Deep Learning for Computer Vision

Self-Supervised Learning

Vineeth N Balasubramanian

Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Unsupervised Learning

Clustering

Group the data into clusters to reveal something meaningful about the data

Dimensionality Reduction

Learn low-dimensional representations of data that are meaningful for a given task

Data Generation

Learn to generate data belonging to a given training distribution

Representation Learning

Learn a distribution that implicitly reveals data representation that helps a downstream task

Unsupervised Learning

Clustering

Group the data into clusters to reveal something meaningful about the data

Dimensionality Reduction

Learn low-dimensional representations of data that are meaningful for a given task

Data Generation

Learn to generate data belonging to a given training distribution

Representation Learning

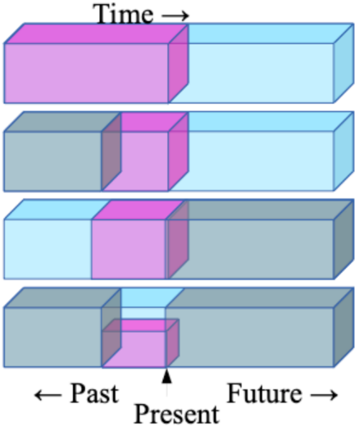
Learn a distribution that implicitly reveals data representation that helps a downstream task
$\rightarrow$ **Self-Supervised Learning**!

What is Self-Supervised Learning?

- Exploit unlabeled data to yield labels
- Design supervised tasks (called **pretext/auxilliary tasks**) that can learn meaningful representations for downstream tasks
- Analogous to filling in the blanks: predict certain part of input from any other part

Self-Supervised Learning

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- **Pretend there is a part of the input you don't know and predict that.**

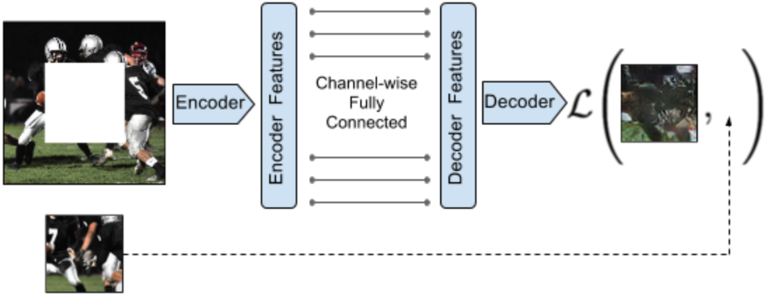


Credit: Yann LeCun

Why Self-Supervised Learning?

- Deep supervised learning works well when labeled data is abundant
- There is a plethora of unlabeled data available; how can we exploit it?
- Humans don't always need supervision to learn, we learn by observation and prediction

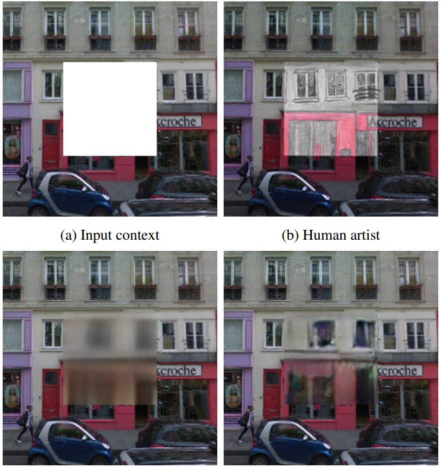
Self-Supervision In Computer Vision: Image Inpainting[1]



- **Context autoencoder** trained to fill in missing parts of an image
- Mask of missing region could be of any shape
- Encoder derived from Alexnet architecture
- Model trained with a combination of L2 loss and adversarial loss

[1] Pathak et al, Context Encoders: Feature Learning by Inpainting, CVPR 2016

Self-Supervision In Computer Vision: Image Inpainting[2]



