Learning to Few-Shot Learn Across Diverse Natural Language Classification Tasks

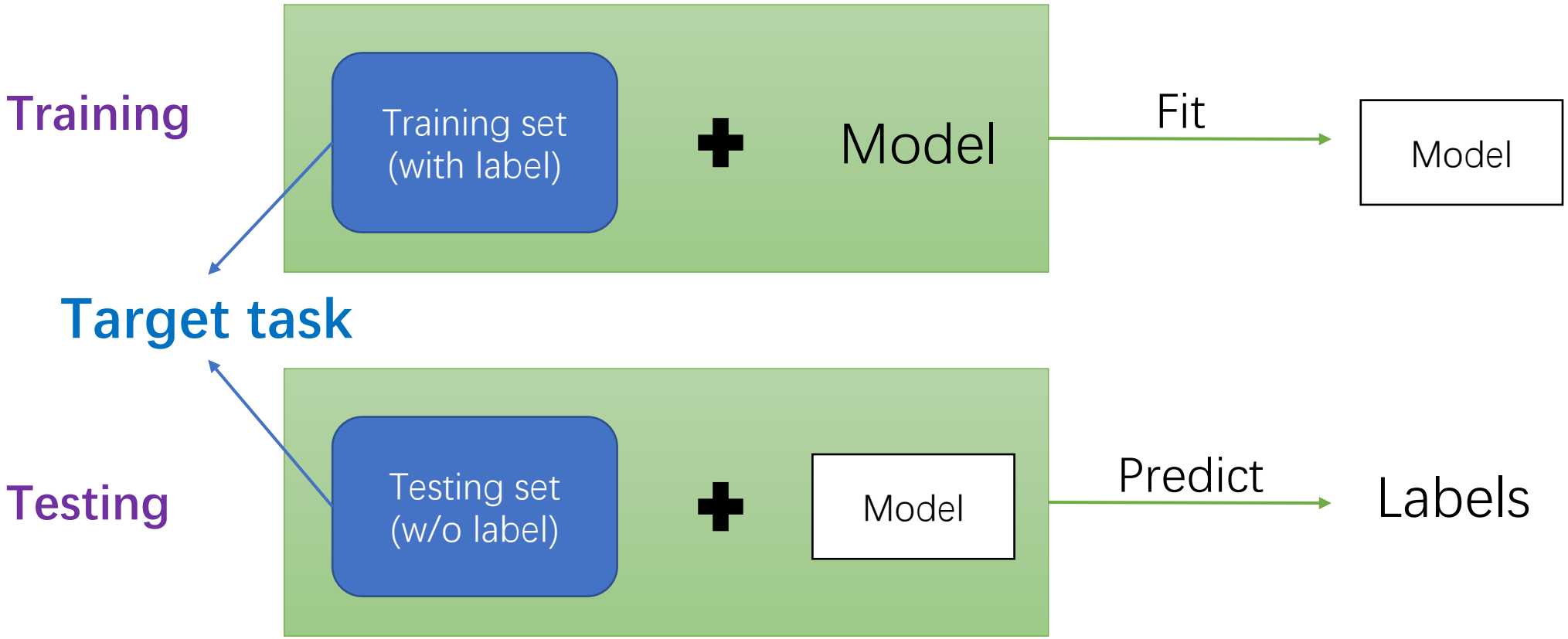
By 李想

2020. 12. 25

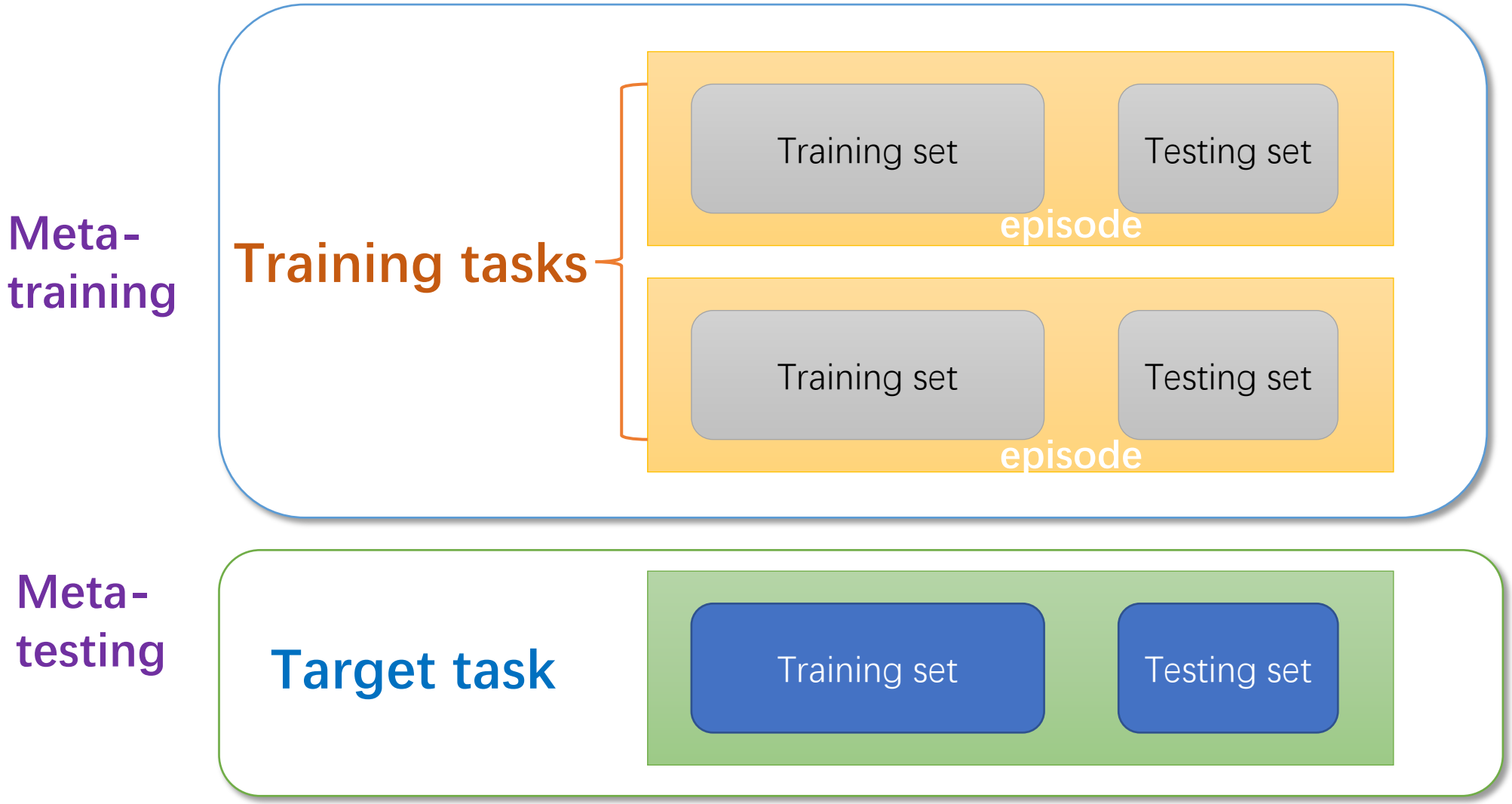
Background: Meta Learning

- Machine Learning = 根据已有数据 找一个函数 f 的能力
- Meta Learning = 根据已有数据找一个 找一个函数 f 的函数 F 的能力（Learn to learn）
  - 授人以鱼不如授人以渔——掌握如何学习的方法

