要点

- memorization effect + critical parameters
- 减弱了early stopping 之前wrong label对模型的影响
- 将网络参数分为关键参数和非关键参数，用不同的方式学习两种参数

方法

关键参数和非关键参数

关键参数和非关键参数的区别

*lottery ticket hypothesis *表明，深层网络可能被过度参数化，并且只有关键参数对于泛化是重要的；非关键参数学的不是简单模式，而是一些小模式（这样会使得模型过拟合）

关键参数和非关键参数的划分

$$G'(1) = \nabla L(\mathcal{W}; S)^{\top}\mathcal{W} = \langle \nabla L(\mathcal{W}; S), \mathcal{W} \rangle.$$

因为损失函数高维不好分析，因此只能分析$G^{'}(1)$。当$G^{'}(1) = 0$时不能推出模型已经最优，但是最优模型可以推出$G^{'}(1) = 0$

对于单个参数：

$$g_i = |\nabla L(\mathrm{w}_i; S) \times \mathrm{w}_i|, i \in [m].$$

因为在早期训练时关键参数负责因此在早期训练时g_i取值大的参数是关键参数，其他是非关键参数。取最大的$(1-\tau)m$个参数作为关键参数，其余为非关键参数（τ为噪声率）

不同的更新

$$\mathcal{W}_c(k+1) \leftarrow \mathcal{W}_c(k) - \eta \left((1-\tau) \frac{\partial L(\mathcal{W}_c(k); \widetilde{S}^{\star})}{\partial \mathcal{W}_c(k)} + \lambda \mathrm{sgn}(\mathcal{W}_c(k)) \right),$$

关键参数正常更新

$$\mathcal{W}_n(k+1) \leftarrow \mathcal{W}_n(k) - \eta\lambda \mathrm{sgn}(\mathcal{W}_n(k)).$$