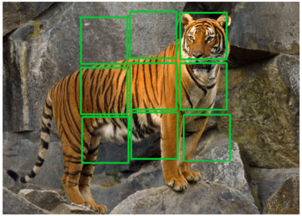
(a) Input context (b) Human artist

(c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

[2]Pathak et al, Context Encoders: Feature Learning by Inpainting, CVPR 2016

Learning Image Representation by Solving Jigsaws[3]

- Used to teach a model that object is made of different parts
- Learns feature mapping of object parts and their spatial arrangement by solving a 9-tiled jigsaw puzzle



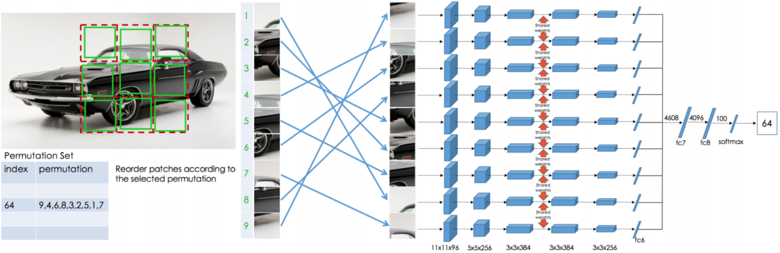
(a)

(b)

[3] Noroozi and Favaro, Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, ECCV 2016

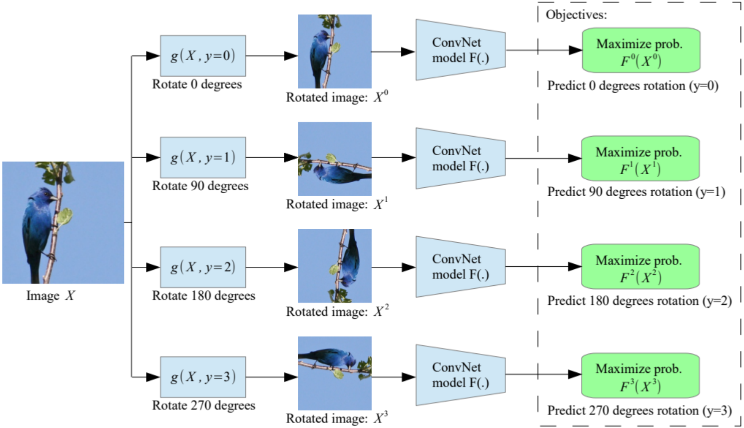
Learning Image Representation by Solving Jigsaws[4]

- 9 tiles shuffled via a randomly chosen permutation from predefined permutation set are fed to network
- Predicts index of permutation applied
- Output vector gives probability of permutation indices used
- Cross entropy loss used for training



[4] Noroozi and Favaro, Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, ECCV 2016

Representation Learning by Predicting Rotations[5]



Learns high level object concepts such as their location in the image, their type, their pose etc.

[5]Gidaris et al, Unsupervised Representation Learning by Predicting Image Rotations, ICLR 2018

Representation Learning by Predicting Rotations[6]

- K rotations are applied, and model outputs a probability distribution over all rotations
- Log loss is used for training
- Loss for an image X is given by:

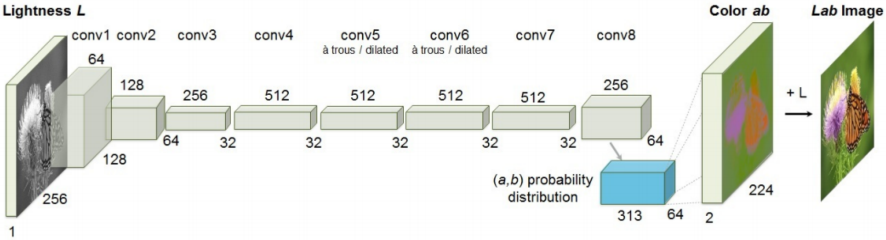
$$\mathcal{L}(X, \theta) = -\frac{1}{K} \Sigma_{y=1}^{K} \log(F(g(X|y)|\theta))$$

