REpresentation **L**earning via **I**nvariant **C**ausal mechanisms (**RELIC**)

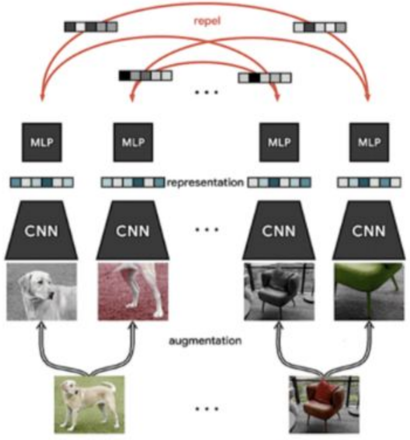
DeepMind, London, UK
Jovana Mitrovic et al.,

ICLR 2021 (under review)

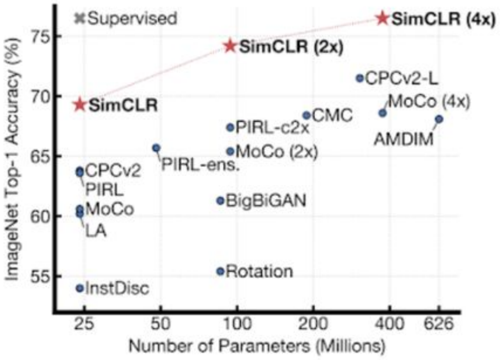
Sungman Cho.

Introduction

- Recently, **self-supervised methods using contrastive objectives** have emerged as one of the **most successful strategies** for unsupervised representation learning.



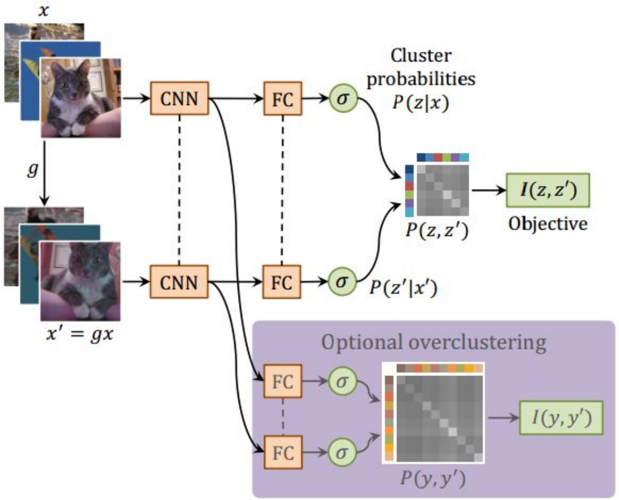
An illustration of SimCLR (from our blog here).



SimCLR, Google research, arxiv:2002.05709

Introduction

- Under assumptions on how the negative examples are sampled, **minimizing the resulting contrastive loss** has been justified as **maximizing a lower bound on the mutual information(MI)** between representations.



W.Ji, Invariant Information Clustering for Unsupervised Image Classification and Segmentation, ICCV 2019

Introduction

- However, (Tschannen et al., 2019, ICLR 2020) has shown that performance on **downstream tasks may be more tightly correlated with the encoder architecture than the achieved MI bound**, highlighting issues with the MI theory of contrastive learning

Introduction

- Contrastive approach has yet to be **theoretically justified.**

- To remedy the theoretical shortcomings, we **analyze** the problem of self-supervised representation learning **through a causal lens.**

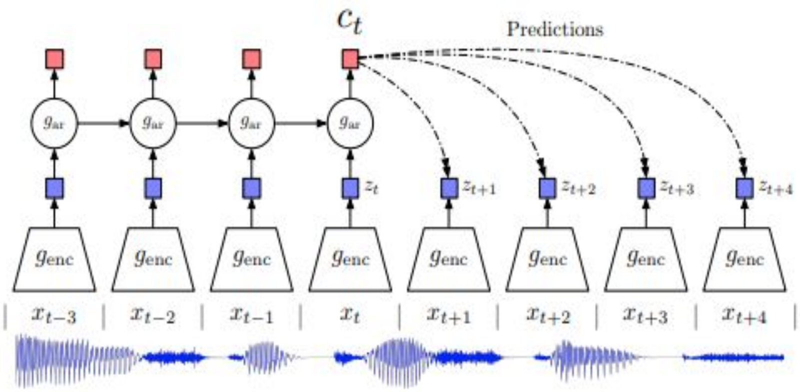
Contributions

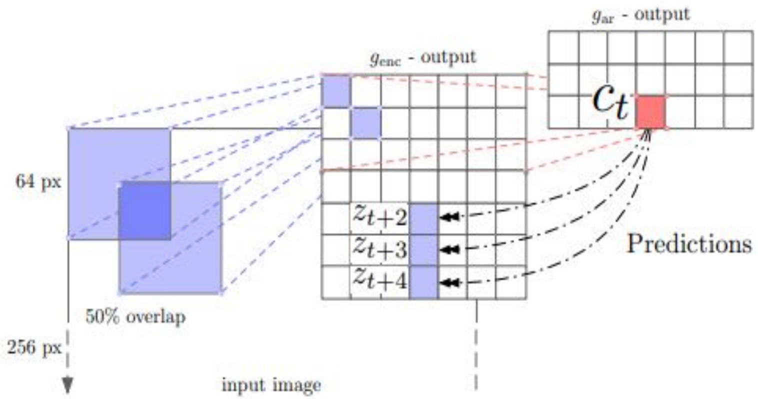
- **Formalize problem of self-supervised representation** learning **using causality** and propose to more effectively leverage data augmentations **through invariant prediction.**

- Propose a new self-supervised objective, RELIC, that **enforces invariant prediction** through an **explicit regularizer** and show **improved generalization guarantees.**

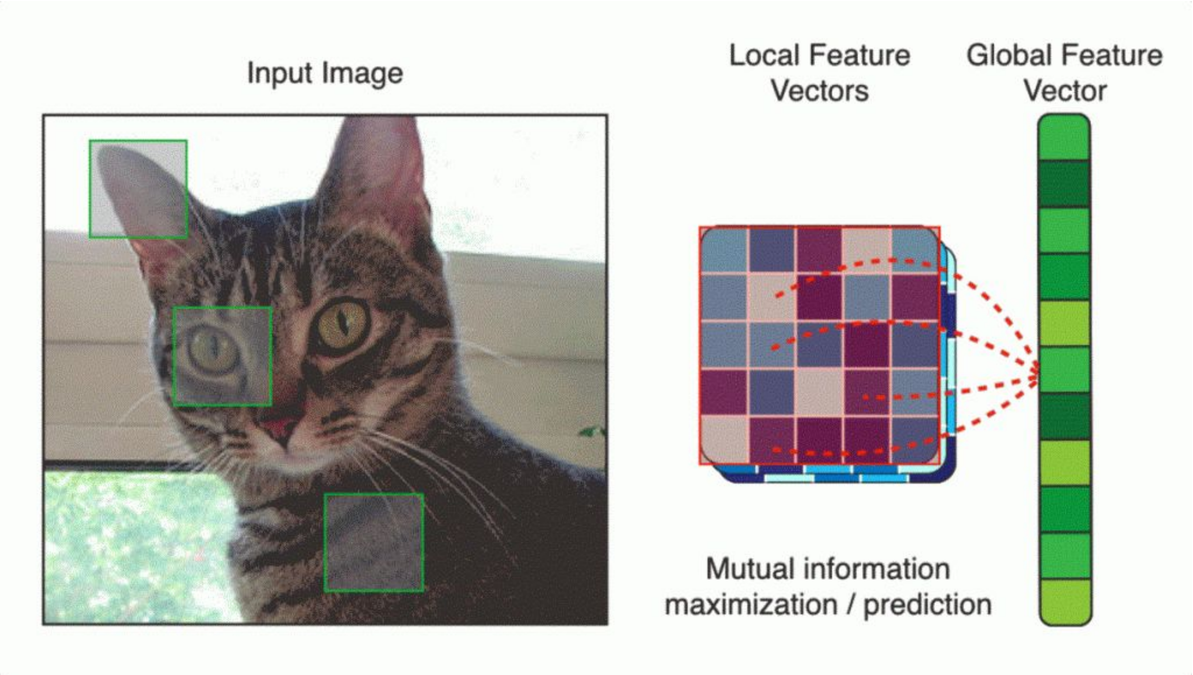
Related Work

- **Contrastive objectives and mutual information maximization**
  - Contrastive Predictive Coding (CPC), arXiv:1807.03748

