Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 522–534,
November 16–20, 2020.

Self-Supervised Meta-Learning for Few-Shot Natural Language Classification Tasks

Trapit Bansal ◇* and **Rishikesh Jha**† and **Tsendsuren Munkhdalai**‡ and **Andrew McCallum**◇
◇University of Massachusetts, Amherst
†Code for Science and Society
‡Microsoft Research, Montréal, Canada

目 录
Contents

问题设定

自监督的预训练语言模型，可以提供一个好的模型参数初始化点，使得到了下游任务微调时，可以有不错的泛化性和效果。

chanllenge 1

微调的数据利用比较低，即在Few-Shot的场景下，利用少量的样本进行微调效果得到不充分的体现。所以这篇文章关注点就在于，如果提高预训练的语言模型在Few-Shot场景下进行分类任务时的泛化能力。

soulution 1

提高模型的泛化能力，可以采用Meta-Learning方式来解决。

challenge 2

任务构建，需要大量的训练任务，以提高泛化性。且对于不同的任务，需要构建不同的模型框架。

soulution 2

利用类似cloze-style的方式，通过Self-Supervised从数据本身来获得标签，构建受监督的元学习任务。通过学习特定任务的Support set来生成任务特定的参数，以支持训练能够适应不同类任务的元学习模型[1]。

[1] Trapit Bansal, Rishikesh Jha, and Andrew McCallum.2019. Learning to few-shot learn across diverse natural language classification tasks.

原因分析—challenge 1

虽然自监督的预训练是有效的，但是它对数据的利用率会比较低，而且微调时也需要大量的数据才能在目标任务上有好的效果[1][2]。

这就是Few-Shot Learning问题，模型只给出几个新任务的例子，并期望在该任务中表现良好。
本文重点关注Few-Shot的问题，并开发了能够更好地对新任务进行Few-Shot 泛化的模型。

1. 此外，对预训练的模型进行微调通常会引入新的随机参数，如Softmax层和重要的超参数（如learning-rate），这些参数很难从少数示例中有效地评估。
2. 大规模的预训练会遭受训练测试不匹配的困扰，因为模型没有经过优化就去学习一个初始点，而该初始点经过少量样本微调后性能较低。

因此，本文提出要消除这种训练测试不匹配，并通过联合学习模型的初始点和超参数，从而允许数据高效微调，即作为一个元学习问题[1][3]。

[2] Dani Yogatama, Cyprien de Masson d’Autume, Jerome Connor, Tom´as Kocisk´y, Mike Chrzanowski, Ling peng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence.
[3] Sebastian Thrun and Lorien Pratt. 2012. Learning to learn. Springer Science & Business Media.

原因分析—Solution 1

1. 大规模的预训练会遭受训练测试失配的困扰，因为模型没有经过优化就去学习一个初始点，而该初始点经过少量样本微调后性能较低。

MAML

$\hat{\theta}^n$: model learned from task n

$\hat{\theta}^n$ depends on ϕ

Loss Function:

$$L(\phi) = \sum_{n=1}^{N} l^n(\hat{\theta}^n)$$

$l^n(\hat{\theta}^n)$: loss of task n on the testing set of task n

How to minimize $L(\phi)$? Gradient De

$$\theta_i' \leftarrow \phi - \alpha \nabla_\phi \mathcal{L}_i(\mathcal{D}^{tr}, \phi)$$