$g(\cdot|y)$ is the y^{th} transformation function, F denotes the ConvNet

[6] Gidaris et al, Unsupervised Representation Learning by Predicting Image Rotations, ICLR 2018

Image Colorization[7]

- Predicts color of a grayscale input image in LAB space
- Maps image to a distribution over 313 AB pairs of quantized color value outputs



Cross-entropy loss of predicted probability distribution over binned color values used to train the network

[7]Zhang et al, Colorful Image Colorization, ECCV 2016

Contrastive Learning-Based SSL

- Learns representations by contrasting positive and negative samples; goal is to learn an encoder f such that:

$$score(f(x), f(x^+)) >> score(f(x), f(x^-))$$

x^+ obtained from same image as x and x^- dissimilar to x; scores given by cosine similarity

Contrastive Learning-Based SSL

- Learns representations by contrasting positive and negative samples; goal is to learn an encoder f such that:

$$score(f(x), f(x^+)) >> score(f(x), f(x^-))$$

  x^+ obtained from same image as x and x^- dissimilar to x; scores given by cosine similarity

- Softmax classifier used to classify positive and negative samples correctly

Contrastive Learning-Based SSL

- Learns representations by contrasting positive and negative samples; goal is to learn an encoder f such that:

$$score(f(x), f(x^+)) >> score(f(x), f(x^-))$$

  x^+ obtained from same image as x and x^- dissimilar to x; scores given by cosine similarity

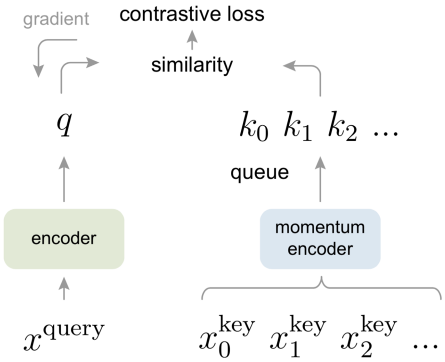
- Softmax classifier used to classify positive and negative samples correctly

- General form of loss function given by:

$$\mathcal{L} = -\mathbb{E}\left[\log \frac{\exp(score(f(x), f(x^+))/\tau)}{\exp(score(f(x), f(x^+))/\tau) + \Sigma_{j=1}^{N-1} \exp(score(f(x), f(x_j^-))/\tau)}\right]$$

  where τ is temperature hyperparameter

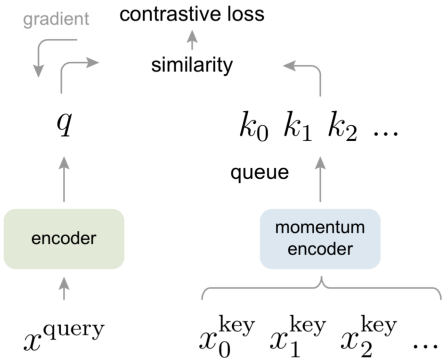
MoCO: Momentum Contrast[8]



- Proposes unsupervised learning of visual representations as a **dynamic dictionary look-up**

[8]He et al, Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020

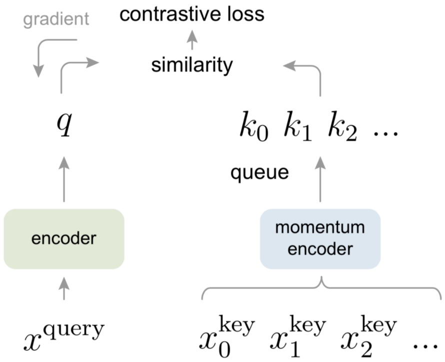
MoCO: Momentum Contrast[8]



- Proposes unsupervised learning of visual representations as a **dynamic dictionary look-up**
- Dictionary structured as a large FIFO queue of encoded representations of data samples

