NEURAL INFORMATION
PROCESSING SYSTEMS

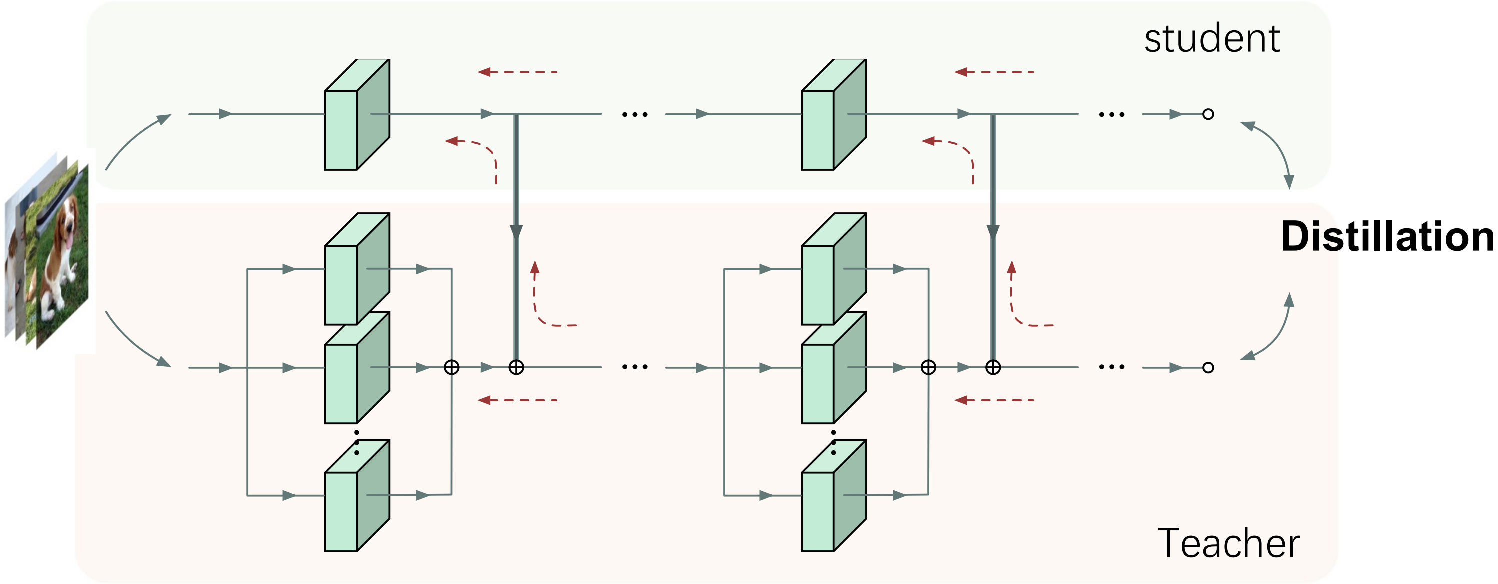
Shadow Knowledge Distillation: Bridging Offline and Online Knowledge Transfer

