Self-supervised Visual Representation Learning

Weidi Xie

Shanghai Jiao Tong University

Agenda

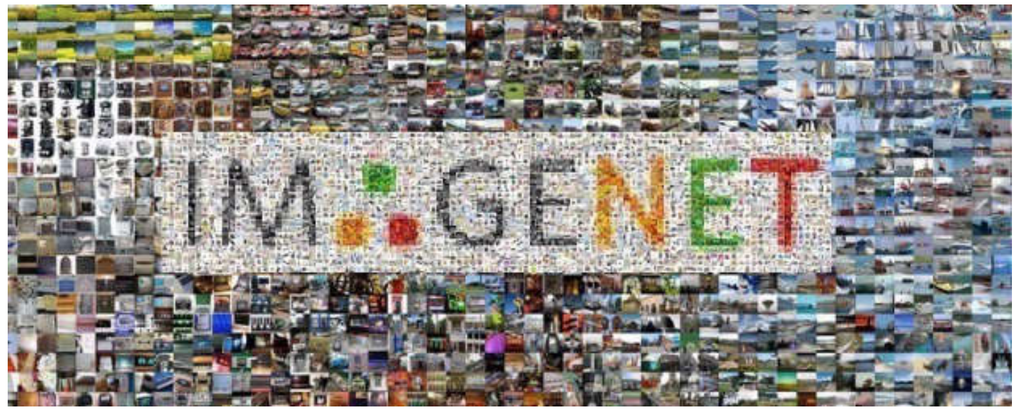
- What is self-supervised learning, and why it is important
- Self-supervised representation learning for images
- Self-supervised representation learning for videos
- Future directions

Disclaimer

- This is huge topic with a vast, multi-disciplinary history
- I will inevitably miss important related work
- Each citation is only meant as a representative example; seed for connectedpapers.com
- There are many views on the literature, and this is only one of them.
- Not necessarily chronological
- Email me with pointers or suggestions and I will update the slides

ImageNet Story

- ImageNet is the large-scale benchmark for image classification
  - training: 1k categories, 1k images per category,
  - testing: 100k images



- To some extent, any visual task can now be solved :
  1) collect a large-scale dataset labelled for the task of interest
  2) specify a training objective and neural network architecture
  3) train the model and deploy
  4) good and sufficient data is all you need

Why self-supervised learning

- Self-supervised learning is an alternative for learning representation:
  1) save cost for manual annotations
  2) more scalable training, e.g. billions of images are generated everyday
  3) more robust visual representation learning

- Human infants adopt very different ways for learning:
  1) mainly from observation and manipulation
  2) only require few explicit supervision
  3) learn from multimodal data, e.g. vision, sound, touch, etc.
  4) learn progressively, e.g. curriculum learning

The Development of Embodied Cognition: Six Lessons from Babies, by Linda Smith and Michael Gasser

Turing Award winners at AAAI 2020

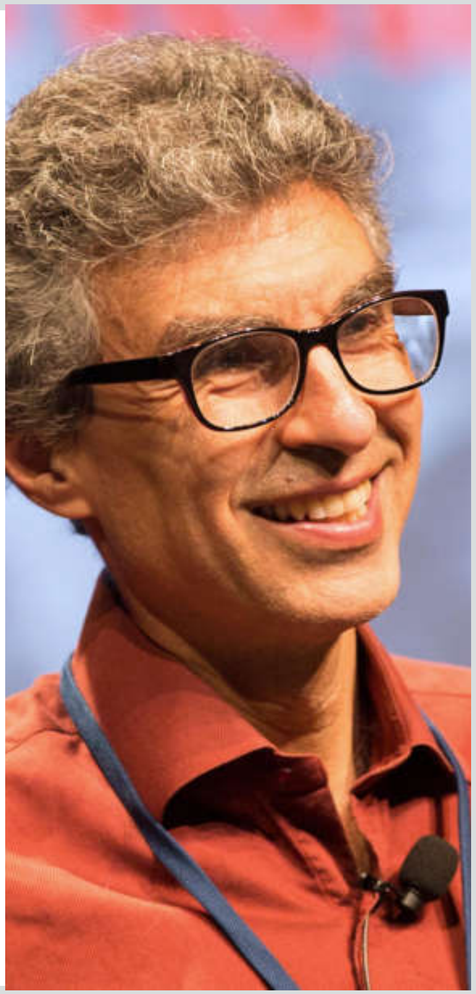
> I always knew unsupervised learning was the right thing to do
>
> — Geoff Hinton

> Basically it's the idea of learning to represent the world before learning a task — and this is what babies do
>
> — Yann LeCun

> And so if we can build models of the world where we have the right abstractions, where we can pin down those changes to just one or a few variables, then we will be able to adapt to those changes because we don't need as much data, as much observation in order to figure out what has changed.
>
> — Yoshua Bengio

Jérémy Barande / Ecole polytechnique Université Paris-Saclay / CC BY-SA 2.0

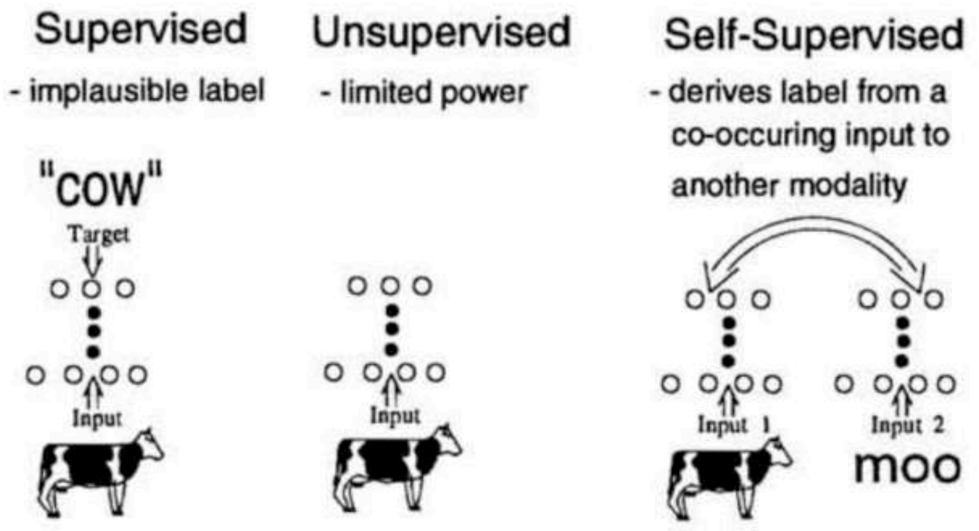
Eviatar Bach / CC BY-SA



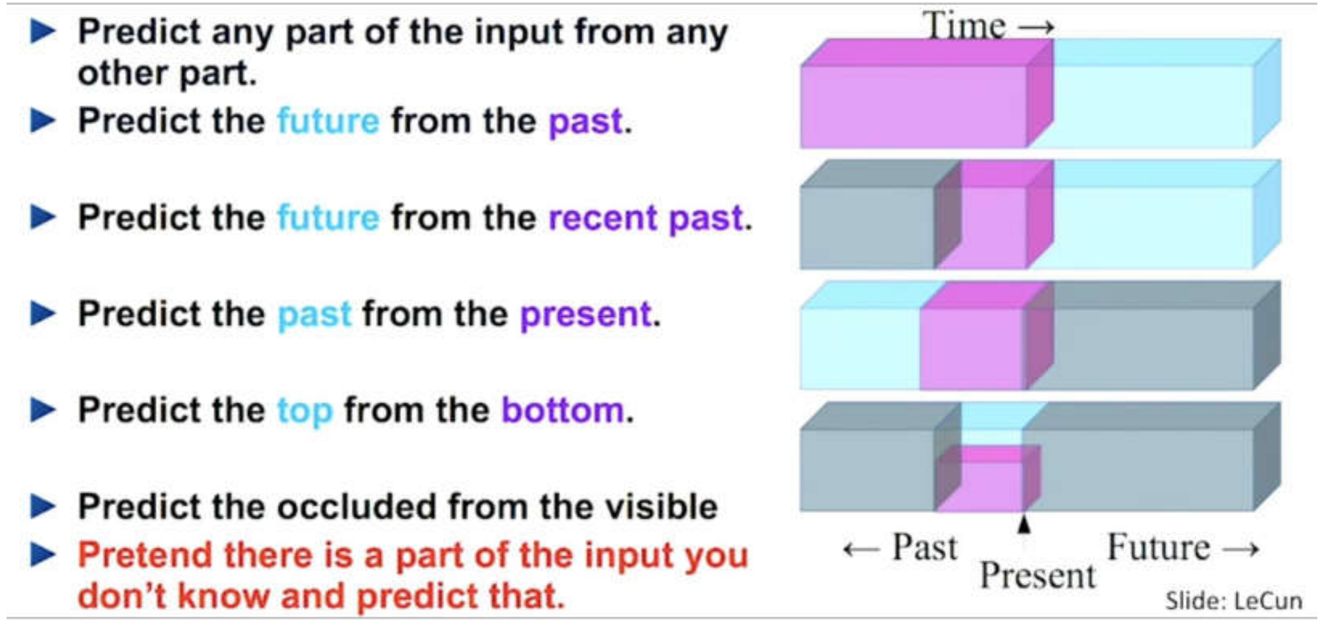
Jérémy Barande / Ecole polytechnique Université Paris-Saclay / CC BY-SA 2.0

What is self-supervised learning

- A form of unsupervised learning where the data itself provides the supervision.
- Withhold some part of the data, and task the network with predicting it, termed as proxy task.
- While solving the proxy task, the network is forced to learn what we really care about, e.g. a semantic representation.