[8] He et al, Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020

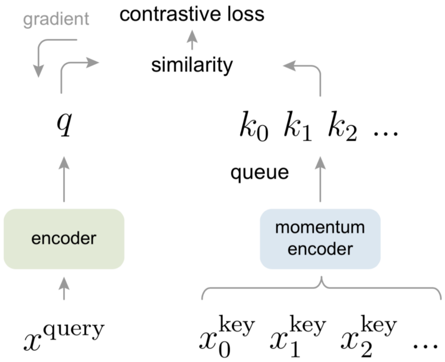
MoCO: Momentum Contrast[8]



- Proposes unsupervised learning of visual representations as a **dynamic dictionary look-up**
- Dictionary structured as a large FIFO queue of encoded representations of data samples
- Given query sample x_q, query representation obtained using an encoder $q = f_q(x_q)$

[8] He et al, Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020

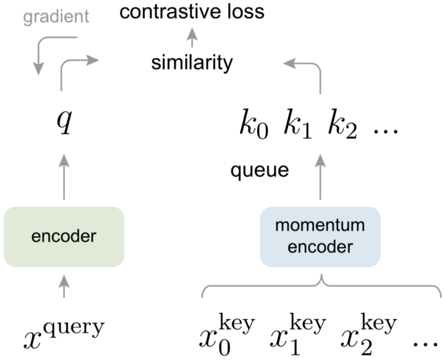
MoCO: Momentum Contrast[8]



- Proposes unsupervised learning of visual representations as a **dynamic dictionary look-up**
- Dictionary structured as a large FIFO queue of encoded representations of data samples
- Given query sample x_q, query representation obtained using an encoder $q = f_q(x_q)$
- Key samples encoded by a momentum encoder $k_i = f_k(x_{k_i})$ gives a set of key representations: $\{k_1, k_2, \cdots\}$ in dictionary

[8]He et al, Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020

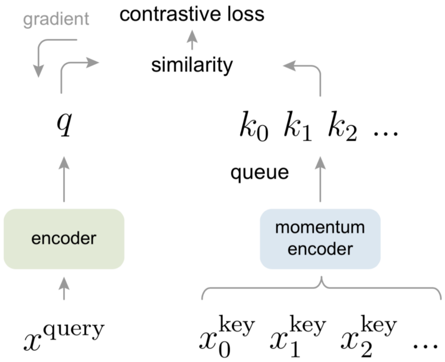
MoCO: Momentum Contrast[8]



- Proposes unsupervised learning of visual representations as a **dynamic dictionary look-up**
- Dictionary structured as a large FIFO queue of encoded representations of data samples
- Given query sample x_q, query representation obtained using an encoder $q = f_q(x_q)$
- Key samples encoded by a momentum encoder $k_i = f_k(x_{k_i})$ gives a set of key representations: $\{k_1, k_2, \cdots\}$ in dictionary
- Positive key k^+ in dictionary created using copy of x_q with different augmentation

[8] He et al, Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020

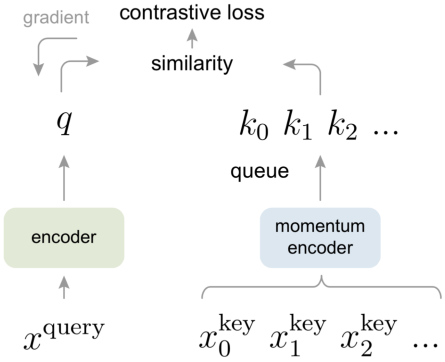
MoCO: Momentum Contrast[8]



- Proposes unsupervised learning of visual representations as a **dynamic dictionary look-up**
- Dictionary structured as a large FIFO queue of encoded representations of data samples
- Given query sample x_q, query representation obtained using an encoder $q = f_q(x_q)$
- Key samples encoded by a momentum encoder $k_i = f_k(x_{k_i})$ gives a set of key representations: $\{k_1, k_2, \cdots\}$ in dictionary
- Positive key k^+ in dictionary created using copy of x_q with different augmentation
- Loss on previous slide (contrastive loss) used to learn

