize (and flip)

(d) Color distort. (drop)

(d) Color distort. (jitter)

(f) Rotate $\{90^{\circ}, 190^{\circ}, 270^{\circ}\}$

(g) Cutout

(h) Gaussian noise

(i) Gaussian blur

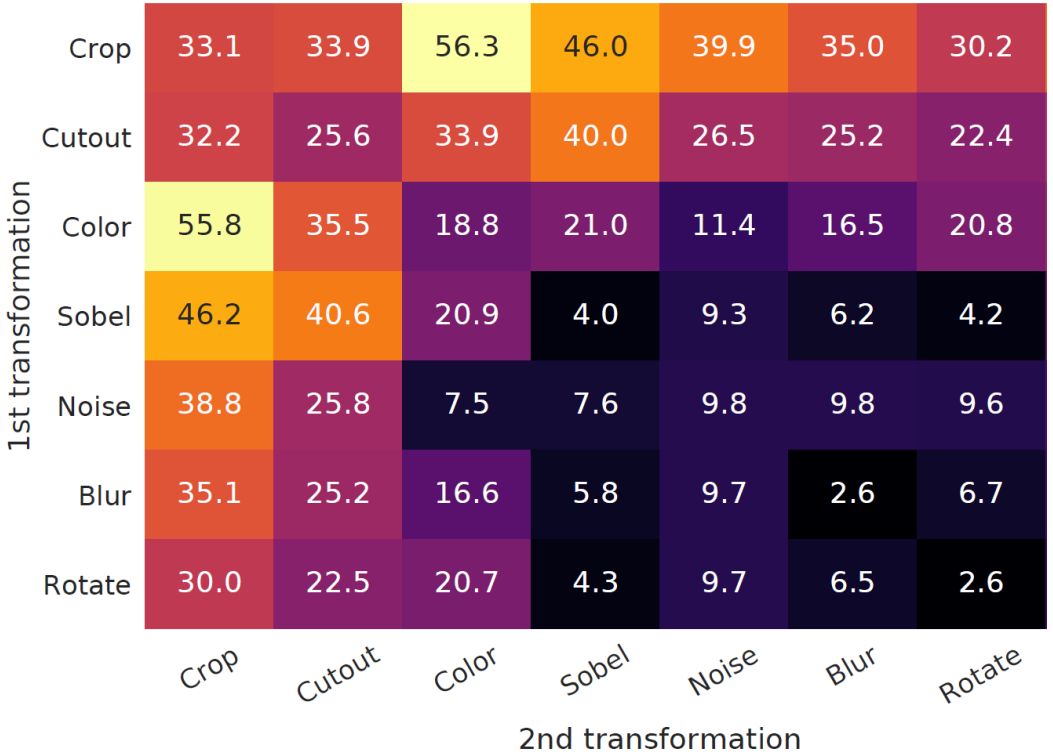
(j) Sobel filtering

Chen, Ting et al. A simple framework for contrastive learning of visual representations. ICML, 2020

Motivation

Some combinations of data augmentation can further improve the performance.

- Combinations of weak augmentations expose unexplored patterns



Methods	Color distortion strength 1/8	1/4	1/2	1	1 (+Blur)	AutoAug
SimCLR	59.6	61.0	62.6	63.2	64.5	61.1
Supervised	77.0	76.7	76.5	75.7	75.4	77.1

Chen, Ting et al. A simple framework for contrastive learning of visual representations. ICML, 2020
Tian, Yonglong, et al. What makes for good views for contrastive learning. arXiv 2020.

Motivation

Contrastive Learning with Stronger Augmentations

- Different combinations of weak augmentations provide different clues for obtaining distinctive feature representations.
- Some useful novel clues may only exist in the stronger augmentations.
- Stronger augmentations greatly boost the performance in supervised learning and semi-supervised learning.

Pave the last mile to close the gap with the fully supervised representation by stronger augmentations.

Challenge

Cannot naively using strongly augmented images in contrastive learning

- A strongly augmented image is perceptually different from the original.
- A strongly augmented image's representation is far apart from the weakly augmented.

Distributional Divergence Minimization (DDM) between Weakly and Strongly Augmented Images.

Method

Another perspective to explain contrastive loss

- Negative pairs' probability

  Likelihood of z_i' being assigned to z_k

  - $$p(z_k|z_i') = \frac{\exp\left(sim(z_i', z_k)/\tau\right)}{\exp\left(sim(z_i', z_i)/\tau\right) + \sum_{k=1}^{K} \exp\left(sim(z_i', z_k)/\tau\right)}$$