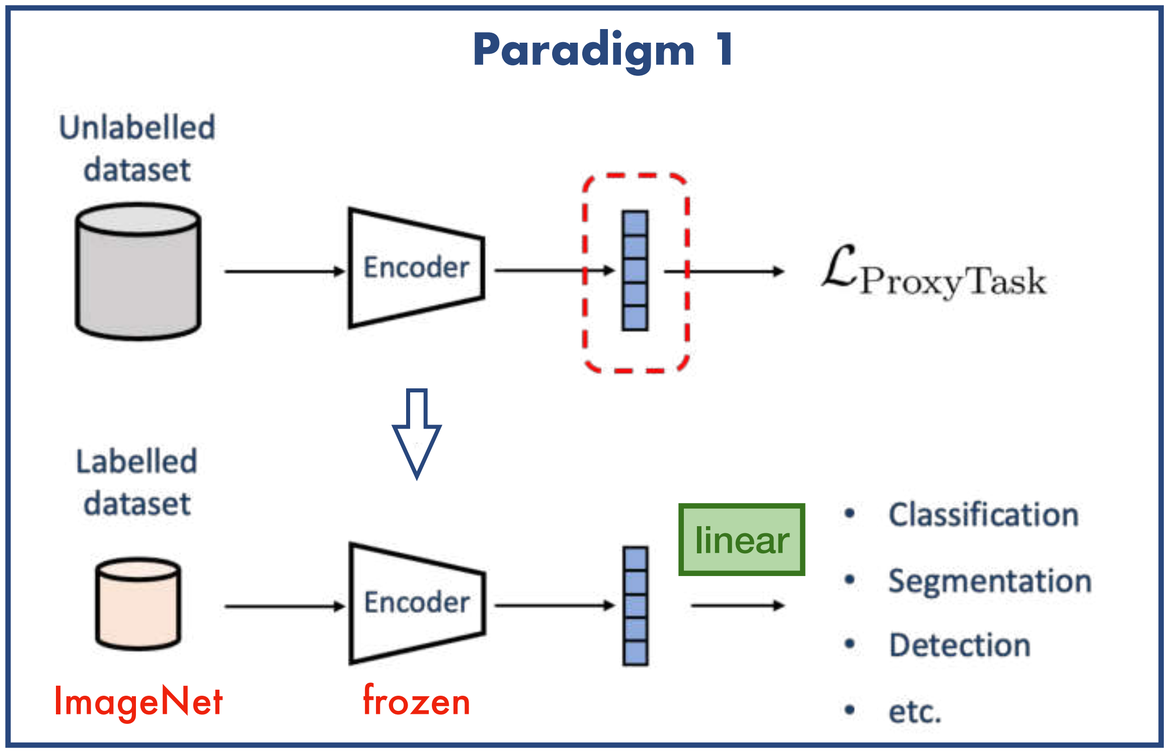


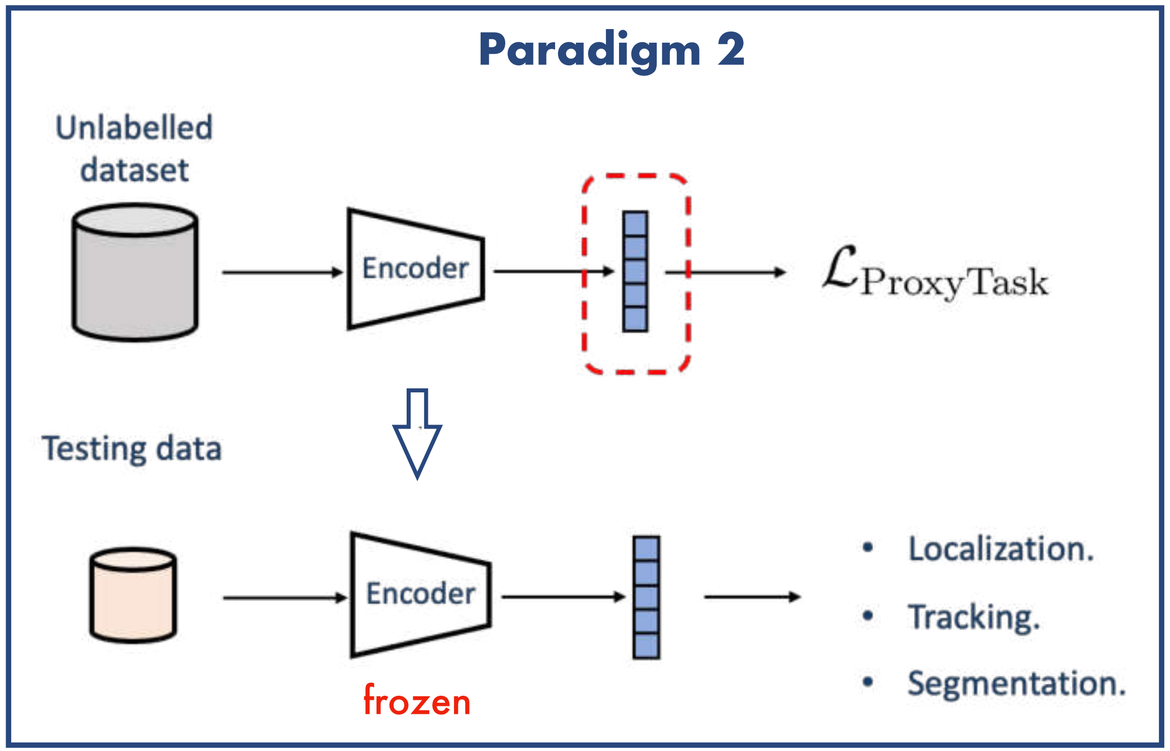
Virginia de Sa, 1994, Image: Learning classification with Unlabeled Data



What is self-supervised learning

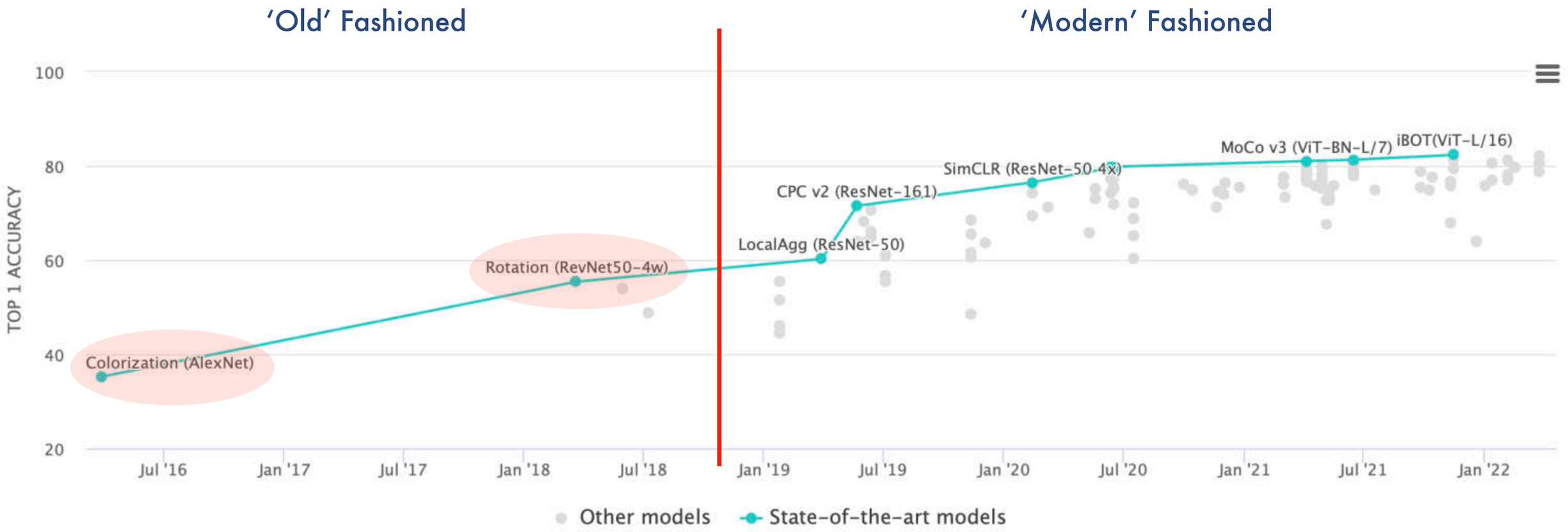
- There are generally two paradigms for evaluating the representation from self-supervised training:
  1) evaluate on ImageNet with linear probing, i.e. freeze the feature encoder, and only train the classifier with labels
  2) evaluate on downstream tasks without fine-tuning, e.g. zero-shot transfer





Results

- Linear probing on ImageNet image classification



Self-supervised Representation Learning from Images

(‘old’ fashioned)

Learn by predicting relative position

Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch[1,2] Abhinav Gupta[1] Alexei A. Efros[2]

[1] School of Computer Science
Carnegie Mellon University

[2] Dept. of Electrical Engineering and Computer Science
University of California, Berkeley

Abstract

This work explores the use of spatial context as a source of free and plentiful supervisory signal for training a rich visual representation. Given only a large, unlabeled image collection, we extract random pairs of patches from each image and train a convolutional neural net to predict the position of the second patch relative to the first. We argue that doing well on this task requires the model to learn to recognize objects and their parts. We demonstrate that the feature representation learned using this within-image context indeed captures visual similarity across images. For example, this representation allows us to perform unsupervised visual discovery of objects like cats, people, and even birds from the Pascal VOC 2011 detection dataset. Furthermore, we show that the learned ConvNet can be used in the R-CNN framework [21] and provides a significant boost over

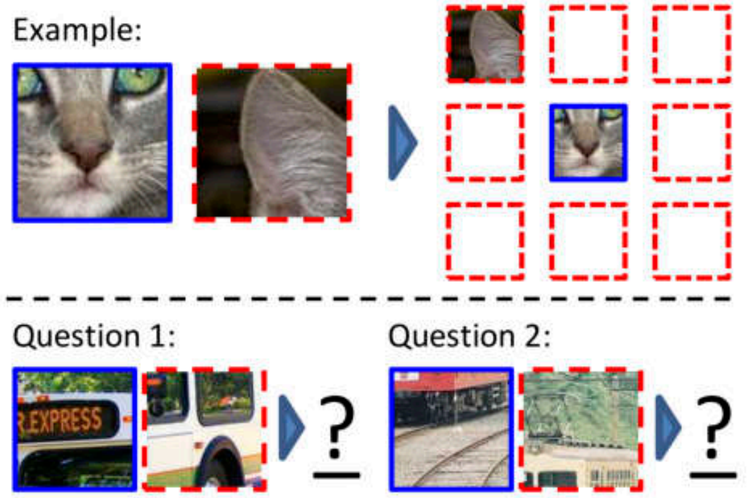
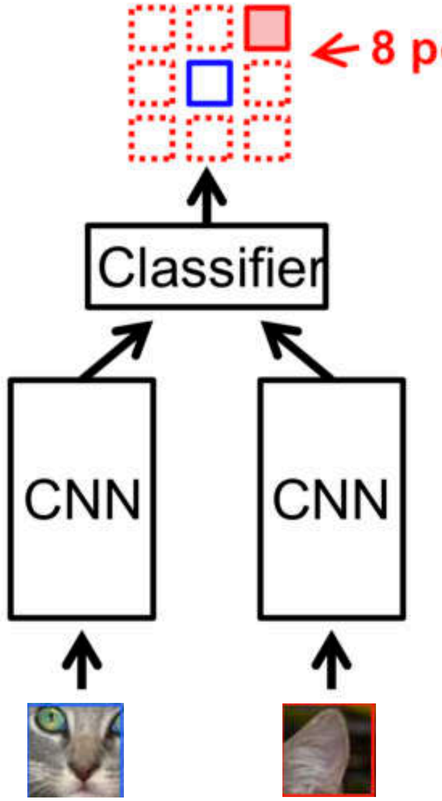
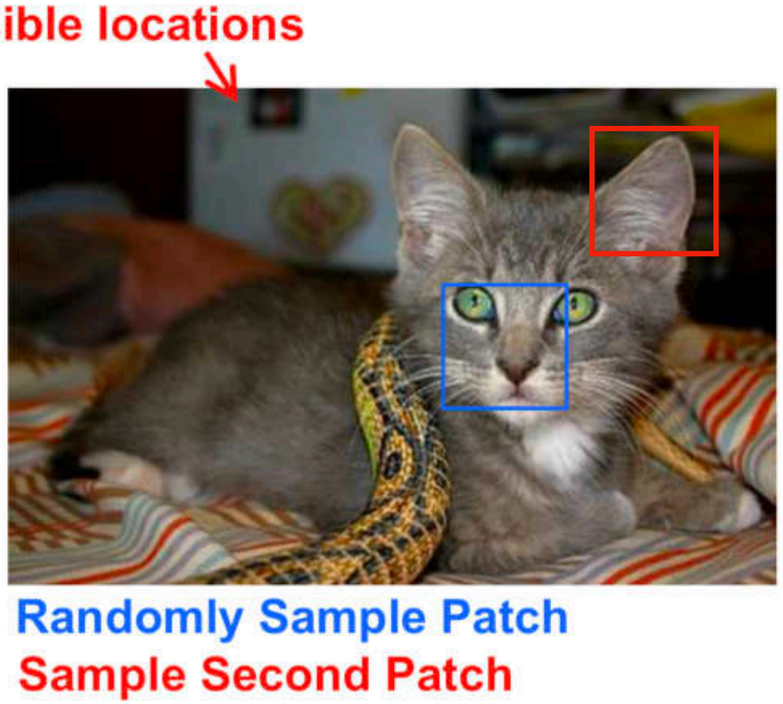


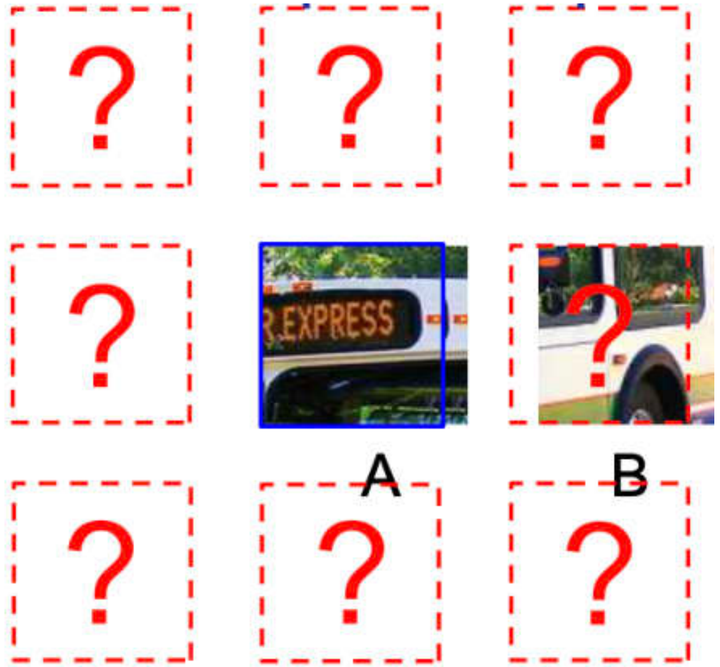
Figure 1. Our task for learning patch representations involves randomly sampling a patch (blue) and then one of eight possible neighbors (red). Can you guess the spatial configuration for the two pairs of patches? Note that the task is much easier once you have recognized the object!