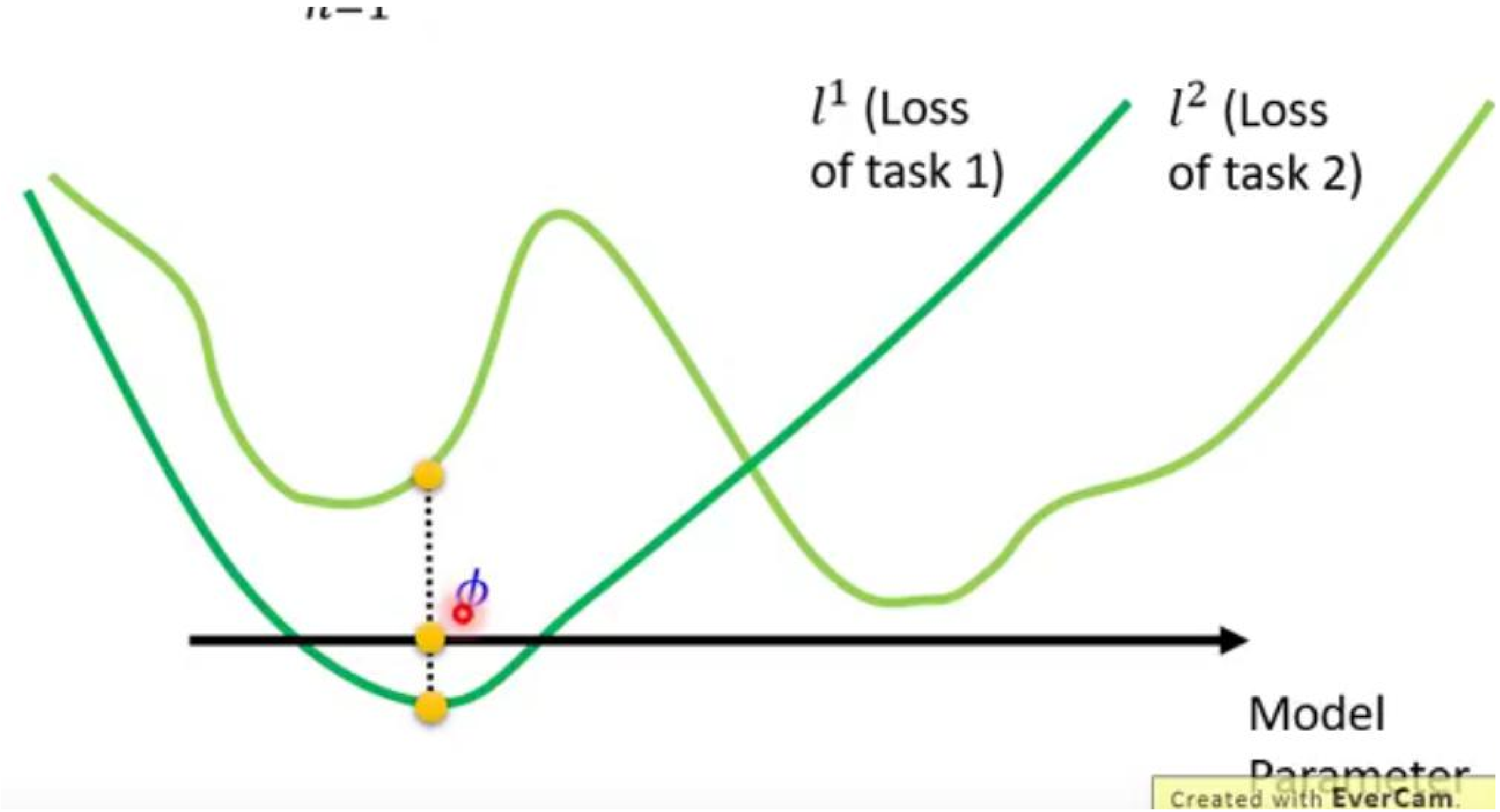
Model Pre-training

Widely used in transfer learning

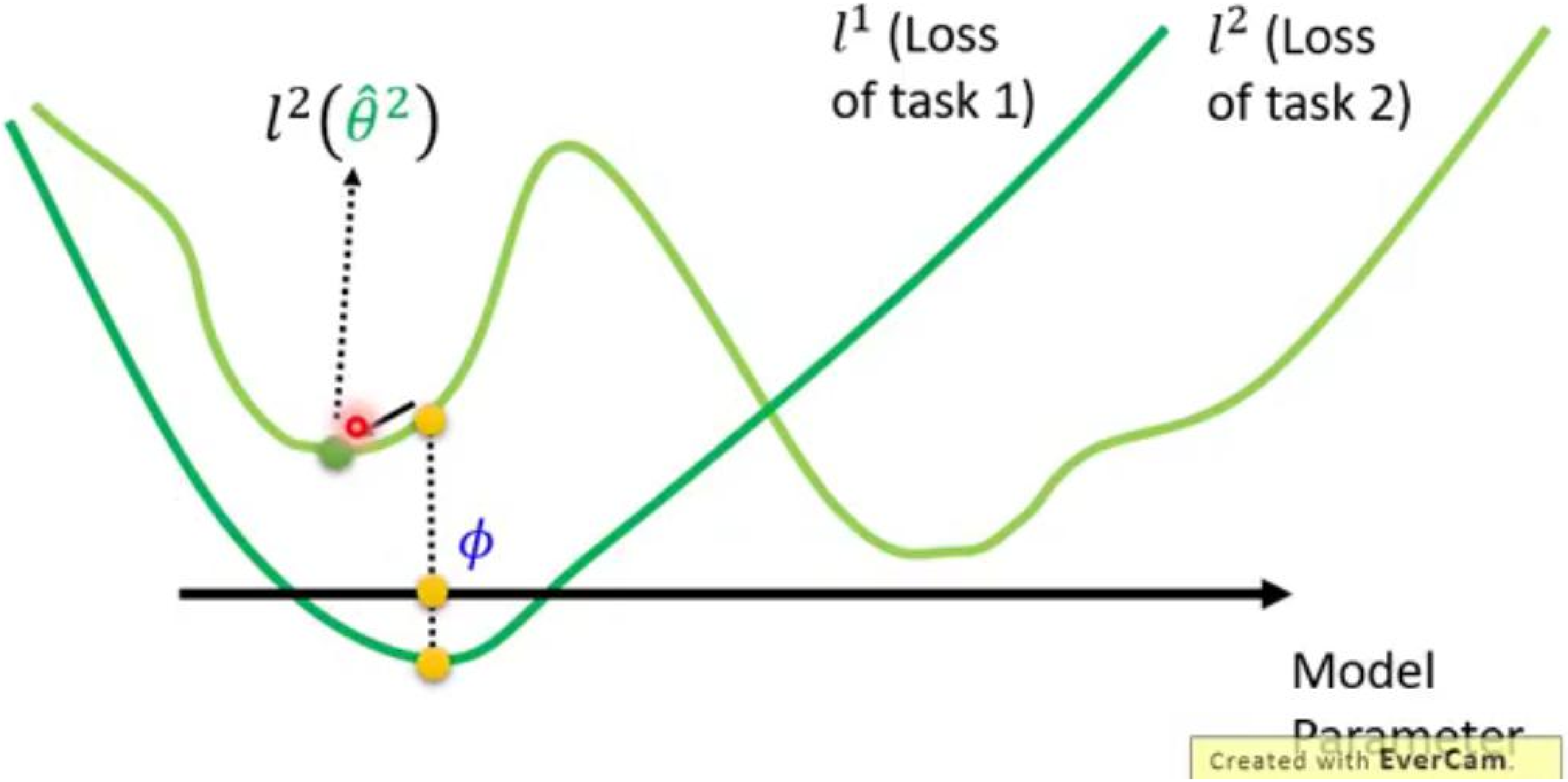
Loss Function:

