The Role of Entropy and Reconstruction for Multi-View Self-Supervised Learning

Borja Rodríguez-Gálvez, Arno Blaas, Pau Rodríguez, Adam Goliński, Xavier Suau, Jason Ramapuram, Dan Busbridge, Luca Zappella ICML 2023





Motivation: MVSSL progressing rapidly

								Exploring S	Simple Siamese	Representation	n Learning		
									Xinlei Chen Facebook AI Re	Kaiming He esearch (FAIR)			
		Emerg	ging Properties in Self-Supervised Visio	on Transformers				Abstract					
		Mathi	lde Caron ^{1,2} Hugo Touvron ^{1,3} Ishan Misra Julien Mairal ² Piotr Bojanowski ¹ Arma	a ¹ Hervé Jegou ¹ and Joulin ¹			S. varie tion twee	iamese networks have become a c ous recent models for unsupervise learning. These models maximiz en two augmentations of one imag	ommon stru d visual rep e the similc e, subject to				
021		1]	Facebook AI Research ² Inria [*] ³ Sorbo	onne University			Normalization of the second se	ditions for avoiding collapsing solu report surprising empirical results vorks can learn meaningful repress e of the following: (i) negative san hes. (iii) momentum encoders. On	tions. In the that simple trations even the pairs, (r, experiment				
May 2					li su		that	allaneina ealutione do sviet for th	ial rol on th	r 2019	Deep Clustering fo of Vis	r Unsupe ual Featu	vised Learnin es
V] 24		. 4							ethod i eam ti e to rei ed rep	18 Ma	Mathilde Caron, Piotr Bojanow Faceb	ski, Armand Jo ook AI Research	ulin, and Matthijs Do
[cs.C	Figure 1: Self-a the [CLS] toke automatically le	ttention from the here arms class-s	om a Vision Transformer with 8×8 patches trained with no sads of the last layer. This token is not attached to any label nor sspecific features leading to unsupervised object segmentations.	supervision. We look at the supervision. These maps sh	Con	trastive Multiv	iview Coding		s in 1	S.CV]	Abstract. Clustering is a clar has been extensively applied ar has been done to adapt it to t on large scale datasets. In	s of unsupervise d studied in com he end-to-end tr	l learning methods that puter vision. Little work aining of visual features
-72		1	Abstract 1. Introduction	a					coura 5, 7]).		tering method that jointly		
6					Yonglong Tian MIT CSAIL	Dilip Krish Google Rese	hnan Phillip Is earch MIT CS.	sola AIL	gene Sian	2	eratively groups the featur		
42	In this j vides new j				yonglong@mit.edu	dilipkay@goog	gle.com phillipi@m	uit.edu	olied o	00	the weights of the network.		
4.1	stand out (Revond the				Abstract		\frown		Rece	52	and YFCC100M. The resul		
10	architectu				Abstract	som channels			of one	02	the art by a significant mar		
5	ing observ. explicit inf				e.g., the long-wavelength light channel, view	ved by the left	1000		e netv		Keywords: unsupervised		
14	image, wh				right ear. Each view is noisy and incomple	te, but impor-		24 50.	ere ha	80			
arX	ViTs, nor w cellent k-N with a smai momentum		A Simple Framework for Contrastive	e Learning of Vis	 to be shared between all views (e.g., a "dog heard, and felt). We investigate the classic h a powerful representation is one that models to factors. We study this hypothesis under the mathieum centering learning under the 	" can be seen, ypothesis that view-invariant framework of	$\{\underbrace{\underbrace{\underbrace{j_{i_1}}}_{e_1^i \in V_1}, \underbrace{j_{i_2}}_{e_2^i \in V_2}, \underbrace{j_{i_2}}$	$\left.\begin{array}{c} \begin{array}{c} & & \\ & & $	3., inst tegativ positiv from 1	$\mathbf{X}_{\mathrm{Pr}}^{\mathrm{1}}$	Introduction re-trained convolutional neura g blocks in most computer visi		Wha
