Tsinghua University

media and network lab

ICML International Conference On Machine Learning

CurBench: Curriculum Learning Benchmark

Yuwei Zhou, Zirui Pan, Xin Wang, Hong Chen, Haoyang Li, Yanwen Huang,

Zhixiao Xiong, Fangzhou Xiong, Peiyang Xu, Shengnan Liu, Wenwu Zhu

Media and Network Lab, Tsinghua University

Background

Curriculum Learning

- Curriculum learning is a training paradigm where machine learning models are trained **in a meaningful order**, inspired by the way humans learn curricula.
- It brings the advantage of **enhancing model generalization** and **accelerating convergence speed**.

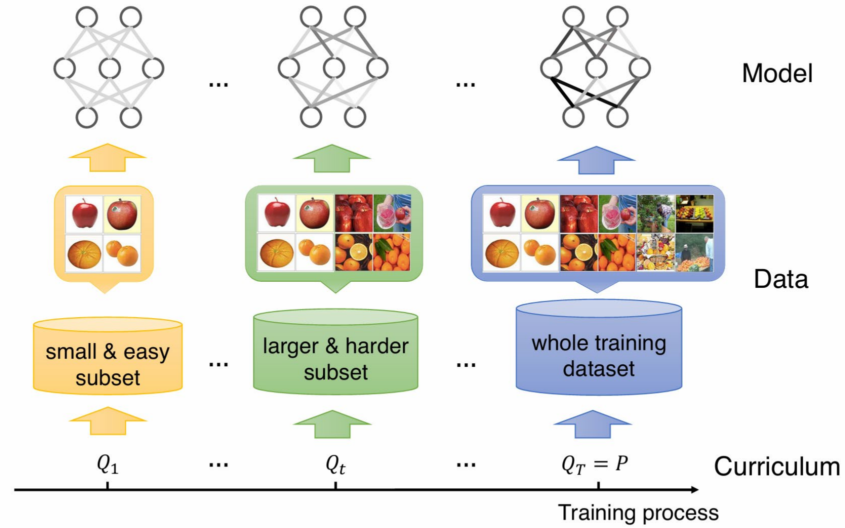


Illustration of Curriculum Learning Concept from [1].

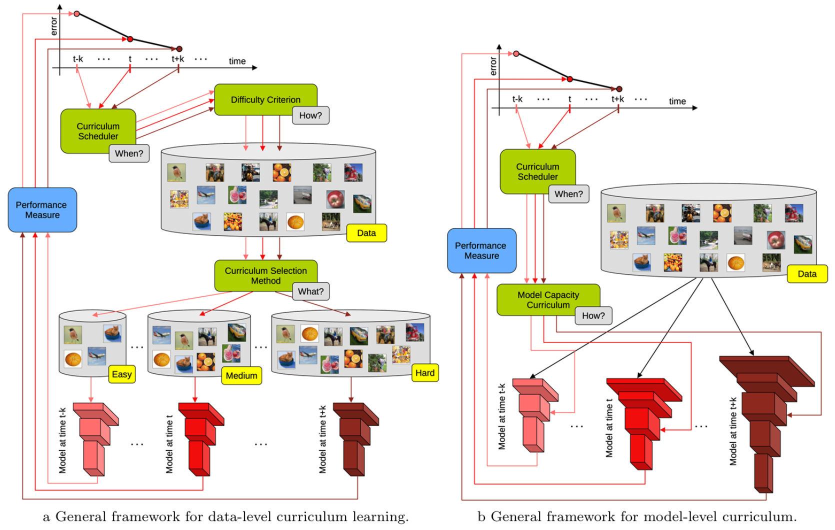


Illustration of Curriculum Learning Framework from [2].

[1] A Survey on Curriculum Learning. TPAMI 2021.
[2] Curriculum Learning: A Survey. IJCV 2022.

Problem

No Benchmark for Curriculum Learning

- As new curriculum learning methods continue to emerge, it remains an open issue to benchmark them.
- The increasing number of works pose challenges in terms of comparison and evaluation, mainly due to **the differences in the experimental setups** including datasets, backbone models, and settings.

