Friedrich-Alexander-Universität
Technische Fakultät

Advanced Deep Learning

Representation Learning

K. Breininger, V. Christlein

Artificial Intelligence in Medical Imaging + Pattern Recognition Lab,

Friedrich-Alexander-Universität Erlangen-Nürnberg SoSe 2023

1. Introduction & Motivation
2. Deep Metric Learning
3. Contrastive and Non-contrastive Learning

1. Introduction & Motivation

2. Deep Metric Learning

3. Contrastive and Non-contrastive Learning

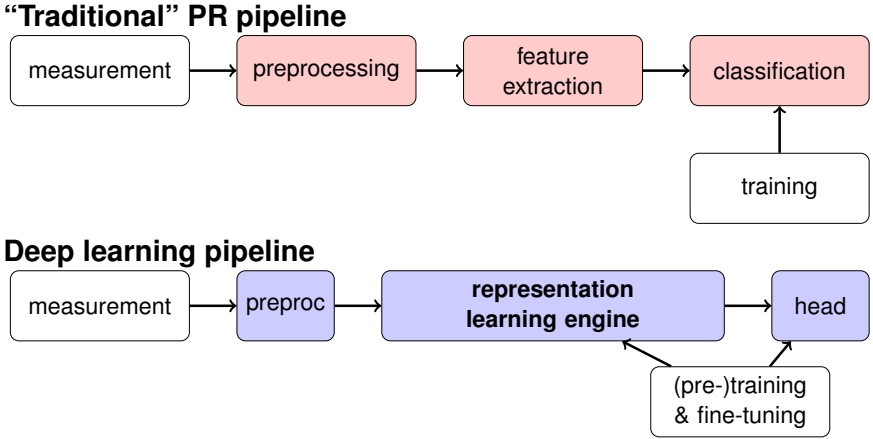
What to do with all this data?

- Supervised learning is ~~data~~ label hungry
- Large amount of “unlabeled” data available

5B-Image-Question:
How can we utilize this data independent of the original task?

Sources:
ImageNet adapted from http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/
LAION: https://laion.ai/blog/laion-5b/

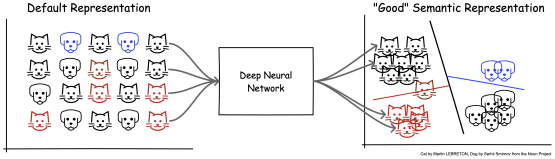
Recap: Pattern Recognition Pipeline



Recap: CNNs & Representation

- Pixel intensities are **bad representations** – for ML
- Approach:
  - Train without labels to find “better” representations
  - Solve downstream task “easily” - (little) labeled data

- Core questions:
  - What is the objective & loss function?
    ➔ must be derived from the data itself
  - What is “better” & “easy”?



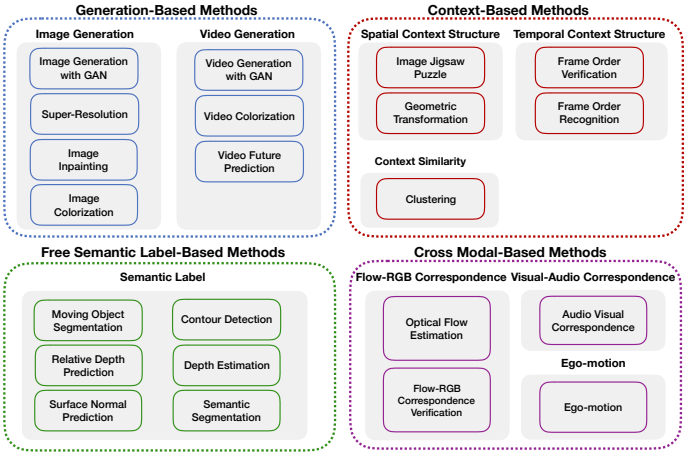
Sources:
https://www.shelma.eu/orig-image/82fc9d0584.jpg, https://blog.fastforwardlabs.com/2020/11/15/representation-learning-101-for-software-engineers.html

Goals for Today

You will be able to ...

- renew your understanding of “deep representation learning”
- explain the underlying concepts of metric and contrastive learning
- discuss the connection between contrastive, non-contrastive and self-supervised learning
- understand the development from contrastive loss to advanced SSL techniques
- discuss current benchmarks and critically comment on strengths / weaknesses of approaches
- **learn some approaches that you can use in this year’s ADL challenge!**

Representation Learning



Source: [1]

