清華大學 Tsinghua University

Advanced Computer Vision

THU×SENSETIME – 80231202

商汤 sensetime

Chapter 2 - Section 13

Representation Learning in Vision Tasks

Dr. Liu Yu

Thursday, May 20, 2021

Acknowledge : Song Guanglu , Liu Boxiao , Zhang Manyuan

13.1 Metric Learning

Dr. Liu Yu

Thursday, May 20, 2021

商汤 sensetime

清華大學 Tsinghua University

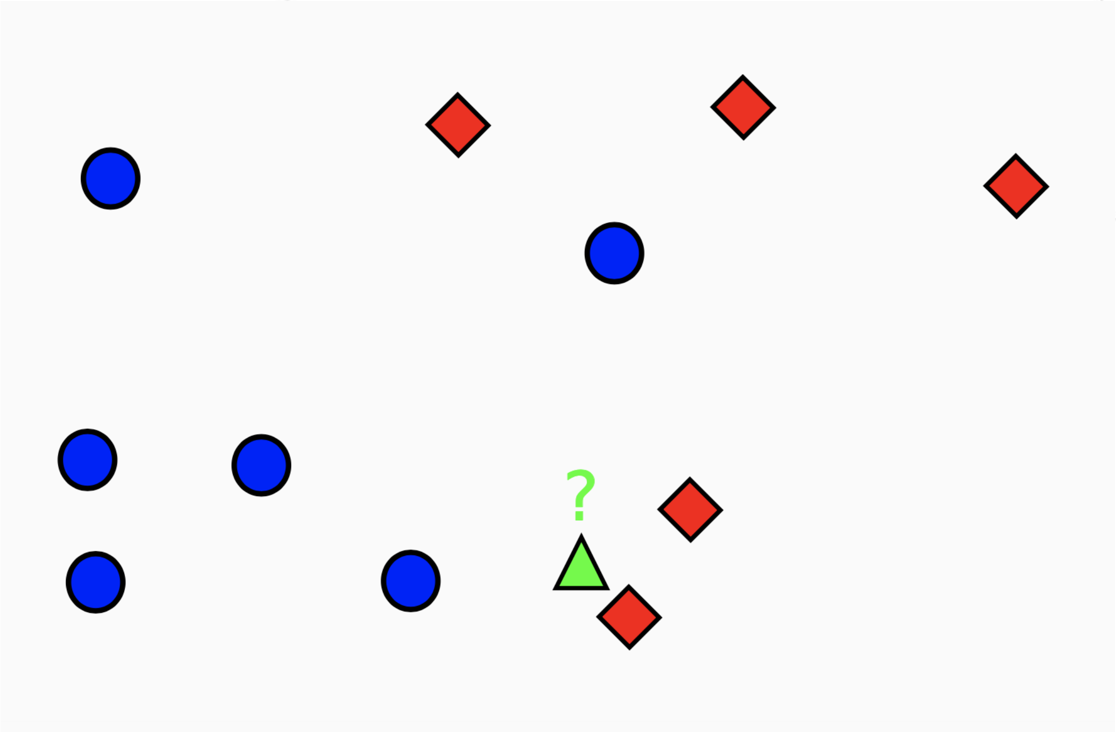
Outline

Introduction

- Similarity / distance judgements are essential components of many human cognitive processes.
  - Compare perceptual or conceptual representations.
  - Perform recognition, categorization.

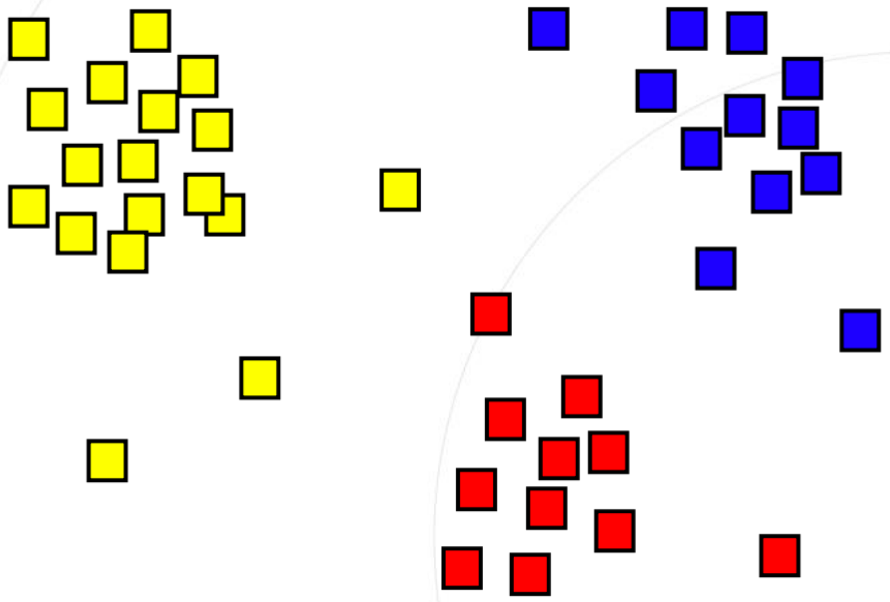
- Underlie most machine learning and data mining techniques.

- Nearest neighbor classification



If you want to find the nearest neighbor, you should calculate the distance between samples.

- Clustering



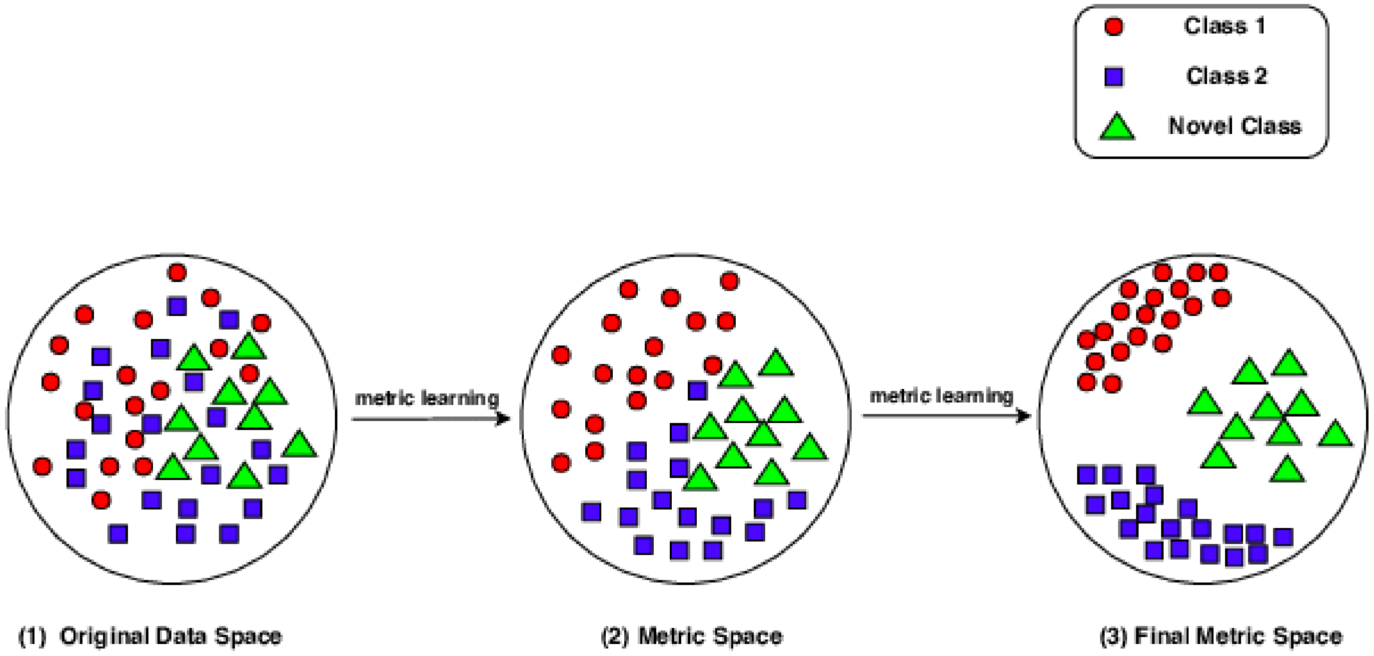
If you want to do clustering, you should calculate the distance between samples.

Introduction

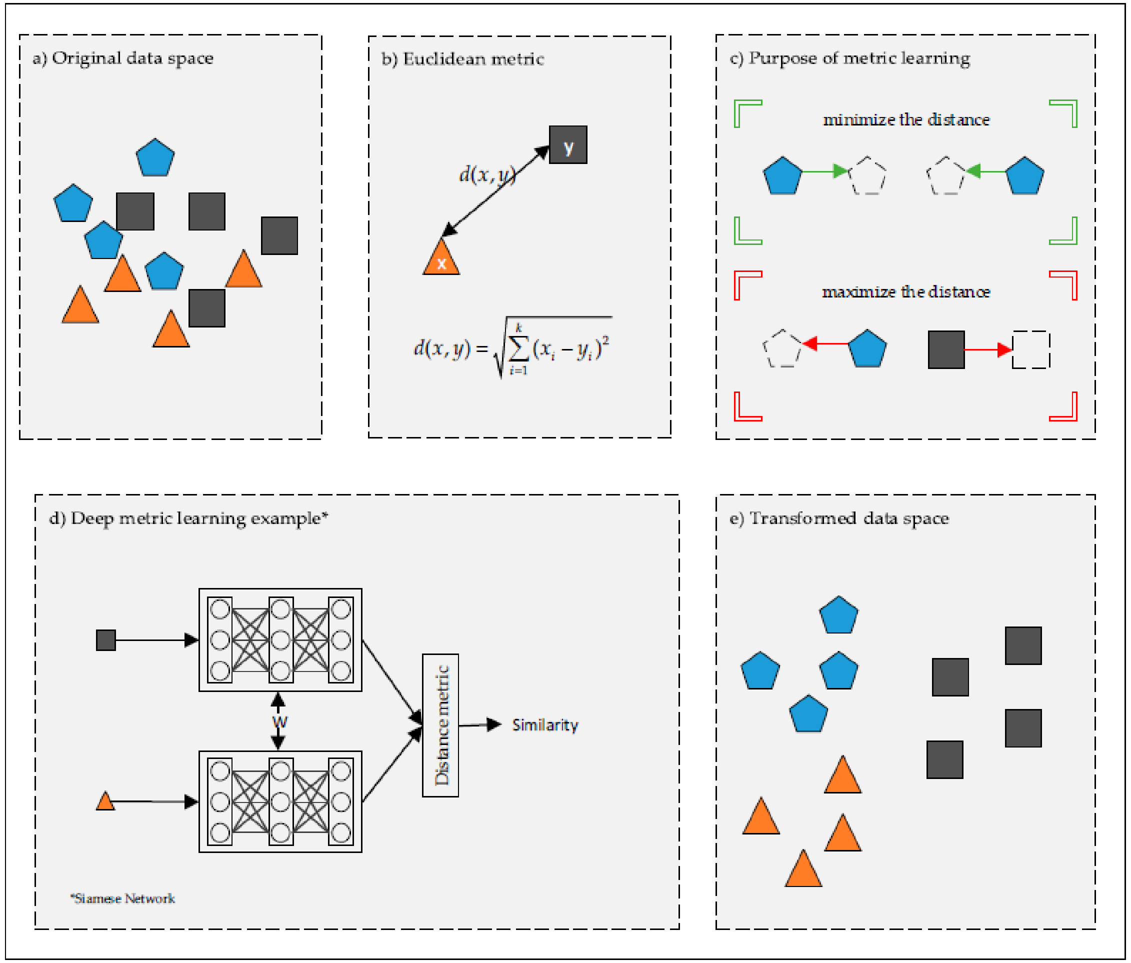
- Choice of similarity is crucial to the performance.
- Humans weight features differently depending on context.
- Fundamental question: how to appropriately measure similarity or distance for a given task?
- Metric learning + infer this automatically from data.
- Note: we will refer to distance or similarity indistinctly as metric.

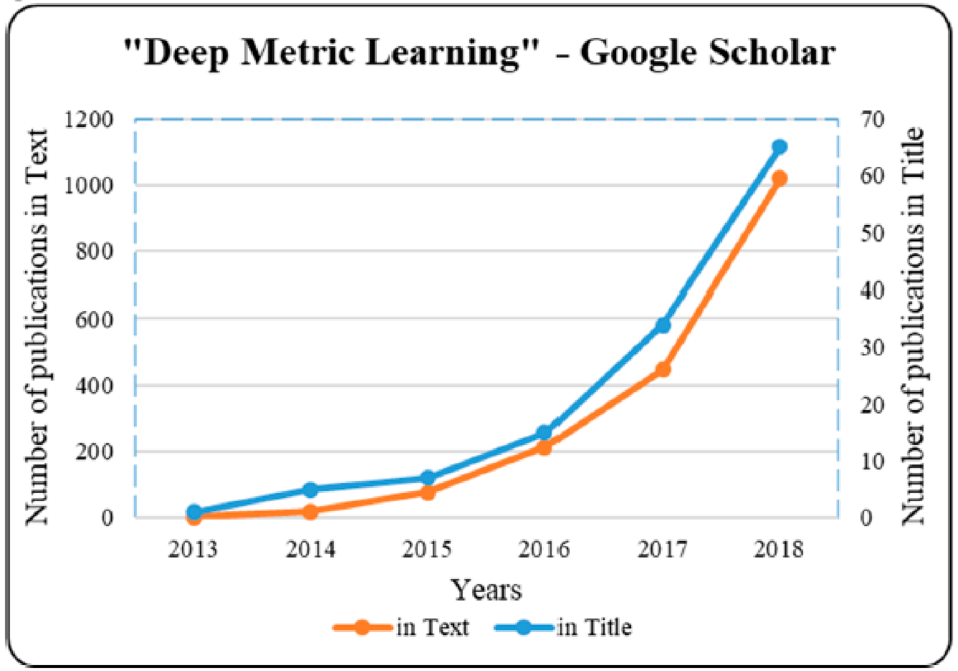
- Measuring Similarity Between Data

  - Similarity: computing distances between data points.
  - Performance: depending on the definitions of similarity.



- Deep Metric Learning





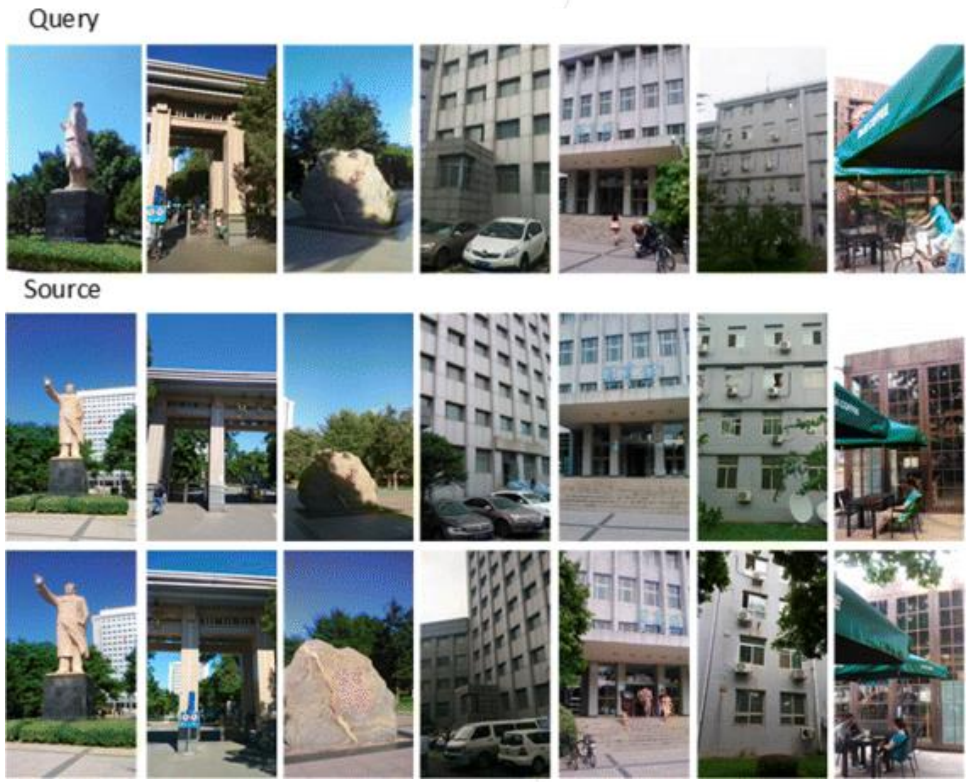
Kaya M, Bilge H Ş. Deep metric learning: A survey[J]. Symmetry, 2019, 11(9): 1066.

