STRUCT

Curriculum Temperature for Knowledge Distillation

AAAI 2023

Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao,
Renjie Song, Lei Luo, Jun Li, Jian Yang

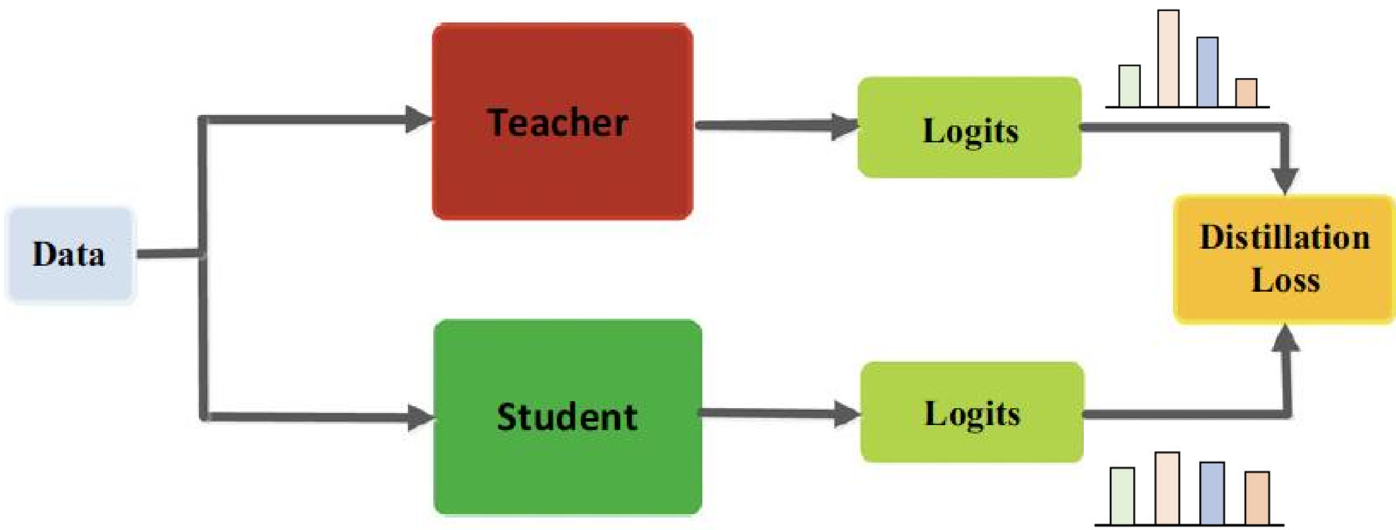
Presented by Yuzhang Hu
2023.1.15

Outline

- Authorship
- Background
- Method
- Experiment
- Conclusion

Background

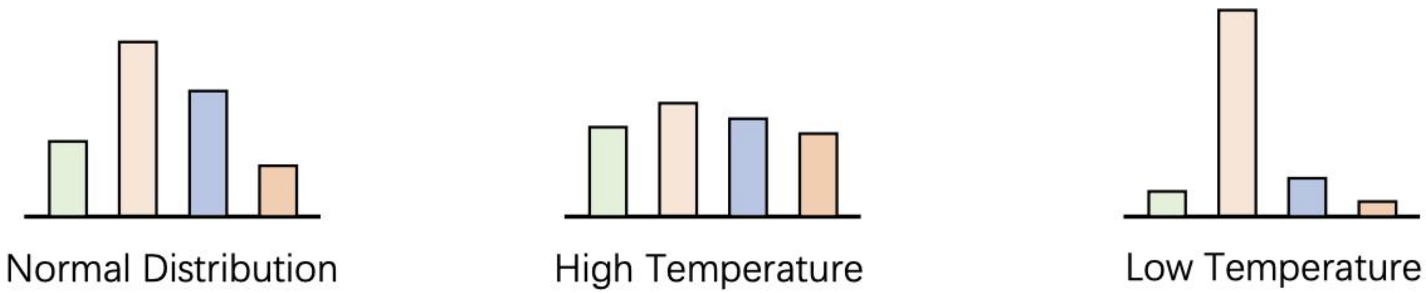
Knowledge Distillation



- Transfer the knowledge from a heavy teacher to a lightweight student
- Minimize distillation loss between two predictions