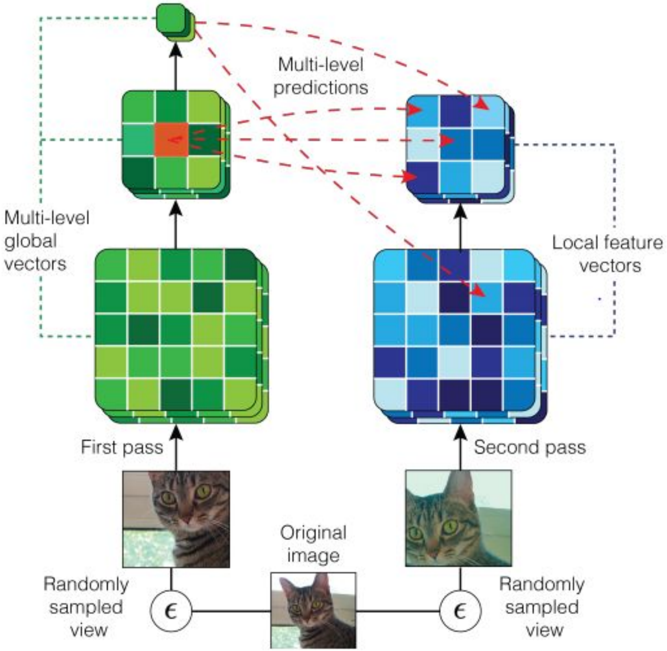




- **Contrastive objectives and mutual information maximization**
    - Deep InfoMax, arXiv:1808.06670



- **Contrastive objectives and mutual information maximization**
  - Learning Representations by Maximizing Mutual Information Across Views (AMDIM), NeurIPS 2019.

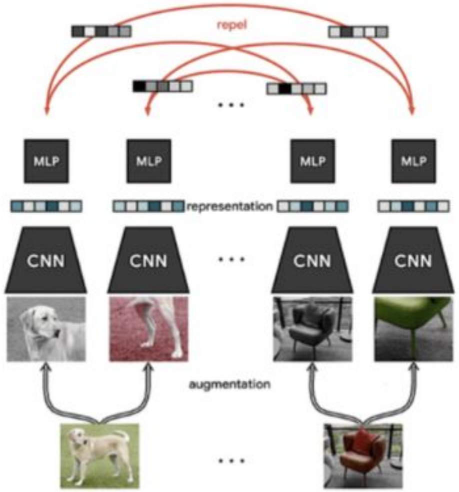


- **Contrastive objectives and mutual information maximization**
  - Contrastive Predictive Coding (CPC)
  - DeepInfomax
  - Augmented Multiscale DIM (AMDIM)

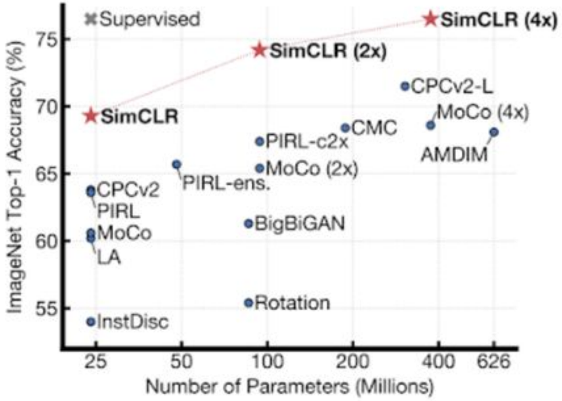
“Performance on downstream tasks is not correlated with the achieved bound on MI, but may be more tightly correlated with encoder architecture and capacity.”

Tschannen et al., 2019. (ICLR 2020)

- **Contrastive estimation with a standard ResNet-50 architecture.**
  - SimCLR
  - MoCo
  - BYOL



An illustration of SimCLR (from our blog here).



- Recently, (Saunshi et al., 2019, ICML 2020) proposed learning theoretic framework to **analyze** the **performance of contrastive objectives.**

- However, without **strong assumptions on intra-class concentration** they note that **contrastive objectives are fundamentally limited in the representations they are able to learn.**

- **RELIC explicitly enforces intra-class concentration via the invariance regularizer**, ensuring that it generalizes under weaker assumptions.

- The reasons for the **improvement in performance** from AMDIM through to SimCLR and BYOL **are not easily explained by either the MI maximization or the learning theoretic viewpoint.**

- In contrast to prior art, the performance of **RELIC is explained by connections to causal theory.**

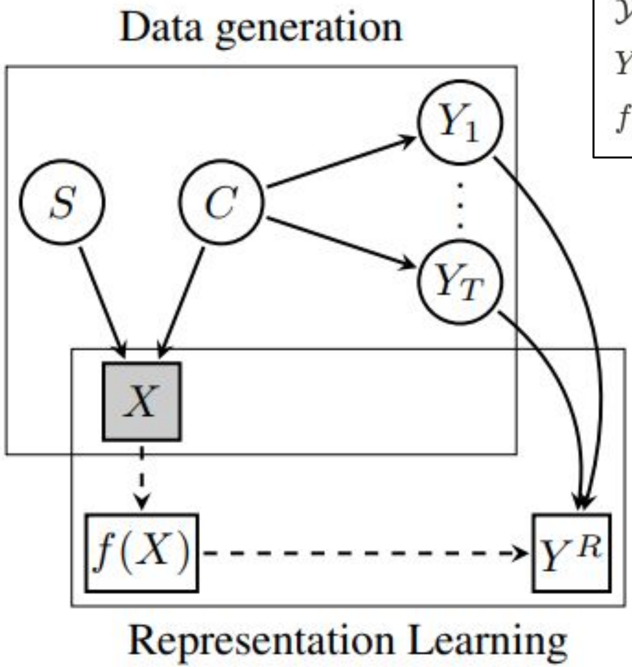
- Furthermore, the use of **invariance penalties** in RELIC as dictated by **causal theory** yields **significantly more robust representations** that generalize better than those learned with SimCLR or BYOL.

- **Causality and invariance**
    - The notion of **invariant prediction** has emerged as an important operational concept in causal inference (Peters et al., 2016)
    - Learn **classifiers which are robust against domain shifts** (Gong et al., 2016)
    - Use **group structure to delineate between different environments by supervised manner**. (Heinze-Deml & Meinshausen, 2017)

- **Causality and invariance (RELIC)**
  - **Enforce invariant prediction within the group by constraining the distance** between distributions resulting from contrasting data across groups.
  - **Does not rely on ground-truth.**