- Examples for deep metric learning
  - Face recognition

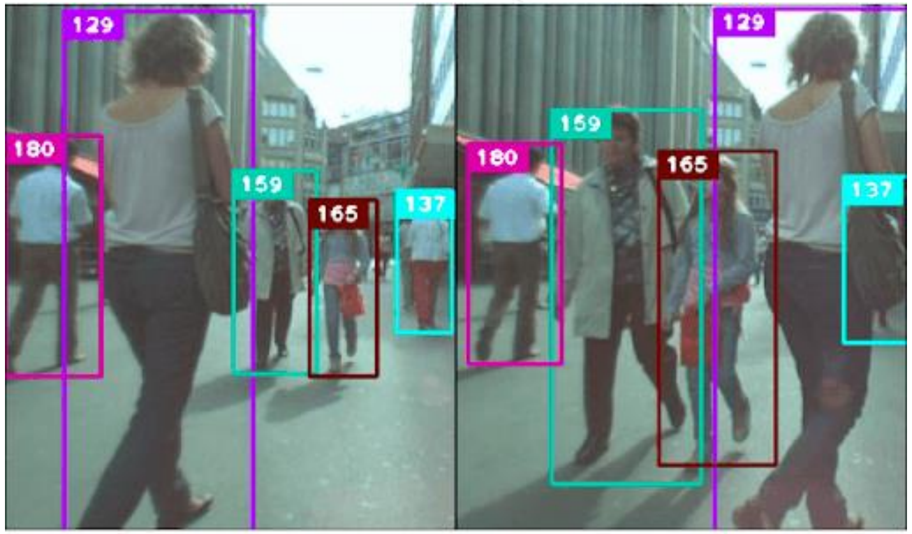
  - Person Re-identification

Introduction

- Examples for deep metric learning
  - Multimedia Searching



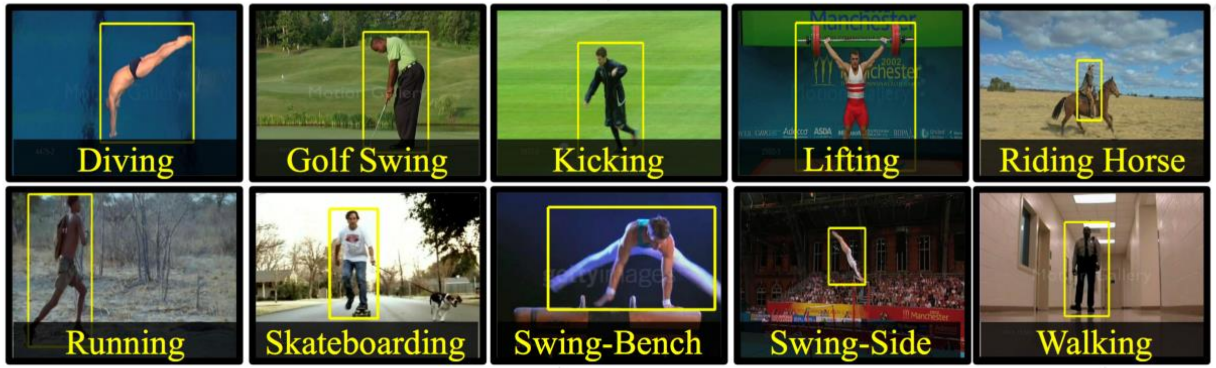
  - Tracking



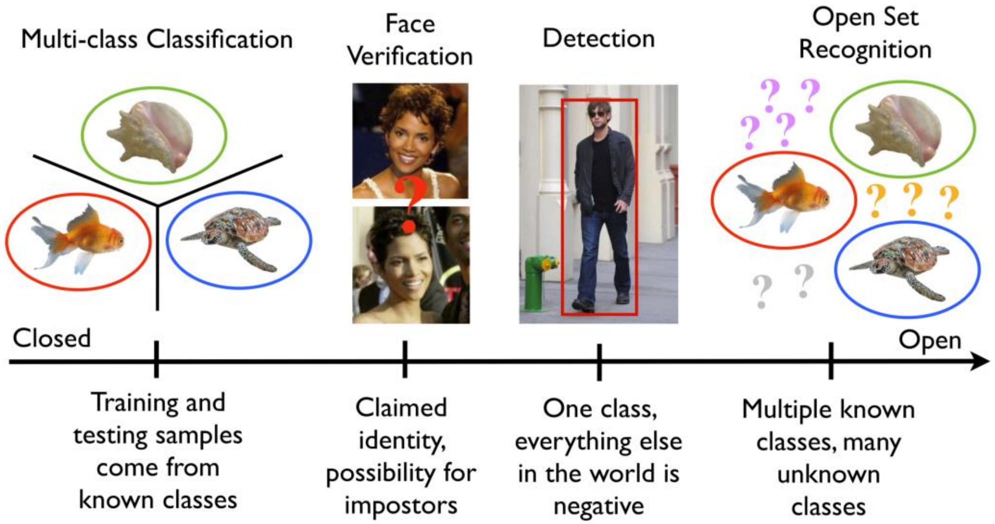
Introduction

- Examples for deep metric learning

  - Activity Recognition



  - Open-set recognition



Introduction

Tsinghua University sensetime

- Measure Similarity: Metric

A **metric** is a function that defines a distance between each pair of elements of a set.

- **Euclidean** or L2:

$$d_{\text{Euclidean}}\left(\bar{x}_1, \bar{x}_2\right) = \|\bar{x}_1 - \bar{x}_2\|_2 = \sqrt{\sum_i \left(x_1^i - x_2^i\right)^2}$$

- **Manhattan** or L1:

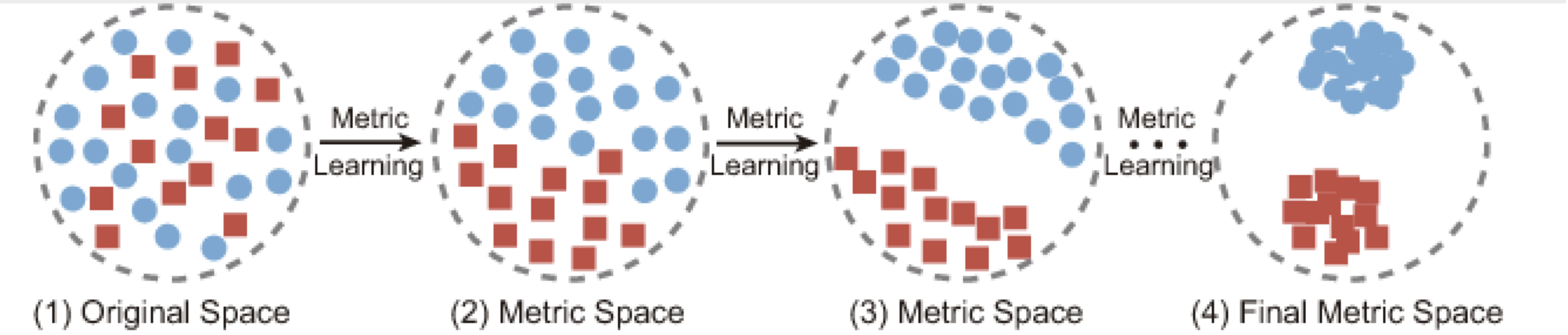
$$d_{\text{Manhattan}}\left(\bar{x}_1, \bar{x}_2\right) = \|\bar{x}_1 - \bar{x}_2\|_1 = \sum_i \left|x_1^{\ell} - x_2^i\right|$$

- **Cosine distance:**

$$d_{\text{Cosine}}\left(\bar{x}_1, \bar{x}_2\right) = 1 - \frac{\bar{x}_1 \cdot \bar{x}_2}{\|\bar{x}_1\|_2 \|\bar{x}_2\|_2}$$

Introduction

- To form compact representations



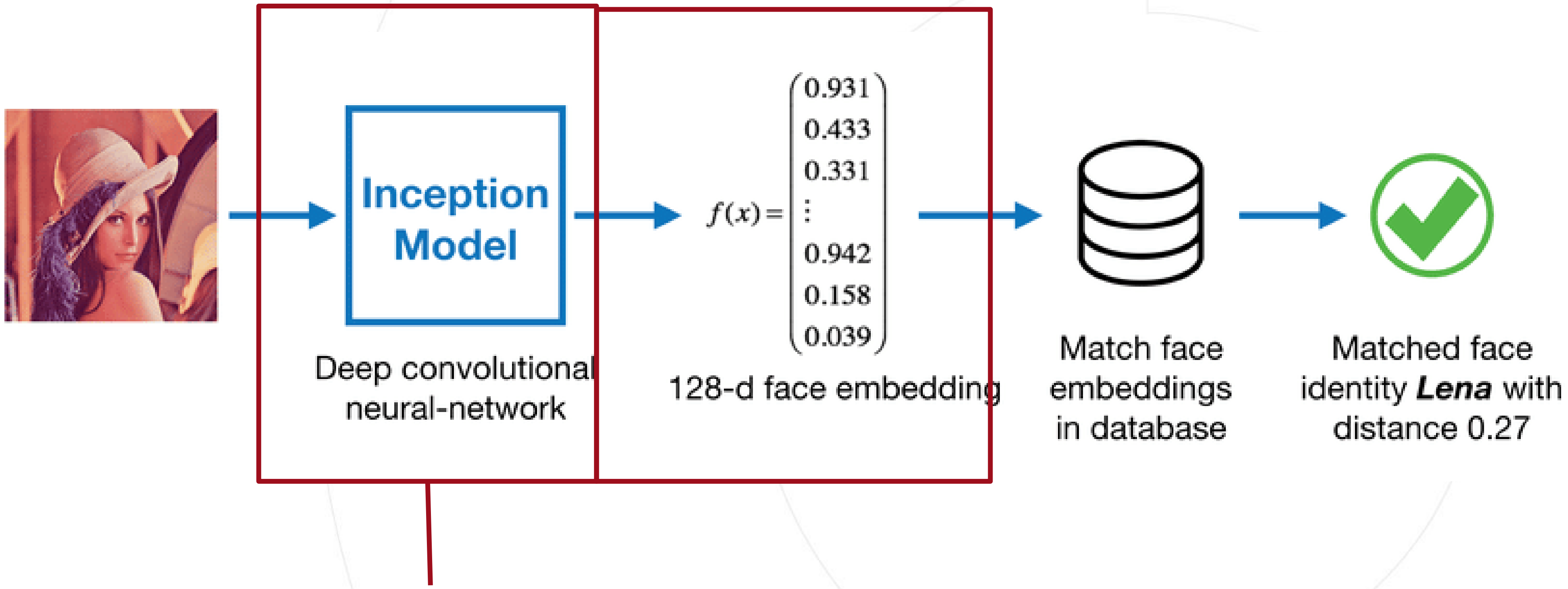
商汤 sensetime

清華大學 Tsinghua University

Outline

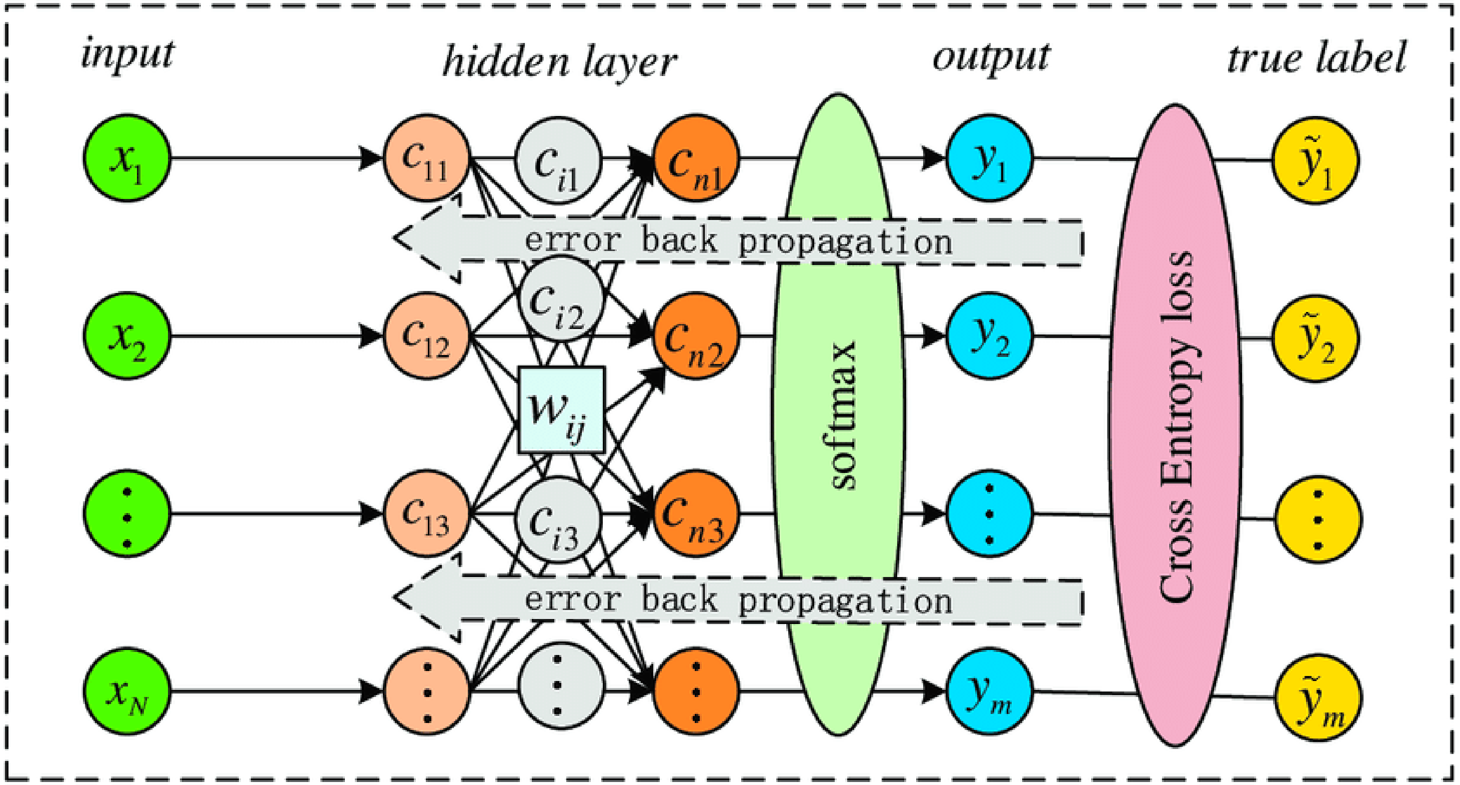
Metric learning for face recognition

- Face recognition pipeline



Training by **loss function**

- Softmax cross-entropy Loss

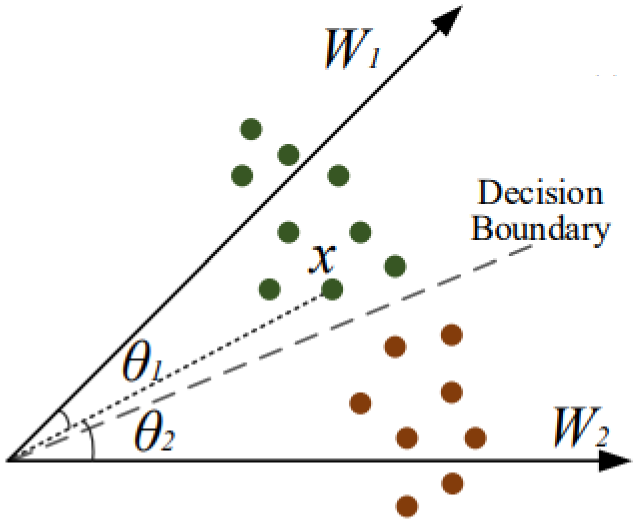


$$L = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right)$$

$$W_1^T x \geq W_2^T x$$

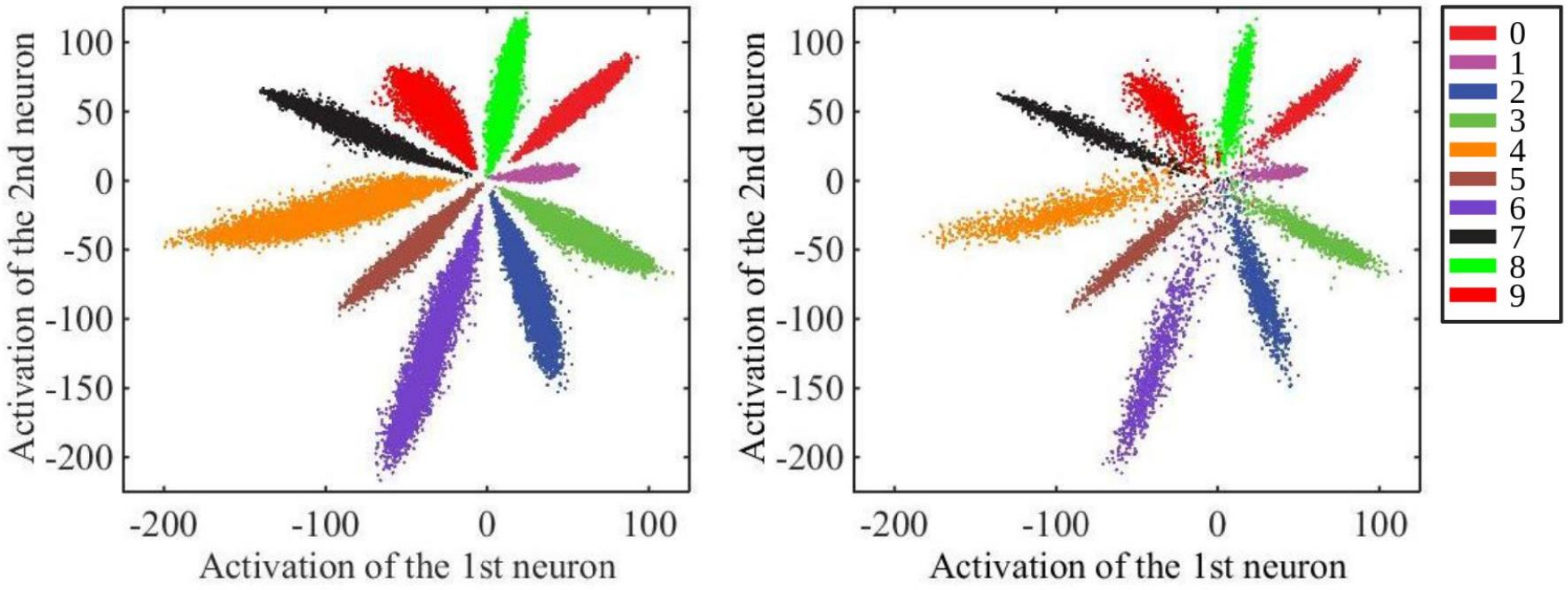
$$\Leftrightarrow$$

$$||W_1||_2||x||_2\cos(\theta_1) \geq ||W_2||_2||x||_2\cos(\theta_2)$$



Metric learning for face recognition

- Is SoftmaxWithLoss good for clustering?

