Ask Me Anything: Dynamic Memory Networks for Natural Language Processing

Ankit Kumar Peter Ondruska Mohit Ivyer James Bradbury
Ishaan Gulrajani Victor Zhong Romain Paulus Richard Socher

MetaMind

ICML, 2017
Presenter: Tianlu Wang

Outline

Introduction

- Tasks in natural language processing can be cast as a question answering problem:
  - Machine Translation $\Rightarrow$ What is the translation into French?
  - Name entity recognition $\Rightarrow$ What are the name entity tags in this sentence?

$$\text{Input sequences}+\text{Questions}$$
$$\downarrow$$
$$\text{Dynamic Memory Network(DMN)}$$
$$\downarrow$$
$$\text{Answers}$$

- State-of-the-art on multiple dataset:
  - Question answering(Facebook bAbI dataset)
  - Text classification for sentiment analysis(Stanford Sentiment Treebank)
  - Sequence modeling for part-of-speech tagging(WSJ-PTB)

Intuition from Neuroscience

- The episodic memory in humans stores specific experiences in their spatial and temporal context.
- Provide a vector representation to capture all relevant information from input sequences and questions.

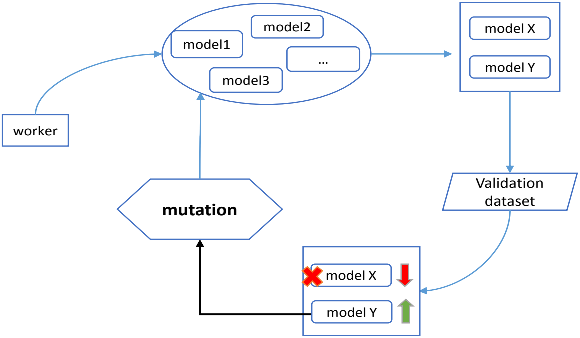
Outline

Model 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

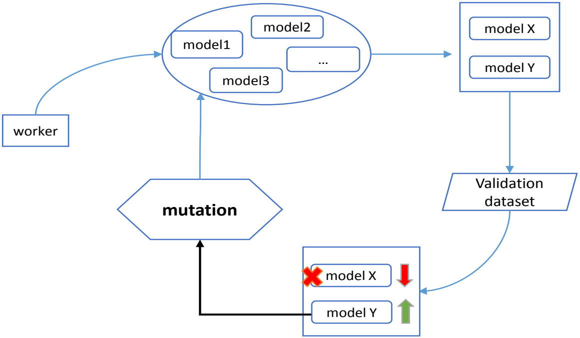
Model Overview

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



Model 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?

Outline

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(image_width * image_height * channels)
  - activations(batch normalization with ReLU or plain linear layer)
- Edges
  - Identity connections
  - Convolutions

Inconsistent input:

- pick and keep primary one
- reshape(interpolation/truncation/padding) non-primary ones

Mutations

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

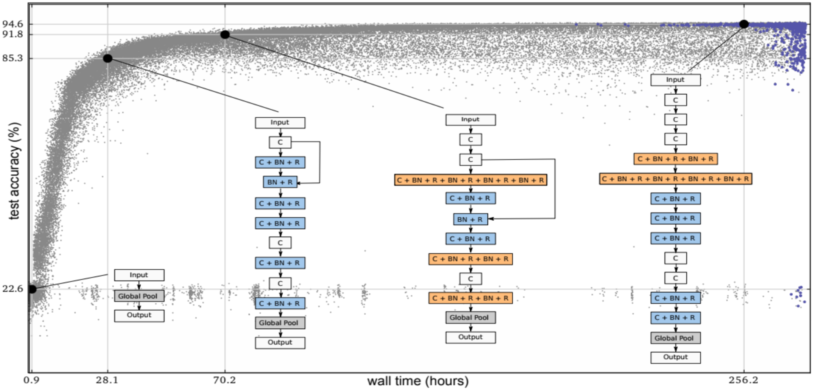
Outline

More Details

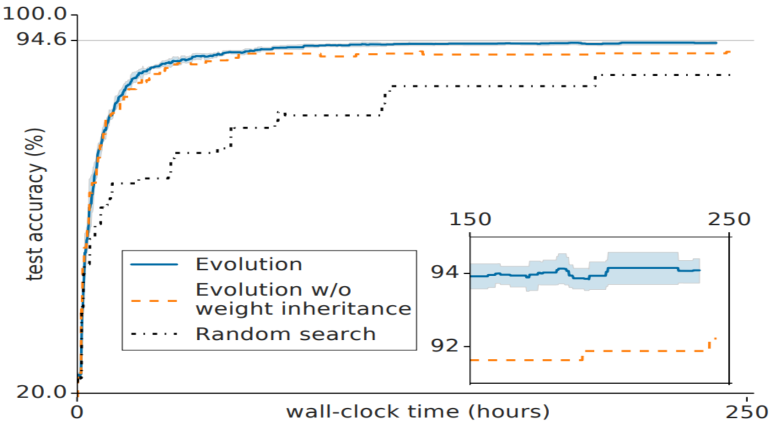
- Poor initial conditions(12th silde)
- 45,000 training; 5,000 validation; 10000 test
- SGD with momentum of 0.9, batch size 50, weight decay 0.0001
- Computation cost: floating-point operations
- Inherit parameters' weights whenever possible

Outline

Progress of an evolution experiment



Repeatability of results and controls



Outline

Compared to hand-designed networks

Study	Params.	C10+	C100+	Reachable?
Maxout (Goodfellow et al., 2013)	–	90.7%	61.4%	No
Network in Network (Lin et al., 2013)	–	91.2%	–	No
All-CNN (Springenberg et al., 2014)	1.3 M	92.8%	66.3%	Yes
Deeply Supervised (Lee et al., 2015)	–	92.0%	65.4%	No
Highway (Srivastava et al., 2015)	2.3 M	92.3%	67.6%	No
ResNet (He et al., 2016)	1.7 M	93.4%	72.8%†	Yes
Evolution (ours)	5.4 M	94.6%		N/A
	40.4 M		77.0%	
Wide ResNet 28-10 (Zagoruyko & Komodakis, 2016)	36.5 M	96.0%	80.0%	Yes
Wide ResNet 40-10+d/o (Zagoruyko & Komodakis, 2016)	50.7 M	96.2%	81.7%	No
DenseNet (Huang et al., 2016a)	25.6 M	96.7%	82.8%	No

Compared to auto-discovered networks

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

Outline

1. Introduction
2. Dynamic Memory Network
   - Model Overview
   - Encoding and Mutations
   - More Details
3. Results
   - Progress of experiments
   - Comparisons
   - Meta-parameters
4. 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