[8]He et al, Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020

MoCO: Momentum Contrast[9]

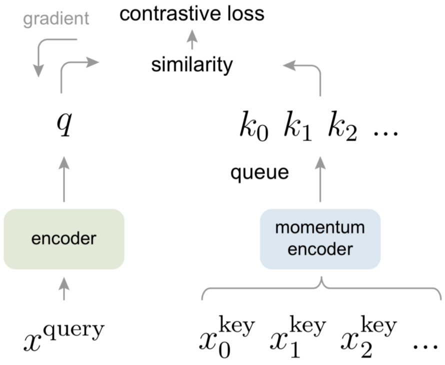


Momentum Contrast

- Query and key encoders both updated based on loss
- Maintains dictionary as queue of data samples
- Allows reuse of encoded keys from immediate preceding mini-batches, decouples dictionary size from batch size
- **Momentum-based update** proposed to keep keys approximately consistent

[9] He et al, Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020

MoCO: Momentum Contrast[10]

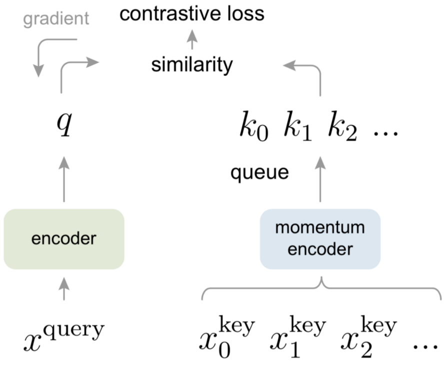


Momentum Contrast

- Using queue as dictionary makes it difficult to update key encoder
- Can we just copy the key encoder from the query encoder?

[10] He et al, Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020

MoCO: Momentum Contrast[10]



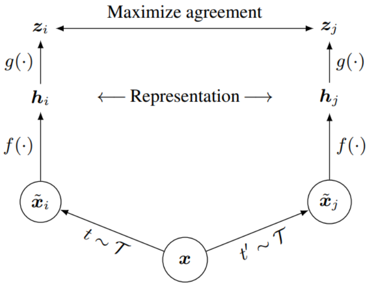
Momentum Contrast

- Using queue as dictionary makes it difficult to update key encoder
- Can we just copy the key encoder from the query encoder? (No! Representation will not be consistent because of rapidly changing query encoder)
- Query encoder (f_q) is updated using backpropagation and key encoder (f_k) is updated using momentum as:

$$\theta_k = m\theta_k + (1-m)\theta_q$$

[10] He et al, Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020

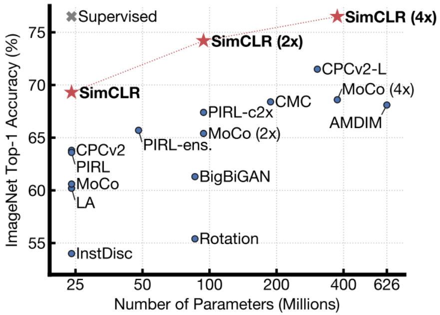
SimCLR: A Simple Framework for Contrastive Learning of Visual Representations[11]



- Learns via maximizing agreement between differently augmented views of same data example in latent space
- Given n images, $2n$ samples obtained by 2 different augmentations. Given one positive pair, there exist $2(n-1)$ negative pairs
- Loss operates on top of an extra projection of the representation via $g(.)$

[11] Chen et al, A Simple Framework for Contrastive Learning of Visual Representations, ICML 2020

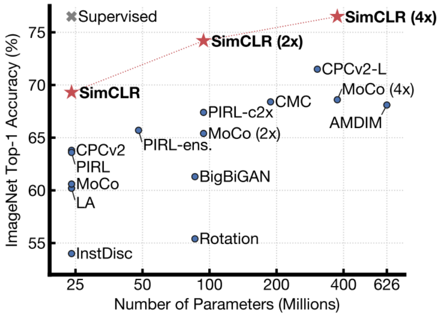
SimCLR vs MoCo[12]