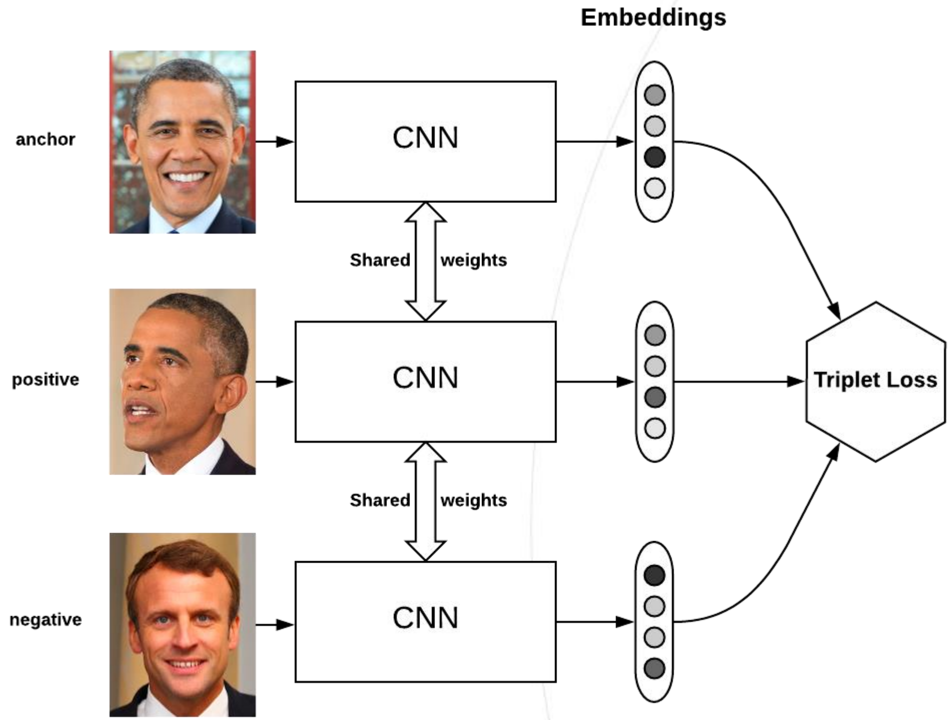


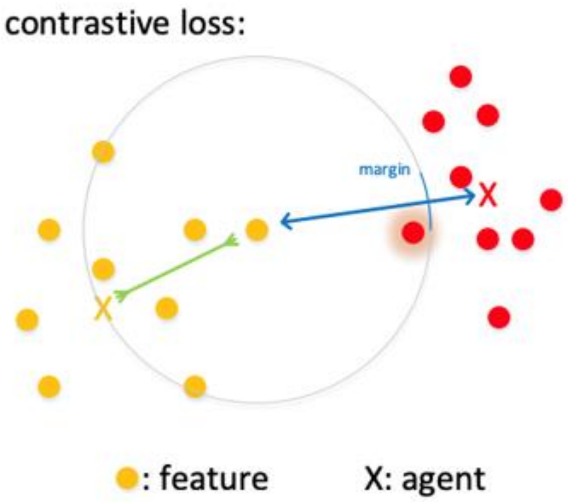
Separable.

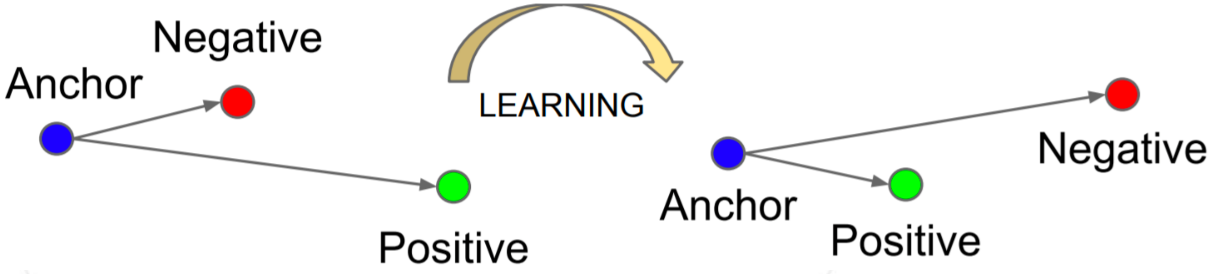
The deep features are not discriminative enough due to intra-class variation

Metric learning for face recognition

- Triplet loss function







Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering [C]// CVPR, 2015.

- Triplet loss function

The goal of the triplet loss is to make sure that:

- Two examples with the **same label** have their embeddings **close** together in the embedding space
- Two examples with **different** labels have their embeddings **far away**.

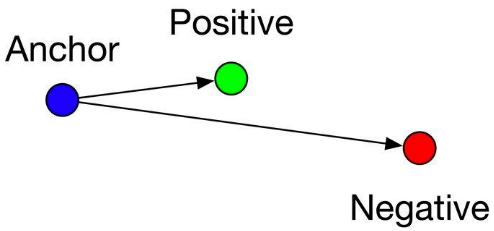
$$\mathcal{L} = max(d(a,p) - d(a,n) + margin, 0)$$

To formalise this requirement, the loss will be defined over triplets of embeddings:

- an anchor
- a positive of the same class as the anchor
- a negative of a different class

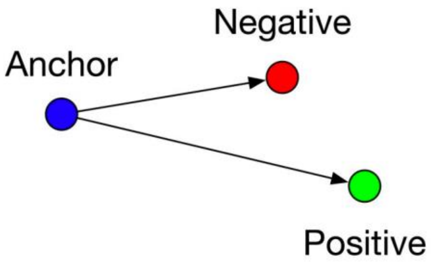
- Hard triplet mining
  - easy triplets

$$d(a,p) + margin < d(a,n)$$



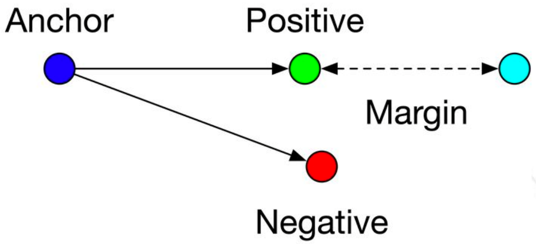
  - hard triplets

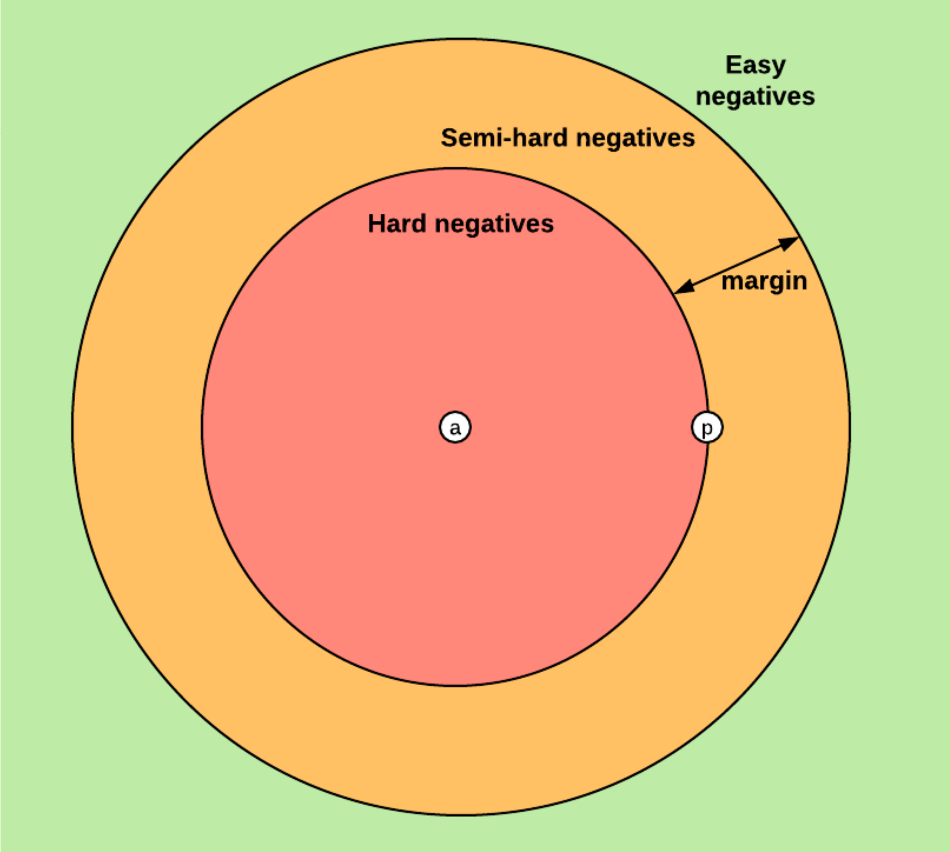
$$d(a,n) < d(a,p)$$



  - semi-hard triplets

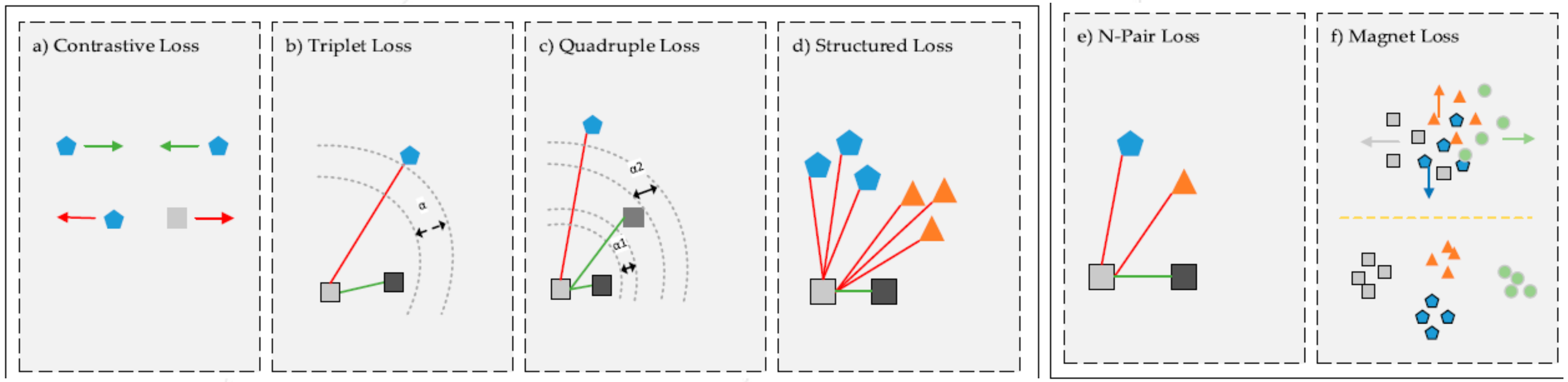
$$d(a,p) < d(a,n) < d(a,p) + margin$$





Metric learning for face recognition

- Metric loss functions



$$D_W(X_1, X_2) = \|G_W(X_1) - G_W(X_2)\|_2$$

(a) $L_{Contrastive} = (1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{\max(0, m - D_W)\}^2$

(b) $L_{Triplet} = \max(0, \|G_W(X) - G_W(X^p)\|_2 - \|G_W(X) - G_W(X^n)\|_2 + \alpha)$

(c) $L_{Quadruple} = \max(0, \|G_W(X) - G_W(X^p)\|_2 - \|G_W(X) - G_W(X^s)\|_2 + \alpha_1) + \max(0, \|G_W(X) - G_W(X^s)\|_2 - \|G_W(X) - G_W(X^n)\|_2 + \alpha_2)$

(d)
$$J = \frac{1}{2|\widehat{\mathcal{P}}|} \sum_{(i,j)\in\widehat{\mathcal{P}}} \max(0, J_{i,j})^2,$$
$$J_{i,j} = \max\left(\max_{(i,k)\in\widehat{\mathcal{N}}} \alpha - D_{i,k}, \max_{(j,l)\in\widehat{\mathcal{N}}} \alpha - D_{j,l} \right) + D_{i,j}$$

(e) $\mathcal{L}_{N\text{-pair-ovo}}(\{(x_i, x_i^+)\}_{i=1}^N; f) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i} \log\left(1 + \exp(f_i^\top f_j^+ - f_i^\top f_i^+)\right).$

(f) $\mathscr{L}(\mathbf{\Theta}) = \frac{1}{N}\sum_{n=1}^{N}\left\{ -\log \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{r}_n - \boldsymbol{\mu}(\mathbf{r}_n)\|_2^2 - \alpha}}{\sum_{c\neq C(\mathbf{r}_n)}\sum_{k=1}^{K} e^{-\frac{1}{2\sigma^2}\|\mathbf{r}_n - \boldsymbol{\mu}_k^c\|_2^2}} \right\}_+$

Kaya M, Bilge H Ş. Deep metric learning: A survey[J]. Symmetry, 2019, 11(9): 1066.

Metric learning for face recognition

- Center loss function

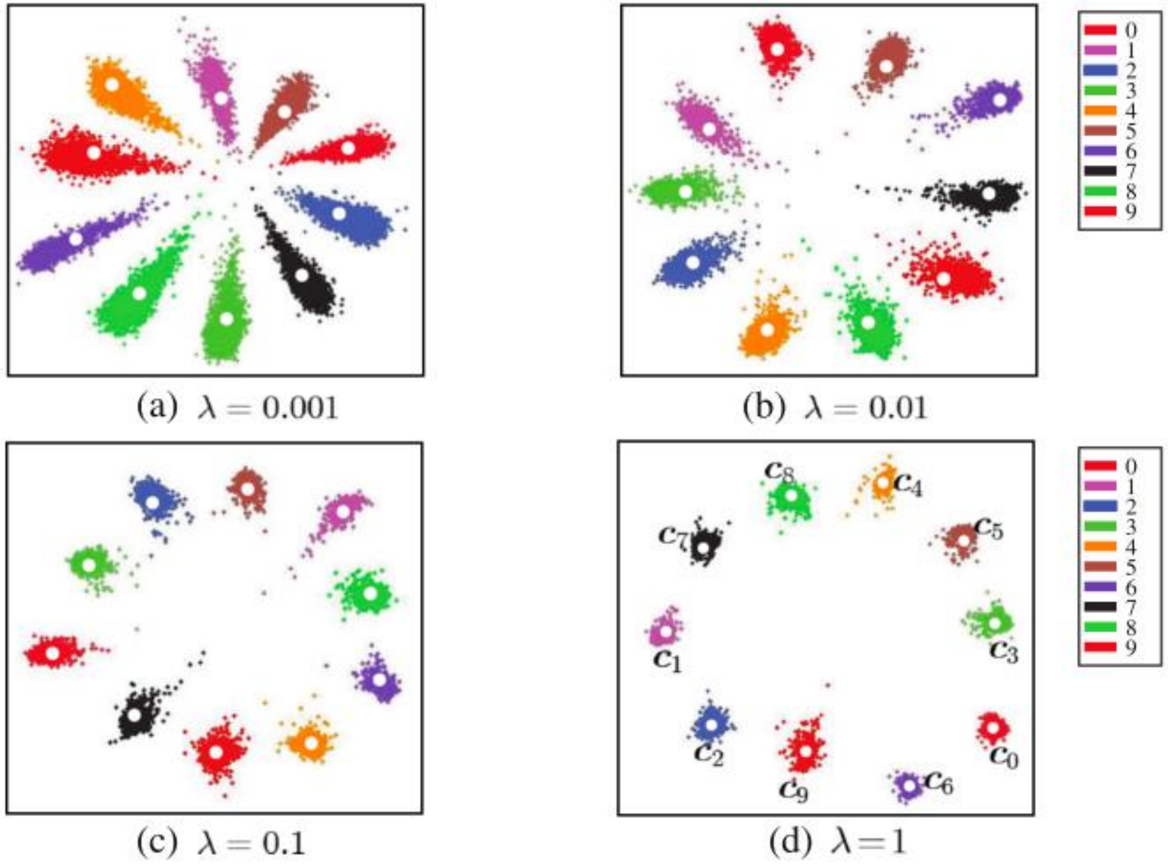
$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$$

$$\frac{\partial \mathcal{L}_C}{\partial \boldsymbol{x}_i} = \boldsymbol{x}_i - \boldsymbol{c}_{y_i}$$

$$\Delta \boldsymbol{c}_j = \frac{\sum_{i=1}^{m} \delta(y_i = j)\cdot(\boldsymbol{c}_j - \boldsymbol{x}_i)}{1 + \sum_{i=1}^{m} \delta(y_i = j)}$$

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

$$= -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \boldsymbol{x}_i + b_j}} + \frac{\lambda}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$$

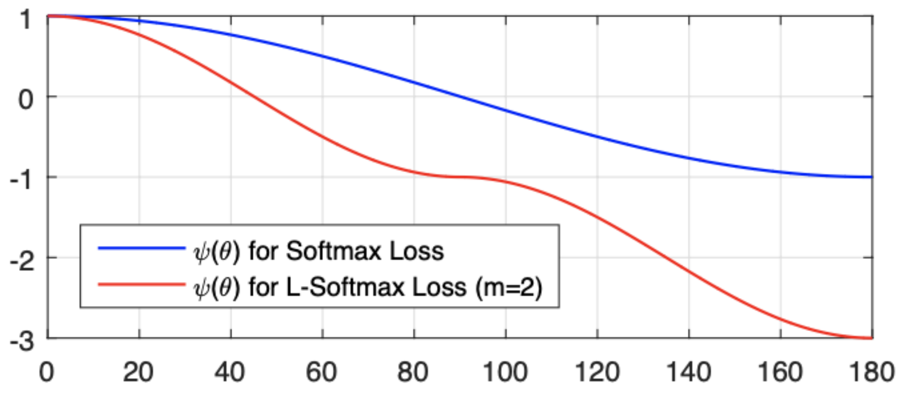


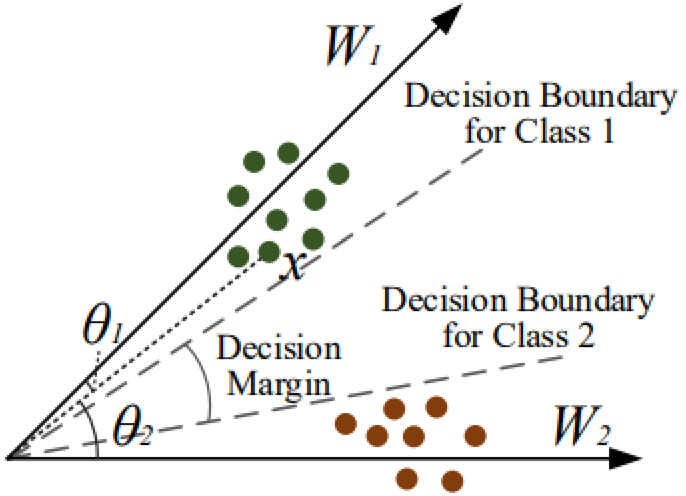
Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition [C]// ECCV, 2016.

