Representation Learning via Invariant Causal Mechanisms

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

Presented by: Christoph Frey
Advisor: Osama Makansi
Examiner: Prof. Dr. Thomas Brox
Contents

• Motivation
• Contrastive Learning
• Content:
  – Causal interpretation
  – RELIC
  – Generalization
• Experiments
• Conclusion
Motivation for Representation Learning

Flying
Non-Flying
Motivation for Representation Learning

- Bovine
- Bird
- Primate
- Cat

- Flying
- Non-Flying
Motivation for Representation Learning

- Robustness
- Generalization
Motivation for Representation Learning

- Robustness
- Generalization
Motivation for Representation Learning

- Robustness
- Generalization
Motivation for Representation Learning
Motivation for Representation Learning
Contents

• Motivation
• Contrastive Learning

• Content:
  – Causal interpretation
  – RELIC
  – Generalization

• Experiments
• Conclusion
Contrastive Learning - SimCLR

![Diagram of SimCLR process]

- **Image**: $X$
- **Augmentation Functions**: $f(\cdot)$, $g(\cdot)$
- **Augmentation Set**: $T$
- **Encoder**: $h_i = f(x_i)$
- **Projector**: $z_i = g(h_i)$
- **Downstream Tasks**: $x_i = t(X)$ for $t \in T$
  - $x_j = t'(x)$ for $t' \in T$

[2] Christoph Frey
Contrastive Learning

- Augmentations ImageNet-C
Contrastive Learning - SimCLR

\[ l_{i,j} = - \log \frac{\exp(sim(z_i, z_j)/J - J)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\sin(z_i, z_k)/\tau)} \]
Contrastive Learning

- Mutual Information

\[ I(X;Y) = KL(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left( \frac{p(x, y)}{p(x) p(y)} \right) \]

- dependent on intersection

- Maximizing lower bound of Mutual Information does not always improve results.[1]
• Motivation
• Contrastive Learning
• RELIC:
  – Causal interpretation
  – RELIC Objective
  – Generalization
• Experiments
• Conclusion
Casuality Example

• Example from Elements of causal inference[4]
• $p(a,t) = p(a|t)p(t) = p(t|a)p(a)$
• Can be represented using causal graphs
Casuality Example

A → T
A ← T

Change temperature
Casuality Example

Change altitude

$p(t|a)p(a)$

$p(t|a)$ and $p(a)$ are invariant
Causal Interpretation

Diagram:

- \( S \) (Source) influences \( C \) (Condition)
- \( C \) influences \( Y_1 \) and \( Y_T \)
- \( X \) (Observed) is a pre-condition to \( C \)

Notes:
- This diagram illustrates a causal relationship where the source \( S \) affects the condition \( C \), and \( C \) in turn affects the outcomes \( Y_1 \) and \( Y_T \).
Independence of variables

$X = f(S, C)$

$S$ and $C$ represent Style and Content, respectively, and $X$ represents the generated data.
Independence of variables

Style

Content

Independence of variables

Data

C $\parallel S$

- Gaussian Noise
- Shot Noise
- Impulse Noise
- Defocus Blur
- Frosted Glass Blur
- Motion Blur
- Zoom Blur
- Snow
- Frost
- Fog
- Brightness
- Contrast
- Elastic
- Pixelate
- JPEG
$X = f(S, C)$

$p(Y_t, C) = p(Y_t|C)p(C)$
Prediction given by $p(Y_t | C = "owl")$ is owl
\[ p(Y_t, C) = p(Y_t|C)p(C) \]

- Prediction given by \( p(Y_t|C = "owl") \) is still owl
- \( p(Y_t|C)p(C) \) is **invariant** under style changes
Assumptions

• Content and Style are independent
• Data generated from Content and Style
• Only Content relevant for unknown tasks
Independence of mechanisms

- \( p(Y_t, C) = p(Y_t|C)p(C) \)
- Prediction given by \( p(Y_t|C) \)

\[
p^{\text{do}(s_i)}(Y_t|f(X)) = p^{\text{do}(s_j)}(Y_t|f(X))
\]