Foreword: Supervised learning

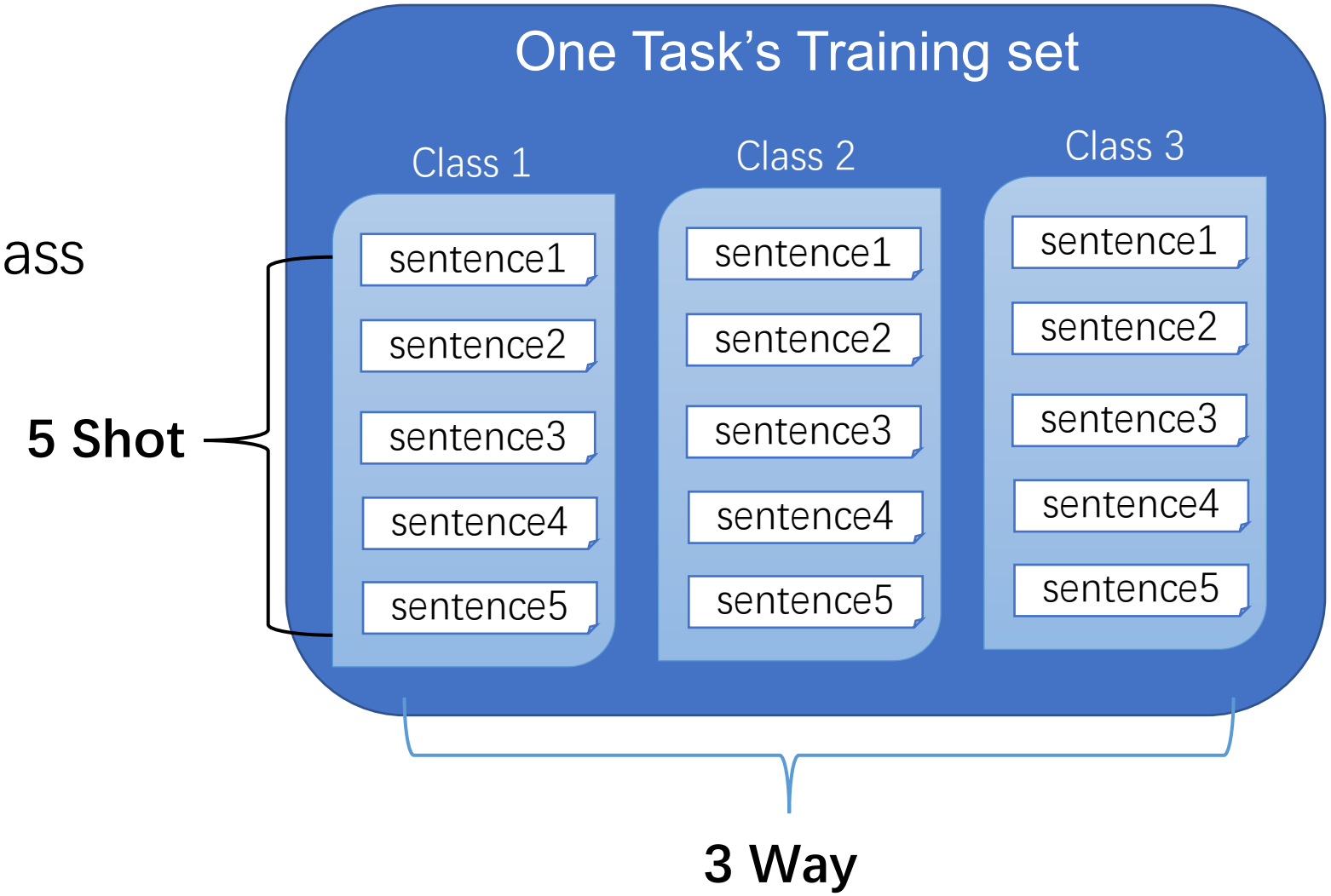


Background: Meta Learning (in few-shot learning)