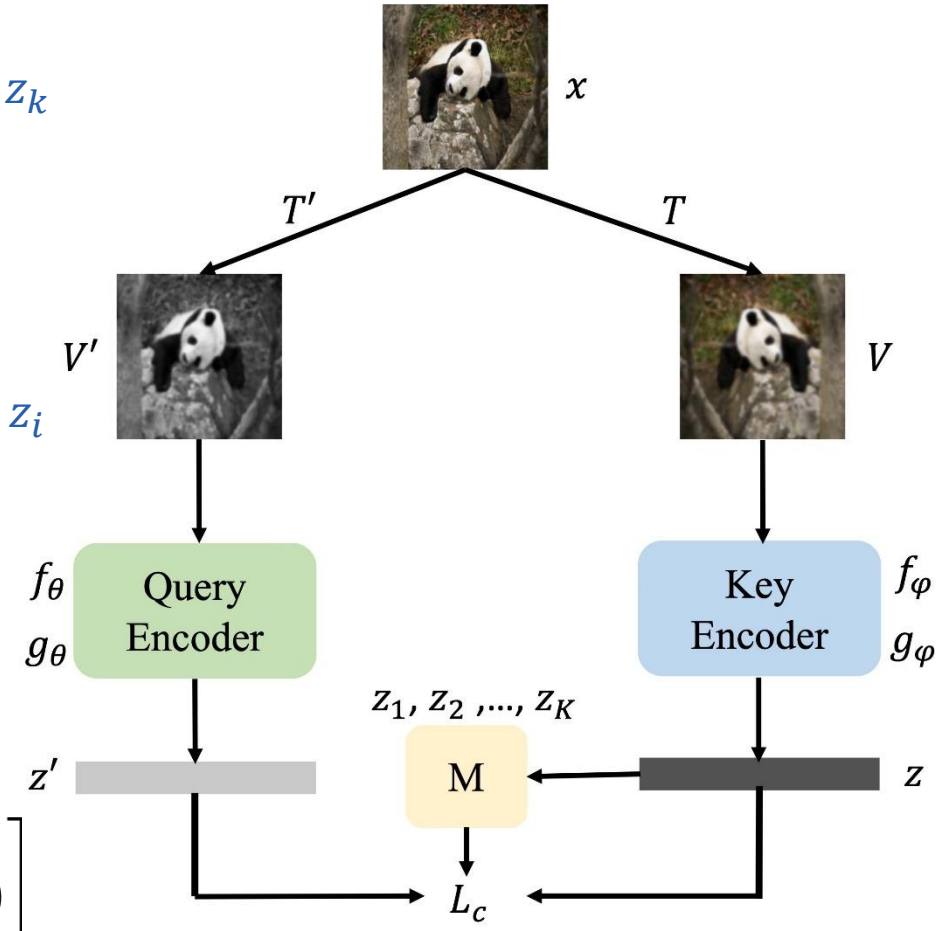
- Positive pairs' probability

  Likelihood of z_i' being assigned to z_i

  - $$p(z_i|z_i') = \frac{\exp\left(sim(z_i', z_i)/\tau\right)}{\exp\left(sim(z_i', z_i)/\tau\right) + \sum_{k=1}^{K} \exp\left(sim(z_i', z_k)/\tau\right)}$$

- Another form of contrastive loss

  - $$\mathcal{L}_c = \mathbb{E}_{i \in B}\left[-q(z_i|z_i')\log p(z_i|z_i') - \sum_{k=1}^{K} q(z_k|z_i')\log p(z_k|z_i')\right]$$



q is the ideal distribution of the likelihood p is the distribution learned by network

Method

Another perspective to explain contrastive loss

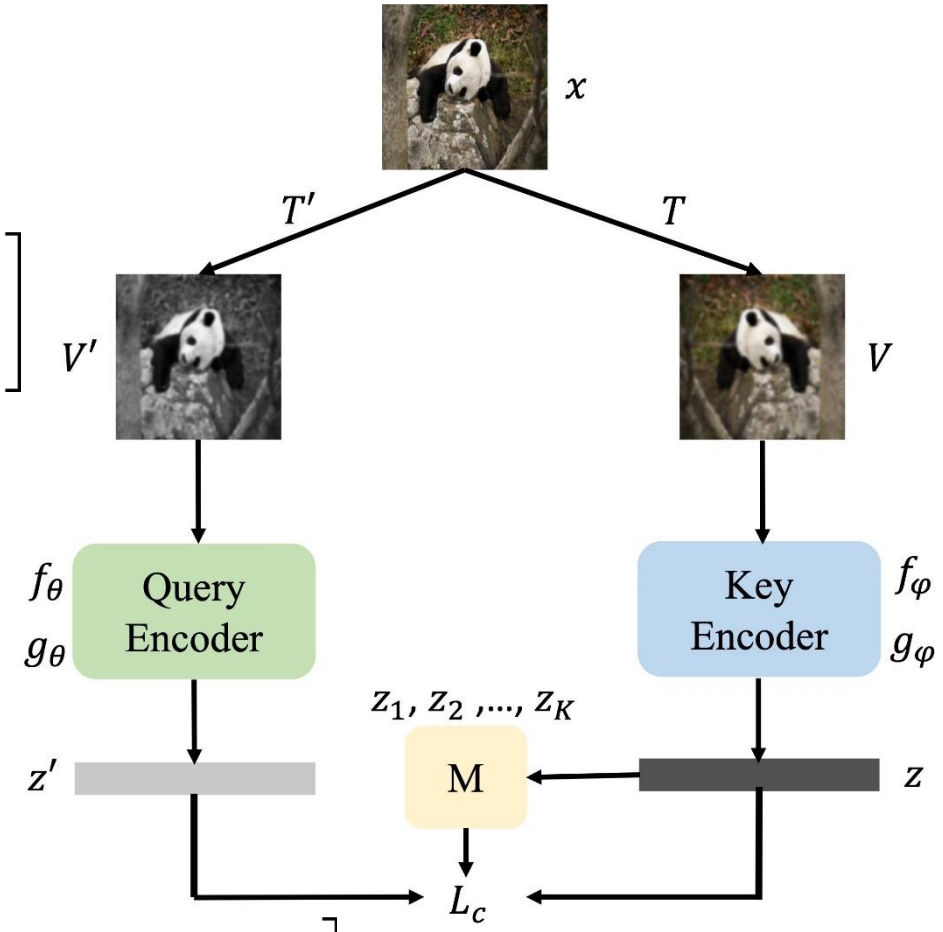
- Another form of contrastive loss
  - $\mathcal{L}_c = \mathbb{E}_{i \in B}\left[-q(z_i|z_i') \log p(z_i|z_i') - \sum_{k=1}^{K} q(z_k|z_i') \log p(z_k|z_i')\right]$