$$L(\phi) = \sum_{n=1}^{N} l^n(\phi)$$

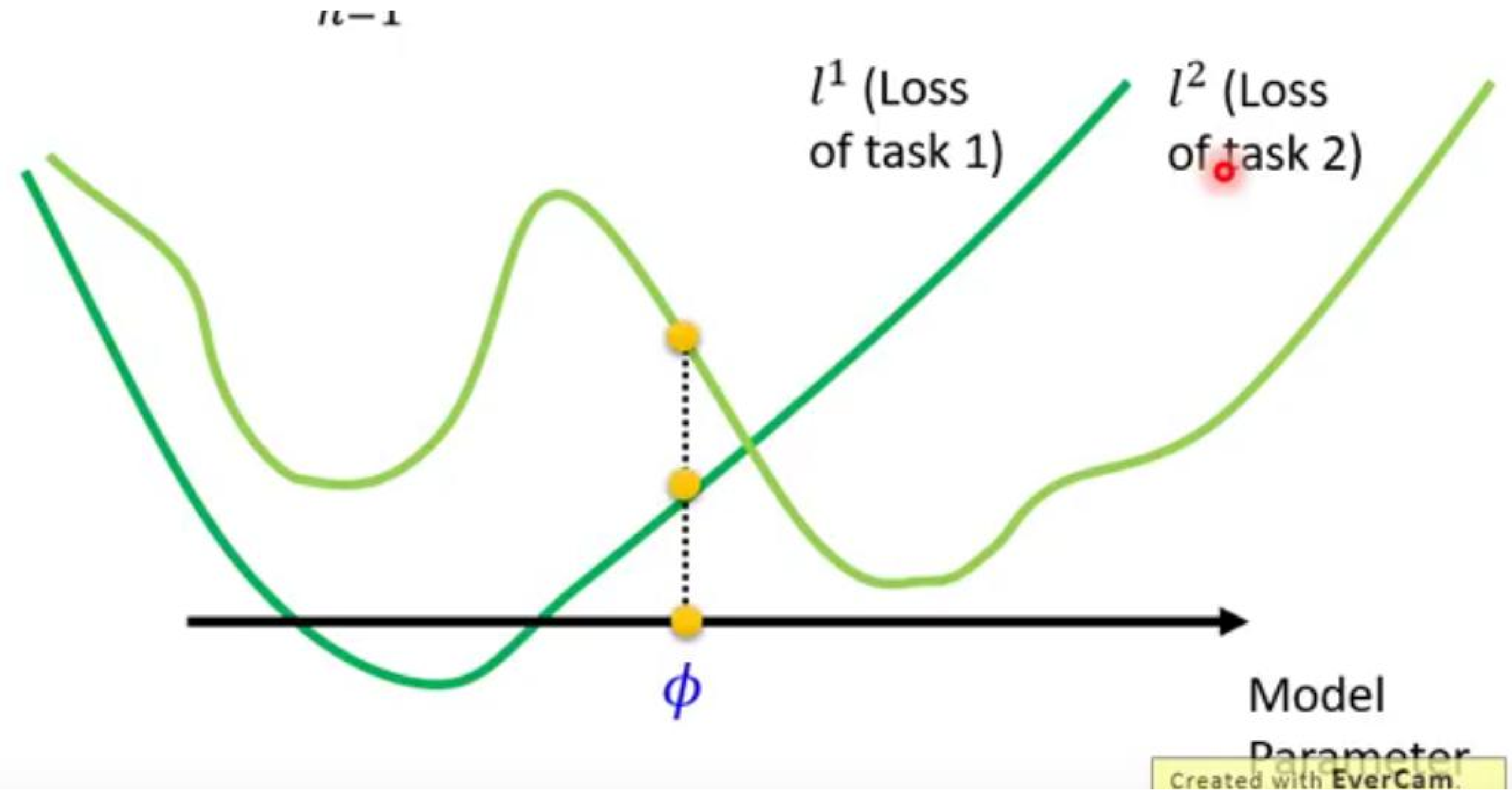
Model Pre-training



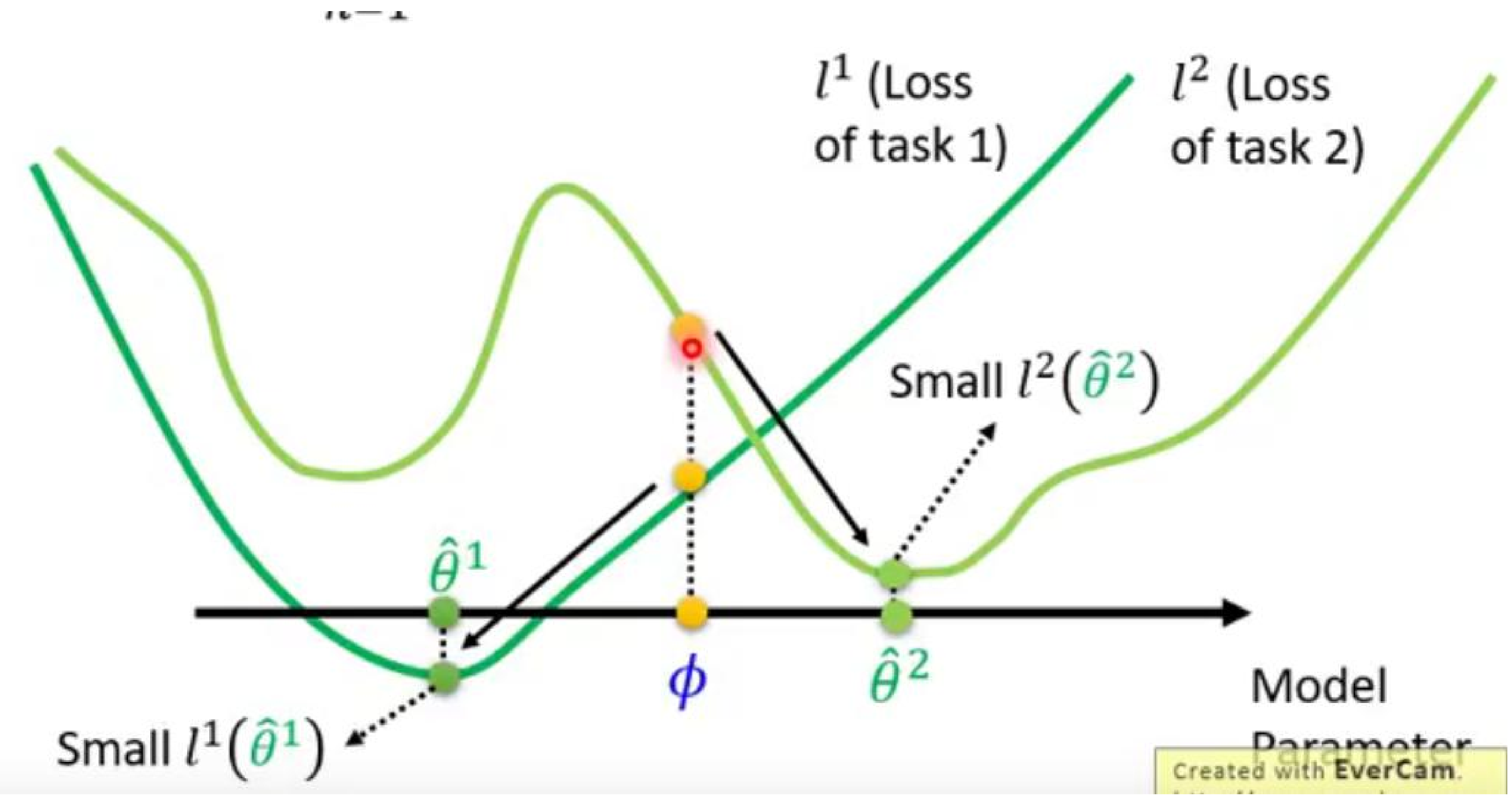
Pre-training Model



MAML


l^1 (Loss of task 1)
l^2 (Loss of task 2)
ϕ
Model Parameter
Created with EverCam.

MAML


l^1 (Loss
of task 1)
l^2 (Loss
of task 2)
Small $l^2(\hat{\theta}^2)$
$\hat{\theta}^1$
ϕ
$\hat{\theta}^2$
Model
Parameter
Small $l^1(\hat{\theta}^1)$
Created with EverCam.

原因分析—challenge 2

Meta-Learning的任务构建，需要大量的训练任务，以提高泛化性。
元学习在相同的任务下，在新的标签下泛化性不错，但是对于不同的任务，需要构建不同的模型框架，使得在不同任务下泛化性较差[4]。

1. 对于不同的任务，需要构建不同的模型框架。
2. Meta Over-Fitting，对于任务的分布过拟合了[4]。

[4] Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. 2020. Meta learning without memorization. In International Conference on Learning Representations.

