DeepMind

Representation Learning via Invariant Causal Mechanisms (ReLIC)

Authors: Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, Charles Blundell

Motivation

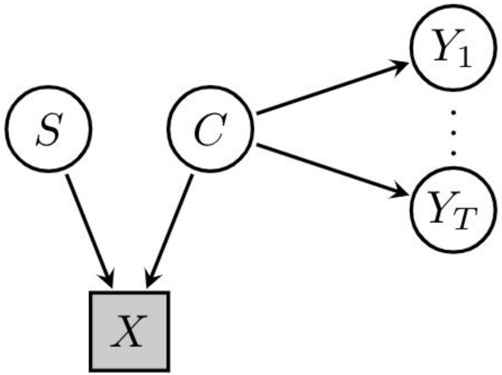
Problem: How to learn useful representations when we don't have access to labels?

Approach: Understand **what** should be learned and then derive **how** to learn it

This work: Use **causality** to formalize self-supervised learning

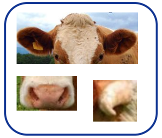
- Provide alternative explanation for contrastive learning (current SoTA)
- New objective based on **invariant prediction**
- Strong **theoretical and empirical** generalization results

Causal Formalization



A representation needs to

1. Capture directly relevant information: **content**



2. Discard spuriously correlated aspects: **style**

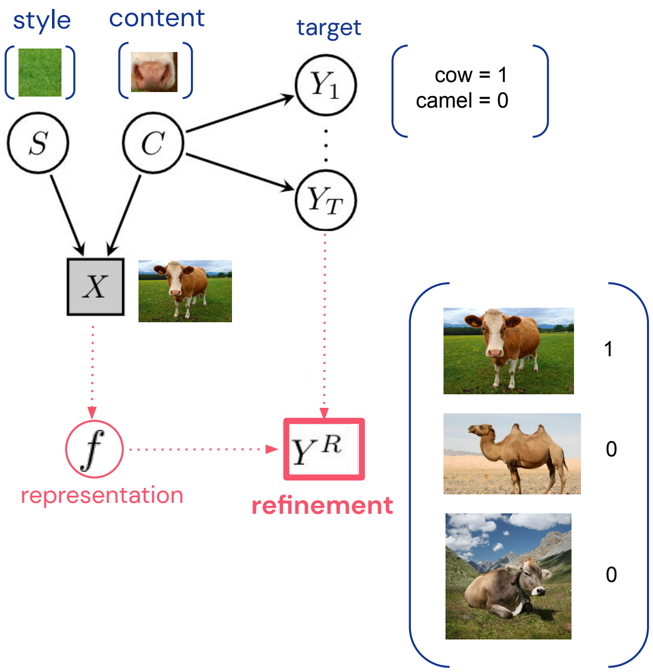
Content is an **invariant predictor** of target under style interventions:

$$p^{do(S=s_i)}(Y_t \mid C) = p^{do(S=s_j)}(Y_t \mid C) \quad \forall\, s_i, s_j \in \mathcal{S}$$

Use data augmentations as simulated interventions on the unobserved style

Defining proxy tasks

Targets are unobserved. How do we define sensible proxy tasks to solve?



Refinements are more fine-grained instances of the original problem.

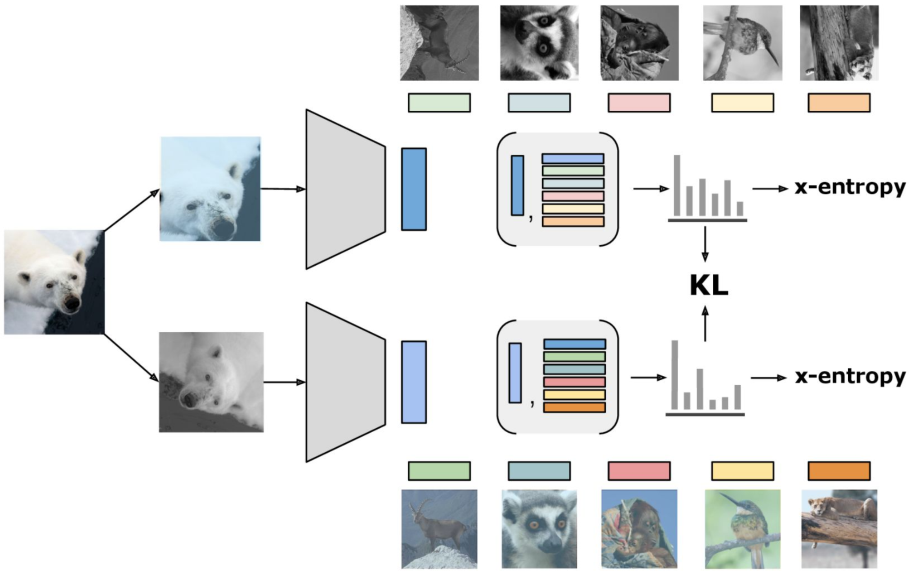
Theorem:

If the proxy task used for learning is a **refinement**, the resulting representation will be useful for downstream tasks

Contrastive learning does **instance discrimination**: the most fine-grained refinement!