Regard q as a one-hot distribution

It fails for stronger augmentations.

$q(z_i|z_i') = 1$

$q(z_k|z_i') = 0 \ \ (k \in [1, K])$



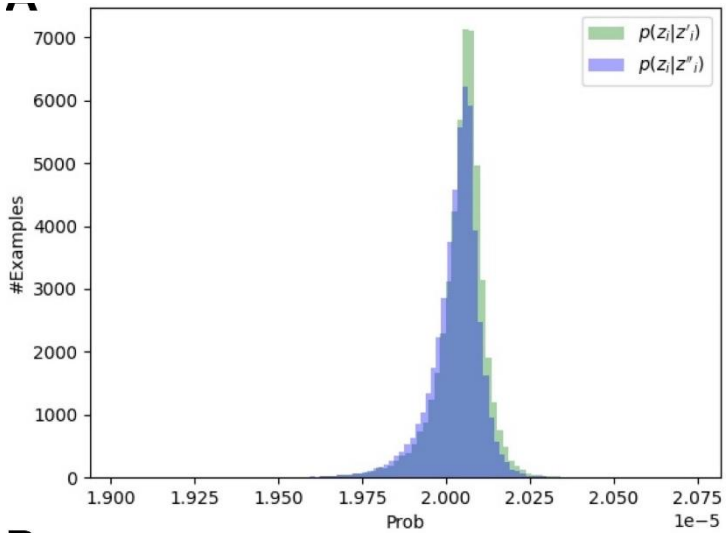
- The original form of contrastive loss
  - $\mathcal{L}_C = \log p(z_i|z_i') = \log\left[\frac{\exp\left(sim(z_i', z_i)/\tau\right)}{\exp\left(sim(z_i', z_i)/\tau\right) + \sum_{k=1}^{K} \exp\left(sim(z_i', z_k)/\tau\right)}\right]$