Learn by predicting relative position

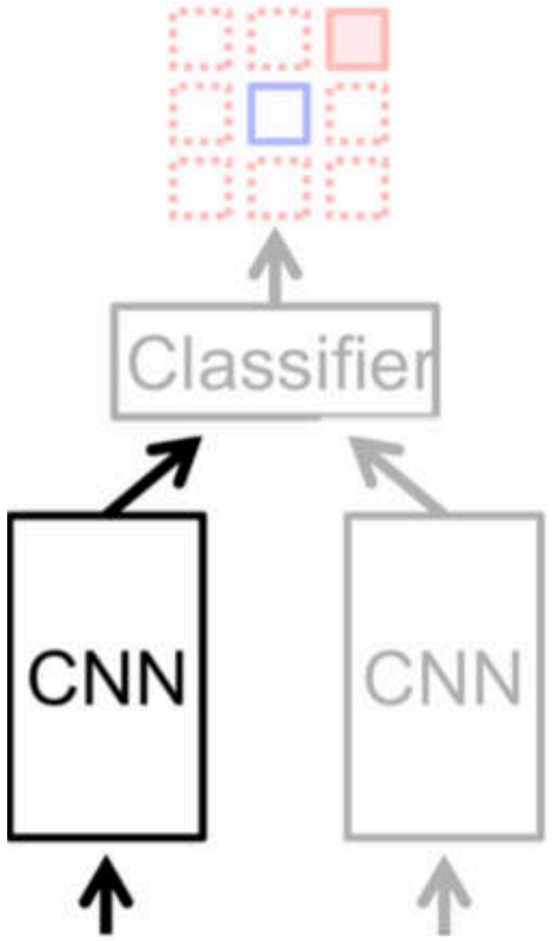
- Randomly sample patches from the image, and task the network to predict their relative position.
- If the model learns to solve this task, it may have learnt useful visual representation on object structure.







Retrieve nearest patches



Learn with Jigsaw Puzzles

Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles

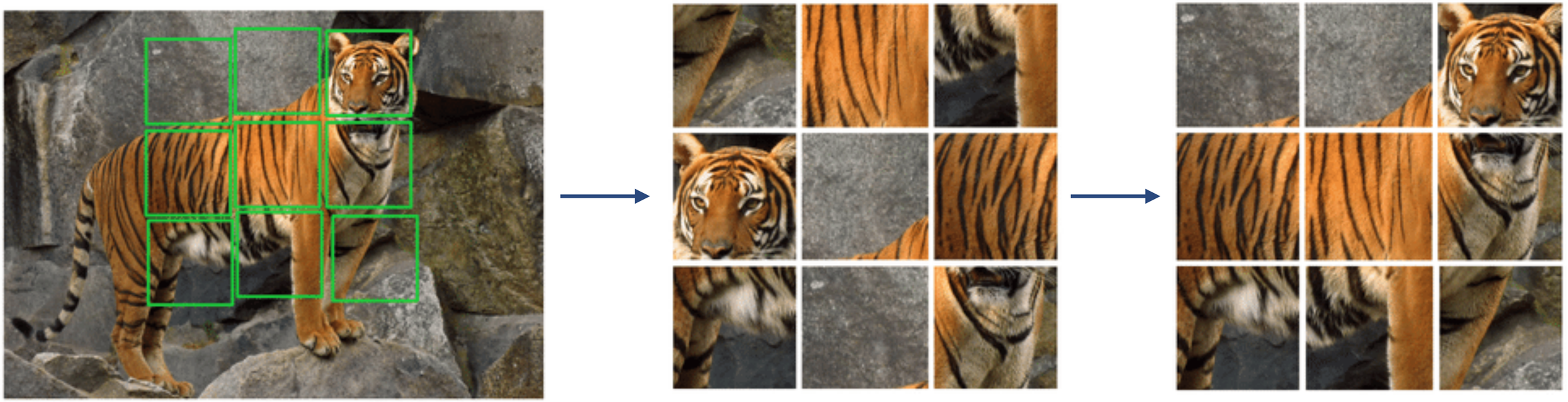
Mehdi Noroozi and Paolo Favaro

Institute for Informatiks
University of Bern
{noroozi,paolo.favaro}@inf.unibe.ch

Abstract. In this paper we study the problem of image representation learning without human annotation. By following the principles of self-supervision, we build a convolutional neural network (CNN) that can be trained to solve Jigsaw puzzles as a *pretext* task, which requires no manual labeling, and then later repurposed to solve object classification

Learn with Jigsaw Puzzles

- Randomly shuffle the image patches.
- Train a network to recover the image, e.g. their relative positions.
- Similar intuition to the previous paper, scale the pairwise relation to multi-patch permutations.



Learning by predicting rotations

UNSUPERVISED REPRESENTATION LEARNING BY PREDICTING IMAGE ROTATIONS

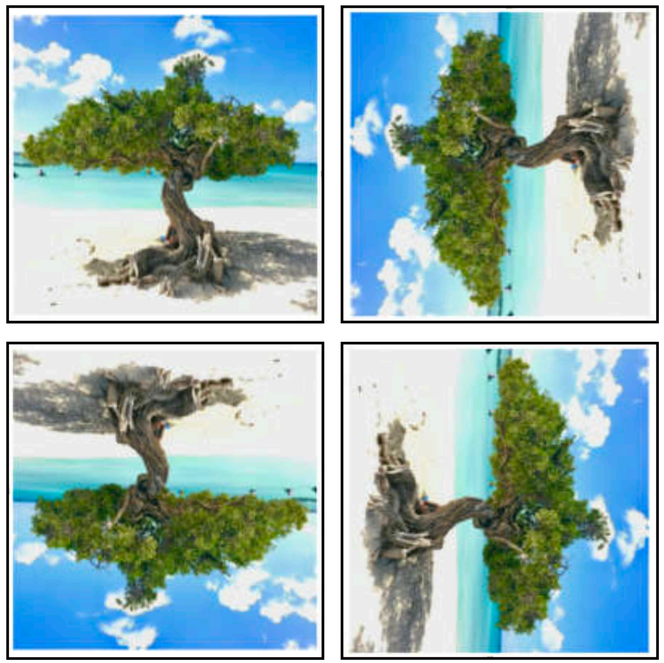
Spyros Gidaris, Praveer Singh, Nikos Komodakis
University Paris-Est, LIGM
Ecole des Ponts ParisTech
{spyros.gidaris,praveer.singh,nikos.komodakis}@enpc.fr

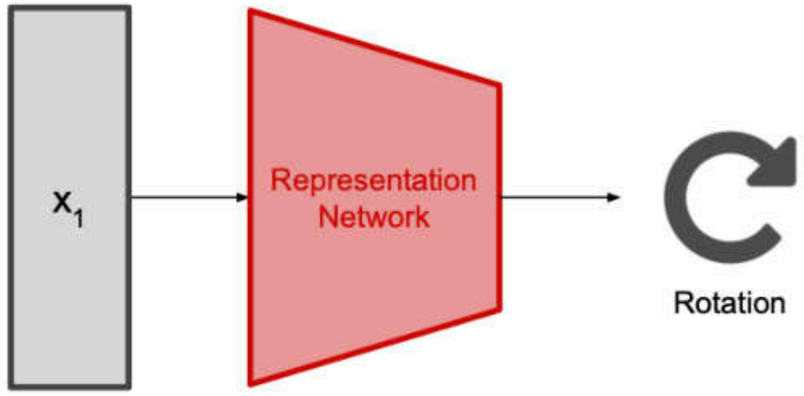
ABSTRACT

Over the last years, deep convolutional neural networks (ConvNets) have transformed the field of computer vision thanks to their unparalleled capacity to learn high level semantic image features. However, in order to successfully learn those features, they usually require massive amounts of manually labeled data, which is both expensive and impractical to scale. Therefore, unsupervised semantic feature learning, i.e., learning without requiring manual annotation effort, is of crucial importance in order to successfully harvest the vast amount of visual data that are available today. In our work we propose to learn image features by training Con

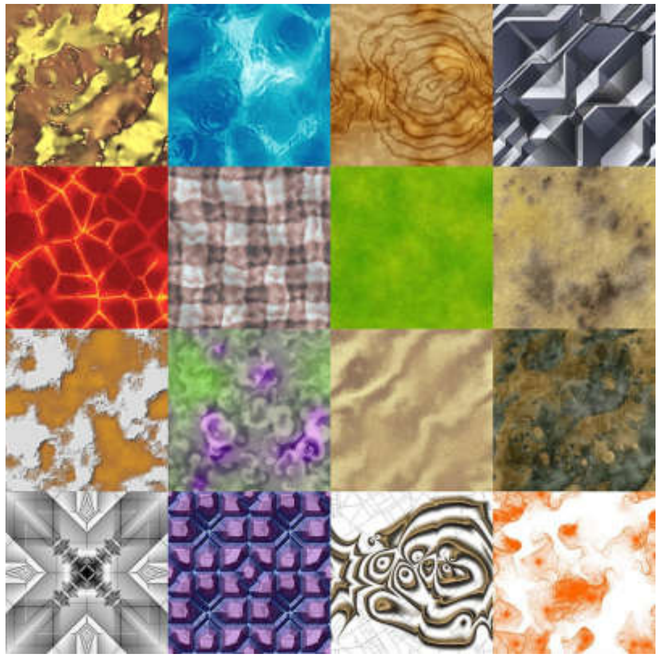
Learning by predicting rotations

- Randomly rotate the image, and task the model to predict the rotation degrees.
- Intuitively, in order to fulfil such task, the network has to rely on the cues from structured object, because textures are supposed to be rotation-invariant.





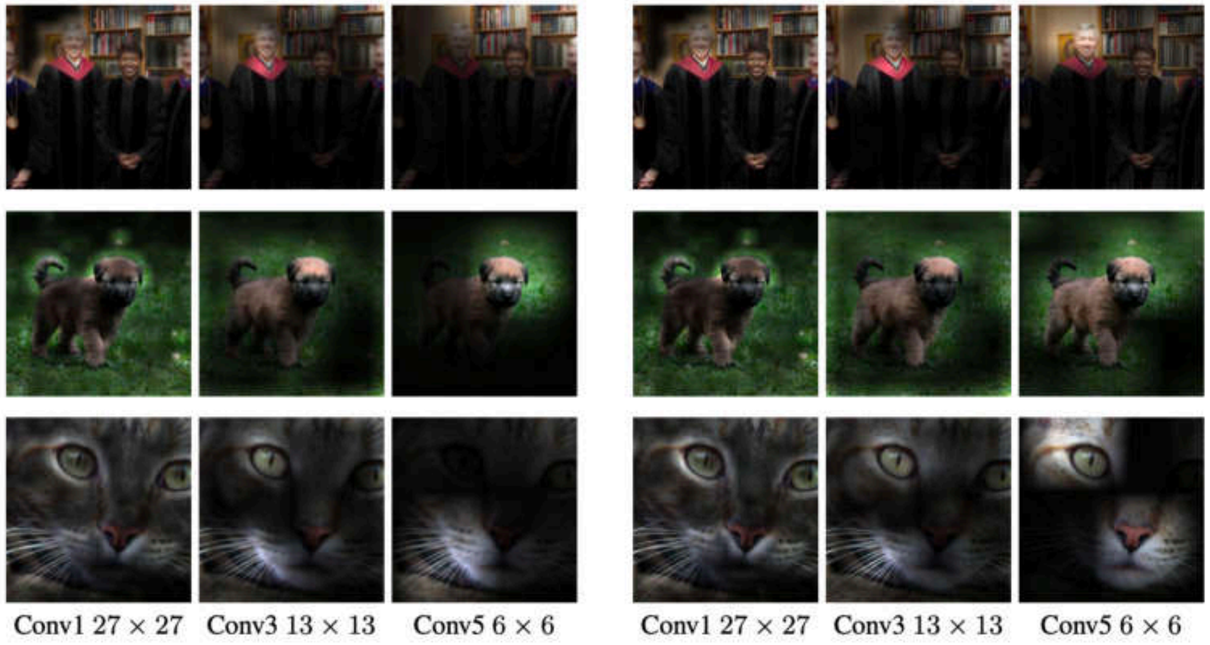
Textures



Learning by predicting rotations

- Results based on AlexNet pretraining:
  - comparable results to supervised learning
  - outperform other ideas by a significant margin
  - the network indeed learns to focus on the structured objects

