Colorization as a Proxy Task for Visual Understanding

Gustav Larsson
University of Chicago
larsson@cs.uchicago.edu

Michael Maire
TTI-Chicago
mmaire@ttic.edu

Gregory Shakhnarovich
TTI-Chicago
greg@ttic.edu

Overview

Problem statement:
• Learning a general-purpose visual representation from unlabelled data

Motivation:
• Reducing reliance on costly data annotation, especially for new problem domains

Approach:
• Training a network for automatic image colorization from scratch
• Use the trained network as a starting point for other visual tasks

Colorization as a Target Task

• Work in automatic colorization was feed-forward networks for per-pixel color predictions (Larsson et al. 2016; Zhang et al. 2018; Buda et al. 2018)
• We use the colorization model with hyperparameters from Larsson et al. (2016)

Colorization as a Proxy Task

• The steps can be reversed in order for colorizations to benefit classification (or other visual tasks)

Colorization as a Proxy Task

The steps can be reversed in order for colorizations to benefit classification (or other visual tasks)

Empirical Study

We summarize some of our findings and describe best practices:

• Model complexity
  • Colorization facilitates scaling up model complexity
  • Conv4_1 is color specific

• Loss
  • Histogram predictions are significantly better than regression

• Training time
  • Longer is better (does not plateau quickly, best model trained for 4 months)

• Learning rate
  • Important to drop during pretraining, even though downstream fine-tuning awaits

End-to-end fine-tuning on downstream task

• Much more important for colorization pretraining than supervised pretraining

Colorization as a Proxy Task

The steps can be reversed in order for colorizations to benefit classification (or other visual tasks)

Model Complexity 1 & 2

Model complexity has significant impact and evaluate their representation learning by using the pretrained models for VOC 2012 Segmentation (val) fine-tuning

End-to-end Fine-tuning

• Features change significantly more for colorization pretraining than classification pretraining
• Colorization (Col) offers large improvement over random initialization (Rnd)
• Fine-tuned layers (VGG-16)

Training Time 1 & Learning Rate 1

Relation between proxy loss (colorization) and downstream loss (classification) for automatic segmentation

Feature Visualization

Examples of top activations (4/7) of a colorization network

Results

Task: Downstream training without supervised pretraining

Initialization: Architecture Simulation Segmentation (val) (s)

<table>
<thead>
<tr>
<th>Method</th>
<th>Architecture</th>
<th>Simulation</th>
<th>Segmentation (val)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>AlexNet</td>
<td>30-125</td>
<td>70-125 (val)</td>
</tr>
<tr>
<td>Autoencoders</td>
<td>AlexNet</td>
<td>40-125</td>
<td>70-125 (val)</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>AlexNet</td>
<td>40-125</td>
<td>70-125 (val)</td>
</tr>
<tr>
<td>VGG-16</td>
<td>AlexNet</td>
<td>40-125</td>
<td>70-125 (val)</td>
</tr>
</tbody>
</table>

Goals:
- VGG-16 is a standard network for classification
- We want to evaluate our method on a standard dataset

Bonne: Revisiting supervised pretraining

Performance: A Table

<table>
<thead>
<tr>
<th>Method</th>
<th>Simulation</th>
<th>Segmentation (val)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>AlexNet</td>
<td>30-125 (val)</td>
</tr>
<tr>
<td>Autoencoders</td>
<td>AlexNet</td>
<td>40-125 (val)</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>AlexNet</td>
<td>40-125 (val)</td>
</tr>
<tr>
<td>VGG-16</td>
<td>AlexNet</td>
<td>40-125 (val)</td>
</tr>
</tbody>
</table>

Goal:
- Baseline: AlexNet
- Autoencoders: AlexNet
- ResNet-152: AlexNet
- VGG-16: AlexNet

VOC Comparison:
- Comparison with other unsupervised pretraining methods on VOC 2012 Segmentation (val)
- Our method: AlexNet
- Baseline: AlexNet
- Autoencoders: AlexNet
- ResNet-152: AlexNet
- VGG-16: AlexNet

Bonus:
- Revisiting supervised pretraining

Performance: A Table

<table>
<thead>
<tr>
<th>Method</th>
<th>Simulation</th>
<th>Segmentation (val)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>AlexNet</td>
<td>30-125 (val)</td>
</tr>
<tr>
<td>Autoencoders</td>
<td>AlexNet</td>
<td>40-125 (val)</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>AlexNet</td>
<td>40-125 (val)</td>
</tr>
<tr>
<td>VGG-16</td>
<td>AlexNet</td>
<td>40-125 (val)</td>
</tr>
</tbody>
</table>

VOC Comparison:
- Comparison with other unsupervised pretraining methods on VOC 2012 Segmentation (val)
- Our method: AlexNet
- Baseline: AlexNet
- Autoencoders: AlexNet
- ResNet-152: AlexNet
- VGG-16: AlexNet