	use of sma into a simį we interpr We show th		Ting Chen ¹ Simon Kornblith ¹ Moh	ammad Norouzi ¹ Geofi	 mainty we contrastive tearning, where we tear tation that aims to maximize mutual informa- different views of the same scene but is other Our approach scales to any number of views agnostic. We analyze key properties of the make it work, finding that the contrastive loss 	ation between ation between wise compact. Fi s, and is view- lest approach that spi s outperforms an	Matching views Figure 1: Given a set of sensory vi earnt by bringing views of the same pace, while pushing views of differe und example of a 4-view dataset (NY	riews, a deep representation is e scene together in embedding ent scenes apart. Here we show (JL BGBD (53)) and its learned	avoid line cl	els la: vr	meral-purpose features that ca s learned on a limited amoun rge fully-supervised dataset, h nets. However, Stock and Ciss		Yonglo
	80.1% top-		ADSIT ACL This paper presents SimCLR: a simple framework	⁷⁵	a popular alternative based on cross-view p that the more views we learn from, the better	rediction, and rep r the resulting to	epresentation. The encodings for ea o form the full representation of a s	ach view may be concatenated scene.		th	at the performance of state-c erestimated, and little error is		MIT
			for contrastive learning of visual representations.		representation captures underlying scene sema proach achieves state-of-the-art results on im	untics. Our ap- age and video				pe	erformance has been saturatin	120	Dili
	*Univ.	20	We simplify recently proposed contrastive self- supervised learning algorithms without requiring	ASINGLA	unsupervised learning benchmarks. Code in http://github.com/HobbitLong/CMC/.	is released at: co	conditions or thermal noise in a	camera's sensor.		po	osed in recent years [289]. A:	20	Goog
	Grenoble, Fra Corresponder	50	specialized architectures or a memory bank. In	65 CPCv2 PIRL-	0 1. Introduction	on	ones that are shared between m	ultiple views of the world,		do	pmain of object classification.	S	
	Code: http:	Ţ	prediction tasks to learn useful representations,		A foundational idea in coding theory is	to learn com- so	or example between multiple ser ound, and touch [70]. Under the	nsory modalities like vision, iis perspective "presence of		ge	er and more diverse dataset, p	Ď	
		ſ	we systematically study the major components of our framework. We show that (1) composition of	AL add	pressed representations that nonetheless can construct the raw data. This idea shows up in	be used to re- do contemporary an	log" is good information, since and felt, but "camera pose" is ba	e dogs can be seen, heard, ad information, since a cam-		th	e expert knowledge in crowds	00	Contras
		_	data augmentations plays a critical role in defining	S5 ●InstDisc	representation learning in the form of autoence	oders [65] and en	era's pose has little or no effect	on the acoustic and tactile This hypothesis corresponds		ye	ears 10. Replacing labels by	_	state of Despite
		Ū,	able nonlinear transformation between the repre-	25 50 Number c	point or distribution as losslessly as possible	e. Yet lossless to	o the inductive bias that the wa	y you view a scene should		ca	in be trained on internet-scale	\geq	this pap
		Ļ	sentation and the contrastive loss substantially im- proves the quality of the learned representations	Figure 1 ImageNet Top	it is trivial to achieve – the raw data itself is	a lossless rep- the	he cognitive science and neuro	science literature that such		m	Unsupervised learning has be unity [12] and algorithms for	<u>O</u>	(MI) be
