DAC: A Double Accelerating Contrastive Learning Framework

Zixuan Pan Zihao Wei
Department of Computer Science Engineering
University of Michigan
Ann Arbor, USA
{zxp,zihaowei}@umich.edu

1. Introduction

Self-supervised pretraining has achieved great success in natural language processing [2, 11, 24]. Recently, these models have been adapted to computer vision and have created new state-of-the-arts results on many tasks. [8, 9, 12, 15]. The two main two main pretraining paradigms are contrastive learning and reconstruction pretraining. Among both paradigms, cropping has been seen as an indispensable augmentation method to improve the model's general performance. Resizing techniques, which are usually used together with cropping, are usually neglected by the researchers.

Resizing, especially down-sampling, has an inherent augmentation effect, which can map high-dimension images to lower-dimension spaces [18]. However, directly using images of different sizes may have many kinds of problems. One problem is the domain shift problem that the resized images' size can be different from the size during testing [26]. Another problem is that the popular CNN or CNN-like architecture will generate different shape representations for input images of different shapes, which makes batch-wise training and testing impossible [17, 19, 20].

In this paper, we proposed a **D**ouble **A**ccelerating **C**ontrastive Learning Framework, called DAC, which can deal with these two problems and make full use of the advantage of resizing.

For the domain-shift problem, we would like to take advantage of contrastive learning [5, 6, 9, 14, 16]. Contrastive learning generally takes two views of a single image , aiming to distinguish views of the same image from views of different images. Since we treat resizing as an augmentation method, a trivial idea is to use the resized image as one view of a contrastive learning framework [8]. For the other view, we are inspired by the mask autoencoder(MAE), where we use masking as another type of augmentation, which is very similar to very strong blurring augmentations [15]. A notable difference of DAC is that by using down-sampling and masking, the two views of the images are both in a lower dimension space compared to the original image, which can greatly accelerate the training process.

For the different representation size, thanks to the vision transformer(ViT), we are able to deal with images of different sizes by introducing a class token, or CLS token, which will give coherent size representations for images [13]. Meanwhile, since the CLS-token is generally stable, which make introducing simsiam contrastive learning loss across different layers of the ViT possible [25].

Inspired by the [27], Wang et al. improve the model's performance by adding reconstruction target to contrastive learning models. We also reserve the reconstruction part of the masking track to enjoy the benefit of both contrastive and reconstructive paradigms.

Our DAC proposed a general idea that lower dimension mapping augmentations, like down-sampling and masking, can boost both the training accuracy and speed of the model. The idea can be plugged into any existing contrastive learning, self-knowledge distillation or reconstruction frameworks [4, 9, 14]. The extensive experiments show the effectiveness of this method, we achieve 0.2% better than MAE on ImageNet-1K top-1 accuracy by using only half of the MAE's training time.

Overall, our main contributions can be summarized as: **1)** We propose a new perspective of two views of an image used in contrastive learning, where two views both have much smaller input dimension compared to the original image, which can greatly accelerate training speed as well as raise model performance. **2)** We propose an asymmetric framework, taking contrastive learning inputs from different transformer layers, which could also make the model better and faster.

2. Related Works

2.1. Masked Vision Modeling

Most recent works in self-supervised learning are focusing on training vision transformers by using masked images to reconstruct the original ones [13]. Researchers have been testing different kinds of reconstruction objectives and the three main categories are token-wise, feature-wise and pixel-wise reconstruction [1, 5, 12, 15, 28, 29]. These kinds of pretraining tasks are called Masked Image Modeling (MIM) [1]. Recently, MIM has also been introduced to other frameworks like self-distillation, autoregressive generation and contrastive learning, which can further improve the model's performance and learn more representative representations [5, 7, 23, 30]. Unlike these methods where reconstruction is used as an accessory, we fully exploit the advantage of reconstruction in our frameworks.

2.2. Contrastive Learning

Contrastive learning is another active self-supervised learning area, where the model will try to distinguish different views of the same image and the other entirely different ones [5, 6, 9, 14, 16]. The different views of the image are generated through different kinds of augmentations like cropping, color jitter and gray-scaling. These augmentations have been proved essential for the success of contrastive learning. Recently, ViT has been introduced to the field of contrastive learning, where they use the class token as the representation of the entire image and achieve better score than those traditional CNN backbones [4, 9]. Unlike previous works, our DAC uses two novel views, which are lower dimension mappings of the original inputs, and introduces an asymmetric architecture to adapt a reconstruction target to the original contrastive model.

3. Method

3.1. Resizing as Data Augmentation

In contrastive learning, we always need to provide two different views of the same input image. In DAC, we proposed a totally new pair of views, which use masking and resizing as augmentation methods. This is based on the fact that both of them are low-rank approximations to the original image. To further improve the model's performance, we also add other data augmentations to the proposed two views, following the Moco V3's convention [9]. An intuition of the proposed views is to view the masked image as one with high probability of Gaussian blur, while the resized image as one with low probability of Gaussian blur. An obvious advantage of the proposed data augmentation is that it implicitly provides multi-scale views of the original image without losing too much information, since we no longer need cropping or downsampling. In practice, we make the resized image have the same number of patches as unmasked parts of the original image.