Background

Distillation Temperature



$$L_{kd}(q^t, q^s, \tau) = \sum_{i=1}^{I} \tau^2 KL(\sigma(q_i^t/\tau), \sigma(q_i^s/\tau))$$

- Control the discrepancy between two distributions
- Determine the difficulty level of the distillation task

Background

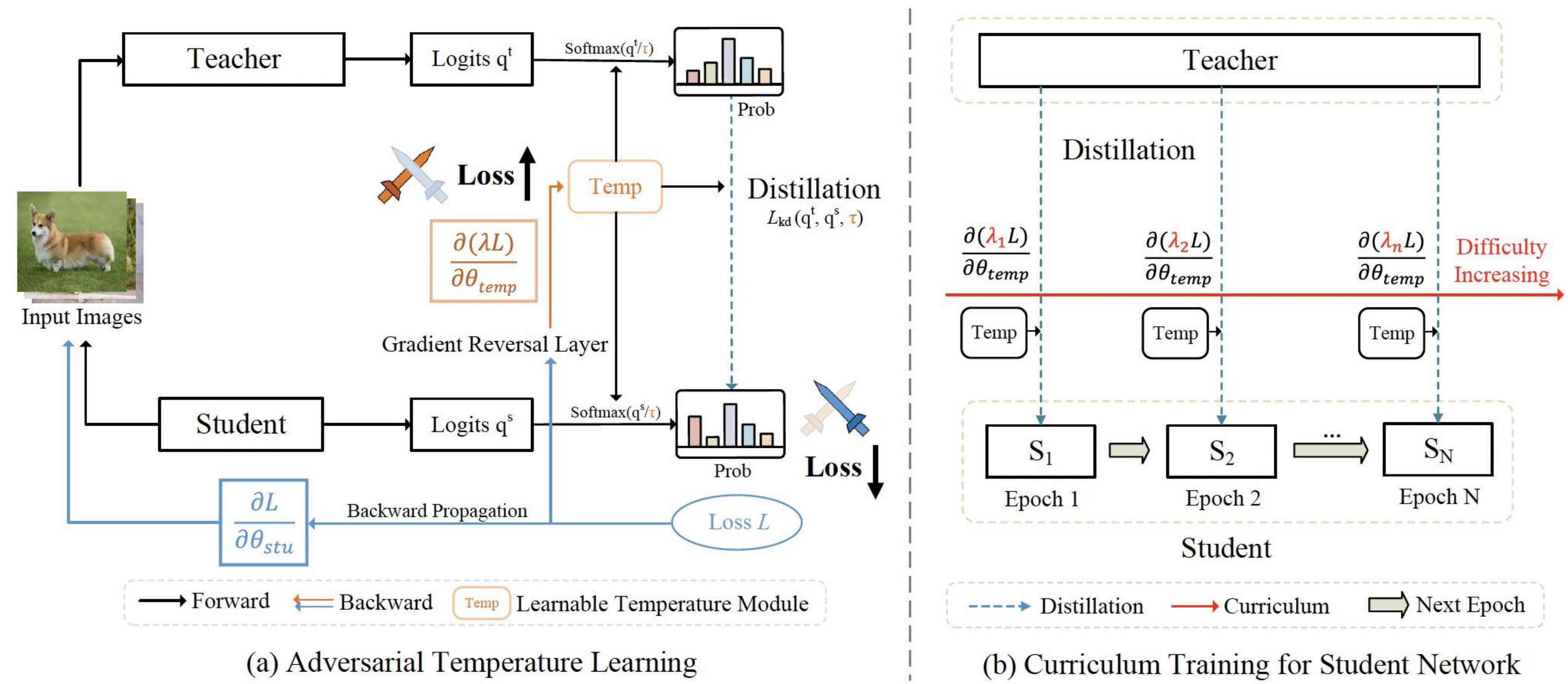
Fixed Temperature

Method	FitNet (ICLR 15)	AT (ICLR 17)	SP (ICCV 19)	Snapshot (CVPR 19)	SSKD (ECCV 20)	FRSKD (CVPR 21)
Temperature	3	4	4	2 or 3	4	4
Method	DML (CVPR 18)	ONE (NIPS 18)	OKDDip (AAAI 20)	KDCL (CVPR 20)	BYOT (ICCV 19)	DCM (ECCV 20)
Temperature	1	3	3	2	1	1

- Fixed temperature is sub-optimal
- Finding optimal temperature is time-consuming