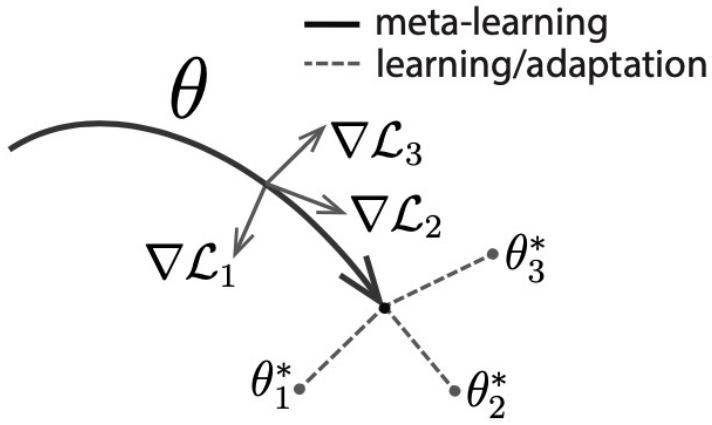


Background: N-way K-shot setting

- N: #class in a task
- K: #samples in a class
- Example:
  3-way 5-shot

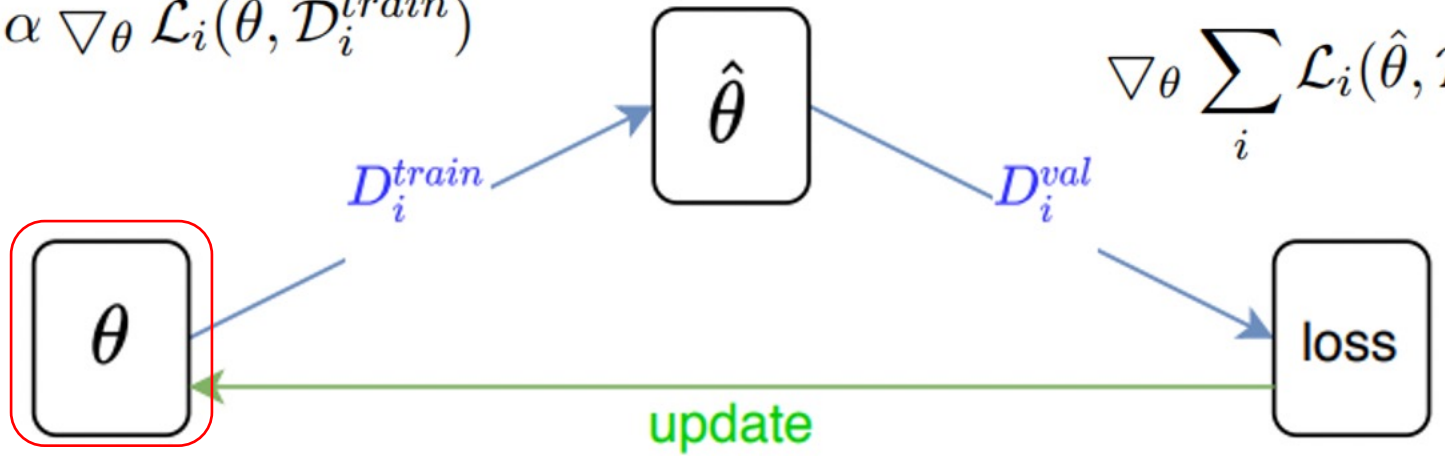


Background: MAML



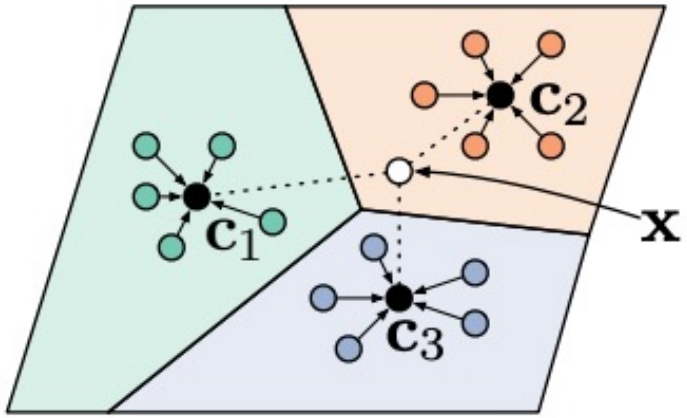
$$\hat{\theta} = \theta - \alpha \bigtriangledown_\theta \mathcal{L}_i(\theta, \mathcal{D}_i^{train})$$

$$\bigtriangledown_\theta \sum_i \mathcal{L}_i(\hat{\theta}, \mathcal{D}_i^{val})$$



$$\theta = \theta - \beta \bigtriangledown_\theta \sum_i \mathcal{L}_i(\hat{\theta}, \mathcal{D}_i^{val})$$

Background: Prototypical network



- More metric-based method:
  - Matching network
  - Relation network
  - ……

Back to this paper: Motivation

- Pre-trained transformer: fine-tuning on a new task still **requires large amounts of task-specific labeled data** to achieve good performance （标注数据耗时耗力耗钱）

- Meta-learning: limited to simulated problems or problems with limited diversity across tasks （任务缺乏多样性）