- **SimCLR Advantages:** Strong data augmentation techniques, MLP projection over the representations

[12] Chen et al, Improved Baselines with Momentum Contrastive Learning, arXiv 2020

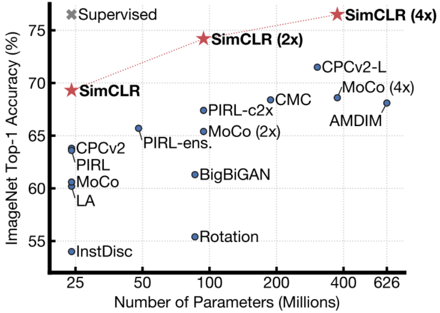
SimCLR vs MoCo[12]



- **SimCLR Advantages:** Strong data augmentation techniques, MLP projection over the representations
- **SimCLR Disadvantages:** Number of negative samples is limited by the batch size

[12] Chen et al, Improved Baselines with Momentum Contrastive Learning, arXiv 2020

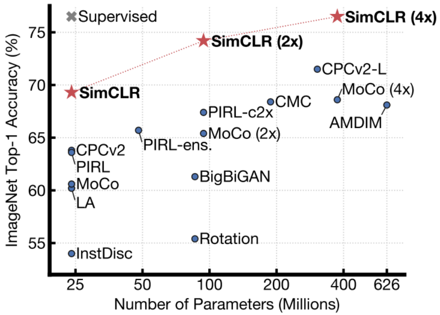
SimCLR vs MoCo[12]



- **SimCLR Advantages:** Strong data augmentation techniques, MLP projection over the representations
- **SimCLR Disadvantages:** Number of negative samples is limited by the batch size
- **MoCo Advantage:** Decouples the batch size from the number of negatives

[12] Chen et al, Improved Baselines with Momentum Contrastive Learning, arXiv 2020

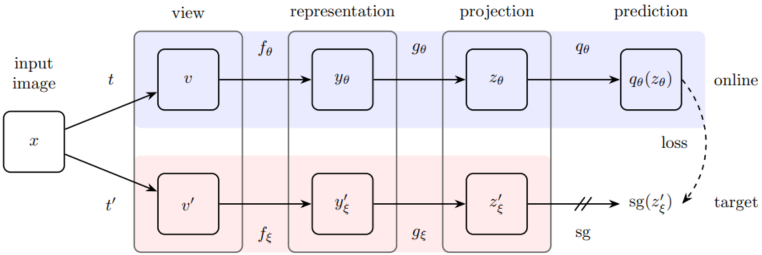
SimCLR vs MoCo[12]



- **SimCLR Advantages:** Strong data augmentation techniques, MLP projection over the representations
- **SimCLR Disadvantages:** Number of negative samples is limited by the batch size
- **MoCo Advantage:** Decouples the batch size from the number of negatives
- Chen et al combined advantages from these two methods in MoCoV2

[12]Chen et al, Improved Baselines with Momentum Contrastive Learning, arXiv 2020

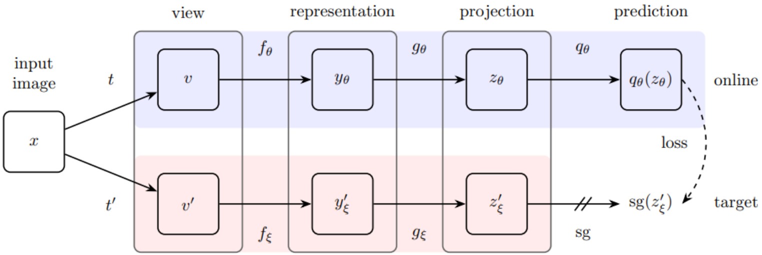
Bootstrap your Own Latent (BYOL)[13]



- Claims to achieve state-of-the-art results without dependency on negative samples