RELIC

RELIC

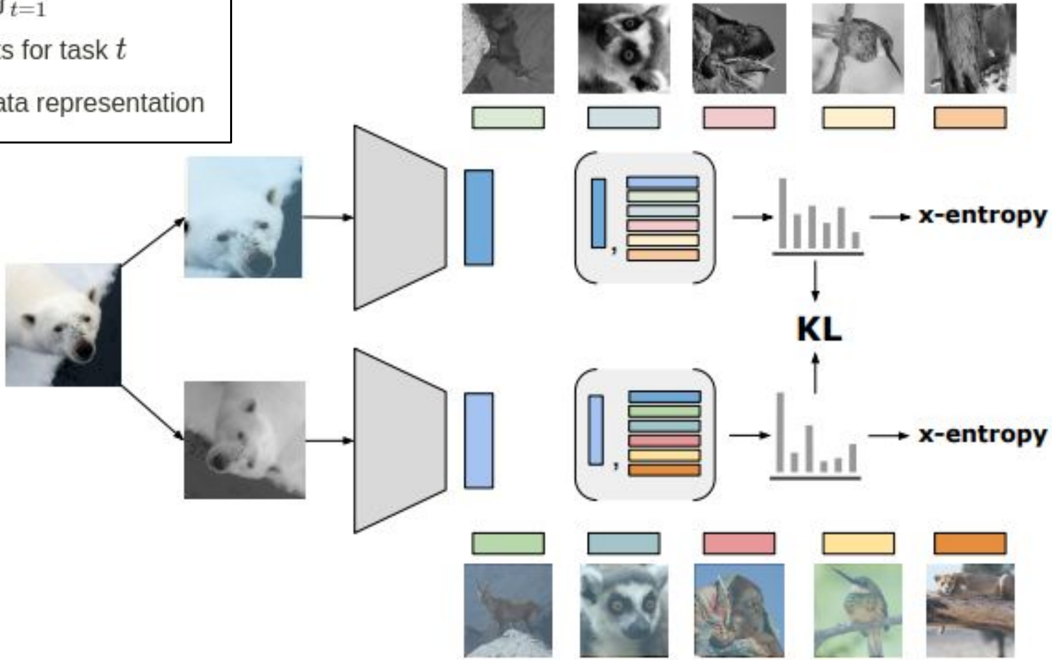


X: unlabeled data

$\mathcal{Y} = \{Y_t\}_{t=1}^{T}$

Y_t : targets for task t

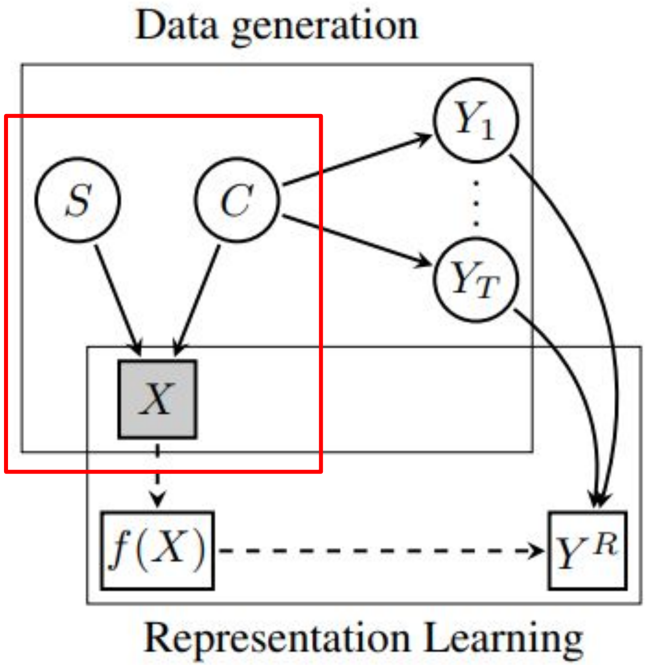
$f(X)$: data representation



Causal graph

RELIC objective

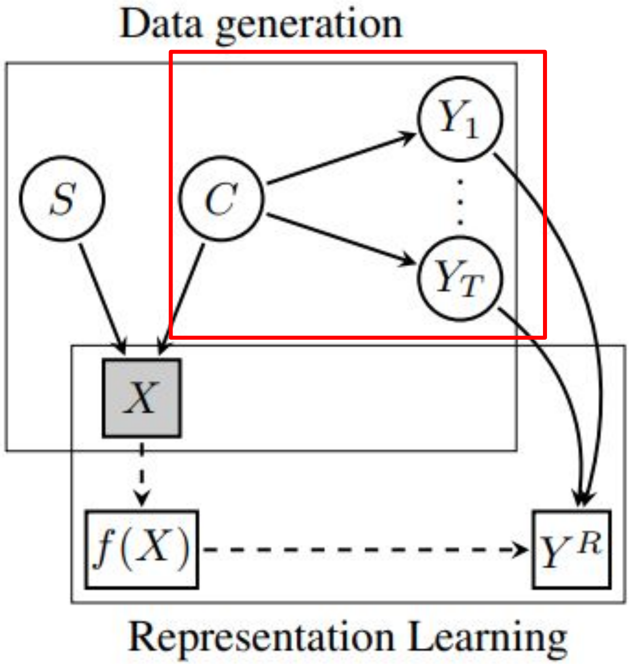
RELIC : Causal Interpretation



Causal graph

- The data is generated from content and style variables.

RELIC : Causal Interpretation

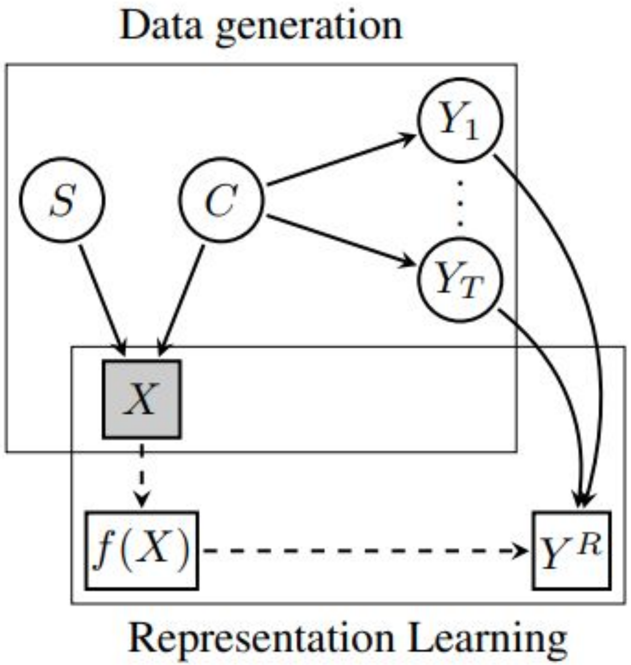


Causal graph

- The data is generated from content and style variables.
- Only content being relevant for the unknown downstream tasks.
- Content and style are independent, i.e. style changes are content-preserving.

Style : background, lighting conditions and camera lens characteristics.
Content : different parts of the animals

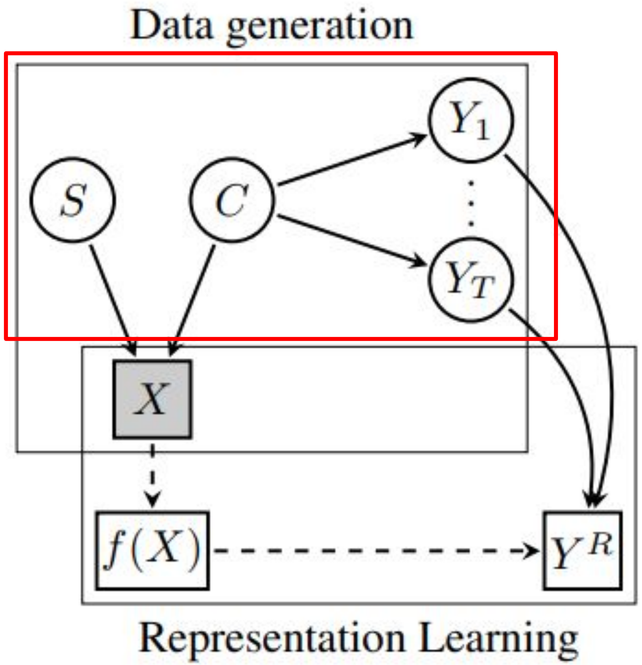
RELIC : Causal Interpretation



Causal graph

- The data is generated from content and style variables.
- Only content being relevant for the unknown downstream tasks.
- Content and style are independent, i.e. style changes are content-preserving.
- Content is a good representation of the data for downstream tasks and we therefore cast the goal of representation learning as estimating content.

