Large-Scale Evolution of Image Classifiers

Esteban Real\textsuperscript{1}  Sherry Moore\textsuperscript{1}  Andrew Selle\textsuperscript{1}  Saurabh Saxena\textsuperscript{1}
Yutaka Leon Suematsu\textsuperscript{2}  Jie Tan\textsuperscript{1}  Quoc V.Le\textsuperscript{1}  Alexey Kurakin\textsuperscript{1}

\textsuperscript{1}Google Brain

\textsuperscript{2}Google Research

ICML, 2017
Presenter: Tianlu Wang
Outline

1 Introduction
   - Motivation
   - Backgrounds

2 Related Work
   - Neuro-evolution
   - Non-evolutionary

3 Methods
   - Algorithm Overview
   - Encoding and Mutations
   - More Details

4 Results
   - Progress of experiments
   - Comparisons
   - Meta-parameters

5 Summary
Outline

1 Introduction
   - Motivation
   - Backgrounds

2 Related Work
   - Neuro-evolution
   - Non-evolutionary

3 Methods
   - Algorithm Overview
   - Encoding and Mutations
   - More Details

4 Results
   - Progress of experiments
   - Comparisons
   - Meta-parameters

5 Summary
Motivation

- AlexNet, GoogleNet, VGG, ResNet...
- Designing neural network architectures can be challenging
- Discover network architectures automatically
Outline

1 Introduction
   • Motivation
   • Backgrounds

2 Related Work
   • Neuro-evolution
   • Non-evolutionary

3 Methods
   • Algorithm Overview
   • Encoding and Mutations
   • More Details

4 Results
   • Progress of experiments
   • Comparisons
   • Meta-parameters

5 Summary
Achievements: evolution algorithm outputs a **fully-trained** model with no human participation

Drawbacks: significant computation

Image classification, CIFAR-10, CIFAR-100
Outline

1 Introduction
   - Motivation
   - Backgrounds

2 Related Work
   - Neuro-evolution
   - Non-evolutionary

3 Methods
   - Algorithm Overview
   - Encoding and Mutations
   - More Details

4 Results
   - Progress of experiments
   - Comparisons
   - Meta-parameters

5 Summary
Neuro-evolution

- Weight evolution: back propagation
- Weight and architecture: *NEAT algorithm* (node and connection)
Outline

1 Introduction
   • Motivation
   • Backgrounds

2 Related Work
   • Neuro-evolution
   • Non-evolutionary

3 Methods
   • Algorithm Overview
   • Encoding and Mutations
   • More Details

4 Results
   • Progress of experiments
   • Comparisons
   • Meta-parameters

5 Summary
Non-evolutionary

- Bayesian optimization
- Reinforcement learning
- Q-learning
Outline

1. Introduction
   - Motivation
   - Backgrounds

2. Related Work
   - Neuro-evolution
   - Non-evolutionary

3. Methods
   - Algorithm Overview
   - Encoding and Mutations
   - More Details

4. Results
   - Progress of experiments
   - Comparisons
   - Meta-parameters

5. Summary
Algorithm Overview

- Input: a population of models, each model is a trained single-layer nonconvolutional model with learning rate $= 0.1$
- Measurement: accuracy on validation dataset
Algorithm Overview

- **Input:** a population of models, each model is a trained single-layer nonconvolutional model with learning rate $= 0.1$
- **Measurement:** accuracy on validation dataset
Algorithm Overview

- Input: a population of models, each model is a trained single-layer nonconvolutional model with learning rate = 0.1
- Measurement: accuracy on validation dataset

When to stop?

Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V.Le, Alexey Kurakin (Google Brain and Google Research)

Large-Scale Evolution of Image Classifiers

ICML, 2017
Outline

1. Introduction
   - Motivation
   - Backgrounds

2. Related Work
   - Neuro-evolution
   - Non-evolutionary

3. Methods
   - Algorithm Overview
   - Encoding and Mutations
   - More Details

4. Results
   - Progress of experiments
   - Comparisons
   - Meta-parameters

5. Summary
Model Encoding

Individual model is encoded as a graph:

- **Vertices**
  - rank-3 tensor(image width * image height * channels)
  - activations(batch normalization with ReLU or plain linear layer)

- **Edges**
  - Identity connections
  - Convolutions
Model Encoding

Individual model is encoded as a graph:

- **Vertices**
  - rank-3 tensor\(\text{image\_width} \times \text{image\_height} \times \text{channels}\)
  - activations\(\text{batch normalization with ReLU or plain linear layer}\)

- **Edges**
  - Identity connections
  - Convolutions

Inconsistent input:

- pick and keep primary one
- reshape\(\text{interpolation/truncation/padding}\) non-primary ones
The worker picks a mutation at random from a set:

- ALTER-LEARNING-RATE
- IDENTITY (effectively means keep training)
- RESET-WEIGHTS
- INSERT/REMOVE CONVOLUTION
- ALTER-STRIDE
- ALTER-NUMBER-OF-CHANNELS
- FILTER-SIZE
- INSERT-ONE-TO-ONE
- INSERT/REMOVE SKIP
Outline

1. Introduction
   - Motivation
   - Backgrounds

2. Related Work
   - Neuro-evolution
   - Non-evolutionary

3. Methods
   - Algorithm Overview
   - Encoding and Mutations
   - More Details

4. Results
   - Progress of experiments
   - Comparisons
   - Meta-parameters

5. Summary
- Poor initial conditions (12th slide)
- 45,000 training; 5,000 validation; 10,000 test
- SGD with momentum of 0.9, batch size 50, weight decay 0.0001
- Computation cost: floating-point operations
- Inherit parameters’ weights whenever possible
Outline