		C	and (3) contrastive learning benefits from larger	on representations learned	"resentation. What we might instead prefer "good" information (signal) and throw away th	is to keep the Vie he rest (noise). [7]	70, 15, 32]). In this paper, we sp	e encoded by the brain (e.g., pecifically study the setting			unity [12], and algorithms for	CS	hypothe
		3	supervised learning. By combining these findings,	ResNet-50. Our method, S	How can we identify what information is sign noise?	al and what is what is as	where the different views are diff as luminance, chrominance, dept	ferent image channels, such h, and optical flow. The fun-				3	tion as a leads to
		90	we are able to considerably outperform previous methods for self-supervised and semi-supervised	However, pixel-level ge	To an autoencoder, or a max likelihood gen a bit is a bit. No one bit is better than any of	therative model, da da da	lamental supervisory signal we e n natural data, of multiple viev	exploit is the co-occurrence, ws of the same scene. For				34	product for Ima
		10	learning on ImageNet. A linear classifier trained	sive and may not be nec Discriminative approach	jecture in this paper is that some bits are in fa others. Some bits code important propertie	act better than ex es like seman- pa	example, we consider an image baired example of the co-occur	in Lab color space to be a rrence of two views of the				54	
		02	CLR achieves 76.5% top-1 accuracy, which is a	tive functions similar to	tics, physics, and geometry, while others code	attributes that sco	cene, the L view and the ab view	w: $\{L, ab\}$. representations that capture				10	1 Introduct
		5	7% relative improvement over previous state-of- the-art, matching the performance of a supervised	puts and labels are deriv	we might consider less important, like mete	icinal lighting	Our goar is increase to rearring	representations that capture				2	It is commonsen
		20(ResNet-50. When fine-tuned on only 1% of the	such approaches have re tasks (Doersch et al., 20		1						8	Luis Borges images becomes bothered
			forming AlexNet with $100 \times$ fewer labels. ¹	Favaro, 2016; Gidaris e									the dog at three f
		X		approaches based on coi								N.	seen before. He
		ar	1. Introduction	have recently shown great pror art results (Hadsell et al., 200	mise, achieving state-of-the- 6; Dosovitskiy et al., 2014;							ar	Most of us, fortu
			Learning effective visual representations without human	Oord et al., 2018; Bachman et	al., 2019).							-	representations i
			supervision is a long-standing problem. Most mainstream approaches fall into one of two classes: generative or dis-	In this work, we introduce a trastive learning of visual rep	simple framework for con- resentations, which we call								representations o paradigm is cont
			criminative. Generative approaches learn to generate or	SimCLR. Not only does SimCL	R outperform previous work								in representation
			2006; Kingma & Welling, 2013; Goodfellow et al., 2014).	(Figure 1), but it is also simple ized architectures (Bachman et	er, requiring neither special- al., 2019; Hénaff et al., 2019)								This is a natura conditions shoul
			¹ Google Research, Brain Team. Correspondence to: Ting Chen	nor a memory bank (Wu et al.,	, 2018; Tian et al., 2019; He								of day then we c
			amtingchen@google.com>.	In order to understand what an	ables good contrastive repre-								ability to track a
			Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by	sentation learning, we systema	tically study the major com-								We therefore see
			the author(s). ¹ Code available at https://github.com/google-research/simclr.	ponents of our framework and	show that:								¹ Project page:
			. · · ·····										34th Conference o