- Large Margin Softmax

$$L_i = -\log\left(\frac{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\psi(\theta_{y_i})}}{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\psi(\theta_{y_i})} + \sum_{j\neq y_i} e^{\|\boldsymbol{W}_j\|\|\boldsymbol{x}_i\|\cos(\theta_j)}}\right)$$

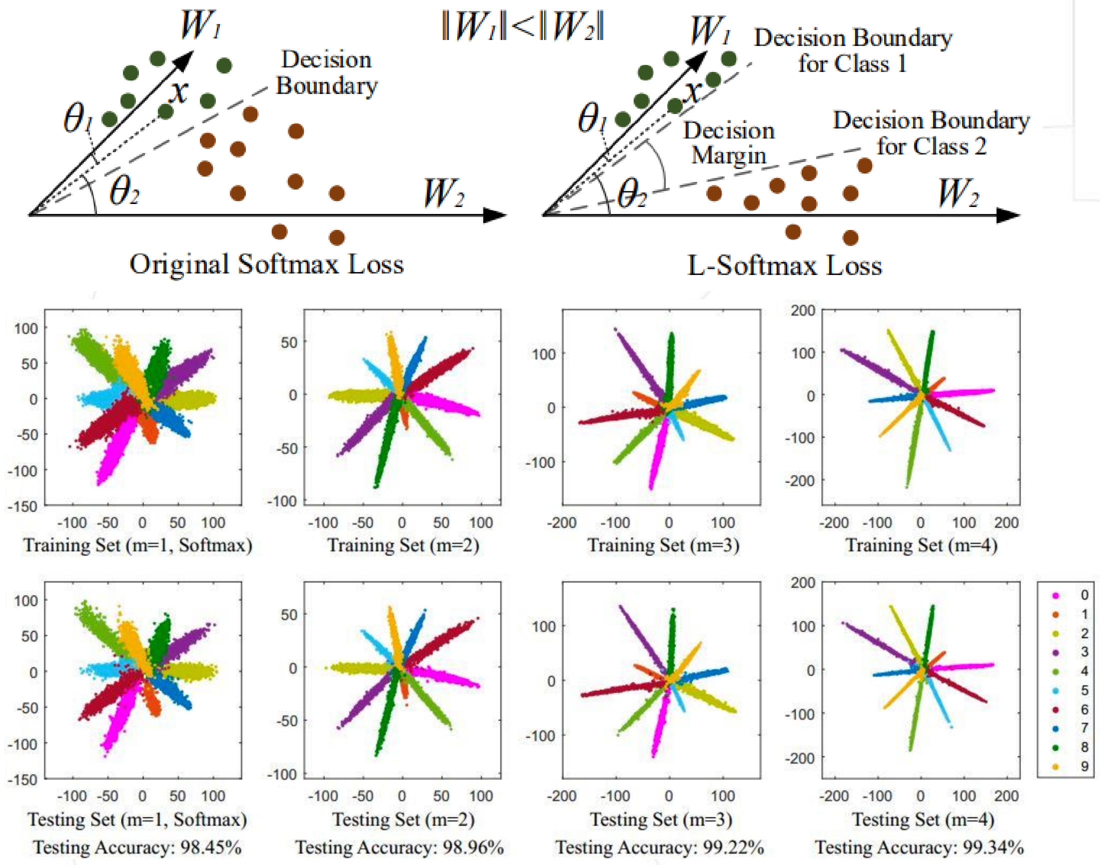
$$\psi(\theta) = (-1)^k \cos(m\theta) - 2k, \quad \theta \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}] \quad k \in [0, m-1] \text{ and } k \text{ is an integer}$$





Liu W, Wen Y, Yu Z, et al. Large-Margin Softmax Loss for Convolutional Neural Networks [C]// ICML, 2016.

- Large Margin Softmax



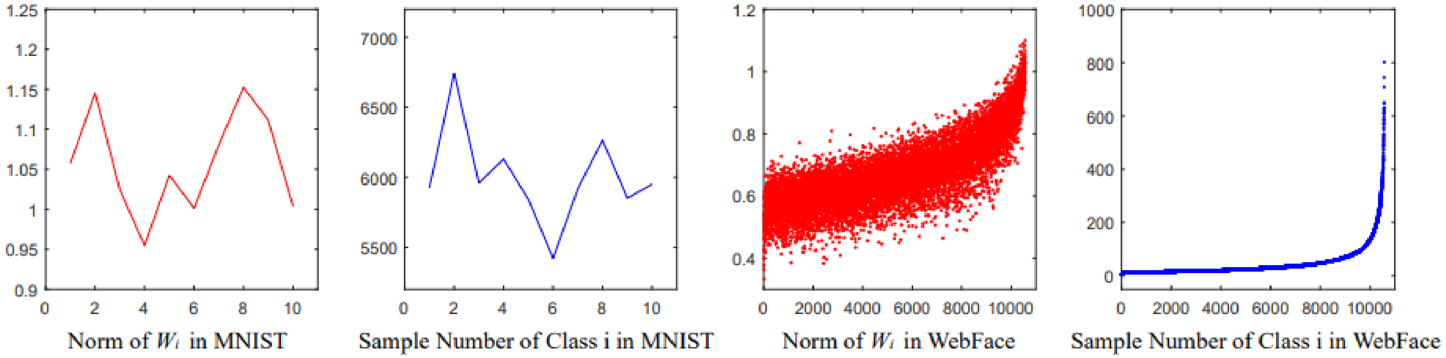
Liu W, Wen Y, Yu Z, et al. Large-Margin Softmax Loss for Convolutional Neural Networks [C]// ICML, 2016.

- SphereFace

$$L_{\text{ang}} = \frac{1}{N}\sum_i -\log\left(\frac{e^{\|\boldsymbol{x}_i\|\psi(\theta_{y_i,i})}}{e^{\|\boldsymbol{x}_i\|\psi(\theta_{y_i,i})} + \sum_{j\neq y_i} e^{\|\boldsymbol{x}_i\|\cos(\theta_{j,i})}}\right)$$

$$\psi(\theta) = (-1)^k \cos(m\theta) - 2k, \quad \theta \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}]$$

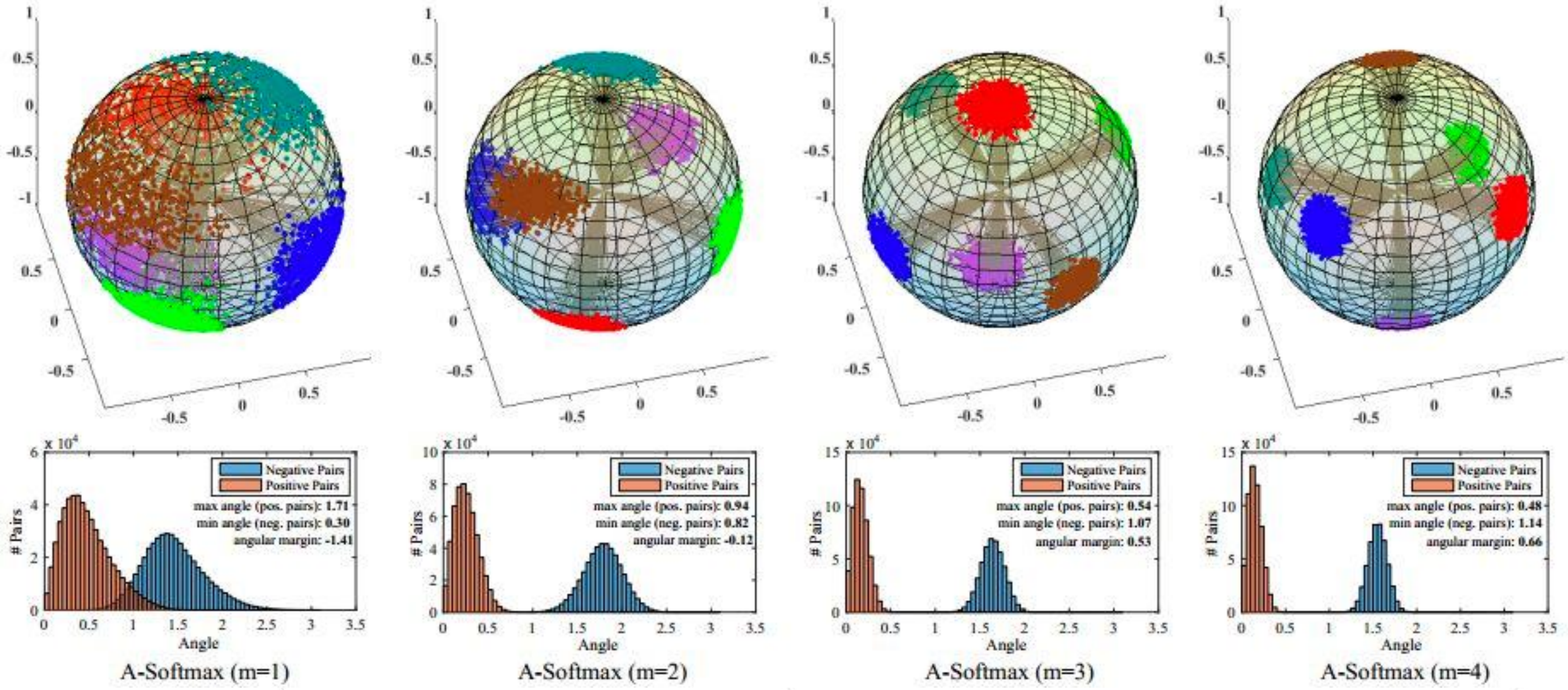
- **Normalizing the weights could reduce the prior caused by the training data imbalance**



Liu W, Wen Y, Yu Z, et al. SphereFace: Deep Hypersphere Embedding for Face Recognition [C]// CVPR. 2017.

Metric learning for face recognition

- SphereFace



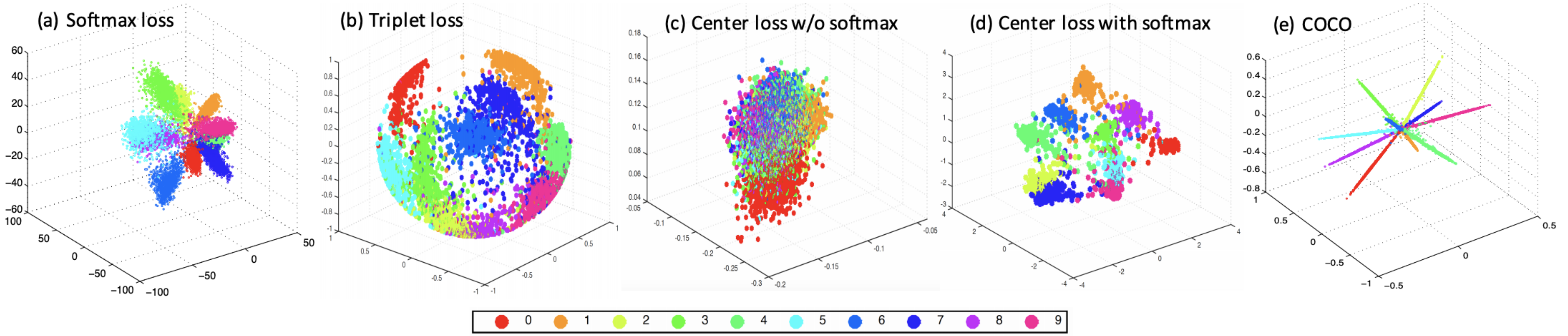
Visualization of features learned with different m.

Liu W, Wen Y, Yu Z, et al. SphereFace: Deep Hypersphere Embedding for Face Recognition [C]// CVPR. 2017.

Metric learning for face recognition

- COCO (Feature Normalization)

$$\mathcal{L}^{COCO}\left(\boldsymbol{f}^{(i)}, \boldsymbol{c}_k\right) = -\sum_{i\in\mathcal{B},k} t_k^{(i)} \log p_k^{(i)} = -\sum_{i\in\mathcal{B}} \log p_{l_i}^{(i)} \quad \hat{\boldsymbol{c}}_k = \frac{\boldsymbol{c}_k}{\|\boldsymbol{c}_k\|}, \hat{\boldsymbol{f}}^{(i)} = \frac{\alpha \boldsymbol{f}^{(i)}}{\left\|\boldsymbol{f}^{(i)}\right\|}, p_k^{(i)} = \frac{\exp\left(\hat{\boldsymbol{c}}_k^T \cdot \hat{\boldsymbol{f}}^{(i)}\right)}{\sum_m \exp\left(\hat{\boldsymbol{c}}_m^T \cdot \hat{\boldsymbol{f}}^{(i)}\right)}$$

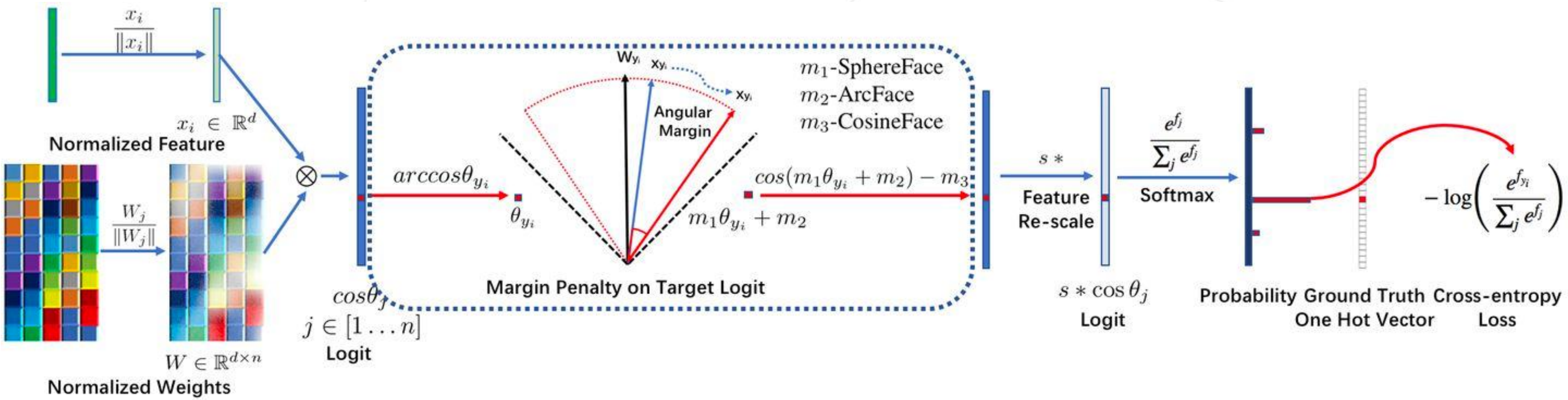


Feature visualization under different loss strategies, trained on MNIST.

Liu W, Wen Y, Yu Z, et al. SphereFace: Deep Hypersphere Embedding for Face Recognition [C]// CVPR. 2017.

- Additive Margin Loss

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}}$$



The overall pipeline for Additive Margin (ArcFace) loss.

Jiankang Deng, Jia Guo, Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition [C]// CVPR, 2019.