[13] Grill et al, Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, arXiv 2020

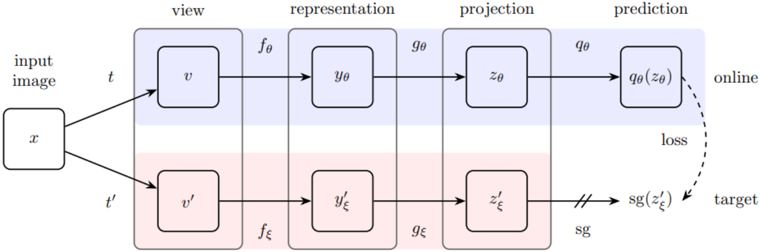
Bootstrap your Own Latent (BYOL)[13]



- Claims to achieve state-of-the-art results without dependency on negative samples
- Bootstraps outputs of a network to serve as targets

[13]Grill et al, Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, arXiv 2020

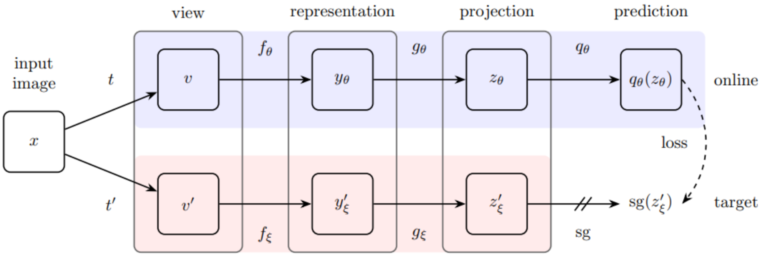
Bootstrap your Own Latent (BYOL)[13]



- Claims to achieve state-of-the-art results without dependency on negative samples
- Bootstraps outputs of a network to serve as targets
- Two networks: **online** and **target**, interact and learn from each other

[13]Grill et al, Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, arXiv 2020

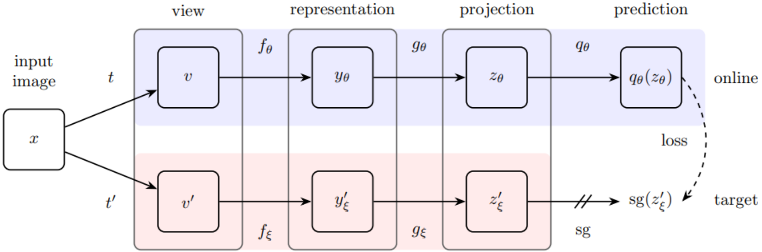
Bootstrap your Own Latent (BYOL)[13]



- Claims to achieve state-of-the-art results without dependency on negative samples
- Bootstraps outputs of a network to serve as targets
- Two networks: **online** and **target**, interact and learn from each other
- Online network predicts target network's representation of another augmented view of same image

[13] Grill et al, Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, arXiv 2020

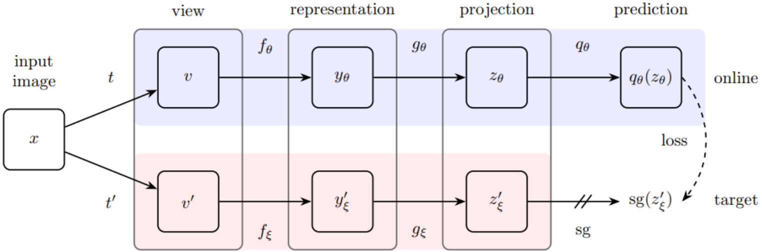
Bootstrap your Own Latent (BYOL)[14]



- $\mathcal{L}_\theta^{BYOL} = \|\bar{q}_\theta(z_\theta) - \bar{z}'_\xi\|_2^2$

[14]Grill et al, Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, arXiv 2020