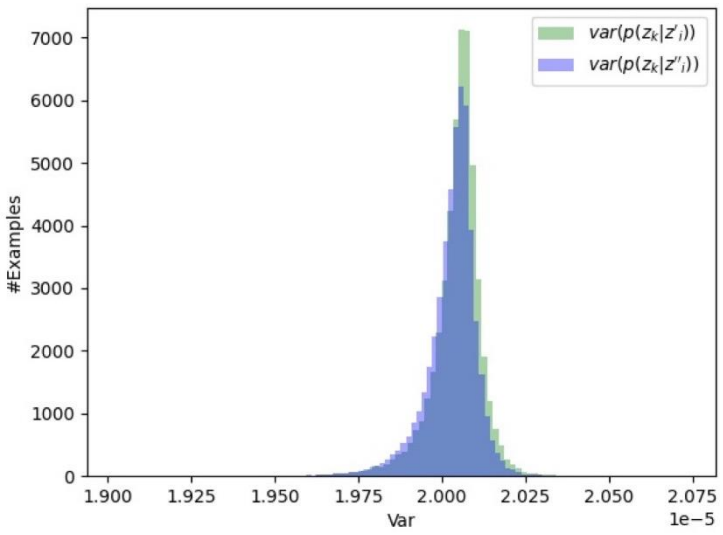
Method

Distribution of positive pairs' probabilities

Variance of negative pairs' probabilities

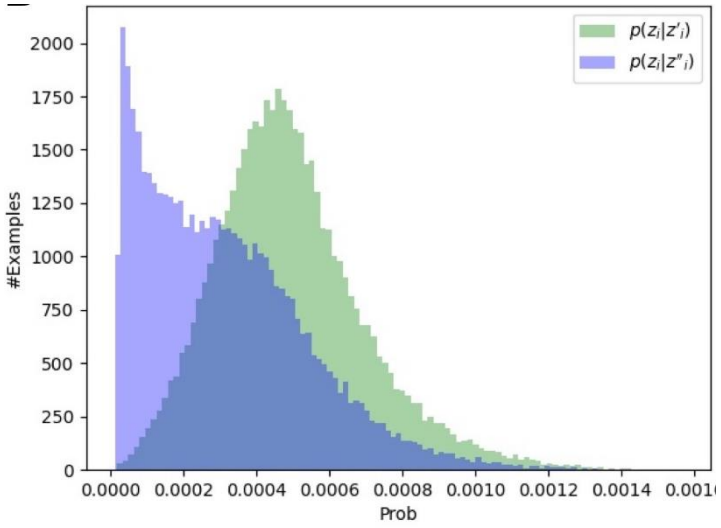
Randomly initialized network

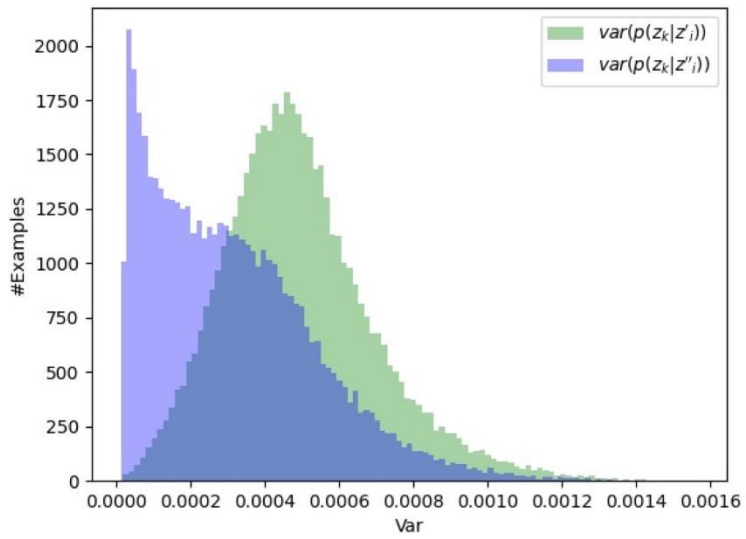




Identical

A pre-trained network by contrastive methods.





Different

Method

Distributional Divergence Minimization between Weakly and Strongly Augmented Images

- Use the distribution of relative similarities of weakly augmented query to supervise that of strongly augmented query.

  - $$p(z_k|z_i'') = \frac{\exp\left(sim(z_i'', z_k)/\tau\right)}{\exp\left(sim(z_i'', z_i)/\tau\right) + \sum_{k=1}^{K} \exp\left(sim(z_i'', z_k)/\tau\right)}$$

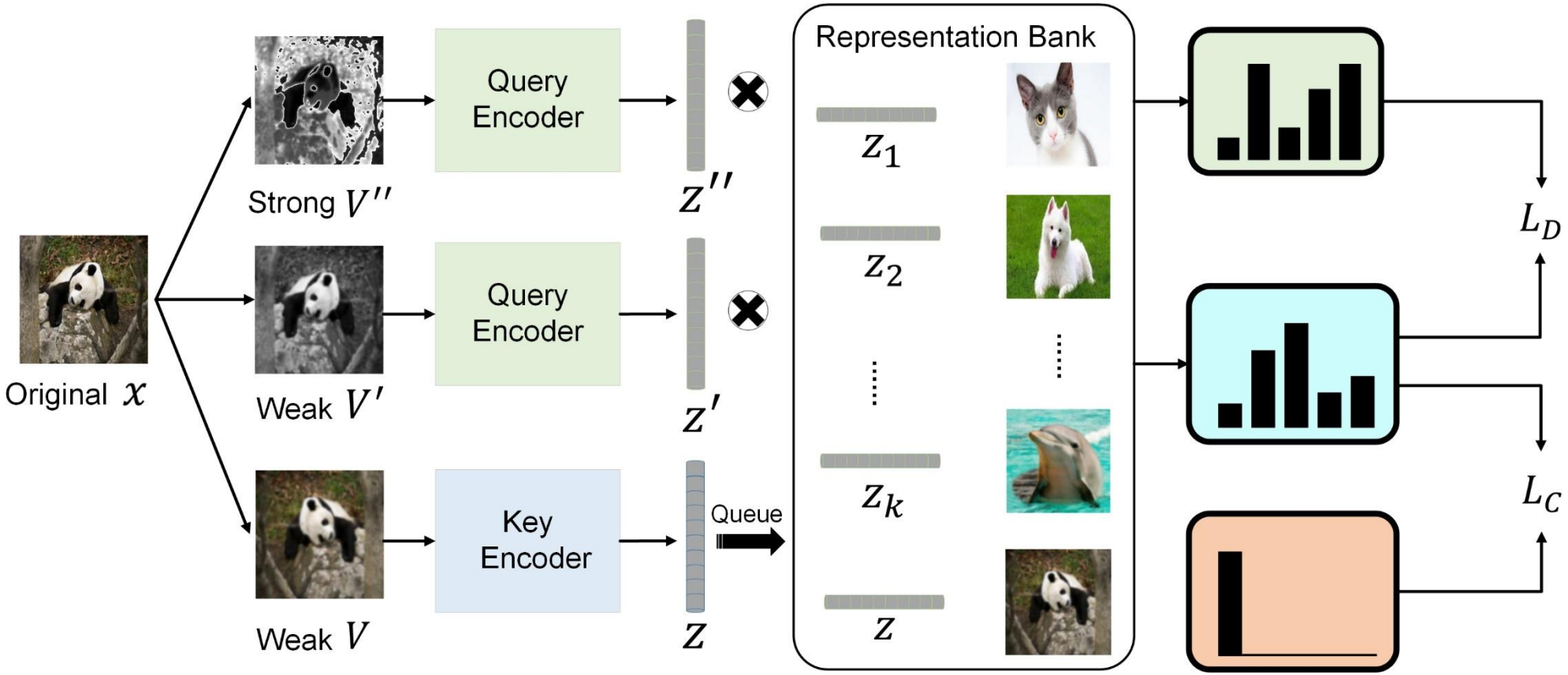
  - $$p(z_i|z_i'') = \frac{\exp\left(sim(z_i'', z_i)/\tau\right)}{\exp\left(sim(z_i'', z_i)/\tau\right) + \sum_{k=1}^{K} \exp\left(sim(z_i'', z_k)/\tau\right)}$$