商汤 sensetime

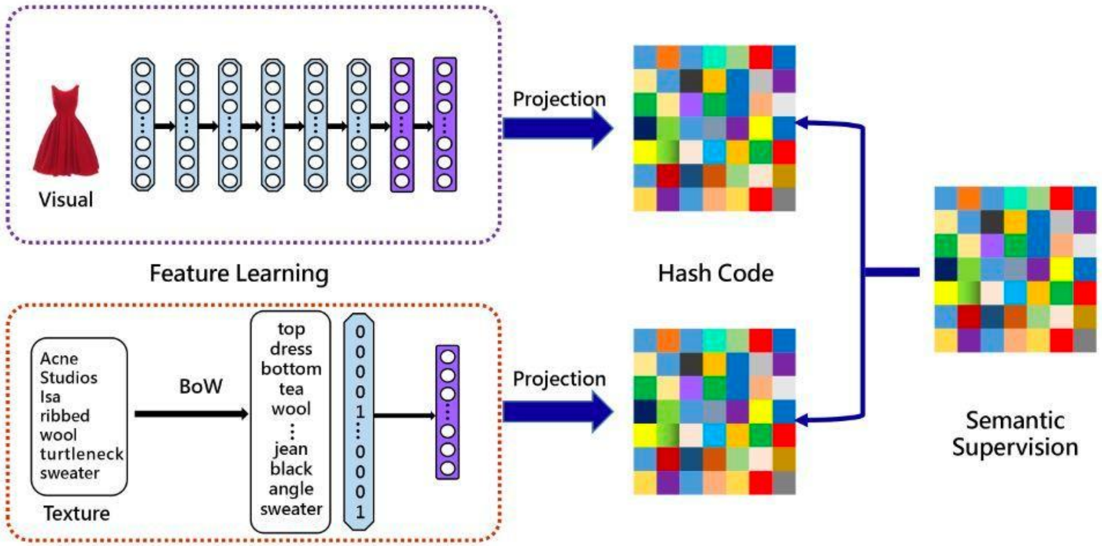
清華大學 Tsinghua University

Outline

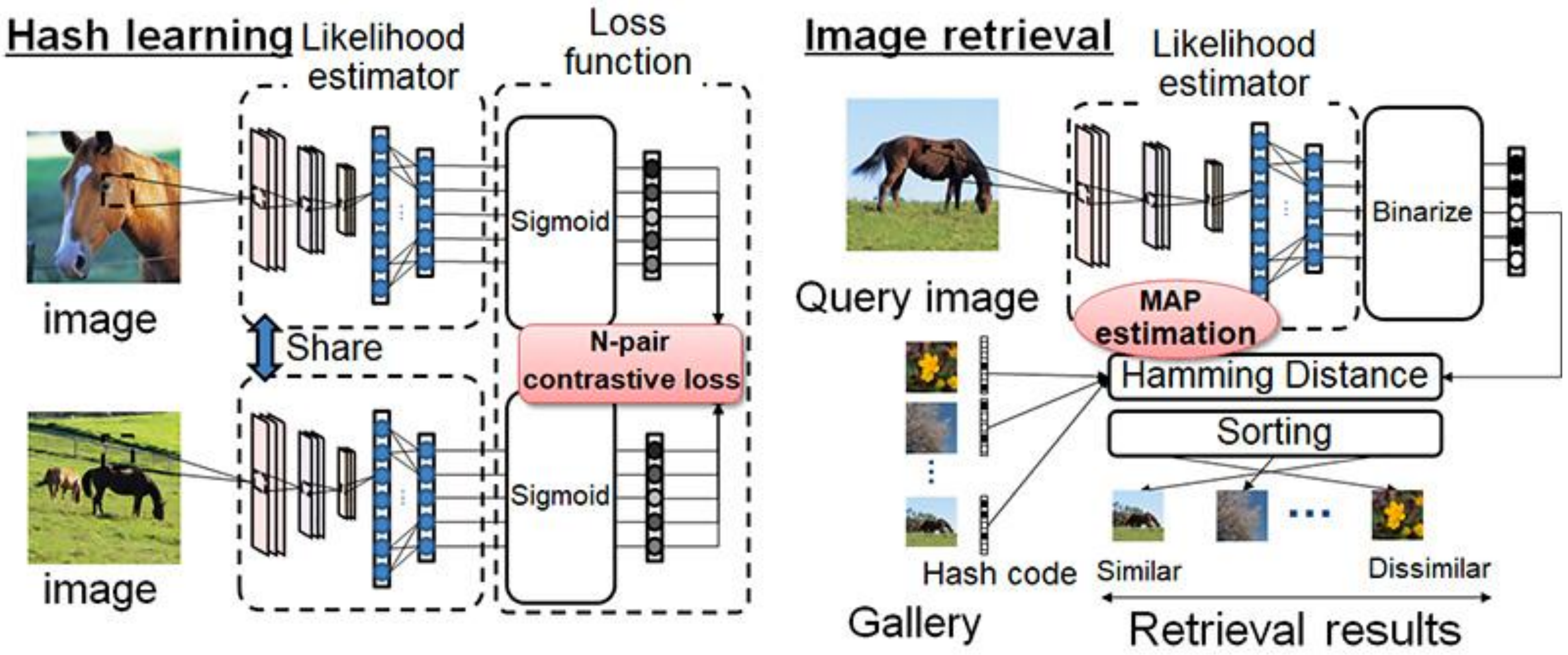
Part 1	Introduction
Part 2	Metric learning for face recongition
Part 3	Hamming Deep Metric Learning

- What is Hamming DML?

  - Hamming deep metric learning
    Input -> Deep neural network -> Binary embedding



- What is Hamming DML?



- Why Binary?

> Considering an online image searching system:
- Offline: training model, gallery features, extraction, storage
- Online: probe feature extraction, matching
- Hamming DML presents high storage efficiency and matching speed

> Lightweight models:
- efficient for training and feature extraction

> Heavyweight models:
- strong discriminative power

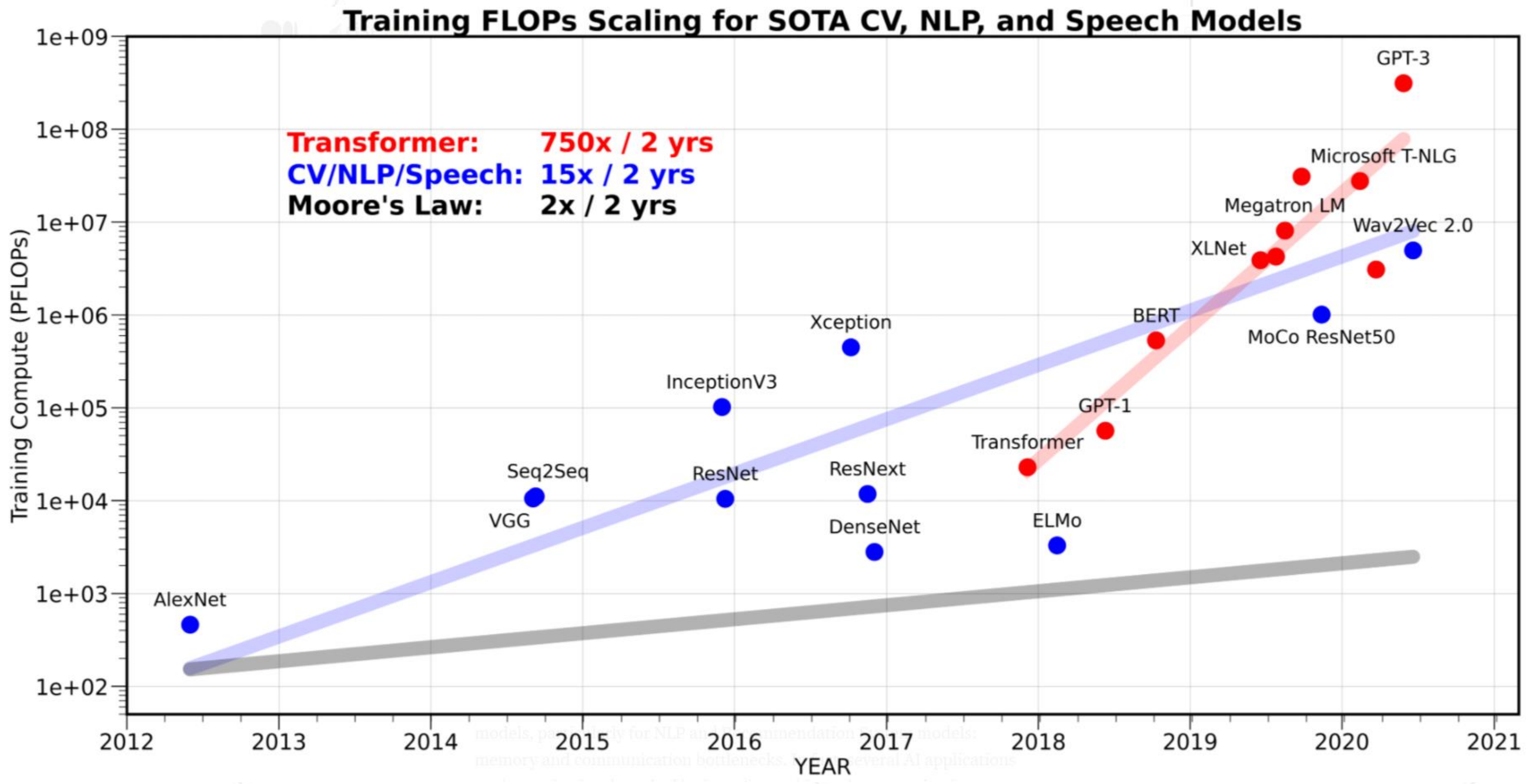
13.2 Self-supervised Learning

Dr. Liu Yu

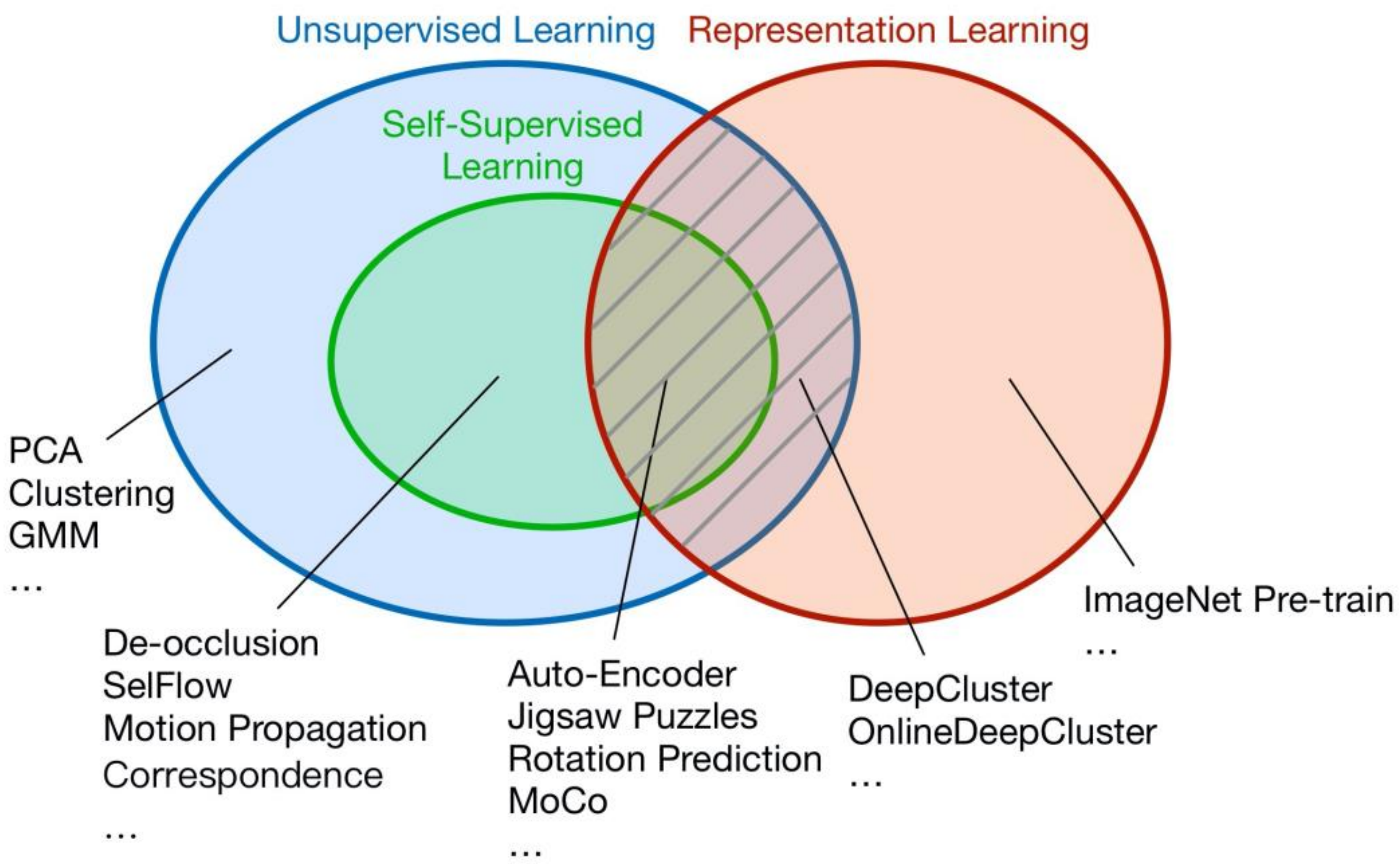
Thursday, May 20, 2021

Outline

- Learn visual representation from images without annotations.
  - Motivated by the success of large-scale pretraining in NLP.

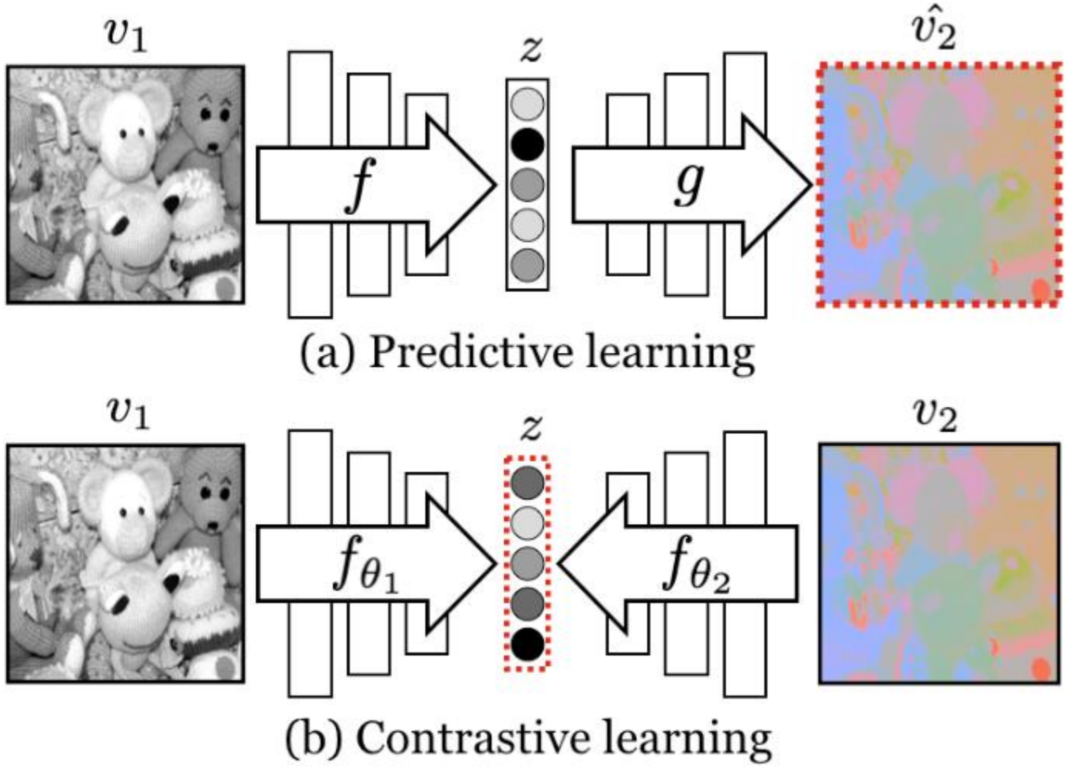


Introduction



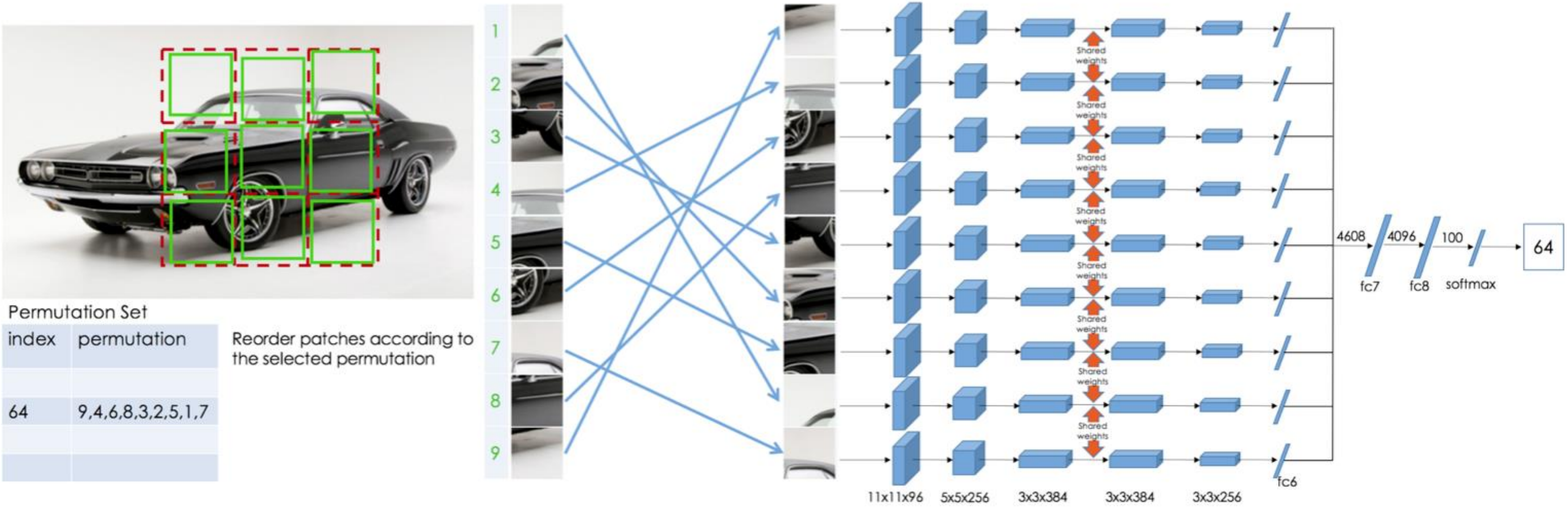
https://github.com/open-mmlab/OpenSelfSup