Method	Conv4	Conv5
ImageNet labels from (Bojanowski & Joulin, 2017)	59.7	59.7
Random from (Noroozi & Favaro, 2016)	27.1	12.0
Tracking Wang & Gupta (2015)	38.8	29.8
Context (Doersch et al., 2015)	45.6	30.4
Colorization (Zhang et al., 2016a)	40.7	35.2
Jigsaw Puzzles (Noroozi & Favaro, 2016)	45.3	34.6
BIGAN (Donahue et al., 2016)	41.9	32.2
NAT (Bojanowski & Joulin, 2017)	-	36.0
(Ours) RotNet	50.0	43.8



(a) **Attention maps of supervised model** (b) **Attention maps of our self-supervised model**

Learning by colorisation

Colorful Image Colorization

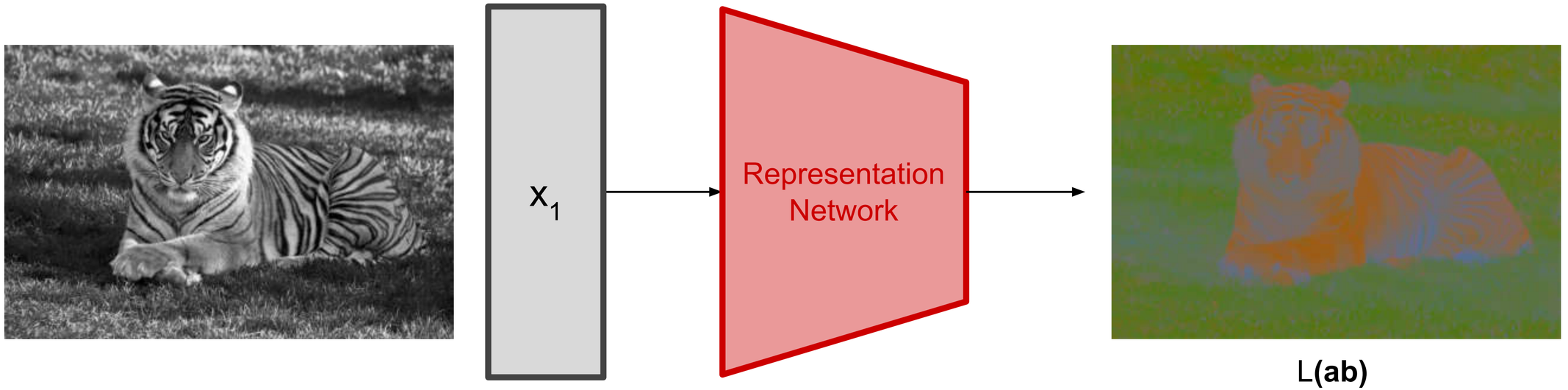
Richard Zhang, Phillip Isola, Alexei A. Efros
{rich.zhang,isola,efros}@eecs.berkeley.edu

University of California, Berkeley

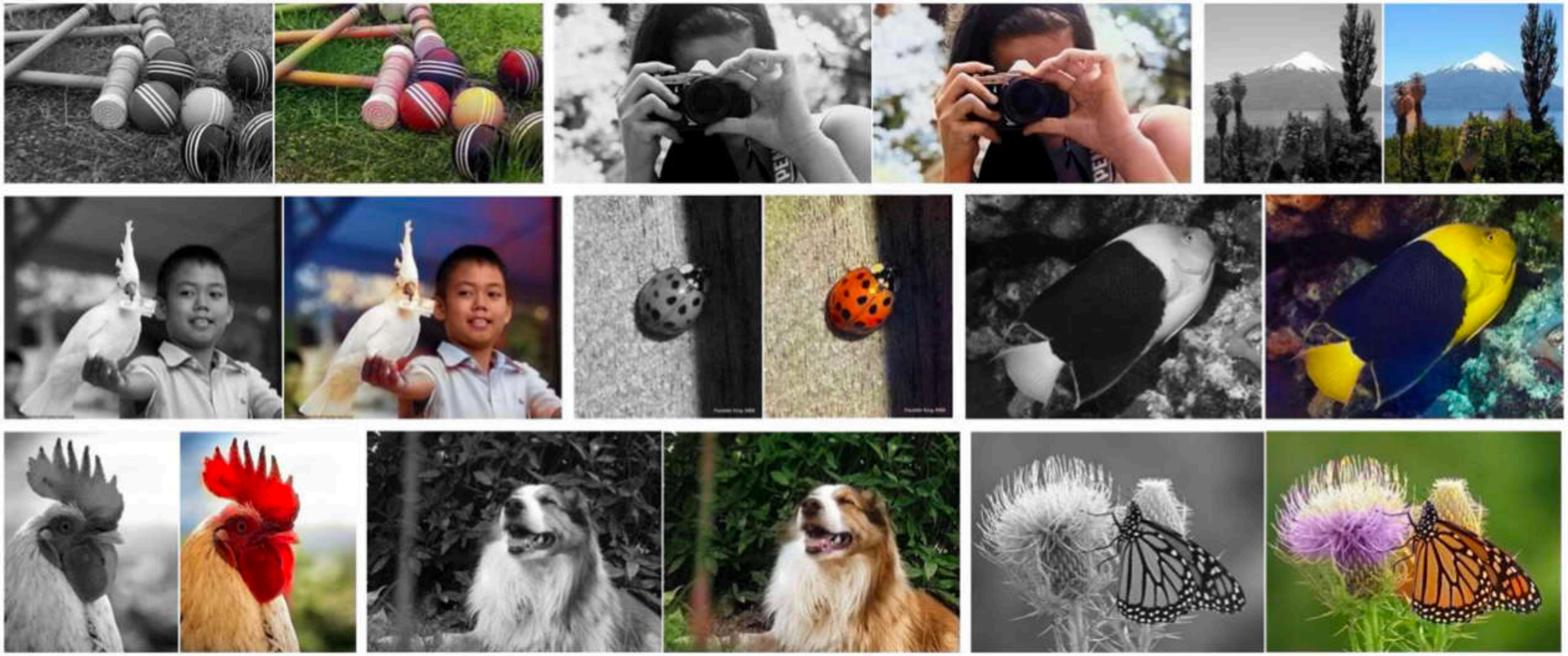
Abstract. Given a grayscale photograph as input, this paper attacks the problem of hallucinating a *plausible* color version of the photograph. This problem is clearly underconstrained, so previous approaches have either relied on significant user interaction or resulted in desaturated colorizations. We propose a fully automatic approach that produces vibrant and realistic colorizations. We embrace the underlying uncertainty of the

Learning by colorisation

- Feed grayscale images as input, and task the model to colorise it.
- Intuitively, if the network can colorise the image, it has understand what the objects look like, e.g. the object patterns.



Learning by colorisation



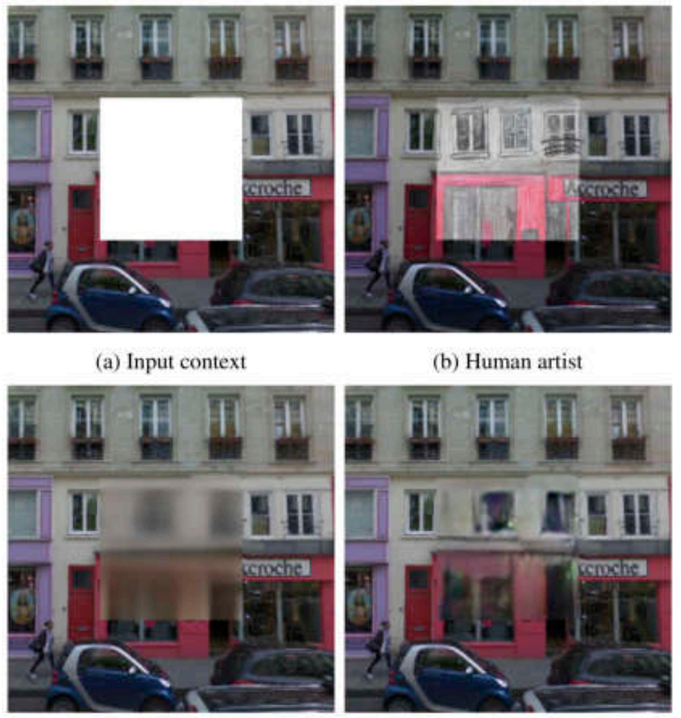
Learning by inpainting

Context Encoders: Feature Learning by Inpainting

Deepak Pathak Philipp Krähenbühl Jeff Donahue Trevor Darrell Alexei A. Efros
University of California, Berkeley
{pathak,philkr,jdonahue,trevor,efros}@cs.berkeley.edu

Abstract

We present an unsupervised visual feature learning algorithm driven by context-based pixel prediction. By analogy with auto-encoders, we propose Context Encoders – a convolutional neural network trained to generate the contents of an arbitrary image region conditioned on its surroundings. In order to succeed at this task, context encoders need to both understand the content of the entire image, as well as produce a plausible hypothesis for the missing part(s). When training context encoders, we have experimented with both a standard pixel-wise reconstruction loss, as well as a reconstruction plus an adversarial loss. The latter produces much sharper results because it can better handle multiple modes in the output. We found that a context encoder learns a representation that captures not just appearance but also the semantics of visual structures. We quantitatively demonstrate the effectiveness of our learned features for CNN pre-training on classification, detection, and segmentation tasks. Furthermore, context encoders can be used for semantic inpainting tasks, either stand-alone or as initialization for non-parametric methods.



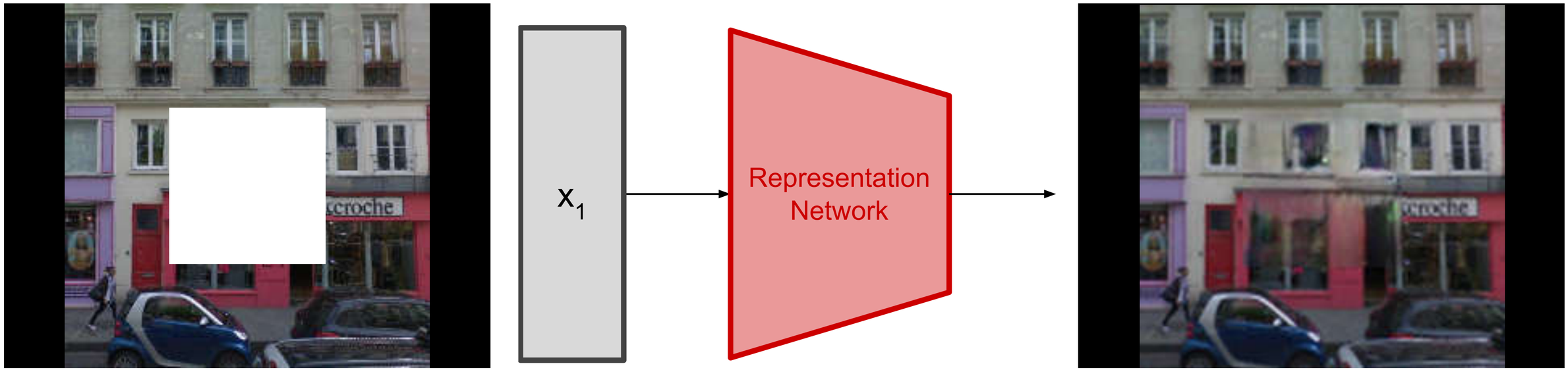
(a) Input context (b) Human artist

(c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

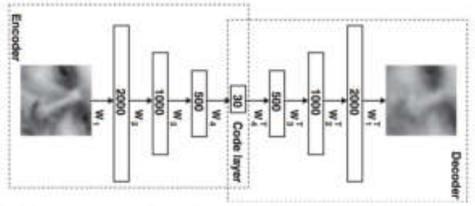
Figure 1: Qualitative illustration of the task. Given an im

Learning by inpainting

- Randomly mask some regions from the image.
- Train a network to recover the missing regions.
- Such training regime turns out to be very good for pre-training Vision Transformers, ie. Masked Autoencoder (MAE)

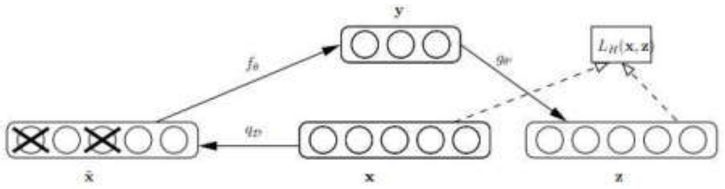


Autoencoders



Hinton & Salakhutdinov. Science 2006.

