Improving Dense Contrastive Learning with Dense Negative Pairs

Berk Iskender, Zhenlin Xu, Simon Kornblith, En-Hung Chu, Maryam Khademi

berki2@illinois.edu, maryamkhademi@google.com

Introduction

- Contrastive learning (CL) performs the pre-text task of instance discrimination
- Global CL (e.g. SimCLR [1]):
  - Trains a single global representation
  - Evaluates on single-label classification
- Global CL can be suboptimal for
  - Multi-label classification
    - Each label ➡ a different object
    - Different region ➡ Different semantic content
  - Dense tasks (e.g. segmentation, detection)
- DenseCL [2] uses dense features to boost performance
  - Dense-Dense positive pairs
  - Dense-Global negative pairs

Motivation & Contributions

- Inspired by DenseCL, we aim to improve the performance by modifying
  - The training scheme
  - The objective function
- Proposed approach **DenseCL++**
  - Dense-Dense negative pairs between the features of augmented views
  - Modified dense contrastive loss
  - Different negative pair formulation alternatives
  - +3.5% and +4% mAP over SimCLR and DenseCL in COCO multi-label classification using ViT-S/16 [3] as encoder
  - +1.8% and +0.7% mIoU over SimCLR in COCO and VOC semantic segmentation

Method

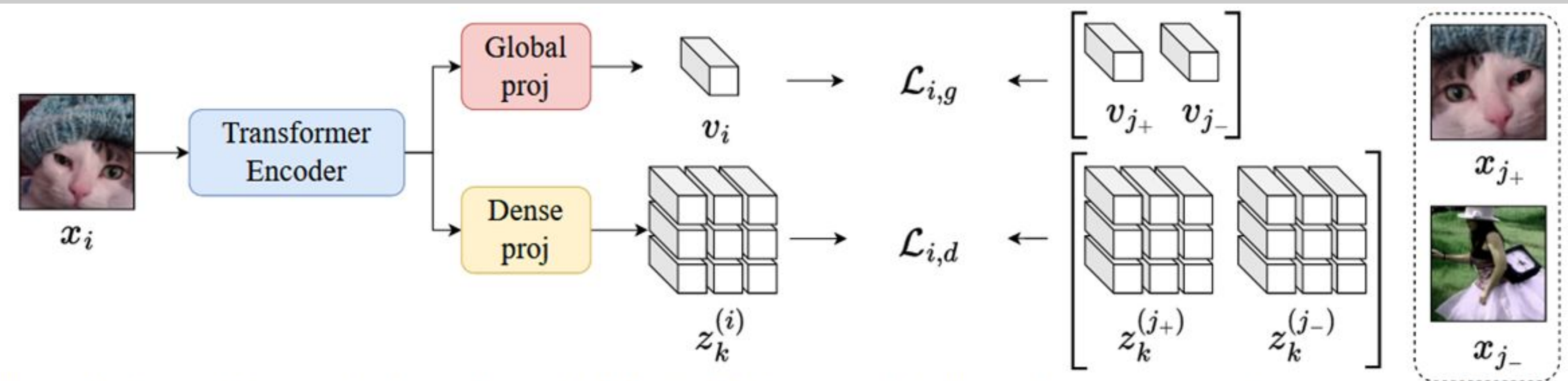


Figure 1: DenseCL++ training scheme. Global and dense positive/negative correspondences are used in the global (top row) and dense (bottom row) loss functions, respectively.

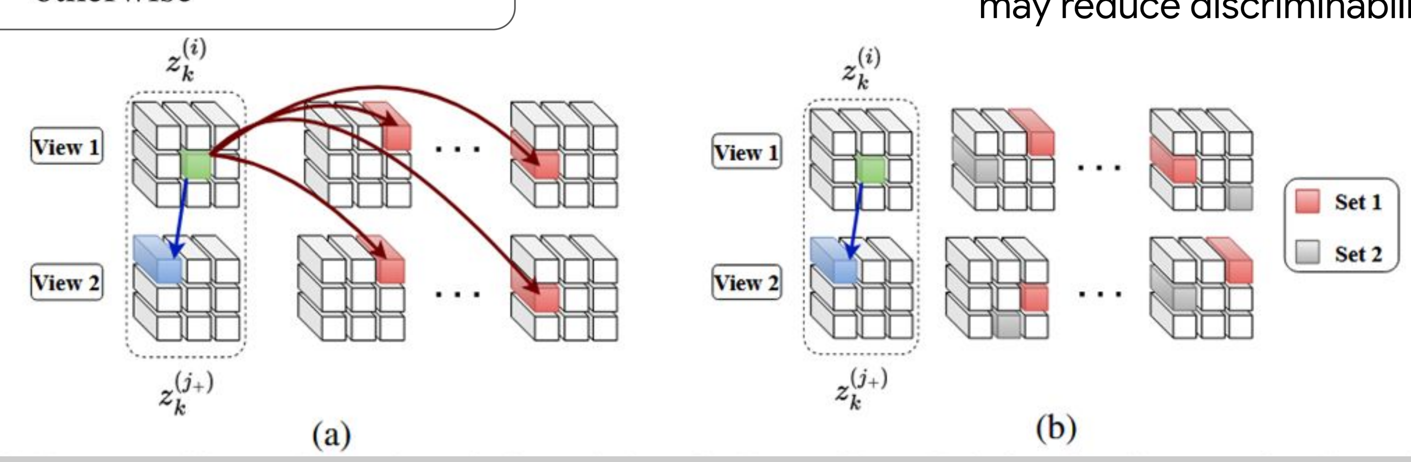
- **Training objective:** $\mathcal{L}_i = (1-\lambda)\mathcal{L}_{i,g} + \lambda\mathcal{L}_{i,d}$
- **Dense Contrastive Loss:**