Methods of Comparison	The Same	The Difference
DCL v.s. DDS	WideResNet-28-10	CIFAR-100 v.s. CIFAR-10
DIHCL v.s. CBS	ImageNet	ResNet-50 v.s. ResNet-18
MCL v.s. LRE	MNIST and LeNet	Standard v.s. Imbalance

Related Works

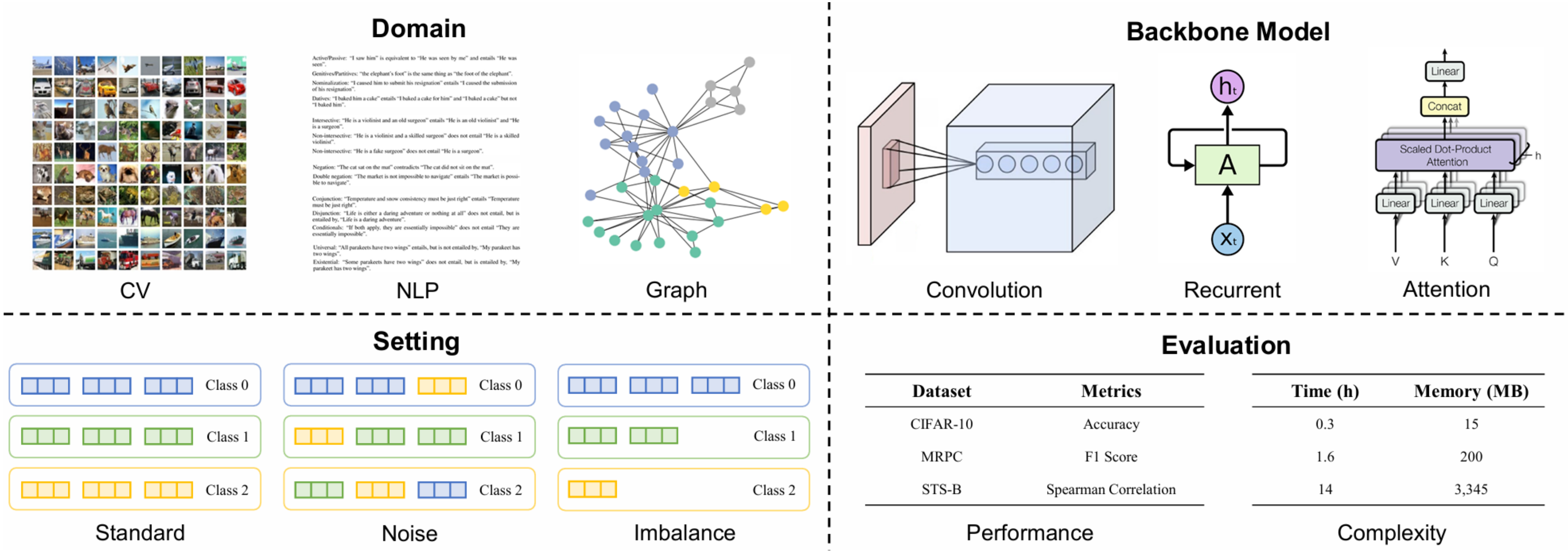
Summative Work on Curriculum Learning

- **From a theoretical perspective:**
  - General Curriculum Learning:
    - A Survey on Curriculum Learning. TPAMI 2021.
    - Curriculum Learning: A Survey. IJCV 2022.
  - Curriculum Learning for Reinforcement Learning:
    - Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey. JMLR 2020.
    - Automatic Curriculum Learning For Deep RL: A Short Survey. IJCAI 2020.
  - Curriculum Learning for Graph Machine Learning:
    - Curriculum graph machine learning: A survey. IJCAI 2023.
- **From an empirical perspective:**
  - Curriculum Learning Library:
    - CurML: A Curriculum Machine Learning Library. ACMMM 2022.

CurBench

Outline

- CurBench includes 15 datasets spanning 3 research domains, 9 backbone models, 3 training settings, and 2 evaluation dimensions, with a toolkit for reproducing 15 core curriculum learning methods.