- Design learning tasks without annotations:
  - Predictive methods
    - VAE, GAN, …
  - Contrastive methods
    - SimCLR, MOCO, …
  - Others
    - predicting rotation
    - solving jigsaw puzzles



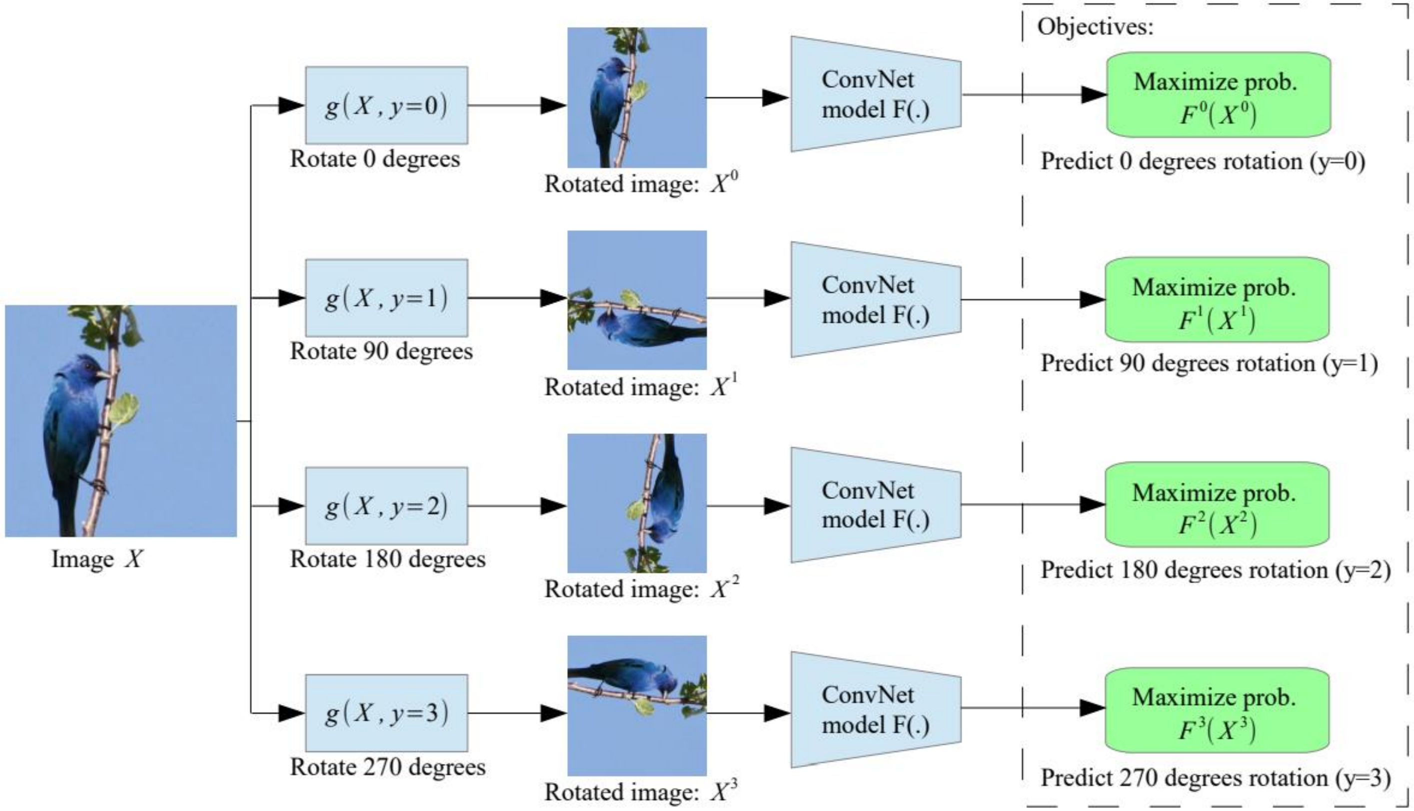
Tian Y, Krishnan D, Isola P. Contrastive multiview coding[J]. arXiv preprint arXiv:1906.05849, 2019.

- Solving jigsaw puzzles.



Noroozi M, Favaro P. Unsupervised learning of visual representations by solving jigsaw puzzles[C]//European conference on computer vision. Springer, Cham, 2016: 69-84.

- Pridicting rotation.



Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations[J]. arXiv preprint arXiv:1803.07728, 2018.

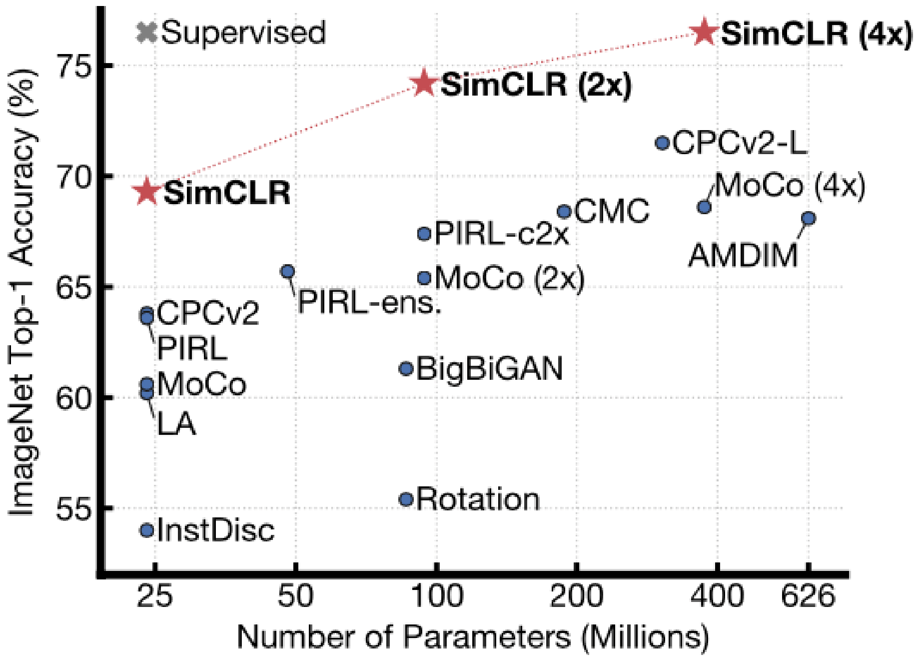
商汤 sensetime

清華大學 Tsinghua University

Outline

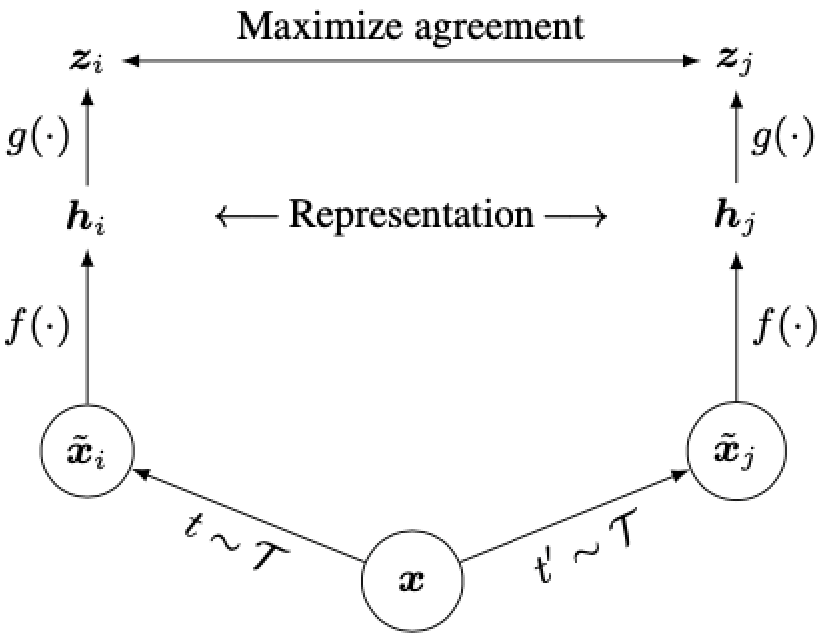
Representative Methods

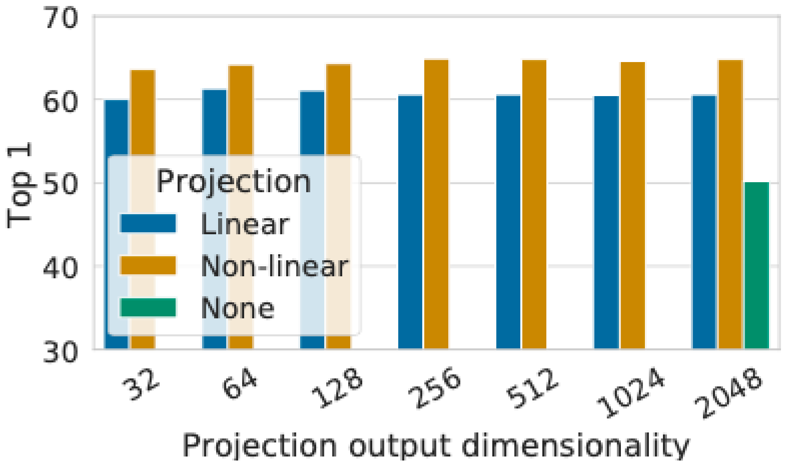
- SimCLR
- A cornerstone for SSL
- Principle
  - The representations from the same image should be near
  - The representations from different images should be far away from each other



Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C] // International conference on machine learning. PMLR, 2020: 1597-1607.

- Key insights:
  - Composition of multiple data augmentation operations
  - Introducing a learnable nonlinear transformation between the representation and the contrastive loss
  - Normalized embeddings and an appropriately adjusted temperature parameter
  - Larger batch sizes and longer training

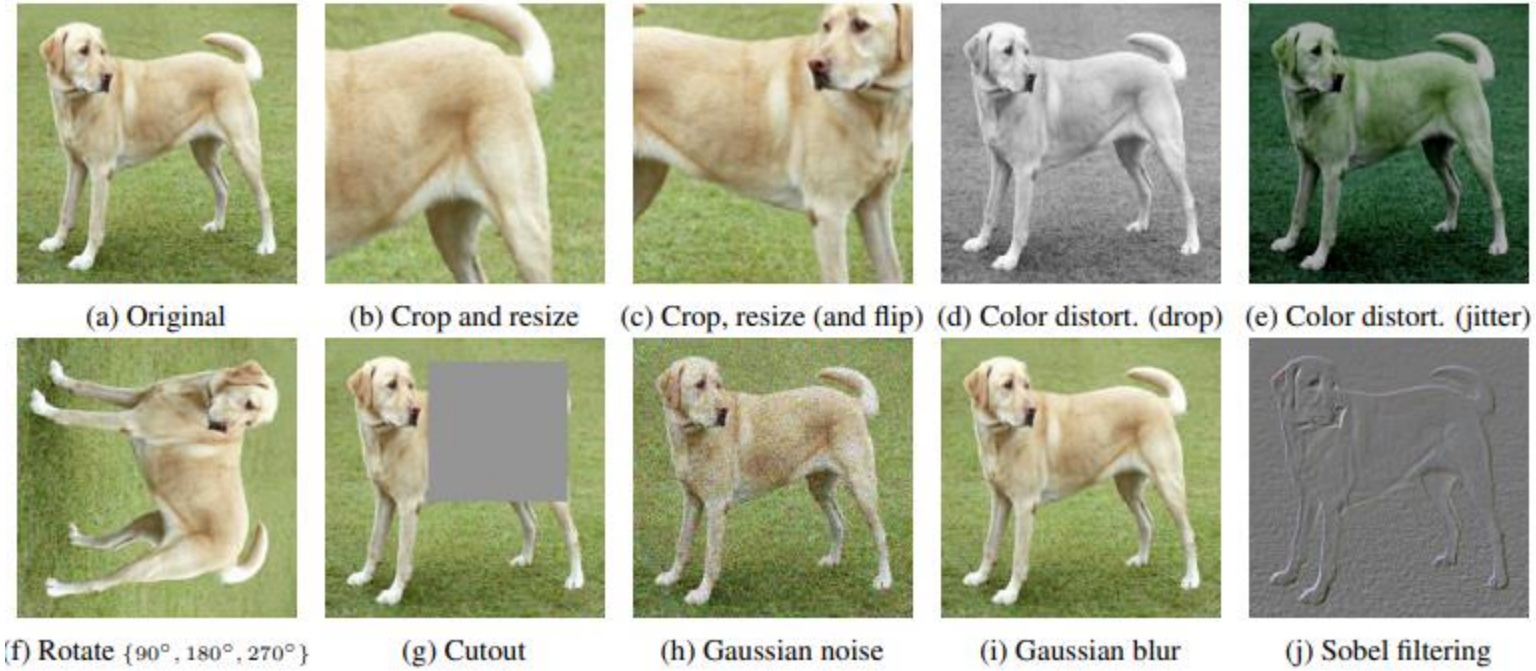




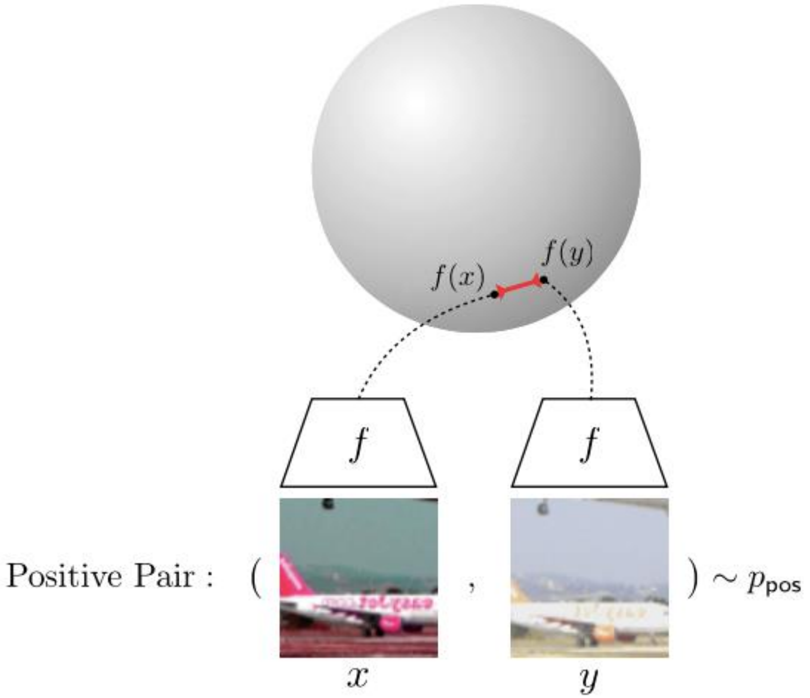
- SimCLR

Then the loss function for a positive pair of examples (i, j) is defined as:

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$



- What the representations look like?