DenseCL:
$$\mathcal{L}_{i,d} = \sum_k -\log \frac{\exp(z_k^{(i)} \cdot z_{k_+}^{(j_+)})/\tau}{\exp(z_k^{(i)} \cdot z_{k_+}^{(j_+)}) + \boxed{\sum_{j_-} \exp(z_k^{(i)} \cdot v_{j_-})/\tau}}$$

DenseCL++:
$$\mathcal{L}_{i,d} = \sum_k -\log \frac{\exp(z_k^{(i)} \cdot z_{k_+}^{(j_+)})/\tau}{\exp(z_k^{(i)} \cdot z_{k_+}^{(j_+)}) + \boxed{\sum_{j_-,m} \exp(z_k^{(i)} \cdot z_m^{(j_-)})/\tau}}$$

- **Dense Negative Pair Formulation Alternatives:**

1. **Random sampling (a) (Baseline):** A random dense feature from each augmented view in the batch
2. **Guided dense negative formulation (b):** Select the most similar set on average to anchor features among M sets
   More similar sets ➡ Potentially harder negatives
3. **Thresholding:** For 2.
   $$\bar{q} = \begin{cases} -1, & q \leq \beta \\ q, & \text{otherwise} \end{cases} \qquad q = \mathrm{sim}(a, b)$$
4. **N cross-view dense negatives:** Only cross-view positives ➡ High similarities may reduce discriminability



(a) (b)

Main Results

Method	Agg.	Pair feature	mAP	F1
DenseCL	CLS	backbone	**59.9**	**38.1**
		proj head	59.8	**38.1**
	GAP	backbone	58.1	37.3
		proj head	57.8	37.5
SimCLR	CLS	-	**59.6**	**37.9**
	GAP	-	58.4	37.7

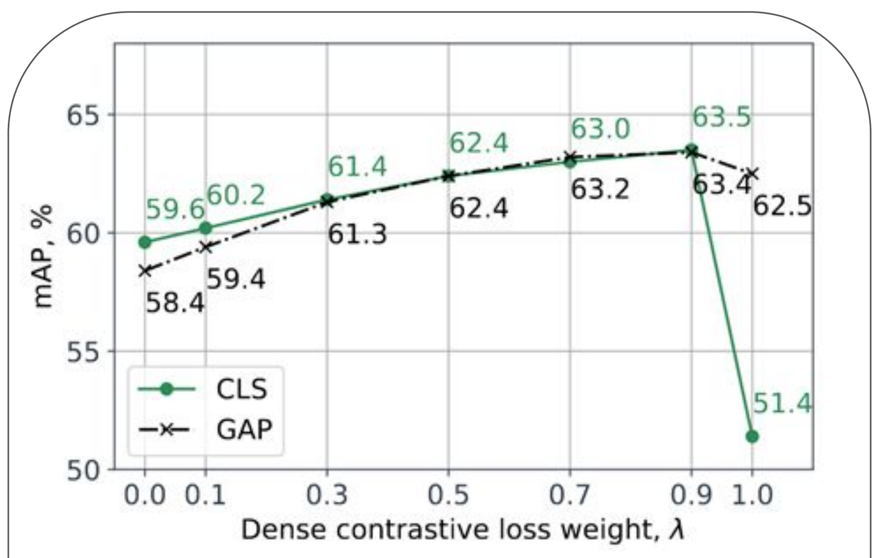
SimCLR and DenseCL multi-label classification results on COCO for different global feature aggregation and dense matching types.

