Workshop 1 小样本 self-supervision + few shot learning

Few shot : transfer, 小样本, 快速迁移
↳ 每类图少
↳ 类内 diversity 大

Does SSL Benefit FSL?

- Self-supervised learning (SSL) aims to learn transferable presentation from a large set of unlabeled data:
  - Pretext task-based SSL
  - Contrastive learning
- Both SSL and FSL aim to transfer knowledge learned from the data of a set of seen tasks to a new unseen one:
  - Integrate SSL into FSL --> learn more transferable knowledge?

Summary and New Results

- New results for meta-learning:
  - ✓ With the PAC-Bayes framework, we derive a fast-rate generalization bound for FSL with DNN. For the first time, we rigorously demonstrate why an FSL model works well on novel tasks. (submitted to AAAI 2022 and ICLR 2022)
  - ✓ Large-scale multi-modal pre-training (like OpenAI CLIP and Wenlan-BriVL) may be a possible solution to FSL. The challenge changes to fine-tuning these pre-trained models for FSL.

第二项工作: 对比学习

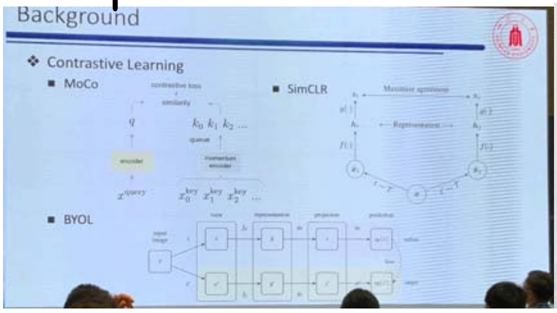
Sub-Conclusions

- We have proposed a novel contrastive prototype learning with augmented embedding (CPLAE) model to address the lack of training data problem in FSL.
- This work shows for the first time that contrastive learning is effective under the supervised and few-shot learning setting.
- Different from existing embedding-based meta-learning methods, we introduce both data augmentation to form an augmented embedding space and a support set prototype centered loss to complement the conventional query centered loss.
- Extensive experiments on three widely used benchmarks demonstrate that our CPLAE achieves new state-of-the-art.

Summary and New Results

- Take-home message about few-shot learning (FSL):
  - ✓ Episode-level self-supervised learning (SSL) is more effective than instance-level SSL.
  - ✓ Contrastive prototype learning brings more benefits to FSL than classic contrastive learning.
  - ✓ Concatenating multiple data augmentations can significantly improve the SOTA methods.

Workshop 2 川大 彭玺

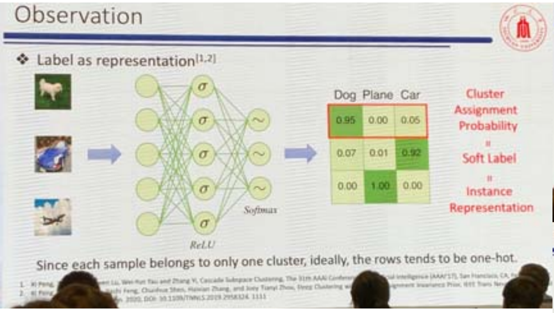


Background

- Contrastive Learning
  - MoCo
  - SimCLR
  - BYOL

~~第二项工作: 聚类~~

cluster 表示形式? 行空间/列空间上 均作对比学习



Observation

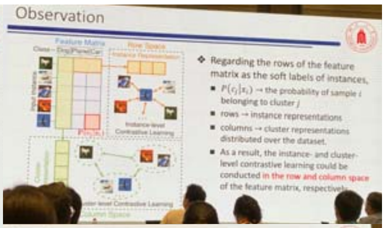
- Label as representation[1,2]

Dog	Plane	Car
0.95	0.00	0.05
0.07	0.01	0.92
0.00	1.00	0.00

Cluster Assignment Probability
=
Soft Label
=
Instance Representation

Since each sample belongs to only one cluster, ideally, the rows tends to be one-hot.

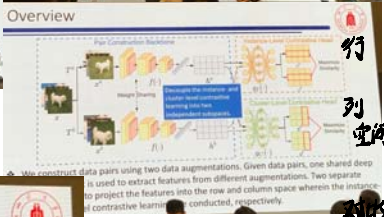
[1] Xi Peng, ... Liu, Wei-Yun Yau and Zhang Yi, Cascade Subspace Clustering, The 31th AAAI Conference ... Intelligence (AAAI'17), San Francisco, CA, ...
[2] Xi Peng, ... Jiashi Feng, Chunhua Shen, Hainan Zhang, and Joey Tianyi Zhou, Deep Clustering w... ...ignment Invariance Prior, IEEE Trans Neu... 2020, DOI: 10.1109/TNNLS.2019.2958324. 1111



Observation

Feature Matrix
Row Space
Instance Representation
Instance-level Contrastive Learning
Cluster-level Contrastive Learning
Column Space

- Regarding the rows of the feature matrix as the soft labels of instances,
  - $P(c_j|x_i)$ → the probability of sample i belonging to cluster j
  - rows → instance representations
  - columns → cluster representations distributed over the dataset.
  - As a result, the instance- and cluster-level contrastive learning could be conducted in the row and column space of the feature matrix, respectively.