Dataset

Domain	Dataset	Setting	Training	Validation	Test	Class	Metrics
CV	CIFAR-10	Standard / Noise-0.4	45,000	5,000	10,000	10	Accuracy
		Imbalance-50	12,536	5,000	10,000	10	Accuracy
	CIFAR-100	Standard / Noise-0.4	45,000	5,000	10,000	100	Accuracy
		Imbalance-50	12,536	5,000	10,000	100	Accuracy
	Tiny-ImageNet	Standard / Noise-0.4	90,000	10,000	10,000	200	Accuracy
		Imbalance-50	22,700	10,000	10,000	200	Accuracy
NLP	RTE	Standard / Noise-0.4	2,490	277	-	2	Accuracy
	MRPC	Standard / Noise-0.4	3,668	408	-	2	F1 Score
	STS-B	Standard / Noise-0.4	5,749	1,500	-	6	Spearman
	CoLA	Standard / Noise-0.4	8,551	1,043	-	2	Matthews
	SST-2	Standard / Noise-0.4	67,349	872	-	2	Accuracy
	QNLI	Standard / Noise-0.4	104,743	5,463	-	2	Accuracy
	QQP	Standard / Noise-0.4	363,846	40,430	-	2	F1 Score
	MNLI-(m/mm)	Standard / Noise-0.4	392,702	9,815/9,832	-	3	Accuracy
Graph	MUTAG	Standard / Noise-0.4	150	19	19	2	Accuracy
	PROTEINS	Standard / Noise-0.4	890	111	112	2	Accuracy
	NCI1	Standard / Noise-0.4	3,288	411	411	2	Accuracy
	ogbg-molhiv	Standard / Noise-0.4	32,901	4,113	4,113	2	ROC-AUC

CurBench

Model

Domain	Model	Mechanism	Parameters
CV	LeNet	Convolution	~ 0.07M
	ResNet-18	Convolution	~ 11.2M
	ViT	Attention	~ 9.6M
NLP	LSTM	Recurrent	~ 10.4M
	BERT	Attention	~ 109M
	GPT2	Attention	~ 124M
Graph	GCN	Convolution	~ 0.01M
	GAT	Attention	~ 0.14M
	GIN	Isomorphism	~ 0.01M

Setting

- Standard: No additional data processing.
- Noise-p: p% data samples are independently attached with random incorrect labels.
- Imbalance-r: A ratio of r between the number of samples in the largest class and that in the smallest class in a long-tailed dataset where the number of samples for each class follows a geometric sequence.

Evaluation

- Performance: We report the average and standard deviation of the metric over 5 runs.
- Complexity: We record the training time and maximum memory consumption on the same GPU device.

CurBench

Toolkit

Data Processing

Dataset	Setting	Ratio
CIFAR-10	Noise	0.4

Model Loading

Model	Input Size	Class
ResNet-18	(32, 32)	10

Objective Fitting

Epoch	Optimizer	Objective
200	Adam	Cross Entropy

Curriculum Learning

via Data Selection

- SPL (NeurIPS, 2010)
- MCL (ICLR, 2018)
- LGL (CVPR, 2019)
- C2F (arXiv)
- TTCL (ICML, 2018)
- DIHCL (NeurIPS, 2020)
- Adaptive CL (ICCV, 2021)
- EfficientTrain (ICCV, 2023)

via Model Adjustment

- CBS (NeurIPS, 2020)

via Loss Reweighting

- ScreenerNet (arXiv)
- MW-Net (NeurIPS, 2019)
- SuperLoss (NeurIPS, 2020)
- LRE (ICML, 2018)
- DCL (NeurIPS, 2019)
- DDS (ICML, 2020)

Evaluation

Performance

- Accuracy
- F1 Score
- …… ……

Complexity

- Training Time
- GPU Memory

Experiment

Main Results on CV and Graph Datasets

- There has been no such method that outperforms others all the time, and the effectiveness depends on specific scenarios.