原因分析—Solution 2

1. Meta-Learning的任务构建，需要大量的训练任务，以提高泛化性。
Subset Masked Language Modeling Tasks (SMLMT)

Subset: {Democratic, Capital}

Support set

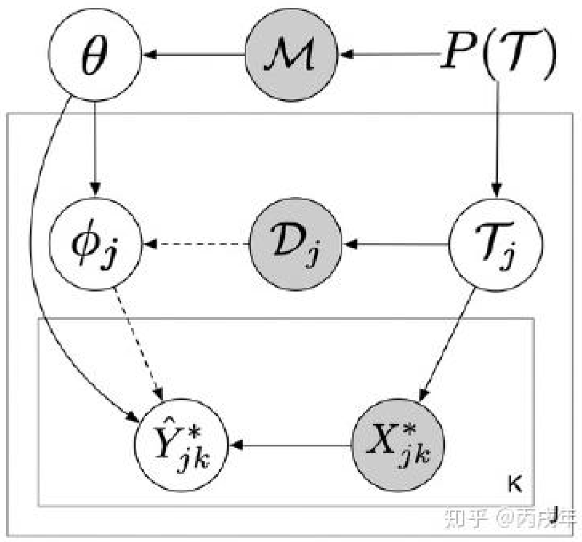
Sentence	Class
A member of the [m] Party, he was the first African American to be elected to the presidency.	1
The [m] Party is one of the two major contemporary political parties in the United States, along with its rival, the Republican Party.	1
Honolulu is the [m] and largest city of the U.S. state of Hawaii.	2
Washington, D.C., formally the District of Columbia and commonly referred to as Washington or D.C., is the [m] of the United States.	2