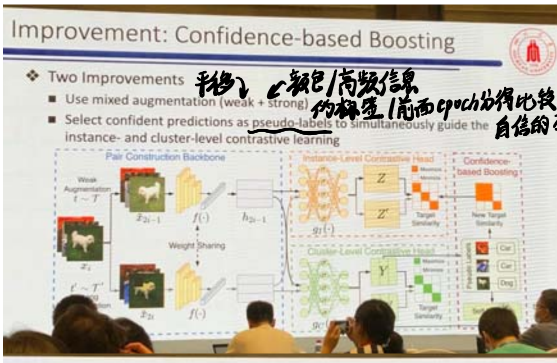


Overview

Pair Construction Backbone
Instance-level Contrastive Head
Cluster-level Contrastive Head
Decouple the instance- and cluster-level contrastive learning into two independent subspaces.
Weight Sharing
Maximize Similarity

- We construct data pairs using two data augmentations. Given data pairs, one shared deep ... is used to extract features from different augmentations. Two separate ... to project the features into the row and column space wherein the instance- ...el contrastive learnin... conducted, respectively.

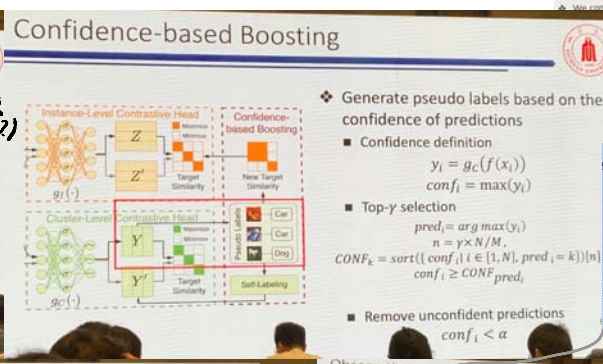
行
列
空间同时
作
对比学习?



Improvement: Confidence-based Boosting

- Two Improvements
  - Use mixed augmentation (weak + strong)
  - Select confident predictions as pseudo-labels to simultaneously guide the instance- and cluster-level contrastive learning

平移) ↙颜色/高频信息
伪标签(前面epoch分得比较自信的?)



Confidence-based Boosting

- Generate pseudo labels based on the confidence of predictions
  - Confidence definition

$$y_i = g_C(f(x_i))$$
$$conf_i = \max(y_i)$$

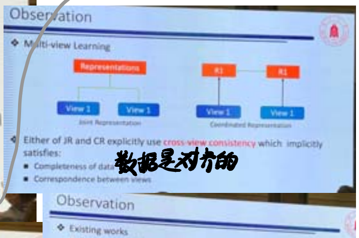
  - Top-γ selection

$$pred_i = arg\ max(y_i)$$
$$n = \gamma \times N/M,$$
$$CONF_k = sort(\{conf_i | i \in [1,N], pred_i = k\})[n]$$
$$conf_i \geq CONF_{pred_i}$$

  - Remove unconfident predictions

$$conf_i < \alpha$$

第二项工作: 多模态方法



Observation

- Multi-view Learning

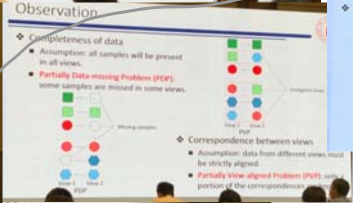
- Either of JR and CR explicitly use cross-view consistency which implicitly satisfies:
  - Completeness of data
  - Correspondence between views

数据是对齐的

Improvement

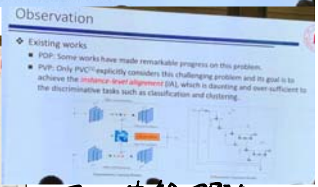
- Deep Clustering (JULE, DC, etc.)
  - Two-stage
    - error accumulation during the alternation
  - Offline
    - need entire dataset to perform clustering
    - limited application on large datasets
- Contrastive Learning (MoCo, SimCLR, BYOL)
  - Instance-level
  - General representation
    - off-the-shelf for downstream tas...

- Contrastive Clustering (Ours)
  - One-stage
    - end-to-end
  - Online 在线聚类
    - batch-wise optimization and cluster prediction
  - Both instance- and cluster-level
    - dual contrastive learning
  - Task-specified
    - clustering-oriented



Observation

- Completeness of data
  - Assumption: all samples will be present in all views.
  - Partially Data-missing Problem (PDP): some samples are missed in some views.
- Correspondence between views
  - Assumption: data from different views must be strictly aligned.
  - Partially View-aligned Problem (PVP): only a portion of the correspondences ...



Observation

- Existing works
  - PDP: Some works have made remarkable progress on this problem.
  - PVP: Only PVC[1] explicitly considers this challenging problem and its goal is to achieve the instance-level alignment (IA), which is daunting and over-sufficient to the discriminative tasks such as classification and clustering.

"最优传输问题"

instance level 对齐
→ category level 对齐

Motivation

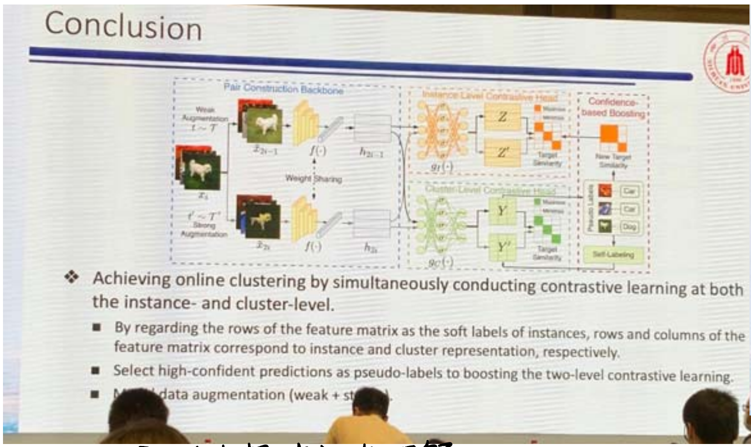
- Our basic idea
  - Category-level Alignment (CA), which embraces higher accessibility. Intuitively, an instance have a probability of 1/K vs. 1/N to be aligned in CA than IA.