1. Introduction
   - Motivation
   - Backgrounds

2. Related Work
   - Neuro-evolution
   - Non-evolutionary

3. Methods
   - Algorithm Overview
   - Encoding and Mutations
   - More Details

4. Results
   - Progress of experiments
   - Comparisons
   - Meta-parameters

5. Summary
Progress of an evolution experiment
Repeatability of results and controls
Outline

1. Introduction
   - Motivation
   - Backgrounds

2. Related Work
   - Neuro-evolution
   - Non-evolutionary

3. Methods
   - Algorithm Overview
   - Encoding and Mutations
   - More Details

4. Results
   - Progress of experiments
   - Comparisons
   - Meta-parameters

5. Summary
Compared to hand-designed networks

<table>
<thead>
<tr>
<th>STUDY</th>
<th>Params</th>
<th>C10+</th>
<th>C100+</th>
<th>REACHABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxout (Goodfellow et al., 2013)</td>
<td>-</td>
<td>90.7%</td>
<td>61.4%</td>
<td>No</td>
</tr>
<tr>
<td>Network in Network (Lin et al., 2013)</td>
<td>-</td>
<td>91.2%</td>
<td>-</td>
<td>No</td>
</tr>
<tr>
<td>All-CNN (Springenberg et al., 2014)</td>
<td>1.3 M</td>
<td>92.8%</td>
<td>66.3%</td>
<td>Yes</td>
</tr>
<tr>
<td>Deeply Supervised (Lee et al., 2015)</td>
<td>-</td>
<td>92.0%</td>
<td>65.4%</td>
<td>No</td>
</tr>
<tr>
<td>Highway (Srivastava et al., 2015)</td>
<td>2.3 M</td>
<td>92.3%</td>
<td>67.6%</td>
<td>No</td>
</tr>
<tr>
<td>ResNet (He et al., 2016)</td>
<td>1.7 M</td>
<td>93.4%</td>
<td>72.8%</td>
<td>Yes</td>
</tr>
<tr>
<td>Evolution (ours)</td>
<td>5.4 M</td>
<td>94.6%</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>40.4 M</td>
<td></td>
<td>77.0%</td>
<td></td>
</tr>
<tr>
<td>Wide ResNet 28-10 (Zagoruyko &amp; Komodakis, 2016)</td>
<td>36.5 M</td>
<td>96.0%</td>
<td>80.0%</td>
<td>Yes</td>
</tr>
<tr>
<td>Wide ResNet 40-10+d/o (Zagoruyko &amp; Komodakis, 2016)</td>
<td>50.7 M</td>
<td>96.2%</td>
<td>81.7%</td>
<td>No</td>
</tr>
<tr>
<td>DenseNet (Huang et al., 2016a)</td>
<td>25.6 M</td>
<td>96.7%</td>
<td>82.8%</td>
<td>No</td>
</tr>
</tbody>
</table>
Compared to auto-discovered networks

<table>
<thead>
<tr>
<th>Study</th>
<th>Starting Point</th>
<th>Constraints</th>
<th>Post-Processing</th>
<th>Params.</th>
<th>C10+</th>
<th>C100+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian</td>
<td>3 layers</td>
<td>fixed architecture, no skips</td>
<td>none</td>
<td>–</td>
<td>90.5%</td>
<td>–</td>
</tr>
<tr>
<td>(Snoek et al., 2012)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Learning</td>
<td>–</td>
<td>discrete params., max. num. layers, no skips</td>
<td>tune, retrain</td>
<td>11.2 M</td>
<td>93.1%</td>
<td>72.9%</td>
</tr>
<tr>
<td>(Baker et al., 2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RL (Zoph &amp; Le, 2016)</td>
<td>20 layers, 50% skips</td>
<td>discrete params., exactly 20 layers</td>
<td>small grid search, retrain</td>
<td>2.5 M</td>
<td>94.0%</td>
<td>–</td>
</tr>
<tr>
<td>RL (Zoph &amp; Le, 2016)</td>
<td>39 layers, 2 pool layers at 13 and 26, 50% skips</td>
<td>discrete params., exactly 39 layers, 2 pool layers at 13 and 26</td>
<td>add more filters, small grid search, retrain</td>
<td>37.0 M</td>
<td>96.4%</td>
<td>–</td>
</tr>
<tr>
<td>Evolution</td>
<td>single layer, zero convs.</td>
<td>power-of-2 strides</td>
<td>none</td>
<td>5.4 M</td>
<td>94.6%</td>
<td>–</td>
</tr>
<tr>
<td>(Ours)</td>
<td></td>
<td></td>
<td></td>
<td>40.4 M</td>
<td>77.0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ENSMB.</td>
<td>95.6%</td>
<td></td>
</tr>
</tbody>
</table>
Outline

1 Introduction
   - Motivation
   - Backgrounds

2 Related Work
   - Neuro-evolution
   - Non-evolutionary

3 Methods
   - Algorithm Overview
   - Encoding and Mutations
   - More Details

4 Results
   - Progress of experiments
   - Comparisons
   - Meta-parameters

5 Summary
Improve the method

- Large population size
- More training steps
- Increase mutation rate
- Reset all weights
Summary

- Neuro-evolution starts from trivial initial conditions and yields fully trained models
- Construct large, accurate networks for two challenging and popular image classification benchmarks
- Large search space and high computation cost