Self-Supervised Learning in Vision

Emilio Zarbali

XITASO GmbH

February 1, 2024

Overview

1 Learning Algorithms

2 Supervised Learning

3 Issues with SL

4 SSL

5 Contrastive Learning

6 MaWis-KI

7 Discussion

XITASO

Taxonomy

- **Reinforcement Learning**
  - Learn model parameters using **active exploration** from sparse rewards
- **Unsupervised Learning**
  - Learn model parameters using **dataset without labels** $\{x_i\}_{i=1}^N$
- **Supervised Learning**
  - Learn model parameters using **dataset of data-label pairs** $\{(x_i, y_i)\}_{i=1}^N$
- **Self-supervised Learning**
  - Learn model parameters using **dataset of data-data pairs** $\{(x_i, x_i')\}_{i=1}^N$

XITASO

Self-supervised Learning

- A form of **unsupervised** learning where the supervision signal is derived from the **data itself**
- For most part we can differentiate between two SSL algorithms:
  - **Discriminative** (SimCLR, MoCo, BYOL, CLIP, ...): some sort of augmentations are applied to achieve learning rich features
  - **Generative** (MAE, oBoW, I-JEPA, ..): some part of the image is withheld and network generates missing part (similar to MLM)

XITASO

Success of SL

- Supervised Learning has shown tremendous capabilities in solving various tasks of learning
  - NLP
  - Computer Vision
  - Autonomous Systems
  - Neural Rendering
- outperforming classical methods

XITASO

Supervised Learning

- Let $\mathcal{D}$ denote the dataframe consisting of $(x_i, y_i)_{i=1}^{N}$ data-label pairs
- Supervised learning aims to learn a mapping function f

$$f : x \rightarrow y$$

- by minimizing some cost function $J(y, \hat{y})$, where y is the ground truth and $\hat{y}$ is the model predictions

XITASO

Drawbacks

- Supervised Learning requires **large amount of annotated data**
  - expensive and time-consuming
  - needs highly balanced dataset
  - struggles with following:
    - adversarial attacks, OOD detection, etc.
- Data distributions shift: Everytime you need **large annotation** campaigns
- Accuracy $\neq$ Robustness

XITASO

Annotation time

Figure: Cityscapes Example

- $\sim$ 90 Min/per Image in Cityscapes

XITASO

Fewer labeled data

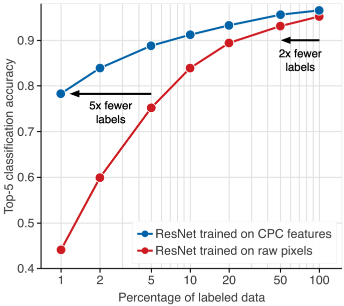
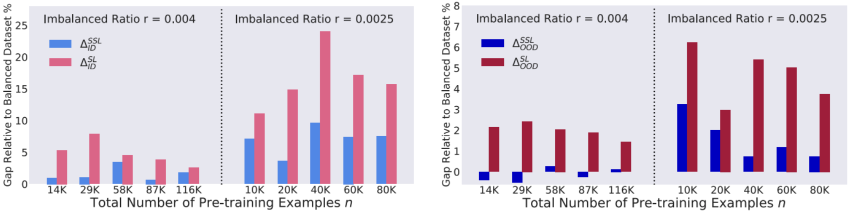


Figure: Label quantity[1] (higher is better)

- SSL performs much better with **fewer** labeled data

[1] Olivier J. Hénaff et al. "Data-Efficient Image Recognition with Contrastive Predictive Coding". In: *CoRR* (2019).

Class imbalance



(a) In Domain (ID). (b) Out of Domain (OOD).

Figure: Class imbalance performance gap[2] (lower is better)

- SSL is more **robust** to class imbalance
- captures richer sets of features that are not limited to **semantic classes**

[2] Hong Liu et al. "Self-supervised Learning is More Robust to Dataset Imbalance". In: *CoRR* (2021).