Key Idea

- We reformulate CA as an Categorial I... (CI) task, which could be achieved by contrastive learning.
- Under the unlabeled setting, we ... using the available aligned data and the Negative Pairs (NPs) using random sampling.

正负样例对

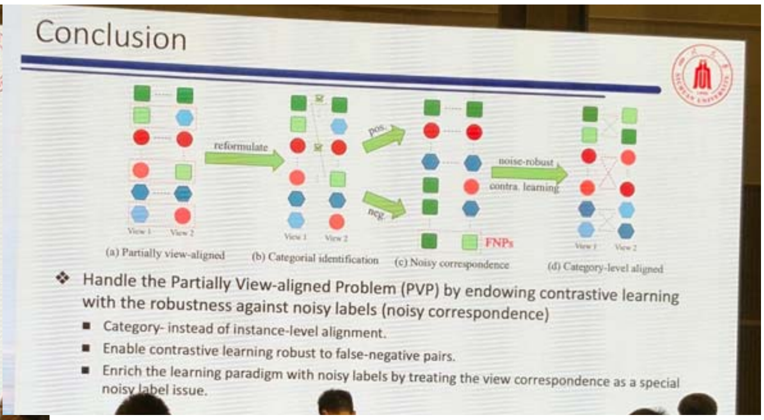
随机→假阴性



Conclusion

Pair Construction Backbone
Weak Augmentation $t \sim \mathcal{T}$
Strong Augmentation $t' \sim \mathcal{T}'$
x_i
x_{2i-1}
x_{2i}
$f(\cdot)$
Weight Sharing
h_{2i-1}
h_{2i}
Instance Level Contrastive Head
Cluster Level Contrastive Head
Z
Z'
Y
Y'
$g_I(\cdot)$
$g_C(\cdot)$
Target Similarity
Confidence-based Boosting
New Target Similarity
Pseudo Labels
Car
Cat
Dog
Self-Labeling

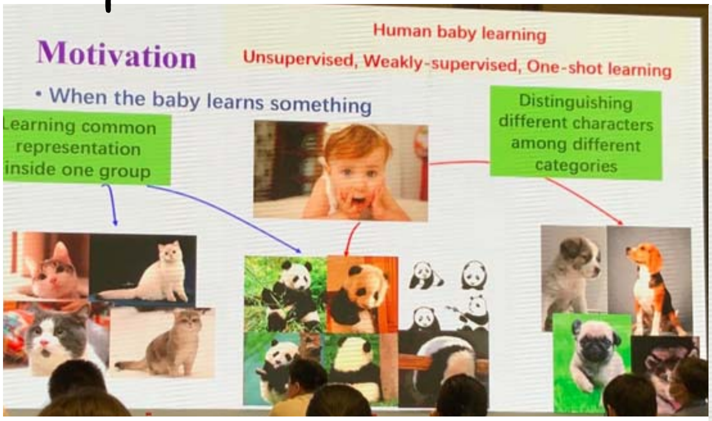
- ❖ Achieving online clustering by simultaneously conducting contrastive learning at both the instance- and cluster-level.
  - ■ By regarding the rows of the feature matrix as the soft labels of instances, rows and columns of the feature matrix correspond to instance and cluster representation, respectively.
  - ■ Select high-confident predictions as pseudo-labels to boosting the two-level contrastive learning.
  - ■ M data augmentation (weak + st).



Conclusion

- ❖ Handle the Partially View-aligned Problem (PVP) by endowing contrastive learning with the robustness against noisy labels (noisy correspondence)
  - ■ Category- instead of instance-level alignment.
  - ■ Enable contrastive learning robust to false-negative pairs.
  - ■ Enrich the learning paradigm with noisy labels by treating the view correspondence as a special noisy label issue.

Workshop 3 自监督 音视频理解.



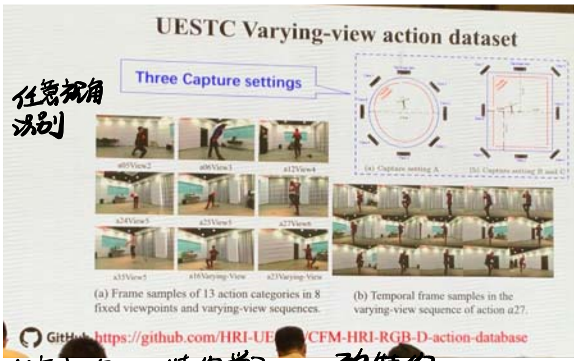
Motivation

Human baby learning

Unsupervised, Weakly-supervised, One-shot learning

- When the baby learns something

Learning common representation inside one group
Distinguishing different characters among different categories



UESTC Varying-view action dataset

任意视角
识别

(a) Frame samples of 13 action categories in 8 fixed viewpoints and varying-view sequences.

(b) Temporal frame samples in the varying-view sequence of action a27.

GitHub https://github.com/HRI-UE /CFM-HRI-RGB-D-action-database

任意视角 → 特征学习 → 一致特征.

1.外向扩展式：GAN 用一个视角生成更多视角 2.内敛 View预处理
→视角转化/统一.