Causality explains why contrastive learning produces useful representations!

ReLIC objective



Learning principle:

$$\textit{(Invariant prediction)} \qquad p^{\text{do}(a_i)}(Y^R | f(X)) = p^{\text{do}(a_j)}(Y^R | f(X)) \quad \forall a_i, a_j \in \mathcal{A}.$$

Linear Evaluation on ImageNet

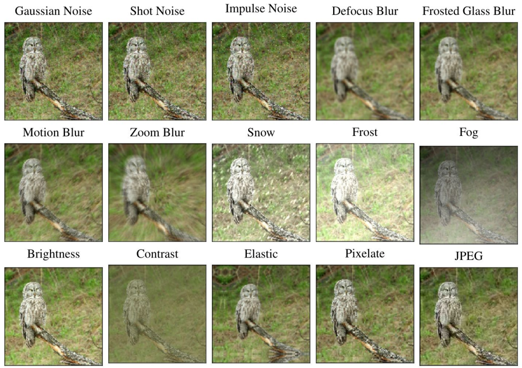
Table 1: Accuracy (in %) under linear evaluation on ImageNet for different self-supervised representation learning methods. Methods with * use SimCLR augmentations. Methods with † use custom, stronger augmentations.

Method		Top-1	Top-5
ResNet-50 architecture			
PIRL		63.6	-
CPC v2		63.8	85.3
CMC		66.2	87.0
SimCLR [4]	*	69.3	89.0
SwAV [2]	*	70.1	-
RELIC (ours)	*	70.3	89.5
InfoMin Aug. [22]	†	73.0	91.1
SwAV [2]	†	75.3	-
ResNet-50 with target network			
MoCo v2 [5]		71.1	-
BYOL [7]	*	74.3	91.6
RELIC (ours)	*	74.8	92.2

* uses standard augmentations

† uses stronger augmentations

Robustness and Out-of-Distribution Generalization

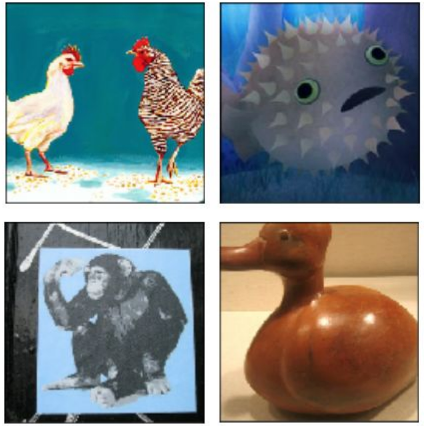


ImageNet-C (corrupted)

Images with diverse corruptions of varying strengths

Tests: robustness of representation

Method	Supervised	SimCLR	RELIC	BYOL	RELIC$_T$
mCE (%)	76.7	87.5	76.4	72.3	70.8



ImageNet-R (rendered)

New renditions of 200 ImageNet classes

Tests: out-of-distribution generalization

Method	Supervised	SimCLR	RELIC	BYOL	RELIC$_T$
Top-1 Error (%)	63.9	81.7	77.4	77.0	76.2

Performance on RL benchmark - Atari

Table 4: Human Normalized Scores of Auxiliary Methods over 57 Atari Games.

Atari Performance	RELIC	SimCLR	CURL	BYOL	Augmentation
Capped mean	**91.46**	88.76	90.72	89.43	80.60
Number of superhuman games	**51**	49	49	49	34
Mean	**3003.73**	2086.16	2413.12	1769.43	503.15
Median	**832.50**	592.83	819.56	483.39	132.17
40% Percentile	356.27	266.07	**409.46**	224.80	94.35
30% Percentile	**202.49**	174.19	190.96	150.21	80.04
20% Percentile	**133.93**	120.84	126.10	118.36	57.95
10% Percentile	**83.79**	37.19	59.09	44.14	32.74
5% Percentile	**20.87**	12.74	20.56	7.75	2.85

DeepMind

Thank you!

Contact: mitrovic@google.com