Method	Agg.	Pair feature	mAP	F1
SimCLR	CLS	–	59.6	37.8
DenseCL	CLS	backbone	59.9	38.1
DenseCL++	GAP	backbone	**63.4**	**39.0**
DenseCL++*	GAP	backbone	**64.1**	**39.1**

Top performing settings for multi-label classification on COCO using different contrastive learning methods. DenseCL: baseline, DenseCL++: M=256, β=0.5, N=64.*

Method	Pair feature	VOC mIoU	COCO mIoU
SimCLR	–	69.3	61.5
DenseCL++	backbone	**70.0**	**63.3**

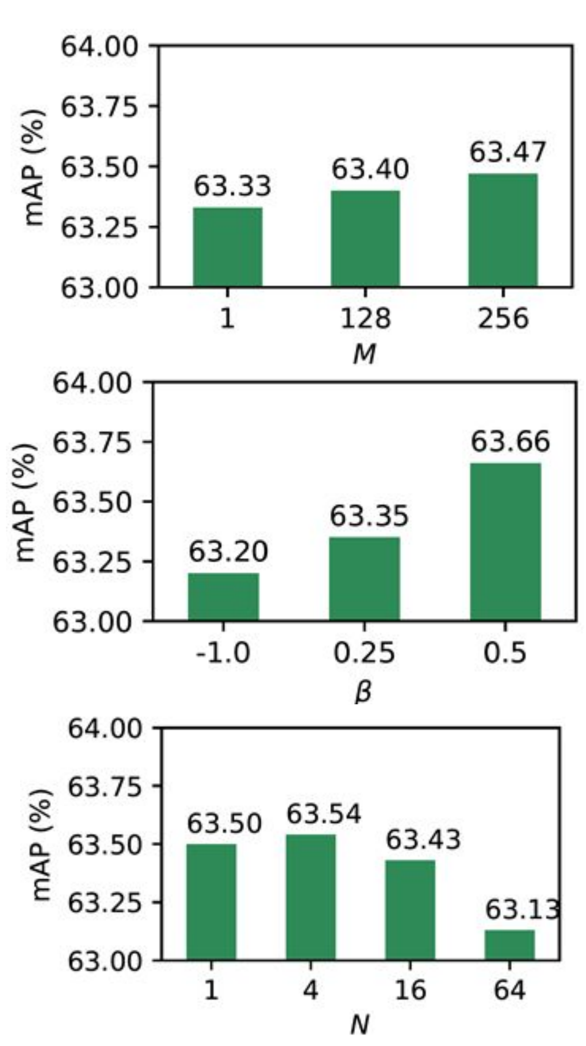
Semantic segmentation on PASCAL VOC and MS COCO with GAP aggregation.



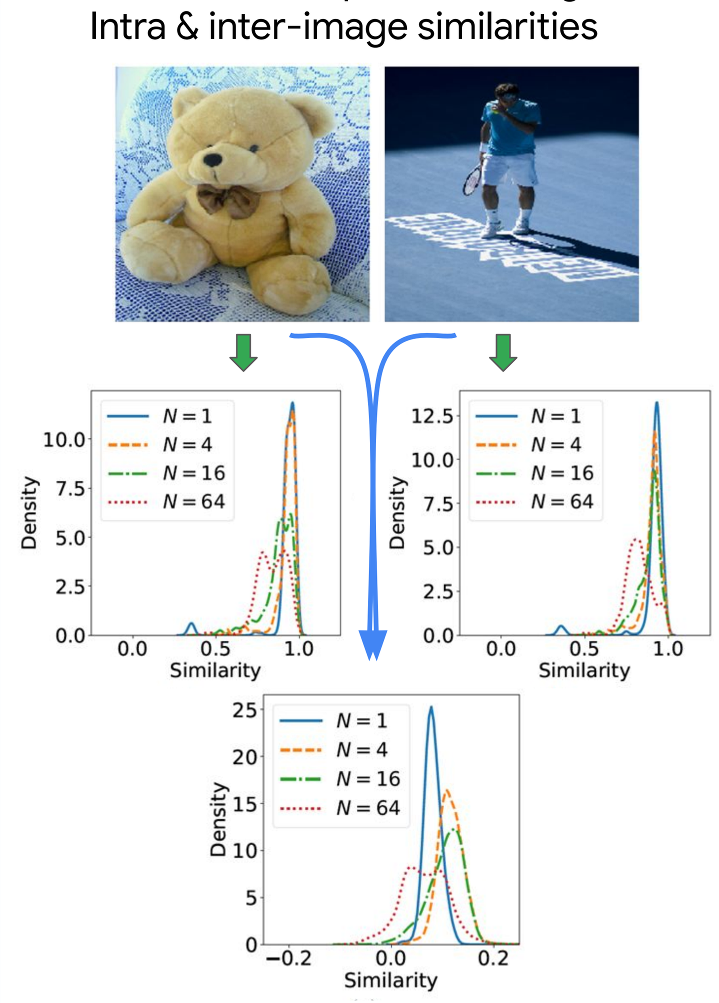
mAP vs. dense contrastive loss weight λ for DenseCL++ for different global feature aggregation settings

