Recent methods for self-supervised learning in computer vision

Damien ROBERT

CSAI, ENGIE Lab CRIGEN

LASTIG, Univ Gustave Eiffel, IGN-ENSG

Introduction

🔥 CVPR22 HOT TOPICS - A SUBJECTIVE PERCEPTION

NeRF Neural Radiance Fields

ViTs Vision Transformers

SSL Self-Supervised Learning

Self-Supervised Learning

Self-Supervised Learning ❓

Self-Supervised Learning ?

I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term. [...]

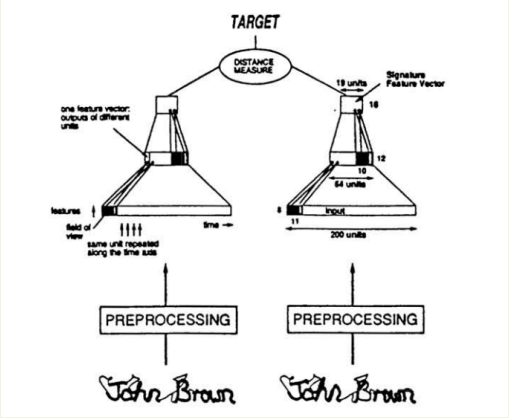
Self-supervised learning uses way more supervisory signals than supervised learning, and enormously more than reinforcement learning. That's why calling it "unsupervised" is totally misleading. [...]

Self-supervised learning has been enormously successful in natural language processing. [...]

credit: Yann LeCun, `https://www.facebook.com/yann.lecun/posts/10155934004262143`

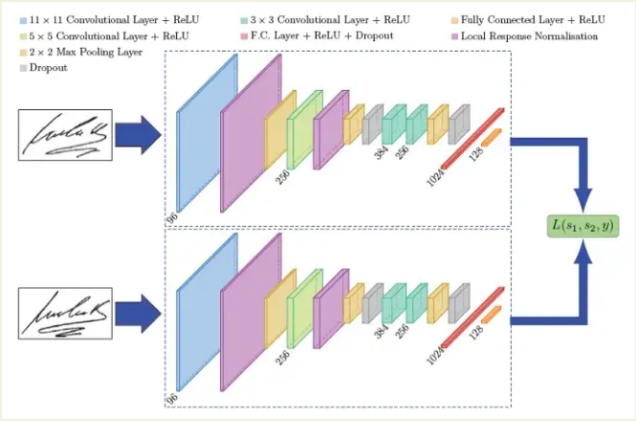
Contrastive Learning

Siamese Networks



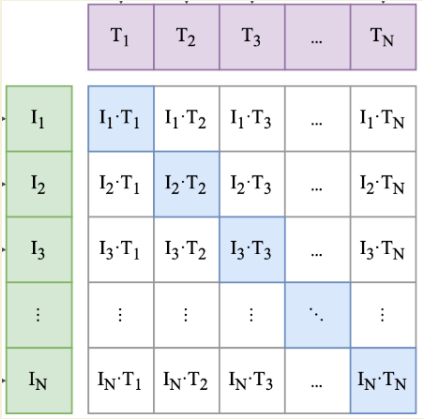
credit: Bromley NeurIPS 1993 [2]

Siamese Networks



credit: https://towardsdatascience.com/a-friendly-introduction-to-siamese-networks-85ab17522942

Contrastive Learning



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

credit: https://lilianweng.github.io/posts/2021-05-31-contrastive

credit: Hadsell CVPR 2006 [6]

Contrastive Learning

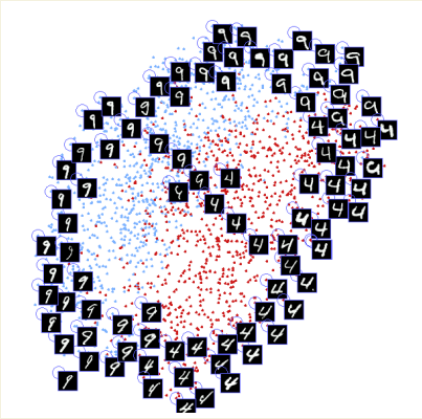
$$D_W(\vec{X_1}, \vec{X_2}) = \|G_W(\vec{X_1}) - G_W(\vec{X_2})\|_2 \quad (1)$$

$$\mathcal{L}(W) = \sum_{i=1}^{P} L(W, (Y, \vec{X_1}, \vec{X_2})^i) \quad (2)$$

$$L(W, Y, \vec{X_1}, \vec{X_2}) = (1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2 \quad (4)$$