Arbitrary-view human action recognition via novel-view action generation

PR, 2021

- Motivation

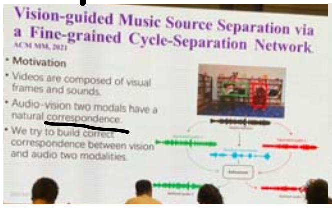
A Common representation

View-invariant Human Action Recognition via View Transformation Network (VTN)

TMM, 2021

- Motivation
  - Larger quantity of possible views provides sufficient samples for self-supervised feature learning.
  - Action features of different views have common representation inside one category.

第二支工作：音视频理解



Vision-guided Music Source Separation via a Fine-grained Cycle-Separation Network

ACM MM, 2021

- Motivation
  - Videos are composed of visual frames and sounds.
  - Audio-vision two modals have a natural correspondence.
  - We try to build correct correspondence between vision and audio two modalities.

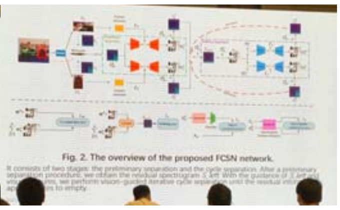
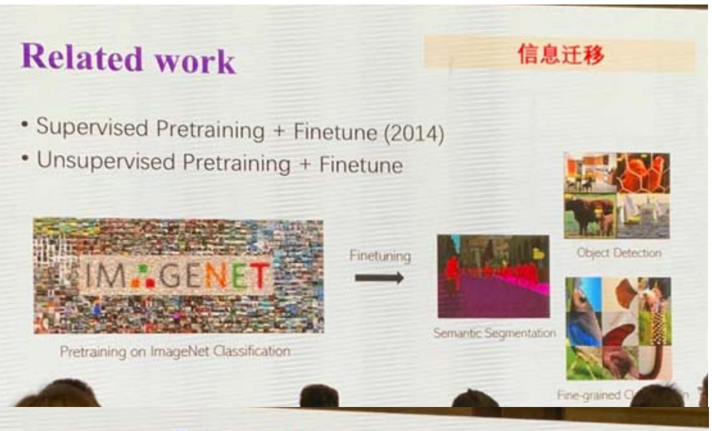


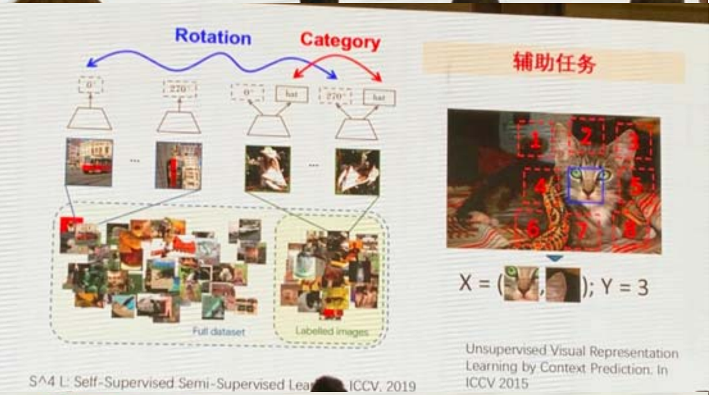
Fig. 2. The overview of the proposed FCSN network.



Related work

信息迁移

- Supervised Pretraining + Finetune (2014)
- Unsupervised Pretraining + Finetune

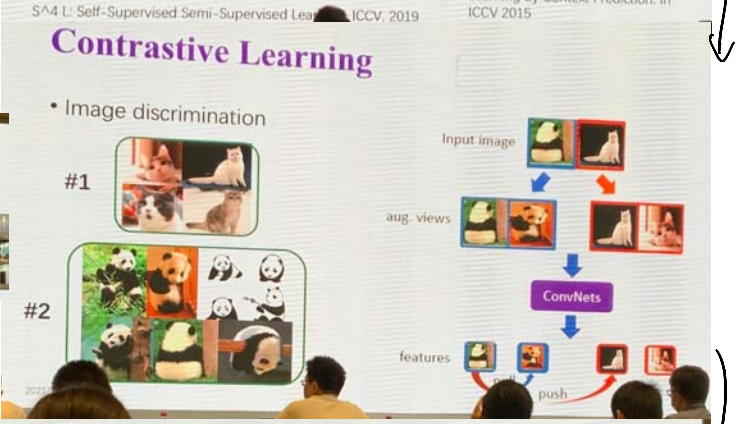


辅助任务

X = (,); Y = 3

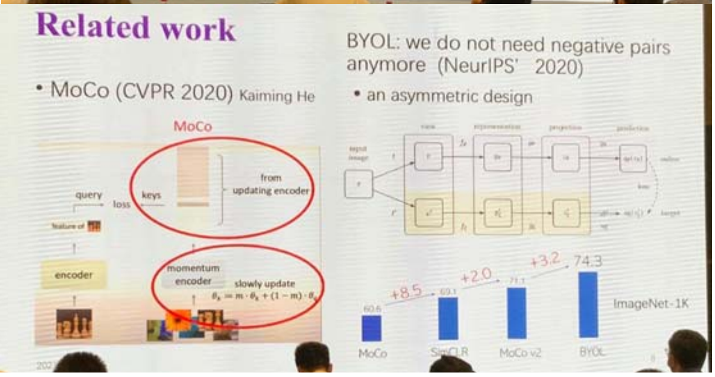
S^4 L: Self-Supervised Semi-Supervised Lea ICCV, 2019

