Data-Efficient Image Recognition with Contrastive Predictive Coding

Seminar Talk

Jonas Kindler

Computer Vision Group, Freiburg University
Table of Contents

• Contrastive predictive coding (CPC)
• InfoNCE
• Paper contributions
• Results
Feature learning

- Basic idea: Information can be learned from images without knowing the labels
- Features unsupervised -> Learned embedding
- Train simple classifier on embedding with few labels
Contrastive predictive coding

Figure 1: Contrastive predictive coding for audio input [1]
Contrastive predictive coding

Figure 2: Neural network structure for CPC [2]
Contrastive predictive coding

Figure 3: Pretraining procedure of CPC [2]
Figure 4: Procedure for linear classification [2]
Figure 5: Procedure for efficient classification [2]
Figure 6: Procedure for Transfer learning [2]
Figure 7: Procedure for supervised learning [2]
Intrinsic vs. Extrinsic Dimension

Figure 8: Intrinsic dimension [6]

Figure 9: Extrinsic high dimension [6]
Figure 10: Mutual Information of two variables [5]
**Mutual Information**

**Figure 11:** Loss measured in the output space: Autoencoder, GAN [6]

**Figure 12:** Loss measured in the representation space: CPC [6]
$$\mathcal{L}_N = - \mathbb{E}_X \left[ \log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right]$$

**Figure 13:** InfoNCE Loss [1]

$$- \sum_{i,j,k} \log \frac{\exp(\hat{z}^T_{i+k,j} z_{i+k,j})}{\exp(\hat{z}^T_{i+k,j} z_{i+k,j}) + \sum_i \exp(\hat{z}^T_{i+k,j} z_i)}$$

**Figure 14:** Network Loss function [2]
Contributions
Model capacity (MC), (LP)

- +5% Top1 performance
- +2% Top1 performance by increasing patch size

\[
\begin{bmatrix}
1 \times 1, 64 \\
3 \times 3, 64 \\
1 \times 1, 256
\end{bmatrix} \times 3
\]

\[
\begin{bmatrix}
1 \times 1, 128 \\
3 \times 3, 128 \\
1 \times 1, 512
\end{bmatrix} \times 4
\]

\[
\begin{bmatrix}
1 \times 1, 256 \\
3 \times 3, 256 \\
1 \times 1, 1024
\end{bmatrix} \times 23
\]

\[
\begin{bmatrix}
1 \times 1, 512 \\
3 \times 3, 512 \\
1 \times 1, 2048
\end{bmatrix} \times 3
\]

\[
\begin{bmatrix}
1 \times 1, 512 \\
3 \times 3, 512 \\
1 \times 1, 2048
\end{bmatrix}
\]
Layer normalization (LN)

- +2% Top1 performance
Prediction lengths and directions (BU),(HP)

- +2% Top1 performance adding bottom-up predictions
- +2.5% Top1 performance adding all four spatial directions

Figure 15: Patch Extraction in CPC [1]
Patch based augmentation (RC),(PA)

- Color dropping: 3% accuracy gain
- Random affine transformations: 4.5% accuracy again

Figure 16: Patch Extraction in CPC [1]
Results – Supervised learning

![Graph showing results for supervised learning](image)

- **ResNet trained on CPC**
- **ResNet trained on pixels**

- **Top-5 classification accuracy**

- **Percentage of labeled data**

- 80% fewer labels
- 50% fewer labels
Influence of spatial direction predictions

![Graph showing influence of spatial direction predictions](image)
Results – Unsupervised learning

• Significantly better than other representation learning methods
• Similar performance to label-propagation methods, but can be combined with them
• Generally smoother performance-label trade-off
## Results – Unsupervised learning

<table>
<thead>
<tr>
<th>Method</th>
<th>Architecture</th>
<th>Top-5 accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled data</td>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>†Supervised baseline</td>
<td>ResNet-200</td>
<td>44.1 75.2* 83.9 93.1 95.2#</td>
</tr>
<tr>
<td><strong>Methods using label-propagation:</strong></td>
<td></td>
<td>5% 10% 50% 100%</td>
</tr>
<tr>
<td>Pseudolabeling [63]</td>
<td>ResNet-50</td>
<td>51.6 - 82.4 - -</td>
</tr>
<tr>
<td>VAT + Entropy Minimization [63]</td>
<td>ResNet-50</td>
<td>47.0 - 83.4 - -</td>
</tr>
<tr>
<td>Unsup. Data Augmentation [61]</td>
<td>ResNet-50</td>
<td>- - 88.5 - -</td>
</tr>
<tr>
<td>Rotation + VAT + Ent. Min. [63]</td>
<td>ResNet-50 ×4</td>
<td>- - 91.2 - 95.0</td>
</tr>
<tr>
<td><strong>Methods using representation learning only:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instance Discrimination [60]</td>
<td>ResNet-50</td>
<td>39.2 - 77.4 - -</td>
</tr>
<tr>
<td>Rotation [63]</td>
<td>ResNet-152 ×2</td>
<td>57.5 - 86.4 - -</td>
</tr>
<tr>
<td>ResNet on BigBiGAN (fixed)</td>
<td>RevNet-50 ×4</td>
<td>55.2 73.7 78.8 85.5 87.0</td>
</tr>
<tr>
<td>ResNet on AMDIM (fixed)</td>
<td>Custom-103</td>
<td>67.4 81.8 85.8 91.0 92.2</td>
</tr>
<tr>
<td>ResNet on CPC v2 (fixed)</td>
<td>ResNet-161</td>
<td>77.1 87.5 90.5 95.0 96.2</td>
</tr>
<tr>
<td>ResNet on CPC v2 (fine-tuned)</td>
<td>ResNet-161</td>
<td>77.9* 88.6 91.2 95.6# 96.5</td>
</tr>
</tbody>
</table>
Results – Transfer learning

• Train on imageNet, then use for other task on other data
• Better performance than all tested methods
Results – Transfer learning

<table>
<thead>
<tr>
<th>Method</th>
<th>Architecture</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transfer from labeled data:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supervised baseline</td>
<td>ResNet-152</td>
<td>74.7</td>
</tr>
<tr>
<td><strong>Transfer from unlabeled data:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instance Discrimination [60]</td>
<td>ResNet-50</td>
<td>65.4</td>
</tr>
<tr>
<td>Deep Cluster [7]</td>
<td>VGG-16</td>
<td>65.9</td>
</tr>
<tr>
<td>Deeper Cluster [8]</td>
<td>VGG-16</td>
<td>67.8</td>
</tr>
<tr>
<td>Local Aggregation [66]</td>
<td>ResNet-50</td>
<td>69.1</td>
</tr>
<tr>
<td>Momentum Contrast [25]</td>
<td>ResNet-50</td>
<td>74.9</td>
</tr>
<tr>
<td>Faster-RCNN trained on CPC v2</td>
<td>ResNet-161</td>
<td><strong>76.6</strong></td>
</tr>
</tbody>
</table>
References

1 Paper: Data-Efficient Image Recognition with Contrastive Predictive Coding

2 Paper: Representation learning with contrastive predictive coding.


4 https://towardsdatascience.com/generating-images-with-autoencoders-77fd3a8dd368

5 https://en.wikipedia.org/wiki/Mutual_information