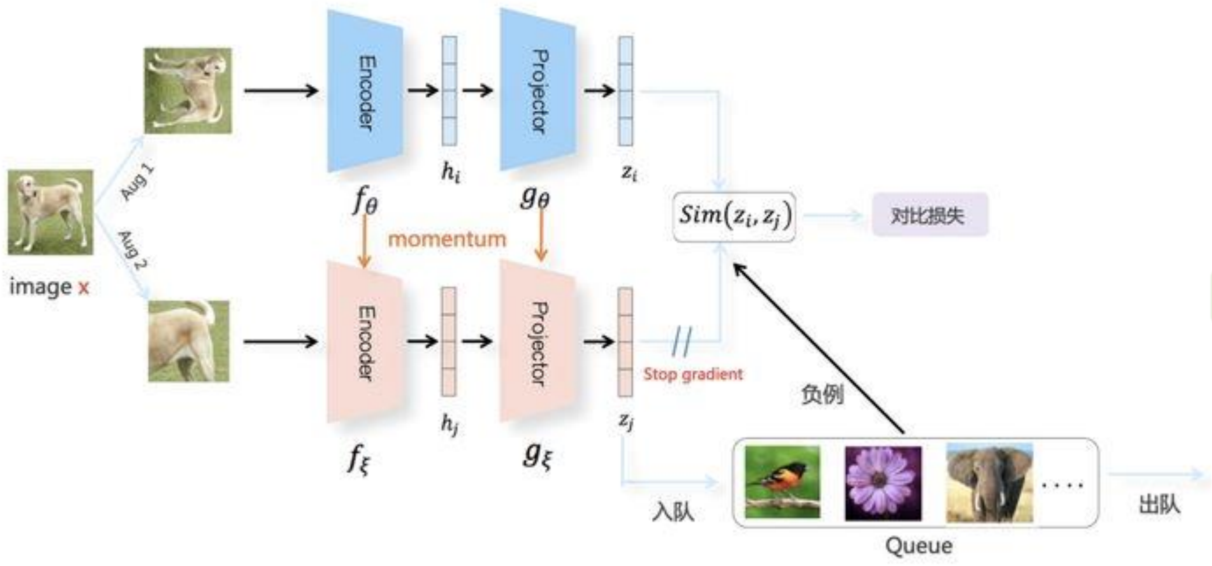


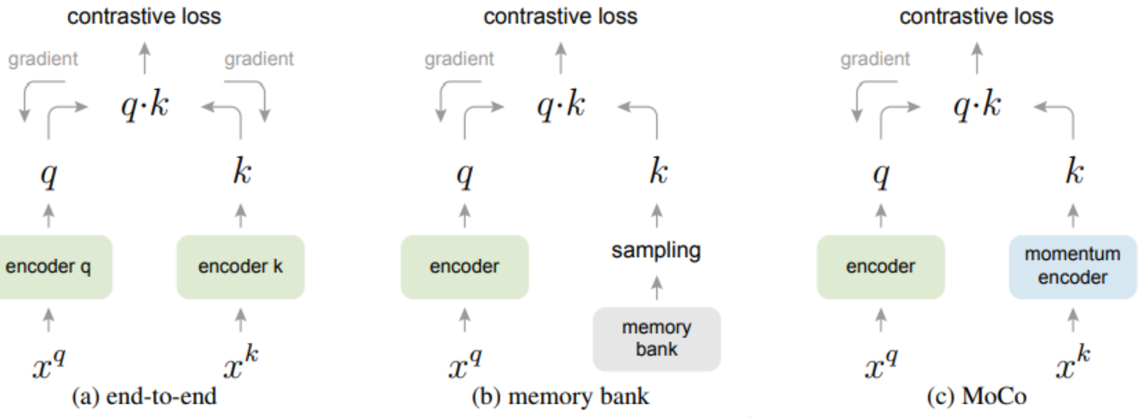
Alignment: Similar samples have similar features.

Uniformity: Preserve maximal information.

Wang T, Isola P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere[C]//International Conference on Machine Learning. PMLR, 2020: 9929-9939.

- MoCo
- Use buffer of representations to harvest more negative pairs

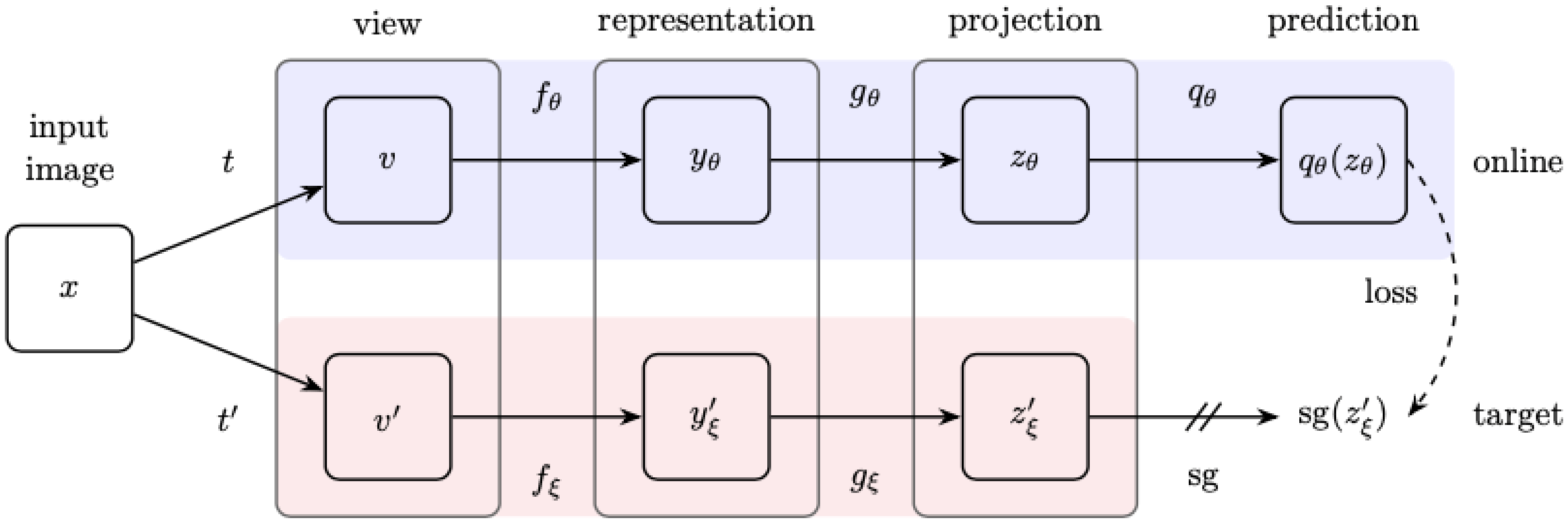




$$\theta_k \leftarrow m\theta_k + (1 - m)\theta_q.$$

Chen X, Fan H, Girshick R, et al. Improved baselines with momentum contrastive learning[J]. arXiv preprint arXiv:2003.04297, 2020.

- BYOL
- Negative samples may be semantically similar
  - How to avoid model collapse after discarding negative pairs?

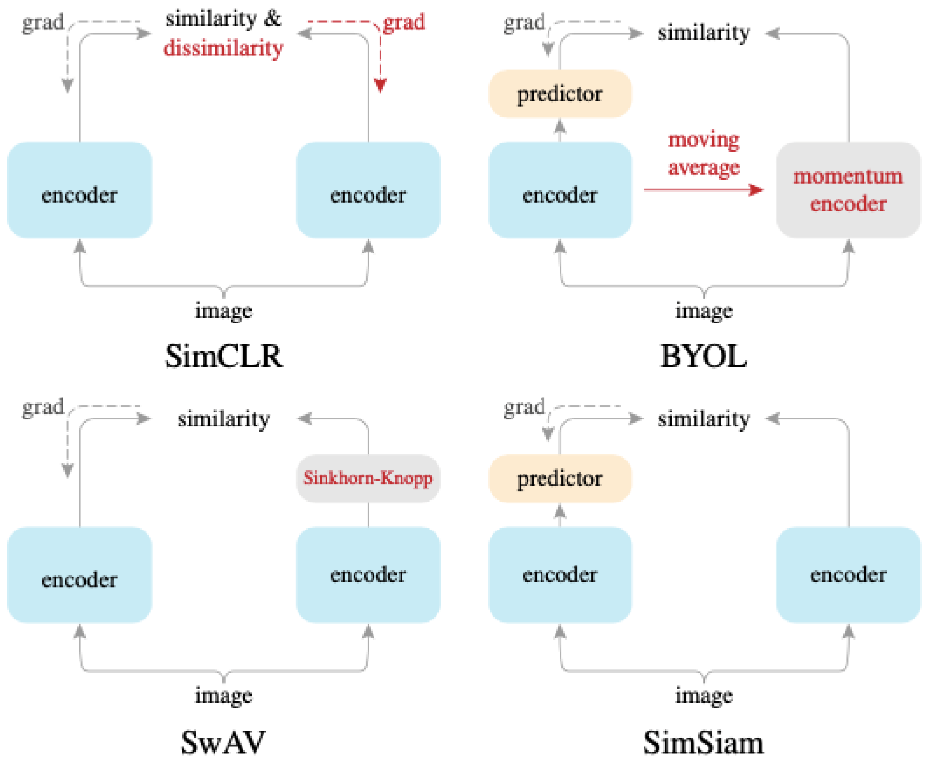


Grill J B, Strub F, Altché F, et al. Bootstrap your own latent: A new approach to self-supervised learning[J]. arXiv preprint arXiv:2006.07733, 2020.

Representative Methods

- SimSiam
- The key component to avoid model collapse is stop-gradient

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



Chen X, He K. Exploring Simple Siamese Representation Learning[J]. arXiv preprint arXiv:2011.10566, 2020.

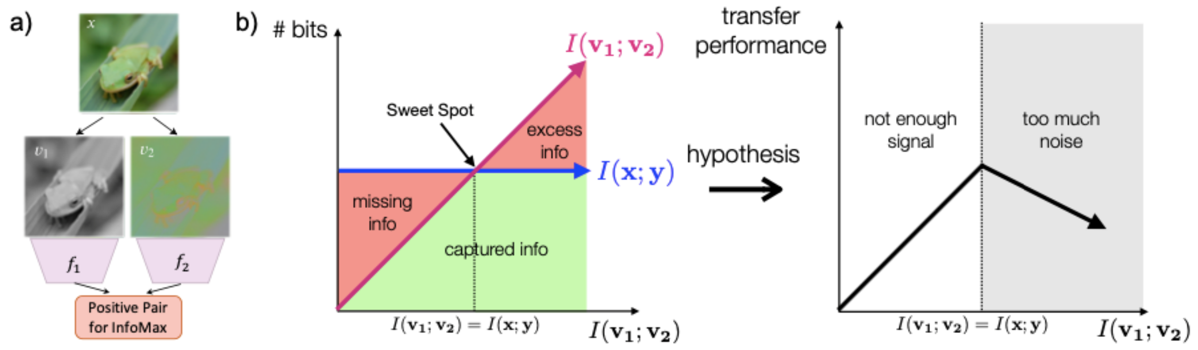
商汤 sensetime

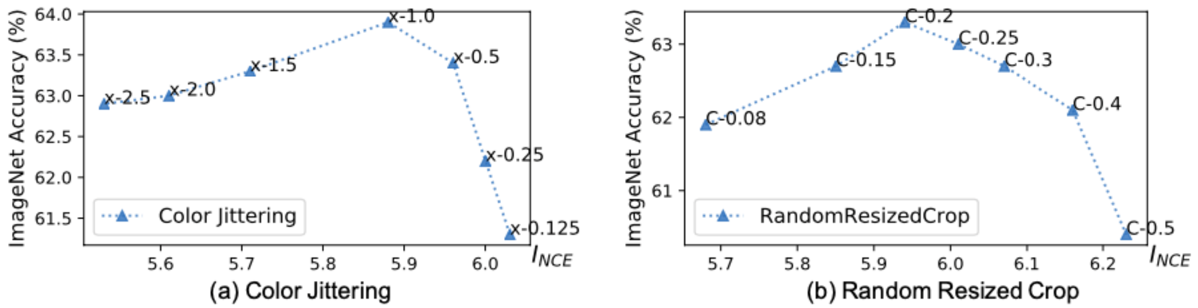
清華大学 Tsinghua University

Outline

Relative Research

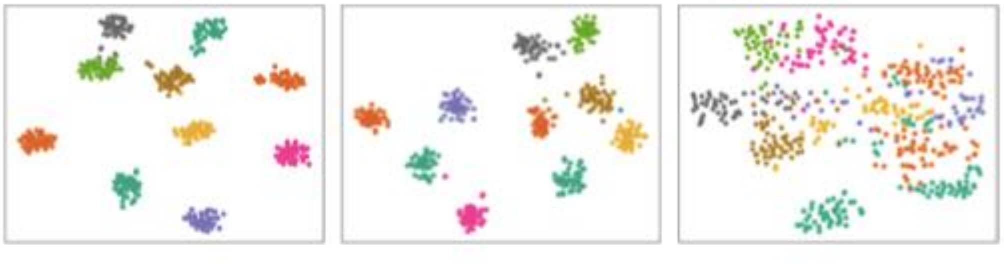
- What augmentation should we use in SSL?
  - A tradeoff between missing info and excess info
  - Learnable augmentation in an adversarial way



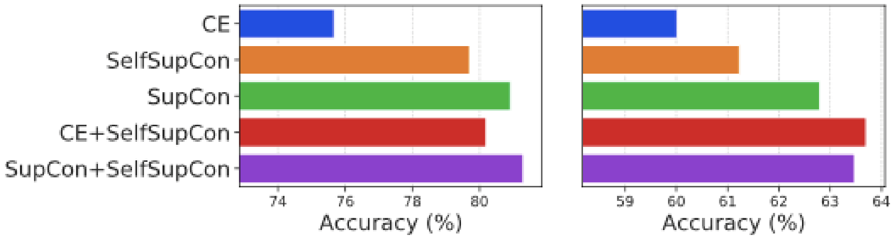


Tian Y, Sun C, Poole B, et al. What makes for good views for contrastive learning[J]. arXiv preprint arXiv:2005.10243, 2020.

- How is the transferability of SSL?
  - Compared with supervised learning, the representations show more intra-variance.
  - Combining SL with SSL improves final performance.



(a) CE (b) CE+SelfSupCon (c) SelfSupCon (d) SupCon (e) SupCon+SelfSupCon

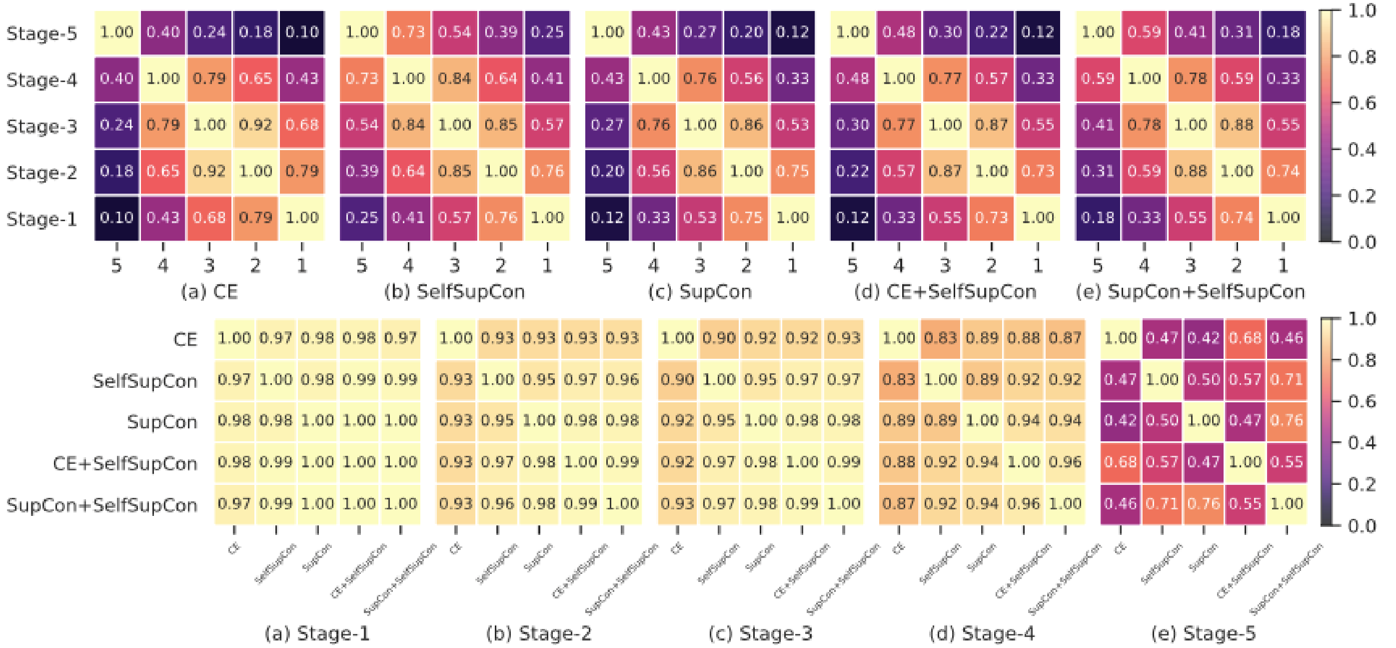


(a) Linear evaluation (b) Few-shot classification

Islam A, Chen C F, Panda R, et al. A Broad Study on the Transferability of Visual Representations with Contrastive Learning[J]. arXiv preprint arXiv:2103.13517, 2021.