Method

Overall pipeline



- Adversarial Learning for dynamic temperature
- Curriculum Training for Easy-to-hard learning

Method

Adversarial Temperature Learning

Fixed T

$$\min_{\theta_{stu}} L(\theta_{stu}) = \min_{\theta_{stu}} \sum_{x \in D} \alpha_1 L_{task}\left(f^s(x;\theta_{stu}), y\right) + \alpha_2 L_{kd}\left(f^t(x;\theta_{tea}), f^s(x;\theta_{stu}), \tau\right)$$

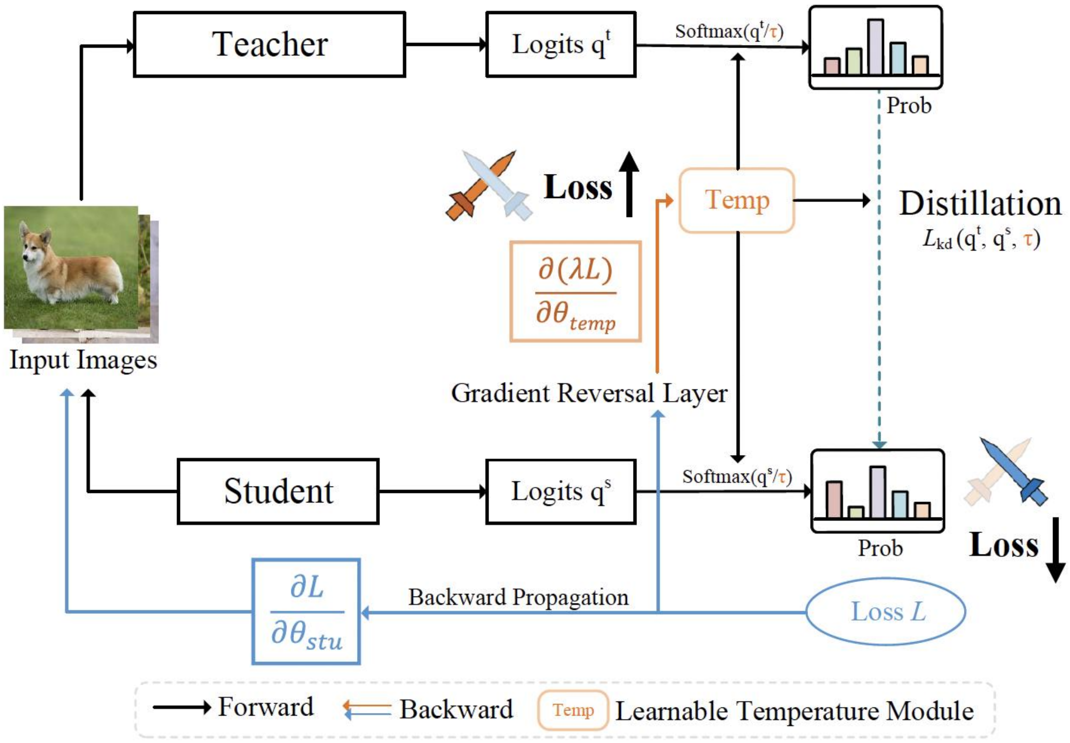
Dynamic T

$$\min_{\theta_{stu}} \max_{\theta_{temp}} L(\theta_{stu}, \theta_{temp})$$

$$= \min_{\theta_{stu}} \max_{\theta_{temp}} \sum_{x \in D} \alpha_1 L_{task}\left(f^s(x;\theta_{stu}), y\right) + \alpha_2 L_{kd}\left(f^t(x;\theta_{tea}), f^s(x;\theta_{stu}), \theta_{temp}\right)$$