Dense Negative Pair Formulation Studies

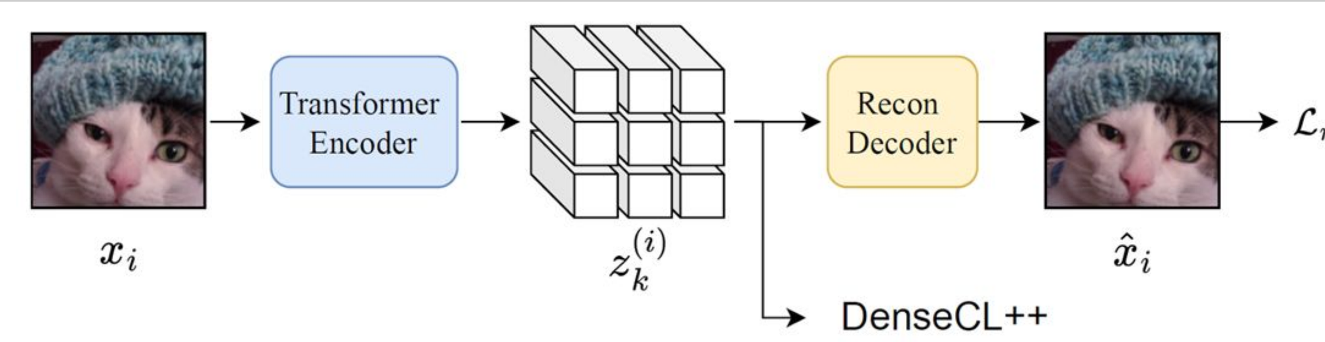
- Averaged mAP for 36 experiments with 3x3x4 configs of MxβxN:



- Effect of multiple cross negatives: Intra & inter-image similarities



Reconstruction as an Auxiliary Task

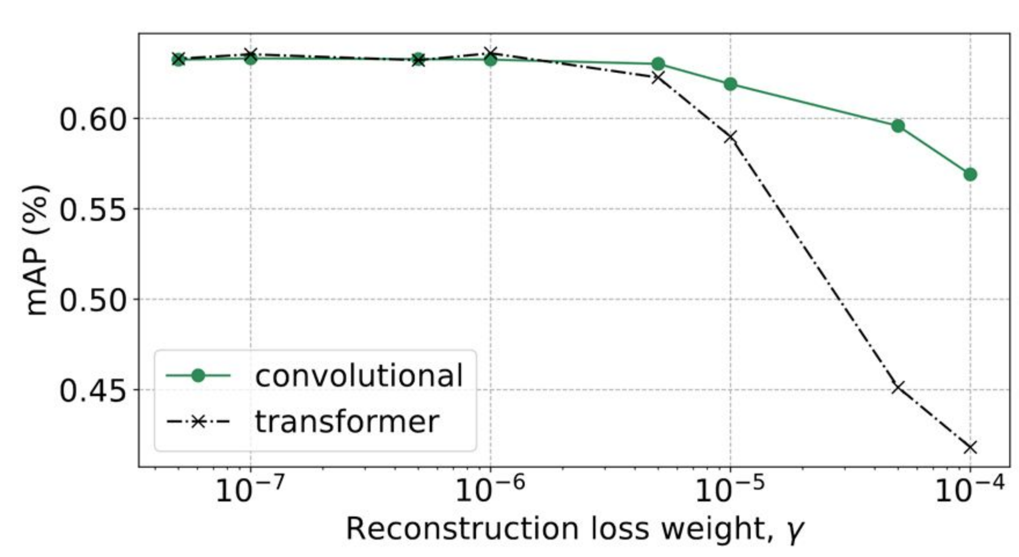


Recon loss: $\mathcal{L}_r = ||x_i - \hat{x}_i||$

Objective: $\mathcal{L} = (1-\lambda)\mathcal{L}_g + \lambda\mathcal{L}_d + \gamma\mathcal{L}_r$

- Tested simple convolutional/transformer decoders

  Prioritizing Accuracy ➡ Degrading eval performance



mAP vs. reconstruction loss weight γ for simple convolutional and transformer-based decoders

Conclusion

- Replacing dense-global negatives with dense-dense counterparts improve evaluation performance of dense contrastive learning for multi-label classification and semantic segmentation
- Various dense negative formulation techniques provide additional improvement for multi-label classification when combined
- Reconstruction as an auxiliary task for DenseCL++
  - Marginal or no improvement
  - Difficult to find an optimal setting
  - Harmful when prioritized

References

[1] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

[2] Wang, X., Zhang, R., Shen, C., Kong, T., & Li, L. (2021). Dense contrastive learning for self-supervised visual pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3024-3033).

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.