Simple Summary of 23Summer Study

Shengjie Niu

Department of Statistics and Data Science, SUSTech

August 10, 2023

1. Deep Metric Learning
2. Two Mainstreams
3. Learning Without Negative Pairs
4. Masked Image Model

1 Deep Metric Learning

2 Two Mainstreams

3 Learning Without Negative Pairs

4 Masked Image Model

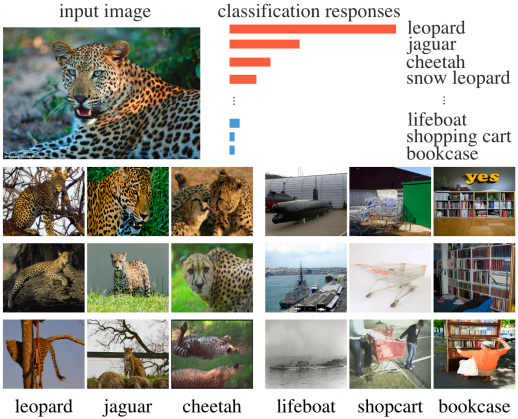
Deep Metric Learning

Learning a distance function over objects.

Given some input pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and their similarities, y_{ij}. Task is to learn a function d which conform the similarities.

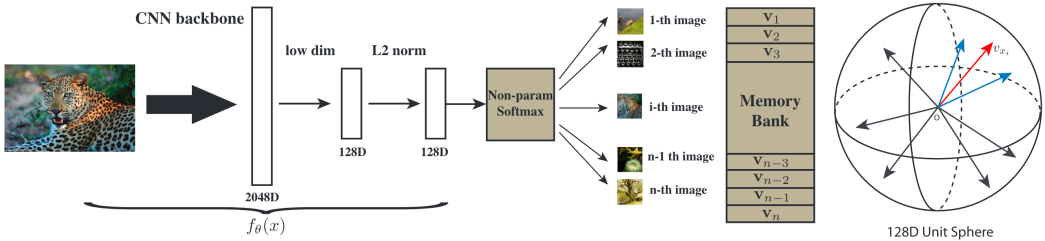
Reflection in Deep Learning

Example: Face verification



Source: Z Wu et al, 1805

InstDisc - 1805

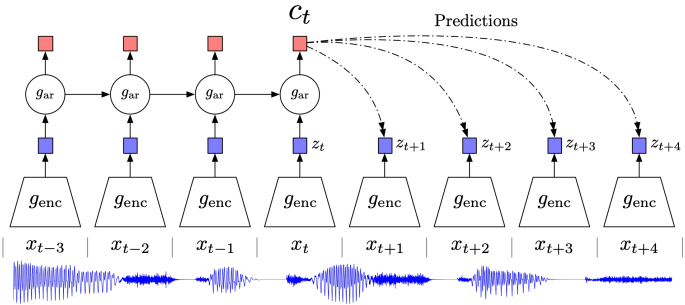


Source: Z Wu et al, 1805

New pretext task: Instance Discrimination Prior to MoCo

Non-Parametric SoftMax Classifier

CPC - 1807



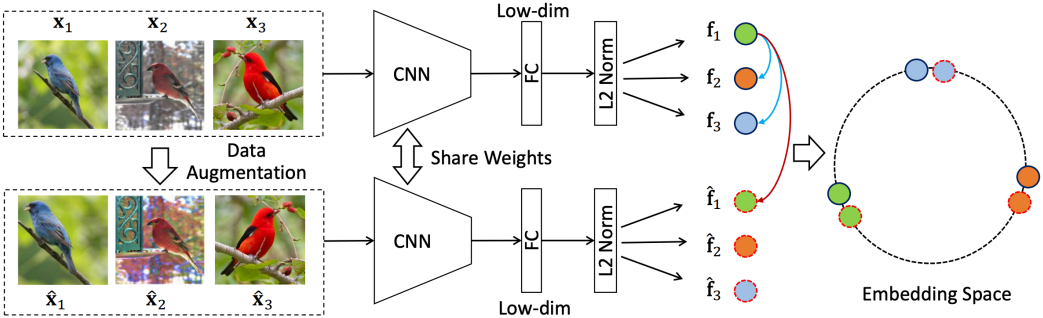
Source: A Oord et al, 1807

Prediction pretext task

Great applicability

Models: Encoder and autoregressive model

InvarSpread - 1904



Source: M Ye et al, 1904

Instance Discrimination

End-to-End without Memory Bank

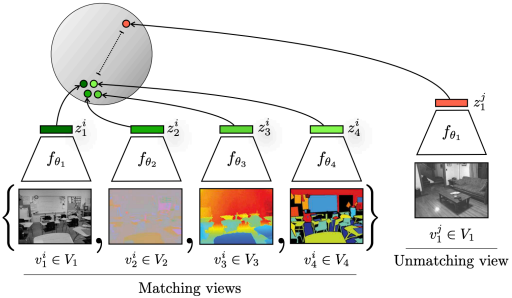
Only an encoder Prior to SimCLR

CMC - 1906

Humans view the world through many sensory channels. Each view is noisy and incomplete, but important factors are shared between all views.
Ex: a dog can be seen, heard, and felt.

Hope to find a powerful representation is one that **models view-invariant factors**.