RELIC : Causal Interpretation



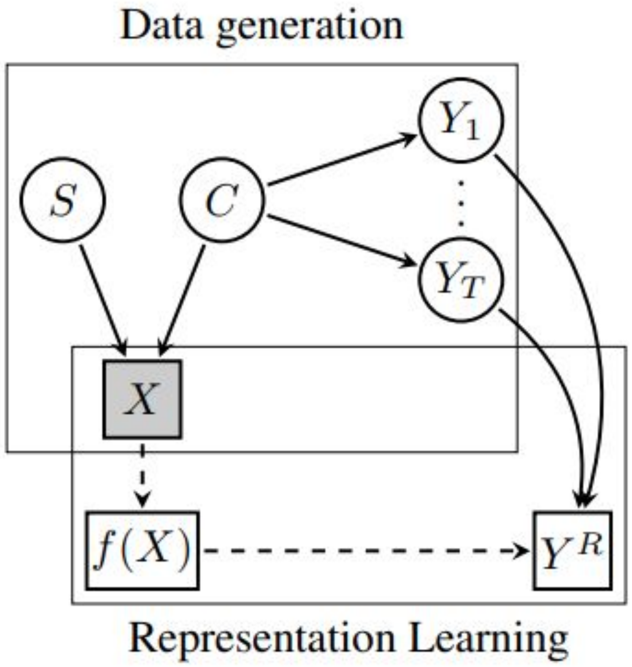
Causal graph

- Content directly influences the target tasks, while style does not.
- **We call content an invariant representation.**

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

$p^{do(S=s)}$: distribution arising from assigning S the value s with $\mathcal{S}$ the domain of S

RELIC : Causal Interpretation



Causal graph

- To learn invariant representations, such as Content, we enforce Equation which requires us to observe data under different style interventions, i.e. we need data that describes the same content under varying style.

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

- We use content-preserving data augmentations (rotation, grayscaling, translation, cropping, …)

RELIC : Objective

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

$\mathcal{A} = \{a_1, \ldots, a_m\}$ is the set of data augmentations

To achieve invariant prediction, we propose to **explicitly enforce invariance** under augmentations through a regularizer.

$$\mathop{\mathbb{E}}_{X} \mathop{\mathbb{E}}_{\substack{a_{lk}, a_{qt} \\ \sim \mathcal{A} \times \mathcal{A}}} \sum_{b \in \{a_{lk}, a_{qt}\}} \mathcal{L}_b(Y^R, f(X)) \quad s.t. \quad KL\left(p^{do(a_{lk})}(Y^R|\, f(X)), p^{do(a_{qt})}(Y^R|\, f(X))\right) \leq \rho$$

Any distance measure on distributions can be used in place of the KL divergence.

RELIC : Objective

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

$\mathcal{A} = \{a_1, \ldots, a_m\}$ is the set of data augmentations

To achieve invariant prediction, we propose to **explicitly enforce invariance** under augmentations through a regularizer.

$$\underset{X}{\mathbb{E}} \; \underset{\substack{a_{lk}, a_{qt} \\ \sim \mathcal{A}\times\mathcal{A}}}{\mathbb{E}} \; \sum_{b \in \{a_{lk}, a_{qt}\}} \underset{\text{original proxy task}}{\boxed{\mathcal{L}_b(Y^R, f(X))}} \; s.t. \;\; KL\left(p^{do(a_{lk})}(Y^R|\, f(X)), p^{do(a_{qt})}(Y^R|\, f(X))\right) \leq \rho$$

Any distance measure on distributions can be used in place of the KL divergence.

RELIC : Objective

Contastive objective (commonly used)

$$p^{do(a_{lk})}(Y^R = j \mid f(x_i)) \propto \exp\left(\phi(f(x_i^{a_l}), h(x_j^{a_k}))/\tau\right) .$$

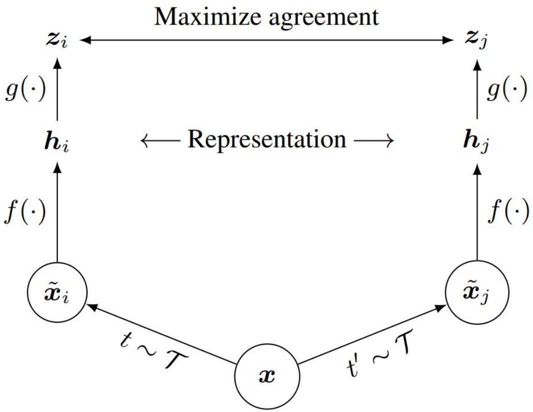
$$\phi(f(x_i), h(x_j)) \;=\; \langle g(f(x_i)), g(h(x_j))\rangle$$

RELIC : Objective

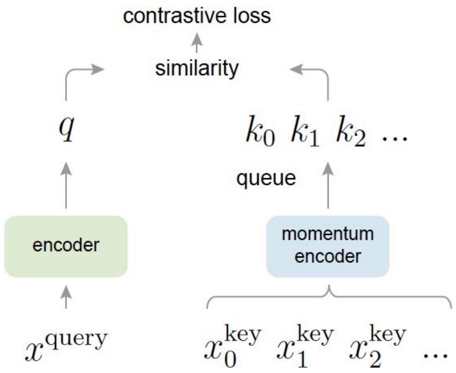
Contastive objective (commonly used)

$$p^{do(a_{lk})}(Y^R = j \mid f(x_i)) \propto \exp\left(\phi(f(x_i^{a_l}), h(x_j^{a_k}))/\tau\right).$$

$$\phi(f(x_i), h(x_j)) = \langle g(f(x_i)), g(h(x_j)) \rangle$$



SimCLR, Google research, arxiv:2002.05709



FAIR, arxiv:1911.05722

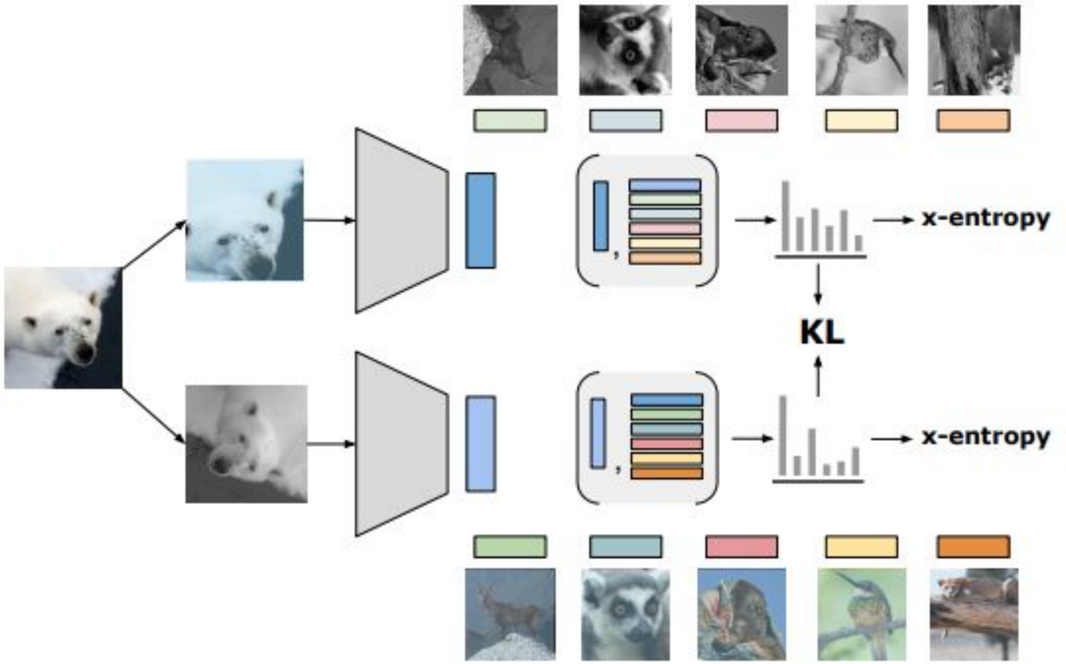
RELIC : Objective

Contrastive objective + RELIC

$$-\sum_{i=1}^{N}\sum_{a_{lk}} \log \frac{\exp\left(\phi(f(x_i^{a_l}), h(x_i^{a_k}))/\tau\right)}{\sum_{m=1}^{M}\exp\left(\phi(f(x_i^{a_l}), h(x_m^{a_k}))/\tau\right)} + \alpha \sum_{a_{lk}, a_{qt}} KL(p^{do(a_{lk})}, p^{do(a_{qt})})$$