1. Introduction & Motivation

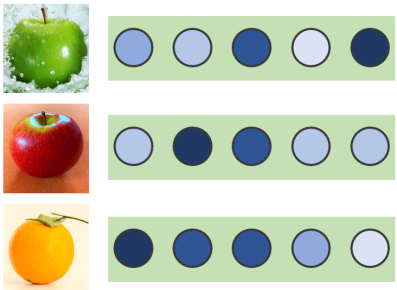
2. Deep Metric Learning

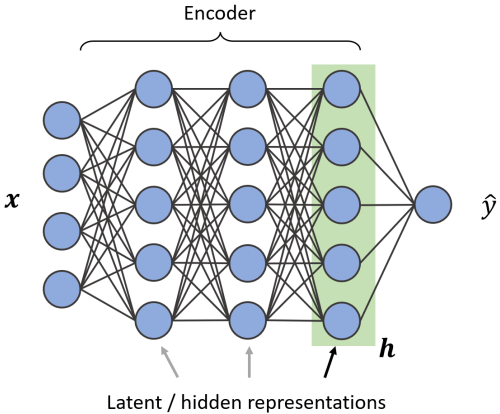
3. Contrastive and Non-contrastive Learning

Core Idea of Contrastive Learning

Supervised learning

- Update weights to minimize $\mathcal{L}(y, \hat{y})$
- Layers represent transformation chain

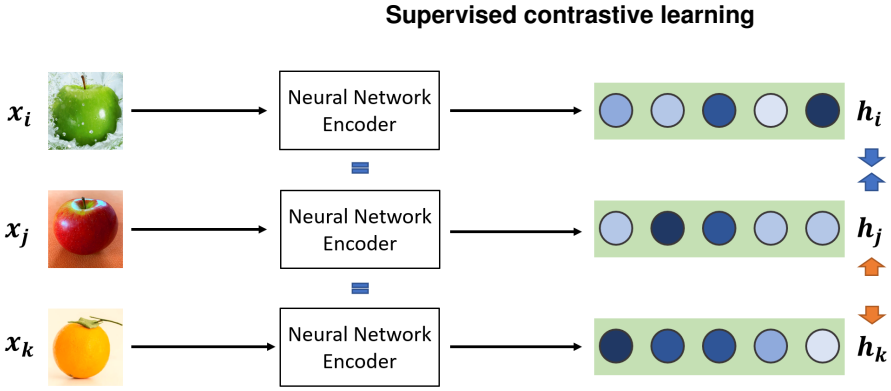




➔ **Implicit** representation learning

Core Idea of Contrastive Learning

(cont.)

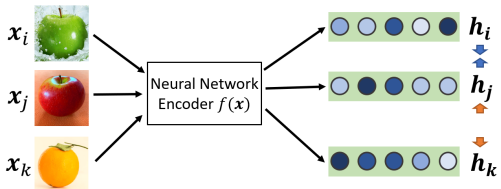


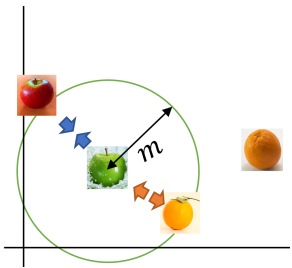
Contrastive Loss [2]

Goal:
Draw pairs and ...
... **maximize** similarity of similar instances,
... **minimize** similarity of unrelated instances.

Contrastive Loss:

$$L(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{1}_{y_i=y_j} \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2^2 + \mathbf{1}_{y_i \neq y_j} \max(0, m - \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2^2)$$





Progressing from Contrastive Loss

Pairs with ...

an **anchor** and **either**

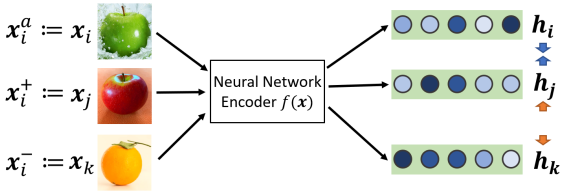
a positive sample **or**

a negative sample

Triplet Loss [3]

Similar goal as before:
Maximize similarity of similar instances,
minimize similarity of unrelated instances.

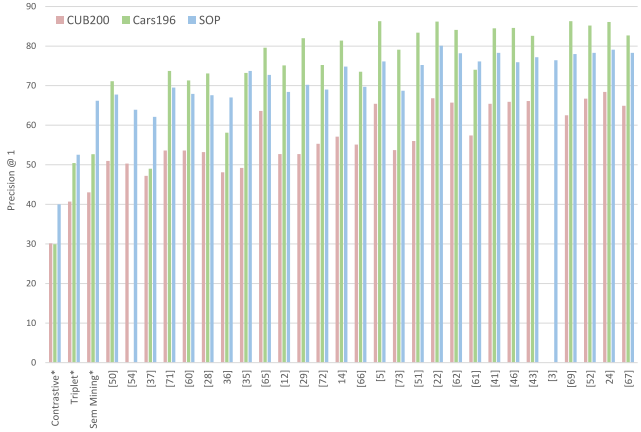
Idea:
For anchor $\mathbf{x}_i^a := \mathbf{x}_i$, use positive and negative samples



$$L(\mathbf{x}_i^a, \mathbf{x}_i^+, \mathbf{x}_i^-) = \max(0, m + \|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^+)\|_2^2 - \|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^-)\|_2^2)$$