Unsupervised Visual Representation Learning by Context Prediction. In ICCV 2015



Contrastive Learning

- Image discrimination



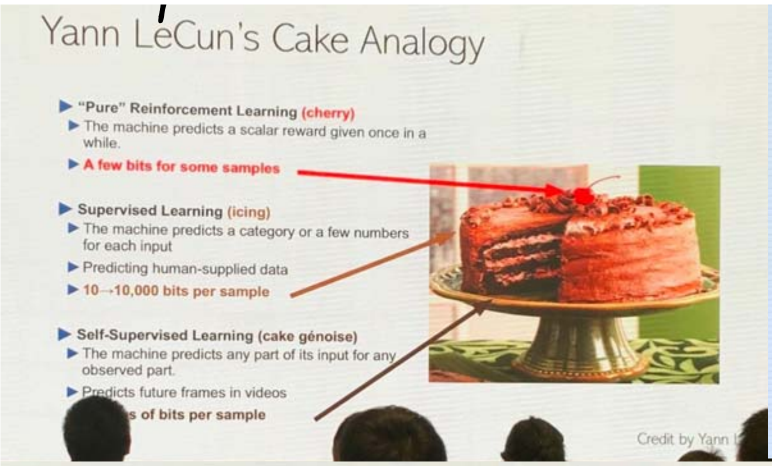
Related work

- MoCo (CVPR 2020) Kaiming He

BYOL: we do not need negative pairs anymore (NeurIPS' 2020)

- an asymmetric design

Workshop 4 胡瀚 视觉自监督.



Yann LeCun's Cake Analogy

- ▶ "Pure" Reinforcement Learning (cherry)
  - ▶ The machine predicts a scalar reward given once in a while.
  - ▶ A few bits for some samples
- ▶ Supervised Learning (icing)
  - ▶ The machine predicts a category or a few numbers for each input
  - ▶ Predicting human-supplied data
  - ▶ 10→10,000 bits per sample
- ▶ Self-Supervised Learning (cake génoise)
  - ▶ The machine predicts any part of its input for any observed part.
  - ▶ Predicts future frames in videos
  - ▶ s of bits per sample

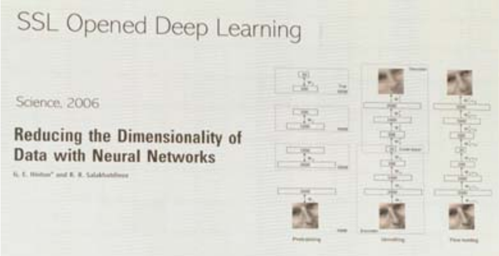
Credit by Yann L

Why Self-Supervised Learning?

- Baby learns to see the world largely by observation

Photos courtesy of Emmanuel Dupoux

Credit by Yann LeCun

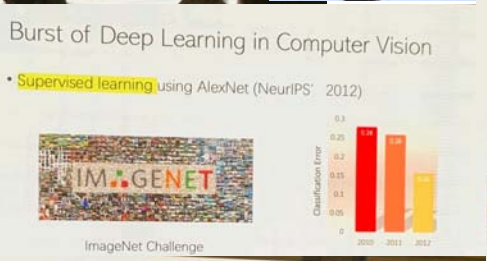


SSL Opened Deep Learning

Science, 2006

Reducing the Dimensionality of Data with Neural Networks

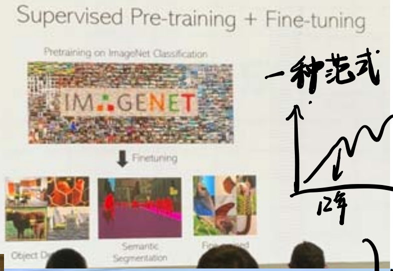
G. E. Hinton* and R. R. Salakhutdinov



Burst of Deep Learning in Computer Vision

- Supervised learning using AlexNet (NeurIPS' 2012)

ImageNet Challenge

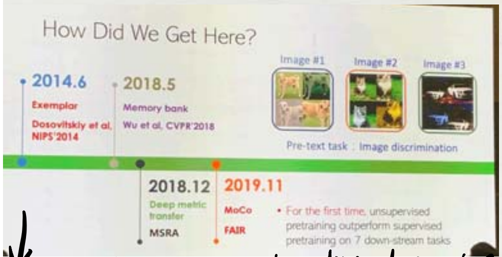


Supervised Pre-training + Fine-tuning

一种范式

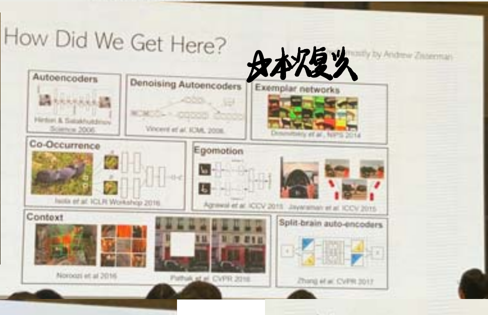
12年

↓文艺复兴



How Did We Get Here?

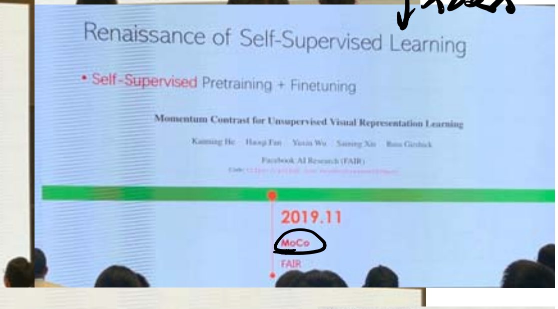
- For the first time, unsupervised pretraining outperform supervised pretraining on 7 down-stream tasks



How Did We Get Here?