Method

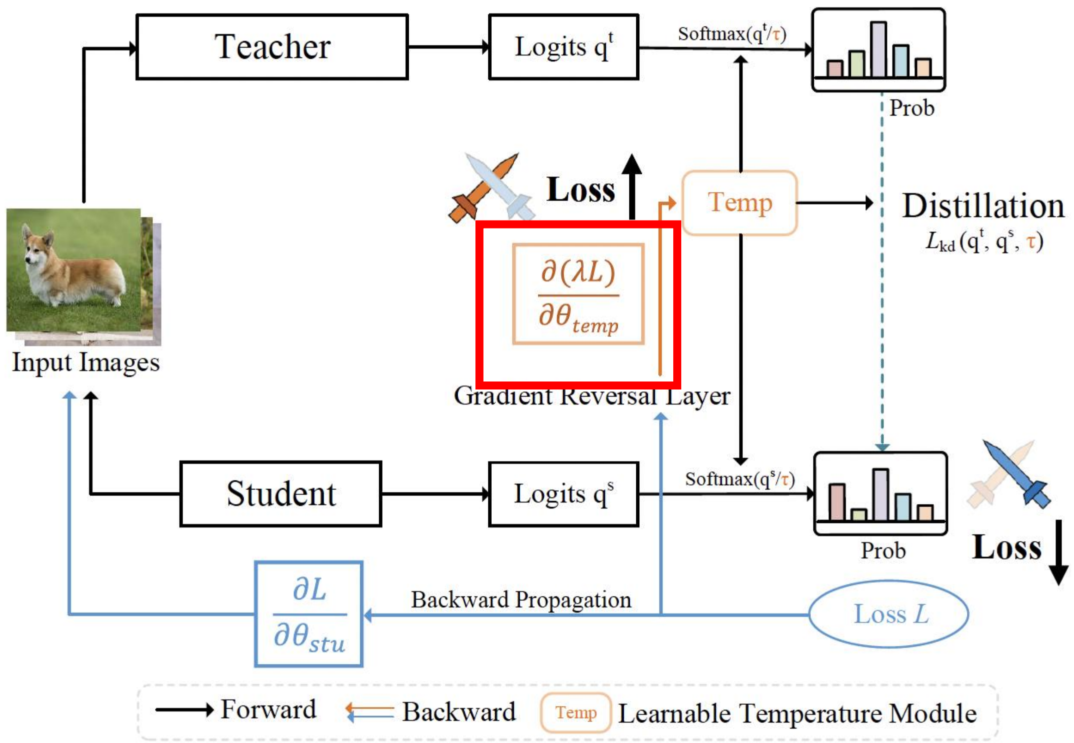
Adversarial Temperature Learning



- GRL reverse the gradient of temperature module
- Update temperature module and student together

Method

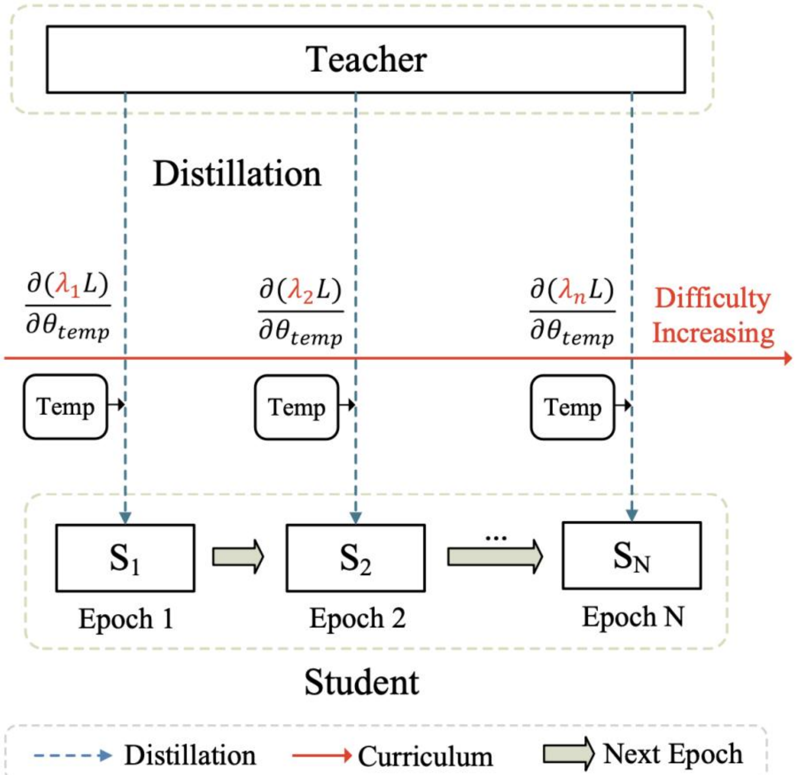
Curriculum Temperature Training



- Easy-to-hard curriculum via scaling temperature gradient

Method

Curriculum Temperature Training



$$\lambda_n = \lambda_{min} + \frac{1}{2}(\lambda_{max} - \lambda_{min})(1 + \cos((1 + \frac{\min(E_n, E_{loops})}{E_{loops}})\pi)$$