$$a_{lk} = (a_l, a_k) \in \mathcal{A} \times \mathcal{A}$$

RELIC : Objective



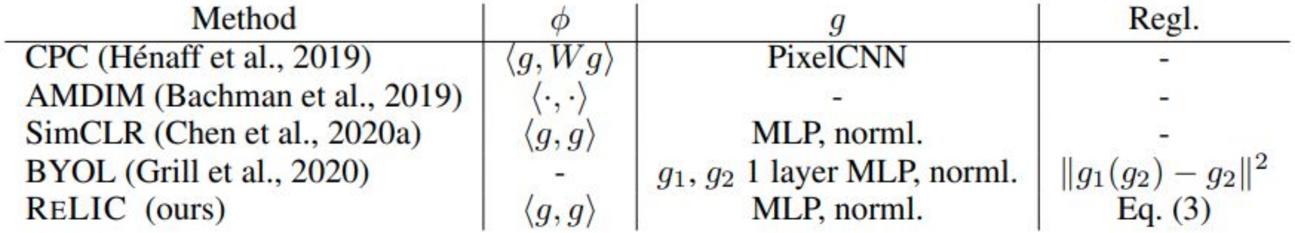
$$-\sum_{i=1}^{N}\sum_{a_{lk}}\log\frac{\exp\left(\phi(f(x_i^{a_l}),h(x_i^{a_k}))/\tau\right)}{\sum_{m=1}^{M}\exp\left(\phi(f(x_i^{a_l}),h(x_m^{a_k}))/\tau\right)}+\alpha\sum_{a_{lk},a_{qt}}KL(p^{do(a_{lk})},p^{do(a_{qt})})$$

RELIC : Objective

The **explicit invariance penalty encourages the within-class distances** of the representations learned by RELIC to be tightly concentrated.

$$-\sum_{i=1}^{N}\sum_{a_{lk}} \log \frac{\exp\left(\phi(f(x_i^{a_l}), h(x_i^{a_k}))/\tau\right)}{\sum_{m=1}^{M}\exp\left(\phi(f(x_i^{a_l}), h(x_m^{a_k}))/\tau\right)} + \alpha \sum_{a_{lk}, a_{qt}} KL(p^{do(a_{lk})}, p^{do(a_{qt})})$$

The above equation recovers SOTA methods depending on design choices.



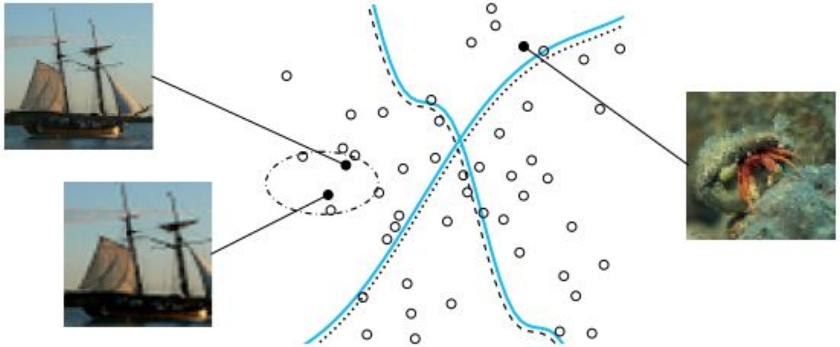
Method	ϕ	g	Regl.
CPC (Hénaff et al., 2019)	$\langle g, Wg \rangle$	PixelCNN	-
AMDIM (Bachman et al., 2019)	$\langle \cdot, \cdot \rangle$	-	-
SimCLR (Chen et al., 2020a)	$\langle g, g \rangle$	MLP, norml.	-
BYOL (Grill et al., 2020)	-	g_1, g_2 1 layer MLP, norml.	$\lVert g_1(g_2) - g_2 \rVert^2$
RELIC (ours)	$\langle g, g \rangle$	MLP, norml.	Eq. (3)

Theoretical Justification

Learning with refinements.

- In contrastive learning, the task of instance discrimination, i.e. classifying the dataset, is used as the proxy task.

- To better understand contrastive learning and motivate this proxy task, we **generalize Instance discrimination using the causal concepts of refinements** (Chalupka et al., 2014).



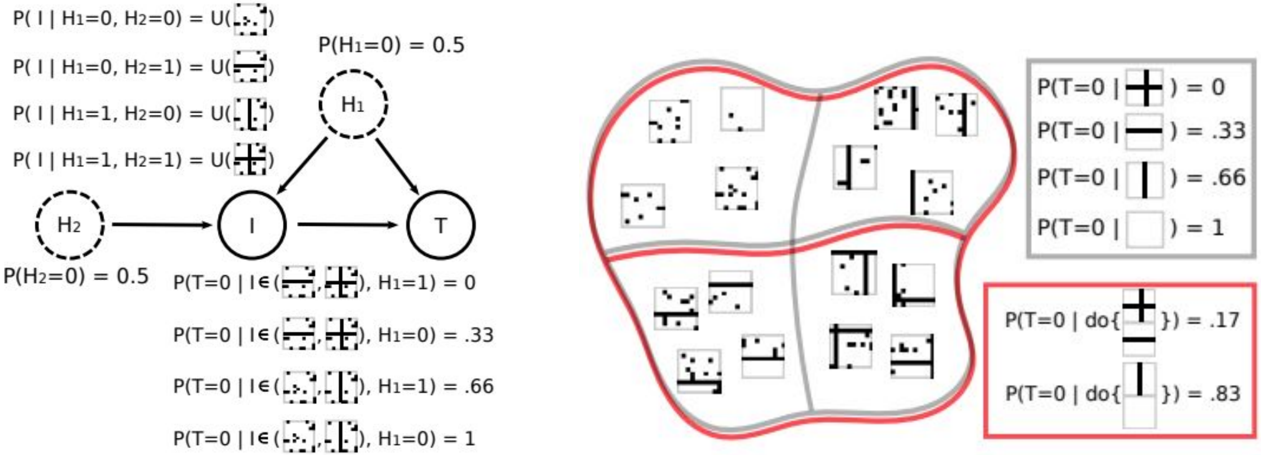
Instance discrimination
- Aquatic vs non-aquatic life
- Animal vs non-animal

Refinements
- Aquatic animal vs aquatic non-animal vs non-aquatic animal vs non-aquatic non-animal

Appendix. Visual Cause

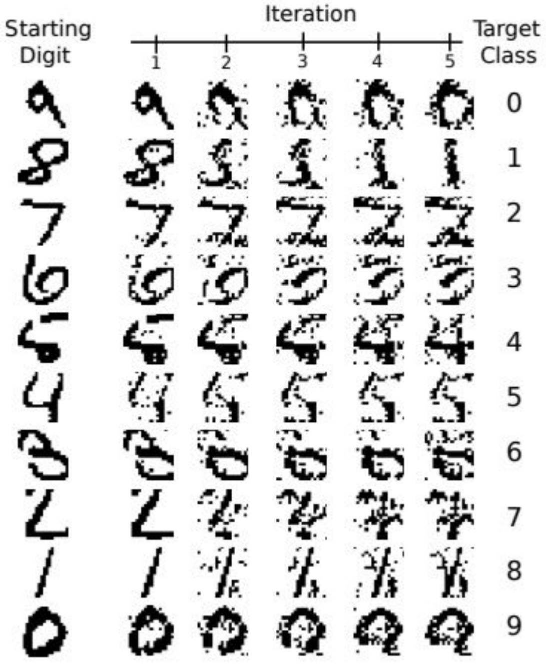
- K.Chalupka, Visual Causal Feature Learning, arXiv:1412.2309
  - “Traffic light turns green”
  - “ It is the increased symmetry of features that leads people to judge faces more attractive than others”

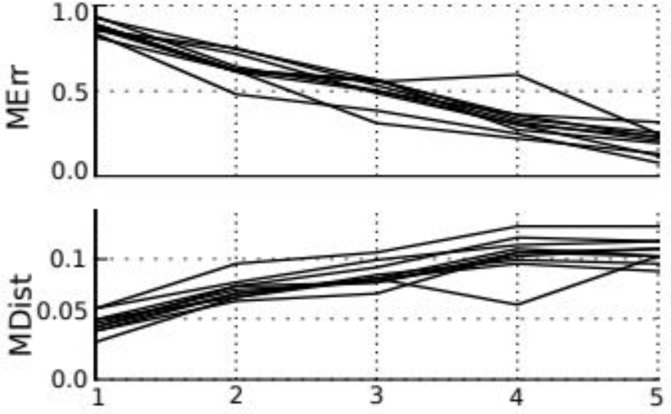
Causal Coarsening Theorem.