Query: New Delhi is an urban district of Delhi which serves as the [m] of India

Correct Prediction: 2

原因分析—Solution 2

2. Meta Over-Fitting，对于任务的分布过拟合了[4][5]。

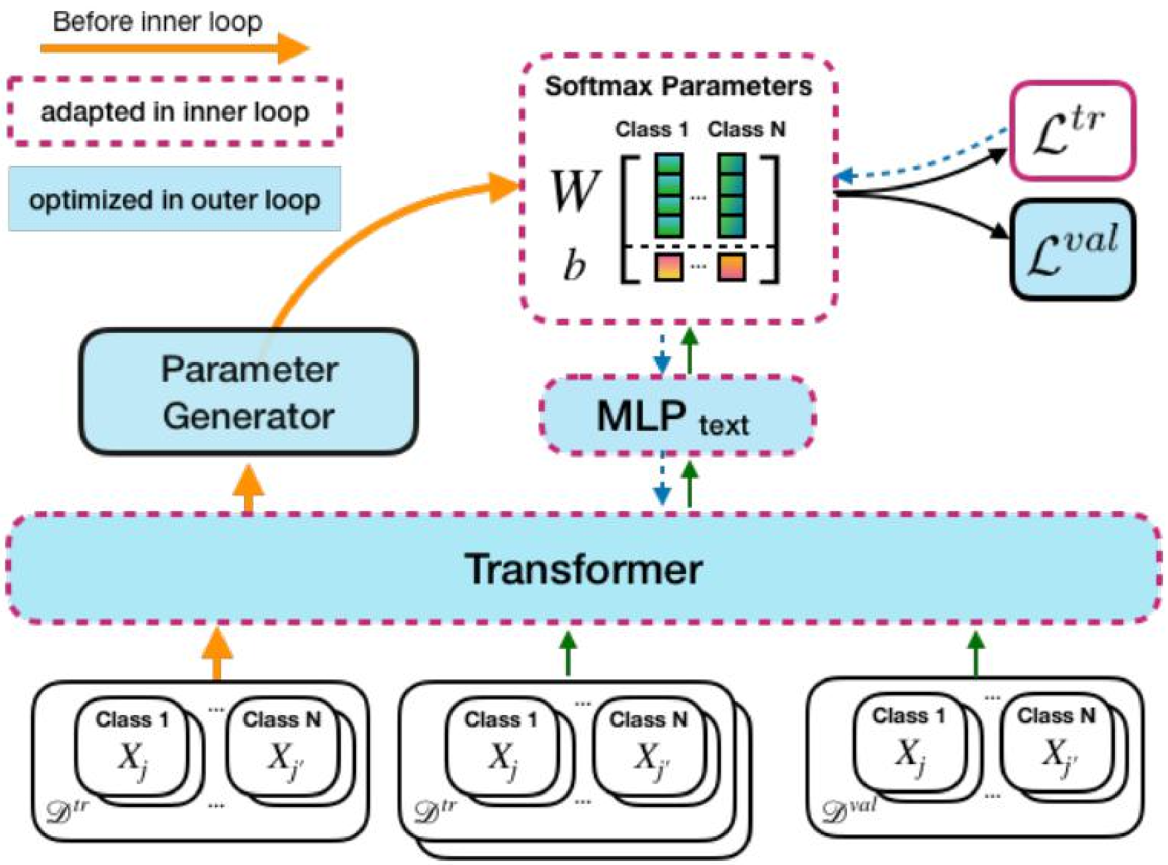


$$-\frac{1}{N}\sum_i \mathbb{E}_{q(\theta|M)q(\phi|D_i^{train},\theta)}\left[\frac{1}{K}\sum_{(x^*,y^*)\in D_i^{test}} logq(\hat{y}^* = y^*|x^*,\theta,\phi)\right],$$

[4] Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. 2020. Meta learning without memorization. In International Conference on Learning Representations.
[5] Luke Metz, Niru Maheswaranathan, Brian Cheung, and Jascha Sohl-Dickstein. 2019. Learning unsupervised learning rules. In International Conference on Learning Representations.

原因分析—Solution 2

3. 对于不同的任务，需要构建不同的模型框架[6]。



$$w_t^n, b_t^n = g_\psi(\{f_\pi(X_{1n}), \ldots, f_\pi(X_{kn})\}) \quad (3)$$

$$p(y|X) = softmax\{\mathbf{W}_t\, h_\phi(f_\pi(X)) + \mathbf{b}_t\} \quad (4)$$

where $\mathbf{W}_t = [w_t^1; \ldots; w_t^N] \in \mathcal{R}^{N\times d}$, $\mathbf{b}_t = [b_t^1; \ldots; b_t^N] \in \mathcal{R}^d$ are the concatenation of the per-class vectors in (3), and h_ϕ is a MLP with parameters ϕ and output dimension d.

[6] Trapit Bansal, Rishikesh Jha, and Andrew McCallum.2019. Learning to few-shot learn across diverse natural language classification tasks.

本文工作

1. 提出了Subset Masked Language Modeling Tasks (SMLMT)，通过自监督的方式创建元学习的任务。并且通过SMLMT创建的任务分布取得了SOTA的结果。

2. 证明了自监督SMLMT也可以与受监督任务的数据相结合，以实现更好的特征学习，同时通过使用SMLMT避免Meta Over-Fitting以保证模型泛化的泛化性。

3. 研究了参数数对Few-Shot Learning的影响，发现规模更大的预训练或元学习模型比小模型泛化性更好，即使对于小模型，元学习也能获得不错的效果。

目 录
Contents

Hybrid SMLMT

来自SMLMT的任务也可以与受监督的任务相结合，以鼓励更好的特征学习[6]，并增加元学习任务的多样性。

使用采样比$\lambda \in (0,1)$，在每个episode中以 λ 的概率选择SMLMT任务或以（$1-\lambda$）的概率选择监督任务。SMLMT与受监督任务的联合使用可以改善Meta Over-Fitting。

[6]Rich Caruana. 1997. Multitask learning. Machine learning, 28(1):41–75.

元训练算法流程

Algorithm 1 Meta-Training

Require: SMLMT task distribution $\mathcal{T}$ and supervised tasks $\mathcal{S}$, model parameters $\{\pi^w, \pi, \phi, \psi, \alpha\}$, adaptation steps G, learning-rate β, sampling ratio λ

Initialize θ with pre-trained BERT-base;

while not converged **do**
  for task_batchsize times **do**
    $t \sim Bernoulli(\lambda)$
    $T \sim t \cdot \mathcal{T} + (1-t) \cdot \mathcal{S}$
    $\mathcal{D}^{tr} = \{(x_j, y_j)\} \sim T$
    $C^n \leftarrow \{x_j | y_j = n\}; \quad N \leftarrow |C^n|$
    $w^n, b^n \leftarrow \frac{1}{|C^n|} \sum_{x_j \in C^n} g_\psi(f_\pi(\mathcal{D}^{tr}))$
    $\mathbf{W} \leftarrow [w^1; \ldots; w^N]; \quad \mathbf{b} \leftarrow [b^1; \ldots; b^N]$
    $\theta \leftarrow \{\pi, \phi, \mathbf{W}, \mathbf{b}\}; \quad \theta^{(0)} \leftarrow \theta$
    $\Theta \leftarrow \{\pi^w, \pi, \psi, \alpha\}$
    $\mathcal{D}^{val} \sim T$
    $q_T \leftarrow 0$
    for $s := 0 \ldots G-1$ **do**
      $\mathcal{D}_s^{tr} \sim T$
      $\theta^{(s+1)} \leftarrow \theta^{(s)} - \alpha \nabla_\theta \mathcal{L}_T(\{\Theta, \theta^{(s)}\}, \mathcal{D}_s^{tr})$
      $q_T \leftarrow q_T + \nabla_\Theta \mathcal{L}_T(\{\Theta, \theta^{(s+1)}\}, \mathcal{D}^{val})$
    end for
  end for
  $\Theta \leftarrow \Theta - \beta \cdot \sum_T \frac{q_T}{G}$
end while