Lujun Li, Zhe Jin

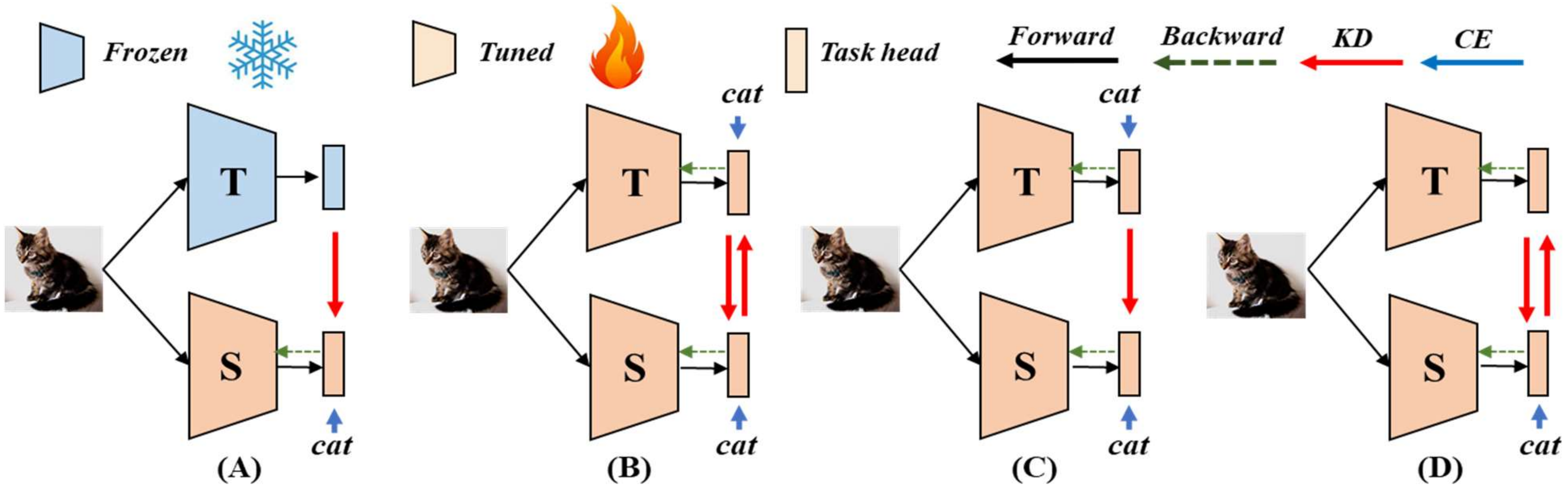
https://lilujunai.github.io/SHAKE/

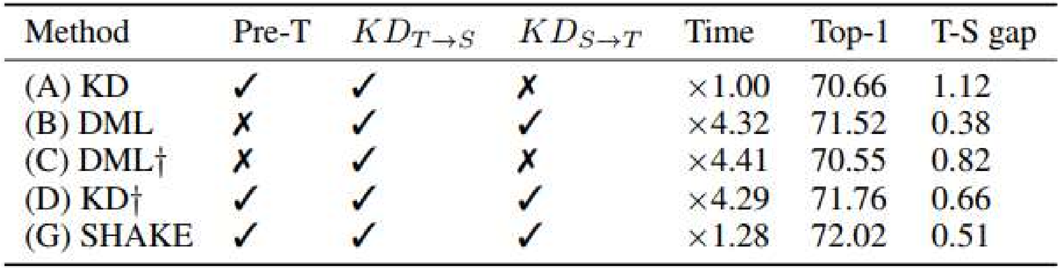
Knowledge Distillation

- Student model minimizes the teacher model output, features, embedding
- Improve lightweight model acc without inference cost
- Offline kd employs existing models yet always with inferior acc than online ones.