But \( S, C \) and \( Y_t \) are unknown
Learning Proxy Task

Data generation

Representation Learning

\[ f(X) \rightarrow Y^R \]

\[ X \rightarrow C \rightarrow Y_1 \rightarrow \cdots \rightarrow Y_T \]
• Motivation
• Contrastive Learning
• Content:
  – Causal interpretation
  – RELIC
  – Generalization
• Experiments
• Conclusion
Learn a representation \( f(X) \)

Simulate styles using augmentations

\[
p^{do(S=s_i)}(Y_t|C) = p^{do(S=s_j)}(Y_t|C)
\]

\[
p^{do(a_i)}(Y^R|f(X)) = p^{do(a_j)}(Y^R|f(X))
\]
\[ -\sum_{i=1}^{N} \sum_{a_{lk}} \log \frac{\exp(\phi(f(x_i^{a_l}), h(x_i^{a_k}))/\tau)}{\sum_{m=1}^{M} \exp(\phi(f(x_i^{a_l}), h(x_m^{a_k}))/\tau)} \]

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

\[+\alpha \sum_{a_{lk}, a_{qt}} KL (p^{do(a_{lk})}, p^{do(a_{qt})})\]

- cross-entropy
- \(\phi \left( f(x_i^a_l), h(x_m^a_k) \right) = \langle g(f(x_i)), g(h(x_i)) \rangle\)
- invariance penalty
- $\text{KL}(p^{do(a_i)}(Y^R|f(X)), p^{do(a_j)}(Y^R|f(X)) < \rho$
RELiC
Data generation

Representation Learning

Proxy task
Contents

- Motivation
- Contrastive Learning
- Content:
  - Causal interpretation
  - RELIC
  - Generalization
- Experiments
- Conclusion
Refinement

- A tasks $Y^R$ is a refinement if:
- Every equivalence class of $Y^R$ is a subset of an equivalence class of $\mathcal{Y} = \{Y_t\}_{t=1}^T$
Refinement
Refinement
Refinement
Refinement - Theorem 1

- $\mathcal{Y} = \{Y_t\}^T_{t=1}$
- $\mathcal{Y}^R$ Refinement of all tasks in $\mathcal{Y}$
- Under style interventions on S:
- If $f(X)$ is an invariant representation for $\mathcal{Y}^R$
- Then $f(X)$ is an invariant representation for all tasks in $\mathcal{Y}$
Refinement - Theorem

- $\mathcal{Y} = \{Y_t\}_{t=1}^T$
- $Y^R$ Refinement of all tasks in $\mathcal{Y}$
- Under style interventions on $S$:
- If $f(X)$ is an invariant representation for $Y^R$
- Then $f(X)$ is an invariant representation for all tasks in $\mathcal{Y}$
- $p^{do(s_i)}(Y^R | f(X)) = p^{do(s_j)}(Y^R | f(X))$
- $p^{do(s_i)}(Y_t | f(X)) = p^{o(s_j)}(Y_t | f(X))$
Contents

• Motivation
• Contrastive Learning
• Content:
  – Causal interpretation
  – RELIC
  – Generalization
• Experiments
• Conclusion
Experiments

• Linear evaluation
• Robustness and generalization
• Reinforcement Learning
Experiments

- 1. Linear Separability
- Fischer’s linear discriminant ratio

![Graph showing linear discriminant ratio distributions for ReLIC, SimCLR, and AMDIM with a distance of 162]
# Experiments

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

### ImageNet-R

<table>
<thead>
<tr>
<th>Method</th>
<th>Supervised</th>
<th>SimCLR</th>
<th>RelIC (ours)</th>
<th>BYOL</th>
<th>RelICₜ (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1 Error (%)</td>
<td>63.9</td>
<td>81.7</td>
<td>77.4</td>
<td>77.0</td>
<td>76.2</td>
</tr>
</tbody>
</table>