	CIFAR-10			CIFAR-100			Tiny-ImageNet		
	Standard	Noise-0.4	Imbalance-50	Standard	Noise-0.4	Imbalance-50	Standard	Noise-0.4	Imbalance-50
LeNet	$69.95_{1.00}$	$65.02_{1.12}$	$44.93_{0.56}$	$35.46_{0.70}$	$29.59_{0.40}$	$19.57_{0.64}$	$22.08_{0.61}$	$18.63_{0.43}$	$11.65_{0.30}$
LeNet + CL	$\mathbf{70.43}_{0.41}$	$\mathbf{65.93}_{0.57}$	$\mathbf{45.28}_{0.56}$	$\mathbf{35.63}_{0.78}$	$\mathbf{30.87}_{0.48}$	$\mathbf{19.74}_{0.17}$	$\mathbf{22.83}_{0.44}$	$\mathbf{19.91}_{0.26}$	$\mathbf{12.36}_{0.47}$
ResNet-18	$92.33_{0.16}$	$82.75_{2.06}$	$75.49_{0.87}$	$69.97_{0.27}$	$52.14_{0.39}$	$42.57_{0.68}$	$51.41_{1.74}$	$39.42_{0.21}$	$28.83_{0.38}$
ResNet-18 + CL	$\mathbf{92.88}_{0.23}$	$\mathbf{86.92}_{0.20}$	$\mathbf{76.43}_{0.96}$	$\mathbf{71.31}_{0.14}$	$\mathbf{58.56}_{0.60}$	$\mathbf{43.47}_{0.43}$	$\mathbf{53.61}_{0.48}$	$\mathbf{43.64}_{0.72}$	$\mathbf{30.82}_{0.36}$
ViT	$79.90_{0.38}$	$64.19_{0.51}$	$52.12_{0.81}$	$51.05_{0.62}$	$35.25_{0.24}$	$26.05_{0.52}$	$38.16_{0.53}$	$24.90_{0.26}$	$17.15_{0.31}$
ViT + CL	$\mathbf{80.66}_{0.27}$	$\mathbf{69.83}_{0.53}$	$\mathbf{52.85}_{0.81}$	$\mathbf{51.93}_{0.64}$	$\mathbf{39.15}_{0.30}$	$\mathbf{26.40}_{0.34}$	$\mathbf{38.92}_{0.53}$	$\mathbf{29.76}_{0.34}$	$\mathbf{17.47}_{0.14}$

	MUTAG		PROTEINS		NCI1		ogbg-molhiv	
	Standard	Noise-0.4	Standard	Noise-0.4	Standard	Noise-0.4	Standard	Noise-0.4
GCN	$73.68_{2.11}$	$66.31_{7.14}$	$70.71_{4.20}$	$63.57_{6.45}$	$69.59_{1.23}$	$55.23_{3.21}$	$75.84_{1.02}$	$64.29_{4.55}$
GCN + CL	$\mathbf{74.74}_{3.94}$	$\mathbf{71.58}_{5.37}$	$\mathbf{73.21}_{4.41}$	$\mathbf{71.61}_{6.62}$	$\mathbf{71.39}_{1.29}$	$\mathbf{67.98}_{2.01}$	$\mathbf{77.41}_{1.15}$	$\mathbf{72.81}_{1.14}$
GAT	$69.47_{6.14}$	$65.26_{5.37}$	$64.46_{2.96}$	$65.71_{9.13}$	$56.74_{2.86}$	$53.77_{2.12}$	$68.07_{2.34}$	$65.37_{2.66}$
GAT + CL	$\mathbf{72.63}_{8.42}$	$\mathbf{69.47}_{10.21}$	$\mathbf{69.82}_{7.13}$	$\mathbf{69.11}_{3.77}$	$\mathbf{59.37}_{1.59}$	$\mathbf{55.67}_{4.70}$	$\mathbf{72.64}_{1.16}$	$\mathbf{66.73}_{1.84}$
GIN	$86.84_{7.90}$	$78.95_{3.72}$	$74.11_{4.24}$	$69.82_{1.73}$	$79.32_{1.40}$	$60.24_{3.92}$	$74.72_{1.36}$	$63.07_{3.73}$
GIN + CL	$\mathbf{88.42}_{2.10}$	$\mathbf{81.58}_{4.56}$	$\mathbf{77.14}_{4.88}$	$\mathbf{73.93}_{1.82}$	$\mathbf{82.04}_{1.90}$	$\mathbf{62.14}_{6.47}$	$\mathbf{76.53}_{1.97}$	$\mathbf{65.53}_{1.61}$

Experiment

Main Results on NLP Datasets

- There has been no such method that outperforms others all the time, and the effectiveness depends on specific scenarios.