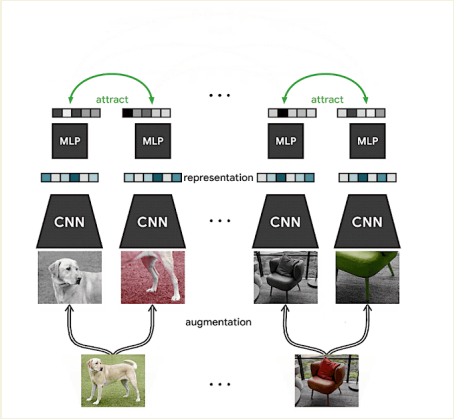
credit: Hadsell CVPR 2006 [6]

Contrastive Learning



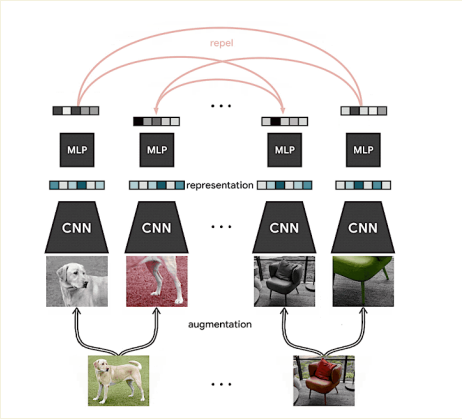
credit: Hadsell CVPR 2006 [6]

SimCLR - Modern framework for Contrastive Learning



credit: https://akichan-f.medium.com/mae-simmim-for-pre-training-like-a-masked-language-model-9b42579e25a9
credit: Chen ICML 2020 [3]

SimCLR - Modern framework for Contrastive Learning



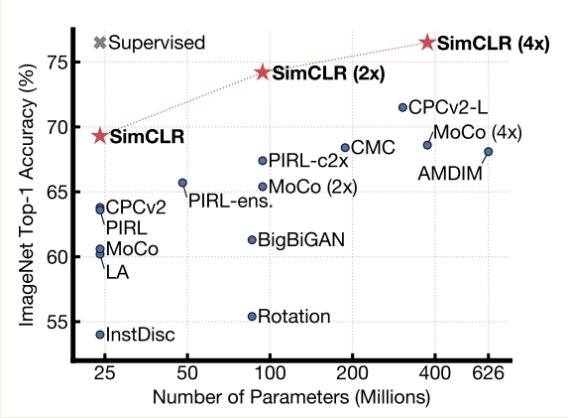
credit: https://akichan-f.medium.com/mae-simmim-for-pre-training-like-a-masked-language-model-9b42579e25a9
credit: Chen ICML 2020 [3]

SimCLR - Modern framework for Contrastive Learning

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

credit: Chen ICML 2020 [3]

SimCLR - Modern framework for Contrastive Learning



credit: Chen ICML 2020 [3]

🤷 So what ?

Few-shot learning

Learn representation with CL

⬇️

❄️ Freeze representation model

⬇️

Learn classifier on few annotations

🤷 So what ?