3.2. Asymmetric Contrastive Learning

We applied SimSiam as our contrastive learning backbone, which does not need any extra target network and can greatly accelerate training speed. Unlike previous methods, where the augmented images will be passed through two feature extraction networks of the same structure, we choose to use features gained from different stages of the networks for contrastive learning.

More specifically, we will use the encoder output for resized image, and the decoder output for the masked image, since we believe they are are the most representative features for the two views. One potential reason is that masked images can only gain global information of the image after the decoder, while resized images have a similar distribution as those used in downstream tasks, which gives the intuition that the encoder, which will be used during fine-tuning, is sufficient to extract high quality features.

The theoretical insights of the comparability between different stages is that image representation is stable throughout the transformer, which makes it possible to compare the CLS token between different layers [3].

3.3. DAC as a whole

As shown in Fig 1,the reconstruction task and contrastive learning will share the same encoder and decoder. Compared to the original MAE model, we only add an extra two-layer MLP projection layer and 2-layer MLP predictor. Note that reconstruction task are only trained on the masked image.

The loss function is given by:

$$\mathcal{L} = \mathcal{L}_C + \lambda \mathcal{L}_R$$

where $\mathcal{L}_C$ is the SimSiam loss given by the contrastive learning, and $\mathcal{L}_R$ is the L2 loss given by the pixel-wise reconstruction.

4. Experiments

Our experiments are carried out on two datasets: ImageNet-1K and ImageNet-tiny [10]. ImageNet is a widely used dataset for image classification. ImageNet-1K contains 1000 class with each class having roughly 1000 training samples and 50 test samples. Imagenet-mini is a subset of ImageNet-1K with each class having 20-30 trainig samples and 3-5 test samples.

4.1. Implementation Details

Our configuration for pretraining on ImageNet-1K is shown in Table 1. We find that a decaying loss weight of contrastive loss generally works better. Our augmentation details are in Table 2,3, generally following previous works [9, 14]. There are two changes: for downsampled image, we didn't perform cropping, and we remove the Gaussian blur of original image. Because mask vision modeling is equivalent to a strong blurring.

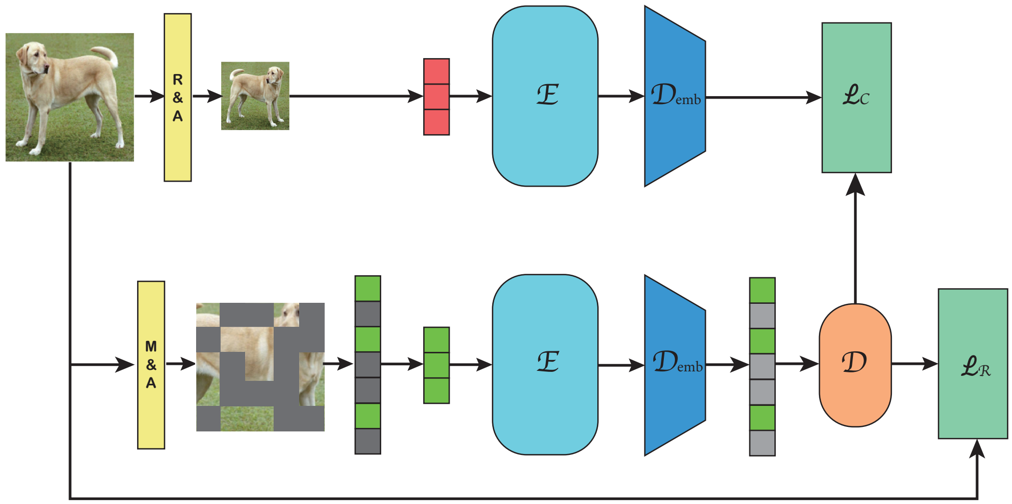

Figure 1. Our model structure, an additional input and contrastive loss are added to the original MAE.

config	value
optimizer	AdamW [22]
batch size	512
learning rate	3e-4
weight decay	0.05
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.5$
learning rate schedule	cosine decay [21]
warmup epochs	40
augmentation	Table 2,3
loss weight λ	$0.24 \times 0.996^{epoch}$

Table 1. Pretraining config.

augmentation	values
resize and crop	scale=(0.2 - 1)
color jitter	strength=(0.4, 0.4, 0.2, 0.1), p=0.8
random grayscale	p=0.2
random horizontal flip	p=0.5

Table 2. Data augmentation for original image.

augmentation	values
resize	scale=(0.5 0.5)
color jitter	strength=(0.4, 0.4, 0.2, 0.1), p=0.8
random grayscale	p=0.2
random horizontal flip	p=0.5
Gaussian blur	strength=(0.1,2), p=0.1
Solarize [14]	p=0.2

Table 3. Data augmentation for downsampled image.