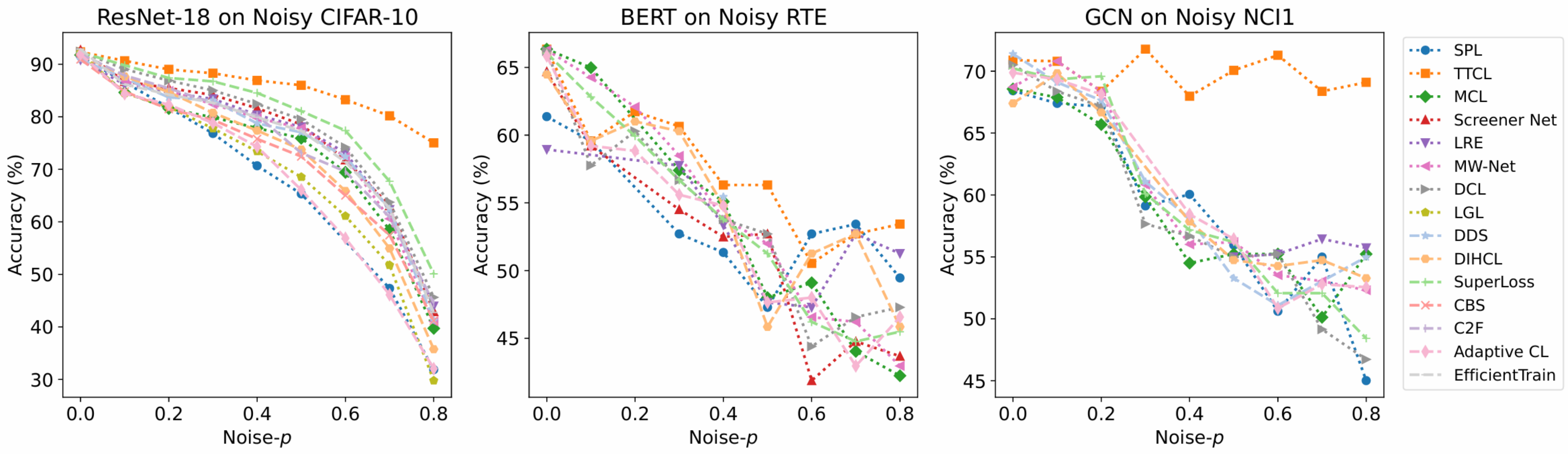
	RTE		MRPC		STS-B		CoLA	
	Standard	Noise-0.4	Standard	Noise-0.4	Standard	Noise-0.4	Standard	Noise-0.4
LSTM	$52.95_{1.34}$	$53.43_{1.77}$	$81.43_{0.14}$	$81.22_{0.00}$	$12.73_{0.72}$	$10.90_{1.19}$	$11.29_{1.27}$	$3.27_{1.68}$
LSTM + CL	$\mathbf{53.07}_{1.29}$	$\mathbf{54.22}_{1.77}$	$\mathbf{81.54}_{0.18}$	$\mathbf{81.24}_{0.05}$	$\mathbf{14.11}_{2.21}$	$\mathbf{11.75}_{1.61}$	$\mathbf{12.65}_{1.21}$	$\mathbf{8.55}_{2.10}$
BERT	$64.62_{3.33}$	$54.22_{3.14}$	$88.54_{0.45}$	$81.89_{0.83}$	$85.26_{0.22}$	$80.71_{1.01}$	$57.39_{1.30}$	$32.35_{0.79}$
BERT + CL	$\mathbf{66.35}_{1.76}$	$\mathbf{56.32}_{5.04}$	$\mathbf{88.69}_{1.24}$	$\mathbf{81.94}_{0.55}$	$\mathbf{85.42}_{0.22}$	$\mathbf{81.31}_{0.25}$	$\mathbf{57.80}_{1.96}$	$\mathbf{45.79}_{1.64}$
GPT2	$65.34_{1.95}$	$52.92_{4.49}$	$85.49_{0.86}$	$78.23_{1.72}$	$76.44_{1.20}$	$69.65_{1.85}$	$37.00_{3.72}$	$5.86_{1.69}$
GPT2 + CL	$\mathbf{66.35}_{2.10}$	$\mathbf{57.40}_{3.39}$	$\mathbf{86.29}_{0.36}$	$\mathbf{82.55}_{0.88}$	$\mathbf{80.82}_{1.39}$	$\mathbf{71.57}_{1.74}$	$\mathbf{39.95}_{3.16}$	$\mathbf{12.54}_{2.75}$