Denoising Autoencoders

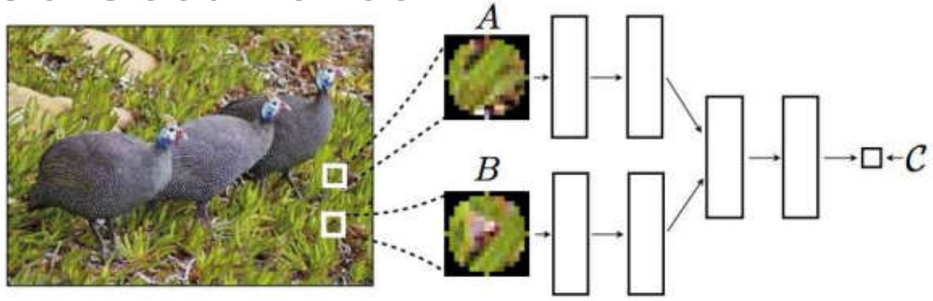


Vincent *et al.* ICML 2008.

Exemplar networks

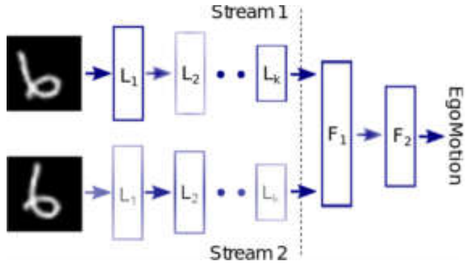
Dosovitskiy *et al.*, NIPS 2014

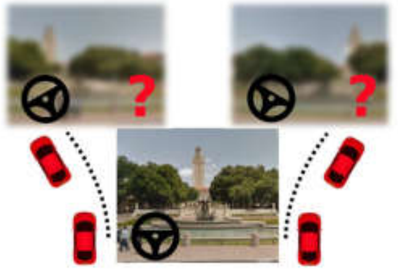
Co-Occurrence



Isola *et al.* ICLR Workshop 2016.

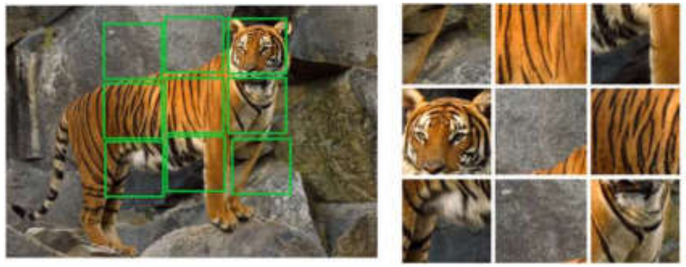
Egomotion





Agrawal *et al.* ICCV 2015 Jayaraman *et al.* ICCV 2015

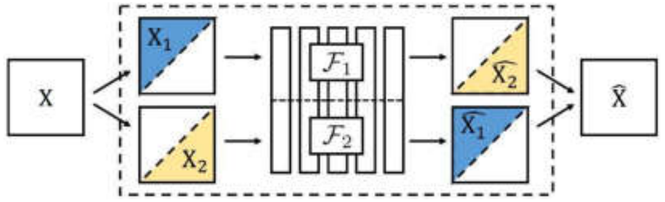
Context



Noroozi et al 2016

Pathak *et al.* CVPR 2016

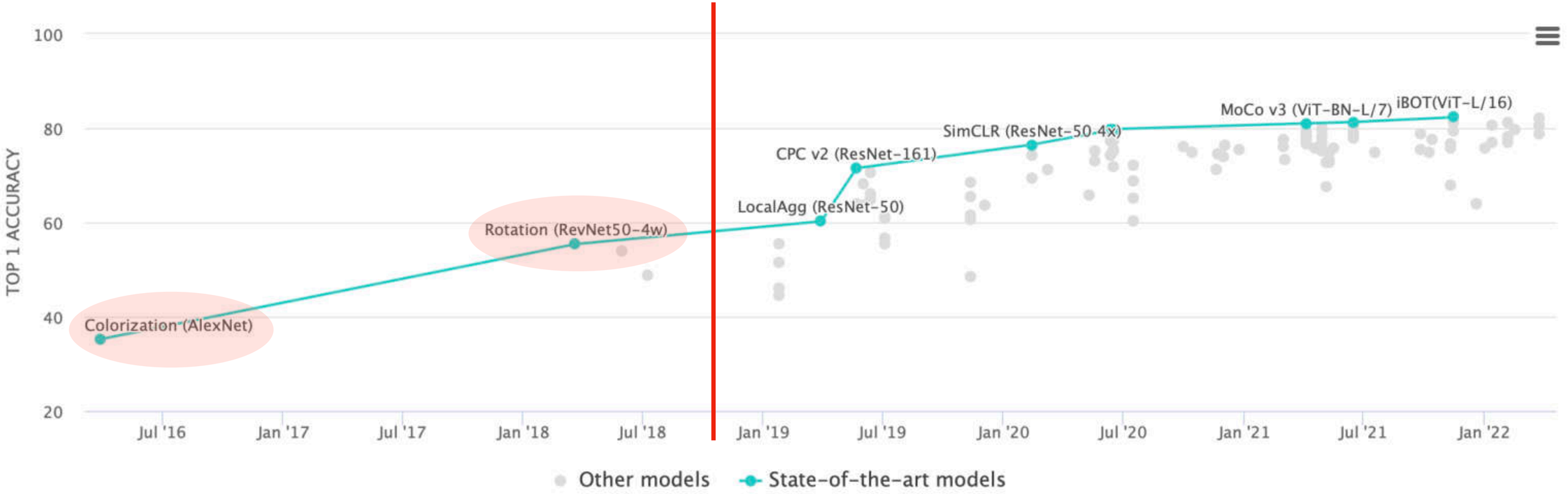
Split-brain auto-encoders



Zhang *et al.* CVPR 2017

Results

- ImageNet Classification : linear probing, *i*.e. freeze the encoder, and only train last classifier with ImageNet.



Self-supervised Representation Learning from Images

(‘modern’ fashioned)

Contrastive Predictive Coding

Representation Learning with Contrastive Predictive Coding

Aaron van den Oord
DeepMind
avdnoord@google.com

Yazhe Li
DeepMind
yazhe@google.com

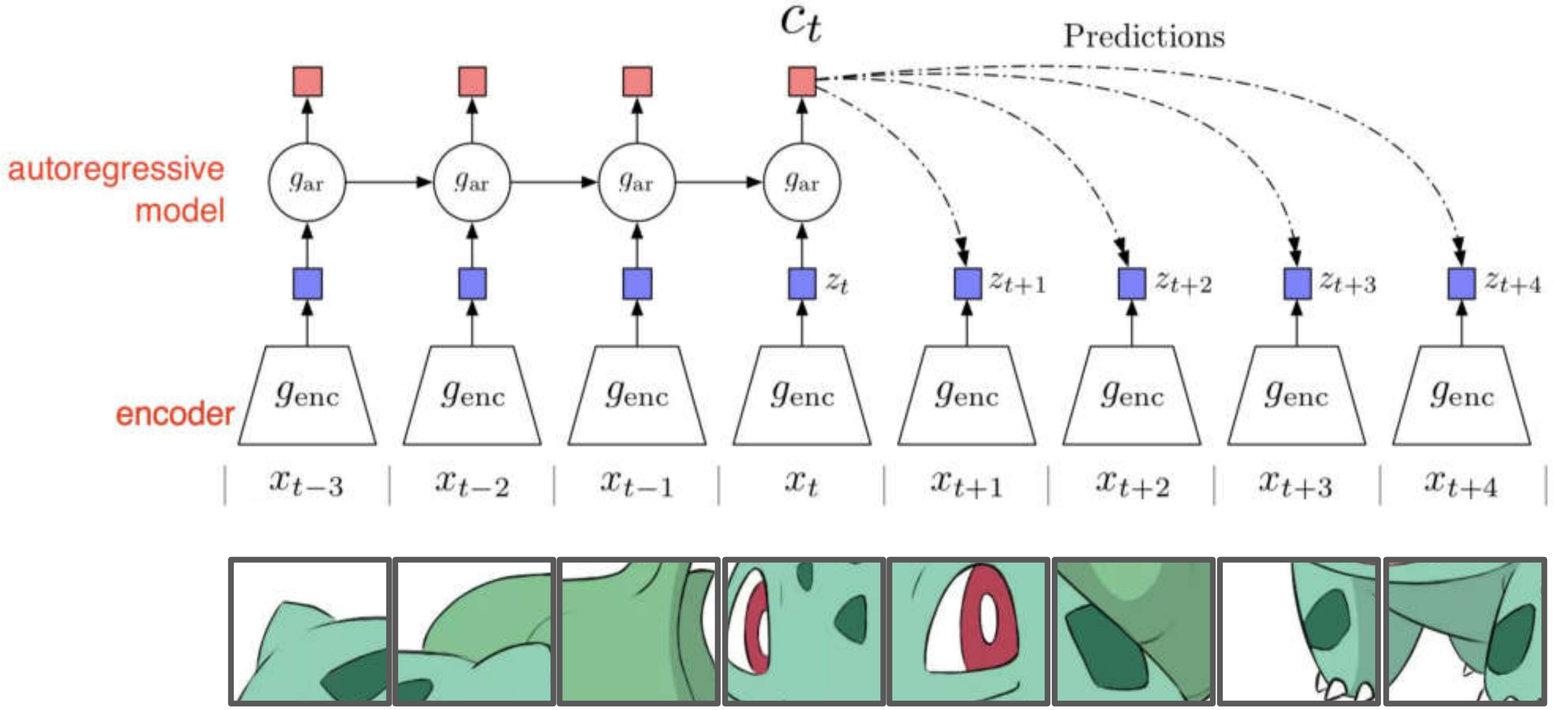
Oriol Vinyals
DeepMind
vinyals@google.com

Abstract

While supervised learning has enabled great progress in many applications, unsupervised learning has not seen such widespread adoption, and remains an important and challenging endeavor for artificial intelligence. In this work, we propose a universal unsupervised learning approach to extract useful representations from high-dimensional data, which we call Contrastive Predictive Coding. The key insight of our model is to learn such representations by predicting the future in *latent* space by using powerful autoregressive models. We use a probabilistic contrastive loss which induces the latent space to capture information that is maximally useful to predict future samples. It also makes the model tractable by using negative sampling. While most prior work has focused on evaluating representations for a particular modality, we demonstrate that our approach is able to learn useful representations achieving strong performance on four distinct domains: speech, images, text and reinforcement learning in 3D environments.