Subsequent works: Clip (OpenAI), Knowledge Distillation, Muse.



Source: Y Tian et al, 1906
Different encoders to sense different signals.

Conclusion to Stage 1

Versatile Developments

- Different Pretext tasks
  - instance-level discrimination
  - Prediction
- Different Objective Function
  - NCE loss
  - Mutual Information
  - InfoNCE loss
- Different views in Positive/Negative Pairs - defined by pretext task

1 Deep Metric Learning

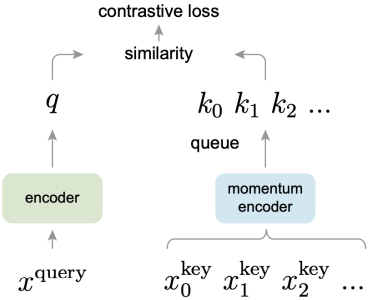
2 Two Mainstreams

3 Learning Without Negative Pairs

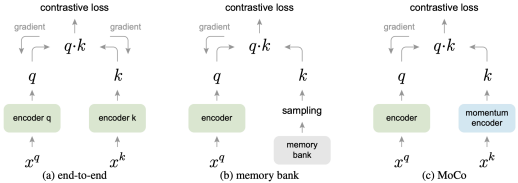
4 Masked Image Model

MoCo - 1911

Contrastive Learning can be viewed as dictionary lookup.

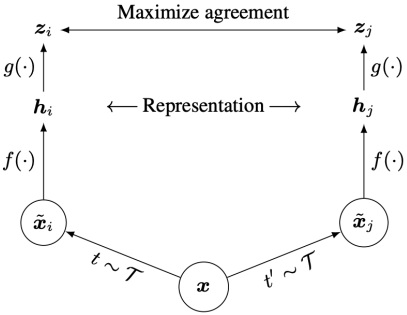


Source: K He et al, 1911



A desirable dictionaries should be
1. Large and 2. Consistent during training
Contributions: Dictionary and Momentum encoder

SimCLR - 2002



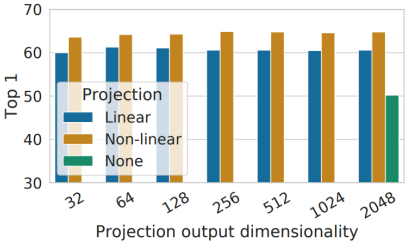
Source: T Chen et all, 2002

Contributions

More augmentations

Larger batch size and longer training time

Learnable nonlinear transformation



MoCo v2 - 2003 & SimCLR v2 - 2006

Updates

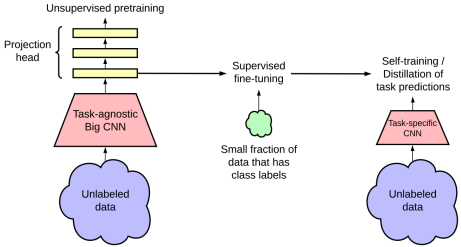
- More augmentations
- Project head

case	unsup. pre-train MLP	aug+	cos	epochs	batch	ImageNet acc.
MoCo v1 [6]				200	256	60.6
SimCLR [2]	✓	✓	✓	200	256	61.9
SimCLR [2]	✓	✓	✓	200	8192	66.6
MoCo v2	✓	✓	✓	200	256	**67.5**
*results of **longer** unsupervised training follow:*						
SimCLR [2]	✓	✓	✓	1000	4096	69.3
MoCo v2	✓	✓	✓	800	256	**71.1**