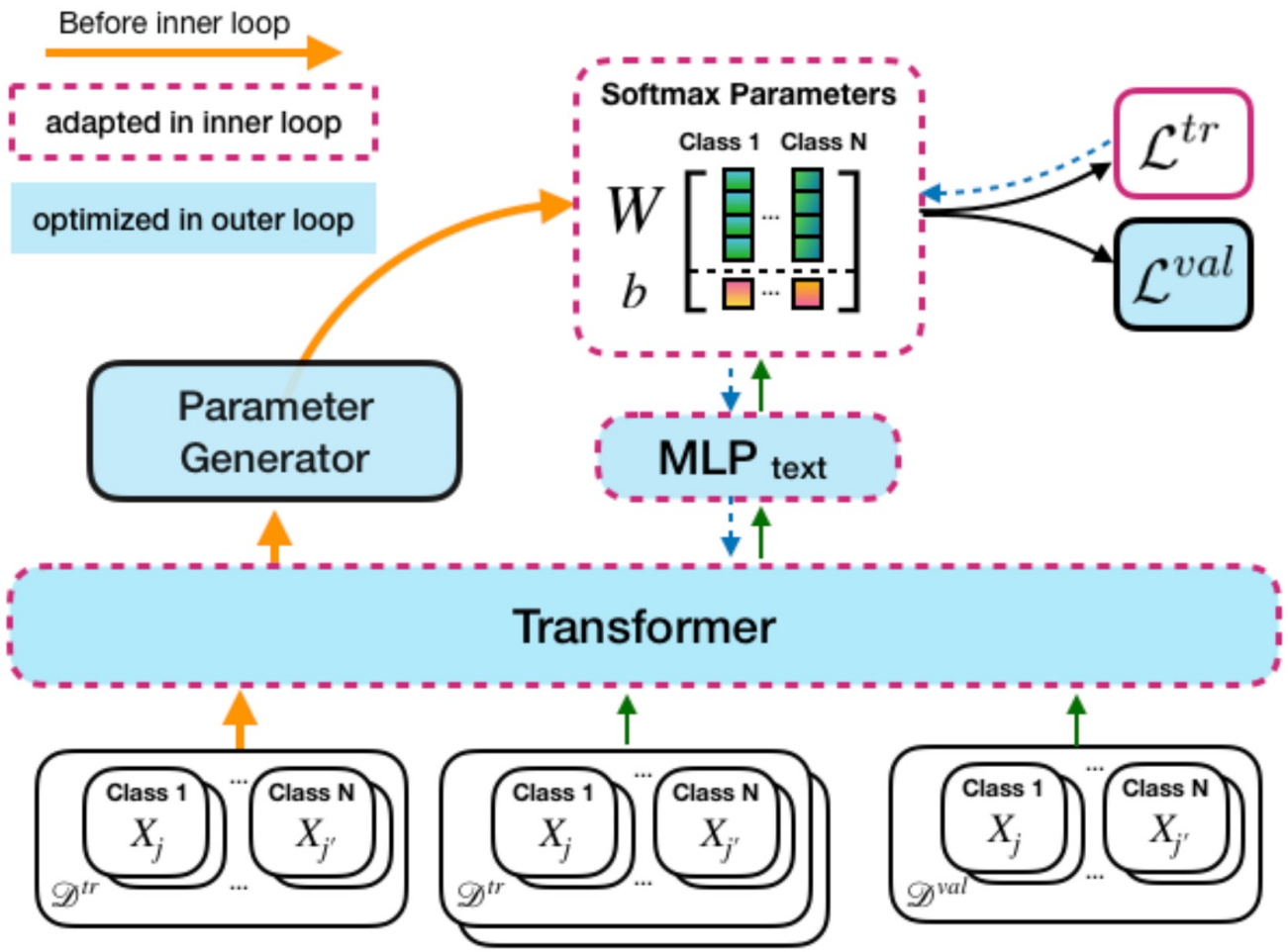
- To enables optimization-based meta-learning across tasks with different number of classes （不同的任务class数不同）
  - MAML: 只考虑了固定的class数
  - Matric-based: 若labeled examples增加，不能适应更大的数据

Method: LEOPARD

$$w_i^n, b_i^n = \frac{1}{|C_i^n|} \sum_{x_j \in C_i^n} g_\psi(f_\theta(\mathbf{x}_j))$$

$$\mathbf{W}_i = [w_i^1; \ldots; w_i^{N_i}] \quad \mathbf{b}_i = [b_i^1; \ldots; b_i^{N_i}]$$

$$p(y|\mathbf{x}^*) = \text{softmax}\left\{\mathbf{W}_i h_\phi(f_\theta(\mathbf{x}^*)) + \mathbf{b}_i\right\}$$



Method: LEOPARD

Algorithm 1 LEOPARD

Require: set of M training tasks and losses $\{(T_1, L_1), \ldots, (T_M, L_M)\}$, model parameters $\Theta = \{\theta, \psi, \alpha\}$, hyper-parameters ν, G, β

Initialize θ with pre-trained BERT-base;

while not converged **do**
  # sample batch of tasks
  for all $T_i \in T$ **do**
    $\mathcal{D}_i^{tr} \sim T_i$ *# sample a batch of train data*
    $C_i^n \leftarrow \{x_j | y_j = n\}$ *# partition data according to class labels*
    $w_i^n, b_i^n \leftarrow \frac{1}{|C_i^n|} \sum_{x_j \in C_i^n} g_\psi(f_\theta(x_j))$ *# generate softmax parameters*
    $\mathbf{W}_i \leftarrow [w_i^1; \ldots; w_i^{N_i}]$; $\mathbf{b}_i \leftarrow [b_i^1; \ldots; b_i^{N_i}]$
    $\Phi_i^{(0)} \leftarrow \theta_{>\nu} \cup \{\phi, \mathbf{W}_i, \mathbf{b}_i\}$ *# task-specific parameters*
    for $s := 0 \ldots G-1$ **do**
      $\mathcal{D}_i^{tr} \sim T_i$ *# sample a batch of train data*
      $\Phi_i^{(s+1)} \leftarrow \Phi_i^{(s)} - \alpha_s \nabla_\Phi \mathcal{L}_i(\{\Theta, \Phi_i\}, \mathcal{D}_i^{tr})$ *# adapt task-specific parameters*
    end for
    $\mathcal{D}_i^{val} \sim T_i$ *# sample a batch of validation data*
    $g_i \leftarrow \nabla_\Theta \mathcal{L}_i(\{\Theta, \Phi_i^{(G)}\}, \mathcal{D}_i^{val})$ *# gradient of task-agnostic parameters on validation*
  end for
  $\Theta \leftarrow \Theta - \beta \cdot \sum_i g_i$ *# optimize task-agnostic parameters*
end while