$\lambda_{max} \rightarrow 1$

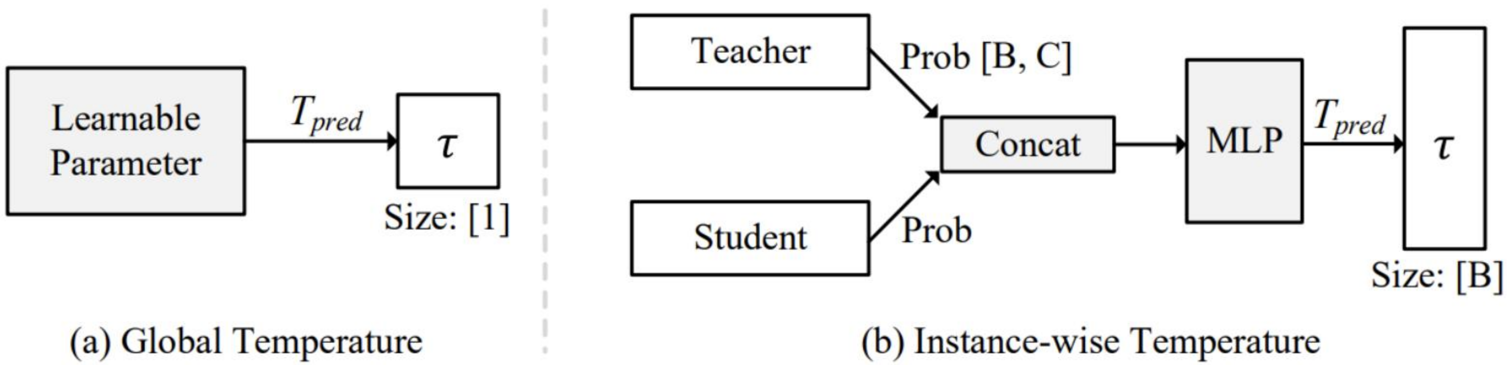
$\lambda_{min} \rightarrow 0$

$E_{loops} \rightarrow 10$

- Increase learning difficulty by gradually increasing λ

Method

Learnable Temperature Module



- **Global Temperature**: one value for all stances
- **Instance-wise Temperature**: takes two predictions as input and outputs a temperature for all instances

Experiment

Quantitative Evaluation

Top-1 accuracy of the student network on CIFAR-100

Teacher Acc	RN-56 72.34	RN-110 74.31	RN-110 74.31	WRN-40-2 75.61	WRN-40-2 75.61	VGG-13 74.64	WRN-40-2 75.61	VGG-13 74.64	RN-50 79.34	RN-32x4 79.42	RN-32x4 79.42
Student Acc	RN-20 69.06	RN-32 71.14	RN-20 69.06	WRN-16-2 73.26	WRN-40-1 71.98	VGG-8 70.36	SN-V1 70.50	MN-V2 64.60	MN-V2 64.60	SN-V1 70.50	SN-V2 71.82
Vanilla KD	70.66	73.08	70.66	74.92	73.54	72.98	74.83	67.37	67.35	74.07	74.45
CTKD	71.19 (+0.53)	73.52 (+0.44)	70.99 (+0.33)	75.45 (+0.53)	73.93 (+0.39)	73.52 (+0.54)	75.78 (+0.95)	68.46 (+1.09)	68.47 (+1.12)	74.48 (+0.41)	75.31 (+0.86)