Contrastive Predictive Coding

- Contrastive Predictive Coding (CPC)
  - One of the most influential self-supervised representation learning ideas
  - Recursively predicting the future, and compare the prediction with true observations
  - Contrastive learning: classify the "future" representation amongst a set of unrelated "negative" samples

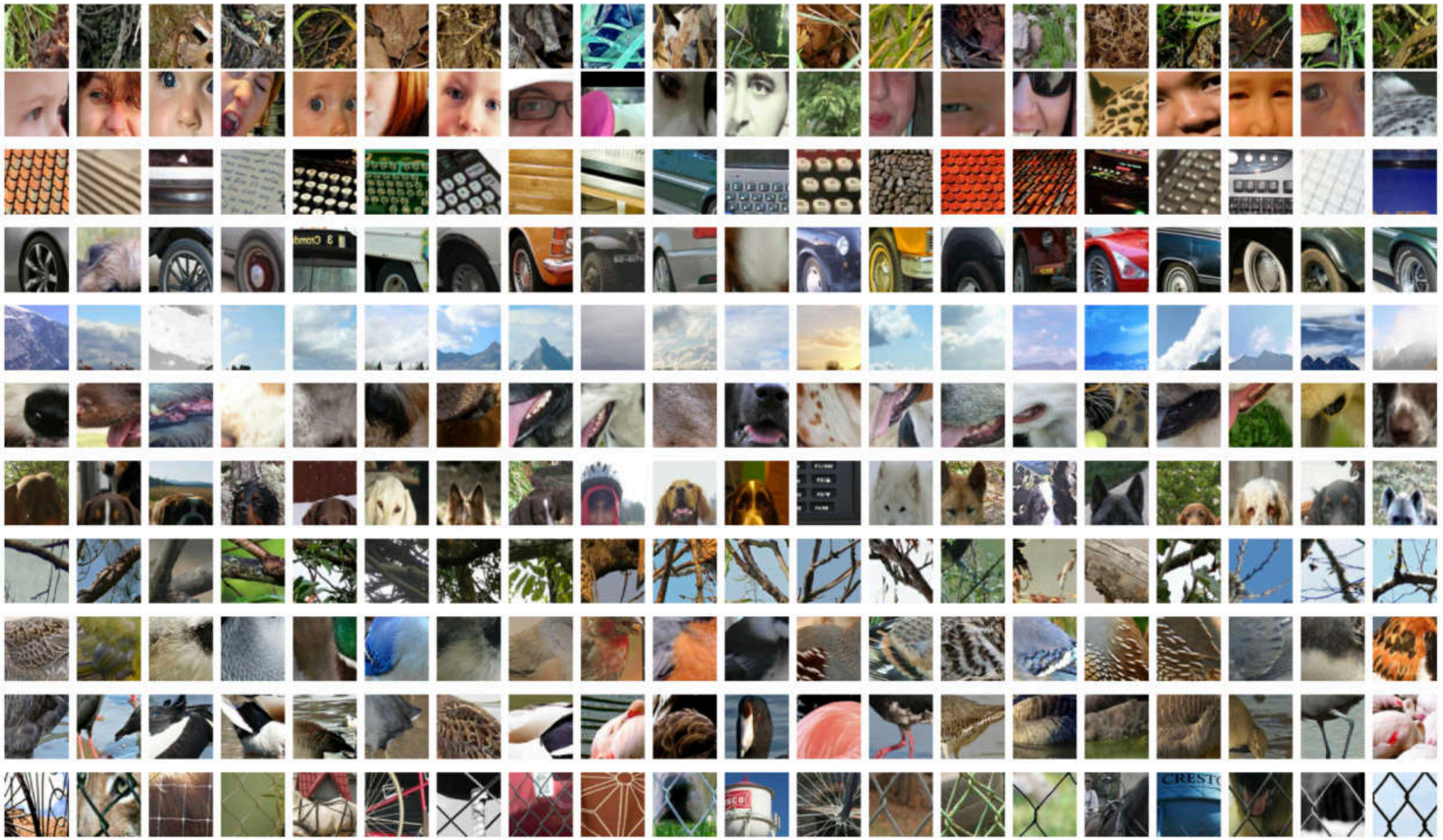


Objective: $\mathcal{D}(\hat{z}_{t+1}, z_{t+1}) << \mathcal{D}(\hat{z}_{t+1}, z_j)$

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(\hat{z}_{t+1}, z_{t+1})/\tau)}{\sum \exp(\text{sim}(\hat{z}_{t+1}, z_j)/\tau)}$$

$$\forall j \neq t+1$$

Contrastive Predictive Coding



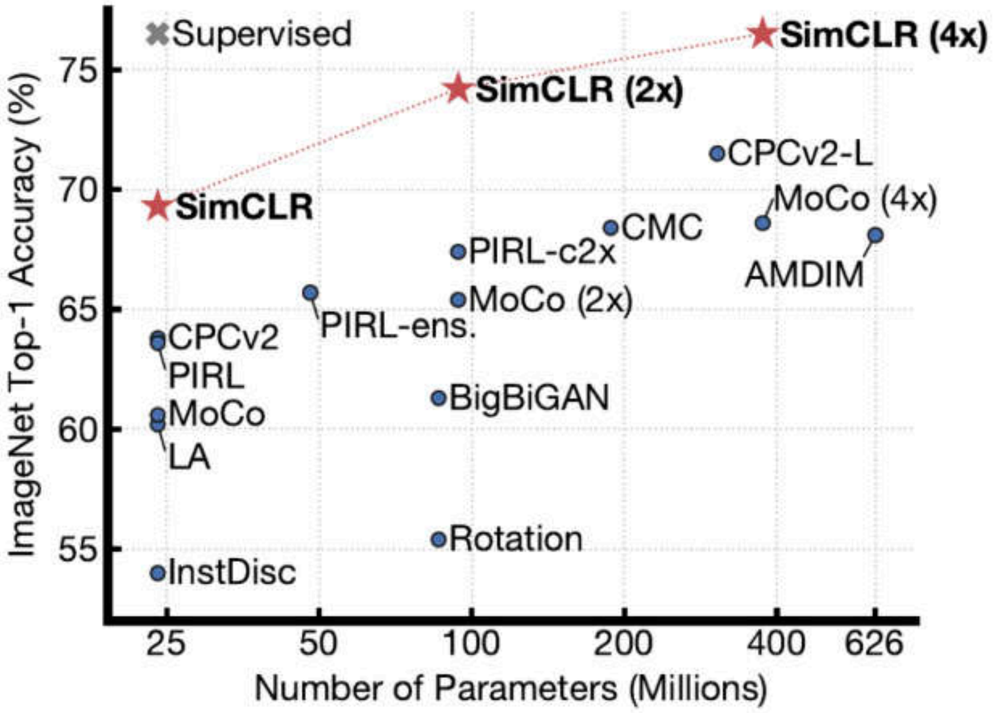
SimCLR

A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen [1] **Simon Kornblith** [1] **Mohammad Norouzi** [1] **Geoffrey Hinton** [1]

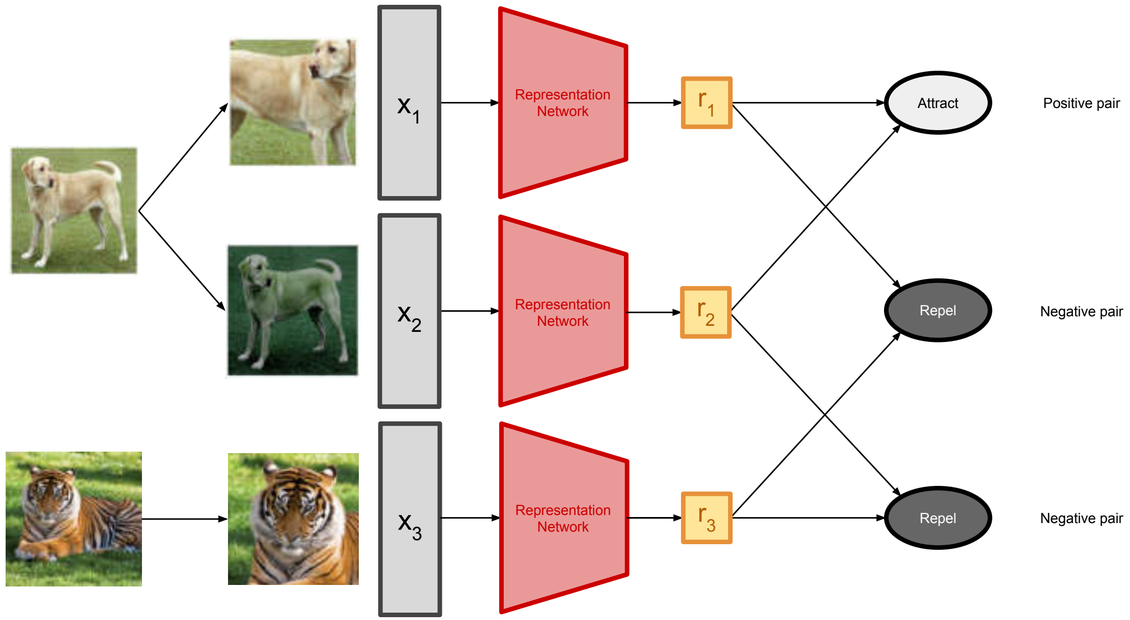
Abstract

This paper presents *SimCLR*: a simple framework for contrastive learning of visual representations. We simplify recently proposed contrastive self-supervised learning algorithms without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of our framework. We show that (1) composition of data augmentations plays a critical role in defining effective predictive tasks, (2) introducing a learnable nonlinear transformation between the repre-

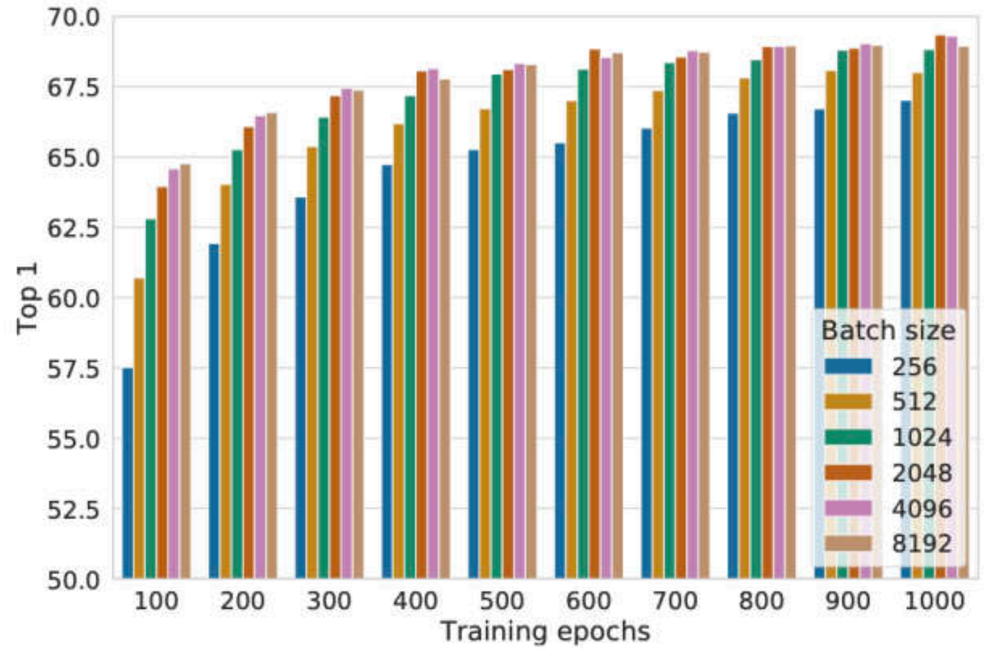


SimCLR

- Predicting future is not necessary
- Still do contrastive learning, find the instance amongst a set of unrelated “negative” samples
- Check more details here : https://www.bilibili.com/video/BV1bD4y1S7nZ?t=23.0



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



It helps to have larger batch size, and training longer.