➔ Additional trick: focus on “hard positive” and “hard negative” samples

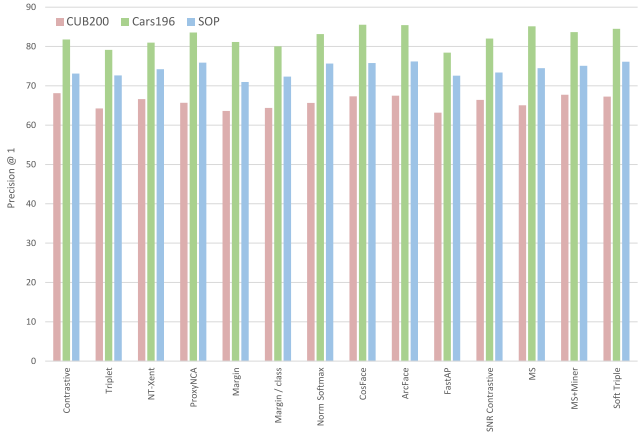
Metric Learning Reality Check [4]



Results acc. to papers

Source: Musgrave et al. 2020 [4]

Metric Learning Reality Check [4]

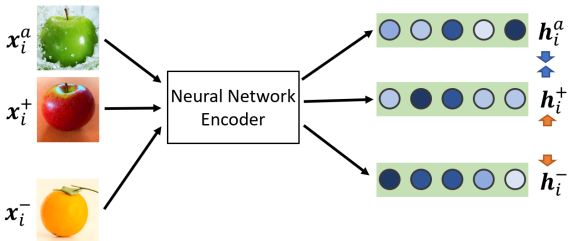


Results acc. to reality

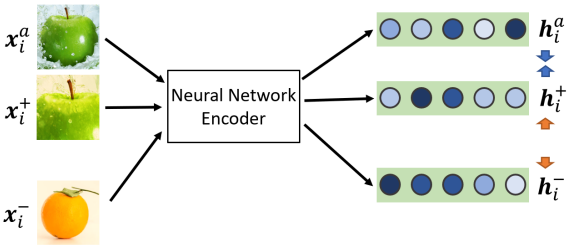
Source: Musgrave et al. 2020 [4]

From Supervised to Self-supervised

Supervised contrastive learning:



Unsupervised contrastive learning:



From Supervised to Self-supervised

(cont.)

Deep Metric Learning		Contrastive SSL
positive/negative pairs come from labels or fixed transforms, e.g., two halves of an image	→	positive pairs come from designed DAs that are continuously sampled, negative pairs: all non-positive pairs regardless of class membership
hard-negative sampling for each mini-batch	→	random sampling
encoder DN	→	encoder DN + projector MLP
small dataset (N<200k)	→	large dataset
zero-shot k-NN validation	→	zero-shot k-NN validation zero/few-shot/fine-tuning linear probing

Source: Balestrerio et al. 2023 [5]

What is a ‘good’ representation?

Evaluation protocols

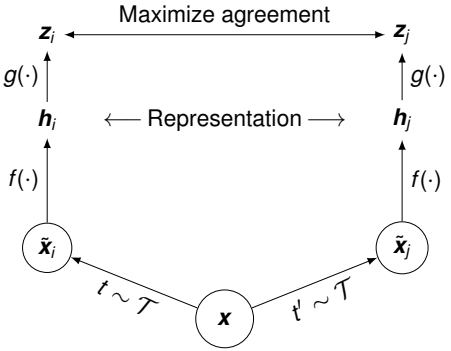
- KNN of embeddings
- Head / linear layer fine-tuning
- Full fine-tuning (re-introduced by MAE [6])
- Problem: (full) fine-tuning often not possible anymore
- Sidenote: Unsupervised alternative for hyper-parameter optimization: RankMe [7]

1. Introduction & Motivation

2. Deep Metric Learning

3. Contrastive and Non-contrastive Learning

Contrastive Learning: SimCLR

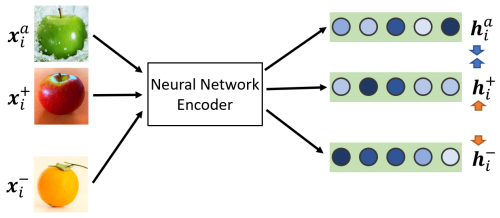


- Mini-batch of n samples
- → $2n$ augmented samples
- One positive pair, $2(n-1)$ negatives
- Contrastive loss (normalized temperature-scaled cross-entropy loss NT-Xent, a variation of InfoNCE):

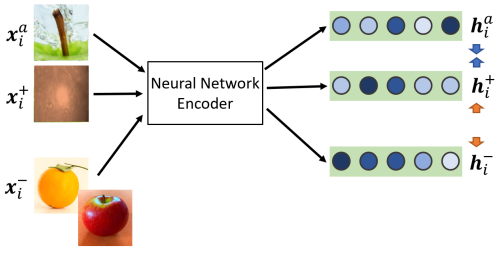
$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(s\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k=1}^{2n} \mathbf{1}_{[k \neq i]} \exp\left(s\left(\mathbf{z}_i, \mathbf{z}_k\right)/\tau\right)}$$