  - $$\mathcal{L}_D = \mathbb{E}_{i\in B}\left[-\underbrace{p(z_i|z_i')}_{q(z_i|z_i'')}\log p(z_i|z_i'') - \sum_{k=1}^{K}\underbrace{p(z_k|z_i')}_{q(z_k|z_i'')}\log p(z_k|z_i'')\right]$$

Method

Diagram of Distribution divergence minimization

- The overall loss to optimize the encoder: $\mathcal{L} = \mathcal{L}_C + \beta * \mathcal{L}_D$



Experiments

Linear Classification on ImageNet

- CLSA: use a single stronger augmentation.
- CLSA*: adopt five different stronger augmentation.

TABLE 2: Top-1 accuracy under the linear evaluation on ImageNet with the ResNet-50 backbone with 200 epochs training.

Method	Top 1
InstDisc [9]	54.0
LocalAgg [13]	58.8
MoCo [5]	60.8
SimCLR [8]	61.9
CPC v2 [15]	63.8
PCL [36]	65.9
MoCo v2 [4]	67.5
InfoMin Aug [16]	70.1
SWAV [6]	72.7
CLSA	69.4
CLSA*	**73.3**
Supervised	76.5

TABLE 3: Top-1 accuracy under the linear evaluation on ImageNet with the ResNet-50 backbone with various numbers of epochs.

Method	Top 1
BigBiGAN [38]	56.6
SeLa-400epochs [39]	61.5
PIRL-800epochs [3]	63.6
CMC [14]	66.2
SimCLR-800epochs [8]	70.0
MoCo v2-800epochs [4]	71.1
InfoMin Aug-800epochs [16]	73.0
BYOL-1000epochs [20]	74.3
SWAV-800epochs [6]	75.3
CLSA-800epochs	72.2
CLSA*-800epochs	**76.2**
Supervised	76.5

Experiments

Transfer Learning Results on Downstream Tasks

- Cross-dataset image classification: VOC07 dataset
- Object detection: VOC dataset and COCO dataset

TABLE 4: Transfer learning results on various downstream tasks.

	Classification	Object Detection		
	VOC07	VOC07+12	COCO	
Measurement	Accuracy	AP_{50}	AP	AP_S
RotNet [40]	64.6	-	-	-
NPID++ [9]	76.6	79.1	-	-
MoCo [5]	79.8	81.5	-	-
PIRL [3]	81.1	80.7	-	-
PCL [36]	84.0	-	-	-
BoWNet [41]	79.3	81.3	-	-
SimCLR [8]	86.4	-	-	-
MoCov2 [4]	87.1	82.5	42.0	20.8
SWAV [6]	88.9	82.6	42.1	19.7
CLSA	**93.6**	**83.2**	**42.3**	**24.4**
Supervised	87.5	81.3	40.8	20.1

Experiments

Ablation Study

- DMM Loss: study the role of DDM loss in the CLSA.
- Running Time: study the extra training time consuming of CLSA compared to MOCO V2.

TABLE 5: Ablation study of the CLSA on ImageNet with 200 epochs of pre-training.