BYOL

Bootstrap Your Own Latent
A New Approach to Self-Supervised Learning

Jean-Bastien Grill[*,1] , **Florian Strub**[*,1] , **Florent Altché**[*,1] , **Corentin Tallec**[*,1] , **Pierre H. Richemond**[*,1,2]
Elena Buchatskaya[1] , **Carl Doersch**[1] , **Bernardo Avila Pires**[1] , **Zhaohan Daniel Guo**[1]
Mohammad Gheshlaghi Azar[1], **Bilal Piot**[1], **Koray Kavukcuoglu**[1] , **Rémi Munos**[1] , **Michal Valko**[1]

[1]DeepMind [2]Imperial College

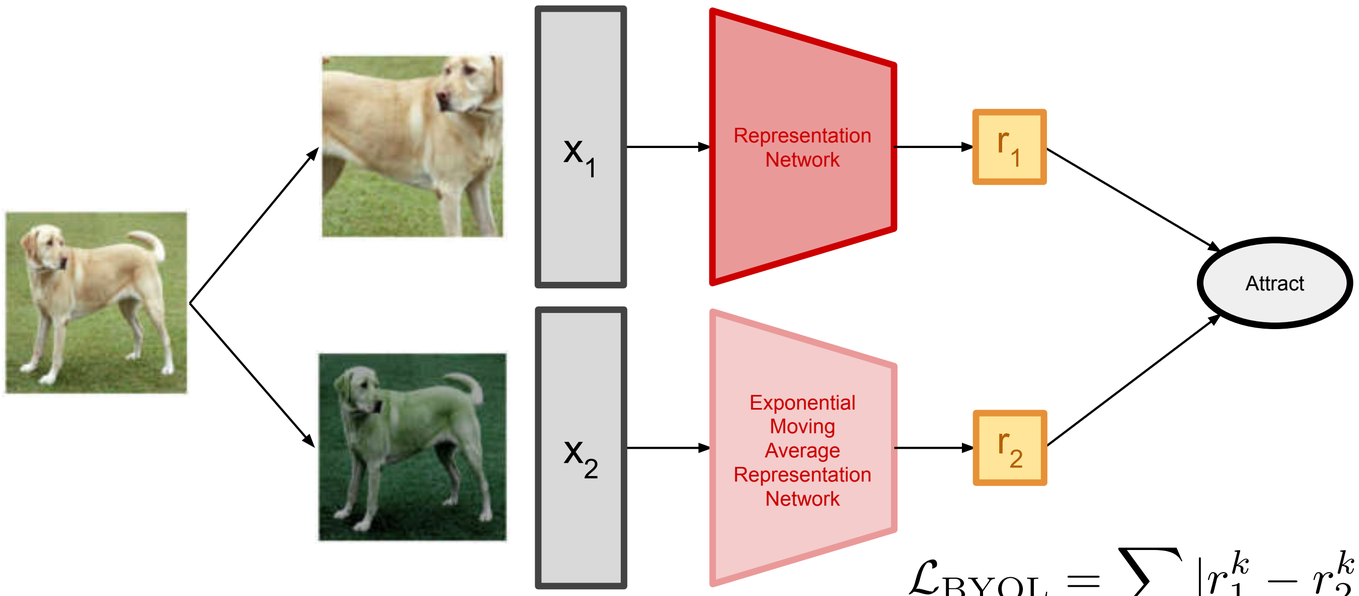
[jbgrill,fstrub,altche,corentint,richemond]@google.com

Abstract

We introduce **B**ootstrap **Y**our **O**wn **L**atent (BYOL), a new approach to self-supervised image representation learning. BYOL relies on two neural networks, referred to as *online* and *target* networks, that interact and learn from each other. From an augmented view of an image, we train the online network to predict the target network representation of the same image under a different augmented view. At the same time, we update the target network with a slow-moving average of the online network. While state-of-the art methods rely on negative pairs, BYOL achieves a new state of the art *without them*. BYOL reaches 74.3% top-1 classification accuracy on ImageNet using a linear evaluation with a ResNet-50 architecture and 79.6% with a larger ResNet. We show that BYOL performs on par or better than the current state of the art on both transfer and semi-supervised benchmarks. Our implementation and pretrained models are given on GitHub.[3]

BYOL

- Contrastive learning is not necessary
- Simply minimise the distance between two augmented view of the same image
- Surprisingly, trivial solution can be avoided with the help of EMA (exponential moving average)
- Check more details here : https://www.bilibili.com/video/BV15D4y1d7GQ?t=2.1



$$\mathcal{L}_{\text{BYOL}} = \sum_k |r_1^k - r_2^k|$$

Method	Architecture	Param.	Top-1	Top-5
SimCLR [8]	ResNet-50 (2×)	94M	74.2	92.0
CMC [11]	ResNet-50 (2×)	94M	70.6	89.7
BYOL (ours)	ResNet-50 (2×)	94M	77.4	93.6
CPC v2 [29]	ResNet-161	305M	71.5	90.1
MoCo [9]	ResNet-50 (4×)	375M	68.6	-
SimCLR [8]	ResNet-50 (4×)	375M	76.5	93.2
BYOL	ResNet-50 (4×)	375M	78.6	94.2
BYOL	ResNet-200 (2×)	250M	**79.6**	**94.8**

Deep Clustering

Deep Clustering for Unsupervised Learning of Visual Features

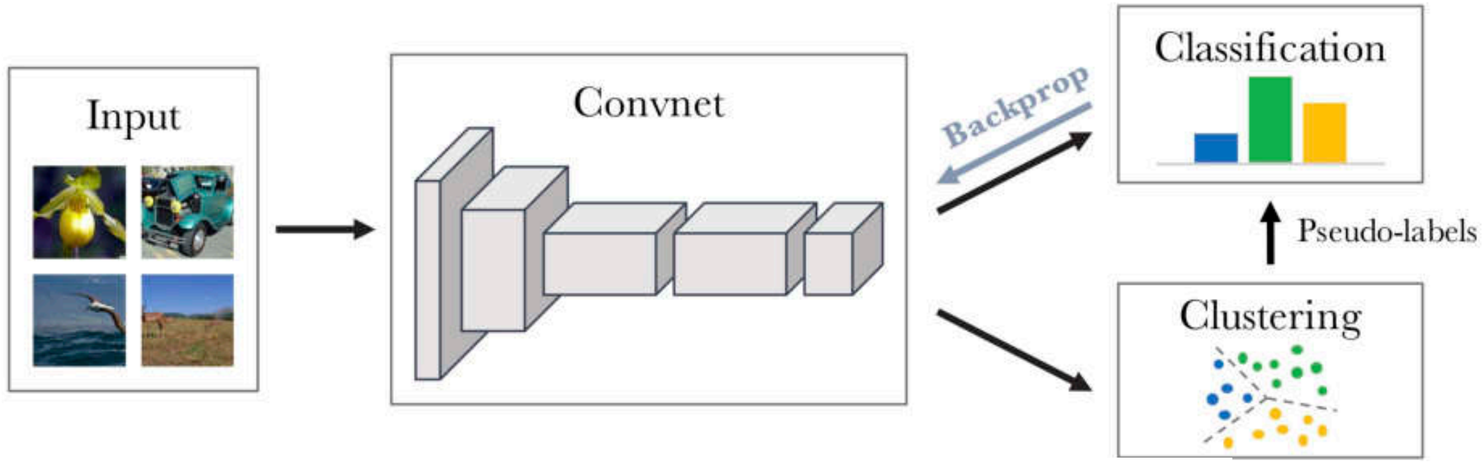
Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze

Facebook AI Research

Abstract. Clustering is a class of unsupervised learning methods that has been extensively applied and studied in computer vision. Little work has been done to adapt it to the end-to-end training of visual features on large scale datasets. In this work, we present DeepCluster, a clustering method that jointly learns the parameters of a neural network and the cluster assignments of the resulting features. DeepCluster iteratively groups the features with a standard clustering algorithm, k-means, and uses the subsequent assignments as supervision to update the weights of the network. We apply DeepCluster to the unsupervised training of convolutional neural networks on large datasets like ImageNet and YFCC100M. The resulting model outperforms the current state of

Deep Clustering

- Iteratively cluster the features via K-means
- Use the cluster assignments as pseudo labels to conduct supervised training
- Resemble to the discriminative K-means



$$\min_{C \in \mathbb{R}^{d \times k}} \frac{1}{N} \sum_{n=1}^{N} \min_{y_n \in \{0,1\}^k} \|f_\theta(x_n) - C y_n\|_2^2 \quad \text{such that} \quad y_n^\top 1_k = 1.$$

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

Mathilde Caron[1,2] **Ishan Misra**[2] **Julien Mairal**[1]

Priya Goyal[2] **Piotr Bojanowski**[2] **Armand Joulin**[2]

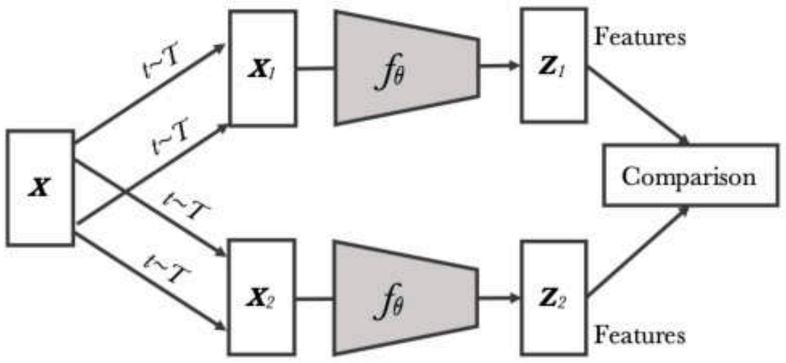
[1] Inria* [2] Facebook AI Research

Abstract

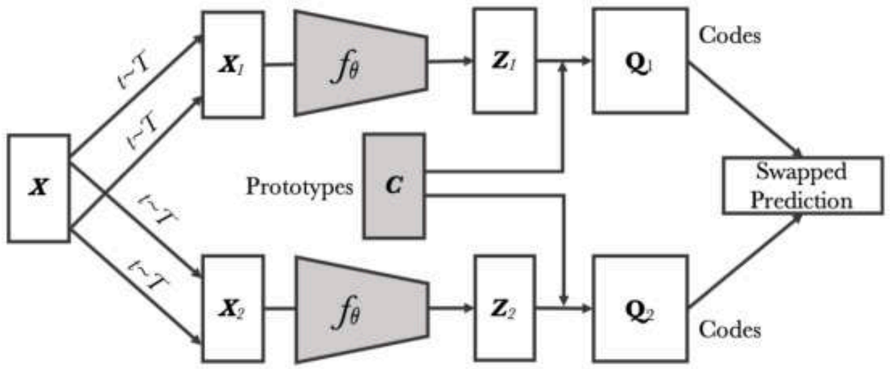
Unsupervised image representations have significantly reduced the gap with su-

SwAV

- Deep Clustering version 2
- Implicit clustering via a learnt *prototype* code ("anchor clusters")
- Cross-predicting the pseudo labels for each of the streams
- With standard ResNet50, SwAV closes the gap between supervised learning



Contrastive instance learning



Swapping Assignments between Views (Ours)

Method	Arch.	Param.	Top1
Supervised	R50	24	76.5
Colorization [65]	R50	24	39.6
Jigsaw [46]	R50	24	45.7
NPID [58]	R50	24	54.0
BigBiGAN [15]	R50	24	56.6
LA [68]	R50	24	58.8
NPID++ [44]	R50	24	59.0
MoCo [24]	R50	24	60.6
SeLa [2]	R50	24	61.5
PIRL [44]	R50	24	63.6
CPC v2 [28]	R50	24	63.8
PCL [37]	R50	24	65.9
SimCLR [10]	R50	24	70.0
MoCov2 [11]	R50	24	71.1
SwAV	R50	24	**75.3**

DINO

Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron[1,2] Hugo Touvron[1,3] Ishan Misra[1] Hervé Jegou[1]
Julien Mairal[2] Piotr Bojanowski[1] Armand Joulin[1]