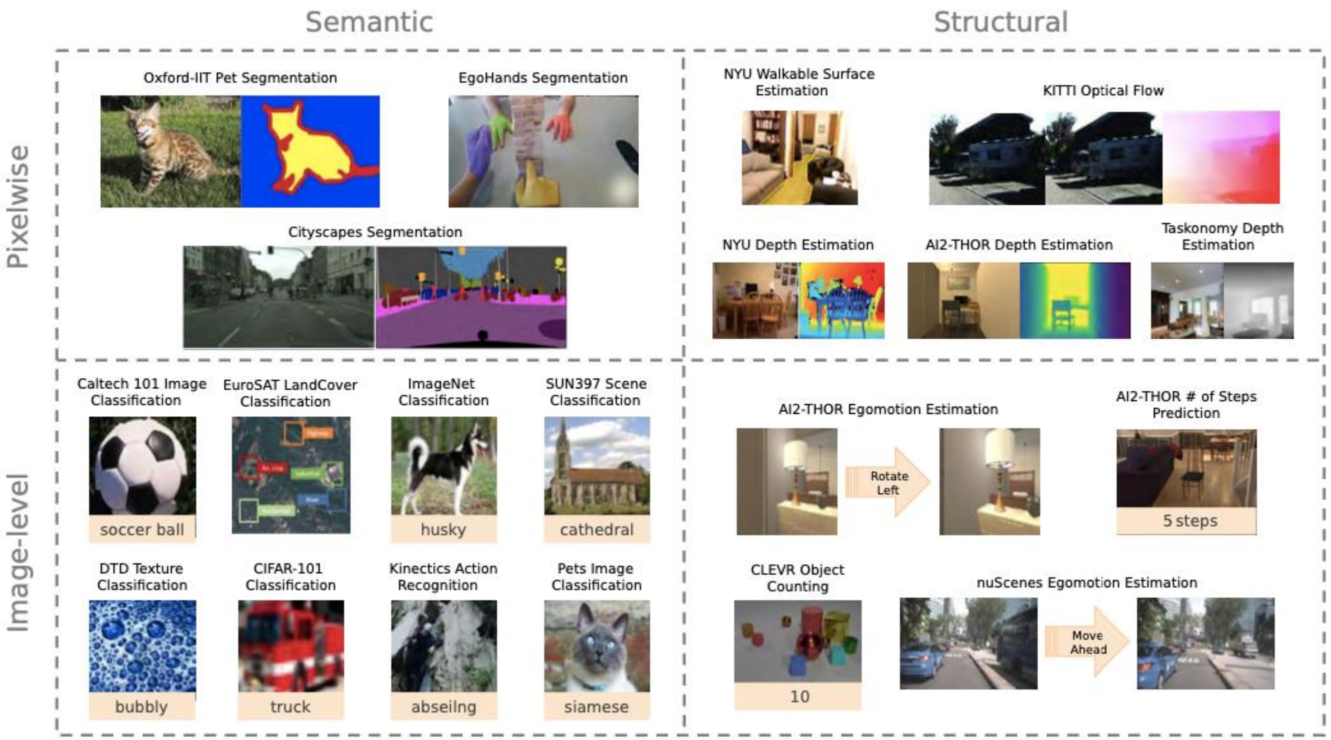
Relative Research

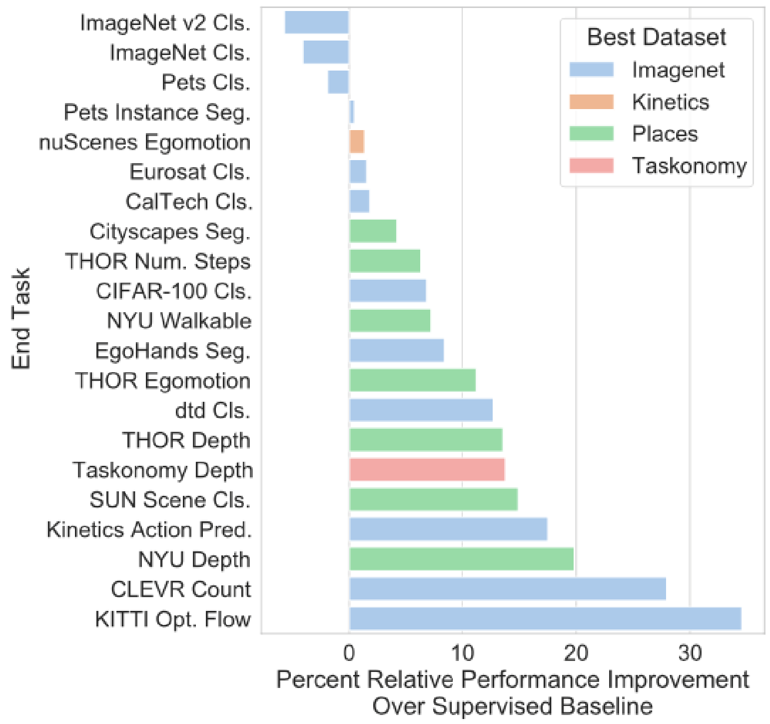
- SSL approaches learn more low/mid-level feature
  - The similarity of different layers' weight learned in SSL is higher
  - The similarity between weights of SL and SSL is low only in stage-5



Relative Research

- SSL show higher transferability in most down-stream tasks.

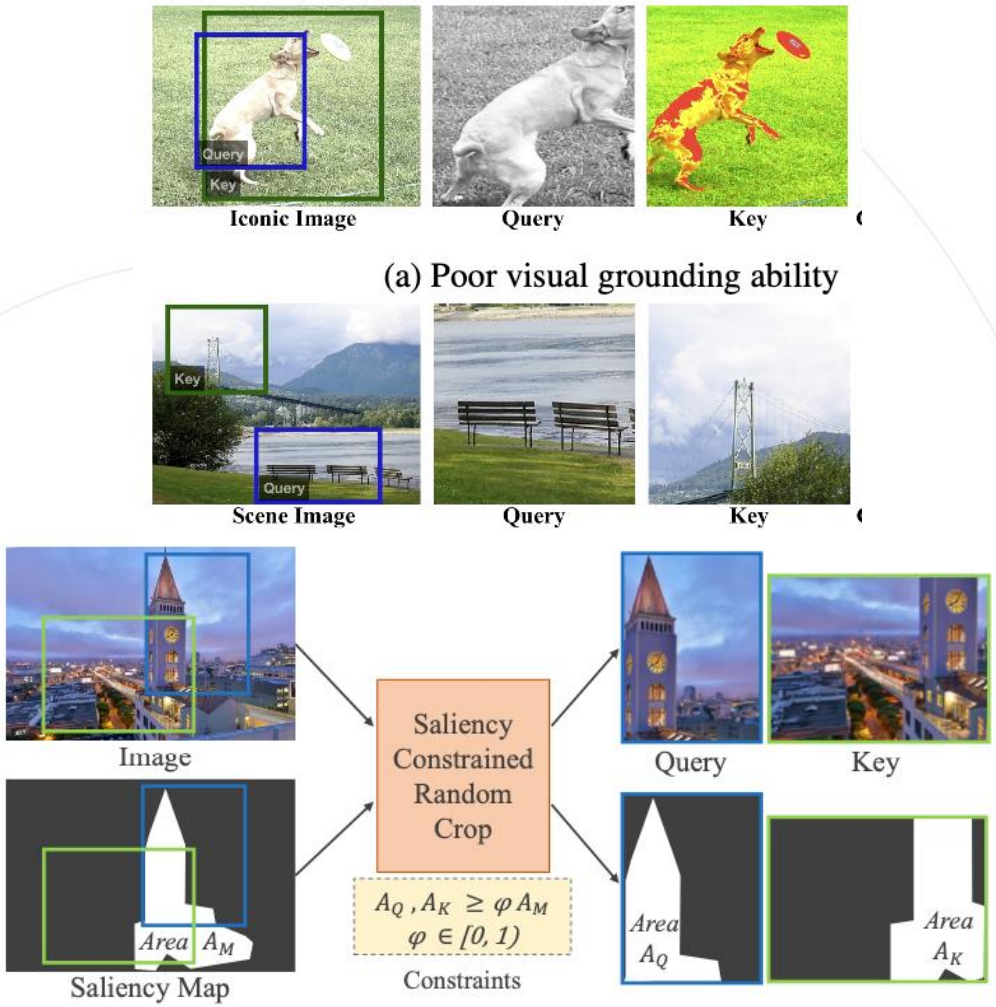




Kotar K, Ilharco G, Schmidt L, et al. Contrasting Contrastive Self-Supervised Representation Learning Models[J]. arXiv preprint arXiv:2103.14005, 2021.

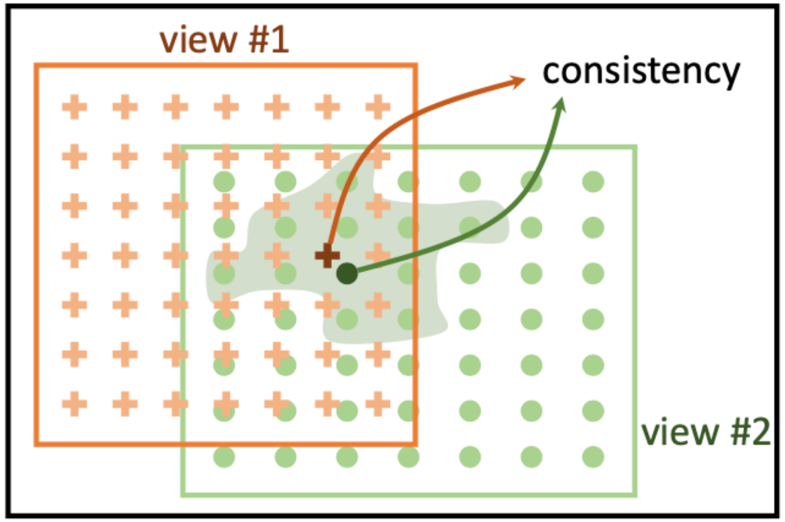
Relative Research

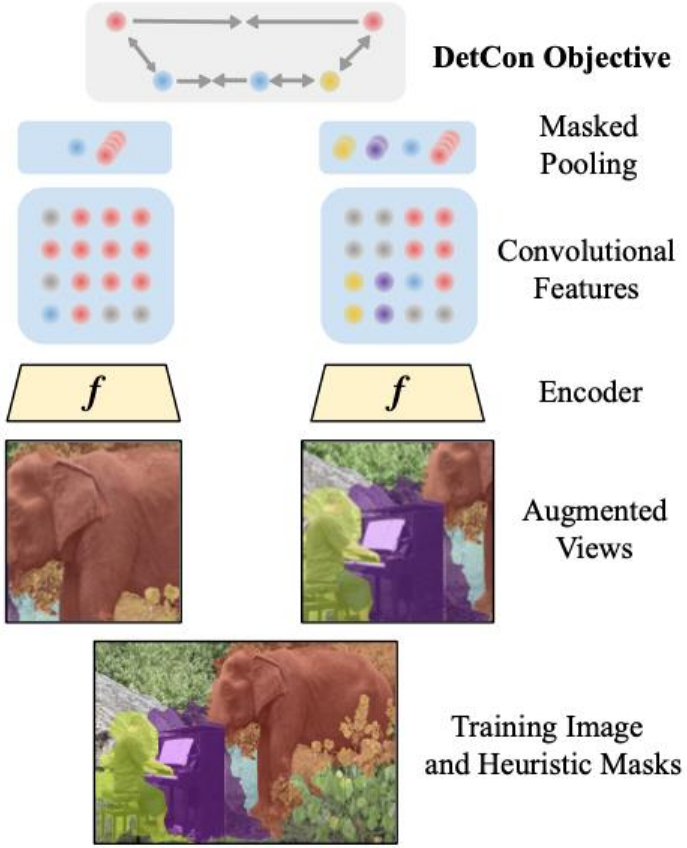
- Dataset bias in SSL
  - Augmented crops from the same image may be semantically different
  - Images in ImageNet are iconic and object-centric
  - Unsupervised saliency map can be used to guide the crops



Selvaraju R R, Desai K, Johnson J, et al. CASTing Your Model: Learning to Localize Improves Self-Supervised Representations[J]. arXiv preprint arXiv:2012.04630, 2020.

- Dense contrastive learning for dense prediction
  - Contrast representations in patch or pixel level





Hénaff O J, Koppula S, Alayrac J B, et al. Efficient Visual Pretraining with Contrastive Detection[J]. arXiv preprint arXiv:2103.10957, 2021.
Xie Z, Lin Y, Zhang Z, et al. Propagate Yourself: Exploring Pixel-Level Consistency for Unsupervised Visual Representation Learning[J]. arXiv preprint arXiv:2011.10043, 2020.

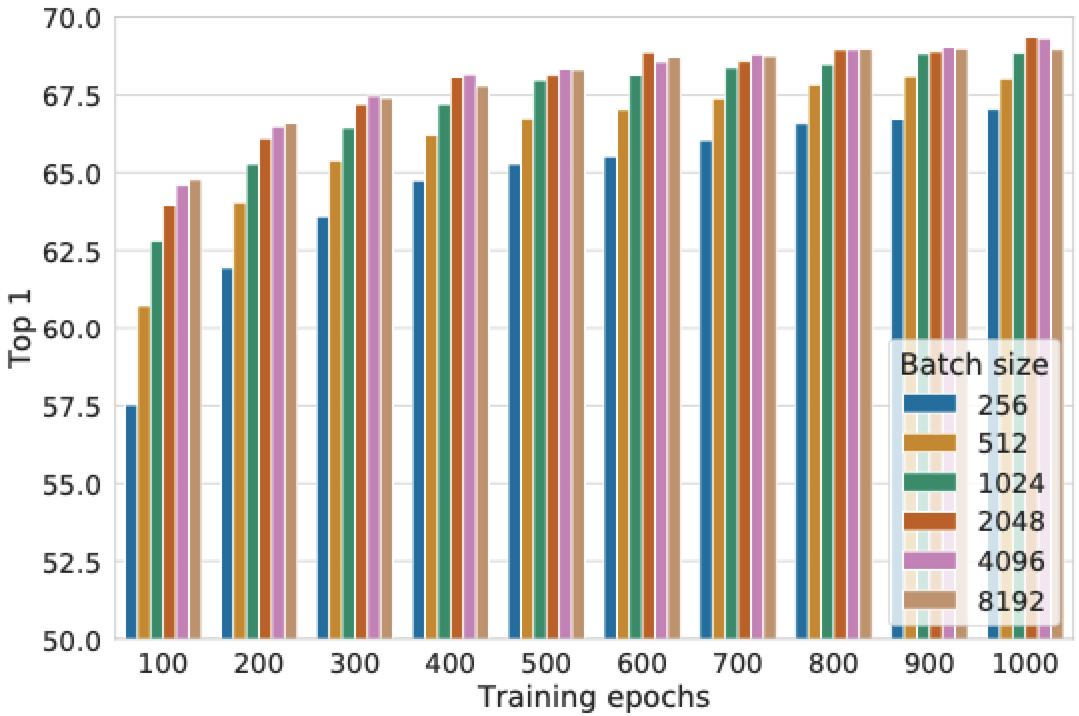
商汤 sensetime

清華大學 Tsinghua University

Outline

Challenges

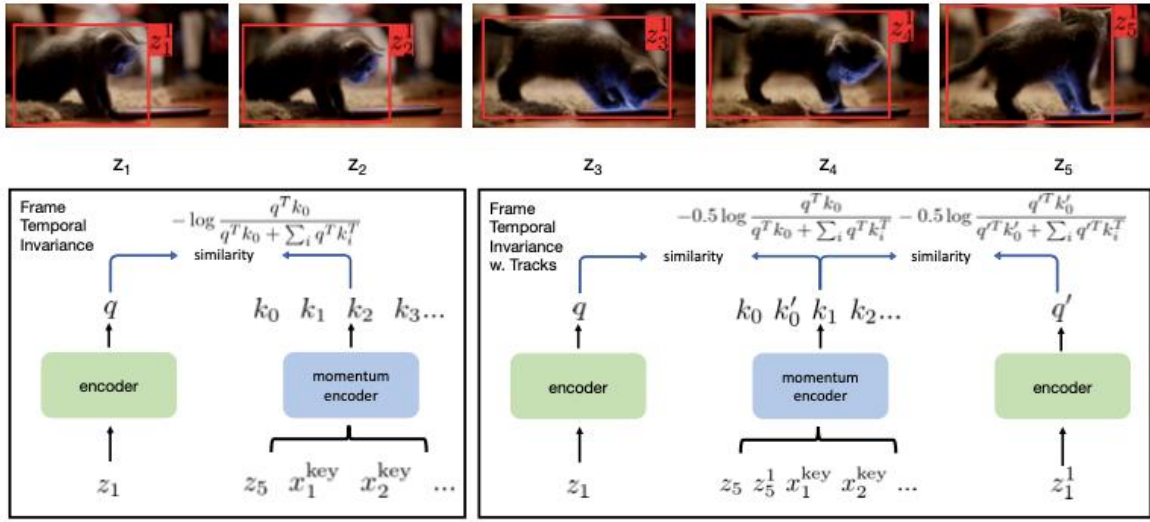
- Performance highly depends on large epochs and batch-size
  - Typically 800 epochs and 4096 batch-size for ImageNet
  - SimCLR:



Hénaff O J, Koppula S, Alayrac J B, et al. Efficient Visual Pretraining with Contrastive Detection[J]. arXiv preprint arXiv:2103.10957, 2021.

Challenges

- Only utilize augmentation-invariance in SSL.
  - SSL is better only at occlusion invariance
  - Utilize video with unsupervised tracking to harvest images under different views.



Dataset	Method	Occlusion Top-10	Occlusion Top-25	Viewpoint Top-10	Viewpoint Top-25	Illumination Dir. Top-10	Illumination Dir. Top-25	Illumination Color Top-10	Illumination Color Top-25	Instance Top-10	Instance Top-25	Instance+Viewpoint Top-10	Instance+Viewpoint Top-25
Imagenet	Sup. R50	80.89	74.21	89.54	82.62	94.63	89.08	99.88	99.38	66.11	59.44	70.17	63.47
Imagenet	MOCOv2	84.19	77.88	85.15	75.08	90.28	80.76	99.66	97.11	62.49	55.01	67.4	60.52
Imagenet	PIRL	84.46	78.38	85.8	76.08	87.7	78.45	99.68	97.19	52.97	46.79	57.01	51.03

Zbontar J, Jing L, Misra I, et al. Barlow twins: Self-supervised learning via redundancy reduction[J]. arXiv preprint arXiv:2103.03230, 2021.