Model	Top-1
MoCo V2	67.5
MoCo V2 with Strong query	67.7
MoCo V2 with Strong query & Strong key	67.0
CLSA with contrastive loss	68.0
CLSA	**69.4**

TABLE 6: Training Time Comparison of CLSA

Model	Time	Epoch	Top-1
MoCo V2	53h	200	67.5
CLSA	35h	100	67.2
CLSA	52.5h	150	68.3
CLSA	70h	200	69.4

Experiments

Ablation Study

- The strength of strong augmentation
- Representations for weak/strong augmented images

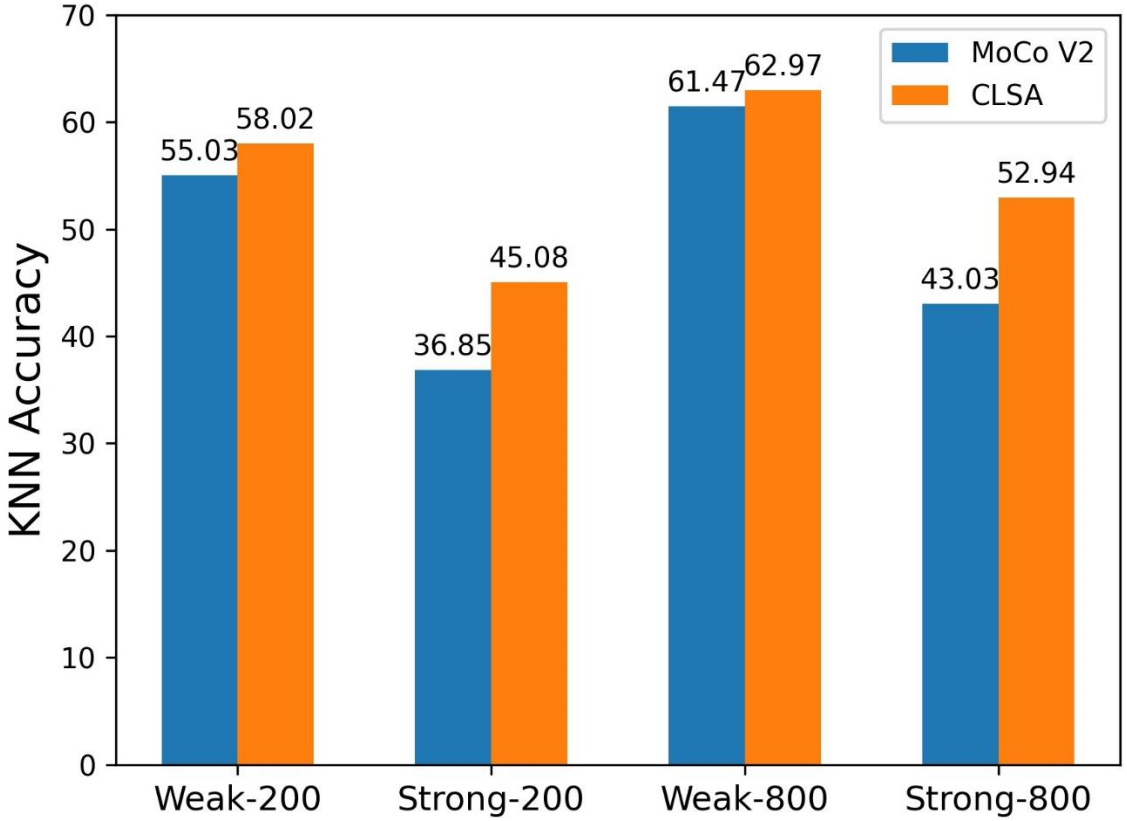


TABLE 7: Comparison of CLSA under different strong augmentations

Strength s	3	5 (default)	7
KNN Acc	52.8	**53.0**	52.6

Fig. 5: Comparison of KNN accuracy of MoCo V2 [4] and CLSA. Both results are compared based on pretrained model with 200/800 epochs with single crop. The comparison used the representation of weakly/strongly augmented images, respectively. The neighbor K for KNN here is set to 20.