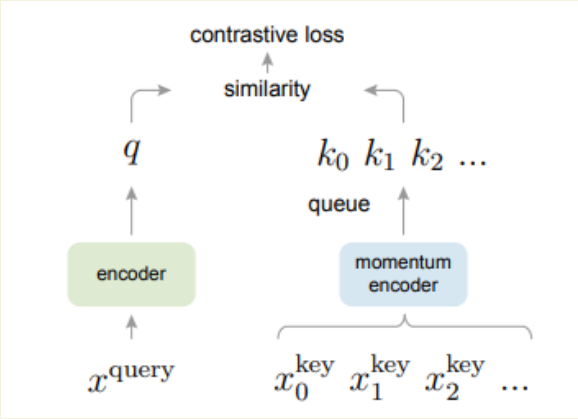
Multimodal learning

Align multimodal representations (📜+🖼️, ☁️+🖼️, ...) with CL

⬇️

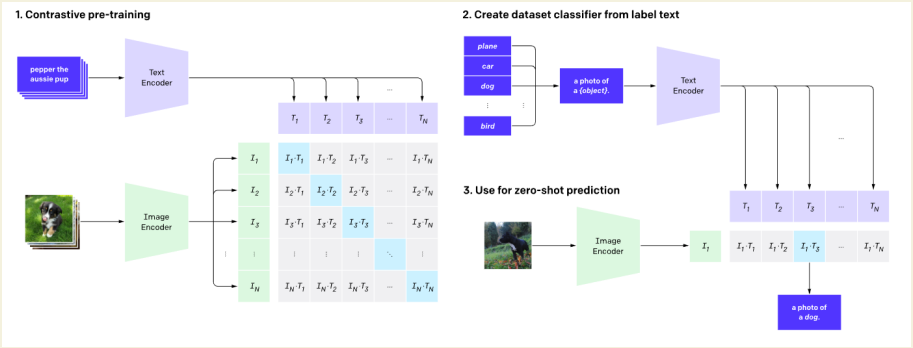
Distillation, single-modality annotation, ...

MoCo



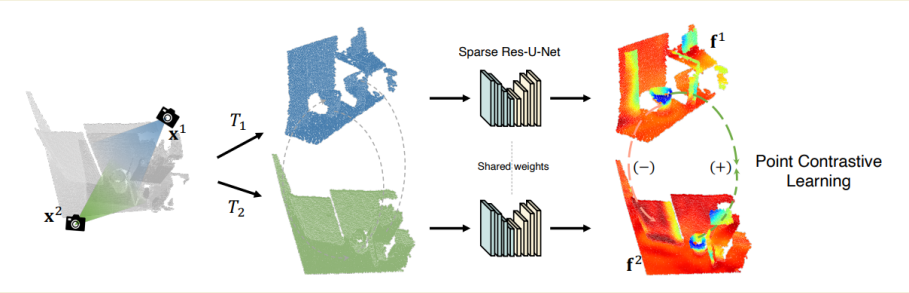
credit: He CVPR 2020 [8]

📜+🖼️ CLIP



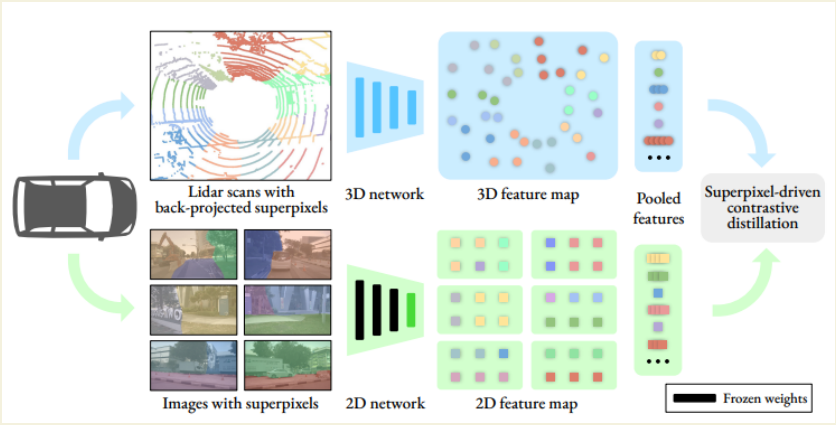
credit: `https://openai.com/blog/clip`
credit: Radford & Kim ICML 2021 [10]

PointContrast



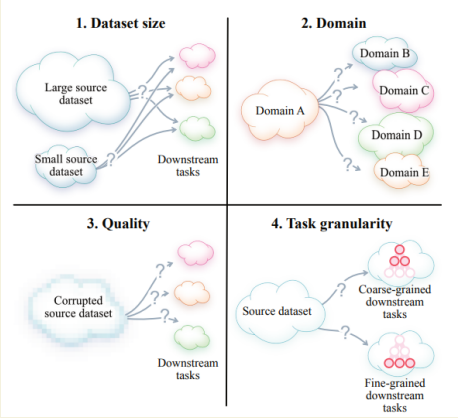
credit: Xie ECCV 2020 [12]

Image-to-Lidar distillation



credit: Sautier CVPR 2022 [11]