	SST-2		QNLI		QQP		MNLI-(m/mm)	
	Standard	Noise-0.4	Standard	Noise-0.4	Standard	Noise-0.4	Standard	Noise-0.4
LSTM	$81.67_{0.85}$	$64.36_{1.12}$	$50.54_{0.00}$	$50.62_{0.16}$	$75.69_{0.27}$	$60.72_{0.79}$	$61.38_{0.30}$ / $61.21_{0.45}$	$44.41_{0.51}$ / $44.83_{0.90}$
LSTM + CL	$\mathbf{82.87}_{0.88}$	$\mathbf{78.58}_{1.64}$	$\mathbf{51.02}_{0.46}$	$\mathbf{50.83}_{0.45}$	$\mathbf{75.73}_{0.21}$	$\mathbf{66.47}_{0.72}$	$\mathbf{62.47}_{0.36}$ / $\mathbf{62.33}_{0.42}$	$\mathbf{58.59}_{0.54}$ / $\mathbf{58.50}_{0.64}$
BERT	$92.66_{0.28}$	$87.22_{0.82}$	$91.21_{0.24}$	$81.21_{0.76}$	$88.05_{0.12}$	$76.23_{0.48}$	$83.89_{0.31}$ / $84.38_{0.29}$	$78.65_{0.70}$ / $79.21_{0.62}$
BERT + CL	$\mathbf{92.82}_{0.16}$	$\mathbf{91.25}_{0.59}$	$\mathbf{91.49}_{0.13}$	$\mathbf{89.45}_{0.44}$	$\mathbf{88.16}_{0.13}$	$\mathbf{84.50}_{0.25}$	$\mathbf{84.27}_{0.07}$ / $\mathbf{84.40}_{0.42}$	$\mathbf{81.73}_{0.31}$ / $\mathbf{82.25}_{0.40}$
GPT2	$91.95_{0.49}$	$85.83_{0.57}$	$87.92_{0.31}$	$78.72_{0.37}$	$86.00_{0.23}$	$75.40_{0.84}$	$81.53_{0.21}$ / $82.40_{0.21}$	$76.56_{0.15}$ / $77.69_{0.15}$
GPT2 + CL	$\mathbf{92.25}_{0.42}$	$\mathbf{90.34}_{0.53}$	$\mathbf{88.17}_{0.67}$	$\mathbf{84.00}_{0.70}$	$\mathbf{86.68}_{0.16}$	$\mathbf{82.16}_{0.35}$	$\mathbf{81.90}_{0.23}$ / $\mathbf{82.59}_{0.35}$	$\mathbf{78.36}_{0.19}$ / $\mathbf{79.62}_{0.44}$

Experiment

Results in Noise Settings

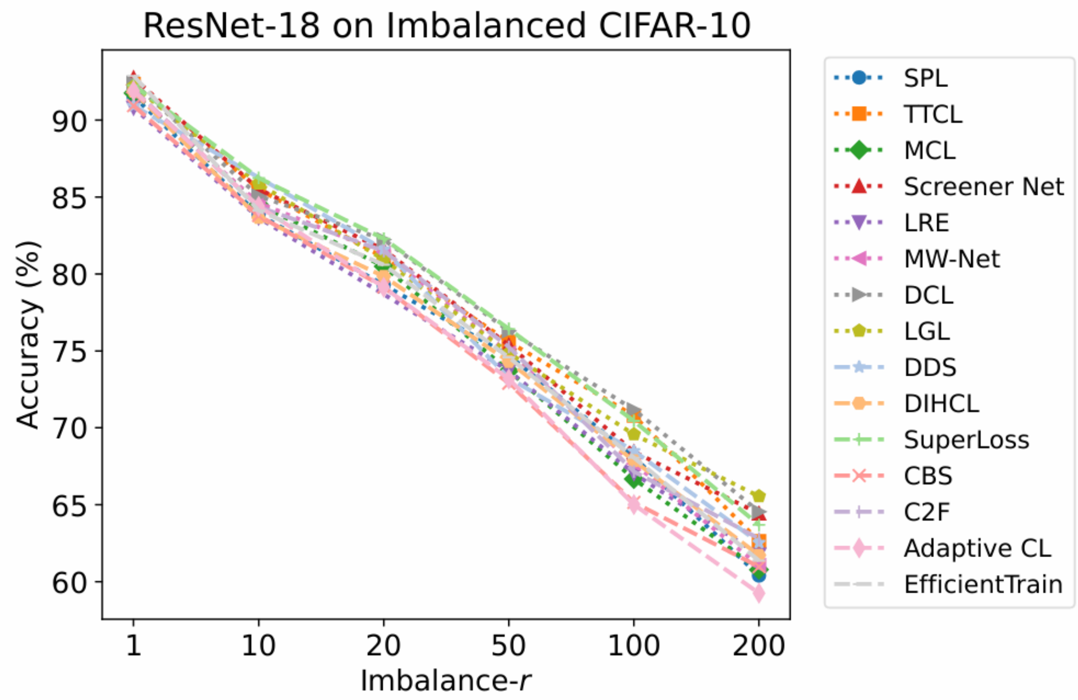
- Methods by teacher transferring have edges in noise settings.