Unsupervised Learning al Features

1 Introduction

It is commonsense that how you look at an object does not change its identity. Nonetheless, Jorge Luis Borges imagined the alternative. In his short story on *Funes the Memorious*, the titular character becomes bothered that a "dog at three fourteen (seen from the side) should have the same name as the dog at three fifteen (seen from the front)" [8]. The curse of Funes is that he has a perfect memory, and every new way he looks at the world reveals a percept minutely distinct from anything he has seen before. He cannot collate the disparate experiences.

Most of us, fortunately, do not suffer from this curse. We build mental representations of identity Most of us, fortunately, do not suffer from this curse. We build mental representations of identity that discard musiances like time of day and viewing angle. The ability to build up view-inwariant representations is central to a rich body of research on multiview learning. These methods seek representations of the world that are invariant to a family of viewing conditions. Currently, a popular paradigm is contrastive multiview learning, where two views of the same scene are brought together in representation space, and two views of different scenes are pushed apart.

What Makes for Good Views for Contrastive

Learning?

Chen Sun Google, Brown University

Cordelia Schmid Google Research

Abstract

Contrastive learning between multiple views of the data has recently achieved state of the art performance in the field of self-supervised representation learning. Despite its success, the influence of different view choices has been less studied. In this paper, we use theoretical and empirical analysis to better understand the impor-tance of view selection, and argue that we should reduce the mutual information (MI) between views while keeping task-relevant information intact. To verify this hypothesis, we devise unsupervised and semi-supervised frameworks that learn effective views by aiming to reduce their MI. We also consider data augmenta-tion as a way to reduce MI, and show that increasing data augmentation indeed leads to decreasing MI and improves downstream classification accuracy. As a by-product, we achieve a new state-of-the-art accuracy on unsupervised pr-training for ImageNet classification (73% top-1 linear readout with a ResNet-50)¹.

Yonglong Tian MIT CSAIL

Dilip Krishnan Google Research

Ben Poole Google Research

Phillip Isola MIT CSAIL

This is a natural and powerful idea but it leaves open an important question: "which viewing conditions should we be invariant to?" It's possible to go too far: if our task is to classify the time of day then we certainly should not use a representation that is invariant to time. Or, like Funes, we could go not far enough: representing each specific viewing angle independently would cripple our ability to track a dog as it moves about a scene.

We therefore seek representations with enough invariance to be robust to inconsequential variations but not so much as to discard information required by downstream tasks. In contrastive learning,

¹Project page: http://hobbitlong.github.io/InfoMir

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada,

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 A	ook AI Research								
Abstract										
Unsupervised image repres	entations have significantly r	educed th								

evements of contr tk online and rely c AV, that takes adv irwise comparison e enforcing consis ntations (or "views contrastive learnir reverpredict the c hod can be trained s of data. Compare efficient since it d vork. In addition, w b, that uses a mix on views, without pur findings by acl s well as surpassi

vised learning, air the performance g of-the-art methods t (or "instance") ar are able to discrim rmations. Recent s wo elements: (i) a c emoves the notion ormations define t of the resulting net ansformations.

38000 Grenoble, Fra urIPS 2020), Vancoi

10

G





In this paper, we introduce Bootstrap Your Own Latent (BYOL), a new algorithm for self-supervised learning of image representations. BYOL achieves higher performance than state-of-the-art contrastive methods (Sup.) baselines [8].

*Equal contribution; the order of first authors was randomly selected. ³https://github.com/deepmind/deepmind-research/tree/mas

... and many more

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} Elena Buchatskaya 1 Carl Doersch 1 Bernardo Avila Pires 1 Zhaohan Daniel Guo 1 Mohammad Gheshlaghi Azar¹ Bilal Piot¹ Koray Kavukcuoglu¹ Rémi Munos¹ Michal Valko¹ ¹DeepMind ²Imperial College

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

Abstract

We introduce Bootstrap Your Own Latent (BYOL), a new approach to self-supervised image We introduce Bootstap Your Own Latent (510L), a new approach to sen-supervised image representation learning. PSUL relies on two neural networks, referred to as *online* and *larget* 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% (no)- classification accuracy on ImageNet using a linear evaluation with a ResNet-50 architecture and 79.6% with a larger ResNet. We show then BYOL netforms on pay or batter than the current state of the art on both transfer and 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

1 Introduction **1** IntrOduction Learning good image representations is a key challenge in computer vision [1, 2, 3] as it allows for efficient training on downstream tasks [4, 5, 6, 7]. Many dif-ferent training approaches have been proposed to learn such representations, usually relying on visual pretext tasks. Among them, state-of-the-art contrastive meth-ods [8, 9, 10, 11, 12] are trained by reducing the dis-tance between representations of different augmented views for different images ('negative pairs'). These methods need careful treatment of negative pairs [13] by either relying on large batch sizes [8, 12], memory banks [9] or customized mining strategies [14, 15] to re-trieve the negative pairs. In this neart, we introduce Bootstrap Your Own In this namer, we introduce Bootstrap Your Own

Motivation: But poorly understood theoretically

- Lens of Information Theory
- \mathbf{V} on the Mutual Information (MI)
 - What about the other MVSSL methods?