Robustness towards distortions

Pre-train Alg	IN Acc	C-10	C-100	STL-10	Car-196	Air-70	Avg Δ ↓
Sup-a	76.1	31.5%	45.3%	31.0%	51.2%	39.9%	39.8%
Sup-b	75.5	32.1%	47.2%	31.9%	53.2%	39.2%	40.7%
BYOL	72.3	29.3%	43.0%	29.0%	42.9%	33.8%	35.6%
SimSiam	68.3	27.8%	40.8%	29.3%	41.5%	32.6%	34.4%
MoCo-v2-a	66.4	28.1%	40.5%	29.4%	36.8%	29.4%	32.8%
MoCo-v2-b	71.1	31.3%	45.2%	31.0%	39.7%	31.3%	35.7%
SimCLR-v2	71.0	31.5%	45.4%	30.8%	43.0%	31.7%	36.5%
BarlowTwins	73.5	26.7%	39.8%	29.7%	43.0%	34.4%	34.7%
DeepCluster-v2	75.2	28.2%	41.1%	28.5%	43.2%	38.9%	36.0%
SwAV-a	72.0	27.0%	39.8%	28.3%	40.6%	33.9%	33.9%
SwAV-b	74.9	26.8%	39.3%	28.6%	41.4%	36.3%	34.5%

Figure: Robustness towards gamma distortions[3] (lower is better)

- **Robustness** allows model to work well in imperfect real-world scenarios

[3] Yuanyi Zhong et al. *Is Self-Supervised Learning More Robust Than Supervised Learning?* 2022. arXiv: 2206.05259.

Data Bias

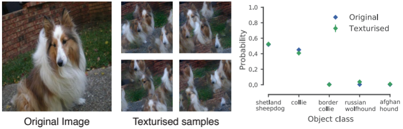


Figure: Pixel Distribution Bias[a]

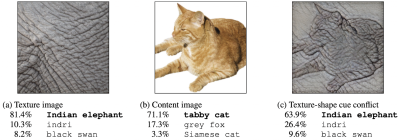


Figure: Texture Bias[a]

[a] Leon A .Gatys et al. "Texture and art with deep neural networks". In: *Current Opinion in Neurobiology* 46 (2017), pp. 178–186.

[a] Robert Geirhos et al. *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness*. 2022.

Summary

- Good datasets for complex tasks are **extremely** costly and **difficult** to collect and label
- Can we learn **useful** and **semantic rich** features only from data alone?

Overview

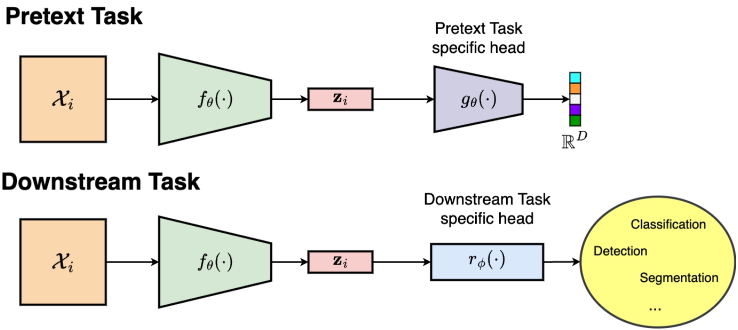


Figure: Self-Supervised Learning

Procedure of SSL

- Goal of pretext task:
  - Learn general knowledge with pretext task
- Pretext task:
  - define an auxiliary task for **pre-training** with large amount of **unlabeled** data
- Drop **projector** $g_\theta(\cdot)$ and use **feature extractor** $f_\theta(\cdot)$ for downstream task with labeled data

Evaluation

- SSL methods are evaluated on **downstream** task performance and not on pretext task
- Evaluation are based on **complexity** and **alignment** of pretext and downstream task
  - k-NN or Linear probe for classification tasks
  - Fine-tuning for tasks like Object detection, Segmentation, etc.