4.2. ImageNet-1K Classification

We use %TODO: ADD model name to carry out self-supervised pretraining on ImageNet-1K training sets [10]. We then finetune our model on supervised classification task with ImageNet-1K. Backbone of both models are ViT-B [13]. Our results are compared with MAE, which is our baseline. Top-1 accuracy with respect to training epochs and relative wall time are given in Table 4. Our model achieved higher accuracy with only about half of training time.

Model	Top1 Acc	Epoch	Wall time
MAE	83.3	1600	2.05×
DAC(ours)	**83.5**	600	**1**×

Table 4. Top 1 Accuracy on ImageNet-1K. 600 epoch of our model has similar wall-time as 800 epoch MAE. Our model is both faster and better than Vanilla MAE.

4.3. ImageNet-mini Classification

We also tested our model on classification with a smaller dataset ImageNet-mini. See Table 5. Both pretrainig and finetuning are on the same dataset. Backbone of both models is ViT-L [13].

Model	Top1 Acc	Epoch	Wall time
MAE	42.5	800	1.08×
DAC(ours)	42.8	550	**1**×
DAC(ours)	**44.8**	800	1.45×

Table 5. Top 1 Accuracy on ImageNet-mini.

DAC finetuning accuracy with epoch is shown in Figure 2. Like MAE, DAC benefits from longer epoch training [15].

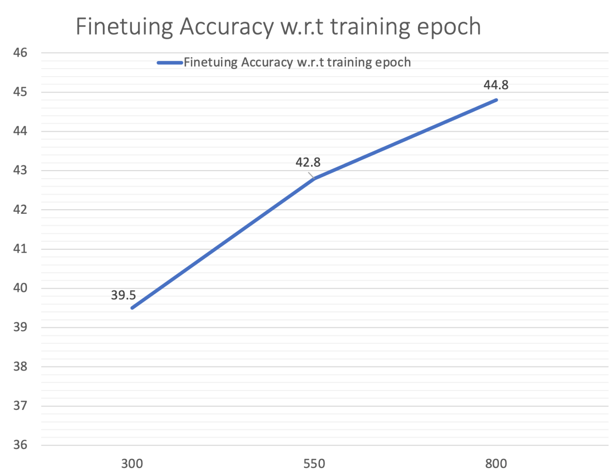

Figure 2. Finetuning accuracy with respect to pretraining epoch.

4.4. Ablation

Ablation studies are done by finetuing a 300 epoch pretrained model on ImageNet-mini.

4.4.1 Data Augmentation

We tested different data augmentation strength during pretraining. Results are in Table 6.

Augmentation	Top1 Acc
resize only	38.9
resize + horizontal flip	39.2
resize + flip + color jitter + grayscale	**39.5**
resize (original image 0.08 - 1)	38.7

Table 6. Ablation on data augmentation on small image. Augmentation on original image is just removing the resize. Value in the bracket indicates resize and crop scale.

Generally, stronger data augmentation can produce better finetuning accuracy. However, the scale of Resize and crop will diminish the performance if the range is set too broad.

4.4.2 Data Augmentation

We ablated the position of small image feature used for contrastive learning. For downsampled image, asymmetric features applies the decoder embedding output, which is the encoder output with an additional linear layer to make shapes match. Original image feature fed into contrastive learning is the decoder output. Symmetric feature apply both decoder outputs to pass through contrastive learning network. The asymmetric input for contrastive learning can not only speed up training but also improve model performance, shown in Table 7.

Augmentation	Top1 Acc
symmetric feature	39.4
asymmetric feature	**39.5**

Table 7. Ablation on the position of small image feature. Asymmetric uses encoder output, while symmetric uses decoder output.

5. Conclusion

In this paper, we proposed a simple but work framework called DAC. By using two novel augmentations, downsampling and masking, we map the original images to two views lies in a lower dimension sample space, which can accelerate the training speed and improve the model's performance. We also use an asymmetric framework and introduce the reconstructive targets to the contrastive objectives. Benefiting from the multitask objectives, we significantly surpasses the previous baselines in training speed. Our DAC is a general idea and can help accelerate the existing contrastive learning, self-knowledge distillation and reconstruction methods. We hope this will inspire future work in related fields.

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[3] Shuhao Cao, Peng Xu, and David A Clifton. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022. 2

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1, 2

[5] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 1, 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2

[7] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 2

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 1

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 1, 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 3

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[12] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 1, 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3, 4

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 1, 2, 3

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 1, 2, 4

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[21] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 3

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[23] Yang Luo, Zhineng Chen, and Xieping Gao. Self-distillation augmented masked autoencoders for histopathological image classification, 2022. 2

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[25] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *CoRR*, abs/2108.08810, 2021. 1

[26] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. 1

[27] Luya Wang, Feng Liang, Yangguang Li, Wanli Ouyang, Honggang Zhang, and Jing Shao. Repre: Improving self-supervised vision transformer with reconstructive pretraining. *arXiv preprint arXiv:2201.06857*, 2022. 1

[28] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature predic-

tion for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021. 2

[29] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*, 2021. 2

[30] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2