Challenges for Contrastive Learning

Supervised contrastive learning:

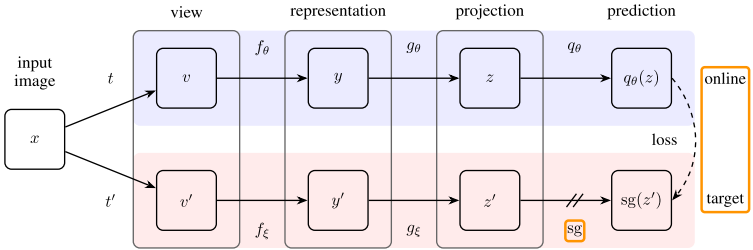


Unsupervised contrastive learning:



- Instance discrimination
- Selection of augmentations
- Global vs. local / whole vs. parts

BYOL [8]



- No negative pair/no contrastive loss
- Two networks: **online** and **target** network → interact and learn from each other
- In theory: trivial solution possible (e.g. zero for all images)

→ use slow-moving average of the online network as target network

- Loss: MSE of ℓ^2-normalized predictions (proportional to cosine distance)
- – Needs large batch sizes

Source: Grill et al. 2020 [8]

SimSiam [9]

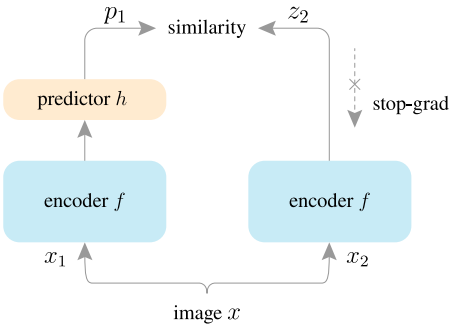
- Shared encoder f
- Prediction MLP head h (3 FC layers w. 2048 units each and BN)

→ Transforms output of one view and matches it to the other view: $p_1 = h(f(x_1)), z_2 = f(x_2)$

- Loss:

$$\mathcal{L} = \frac{1}{2}\mathcal{D}\left(p_1, \text{stopgrad}(z_2)\right) + \frac{1}{2}\mathcal{D}\left(p_2, \text{stopgrad}(z_1)\right)$$

with cosine similarity: $\mathcal{D}\left(p_1, z_2\right) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}$

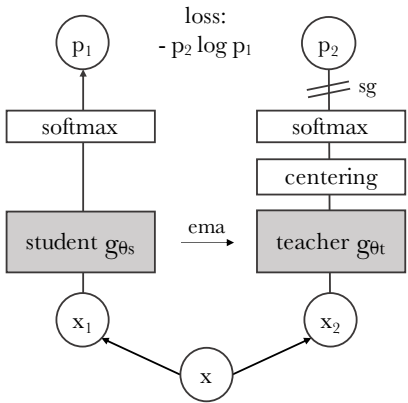


Source: Chen & He 2021 [9]

DINO [10]

- 2-branch network similar to BYOL (teacher,student)
- Teacher output is centered w. batch mean
- Temperature softmax is applied: $\frac{\exp\left(g_{\theta_s}(x)^{(i)}/\tau_s\right)}{\sum_{k=1}^{K}\exp\left(g_{\theta_s}(x)^{(k)}/\tau_s\right)}$
- Gradient only through student network
- Teacher parameters updated w. exponential moving average (EMA) of student
- Cross-entropy between views V (2 global views x_1^g, x_2^g and several local views 96×96), all views pass through student, teacher gets only global views

$$\min_{\theta_s} \sum_{x \in \left\{x_1^g, x_2^g\right\}} \sum_{\substack{x' \in V \\ x' \neq x}} H\left(P_t(x), P_s\left(x'\right)\right)$$

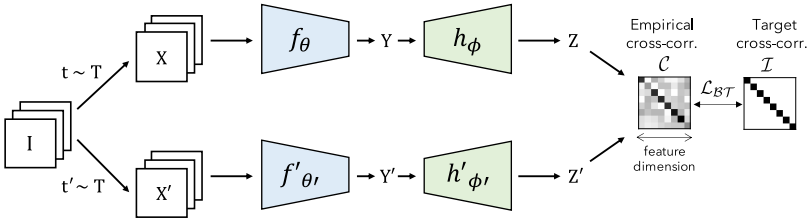


Source: Caron et al. 2021 [10]

DINO [10]

Source: Caron et al. 2021 [10]

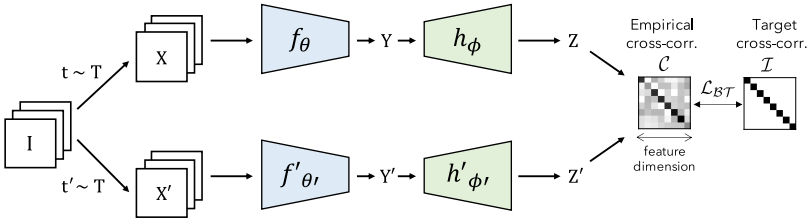
Barlow Twins [12]



- Dual-branch (Siamese) Network w. shared weights
- Encoder f_θ: ResNet-50 (output: 2048 units)
- Projector h_ϕ: 3 FC layers (first 2: w. BN + ReLU, last: linear) w. 8192 nodes

Source: Mix of [11], [12]

Barlow Twins [12] (cont.)