Challenges in SSL

Problems:

- Designing good pretext tasks are tedious and have no underlying theory behind it
- representations may not be general
- Mode Collapse

Introduction to Contrastive Learning

- SL and Metric Learning
  - Anchor and Positive: same **class**
  - Negative: **random different class**
- SSL
  - Anchor and Positive: same **image**
  - Negative: **random different image**

Motivation

- Idea behind contrastive learning is to make images from different views **close** in the feature space and all the other images **far away**
- Given a **score function** $s(\cdot, \cdot)$, we want to learn an encoder $f(\cdot)$ that yields **high score** for positive pairs (x, x^+) and low score for negative pairs (x, x^-)

$$s(f(x), f(x^+)) \gg s(f(x), f(x^-))$$

Another perspective

- Maximizing the **mutual information** between features extracted from different views forces encoder $f(\cdot)$ to capture information about higher-level factors

$$MI(x, x^{+}) \geq log(N) - \mathcal{L}$$

XITASO

Mutual Information

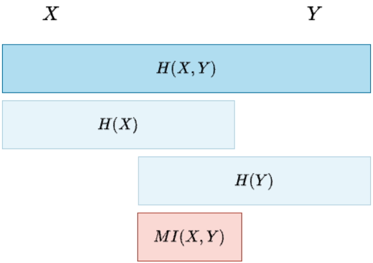
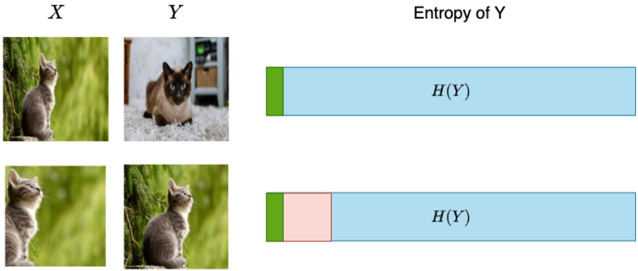
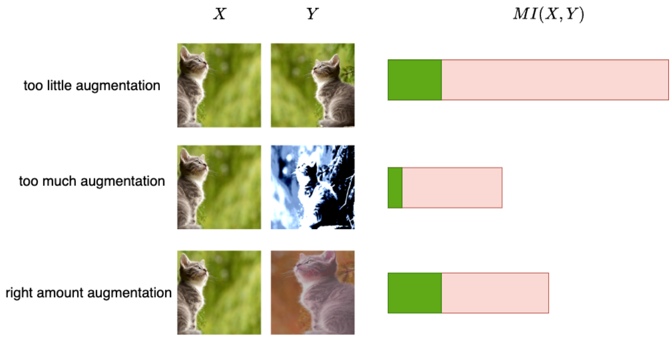


Figure: X and Y are two different images

Useful Mutual Information



Augmentations



XITASO

InfoMIN

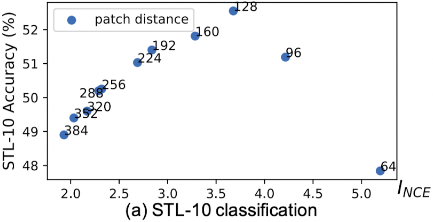


Figure: Performance vs Mutual Information[4]

[4] Yonglong Tian et al. *What Makes for Good Views for Contrastive Learning?* 2020. arXiv: 2005.10243 [cs.CV].

SimCLR

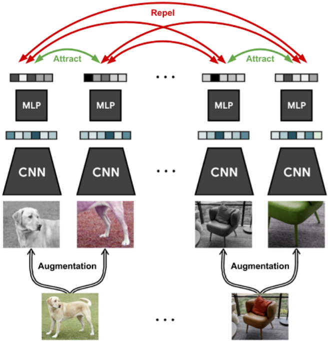


Figure: SimCLR[5]

- Augmentations of the same image are viewed as positives
- the rest of the batch is seen as negatives

XITASO

[5] Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *CoRR* (2020).