Task-agnostic param:

$$\Theta = \theta_{\leq \nu} \cup \{\psi\}$$

Task-specific param:

$$\Phi_i = \theta_{>\nu} \cup \{\phi, \mathbf{W}_i, \mathbf{b}_i\}$$

Experiments: setting

- Training Tasks
  - GLUE benchmark tasks:
    MNLI (m/mm), SST-2, QNLI, QQP, MRPC, RTE, and the SNLI

- Evaluation
  - 17 target NLP tasks
    - Generalization Beyond Training Tasks (new tasks unseen at training)
    - Domain transfer tasks (new domains of tasks seen at training time)

  - Evaluated on the entire test set for the task.

Experiments: Generalization Beyond Training Tasks

	N	k	$\text{BERT}_{\text{base}}$	$\text{MT-BERT}_{\text{softmax}}$	MT-BERT	Proto-BERT	LEOPARD
Entity Typing							
CoNLL	4	4	50.44 ± 08.57	52.28 ± 4.06	**55.63** ± 4.99	32.23 ± 5.10	54.16 ± 6.32
		8	50.06 ± 11.30	65.34 ± 7.12	58.32 ± 3.77	34.49 ± 5.15	**67.38** ± 4.33
		16	74.47 ± 03.10	71.67 ± 3.03	71.29 ± 3.30	33.75 ± 6.05	**76.37** ± 3.08
MITR	8	4	49.37 ± 4.28	45.52 ± 5.90	**50.49** ± 4.40	17.36 ± 2.75	49.84 ± 3.31
		8	49.38 ± 7.76	58.19 ± 2.65	58.01 ± 3.54	18.70 ± 2.38	**62.99** ± 3.28
		16	69.24 ± 3.68	66.09 ± 2.24	66.16 ± 3.46	16.41 ± 1.87	**70.44** ± 2.89
Text Classification							
Airline	3	4	42.76 ± 13.50	43.73 ± 7.86	46.29 ± 12.26	40.27 ± 8.19	**54.95** ± 11.81
		8	38.00 ± 17.06	52.39 ± 3.97	49.81 ± 10.86	51.16 ± 7.60	**61.44** ± 03.90
		16	58.01 ± 08.23	58.79 ± 2.97	57.25 ± 09.90	48.73 ± 6.79	**62.15** ± 05.56
Disaster	2	4	**55.73** ± 10.29	52.87 ± 6.16	50.61 ± 8.33	50.87 ± 1.12	51.45 ± 4.25
		8	**56.31** ± 09.57	56.08 ± 7.48	54.93 ± 7.88	51.30 ± 2.30	55.96 ± 3.58
		16	64.52 ± 08.93	**65.83** ± 4.19	60.70 ± 6.05	52.76 ± 2.92	61.32 ± 2.83
Emotion	13	4	09.20 ± 3.22	09.41 ± 2.10	09.84 ± 2.14	09.18 ± 3.14	**11.71** ± 2.16
		8	08.21 ± 2.12	11.61 ± 2.34	11.21 ± 2.11	11.18 ± 2.95	**12.90** ± 1.63
		16	13.43 ± 2.51	**13.82** ± 2.02	12.75 ± 2.04	12.32 ± 3.73	13.38 ± 2.20
Political Bias	2	4	54.57 ± 5.02	54.32 ± 3.90	54.66 ± 3.74	56.33 ± 4.37	**60.49** ± 6.66
		8	56.15 ± 3.75	57.36 ± 4.32	54.79 ± 4.19	58.87 ± 3.79	**61.74** ± 6.73
		16	60.96 ± 4.25	59.24 ± 4.25	60.30 ± 3.26	57.01 ± 4.44	**65.08** ± 2.14