☆本次复兴

Mostly by Andrew Zisserman



Renaissance of Self-Supervised Learning

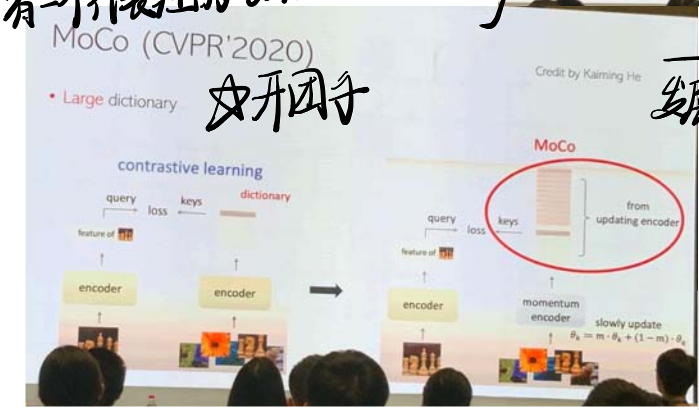
- Self-Supervised Pretraining + Finetuning

Momentum Contrast for Unsupervised Visual Representation Learning

Kaiming He Haoqi Fan Yuxin Wu Saining Xie Ross Girshick

Facebook AI Research (FAIR)

↓有一个代表性的 contrastive learning



MoCo (CVPR'2020)

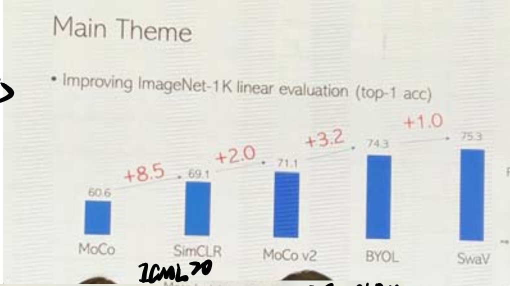
☆开团手

Credit by Kaiming He

- Large dictionary

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

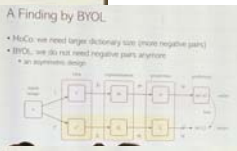
→发展变化



Main Theme

- Improving ImageNet-1K linear evaluation (top-1 acc)

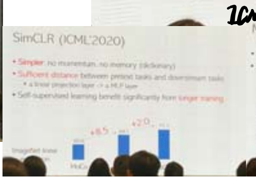
ICML'20



A Finding by BYOL

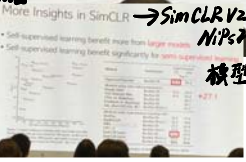
- MoCo: we need larger dictionary size (more negative pairs)
- BYOL: we do not need negative pairs anymore
  - an asymmetric design

PIC: a Single-Branch Method (NeurIPS'2020)



SimCLR (ICML'2020)

- Simpler: no momentum, no memory (dictionary)
- Sufficient distance between pretext tasks and downstream tasks
  - a linear projection layer -> a MLP layer
- Self-supervised learning benefit significantly from longer training



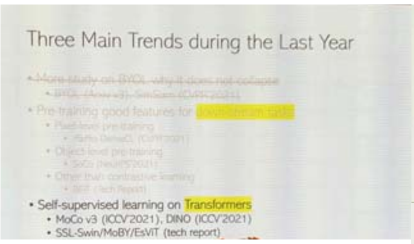
More Insights in SimCLR →SimCLR V2 NiPs'20

- Self-supervised learning benefit more from larger models
- Self-supervised learning benefit significantly for semi-supervised learning

模型越大←数据越少（半监督效果越好）

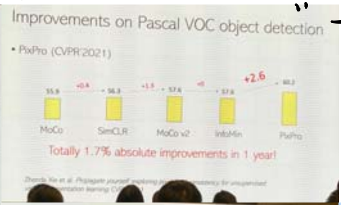
New Trend:

改善下游任务性能.



Three Main Trends during the Last Year

- More study on BYOL: why it does not collapse
  - BYOL (Arxiv v3), SimSiam (CVPR'2021)
- Pre-training good features for downstream tasks
  - Pixel-level pre-training
    - PixPro, DenseCL (CVPR'2021)
  - Object-level pre-training
    - SoCo (NeurIPS'2021)
  - Other than contrastive learning
    - BEiT (Tech Report)
- Self-supervised learning on Transformers
  - MoCo v3 (ICCV'2021), DINO (ICCV'2021)
  - SSL-Swin/MoBY/EsViT (tech report)



Improvements on Pascal VOC object detection

- PixPro (CVPR'2021)

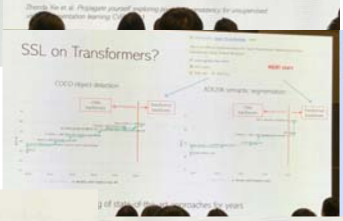
Totally 1.7% absolute improvements in 1 year!

Zhenda Xie et al. Propagate yourself: exploring pixel-level consistency for unsupervised visual representation learning CVPR 2021

"→像素级别预训练→目标级别预训练

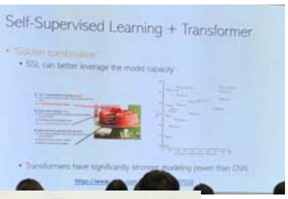
Bert for transformer?

CV → Contrast
CV → MM?