非关键参数用权重衰减抑制

实验

t%

Dataset	Method	Symmetric		Asymmetric		Pairflip		Instance	
		20%	40%	20%	40%	20%	40%	20%	40%
MNIST	CE	98.60±0.07	98.18±0.16	99.00±0.08	98.31±0.28	98.74±0.17	94.08±0.93	98.14±0.02	92.76±0.21
	GCE	98.84±0.12	98.12±0.33	98.92±0.09	98.31±0.27	98.94±0.05	97.39±0.62	98.17±0.08	94.97±0.32
	DMI	98.94±0.02	98.62±0.17	99.07±0.10	98.61±0.26	98.96±0.12	97.27±0.39	98.34±0.17	95.34±0.28
	APL	98.74±0.09	97.04±0.35	98.90±0.07	97.23±0.73	98.24±0.39	95.24±1.49	97.03±0.29	90.04±3.93
	MentorNet	97.21±0.13	93.96±0.76	98.51±0.09	93.47±0.80	97.25±0.32	93.27±0.73	95.17±0.26	90.05±1.43
	Co-teaching	97.22±0.18	94.64±0.33	98.63±0.12	93.62±1.27	97.44±0.26	94.81±0.45	97.32±0.15	92.45±0.59
	Co-teaching+	98.11±0.07	95.87±0.27	98.83±0.08	96.65±1.73	98.81±0.12	95.42±0.33	98.07±0.12	94.37±0.48
	S2E	98.93±0.39	93.23±2.37	99.23±0.07	98.31±0.13	99.10±0.04	80.15±3.78	98.42±0.47	83.38±0.94
	Forward	98.10±0.12	96.83±0.28	98.79±0.28	97.94±0.47	98.62±0.16	95.37±0.70	97.87±0.21	92.30±0.18
	T-Revision	98.93±0.07	98.40±0.53	99.05±0.16	98.23±0.54	98.82±0.07	97.43±0.19	98.33±0.15	**95.64±0.34**
	Joint	98.54±0.13	98.30±0.28	98.96±0.05	98.40±0.11	98.70±0.08	96.33±0.82	98.11±0.13	93.15±0.43
	CDR	**99.00±0.04**	**98.80±0.12**	**99.30±0.04**	**98.80±0.17**	**99.17±0.08**	**98.12±0.23**	**98.49±0.09**	94.45±1.04
F-MNIST	CE	90.36±0.21	87.61±0.72	91.31±0.13	86.43±2.01	91.65±0.12	76.42±4.13	88.81±0.67	78.62±2.92
	GCE	91.77±0.13	90.02±0.37	91.45±0.29	73.62±2.92	91.99±0.36	84.21±2.05	91.06±0.55	74.82±0.94
	DMI	91.87±0.26	88.65±0.37	92.33±0.11	89.62±0.48	91.33±0.37	83.93±0.92	90.87±0.15	80.51±0.66
	APL	87.23±0.19	73.62±0.88	86.21±0.24	81.03±3.09	84.52±0.73	76.39±2.85	84.38±0.70	60.38±8.37
	MentorNet	88.12±0.12	86.05±0.27	89.76±0.18	68.93±3.20	87.39±0.57	76.90±5.72	86.50±0.26	78.37±0.95
	Co-teaching	89.03±0.32	87.04±0.69	92.03±0.16	72.23±4.38	89.63±0.78	84.10±0.92	89.27±0.86	83.49±1.27
	Co-teaching+	91.34±0.17	90.23±0.21	83.98±1.05	66.27±3.01	91.08±0.25	72.65±0.49	83.78±0.73	38.79±9.93
	S2E	90.89±0.27	75.68±3.73	91.20±0.31	87.06±0.50	91.52±0.19	72.09±2.15	89.17±0.32	72.62±2.73
	Forward	90.72±0.19	88.05±0.73	92.05±0.21	85.42±0.74	90.02±0.87	83.06±0.79	87.95±0.75	75.34±1.89
	T-Revision	91.95±0.20	90.35±0.28	92.07±0.11	88.53±0.32	91.06±0.19	85.67±0.88	91.05±0.28	84.34±1.37
	Joint	82.01±0.77	72.36±2.84	85.92±0.83	73.09±0.91	86.04±0.99	70.87±3.95	82.07±0.94	50.62±4.77
	CDR	**92.24±0.11**	**90.91±0.27**	**93.01±0.14**	**90.37±0.32**	**93.06±0.19**	**87.55±1.07**	**91.52±0.17**	**85.04±1.02**
CIFAR-10	CE	89.14±0.41	86.25±1.32	88.21±0.19	86.37±1.03	89.68±0.72	86.53±0.37	86.73±0.36	75.33±2.72
	GCE	89.48±0.28	86.07±0.41	89.03±0.21	84.12±1.24	88.58±0.34	83.23±3.98	88.02±0.34	76.89±0.96
	DMI	89.29±0.30	86.89±1.07	89.37±0.82	86.32±1.17	88.41±1.01	84.02±1.73	88.93±0.29	79.35±2.17
	APL	88.21±0.32	81.07±1.36	89.03±0.75	85.10±2.42	87.34±1.44	80.12±3.65	76.31±2.24	50.73±4.89
	MentorNet	83.26±0.72	78.37±1.73	84.07±0.59	60.22±3.47	78.73±0.89	69.37±3.28	83.06±0.92	73.40±2.19
	Co-teaching	88.20±0.27	84.45±0.68	87.42±0.38	64.03±0.73	82.66±0.32	73.68±0.62	86.71±0.79	81.14±1.32
	Co-teaching+	86.47±0.92	78.93±0.74	85.37±0.47	63.17±3.48	84.01±1.01	70.17±1.37	85.92±0.26	57.95±3.17
	S2E	**90.26±0.24**	75.20±2.05	90.73±0.32	87.83±0.97	89.92±0.37	76.18±1.93	90.32±0.21	68.93±1.86
	Forward	88.36±0.34	86.47±0.98	89.30±0.71	85.33±1.48	87.62±0.24	83.23±1.30	85.39±0.23	76.88±1.26
	T-Revision	89.43±0.62	86.98±0.87	89.94±0.74	88.11±1.22	91.01±0.29	87.10±1.38	90.43±0.38	85.46±1.04
	Joint	89.94±0.25	87.17±0.35	90.83±0.18	88.24±0.79	91.31±0.73	85.62±1.75	90.13±0.34	85.23±0.74
	CDR	**90.26±0.31**	**87.19±0.43**	**92.00±0.27**	**88.68±0.67**	**92.11±0.23**	**88.58±0.39**	**91.14±0.23**	**86.25±0.57**
CIFAR-100	CE	63.93±0.72	56.82±0.82	64.12±0.54	52.86±0.92	64.10±0.46	52.77±0.79	63.33±0.29	50.84±0.89
	GCE	65.62±0.82	57.97±1.21	65.34±0.64	54.35±1.28	62.32±1.04	55.03±1.25	66.67±0.40	55.14±1.77
	DMI	62.77±0.92	57.42±0.53	64.30±0.84	51.31±2.73	58.77±0.64	42.89±0.77	59.04±0.35	46.99±0.62
	APL	59.37±0.82	51.03±1.04	54.31±0.84	48.22±1.35	59.77±0.74	53.25±0.92	49.17±2.72	38.18±4.04
	MentorNet	57.27±1.32	49.01±2.09	54.10±0.92	33.21±1.82	54.73±1.26	45.31±2.93	50.02±0.73	36.27±1.64
	Co-teaching	61.47±0.41	53.44±0.40	57.35±0.82	37.62±1.77	58.11±0.47	48.46±0.64	57.73±0.37	43.28±0.55
	Co-teaching+	64.13±0.32	55.92±0.81	58.97±1.19	40.16±2.74	56.31±0.41	38.03±0.55	55.45±0.57	41.11±1.32
	S2E	64.21±0.72	43.12±2.77	63.92±0.46	42.45±1.73	58.21±0.43	41.74±2.09	61.08±0.59	47.06±1.93
	Forward	54.88±0.92	45.64±1.77	64.07±1.02	53.84±2.71	58.37±0.56	39.82±0.73	58.55±0.31	46.42±0.95
	T-Revision	64.67±0.38	57.15±1.02	68.02±0.53	54.93±1.09	62.69±0.73	52.31±1.46	60.22±0.68	50.23±1.79
	Joint	66.12±0.42	59.45±0.68	68.29±0.25	55.53±0.47	67.35±0.31	52.22±1.85	65.91±0.43	55.09±0.93
	CDR	**68.68±0.33**	**62.72±0.38**	**70.64±0.51**	**55.58±0.78**	**71.93±0.57**	**56.94±1.30**	**69.82±0.42**	**61.03±0.77**