$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{redundancy reduction}}$$

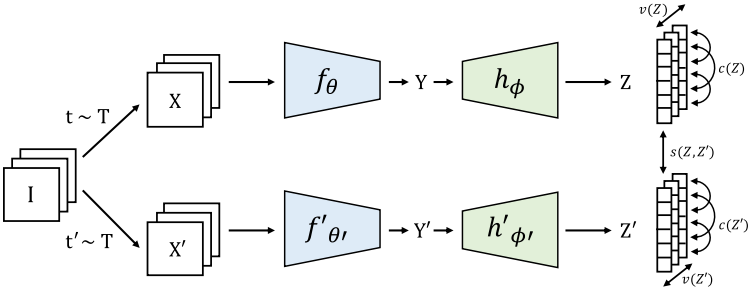
- C: cross-correlation matrix between the two branches along batch dimension
- Invariance term → forces embedding invariant to the distortions

- Redundancy reduction term: decorrelates the different vector components of the embedding

→ Reduces redundancy between output units

Source: Mix of [11], [12]

VICReg [11]



- Same architecture as Barlow Twins
+ Flexible architecture (possible: different branch archs, different inputs)

Source: Bardes et al. 2022 [11]

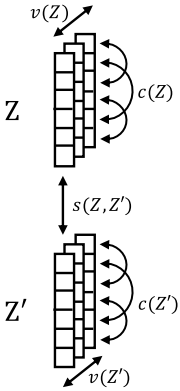
VICReg [11] (cont.)

Three losses:

1. Hinge loss maintaining variance for each dimension d_j in batch Z

$$v(Z) = \frac{1}{d}\sum_{j=1}^{d} \max(0, \gamma - \text{std}(\mathbf{z}^j))$$

- → Forces variance inside current batch to be γ ($\gamma = 1$)
- → Forces embedding vectors of samples within a batch to be different
- → Prevents collapse due to a shrinkage of the embedding vectors towards zero



Source: Bardes et al. 2022 [11]

VICReg [11] (cont.)

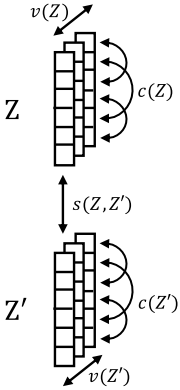
Three losses:

2. Regularization of covariance C of Z (similar to Barlow Twins):

$$c(Z) = \frac{1}{d} \sum_{j=1}^{d} \sum_{\substack{i=0 \\ i \neq j}}^{d} \left[C(Z)\right]_{i,j}^2$$

3. Invariance criterion:

$$s(Z, Z') = \frac{1}{n} \sum_{i} \|z_i - z'_i\|_2^2$$



Final loss:

$$\mathcal{L}(Z, Z') = \lambda s(Z, Z') + \mu\left[v(Z) + v(Z')\right] + \nu\left[c(Z) + c(Z')\right]$$

Source: Bardes et al. 2022 [11]

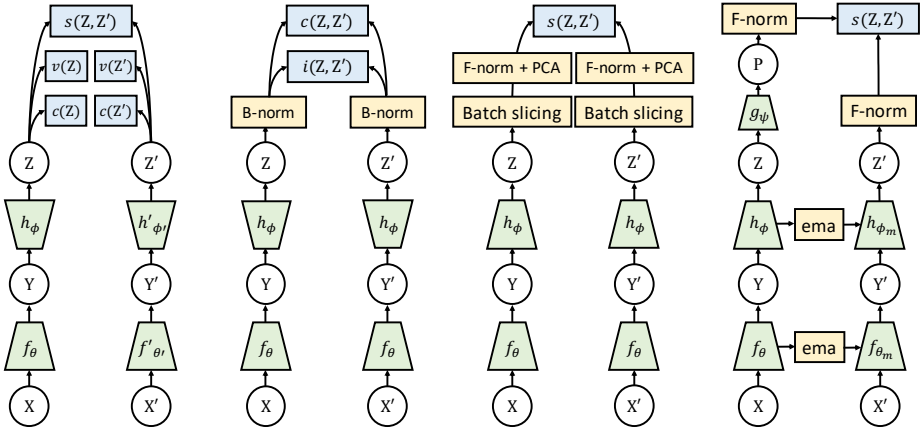
VICReg [11] (cont.)