P(I | H1=0, H2=0) = U( )
P(I | H1=0, H2=1) = U( )
P(I | H1=1, H2=0) = U( )
P(I | H1=1, H2=1) = U( )
P(H1=0) = 0.5
H1
H2
I
T
P(H2=0) = 0.5
P(T=0 | I∈(,), H1=1) = 0
P(T=0 | I∈(,), H1=0) = .33
P(T=0 | I∈(,), H1=1) = .66
P(T=0 | I∈(,), H1=0) = 1
P(T=0 |) = 0
P(T=0 |) = .33
P(T=0 |) = .66
P(T=0 |) = 1
P(T=0 | do{ }) = .17
P(T=0 | do{ }) = .83

Appendix. Visual Cause

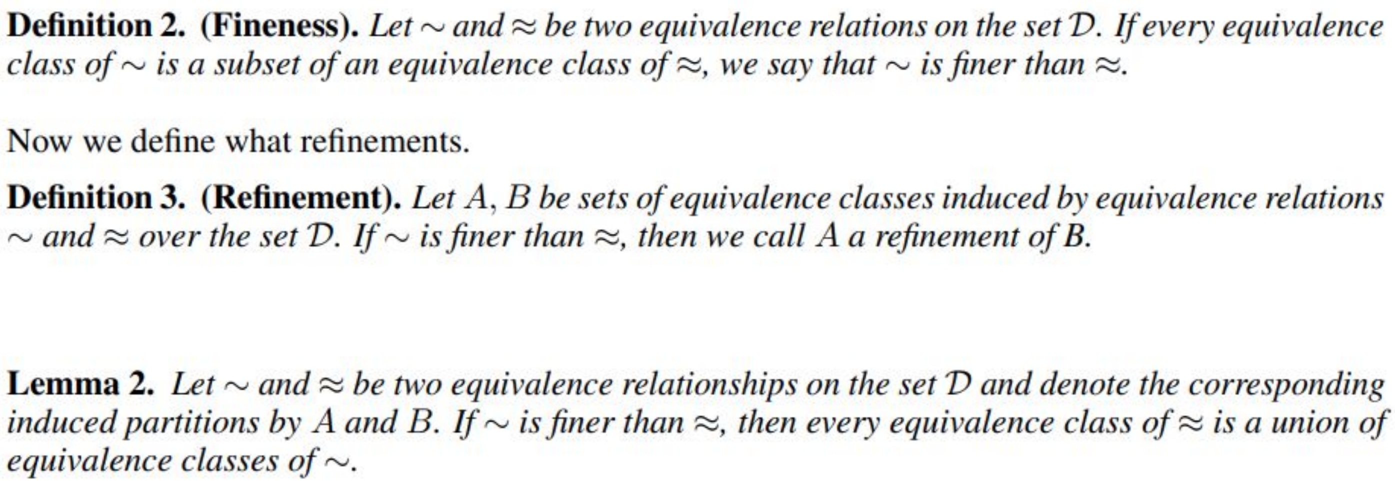
- K.Chalupka, Visual Causal Feature Learning, arXiv:1412.2309





Learning with refinements.

- To mathematically define refinements,



Definition 2. (Fineness). *Let $\sim$ and $\approx$ be two equivalence relations on the set $\mathcal{D}$. If every equivalence class of $\sim$ is a subset of an equivalence class of $\approx$, we say that $\sim$ is finer than $\approx$.*

Now we define what refinements.

Definition 3. (Refinement). *Let A, B be sets of equivalence classes induced by equivalence relations $\sim$ and $\approx$ over the set $\mathcal{D}$. If $\sim$ is finer than $\approx$, then we call A a refinement of B.*

Lemma 2. *Let $\sim$ and $\approx$ be two equivalence relationships on the set $\mathcal{D}$ and denote the corresponding induced partitions by A and B. If $\sim$ is finer than $\approx$, then every equivalence class of $\approx$ is a union of equivalence classes of $\sim$.*

Learning with refinements.

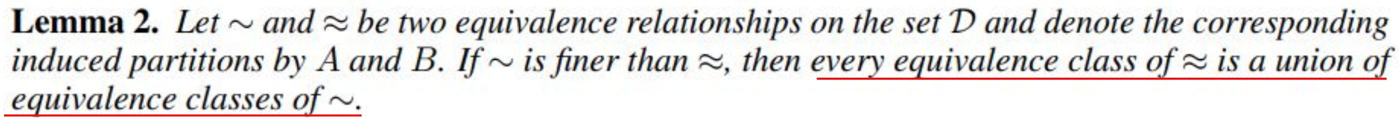
- **Dogs ≈ Dogs**

- **Poodles ~ Poodles**

Learning with refinements.

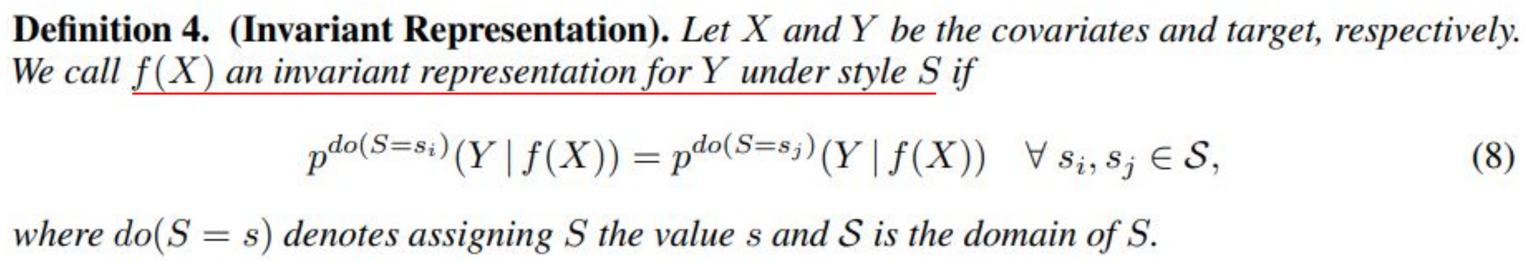
- **Dogs ≈ Dogs**

- **Poodles ~ Poodles**



Lemma 2. *Let $\sim$ and $\approx$ be two equivalence relationships on the set $\mathcal{D}$ and denote the corresponding induced partitions by A and B. If $\sim$ is finer than $\approx$, then every equivalence class of $\approx$ is a union of equivalence classes of $\sim$.*

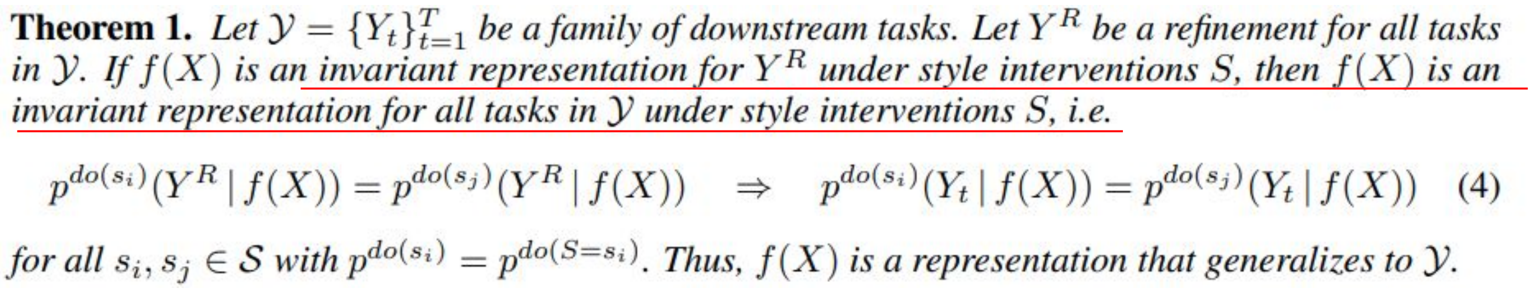
Learning with refinements.