SSL on Transformers?

g of state-of-the-art approaches for years

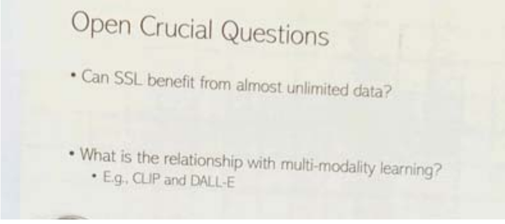


Self-Supervised Learning + Transformer

- "Golden combination"
  - SSL can better leverage the model capacity
- Transformers have significantly stronger modeling power than CNN

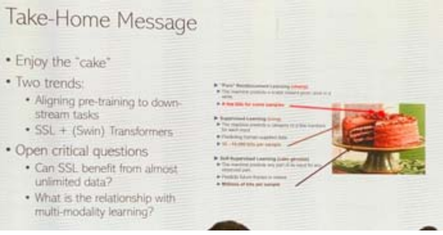
MoCo V3
DINO : segmentation
SSL-SWIN

OPEN Questions:



Open Crucial Questions

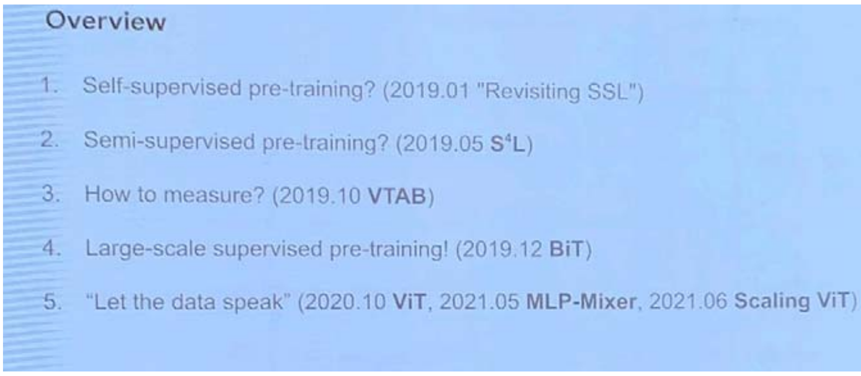
- Can SSL benefit from almost unlimited data?
- What is the relationship with multi-modality learning?
  - E.g., CLIP and DALL-E



Take-Home Message

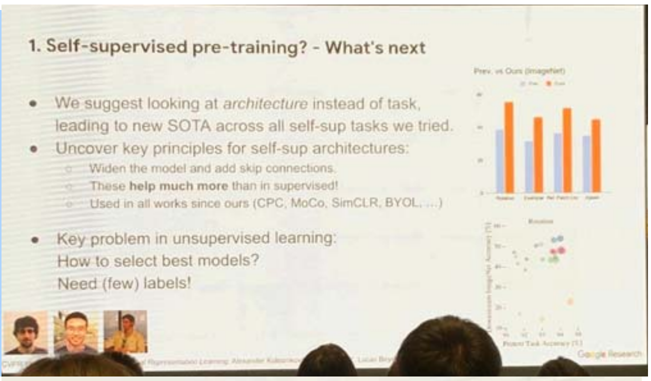
- Enjoy the "cake"
- Two trends:
  - Aligning pre-training to down-stream tasks
  - SSL + (Swin) Transformers
- Open critical questions
  - Can SSL benefit from almost unlimited data?
  - What is the relationship with multi-modality learning?

Workshop5 large scale ~
通过预训练来提升样本有效性?



Overview

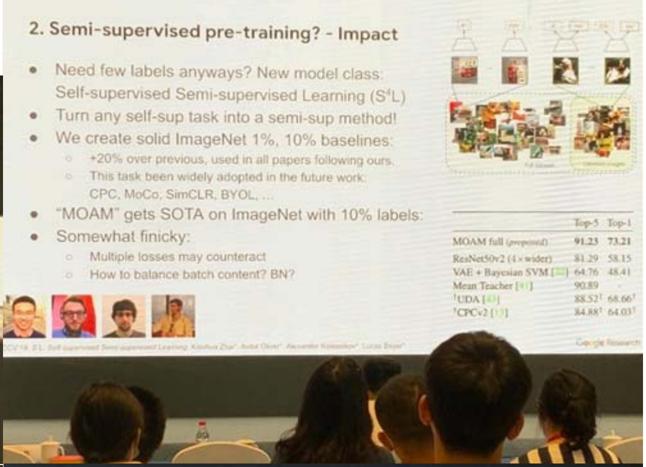
1. Self-supervised pre-training? (2019.01 "Revisiting SSL")
2. Semi-supervised pre-training? (2019.05 S⁴L)
3. How to measure? (2019.10 VTAB)
4. Large-scale supervised pre-training! (2019.12 BiT)
5. "Let the data speak" (2020.10 ViT, 2021.05 MLP-Mixer, 2021.06 Scaling ViT)



1. Self-supervised pre-training? - What's next

- We suggest looking at *architecture* instead of task, leading to new SOTA across all self-sup tasks we tried.
- Uncover key principles for self-sup architectures:
  - Widen the model and add skip connections.
  - These **help much more** than in supervised!
  - Used in all works since ours (CPC, MoCo, SimCLR, BYOL, ...)
- Key problem in unsupervised learning:
  How to select best models?
  Need (few) labels!