Top-1 accuracy improvement when applied to existing distillation methods

Teacher Acc	ResNet-56 72.34	ResNet-110 74.31	ResNet-110 74.31	WRN-40-2 75.61	WRN-40-2 75.61	ResNet32x4 79.42	ResNet32x4 79.42
Student Acc	ResNet-20 69.06	ResNet-32 71.14	ResNet-20 69.06	WRN-16-2 73.26	WRN-40-1 71.98	ShuffleNet-V1 70.70	ShuffleNet-V2 71.82
PKT	70.85 ± 0.22	73.36 ± 0.15	70.88 ± 0.16	74.82 ± 0.19	74.01 ± 0.23	74.39 ± 0.16	75.10 ± 0.11
+CTKD	71.16 ± 0.08 (+0.31)	73.53 ± 0.05 (+0.17)	71.15 ± 0.09 (+0.27)	75.32 ± 0.11 (+0.52)	74.11 ± 0.20 (+0.10)	74.68 ± 0.16 (+0.29)	75.47 ± 0.19 (+0.37)
SP	70.84 ± 0.25	73.09 ± 0.18	70.74 ± 0.23	74.88 ± 0.28	73.77 ± 0.20	74.97 ± 0.28	75.59 ± 0.15
+CTKD	71.27 ± 0.10 (+0.43)	73.39 ± 0.11 (+0.30)	71.13 ± 0.13 (+0.39)	75.33 ± 0.14 (+0.45)	74.00 ± 0.15 (+0.23)	75.37 ± 0.17 (+0.40)	75.82 ± 0.18 (+0.23)
VID	70.62 ± 0.08	73.02 ± 0.10	70.59 ± 0.19	74.89 ± 0.16	73.60 ± 0.26	74.81 ± 0.17	75.24 ± 0.05
+CTKD	70.75 ± 0.11 (+0.13)	73.38 ± 0.24 (+0.36)	71.09 ± 0.24 (+0.50)	75.22 ± 0.20 (+0.33)	73.81 ± 0.24 (+0.21)	75.19 ± 0.14 (+0.38)	75.52 ± 0.11 (+0.28)
CRD	71.69 ± 0.15	73.63 ± 0.19	71.38 ± 0.04	75.53 ± 0.10	74.36 ± 0.10	75.13 ± 0.33	75.90 ± 0.15
+CTKD	72.11 ± 0.15 (+0.42)	74.10 ± 0.20 (+0.47)	72.02 ± 0.10 (+0.64)	75.75 ± 0.27 (+0.22)	74.69 ± 0.05 (+0.33)	75.47 ± 0.22 (+0.34)	76.21 ± 0.19 (+0.31)
SRRL	71.13 ± 0.18	73.48 ± 0.16	71.09 ± 0.21	75.69 ± 0.19	74.18 ± 0.03	75.36 ± 0.25	75.90 ± 0.09
+CTKD	71.45 ± 0.15 (+0.32)	73.75 ± 0.30 (+0.27)	71.48 ± 0.14 (+0.39)	75.96 ± 0.06 (+0.27)	74.40 ± 0.13 (+0.22)	75.70 ± 0.22 (+0.34)	76.00 ± 0.22 (+0.10)
DKD	71.43 ± 0.13	73.66 ± 0.15	71.28 ± 0.20	75.70 ± 0.06	74.54 ± 0.12	75.44 ± 0.20	76.48 ± 0.08
+CTKD	71.65 ± 0.24 (+0.27)	74.02 ± 0.29 (+0.36)	71.70 ± 0.10 (+0.42)	75.81 ± 0.14 (+0.11)	74.59 ± 0.08 (+0.05)	75.93 ± 0.29 (+0.49)	76.94 ± 0.04 (+0.46)

Experiment

Quantitative Evaluation

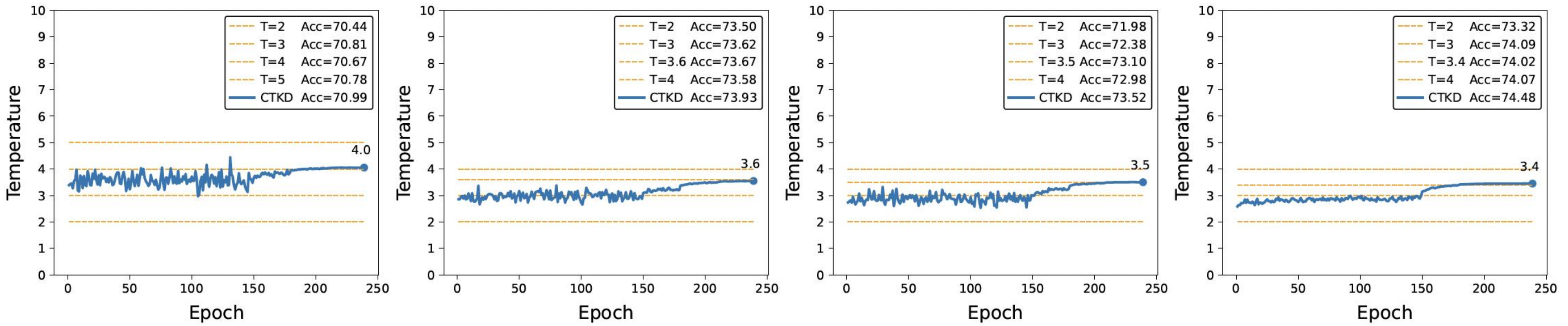
Comparison of global and instance-wise CTKD