Definition 4. (Invariant Representation). *Let X and Y be the covariates and target, respectively. We call $f(X)$ an invariant representation for Y under style S if*

$$p^{do(S=s_i)}(Y \mid f(X)) = p^{do(S=s_j)}(Y \mid f(X)) \quad \forall\, s_i, s_j \in \mathcal{S}, \tag{8}$$

where $do(S = s)$ denotes assigning S the value s and $\mathcal{S}$ is the domain of S.

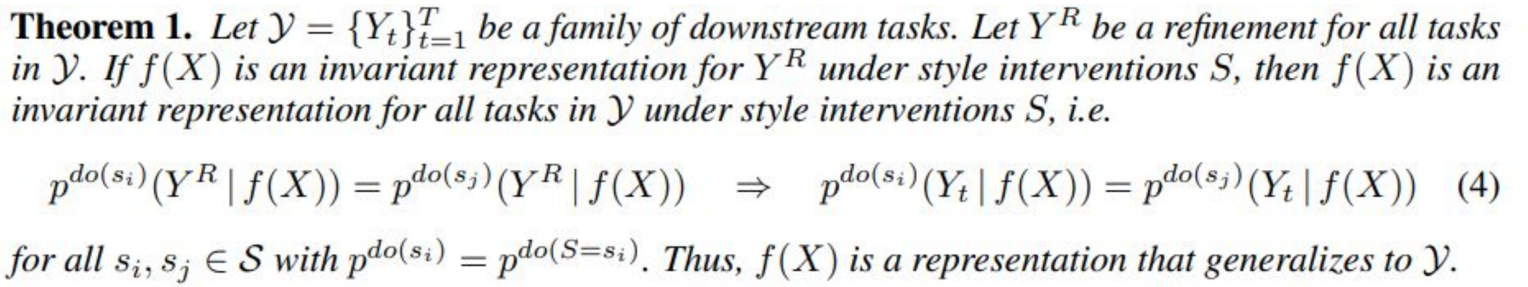


Theorem 1. *Let $\mathcal{Y} = \{Y_t\}_{t=1}^{T}$ be a family of downstream tasks. Let Y^R be a refinement for all tasks in $\mathcal{Y}$. If $f(X)$ is an invariant representation for Y^R under style interventions S, then $f(X)$ is an invariant representation for all tasks in $\mathcal{Y}$ under style interventions S, i.e.*

$$p^{do(s_i)}(Y^R \mid f(X)) = p^{do(s_j)}(Y^R \mid f(X)) \quad \Rightarrow \quad p^{do(s_i)}(Y_t \mid f(X)) = p^{do(s_j)}(Y_t \mid f(X)) \tag{4}$$

for all $s_i, s_j \in \mathcal{S}$ with $p^{do(s_i)} = p^{do(S=s_i)}$. Thus, $f(X)$ is a representation that generalizes to $\mathcal{Y}$.

Learning with refinements.



Theorem 1. *Let $\mathcal{Y} = \{Y_t\}_{t=1}^{T}$ be a family of downstream tasks. Let Y^R be a refinement for all tasks in $\mathcal{Y}$. If $f(X)$ is an invariant representation for Y^R under style interventions S, then $f(X)$ is an invariant representation for all tasks in $\mathcal{Y}$ under style interventions S, i.e.*

$$p^{do(s_i)}(Y^R \mid f(X)) = p^{do(s_j)}(Y^R \mid f(X)) \quad \Rightarrow \quad p^{do(s_i)}(Y_t \mid f(X)) = p^{do(s_j)}(Y_t \mid f(X)) \quad (4)$$

for all $s_i, s_j \in \mathcal{S}$ with $p^{do(s_i)} = p^{do(S=s_i)}$. Thus, $f(X)$ is a representation that generalizes to $\mathcal{Y}$.

Y^R is a refinement of $\mathcal{Y}$ then learning a representation on Y^R is a *sufficient condition* for this representation to be useful on $\mathcal{Y}$

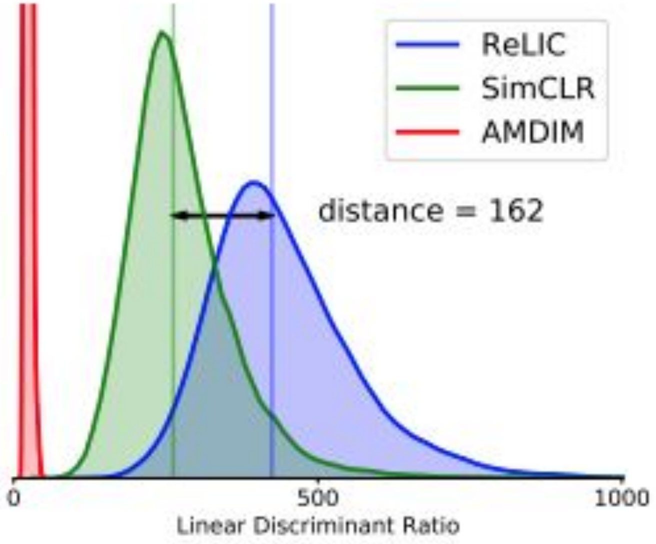
Leveraging causal tools,

We **connect** learning on **refinements** to learning on **downstream tasks.**

Experiments

Linear evaluation.

- Fischer’s linear discriminant ratio

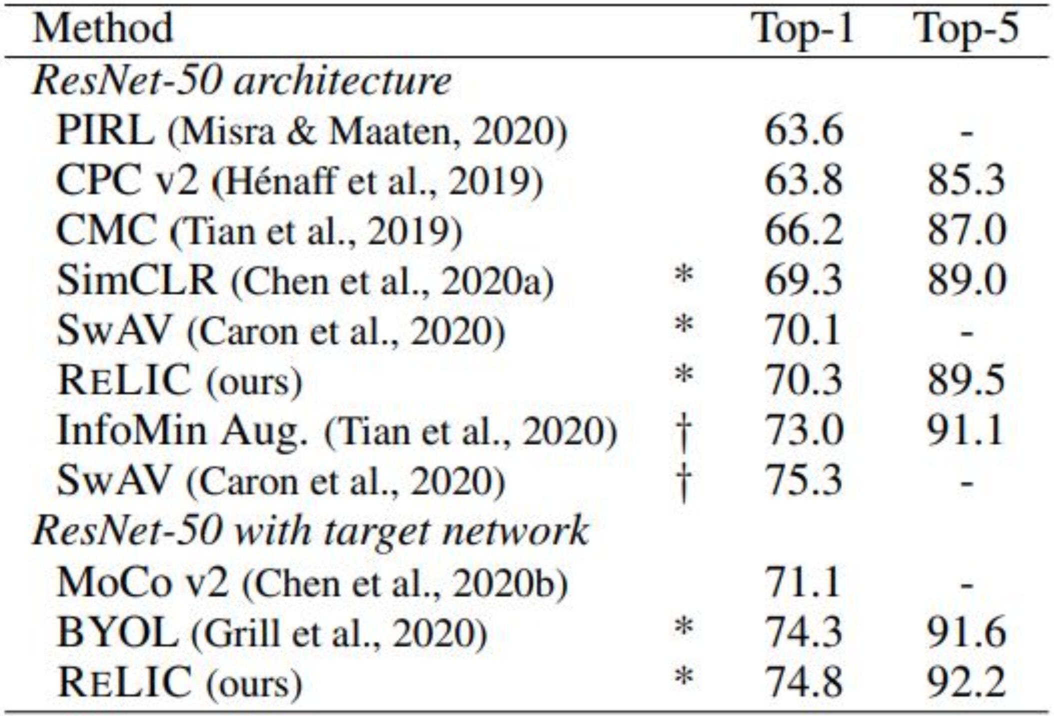


$$F_{\text{LDA}} = \|\mu_k - \mu_{k'}\|^2 / \sum_{i,j \in \mathcal{C}_k} \|f(x_i) - f(x_j)\|^2$$

$$\mu_k = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} f(x_i)$$

"A larger F_{LDA} implies that classes are more easily separated with a linear classifier."

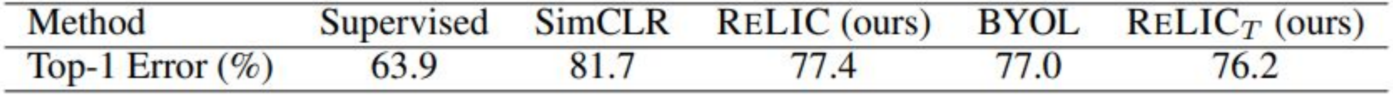
Linear evaluation.