π^ω 是warp layers的参数

π 是Transformer除π^ω外的参数

ϕ 是MLP的参数

ψ 是类别编码器参数

α 是学习率

the parameters are meta-trained using the MAML algorithm (Finn et al., 2017). Concretely, set $\theta := \{\pi, \phi, \mathbf{W}_t, \mathbf{b}_t\}$ for the task-specific inner loop gradient updates in (1) and set $\Theta := \{\pi, \psi, \alpha\}$ for the outer-loop updates in (2). Note that we do multiple steps of gradient descent in the inner loop.

目 录
Contents

数据集统计

Dataset	Labels	Train	Validation	Test
CoLA	2	8551	1042	—
MRPC	2	3669	409	—
QNLI	2	104744	5464	—
QQP	2	363847	40431	—
RTE	2	2491	278	—
SNLI	3	549368	9843	—
SST-2	2	67350	873	—
MNLI (m/mm)	3	392703	19649	—
Scitail	2	23,596	1,304	2,126
Amazon Sentiment Domains	2	800	200	1000
Airline	3	7320	—	7320
Disaster	2	4887	—	4887
Political Bias	2	2500	—	2500
Political Audience	2	2500	—	2500
Political Message	9	2500	—	2500
Emotion	13	20000	—	20000
CoNLL	4	23499	5942	5648
MIT-Restaurant	8	12474	—	2591

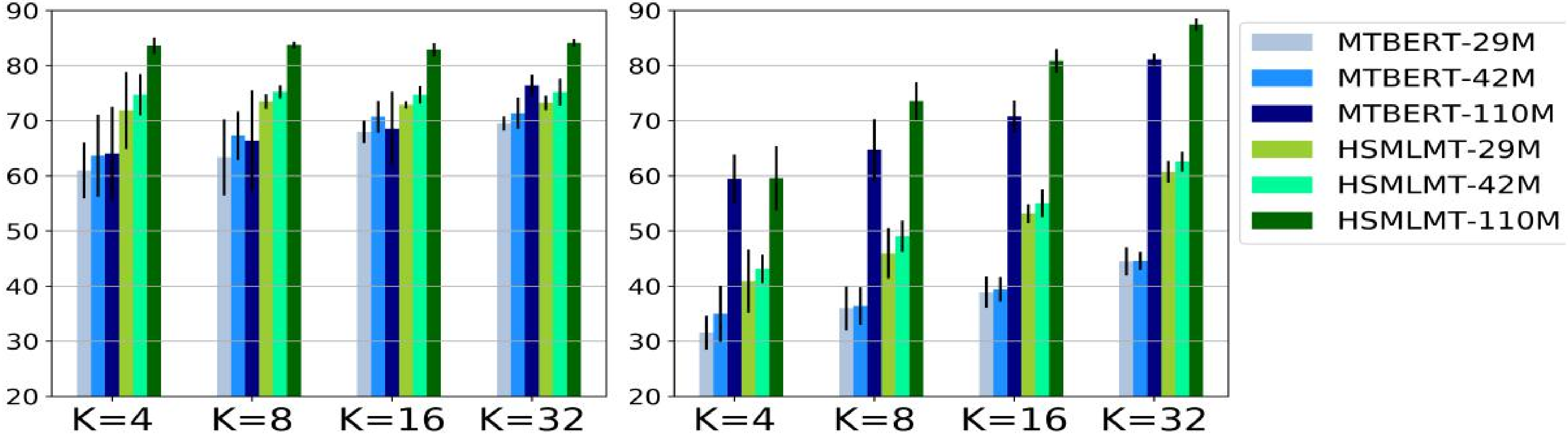
k-shot accuracy on novel tasks not seen in training.