InfoNCE Loss

- SimCLR uses InfoNCE loss

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

- $sim(\cdot, \cdot)$ is typically cosine similarity
- $\mathcal{D}$ is of size 2N, as we obtain two views per image in dataset

Performance of SimCLR

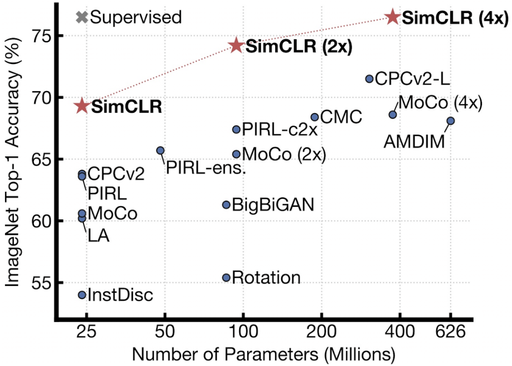


Figure: Performance on ImageNet (higher is better)

Drawbacks of SimCLR

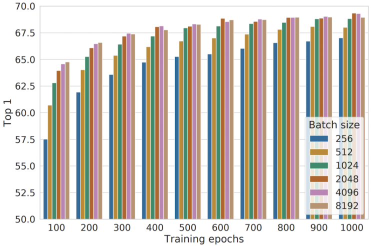


Figure: Batchsize used in SimCLR

- SimCLR rely shines under two following criterias
  - Large negatives: Bound is tighter with more negatives
  - Consequence: Large batch size

XITASO

MoCo Framework

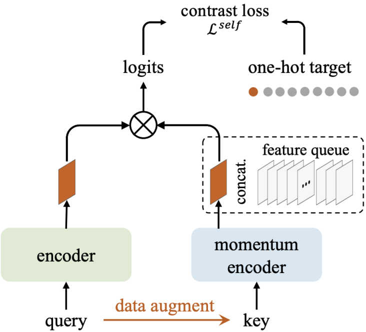


Figure: Momentum Contrast[6]

- Main objective: leverage contrastive learning with a smaller batch size
- BUT: More negatives are necessary for tighter bound

XITASO

[6] Kaiming He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *CoRR* (2019).

Query and Key Encoder

- Use two identical networks f_q and f_k, one query and key encoder respectively
- f_q is updated with gradient descent
- in order to keep memory consistent, He et al. used following trick:
- f_k is updated with $\theta_k \leftarrow m\theta_k + (1 - m)\theta_q$

XITASO

Results

case	MLP	aug+	cos	epochs	batch	ImageNet acc.
	unsup. pre-train					
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**

Figure: Results on ImageNet Evaluation[7]

[7] Kaiming He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *CoRR* (2019).

Bootstrap Your Own Latent (BYOL)

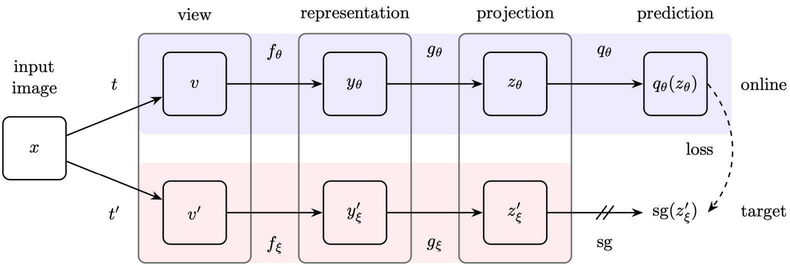


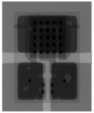
Figure: BYOL[8]

- MSE-Loss between Online representations and Target representations
- f_{online} is updated with gradient descent
- f_{target} is updated with $\theta_{target} \leftarrow m\theta_{target} + (1-m)\theta_{online}$

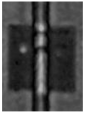
[8] Jean-Bastien Grill et al. *Bootstrap your own latent: A new approach to self-supervised Learning*. 2020.