Experiments

Ablation Study

- Distribution of Relative Similarity by CLSA

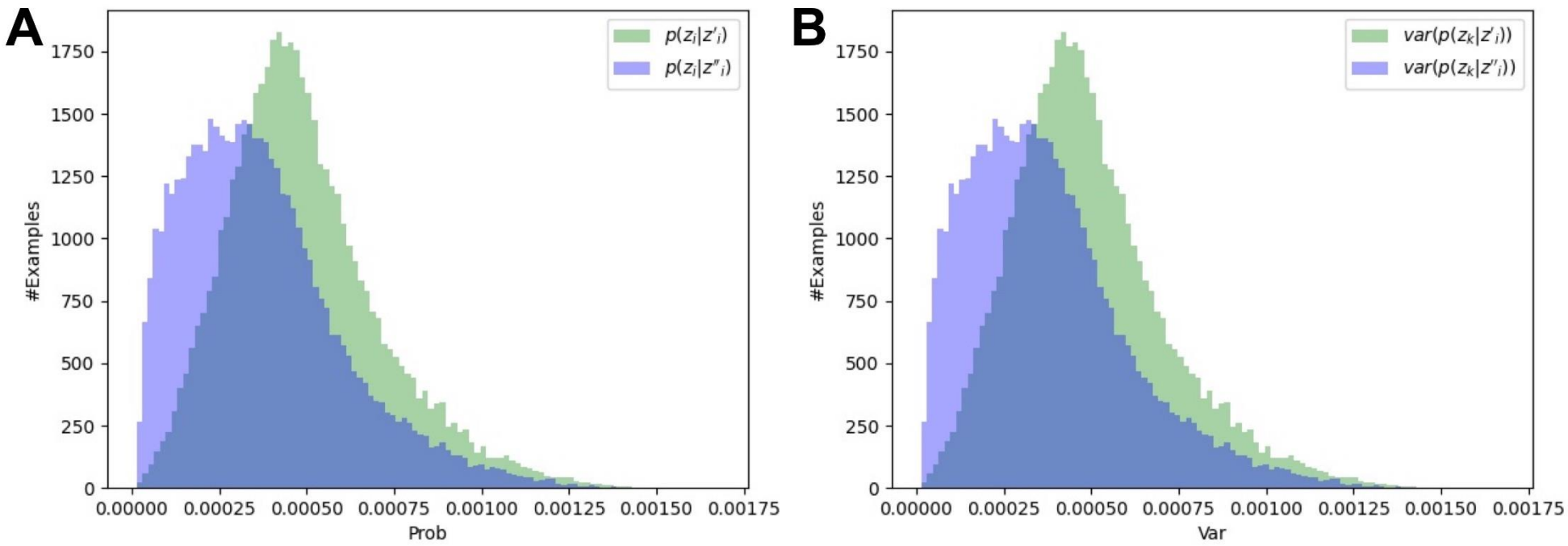


Fig. 6: The comparison of the distribution of positive pair's probabilities and variance of negative pairs' probabilities with a pre-trained network by CLSA. **A.** The distribution comparison of positive pair's probabilities. **B.** The distribution comparison of variance of negative pairs' probabilities.

Reviews

Final Decision

ICLR 2021 Conference Program Chairs

08 Jan 2021 (modified: 13 Jan 2021) ICLR 2021 Conference Paper1274 Decision Readers: Everyone

Decision: Reject

Comment:

This paper improves MoCo-based contrastive learning frameworks by enabling stronger views via an additional divergence loss to the standard (weaker) views. Three reviewers suggested acceptance, and one did rejection. Positive reviewers found the proposed method is novel and shows promising empirical results. However, as pointed out by the negative reviewer, the paper should have clarified about computational overheads of the method compared to the baseline (MoCo) in several aspects, e.g., their effective batch sizes or training costs, for the readers' better understanding. As the concern was not fully resolved during the discussion phase, AC is a bit toward for reject. AC thinks the paper would be stronger if the authors include more ablations (and the respective discussions) regarding this point, e.g., CLSA-multi (and -single) vs. MoCo-v2 under the same training time, both at early epochs (~200; as reported in the author response) and longer epochs (after convergence; ~1000 and even more).

Rejection

Reviews

The experimental evaluations are not convincing

ICLR 2021 Conference Paper1274 AnonReviewer1

03 Nov 2020 (modified: 11 Nov 2020) ICLR 2021 Conference Paper1274 Official Review Readers: Everyone

Review:

This paper focuses on designing more effective ways for contrastive learning. The author claims that stronger augmentations are beneficial for better representation learning. Different from directly applying the stronger augmentations to minimize the contrastive loss, the author proposes to minimize the distribution divergence between the weakly and strongly augmented images. The experimental evaluations are conducted on ImageNet classification and related downstream tasks, and the results are promising.

Clarity:

1. The method is very simple and straightforward. My main concern is the experimental comparisons. As we all know, contrastive learning algorithms like MOCO and SimCLR benefits from longer training epochs a lot (for example, training with 800 epochs is much better than with 400 epochs). Thus I think the comparisons in Table 2 are not convincing. From algorithm 1, we can find that the equivalent batch size of the proposed CLSA method is two times as classical MOCO method. Thus I would prefer to check the results of CLSA at epochs 100 and 400 for fair comparisons.
2. What is the value of the balancing coefficient? It would be nice if some ablation results are provided.

Rating: 4: Ok but not good enough - rejection

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Rejection

Reviews

CONTRASTIVE LEARNING WITH STRONGER AUGMENTATIONS

ICLR 2021 Conference Paper1274 AnonReviewer4

28 Oct 2020 (modified: 25 Nov 2020) ICLR 2021 Conference Paper1274 Official Review Readers: Everyone

Review:

This paper proposes the better utilization of strong data augmentations for contrastive loss functions in unsupervised learning. In Moco set up, typically, weaker augmentations such as color jittering, cropping is applied to construct positive pairs from the same image. In this study, by proposing a modified objective, the authors leverage stronger data augmentations to construct more challenging positives and negatives pairs to improve the quality of the representations. The paper delivers a novel objective together with leveraging existing strong augmentations to improve downstream performance. The authors can find my questions/concerns listed below.