Shadow Knowledge Distillation: Motivation

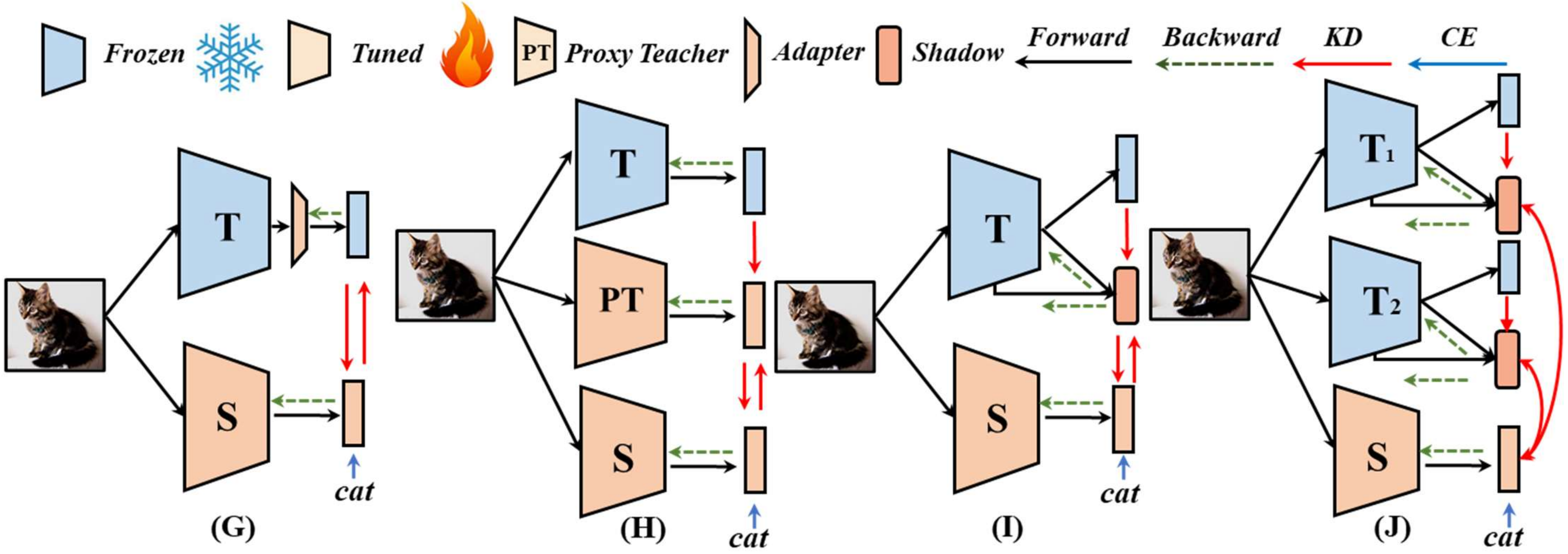




Method	Pre-T	$KD_{T\rightarrow S}$	$KD_{S\rightarrow T}$	Time	Top-1	T-S gap
(A) KD	✓	✓	✗	×1.00	70.66	1.12
(B) DML	✗	✓	✓	×4.32	71.52	0.38
(C) DML†	✗	✓	✗	×4.41	70.55	0.82
(D) KD†	✓	✓	✓	×4.29	71.76	0.66
(G) SHAKE	✓	✓	✓	×1.28	72.02	0.51

Intriguing observation: fashion may not affect the distillation performance. Instead, the reversed distillation from the student model yields significant accuracy gains

Shadow Knowledge Distillation: Method



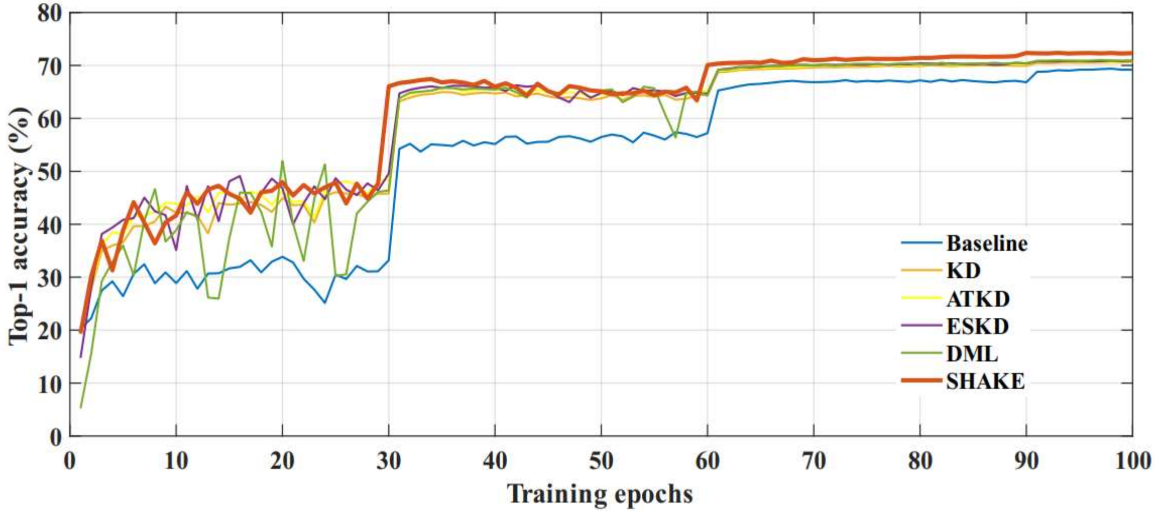
Evolution of SHAKE. (G) KD† with an additional adaptation layer for teacher fine-tuning. (H) We build a proxy teacher model to inherit knowledge from pre-trained models. (I) This proxy teacher model could reuse the backbone. (J) Our SHAKE for multiple teachers: SHAKE leverages multiple shadow heads to individually follow various teacher models.