Source: X Chen et all, 2003

Updates

- Deeper ResNet
- Two hidden layer MLP
- Momentum encoder

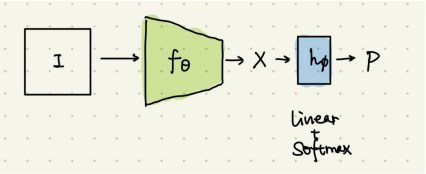


Source: T Chen et all, 2006

SeLa - 1911 & SwAV - 2006

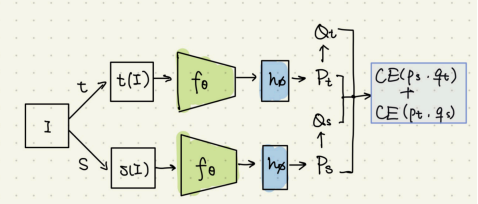
IDEA: Get an analytics solution of pseudo-label and use p to predict the pseudo-label.

q should be K-dimension determined by # of Clusters.



Ref: Y Asano et all, 1911

IDEA: Applied constrastive concept into the SeLa.



Ref: M Caron et all, 2006

Clustering SSL is similar to constrative loss

MoCo: 6w Negative v.s. 3k clusterings

Supervised: 76.5% v.s. SwAV: 75.3%

Trick: Mult-crop

Conclusion to Stage 2

Close to Unified Architecture

- Objective: InfoNCE-related
- Similar Architecture
  - Project head
  - Momentum encoder
  - Deeper backbone

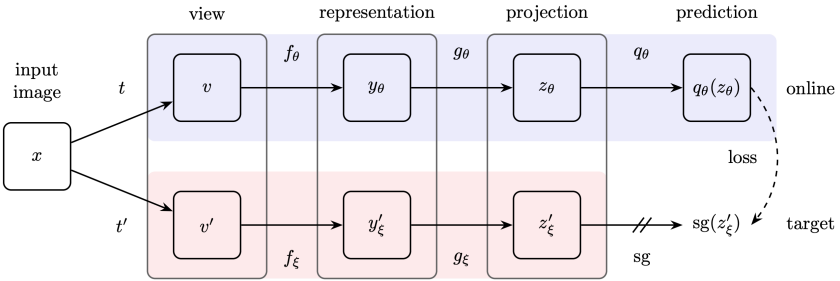
1 Deep Metric Learning

2 Two Mainstreams

3 Learning Without Negative Pairs

4 Masked Image Model

BYOL - 2006

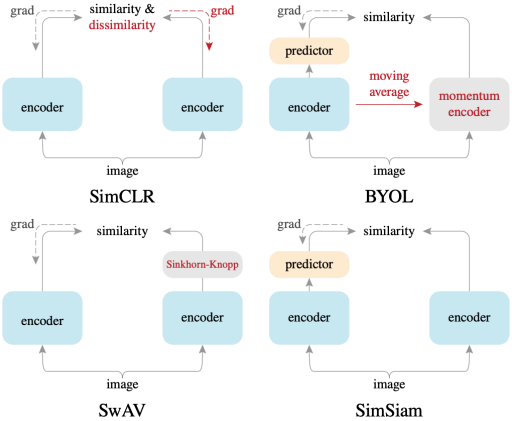


Source: J Grill et all, 2006

Combination of MoCo, SimCLR and SwAV.

Appply the MSE loss instead of InfoNCE.

SimSiam - 2011



Source: X Chen et all, 2011

IDEA: Applied constrastive concept into the SeLa.

method	batch size	negative pairs	momentum encoder	100 ep	200 ep	400 ep	800 ep
SimCLR (repro.+)	4096	✓		66.5	68.3	69.8	70.4
MoCo v2 (repro.+)	**256**	✓	✓	67.4	69.9	71.0	72.2
BYOL (repro.)	4096		✓	66.5	**70.6**	**73.2**	**74.3**
SwAV (repro.+)	4096			66.5	69.1	70.7	71.8
SimSiam	**256**			**68.1**	70.0	70.8	71.3

Why not collapse: stop gradient.

