Optimization as a model for few-shot learning

Sachin Ravi[1] Hugo Larochelle[1]

[1]Twitter

ICLR, 2017
Presenter: Beilun Wang

Outline

1. Introduction
   - Motivation
   - Previous Solutions
   - Contributions

2. Proposed Methods
   - gradient descent and LSTM
   - The Proposed Method

3. Summary

Outline

Motivation

Motivation:

- Deep Learning has shown great success in a variety of tasks with large amounts of labeled data.
- Perform poorly on few-shot learning tasks
- This paper uses an LSTM based *meta-learner* model to learn the exact optimization algorithm.

Problem Setting:

Problem Setting:

- Input: meta-sets $\mathscr{D}$. For each $D \in \mathscr{D}$ has a split of D_{train} and D_{test}.
- Target: an LSTM-based *meta-learner*.
- Output: a neural network

Outline

1. Introduction
   - Motivation
   - Previous Solutions
   - Contributions

2. Proposed Methods
   - gradient descent and LSTM
   - The Proposed Method

3. Summary

Previous Solutions

- gradient-based optimization
  - momentum
  - adagrad
  - Adadelta
  - ADAM
- learning to learn
- no strong guarantees of speed of convergence
- meta-learning
  - quick acquisition of knowledge within each separate task presented
  - slower extraction of information learned across all the tasks.

Outline

1. Introduction
   - Motivation
   - Previous Solutions
   - Contributions
2. Proposed Methods
   - gradient descent and LSTM
   - The Proposed Method
3. Summary

Contributions

- An LSTM based meta-learner model
- Achieve better performance in few-shot learning task

Outline

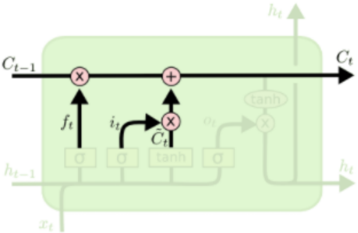
1. Introduction
   - Motivation
   - Previous Solutions
   - Contributions
2. Proposed Methods
   - gradient descent and LSTM
   - The Proposed Method
3. Summary

Gradient-based method

- $\theta_t = \theta_{t-1} - \alpha_t \nabla_{\theta_{t-1}} \mathcal{L}_t$

the Update for the cell state in an LSTM

- $c_t = f_t \odot c_{t-1} + i_t \cdot \tilde{c}_t$



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

- if $f_t = 1$, $c_{t-1} = \theta_{t-1}, i_t = \alpha_t$, and $\tilde{c}_t = -\nabla_{\theta_{t-1}} \mathcal{L}_t$
- Then it equals to gradient-based approach.

Outline

1. Introduction
   - Motivation
   - Previous Solutions
   - Contributions

2. Proposed Methods
   - gradient descent and LSTM
   - The Proposed Method

3. Summary

The formulation of the meta-learner

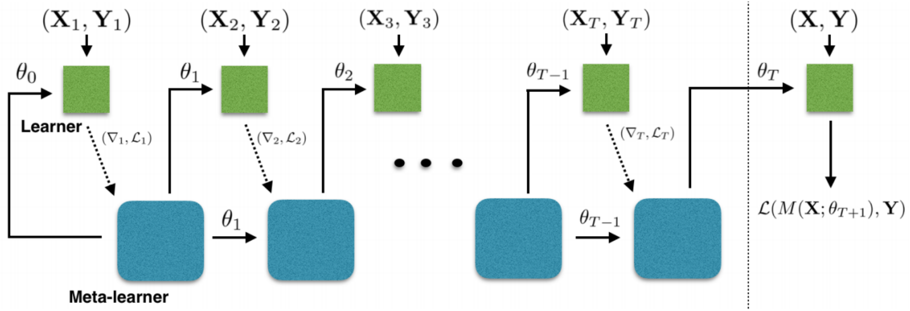
- learning rate i_t:
- $i_t = \sigma(\mathbf{W}_I \cdot [\nabla_{\theta_{t-1}} \mathcal{L}_t, \mathcal{L}_t, \theta_{t-1}, i_{t-1}] + \mathbf{b}_I)$
- f_t:
- $f_t = \sigma(\mathbf{W}_F \cdot [\nabla_{\theta_{t-1}} \mathcal{L}_t, \mathcal{L}_t, \theta_{t-1}, f_{t-1}] + \mathbf{b}_F)$

Parameter sharing and Normalization

- Share parameters across the coordinates of the learner gradient
- Each dimension has its own hidden and cell state values but the LSTM parameters are the same across all coordinates.
- Normalization the gradients and the losses across different dimensions

$$x \to \begin{cases} \left(\frac{\log(|x|)}{p}, \text{sign}(x)\right) \text{ if } |x| \geq e^{-p} \\ (-1, e^{p}x) \text{ otherwise} \end{cases} \quad (1)$$

Summary Figure



Experiment Results–average classification accuracy

Model	5-class	
	1-shot	5-shot
Baseline-finetune	$28.86 \pm 0.54\%$	$49.79 \pm 0.79\%$
Baseline-nearest-neighbor	$41.08 \pm 0.70\%$	$51.04 \pm 0.65\%$
Matching Network	$\mathbf{43.40 \pm 0.78\%}$	$51.09 \pm 0.71\%$
Matching Network FCE	$\mathbf{43.56 \pm 0.84\%}$	$55.31 \pm 0.73\%$
Meta-Learner LSTM (OURS)	$\mathbf{43.44 \pm 0.77\%}$	$\mathbf{60.60 \pm 0.71\%}$

Table 1: Average classification accuracies on Mini-ImageNet with 95% confidence intervals. Marked in bold are the best results for each scenario, as well as other results with an overlapping confidence interval.

Summary

- This paper proposes an LSTM based meta-learner model.
- It improves the performance of training deep Neural networks in few-shot learning tasks.