Prev. vs Ours (ImageNet)
Rotation
Pretext Task Accuracy [%]
Google Research

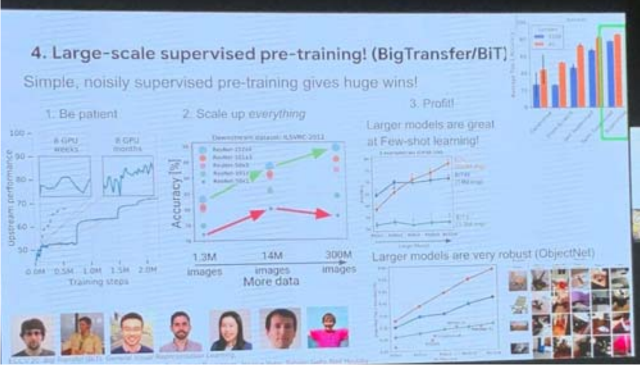


2. Semi-supervised pre-training? - Impact

- Need few labels anyways? New model class:
  Self-supervised Semi-supervised Learning (S⁴L)
- Turn any self-sup task into a semi-sup method!
- We create solid ImageNet 1%, 10% baselines:
  - +20% over previous, used in all papers following ours.
  - This task been widely adopted in the future work:
    CPC, MoCo, SimCLR, BYOL, ...
- "MOAM" gets SOTA on ImageNet with 10% labels:
- Somewhat finicky:
  - Multiple losses may counteract
  - How to balance batch content? BN?

	Top-5	Top-1
MOAM full (proposed)	**91.23**	**73.21**
ResNet50v2 (4×wider)	81.29	58.15
VAE + Bayesian SVM [32]	64.76	48.41
Mean Teacher [41]	90.89	-
†UDA [43]	88.52†	68.66†
†CPCv2 [17]	84.88†	64.03†

ICCV'19. S⁴L: Self-supervised Semi-supervised Learning. Xiaohua Zhai*, Avital Oliver*, Alexander Kolesnikov*, Lucas Beyer*



4. Large-scale supervised pre-training! (BigTransfer/BiT)

Simple, noisily supervised pre-training gives huge wins!

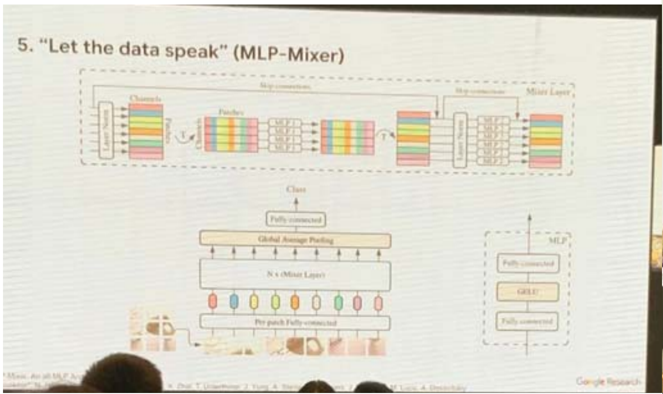
1. Be patient

2. Scale up *everything*

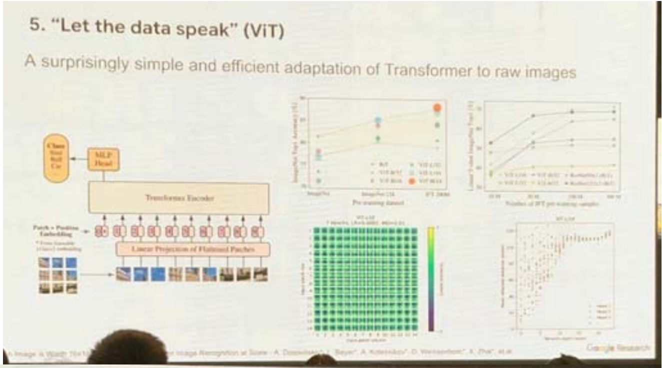
3. Profit!

Larger models are great at Few-shot learning!

Larger models are very robust (ObjectNet)

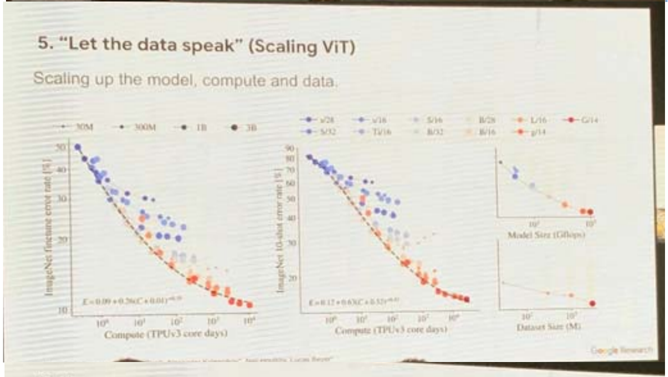


5. "Let the data speak" (MLP-Mixer)



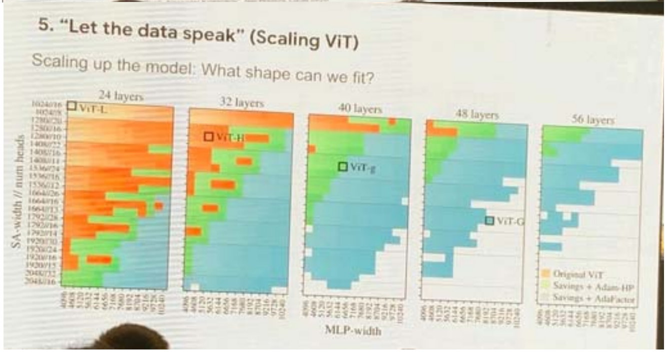
5. "Let the data speak" (ViT)

A surprisingly simple and efficient adaptation of Transformer to raw images



5. "Let the data speak" (Scaling ViT)

Scaling up the model, compute and data.



5. "Let the data speak" (Scaling ViT)

Scaling up the model: What shape can we fit?