MoCo v3 - 2104 & DINO - 2104

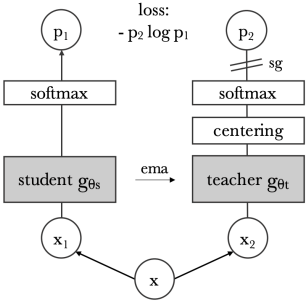
ViT bring a drastic technological changes, replace CNN with ViT

Algorithm 1 MoCo v3: PyTorch-like Pseudocode

```
# f_q: encoder: backbone + proj mlp + pred mlp
# f_k: momentum encoder: backbone + proj mlp
# m: momentum coefficient
# tau: temperature

for x in loader: # load a minibatch x with N samples
   x1, x2 = aug(x), aug(x) # augmentation
   q1, q2 = f_q(x1), f_q(x2) # queries: [N, C] each
   k1, k2 = f_k(x1), f_k(x2) # keys: [N, C] each

   loss = ctr(q1, k2) + ctr(q2, k1) # symmetrized
   loss.backward()

   update(f_q) # optimizer update: f_q
   f_k = m*f_k + (1-m)*f_q # momentum update: f_k

# contrastive loss
def ctr(q, k):
   logits = mm(q, k.t()) # [N, N] pairs
   labels = range(N) # positives are in diagonal
   loss = CrossEntropyLoss(logits/tau, labels)
   return 2 * tau * loss
```

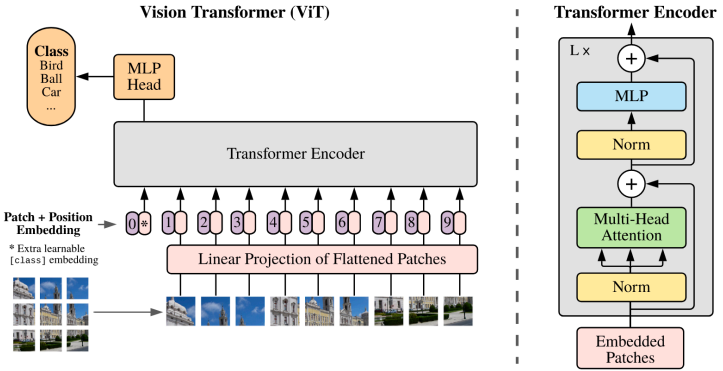
Notes: `mm` is matrix multiplication. `k.t()` is `k`'s transpose. The prediction head is excluded from `f_k` (and thus the momentum update).

Source: M Caron et all, 2006



Source: M Caron et all, 2104

ViT - 2010


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0* 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Encoder
L ×
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

1 Deep Metric Learning

2 Two Mainstreams

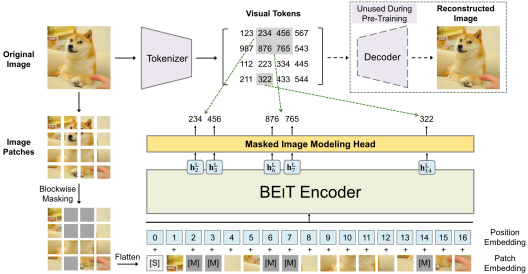
3 Learning Without Negative Pairs

4 Masked Image Model

BEiT - 2106

Random masking some percentage of image patches, and then predict the discrete visual tokens corresponding to the masked patches.

Image patch & Visual token



Source: H Bao et al, 2106

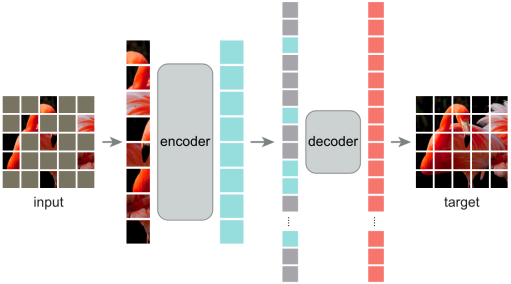
MAE - 2111

Random masking with a **high mask ratio**.

Encoder operate only on a small subset of all set.

Decoder: Full set of tokens consisting of (i) encoded visible patches, and (ii) mask tokens. Each mask token is a shared, learned vector.

Reconstruction target: MSE between the reconstructed and original images only on masked patches.

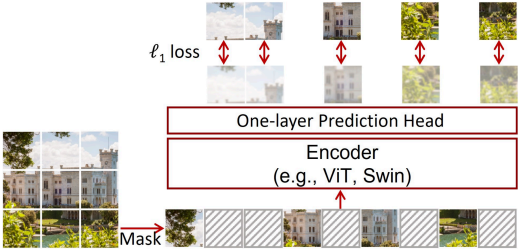


Source: K He et al, 2111

SimMIM - 2111

Observations

- Masking strategy: Random masking with relatively large patch size (eg. 32 $\times$ 32).
- Prediction heads: An extremely lightweight prediction head (linear layer).
- Prediction tragets: A simple raw pixel regression task.
- Encoder architecures: ViT, Swin and ResNet could all benefit from SimMIM.



Source: Z Xie et al, 2111