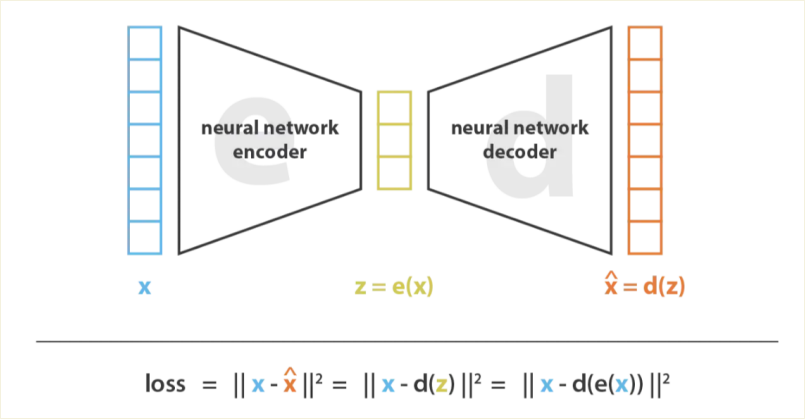
🤔 When Does Contrastive Visual Representation Learning Work ?



credit: Cole CVPR 2022 [4]

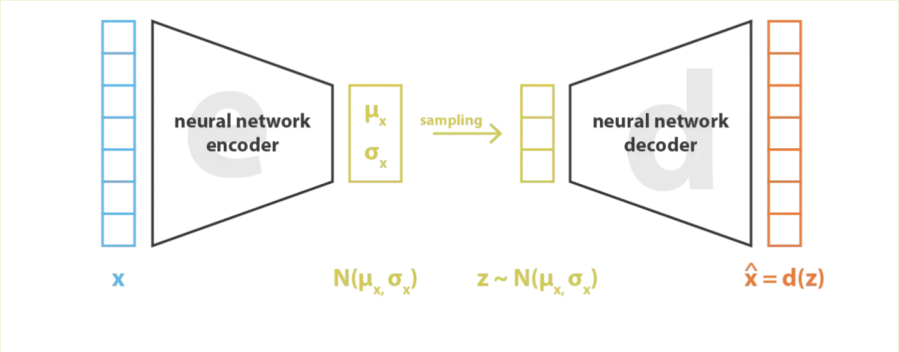
Masked Auto-Encoders

Autoencoder



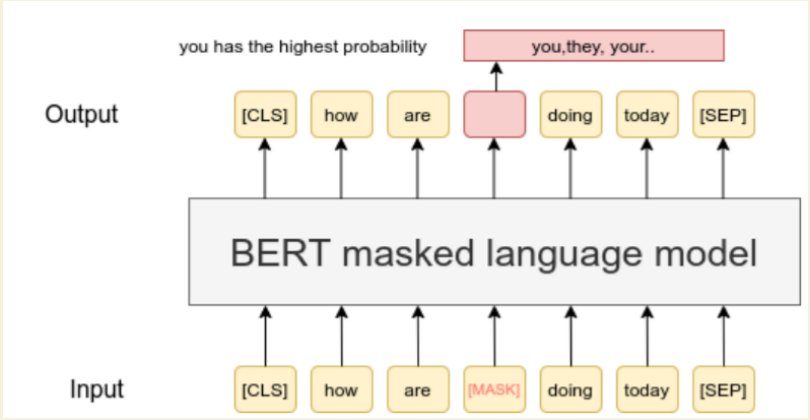
credit: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Variational Autoencoder



