Is the Meta-Learning Idea Able to Improve the Generalization of Deep Neural Networks on the Standard Supervised Learning?

Xiang Deng, Zhongfei Zhang
Computer Science Department, State University of New York at Binghamton

Motivation

- Meta-learning approaches exhibit powerful generalization in few-shot learning.
- Intuitively, few-shot learning is more challenging than the standard supervised learning as each class only has a very few or no training samples.

The natural question that arises is whether the meta-learning idea can be used for improving the generalization of deep neural networks on standard supervised learning.

Key Ideas

- We propose a novel meta-learning based training procedure (MLTP) for DNNs and demonstrate that the meta-learning idea can indeed improve the generalization abilities of DNNs on standard supervised learning.

- The key idea of MLTP is that the gradient descent step for improving the current task performance should also improve a new task performance, which is ignored by the current standard procedure for training DNNs.

MLTP

- In every gradient descent iteration, MLTP randomly takes two different batches of training samples (x^i_{bat}, y^i_{bat}) and (x^j_{bat}, y^j_{bat}) as two tasks $task_i$ and $task_j$, respectively.

- The loss on $task_i$ is written as:

$$C(w, x^i_{bat}, y^i_{bat}) = L(f(w, x^i_{bat}), y^i_{bat})$$

- MLTP requires the parameters w after one gradient descent on the current task to also work well on a new task. The loss on the new task $task_j$ is written as:

$$C(w - \alpha\frac{\partial L(f(w, x^i_{bat}), y^i_{bat})}{\partial w}, x^j_{bat}, y^j_{bat}) = L(f(w - \alpha\frac{\partial L(f(w, x^i_{bat}), y^i_{bat})}{\partial w}, x^j_{bat}), y^j_{bat})$$

where α is an online adapted hyperparameter.

- The final objective function is the sum of the weighted losses from $task_i$ and $task_j$:

$$J = C(w, x^i_{bat}, y^i_{bat}) + \eta C(w - \alpha\frac{\partial L(f(w, x^i_{bat}), y^i_{bat})}{\partial w}, x^j_{bat}, y^j_{bat})$$

MLTP Framework

Algorithm 1 MLTP

Input: Training data (X_{tra}, Y_{tra}), a neural network f with parameters w

Output: The optimal parameters w

for iterations = 1, 2, ..., n **do**

  Randomly take two different batches of samples (x^i_{bat}, y^i_{bat}) and (x^j_{bat}, y^j_{bat}) as two tasks

  Compute the loss of the first task (x^i_{bat}, y^i_{bat}): $L(f(w, x^i_{bat}), y^i_{bat})$

  Do one gradient step to w: $w' = w - \alpha \frac{\partial L(f(w, x^i_{bat}), y^i_{bat})}{\partial w}$ where α are the online adapted inner step sizes

  Apply w' to the second task (x^j_{bat}, y^j_{bat}) to obtain the loss: $L(f(w - \alpha \frac{\partial L(f(w, x^i_{bat}), y^i_{bat})}{\partial w}, x^j_{bat}), y^j_{bat})$

  Obtain the final objective function: $J = L(f(w, x^i_{bat}), y^i_{bat}) + \eta L(f(w - \alpha \frac{\partial L(f(w, x^i_{bat}), y^i_{bat})}{\partial w}, x^j_{bat}), y^j_{bat})$

  Update w and α: $w = w - r\frac{\partial J}{\partial w}$; $\alpha = \alpha - r\frac{\partial J}{\partial \alpha}$ where r is the learning rate

end for

Theoretical Analysis of MLTP

- We provide the first-order Taylor expansion of the objective function:

$$J = C(w, x^i_{bat}, y^i_{bat}) + \eta C(w, x^j_{bat}, y^j_{bat}) - \eta\alpha \frac{\partial C(w, x^i_{bat}, y^i_{bat})}{\partial w} \cdot \frac{\partial C(w, x^j_{bat}, y^j_{bat})}{\partial w}$$

where . denotes the inner product operation.

- The first two terms on the right hand side minimize the losses on both $task_i$ and $task_j$ while the third term maximizes the similarity between the gradients on the two tasks.
- The third term is the main difference between MLTP and the standard training procedure.

MLTP Variates

Minimizing the objective requires the second derivatives with respect to w, which may be computationally expensive, especially for large neural networks. To address this issue, we introduce three alternative MLTP variants:

- $MLTP_{conv}$: it only applies MLTP to the convolutional layers of a DNN.
- $MLTP_{fc}$: it only applies MLTP to the fully connected layers.
- $MLTP_{FO}$: it only uses the first-order derivatives of the objective to update w by ignoring the second derivatives (similar to the case in first-order MAML [1] or Reptile [2]).

Experiments

Test Accuracies on CIFAR-10

	Standard Training	Ours			
		MLTP	$MLTP_{conv}$	$MLTP_{fc}$	$MLTP_{FO}$
CNet1	81.9±0.29	82.4±0.26	82.3±0.17	82.3±0.20	**82.6±0.24**
CNet2	86.0±0.24	86.4±0.19	86.3±0.27	86.4±0.23	**86.7±0.20**
CNet3	85.9±0.19	86.5±0.15	86.5±0.22	86.6±0.15	**86.7±0.17**
CNet4	93.3±0.22	-*	-*	-*	**93.6±0.16**

Test Accuracies on CIFAR-100

	Standard Training	Ours			
		MLTP	$MLTP_{conv}$	$MLTP_{fc}$	$MLTP_{FO}$
CCNet1	55.0±0.24	55.3±0.21	55.5±0.25	**55.7±0.22**	55.4±0.19
CCNet2	58.8±0.18	**59.7±0.25**	59.1±0.20	59.2±0.17	59.5±0.23
CCNet3	58.4±0.22	58.5±0.26	59.0±0.19	**59.5±0.24**	59.0±0.20
CCNet4	71.9±0.19	-*	-*	-*	**72.4±0.18**

Test Accuracies on Tiny ImageNet

	Standard Training		$MLTP_{FO}$	
	TOP1	TOP5	TOP1	TOP5
ResNet-18	53.2±0.27	76.5±0.24	**54.5±0.22**	**77.2±0.23**
ResNet-34	54.3±0.21	77.1±0.17	**54.9±0.20**	**77.2±0.18**

Conclusion

- Considering that meta-learning has shown excellent generalization abilities on few-shot learning, we study the question of whether meta-learning can be used to further tap the potential generalization abilities of DNNs on standard supervised learning.

- We have proposed a meta-learning based training procedure (MLTP) and have demonstrated that meta-learning can indeed improve the generalization abilities of DNNs on standard supervised learning.

- Experimental results with DNNs of various sizes on three benchmark datasets have demonstrated the effectiveness of MLTP.

- To the end, we bridge the gap between meta-learning and the generalization of DNNs on standard supervised learning by MLTP.

References:

[1] Model-agnostic meta-learning for fast adaptation of deep networks, Finn, Chelsea and Abbeel, Pieter and Levine, Sergey, ICML 2017

[2] On first-order meta-learning algorithms, Nichol, Alex and Achiam, Joshua and Schulman, John, arXiv preprint arXiv:1803.02999, 2018

[3] Is the Meta-Learning Idea Able to Improve the Generalization of Deep Neural Networks on the Standard Supervised Learning?
Deng, Xiang and Zhang, Zhongfei, ICPR 2020

Thank you for listening!