### ImageNet-C

<table>
<thead>
<tr>
<th>Method</th>
<th>mCE</th>
<th>mrCE</th>
<th>Gaussian</th>
<th>Shot</th>
<th>Impulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>76.7</td>
<td>105.0</td>
<td>80.0</td>
<td>82.0</td>
<td>83.0</td>
</tr>
<tr>
<td><strong>ResNet-50 architecture:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SimCLR</td>
<td>87.5</td>
<td>111.9</td>
<td>79.4</td>
<td>81.9</td>
<td>89.6</td>
</tr>
<tr>
<td>RelIC (ours)</td>
<td>76.4</td>
<td><strong>87.7</strong></td>
<td>67.8</td>
<td>70.7</td>
<td>77.0</td>
</tr>
<tr>
<td><strong>ResNet-50 with target network:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BYOL</td>
<td>72.3</td>
<td>90.0</td>
<td>65.9</td>
<td>68.4</td>
<td>73.7</td>
</tr>
<tr>
<td>RelIC (ours)</td>
<td><strong>70.8</strong></td>
<td>88.4</td>
<td><strong>63.6</strong></td>
<td><strong>65.7</strong></td>
<td><strong>69.2</strong></td>
</tr>
</tbody>
</table>
### Experiments

<table>
<thead>
<tr>
<th>Atari Performance</th>
<th>RELIC</th>
<th>SimCLR</th>
<th>CURL</th>
<th>BYOL</th>
<th>Augmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capped mean</td>
<td>91.46</td>
<td>88.76</td>
<td>90.72</td>
<td>89.43</td>
<td>80.60</td>
</tr>
<tr>
<td>Number of superhuman games</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>51</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>34</td>
</tr>
<tr>
<td>Median</td>
<td>3003.73</td>
<td>2086.16</td>
<td>2413.12</td>
<td>1769.43</td>
<td>503.15</td>
</tr>
<tr>
<td>40% Percentile</td>
<td>832.50</td>
<td>592.83</td>
<td>819.56</td>
<td>483.39</td>
<td>132.17</td>
</tr>
<tr>
<td>30% Percentile</td>
<td>356.27</td>
<td>266.07</td>
<td>409.46</td>
<td>224.80</td>
<td>94.35</td>
</tr>
<tr>
<td>20% Percentile</td>
<td>202.49</td>
<td>174.19</td>
<td>190.96</td>
<td>150.21</td>
<td>80.04</td>
</tr>
<tr>
<td>10% Percentile</td>
<td>133.93</td>
<td>120.84</td>
<td>126.10</td>
<td>118.36</td>
<td>57.95</td>
</tr>
<tr>
<td>5% Percentile</td>
<td>83.79</td>
<td>37.19</td>
<td>59.09</td>
<td>44.14</td>
<td>32.74</td>
</tr>
</tbody>
</table>
Conclusion

• Causal Interpretation
• ReLIC objective
• Condition for generalization
References


Backup Slides - Refinement

• Fineness:

• Every equivalence class of $Y^R$ is a subset of an equivalence class of $\mathcal{Y}$

• Lemma if $\sim$ is finer then $\cong$ then every equivalence class of $\cong$ is a union of equivalence classes of $\sim$
Proof for Theorem 1

\[ p^{do(s_i)}(Y_t|f(X)) = \int p^{do(s_i)}(Y_t|Y^R)p^{d_0(s_i)}(Y^R|f(X)) \, dY^R = \]

\[ \int p(Y_t|Y^R)p^{d_0(s_i)}(Y^R|f(X)) \, dY^R = \]

\[ \int P(Y_t|Y^R)p^{d_0(s_j)}(Y^R|f(x)) \, dY^R = \]

\[ p^{do(s_j)}(Y_t|f(X)) \]

as \( Y^R \) is a Refinement of \( Y_t \): \( P(Y_t|Y^R) = p(Y_t|Y^R, f(X)) \)