credit: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

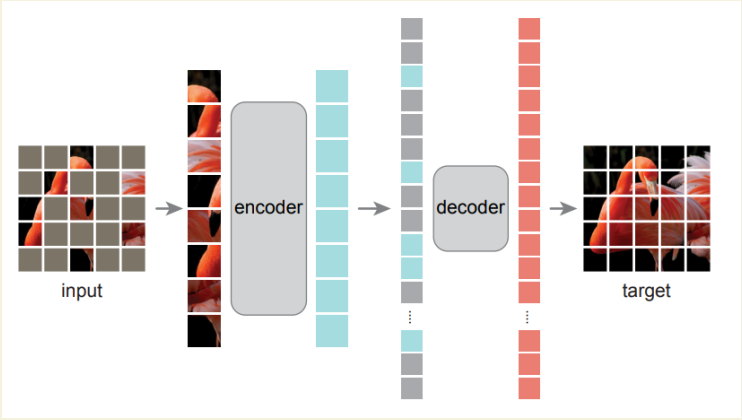
BERT - Masked Language Modeling



credit: https://www.sbert.net/examples/unsupervised_learning/MLM/README.html

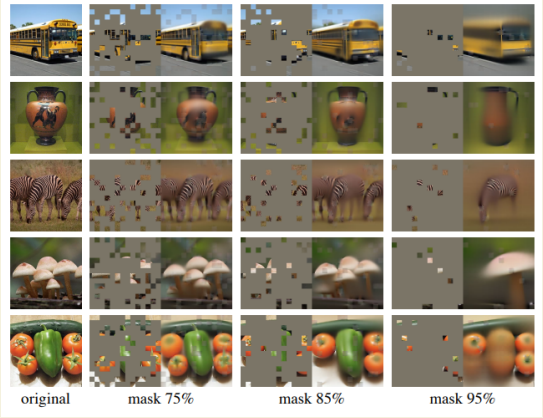
credit: Devlin arxiv 2018 [5]

Masked (image) Autoencoder



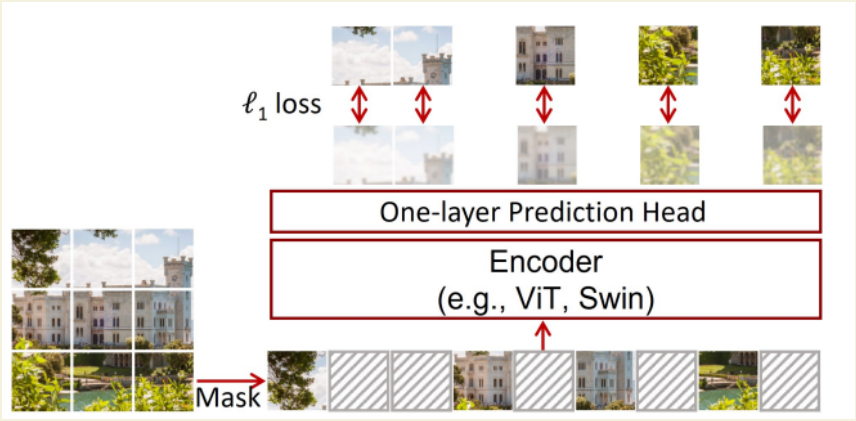
credit: He CVPR 2022 [7]

Masked (image) Autoencoder



credit: He CVPR 2022 [7]

SimMIM



credit: Xie CVPR 2022 [13]

🌊 The MAE wave

BEiT BEIT: BERT Pre-Training of Image Transformers [1]

MeshMAE Masked Autoencoders for 3D Mesh Data Analysis [9]

3D points Masked Autoencoders in 3D Point Cloud Representation Learning
`https://arxiv.org/abs/2207.01545`

Loads more 🚀 `https://github.com/EdisonLeeeee/Awesome-Masked-Autoencoders`

Questions ?

References

References I


H. Bao, L. Dong, and F. Wei.
Beit: Bert pre-training of image transformers.
arXiv, 2021.


J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah.
Signature verification using a" siamese" time delay neural network.
NeurIPS, 1993.


T. Chen, S. Kornblith, M. Norouzi, and G. Hinton.
A simple framework for contrastive learning of visual representations.
In *ICML*, 2020.


E. Cole, X. Yang, K. Wilber, O. Mac Aodha, and S. Belongie.
When does contrastive visual representation learning work?
In *CVPR*, 2022.

References II


J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova.
Bert: Pre-training of deep bidirectional transformers for language understanding.
arXiv, 2018.


R. Hadsell, S. Chopra, and Y. LeCun.
Dimensionality reduction by learning an invariant mapping.
In *CVPR*, 2006.


K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick.
Masked autoencoders are scalable vision learners.
In *CVPR*, 2022.


K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick.
Momentum contrast for unsupervised visual representation learning.
In *CVPR*, 2020.

References III


Y. Liang, S. Zhao, B. Yu, J. Zhang, and F. He.
Meshmae: Masked autoencoders for 3d mesh data analysis.
In *European Conference on Computer Vision*, pages 37–54. Springer, 2022.


A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al.
Learning transferable visual models from natural language supervision.
In *ICML*, 2021.


C. Sautier, G. Puy, S. Gidaris, A. Boulch, A. Bursuc, and R. Marlet.
Image-to-lidar self-supervised distillation for autonomous driving data.
In *CVPR*, 2022.


S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany.
Pointcontrast: Unsupervised pre-training for 3d point cloud understanding.
In *ECCV*, 2020.

References iv


Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu.
Simmim: A simple framework for masked image modeling.
In *CVPR*, 2022.