MaWis-KI

X-Ray



SAM



TTA

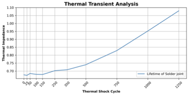


Figure: Data

- Goal: Reliable lifetime prediction of solder joints using data-driven methods
  - automotive electronics
  - Voids and cracks have big impact on quality

XITASO

Sampling rate

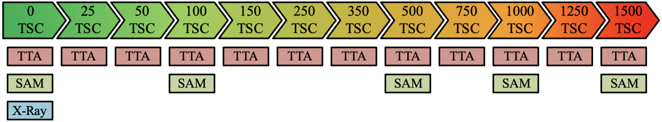


Figure: Sampling Rate

Model

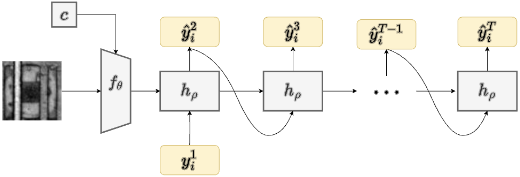


Figure: Model Architecture

Visualization of the Embedding Space

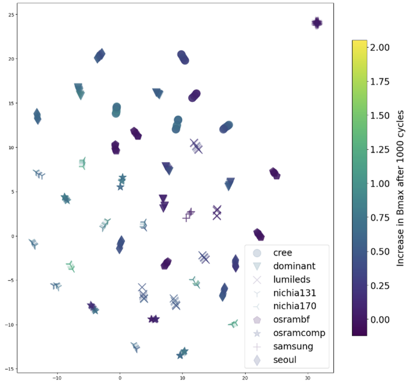


Figure: Embedding Space of SL

XITASO

Pretext Task

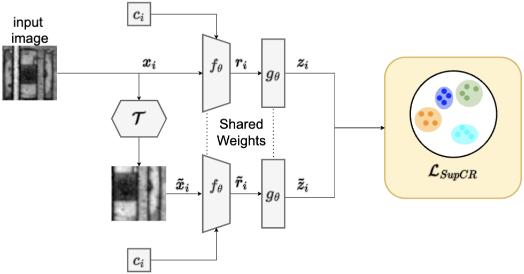


Figure: Contrastive Learning Pipeline

XITASO

Visualization of the Embedding Space

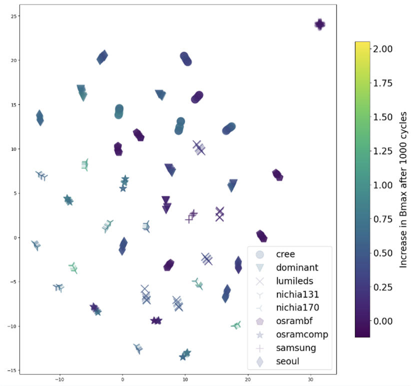


Figure: Embedding Space of SL

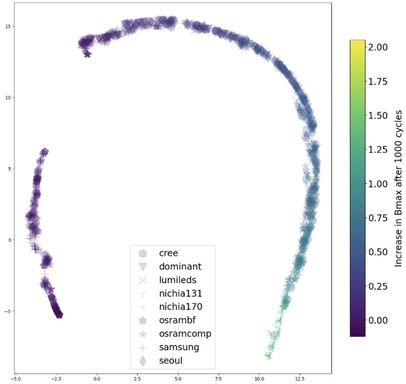


Figure: Embedding Space of SSL

XITASO

Learning Features

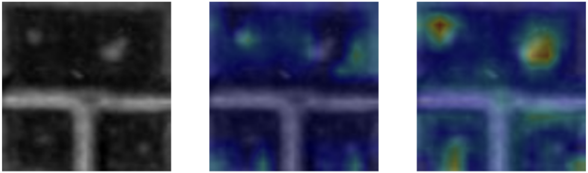


Figure: GT vs SL vs SSL[9]

[9] Emilio Zarbali et al. *Contrastive pretraining of regression tasks in automotive electronics*. 2023.

Takeaway

- Self-supervised learning should not be seen as a new technique to compete against supervised learning
- rather in conjunction with supervised learning as seen in
  - NLP: BERT, GPT, LLaMa, etc.
  - Multitask Learning: CLIP, Flamingo, etc.
  - Vision: SimCLR, DINO, etc.
- Pretext task has to be carefully designed with respect to
  - Goals of downstream task
  - Invariance and equivariance of downstream task
- especially **Transformer** architecture benefits from SSL pretraining

Discussion!

emilio.zarbali@xitaso.com

XITASO