Method		Top-1	Top-5
ResNet-50 architecture			
PIRL (Misra & Maaten, 2020)		63.6	-
CPC v2 (Hénaff et al., 2019)		63.8	85.3
CMC (Tian et al., 2019)		66.2	87.0
SimCLR (Chen et al., 2020a)	*	69.3	89.0
SwAV (Caron et al., 2020)	*	70.1	-
RELIC (ours)	*	70.3	89.5
InfoMin Aug. (Tian et al., 2020)	†	73.0	91.1
SwAV (Caron et al., 2020)	†	75.3	-
ResNet-50 with target network			
MoCo v2 (Chen et al., 2020b)		71.1	-
BYOL (Grill et al., 2020)	*	74.3	91.6
RELIC (ours)	*	74.8	92.2

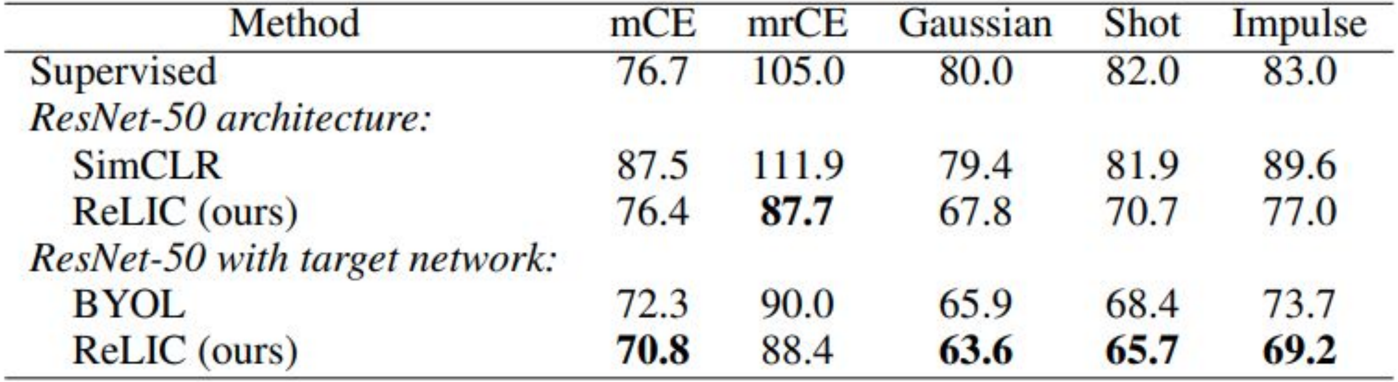
Robustness and generalization.

Trained : clean ImageNet **/ Tested :** ImageNet-R



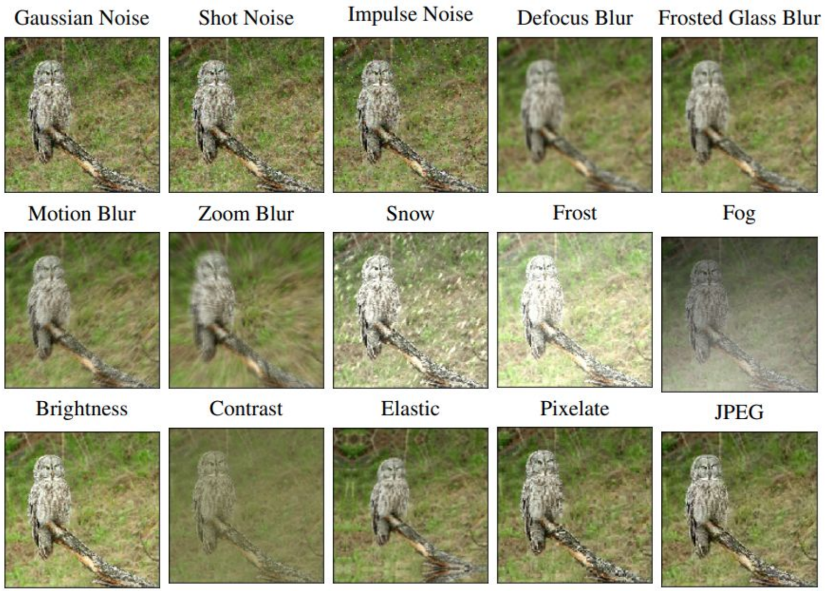
Method	Supervised	SimCLR	ReLIC (ours)	BYOL	ReLIC$_T$ (ours)
Top-1 Error (%)	63.9	81.7	77.4	77.0	76.2

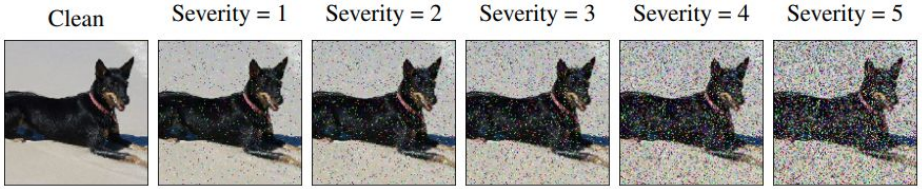
Trained : clean ImageNet **/ Tested :** ImageNet-C



Method	mCE	mrCE	Gaussian	Shot	Impulse
Supervised	76.7	105.0	80.0	82.0	83.0
ResNet-50 architecture:					
SimCLR	87.5	111.9	79.4	81.9	89.6
ReLIC (ours)	76.4	**87.7**	67.8	70.7	77.0
ResNet-50 with target network:					
BYOL	72.3	90.0	65.9	68.4	73.7
ReLIC (ours)	**70.8**	88.4	**63.6**	**65.7**	**69.2**

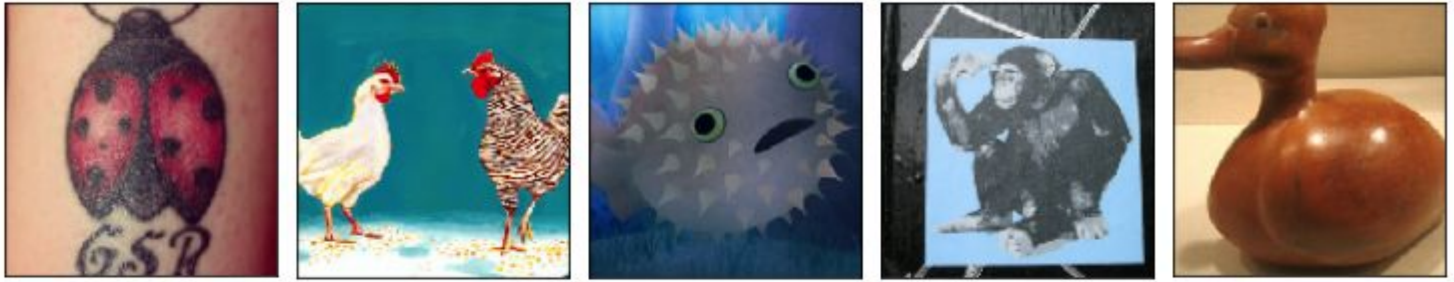
Appendix. ImageNet-C (for semantic robustness)





ImageNet-C dataset consists of **15 types of corruptions** from noise, blur, weather, and digital categories. Each type of corruption has **five levels** of severity, resulting in **75 distinct corruption.**

Appendix. ImageNet-R (for O.O.D generalization)



ImageNet-R contains **30,000 images of 200 ImageNet classes**. This dataset **emphasizes shape over texture** and has different textures and local image statistics to those of ImageNet.

Conclusion

- Have analyzed **self-supervised** learning using a **causal framework.**
  - Using a causal graph, we have formalized the problem of self-supervised representation learning and derived properties of the optimal representation.
  - Representations need to be invariant predictors of proxy targets under interventions on features that are only correlated, but not causally related to the downstream tasks.
- Proposed a new self-supervised objective, RELIC, that **enforces invariant prediction** of proxy targets across augmentations **using an invariance regularizer.**

Thank You.