1. The paper is overall well-written, however, it is disappointing to see many typos grammar mistakes throughout the paper. Some examples are in "Thus we proposed the CLSA (Contrastive Learning with Stronger Augementations)", "to train an unsupervised representation", "The contrastive learning (Hadsell et al. (2006)) is a popular self-supervised idea".
2. In section 3.1, the authors mention that the keys in the memory bank is managed with first in first out method. Is it not supposed to be first in last out? I would like to see some clarification on this.
3. The numerator in Equation 3 should be z_i' vs. z_i not z_k.
4. The authors claim that in He et al. an input image is resized and cropped to 224×224 pixels. It should be "an image is first cropped from an input image and resized to 224x224 pixels."
5. In the experiments section, the authors list other methods including MoCo, SimCLR, MoCo-v2, BYOL and compare to what they propose. As a baseline, it would be nice to directly use the stronger augmentations in MoCo-v2 objective and perform comparison to their method. Throughout the paper, the authors claim that strong augmentations hurt the learned representations due to distorted images. It would be meaningful to show this experimentally as well.
6. The authors explain that they choose a strong transformation randomly from the given 14 transformations and repeat it 5 times to strongly augment an image. Is the sampling done without replacement? In other words, do the authors choose 5 unique transformations with the corresponding magnitude and apply those transformations to a single image?
7. I like how the authors point the similarity of their objective to knowledge distillation. In this case, strong augmentations are assigned probability of being a positive pair from the positive pair constructed with weak augmentations. It helps to understand the full picture for the proposed method.
8. Finally, I think the figure 3 is confusing rather than being helpful. Both weak and strong augmentations go to the memory bank and it looks like two distributions come out of nowhere in the figure. It would be more clear to point out that there is distribution of the representations from the strong augmentations and weak augmentations and they supervise the assignment for strong augmentations given predictions on the weak augmentations.

Rating: 7: Good paper, accept

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Accept

Reviews

Official Blind Review #3

ICLR 2021 Conference Paper1274 AnonReviewer3

27 Oct 2020 (modified: 11 Nov 2020) ICLR 2021 Conference Paper1274 Official Review Readers: Everyone

Review:

This paper presents a method to incorporate stronger augmentations into the visual representation contrastive learning framework. Specifically, three correlated views of an image are first generated by using two weak and one strong augmentation operations on the same image. Then, the networks are trained to maximize the agreement between the two weak views and also to minimize the distribution divergence between a weak view and the strong view. The method is evaluated on several visual tasks including classification, transfer learning, and object detection, with the standard evaluation protocol for self-supervised learning, and the results are promising.

Pros:

1. This paper is well-structured and easy-to-follow.
2. The idea of utilizing strong augmentations for contrastive learning is interesting and novel to me, and the results are promising.
3. The proposed framework seems general which might be easily incorporated into the existing contrastive learning frameworks.

Cons:

1. The motivation about using stronger augmentations is not well justified. Specifically, the authors propose to use stronger augmentations based on two reasons: (1) stronger augmentations can expose some novel useful patterns; (2) the effectiveness of stronger augmentations is proved in the semi-supervised learning and supervised learning field. However, no related papers are provided to support the first point, while the papers (Cubuk et al. (2018)); Qi et al. (2019); Wang et al. (2019)) that are cited to support the second point do not explicitly make relevant conclusions. (Chen et al. (2020a)) even demonstrate that when training supervised models, stronger color augmentation hurts their performance. I would like to see a more comprehensive review of related works to clarify the motivation.
2. In addition, some important ablation studies are missing in the experiment. E.g., how does the performance change as the magnitude or usage times of stronger augmentations changes?
3. The proposed DDM loss seems general for different contrastive learning frameworks. I would like to see if it still works when applied to other frameworks, e.g., SimCLR, InfoMin?

Overall, given the novelty and strong results of the proposed framework, I remain positive towards this paper. I will be happy to increase my rating if my concerns are addressed in the rebuttal period.

Rating: 6: Marginally above acceptance threshold

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Weak Accept

Reviews

Official Blind Review #2

ICLR 2021 Conference Paper1274 AnonReviewer2

25 Oct 2020 (modified: 11 Nov 2020) ICLR 2021 Conference Paper1274 Official Review Readers: Everyone

Review:

Summary:
This work investigate the recent popular direction of unsupervised representation learning using contrastive loss between augmented images. Authors propose to minimize the divergence between the distributions of strongly augmented vs. weakly augmented images. The method reaches competitive performance in recognition and object detection.

+Strengths
+The main idea is well motivated: that strong augmentation reveal useful cues in visual representation learning but has not been successfully exploited in unsupervised learning.
+The proposed solution is novel within contrastive learning to my best knowledge.
+Results are extremely strong.

-Concerns
-The divergence between the two conditional distributions can be a moving target since they are trained jointly. It is not clear if this is will result in stable learning for the unsupervised setting, and what effect that may have on the performance and quality of the representations.
-Evaluation only focus on final result and lacks analysis of the proposed method, especially when compared to recent paper of similar nature published in top conferences. For example, strong augmentation is a focus of this paper, but there are no ablation regarding the augmentations. Is the performance sensitive to the choice of strong augmentation?
-The paper could also use some more theoretical analysis to address some of the weaknesses stated above.

Recommendation
I like the proposed idea. It is novel and interesting and seems to achieve good results. However the lack of both theoretical and empirical analysis beyond results on performance raises many questions. As a result I am on the fence but leaning towards accept.

Rating: 6: Marginally above acceptance threshold

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Weak Accept

THANKS