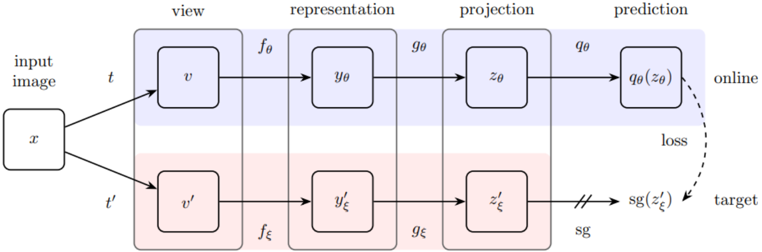
Bootstrap your Own Latent (BYOL)[14]



- $\mathcal{L}_\theta^{BYOL} = \|\bar{q}_\theta(z_\theta) - \bar{z}'_\xi\|_2^2$
- $\bar{q}_\theta(z_\theta)$ and $\bar{z}'_\xi$ are L_2-normalized

[14] Grill et al, Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, arXiv 2020

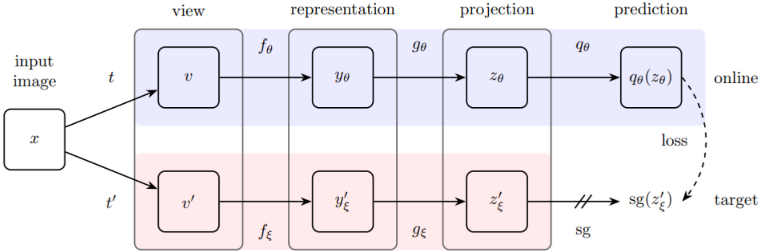
Bootstrap your Own Latent (BYOL)[14]



- $\mathcal{L}_\theta^{BYOL} = \|\bar{q}_\theta(z_\theta) - \bar{z}'_\xi\|_2^2$
- $\bar{q}_\theta(z_\theta)$ and $\bar{z}'_\xi$ are L_2-normalized
- $\bar{\mathcal{L}}_\theta^{BYOL}$ obtained by switching v' and v

[14] Grill et al, Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, arXiv 2020

Bootstrap your Own Latent (BYOL)[14]



- $\mathcal{L}_\theta^{BYOL} = \|\bar{q}_\theta(z_\theta) - \bar{z}'_\xi\|_2^2$
- $\bar{q}_\theta(z_\theta)$ and $\bar{z}'_\xi$ are L_2-normalized
- $\bar{\mathcal{L}}_\theta^{BYOL}$ obtained by switching v' and v
- **Final Loss:** $\mathcal{L}_{final} = \mathcal{L}_\theta^{BYOL} + \bar{\mathcal{L}}_\theta^{BYOL}$

[14] Grill et al, Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, arXiv 2020

Homework

Readings

- Lilian Weng, Self-Supervised Representation Learning

References I

- Mehdi Noroozi and Paolo Favaro. “Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles”. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. 2016.
- Deepak Pathak et al. “Context Encoders: Feature Learning by Inpainting”. In: 2016.
- Richard Zhang, Phillip Isola, and Alexei A Efros. “Colorful Image Colorization”. In: *ECCV*. 2016.
- Spyros Gidaris, Praveer Singh, and Nikos Komodakis. “Unsupervised Representation Learning by Predicting Image Rotations”. In: *International Conference on Learning Representations*. 2018.
- Kaiming He et al. “Momentum Contrast for Unsupervised Visual Representation Learning”. In: *arXiv preprint arXiv:1911.05722* (2019).
- Lilian Weng. “Self-Supervised Representation Learning”. In: *lilianweng.github.io/lil-log* (2019).
- Ankesh Anand. *Contrastive Self-Supervised Learning*. 2020.

References II

- Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *arXiv preprint arXiv:2002.05709* (2020).
- Jean-Bastien Grill et al. *Bootstrap your own latent: A new approach to self-supervised Learning*. 2020. arXiv: 2006.07733 [cs.LG].
- Jeremy Howard. *Self-supervised learning and computer vision*. 2020.