Some contrastive MVSSL methods optimize InfoNCE, a lower bound

Image X





View V_2



View V_2

Projection Z_1 Projection Z_2



View V_2







Contrastive methods SimCLR, CMC, MoCo



Contrastive methods SimCLR, CMC, MoCo

- **Clustering-based methods**
 - SwAV, DeepCluster



Contrastive methods SimCLR, CMC, MoCo



- Clustering-based methods
 - SwAV, DeepCluster

BYOL, DINO

Prior work: What role does MI optimization play in MVSSL?



Contrastive methods SimCLR, CMC, MoCo



- Clustering-based methods
 - SwAV, DeepCluster

- **Distillation-based methods**
 - BYOL, DINO

Prior work: What role does MI optimization play in MVSSL?



Contrastive methods SimCLR, CMC, MoCo

Some optimize $I(Z_1; Z_2)$



- Clustering-based methods
 - SwAV, DeepCluster

- **Distillation-based methods**
 - BYOL, DINO

Prior work: What role does MI optimization play in MVSSL?



Contrastive methods SimCLR, CMC, MoCo

Some optimize $I(Z_1; Z_2)$



- Clustering-based methods
 - SwAV, DeepCluster

- **Distillation-based methods**
 - BYOL, DINO



Analysis: Using a different bound on MI w.r.t. prior works

$$\begin{split} I(Z_1; Z_2) &= H(Z_2) - H(Z_2 | Z_1) \\ &\geq H(Z_2) - \mathbb{E}_{Z_1, Z_2} \Big[-\log q_{Z_2 | Z_1}(Z_2) \Big] := I_{\mathsf{ER}}(Z_1; Z_2) \end{split}$$

Analysis: Using a different bound on MI w.r.t. prior works

$I(Z_1; Z_2) = H(Z_2) - H(Z_2 | Z_1)$ $\geq H(Z_2) - \mathbb{E}_{Z_1, Z_2} \left[-\log q_{Z_2|Z_1}(Z_2) \right] := I_{\mathsf{ER}}(Z_1; Z_2)$

Entropy: How much information can be learnt



Analysis: Using a different bound on MI w.r.t. prior works

$I(Z_1; Z_2) = H(Z_2) - H(Z_2 | Z_1)$ $\ge H(Z_2) - \mathbb{E}_{Z_1, Z_2} [-10]$

Entropy: How much information *can* be learnt
 Reconstruction: How much information *is* learnt

$$\log q_{Z_2|Z_1}(Z_2)] := I_{\mathsf{ER}}(Z_1; Z_2)$$





Contrastive methods SimCLR, CMC, MoCo



- Clustering-based methods
 - SwAV, DeepCluster

Distillation-based methods

BYOL, DINO



Contrastive methods SimCLR, CMC, MoCo

Some optimize $I(Z_1; Z_2)$ exactly, some not exactly



- Clustering-based methods
 - SwAV, DeepCluster

Distillation-based methods

BYOL, DINO



Contrastive methods SimCLR, CMC, MoCo

SwAV, DeepCluster

Some optimize $I(Z_1; Z_2)$ exactly, some not exactly



- Clustering-based methods
- BYOL, DINO

Optimize $I(Z_1; Z_2)$ exactly



Contrastive methods SimCLR, CMC, MoCo Clustering-based methods SwAV, DeepCluster

Some optimize $I(Z_1; Z_2)$ exactly, some not exactly

Optimize $I(Z_1; Z_2)$ exactly

 $f_{\theta} \circ \pi_{\theta}$ Z_1 predict X V_2 ι_2 $f_{\xi} \circ \pi_{\xi}$

Distillation-based methods BYOL, DINO

Maximize reconstruction, but the entropy is only maintained stable



















EMA 0.99 - EMA 0.8





TM and © 2023 Apple Inc. All rights reserved.