Task	N	k	BERT	**SMLMT**	MT-BERT$_{\text{softmax}}$	MT-BERT	LEOPARD	**Hybrid-SMLMT**
CoNLL	4	4	50.44 ± 08.57	46.81 ± 4.77	52.28 ± 4.06	55.63 ± 4.99	54.16 ± 6.32	**57.60** ± 7.11
		8	50.06 ± 11.30	61.72 ± 3.11	65.34 ± 7.12	58.32 ± 3.77	67.38 ± 4.33	**70.20** ± 3.00
		16	74.47 ± 03.10	75.82 ± 4.04	71.67 ± 3.03	71.29 ± 3.30	76.37 ± 3.08	**80.61** ± 2.77
		32	83.27 ± 02.14	84.01 ± 1.73	73.09 ± 2.42	79.94 ± 2.45	83.61 ± 2.40	**85.51** ± 1.73
MITR	8	4	49.37 ± 4.28	46.23 ± 3,90	45.52 ± 5.90	50.49 ± 4.40	49.84 ± 3.31	**52.29** ± 4.32
		8	49.38 ± 7.76	61.15 ± 1.91	58.19 ± 2.65	58.01 ± 3.54	62.99 ± 3.28	**65.21** ± 2.32
		16	69.24 ± 3.68	69.22 ± 2.78	66.09 ± 2.24	66.16 ± 3.46	70.44 ± 2.89	**73.37** ± 1.88
		32	78.81 ± 1.95	78.82 ± 1.30	69.35 ± 0.98	76.39 ± 1.17	78.37 ± 1.97	**79.96** ± 1.48
Airline	3	4	42.76 ± 13.50	42.83 ± 6.12	43.73 ± 7.86	46.29 ± 12.26	54.95 ± 11.81	**56.46** ± 10.67
		8	38.00 ± 17.06	51.48 ± 7.35	52.39 ± 3.97	49.81 ± 10.86	61.44 ± 03.90	**63.05** ± 8.25
		16	58.01 ± 08.23	58.42 ± 3.44	58.79 ± 2.97	57.25 ± 09.90	62.15 ± 05.56	**69.33** ± 2.24
		32	63.70 ± 4.40	65.33 ± 3.83	61.06 ± 3.89	62.49 ± 4.48	67.44 ± 01.22	**71.21** ± 3.28
Disaster	2	4	55.73 ± 10.29	**62.26** ± 9.16	52.87 ± 6.16	50.61 ± 8.33	51.45 ± 4.25	55.26 ± 8.32
		8	56.31 ± 09.57	**67.89** ± 6.83	56.08 ± 7.48	54.93 ± 7.88	55.96 ± 3.58	63.62 ± 6.84
		16	64.52 ± 08.93	**72.86** ± 1.70	65.83 ± 4.19	60.70 ± 6.05	61.32 ± 2.83	70.56 ± 2.23
		32	73.60 ± 01.78	**73.69** ± 2.32	67.13 ± 3.11	72.52 ± 2.28	63.77 ± 2.34	71.80 ± 1.85

k-shot domain transfer accuracy

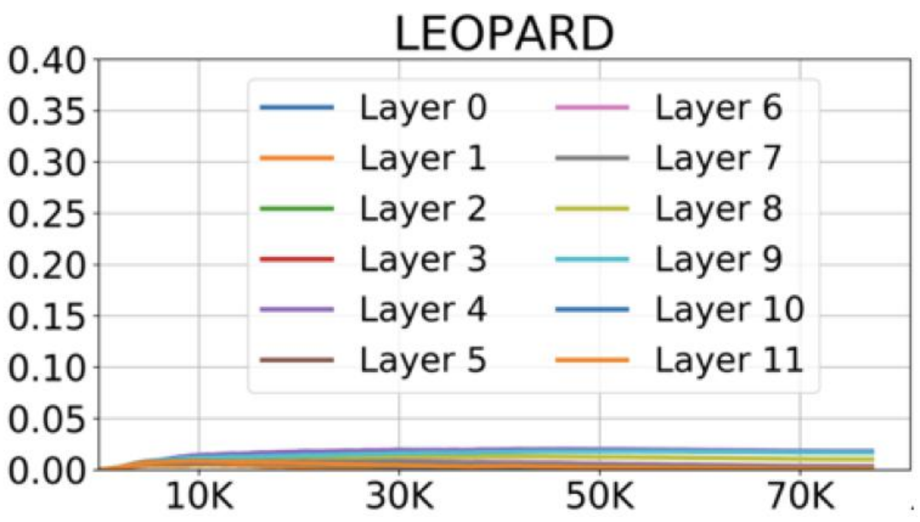
Task	k	BERT$_{\text{base}}$	**SMLMT**	MT-BERT$_{\text{softmax}}$	MT-BERT	MT-BERT$_{\text{reuse}}$	LEOPARD	**Hybrid-SMLMT**
Scitail	4	58.53 ± 09.74	50.68 ± 4.30	74.35 ± 5.86	63.97 ± 14.36	76.65 ± 2.45	69.50 ± 9.56	**76.75** ± 3.36
	8	57.93 ± 10.70	55.60 ± 2.40	**79.11** ± 3.11	68.24 ± 10.33	76.86 ± 2.09	75.00 ± 2.42	**79.10** ± 1.14
	16	65.66 ± 06.82	56.51 ± 3.78	79.60 ± 2.31	75.35 ± 04.80	79.53 ± 2.17	77.03 ± 1.82	**80.37** ± 1.44
	32	68.77 ± 6.27	62.38 ± 3.22	**82.23** ± 1.12	74.87 ± 3.62	81.77 ± 1.13	79.44 ± 1.99	**82.20** ± 1.34
Amazon Books	4	54.81 ± 3.75	55.68 ± 2.56	68.69 ± 5.21	64.93 ± 8.65	74.79 ± 6.91	82.54 ± 1.33	**84.70** ± 0.42
	8	53.54 ± 5.17	60.23 ± 5.28	74.86 ± 2.17	67.38 ± 9.78	78.21 ± 3.49	83.03 ± 1.28	**84.85** ± 0.52
	16	65.56 ± 4.12	62.92 ± 4.39	74.88 ± 4.34	69.65 ± 8.94	78.87 ± 3.32	83.33 ± 0.79	**85.13** ± 0.66
	32	73.54 ± 3.44	71.49 ± 4.74	77.51 ± 1.14	78.91 ± 1.66	82.23 ± 1.10	83.55 ± 0.74	**85.27** ± 0.36
Amazon DVD	4	54.98 ± 3.96	52.95 ± 2.51	63.68 ± 5.03	66.36 ± 7.46	71.74 ± 8.54	80.32 ± 1.02	**83.28** ± 1.85
	8	55.63 ± 4.34	54.28 ± 4.20	67.54 ± 4.06	68.37 ± 6.51	75.36 ± 4.86	80.85 ± 1.23	**83.91** ± 1.14
	16	58.69 ± 6.08	57.87 ± 2.69	70.21 ± 1.94	70.29 ± 7.40	76.20 ± 2.90	81.25 ± 1.41	**83.71** ± 1.04
	32	66.21 ± 5.41	65.09 ± 4.37	70.19 ± 2.08	73.45 ± 4.37	79.17 ± 1.71	81.54 ± 1.33	**84.15** ± 0.94