Teacher	ResNet-56	ResNet-110	WRN-40-2
Acc	72.34	74.31	75.61
Student	ResNet-20	ResNet-32	WRN-40-1
Acc	69.06	71.14	71.98
Vanilla KD	70.66	73.08	73.54
MACs	41.6M	70.4M	84.7M
Time	10s	15s	17s
Global-T	71.19	73.52	73.93
MACs	41.6M	70.4M	84.7M
Time	10s	15s	17s
Instance-T	71.32	73.61	74.10
MACs	41.7M	70.5M	84.8M
Time	11s	17s	18s

- **Global Temperature**: no extra complexity
- **Instance-wise Temperature**: negligible extra complexity and better performance

Experiment

Curve Visualization



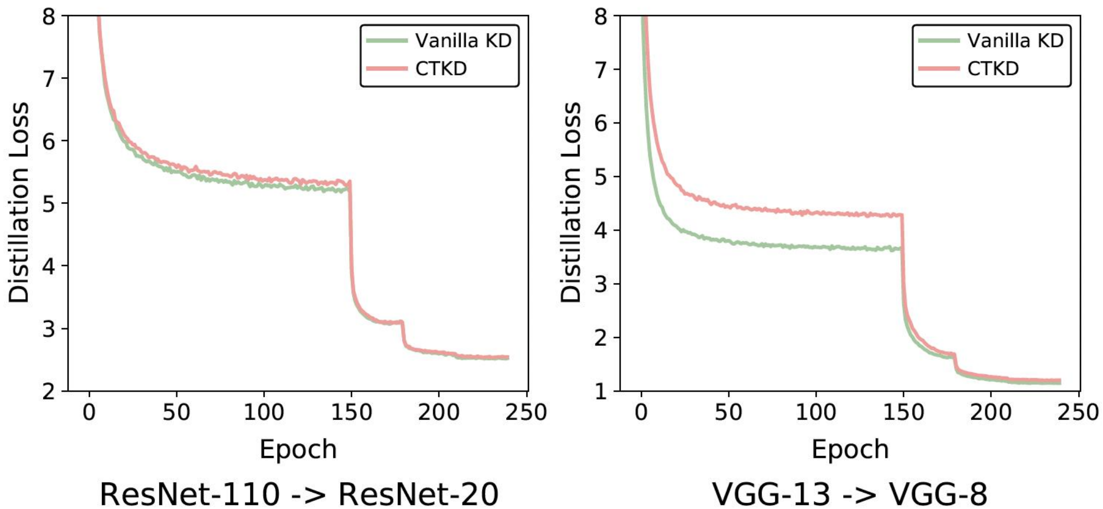
ResNet-110 -> ResNet-20 | WRN-40-2 -> WRN-40-1 | VGG-13 -> VGG-8 | ResNet-32x4 -> ShuffleNet-V1

The learning curves of temperature during training

- Dynamic curriculum temperature outperforms the static method
- Temperature increase in the training process

Experiment

Curve Visualization



The curves of distillation loss during training

The adversarial distillation technique makes the optimization process harder than the vanilla method as expected

Experiment

Hyper-parameter Discussion

E_{loops}	$[\lambda_{min}, \lambda_{max}]$				
	[0, 1]	[0, 2]	[0, 5]	[0, 10]	[1, 10]
10 Epoch	**73.52**	73.16	73.12	73.05	72.58
20 Epoch	73.44	73.48	73.01	73.00	72.88
40 Epoch	73.26	73.40	<u>73.50</u>	73.15	72.95
80 Epoch	73.35	73.46	**73.52**	73.41	73.12
120 Epoch	73.31	73.39	73.16	73.36	73.04
240 Epoch	73.23	73.29	73.20	73.42	73.08

Distillation performance under different Range of dynamic curriculum

Two trends to hurt the distillation performance

- Directly start with a fixed high-difficulty temperature
- Increase temperature in a short time

Conclusion

- Adversarially learn a Dynamic Temperature during the distillation process
- Organize the distillation task from easy to hard with the curriculum temperature training scheme

Thanks!