Table 3: Classification accuracy (percentage) on *Food-101* dataset. The best result is in bold.

CE	GCE	DMI	APL	MentorNet	Co-teaching
84.03	84.96	85.52	82.17	81.25	83.73
Co-teaching+	S2E	Forward	T-Revision	Joint	CDR
76.89	84.97	85.52	85.97	83.10	**86.36**

Table 4: Top-1 validation accuracies (percentage) on clean *ILSVRC12* validation set of Inception-ResNet v2 models trained on *WebVision* dataset, under the "Mini" setting in (Jiang et al., 2018; Chen et al., 2019; Ma et al., 2020). The best result is in bold.

CE	GCE	DMI	APL	MentorNet	Co-teaching
57.34	55.62	56.93	61.27	57.66	61.22
Co-teaching+	S2E	Forward	T-Revision	Joint	CDR
33.26	54.33	56.39	60.58	47.60	**61.85**

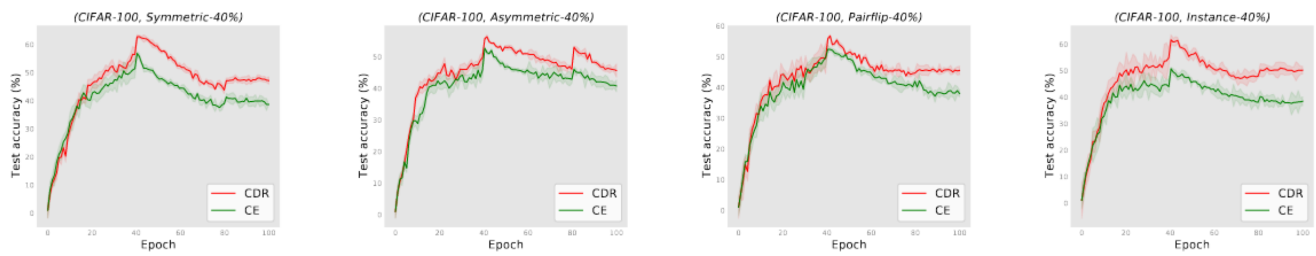
测试如何估计划分系数(默认为$(1-\tau)$)

Table 2: The located constant on synthetic noisy datasets with different noise levels. The result with an underline means that the located constant and the noise rate are numerically equal.

Dataset	Symmetric		Asymmetric		Pairflip		Instance	
	20%	40%	20%	40%	20%	40%	20%	40%
MNIST	$\underline{0.20}$	0.50	$\underline{0.20}$	$\underline{0.40}$	$\underline{0.20}$	0.50	0.30	0.60
F-MNIST	$\underline{0.20}$	$\underline{0.40}$	$\underline{0.20}$	0.50	0.30	0.50	$\underline{0.20}$	0.60
CIFAR-10	0.30	0.50	0.30	0.30	0.30	0.50	$\underline{0.20}$	$\underline{0.40}$

在同一数据集上找出最好的划分系数与(1-τ)作比较

t%e



(a) (b) (c) (d)

证明在 early training stage时这种方法确实可以增强泛化性