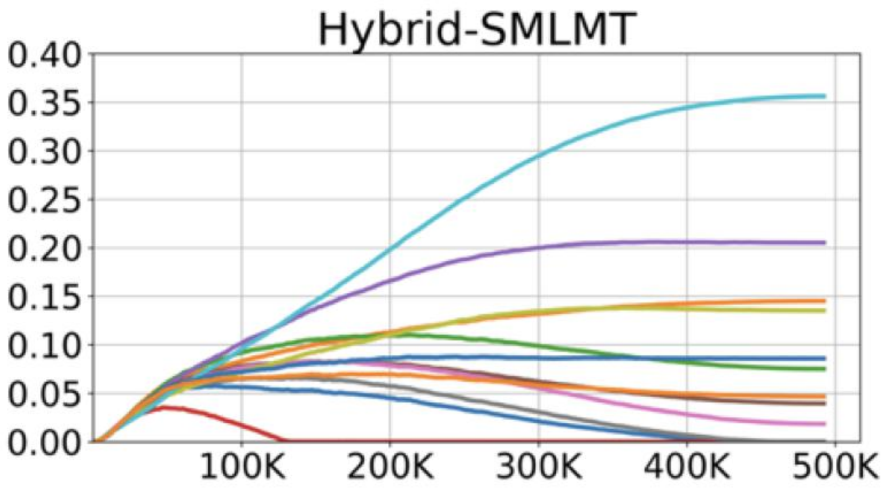
k-shot performance with number of parameters on Amazon DVD (left), and CoNLL(right)



Larger models generalize better and Hybrid-SMLMT provides accuracy gains for all parameter sizes

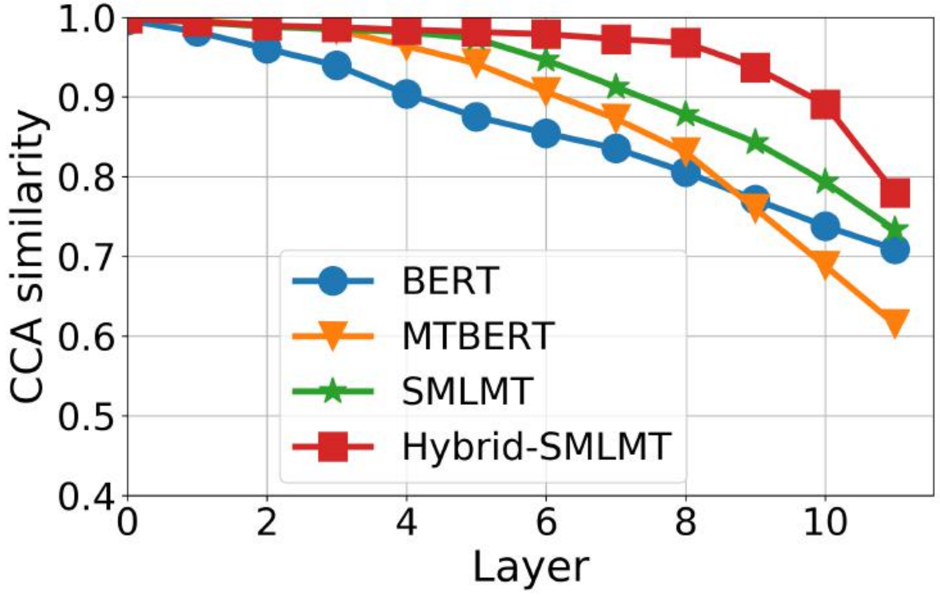
元训练中的学习率轨迹

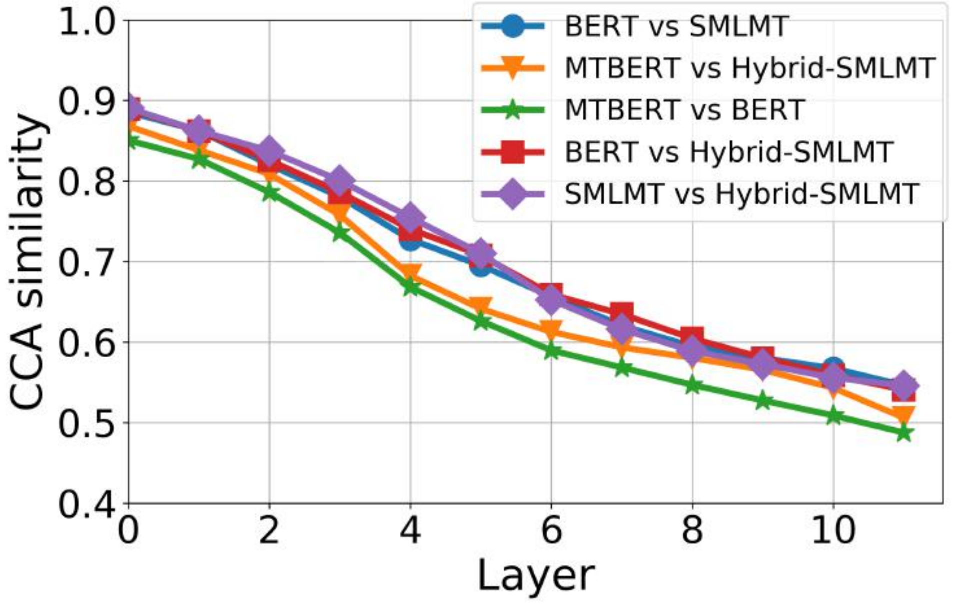




LEOPARD学习率在许多层都趋向于0，这表明元拟合过度

Transformer每层的CCA相似度





左图：对相同模型进行微调之前和之后的相似性。
右图：微调后不同模型对之间的相似性。