Experiment

Results in Imbalance Settings

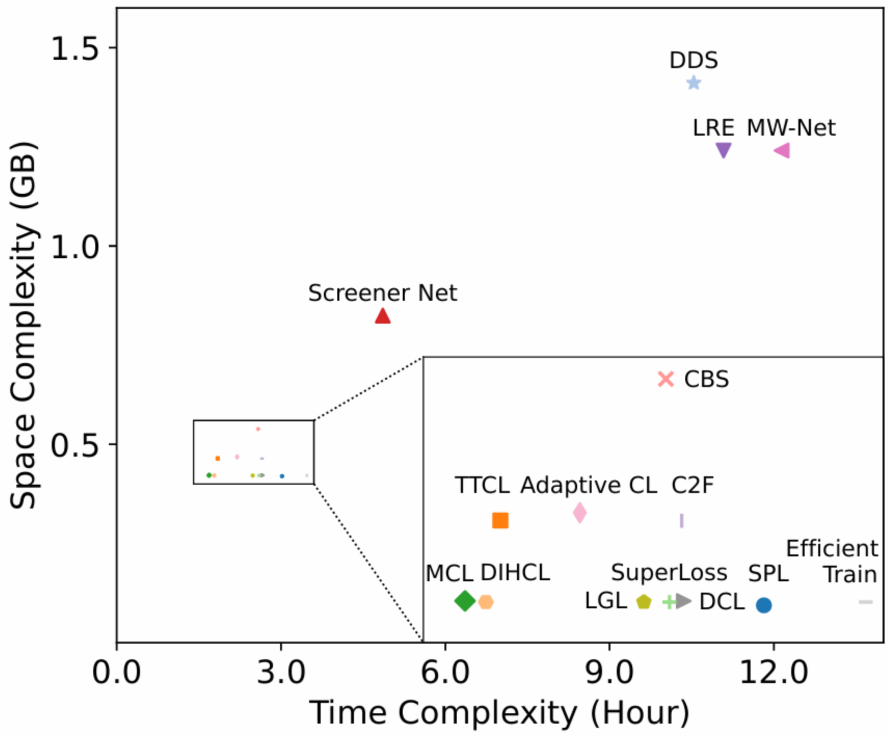
- All methods achieve similar performances under different imbalance ratios.
- Methods by reweighting perform relatively well in imbalance settings.



Experiment

Time and Space Complexity

- Methods involving gradient calculation and extra learnable networks generally have higher time and space complexity.



Summary

Findings

- 1) There has been no such method that outperforms others all the time, and the effectiveness depends on specific scenarios.
- 2) Curriculum learning brings more significant improvements in noise settings than in standard and imbalance ones.
- 3) Methods by teacher transferring have edges in noise settings, while methods by reweighting perform relatively well in imbalance settings.
- 4) Methods involving gradient calculation and extra learnable networks generally have higher time and space complexity.

Summary

Contributions

- 1) We propose CurBench, the first benchmark on curriculum learning to the best of our knowledge.
- 2) We conduct extensive experiments to impartially evaluate and compare the performance and complexity of existing curriculum learning methods under various experimental setups.
- 3) We make in-depth analyses and demonstrate intriguing observations on curriculum learning based on empirical results derived from CurBench.

CurBench: Curriculum Learning Benchmark

Yuwei Zhou, Zirui Pan, Xin Wang, Hong Chen, Haoyang Li, Yanwen Huang,

Zhixiao Xiong, Fangzhou Xiong, Peiyang Xu, Shengnan Liu, Wenwu Zhu

Media and Network Lab, Tsinghua University