ME: Momentum encoder | SG: Stop gradient | PR: Predictor | BN: Batch normalization

Method	ME	SG	PR	BN	No Reg	Var Reg	Var/Cov Reg
BYOL	✓	✓	✓	✓	69.3†	70.2	69.5
SimSiam		✓	✓	✓	67.9†	68.1	67.6
SimSiam		✓	✓		35.1	67.3	67.1
SimSiam		✓			collapse	56.8	66.1
VICReg			✓		collapse	56.2	67.3
VICReg			✓	✓	collapse	57.1	68.7
VICReg				✓	collapse	57.5	68.6†
VICReg					collapse	56.5	67.4

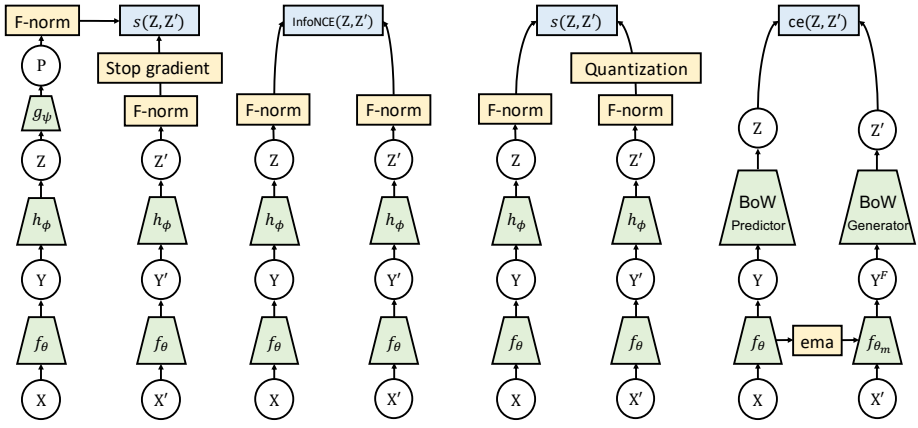
+ No memory-bank, no large batch sizes, no stop gradient
+ No batch-wise/feature-wise normalization necessary
+ No predictor module needed
– Often slightly behind SOTA

Architecture comparison



Source: Bardes et al. 2022 [11]

Architecture comparison (cont.)



(e) SimSiam (f) SimCLR (g) SwAV (h) OBoW

Source: Bardes et al. 2022 [11]

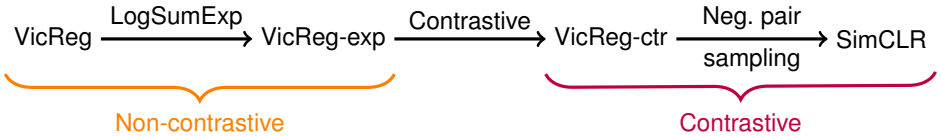
Contrastive vs. Non-contrastive [13]

Non-contrastive (=dimension contrastive) (L_{nc}) equivalent to (sample-)contrastive (L_c) methods up to normalization:

$$L_{nc} + \sum_{j=1}^{M} \|\mathcal{K}_{j,\cdot}\|_2^4 = L_c + \sum_{i=1}^{N} \|\mathcal{K}_{\cdot,i}\|_2^4$$

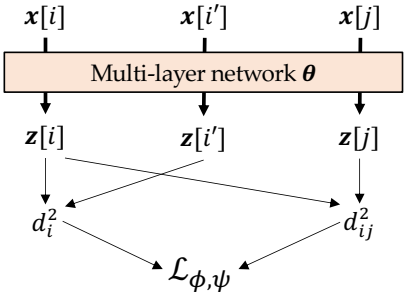
K	embedding matrix
M	batch size
N	embedding dimension

- Possible to interpolate between VICReg and SimCLR



- No performance difference between different variations
- Performance gaps mainly due to projection dimensions and hyper-parameter tuning

Contrastive Loss Generalization [21]



$$\mathcal{L}_{\phi,\psi}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \phi\Big(\sum_{j \neq i} \psi\left(d_i^2 - d_{ij}^2\right) \Big)$$

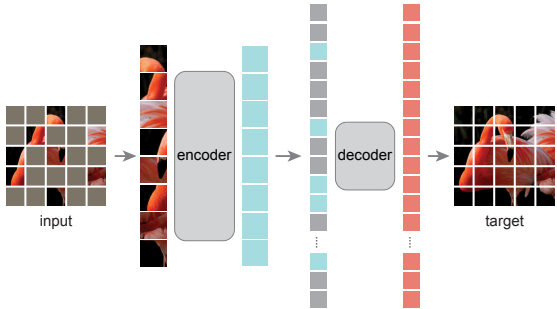
Contrastive Loss	$\phi(x)$	$\psi(x)$
InfoNCE [14]	$\tau \log(\epsilon + x)$	$e^{x/\tau}$
MINE [15]	$\log(x)$	e^x
Triplet [3]	x	$[x + \epsilon]_+$
Soft Triplet [16]	$\tau \log(1 + x)$	$e^{x/\tau+\epsilon}$
N+1 Tuplet [17]	$\log(1 + x)$	e^x
Lifted Structured [18]	$[\log(x)]_+^2$	$e^{x+\epsilon}$
Modified Triplet [19]	x	$\mathrm{sigmoid}(cx)$
Triplet Contrastive [20]	linear	linear