	N	k	$\text{BERT}_{\text{base}}$	$\text{MT-BERT}_{\text{softmax}}$	MT-BERT	Proto-BERT	LEOPARD
Political Audience	2	4	51.89 ± 1.72	51.50 ± 2.72	51.53 ± 1.80	51.47 ± 3.68	**52.60** ± 3.51
		8	52.80 ± 2.72	53.53 ± 2.26	**54.34** ± 2.88	51.83 ± 3.77	54.31 ± 3.95
		16	**58.45** ± 4.98	56.37 ± 2.19	55.14 ± 4.57	53.53 ± 3.25	57.71 ± 3.52
Political Message	9	4	15.64 ± 2.73	13.71 ± 1.10	14.49 ± 1.75	14.22 ± 1.25	**15.69** ± 1.57
		8	13.38 ± 1.74	14.33 ± 1.32	15.24 ± 2.81	15.67 ± 1.96	**18.02** ± 2.32
		16	**20.67** ± 3.89	18.11 ± 1.48	19.20 ± 2.20	16.49 ± 1.96	18.07 ± 2.41
Rating Books	3	4	39.42 ± 07.22	44.82 ± 9.00	38.97 ± 13.27	48.44 ± 7.43	**54.92** ± 6.18
		8	39.55 ± 10.01	51.14 ± 6.78	46.77 ± 14.12	52.13 ± 4.79	**59.16** ± 4.13
		16	43.08 ± 11.78	54.61 ± 6.79	51.68 ± 11.27	57.28 ± 4.57	**61.02** ± 4.19
Rating DVD	3	4	32.22 ± 08.72	45.94 ± 7.48	41.23 ± 10.98	47.73 ± 6.20	**49.76** ± 9.80
		8	36.35 ± 12.50	46.23 ± 6.03	45.24 ± 9.76	47.11 ± 4.00	**53.28** ± 4.66
		16	42.79 ± 10.18	49.23 ± 6.68	45.19 ± 11.56	48.39 ± 3.74	**53.52** ± 4.77
Rating Electronics	3	4	39.27 ± 10.15	39.89 ± 5.83	41.20 ± 10.69	37.40 ± 3.72	**51.71** ± 7.20
		8	28.74 ± 08.22	46.53 ± 5.44	45.41 ± 09.49	43.64 ± 7.31	**54.78** ± 6.48
		16	45.48 ± 06.13	48.71 ± 6.16	47.29 ± 10.55	44.83 ± 5.96	**58.69** ± 2.41
Rating Kitchen	3	4	34.76 ± 11.20	40.41 ± 5.33	36.77 ± 10.62	44.72 ± 9.13	**50.21** ± 09.63
		8	34.49 ± 08.72	48.35 ± 7.87	47.98 ± 09.73	46.03 ± 8.57	**53.72** ± 10.31
		16	47.94 ± 08.28	52.94 ± 7.14	53.79 ± 09.47	49.85 ± 9.31	**57.00** ± 08.69
Overall Average		4	38.13	40.13	40.10	36.29	**45.99**
		8	36.99	45.89	44.25	39.15	**50.86**
		16	48.55	49.93	49.07	39.85	**55.50**

Experiments: Few-Shot Domain Transfer

	k	$\text{BERT}_{\text{base}}$	$\text{MT-BERT}_{\text{softmax}}$	MT-BERT	$\text{MT-BERT}_{\text{reuse}}$	Proto-BERT	LEOPARD
Natural Language Inference							
Scitail	4	58.53 ± 09.74	74.35 ± 5.86	63.97 ± 14.36	**76.65 ± 2.45**	76.27 ± 4.26	69.50 ± 9.56
	8	57.93 ± 10.70	**79.11 ± 3.11**	68.24 ± 10.33	76.86 ± 2.09	78.27 ± 0.98	75.00 ± 2.42
	16	65.66 ± 06.82	**79.60 ± 2.31**	75.35 ± 04.80	79.53 ± 2.17	78.59 ± 0.48	77.03 ± 1.82
Amazon Review Sentiment Classification							
Books	4	54.81 ± 3.75	68.69 ± 5.21	64.93 ± 8.65	74.79 ± 6.91	73.15 ± 5.85	**82.54 ± 1.33**
	8	53.54 ± 5.17	74.86 ± 2.17	67.38 ± 9.78	78.21 ± 3.49	75.46 ± 6.87	**83.03 ± 1.28**
	16	65.56 ± 4.12	74.88 ± 4.34	69.65 ± 8.94	78.87 ± 3.32	77.26 ± 3.27	**83.33 ± 0.79**
Kitchen	4	56.93 ± 7.10	63.07 ± 7.80	60.53 ± 9.25	75.40 ± 6.27	62.71 ± 9.53	**78.35 ± 18.36**
	8	57.13 ± 6.60	68.38 ± 4.47	69.66 ± 8.05	75.13 ± 7.22	70.19 ± 6.42	**84.88 ± 01.12**
	16	68.88 ± 3.39	75.17 ± 4.57	77.37 ± 6.74	80.88 ± 1.60	71.83 ± 5.94	**85.27 ± 01.31**