[1] Facebook AI Research [2] Inria* [3] Sorbonne University

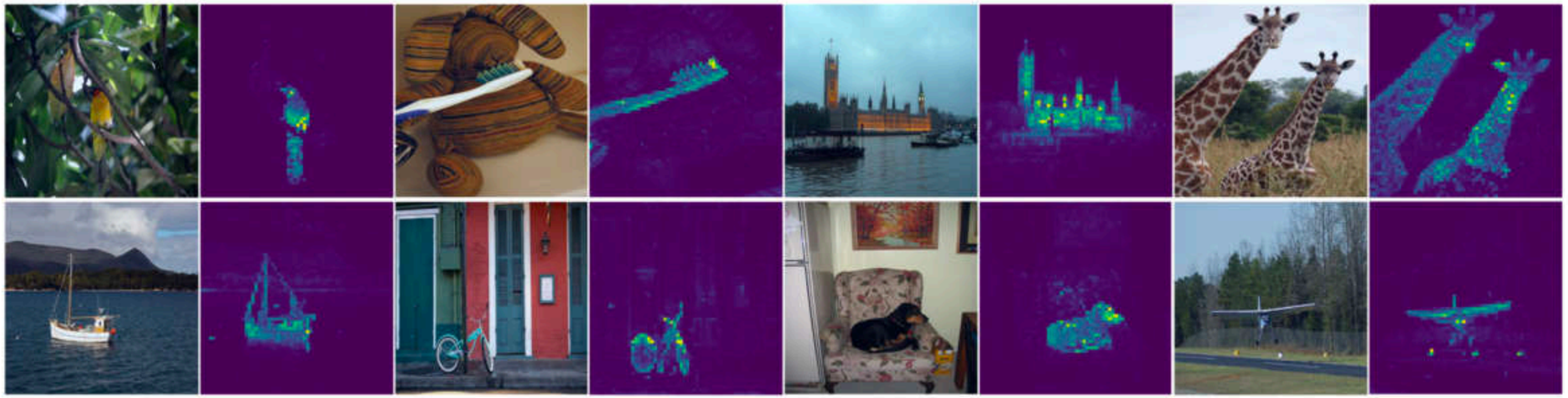
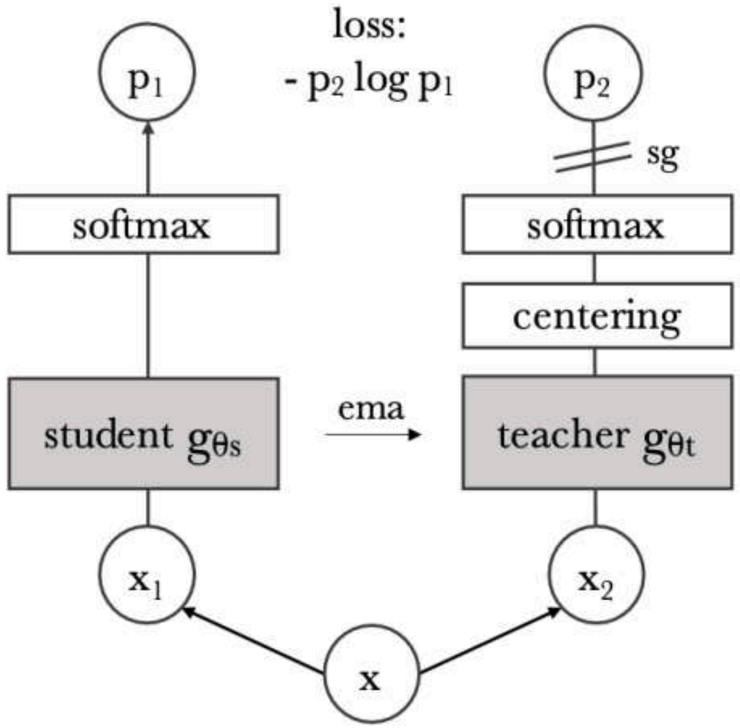


Figure 1: **Self-attention from a Vision Transformer with 8×8 patches trained with no supervision.** We look at the self-attention of the `[CLS]` token on the heads of the last layer. This token is not attached to any label nor supervision. These maps show that the model automatically learns class-specific features leading to unsupervised object segmentations.

DINO

- Deep Clustering version 3……
- We are coming to the Transformer era, use Vision Transformer (ViT).
- ViT outperforms ResNet, on linear probing, SOTA methods achieve 80.1 Top1 ACC on ImageNet.



Method	Arch.	Param.	im/s	Linear	k-NN
Supervised	RN50	23	1237	79.3	79.3
SCLR [12]	RN50	23	1237	69.1	60.7
MoCov2 [15]	RN50	23	1237	71.1	61.9
InfoMin [67]	RN50	23	1237	73.0	65.3
BarlowT [81]	RN50	23	1237	73.2	66.0
OBoW [27]	RN50	23	1237	73.8	61.9
BYOL [30]	RN50	23	1237	74.4	64.8
DCv2 [10]	RN50	23	1237	75.2	67.1
SwAV [10]	RN50	23	1237	**75.3**	65.7
DINO	RN50	23	1237	**75.3**	**67.5**
Supervised	ViT-S	21	1007	79.8	79.8
BYOL* [30]	ViT-S	21	1007	71.4	66.6
MoCov2* [15]	ViT-S	21	1007	72.7	64.4
SwAV* [10]	ViT-S	21	1007	73.5	66.3
DINO	ViT-S	21	1007	**77.0**	**74.5**

Comparison across architectures					
SCLR [12]	RN50w4	375	117	76.8	69.3
SwAV [10]	RN50w2	93	384	77.3	67.3
BYOL [30]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [10]	RN50w5	586	76	78.5	67.1
BYOL [30]	RN50w4	375	117	78.6	–
BYOL [30]	RN200w2	250	123	79.6	73.9
DINO	ViT-S/8	21	180	79.7	**78.3**
SCLRv2 [13]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	**80.1**	77.4

CLIP

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] **Jong Wook Kim** [*1] **Chris Hallacy** [1] **Aditya Ramesh** [1] **Gabriel Goh** [1] **Sandhini Agarwal** [1]
Girish Sastry [1] **Amanda Askell** [1] **Pamela Mishkin** [1] **Jack Clark** [1] **Gretchen Krueger** [1] **Ilya Sutskever** [1]

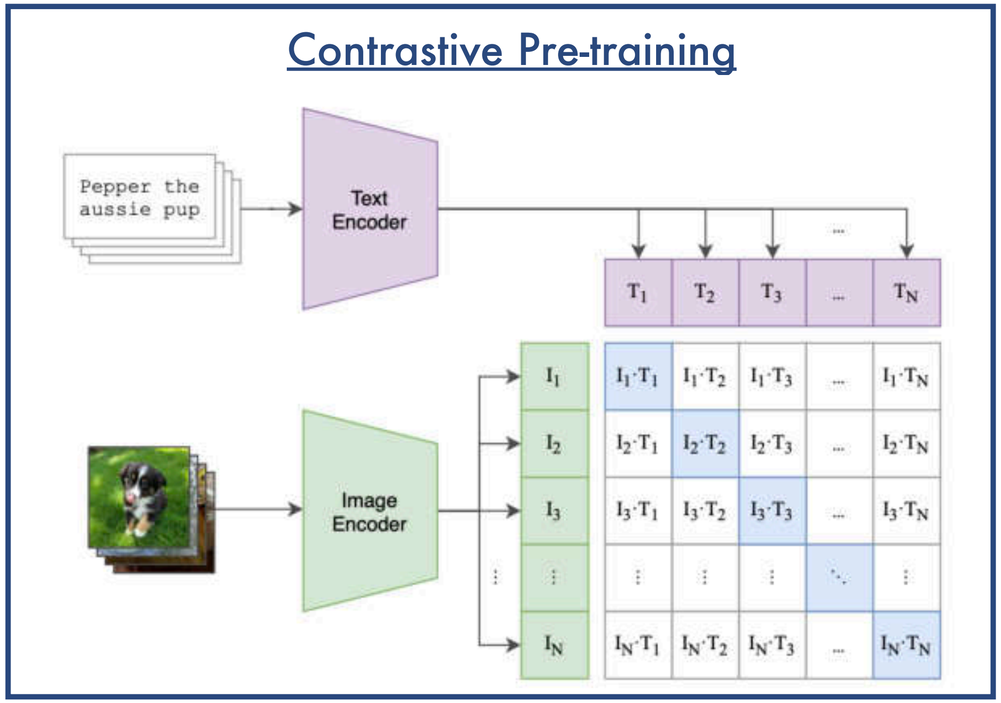
Abstract

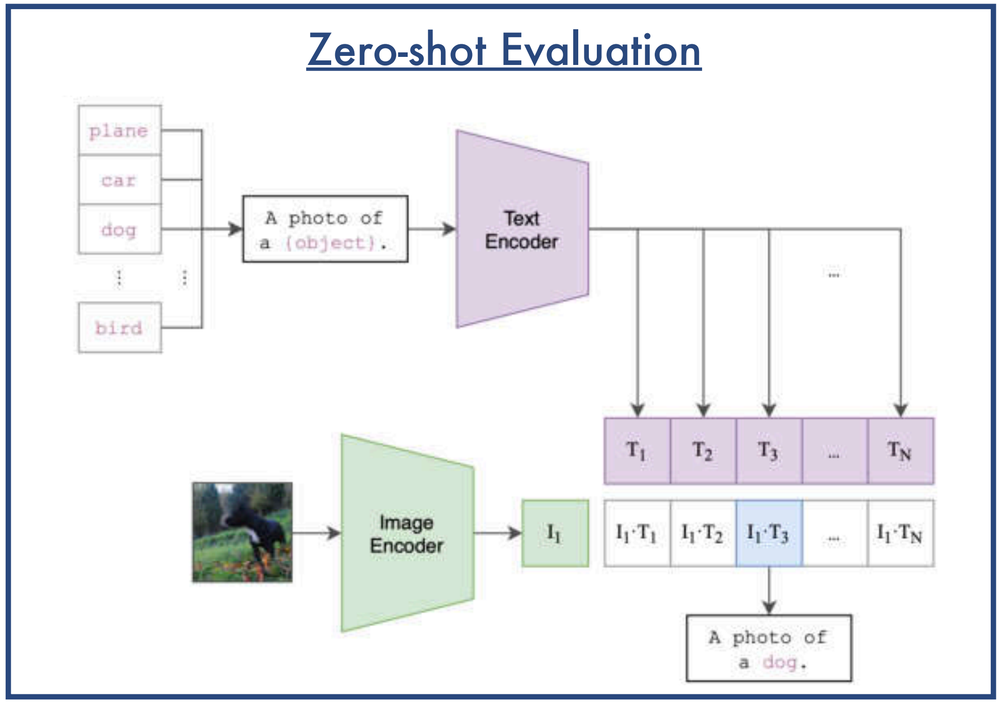
State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since

Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of "text-to-text" as a standardized input-output interface (McCann et al., 2018;

CLIP

- A true game changer, from visual pertaining to visual-language representation pre-training
- Classifiers can be generated from text embedding, take benefit from the language structure, e.g. fine-grained and shared attributes between different semantic classes, thus enable zero-shot generalisation
- For the first time, the community finds a scalable way to train the **Foundation Models**
- Check more details : https://openai.com/blog/clip/





CLIP

		Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Stanford Cars	FGVC Aircraft	VOC2007	DTD	Oxford Pets	Caltech101	Flowers102	MNIST	FER2013	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCam	UCF101	Kinetics700	CLEVR	HatefulMemes	Rendered SST2	ImageNet
CLIP-ResNet	RN50	81.1	75.6	41.6	32.6	59.6	55.8	19.3	82.1	41.7	85.4	82.1	65.9	66.6	42.2	94.3	41.1	54.2	35.2	42.2	16.1	57.6	63.6	43.5	20.3	59.7	56.9	59.6
	RN101	83.9	81.0	49.0	37.2	59.9	62.3	19.5	82.4	43.9	86.2	85.1	65.7	59.3	45.6	96.7	33.1	58.5	38.3	33.3	16.9	55.2	62.2	46.7	28.1	61.1	64.2	62.2
	RN50x4	86.8	79.2	48.9	41.6	62.7	67.9	24.6	83.0	49.3	88.1	86.0	68.0	75.2	51.1	96.4	35.0	59.2	35.7	26.0	20.2	57.5	65.5	49.0	17.0	58.3	66.6	65.8
	RN50x16	90.5	82.2	54.2	45.9	65.0	72.3	30.3	82.9	52.8	89.7	87.6	71.9	80.0	56.0	97.8	40.3	64.4	39.6	33.9	24.0	62.5	68.7	53.4	17.6	58.9	67.6	70.5
	RN50x64	91.8	86.8	61.3	48.9	66.9	76.0	35.6	83.8	53.4	93.4	90.6	77.3	90.8	61.0	98.3	59.4	69.7	47.9	33.2	29.6	65.0	74.1	56.8	27.5	62.1	70.7	73.6
CLIP-ViT	B/32	84.4	91.3	65.1	37.8	63.2	59.4	21.2	83.1	44.5	87.0	87.9	66.7	51.9	47.3	97.2	49.4	60.3	32.2	39.4	17.8	58.4	64.5	47.8	24.8	57.6	59.6	63.2
	B/16	89.2	91.6	68.7	39.1	65.2	65.6	27.1	83.9	46.0	88.9	89.3	70.4	56.0	52.7	98.2	54.1	65.5	43.3	44.0	23.3	48.1	69.8	52.4	23.4	61.7	59.8	68.6
	L/14	92.9	96.2	77.9	48.3	67.7	77.3	36.1	84.1	55.3	93.5	92.6	78.7	87.2	57.5	99.3	59.9	71.6	50.3	23.1	32.7	58.8	76.2	60.3	24.3	63.3	64.0	75.3
	L/14-336px	93.8	95.7	77.5	49.5	68.4	78.8	37.2	84.3	55.7	93.5	92.8	78.3	88.3	57.7	99.4	59.6	71.7	52.3	21.9	34.9	63.0	76.9	61.3	24.8	63.3	67.9	76.2

Table 11. Zero-shot performance of CLIP models over 27 datasets.