- Unified contrastive learning framework
- With deep linear networks: representation learning equivalent to PCA

Masked AE [6]

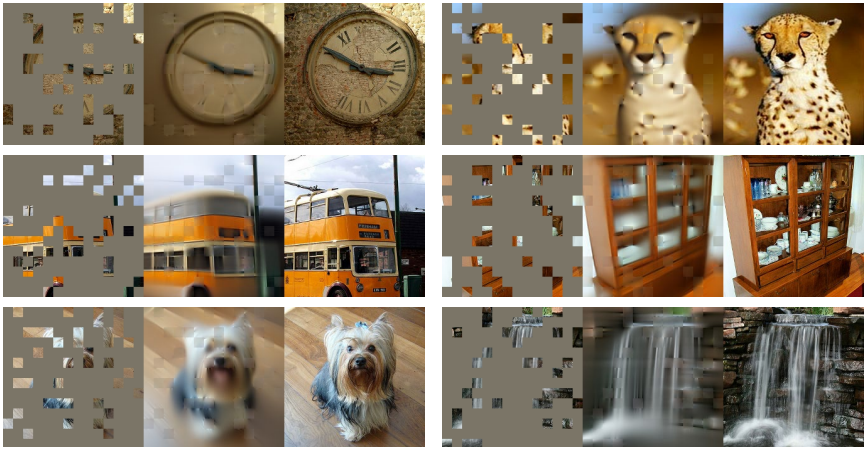
1. Divide image into non-overlapping patches
2. Mask large set of patches out (75 %)
3. Encoder: ViT applied only on visible/unmasked patches + positional encoding
4. Decoder (similar to a “projector”)
   - Input: full set of tokens of both visible + mask tokens
   - Each mask token: learned vector
   - Add positional embedding
   - Only used during pre-training

Loss: mean-squared error



Source: He et al. 2022 [6]

Masked AE [6] (cont.)



Source: He et al. 2022 [6]

SSL Cookbook [5]

- Augmentation strategy
  - Or: use reconstruction loss in pixel space (e.g., MAE [6]) or representation space (e.g., I-JEPA [22])
  - Or: enforce equivariance, s. [5]
- Multi-crop improves performance but increases training costs (e.g., SwAV [23])
- A large projector head beneficial

→ Handles noisy image augmentations [5]

- Imbalanced class-distribution affects performance (possible solution: MSN reguluarization [24])
- Teacher-student often beneficial (BYOL [8])
- Hyper-parameter tuning important (Mini-batch size, lr schedulers, optimizers, weight decay), especially for ViT (increase mini-batch size, decrease patch-size, ...)

→ “DINO v2” [25]

Summary & Other Directions

Summary

- From deep metric learning to SSL
- Contrastive methods vs. non-contrastive
- No clear winner among DML and SSL methods

Other directions

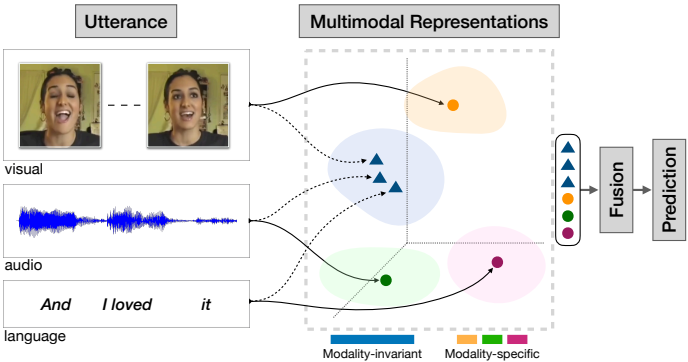
- Ranking-based learning, a. k. a. list-based losses
- Other pre-text tasks possible, e. g. cluster-based self-supervised learning (SwAV [23])
- Features for registration (SuperPoint [26])
- Domain contrastive/adversarial learning

NEXT TIME

ADVANCED

ON DEEP LEARNING

Multi-modal learning



Source: Hazarika et al. 2020 [27]

WIP: Comprehensive Questions

- What is self-supervised learning, and how does it differ from supervised and unsupervised learning methods?
- What are the main advantages of using self-supervised learning techniques compared to traditional supervised learning?
- Can you explain the concept of pretext tasks in self-supervised learning, and provide examples of how these tasks are used to generate meaningful representations?
- What are important hyper-parameters for self-supervised learning?
- What is the role of the projection head?

Further Reading

- K. Musgrave: PyTorch Metric Learning Library: https://kevinmusgrave.github.io/pytorch-metric-learning/
- SSL Cookbook by Balestriero et al. [5]

References

References I

[1] Longlong Jing and Yingli Tian. “Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 43.11 (2021), pp. 4037–4058.