Ablation Study:

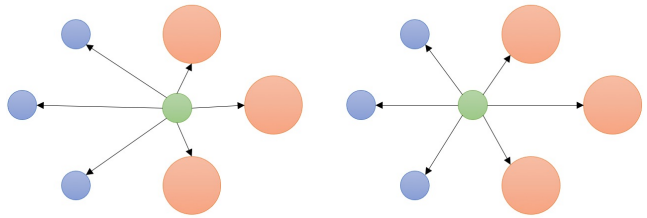
- Importance of softmax parameters
- Parameter efficiency

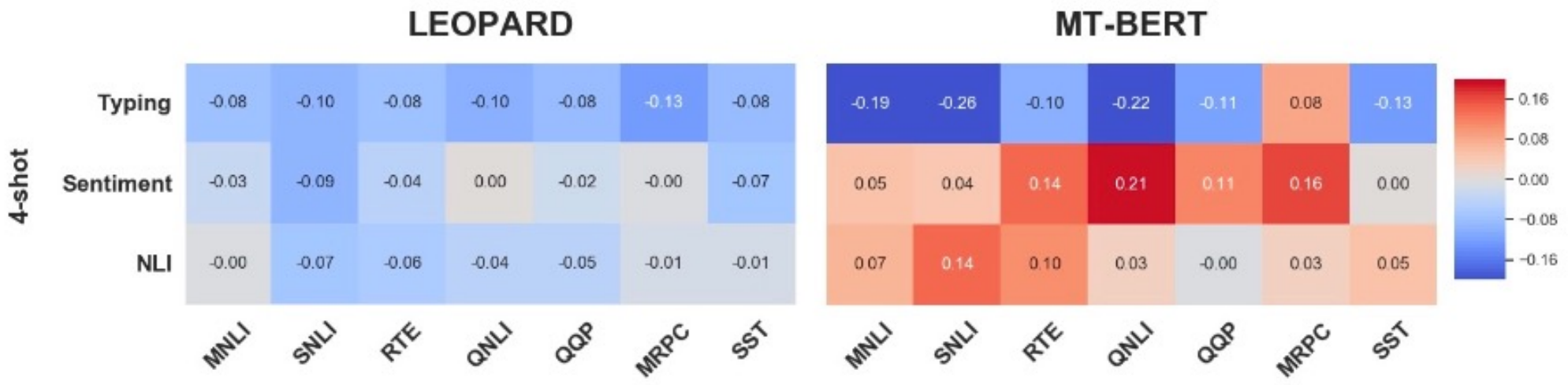
k	Model	Entity Typing	Sentiment Classification	NLI
16	LEOPARD $_{10}$	37.62 $\pm$ 7.37	58.10 $\pm$ 5.40	78.53 $\pm$ 1.55
	LEOPARD $_{5}$	62.49 $\pm$ 4.23	71.50 $\pm$ 5.93	73.27 $\pm$ 2.63
	LEOPARD	69.00 $\pm$ 4.76	76.65 $\pm$ 2.47	76.10 $\pm$ 2.21
	LEOPARD-ZERO	44.79 $\pm$ 9.34	74.45 $\pm$ 3.34	74.36 $\pm$ 6.67

MAML →

Table 3: Ablations: LEOPARD$_\nu$ does not adapt layers $0-\nu$ (inclusive) in the inner loop (and fine-tuning), while LEOPARD adapts all parameters. Note that the outer loop still optimizes all parameters. For new tasks (like entity typing) adapting all parameters is better while for tasks seen at training time (like NLI) adapting fewer parameters is better. LEOPARD-ZERO is a model trained without the softmax-generator and a zero initialized softmax classifier, which shows the importance of softmax generator in LEOPARD.

Ablation Study:
Importance of training tasks





- 泛化能力，受training tasks影响
- Initial point经过很少的training samples后可以迅速适应新task

Reference

- http://speech.ee.ntu.edu.tw/~tlkagk/courses_ML20.html
- Bansal, T., Jha, R., & McCallum, A. (2019). Learning to few-shot learn across diverse natural language classification tasks. *arXiv preprint arXiv:1911.03863*.
- Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. In EMNLP-IJCNLP, pages 1192–1197
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In ICML, pages 1126–1135.

Thanks!

Q&A