[2] S. Chopra, R. Hadsell, and Y. LeCun. “Learning a similarity metric discriminatively, with application to face verification”. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05). Vol. 1. 2005, pp. 539–546.

[3] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “Facenet: A unified embedding for face recognition and clustering”. In: CVPR. 2015.

[4] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. “A Metric Learning Reality Check”. In: Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020, pp. 681–699.

[5] Randall Balestriero, Mark Ibrahim, Vlad Sobal, et al. A Cookbook of Self-Supervised Learning. 2023. arXiv: 2304.12210 [cs.LG].

References II

[6] Kaiming He, Xinlei Chen, Saining Xie, et al. “Masked Autoencoders Are Scalable Vision Learners”. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2022, pp. 15979–15988.

[7] Quentin Garrido, Randall Balestriero, Laurent Najman, et al. RankMe: Assessing the downstream performance of pretrained self-supervised representations by the 2023. arXiv: 2210.02885 [cs.LG].

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, et al. “Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning”. In: Advances in Neural Information Processing Systems. Vol. 33. Curran Associates, Inc., 2020, pp. 21271–21284.

[9] Xinlei Chen and Kaiming He. “Exploring Simple Siamese Representation Learning”. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021, pp. 15745–15753.

References III

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, et al. “Emerging Properties in Self-Supervised Vision Transformers”. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Oct. 2021, pp. 9650–9660.

[11] Adrien Bardes, Jean Ponce, and Yann LeCun. “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning”. In: International Conference on Learning Representations. 2022.

[12] Jure Zbontar, Li Jing, Ishan Misra, et al. “Barlow Twins: Self-Supervised Learning via Redundancy Reduction”. In: Proceedings of the 38th International Conference on Machine Learning. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 12310–12320.

[13] Quentin Garrido, Yubei Chen, Adrien Bardes, et al. “On the duality between contrastive and non-contrastive self-supervised learning”. In: The Eleventh International Conference on Learning Representations. 2023.

References IV

[14] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. “Representation learning with contrastive predictive coding”. In: arXiv preprint arXiv:1807.03748 (2018).

[15] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, et al. “Mutual information neural estimation”. In: International Conference on Machine Learning. PMLR. 2018, pp. 531–540.

[16] Yuandong Tian, Lantao Yu, Xinlei Chen, et al. “Understanding self-supervised learning with dual deep networks”. In: arXiv preprint arXiv:2010.00578 (2020).

[17] Kihyuk Sohn. “Improved deep metric learning with multi-class n-pair loss objective”. In: Advances in neural information processing systems. 2016, pp. 1857–1865.

[18] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, et al. “Deep metric learning via lifted structured feature embedding”. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 4004–4012.

References V

[19] Juan M Coria, Hervé Bredin, Sahar Ghannay, et al. “A comparison of metric learning loss functions for end-to-end speaker verification”. In: International Conference on Statistical Language and Speech Processing. Springer. 2020, pp. 137–148.

[20] Wenlong Ji, Zhun Deng, Ryumei Nakada, et al. “The Power of Contrast for Feature Learning: A Theoretical Analysis”. In: arXiv preprint arXiv:2110.02473 (2021).

[21] Yuandong Tian. Understanding Deep Contrastive Learning via Coordinate-wise Optimization. 2022. arXiv: 2201.12680 [cs.LG].

[22] Mahmoud Assran, Quentin Duval, Ishan Misra, et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. 2023. arXiv: 2301.08243 [cs.CV].

References VI

[23] Mathilde Caron, Ishan Misra, Julien Mairal, et al. “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments”. In: Advances in Neural Information Processing Systems. Vol. 33. Curran Associates, Inc., 2020, pp. 9912–9924.

[24] Mahmoud Assran, Mathilde Caron, Ishan Misra, et al. “Masked Siamese Networks for Label-Efficient Learning”. In: Computer Vision – ECCV 2022. Cham: Springer Nature Switzerland, 2022, pp. 456–473.

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning Robust Visual Features without Supervision. 2023. arXiv: 2304.07193 [cs.CV].

[26] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. “SuperPoint: Self-Supervised Interest Point Detection and Description”. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). June 2018, pp. 337–33712.

References VII

[27] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. “MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis”. In: Proceedings of the 28th ACM International Conference on Multimedia. MM ’20. Seattle, WA, USA: Association for Computing Machinery, 2020, pp. 1122–1131.

[28] Ting Chen, Simon Kornblith, Mohammad Norouzi, et al. “A Simple Framework for Contrastive Learning of Visual Representations”. In: Proceedings of the 37th International Conference on Machine Learning. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 1597–1607.

[29] Shikun Liu, Edward Johns, and Andrew J. Davison. “End-To-End Multi-Task Learning With Attention”. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2019, pp. 1871–1880.

References VIII

[30] Chaoning Zhang, Kang Zhang, Chenshuang Zhang, et al. “How Does SimSiam Avoid Collapse Without Negative Samples? A Unified Understanding with Self-supervised Contrastive Learning”. In: International Conference on Learning Representations. 2022.