Shadow Knowledge Distillation: Results

	Same architectural style					Different architectural style		
Teacher	W40-2	R56	R110	R32x4	VGG13	VGG13	R50	W40-2
Student	W16-2	R20	R20	R8x4	VGG8	MV2	VGG8	SV1
Teacher	75.61	72.34	74.31	79.42	74.64	74.64	79.34	75.61
Student	73.26	69.06	69.06	72.5	70.36	64.6	70.36	70.5
DML	75.33	71.35	71.52	74.3	73.64	68.52	74.22	75.58
DML+	74.83	69.95	70.04	73.15	72.86	66.3	73.34	74.52
KD	74.92	70.66	70.67	73.33	72.98	67.37	73.81	74.83
KD+(ours)	**75.78**	**71.52**	**71.76**	**74.91**	**73.85**	**68.81**	**74.4**	**76.42**
SHAKE(ours)	**76.62**	**71.98**	**72.02**	**77.35**	**74.87**	**70.03**	**75.06**	**77.25**
KD+FitNets	75.12	71.12	71.24	74.66	73.49	67.73	73.91	77.42
SHAKE+FitNets	**76.91**	**72**	**72.15**	**78.06**	**74.78**	**70.38**	**75.27**	**78.04**
KD+CRD	75.89	70.9	71.6	75.46	74.08	69.94	74.22	76.27
SHAKE+CRD	**77.17**	**72.29**	**71.87**	**76.57**	**74.65**	**70.04**	**75.22**	**77.61**
KD+Mixup	75.28	71.66	71.33	75.2	74.07	67.31	73.91	76.49
SHAKE+Mixup	**76.91**	**71.82**	**72.07**	**77.39**	**75.53**	**70.25**	**75.66**	**78.17**
KD+CutMix	75.66	70.9	70.69	75.39	74.78	66.39	75.04	77.44
SHAKE+CutMix	**76.29**	**70.92**	**70.9**	**78.28**	**75.11**	**69.44**	**75.98**	**78.27**
KD+AVER	75.22	71.08	71.24	74.99	74.9	68.91	73.26	76.3
SHAKE+AVER	**76.82**	**72.28**	**72.22**	**78.59**	**75.6**	**70.35**	**75.51**	**77.52**
KD+AEKD	75.68	71.25	71.36	74.75	74.75	68.39	73.11	76.34
SHAKE+AEKD	**76.88**	**72.32**	**72.35**	**78.9**	**76.26**	**70.42**	**75.67**	**77.6**

Shadow Knowledge Distillation: Results

T	S	Acc	T	S	KD	ESKD	ATKD	ONE	DML	CRD	SHAKE	SHAKE†
R34	R18	Top-1 Top-5	73.40 91.42	69.75 89.07	70.66 89.88	70.89 90.06	70.78 89.99	70.55 89.59	71.03 90.28	71.17 90.32	**72.07**$_{\pm 0.31}$ **91.05**$_{\pm 0.22}$	**72.53**$_{\pm 0.15}$ **91.26**$_{\pm 0.25}$
T	S	Acc	T	S	KD	AT	RKD	OFD	DML	CRD	**SHAKE**	**SHAKE†**
R50	MV1	Top-1 Top-5	76.16 92.86	70.13 89.49	70.68 90.30	70.72 90.03	71.32 90.62	71.25 90.34	71.13 90.22	71.40 90.42	**72.66**$_{\pm 0.35}$ **91.35**$_{\pm 0.25}$	**73.02**$_{\pm 0.32}$ **91.62**$_{\pm 0.21}$



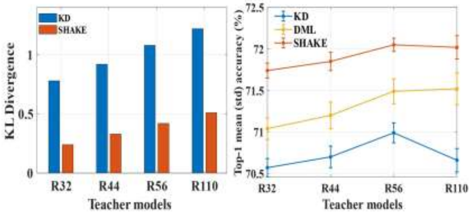


Figure 4: KL-divergence and Top-1 accuracy (%).

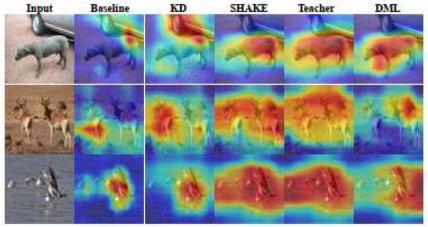


Figure 5: Grad-CAM++[4] visualization.

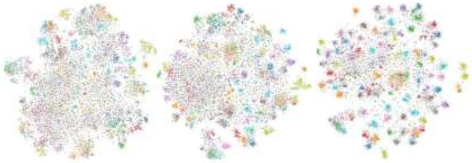


Figure 6: The penultimate layer visualization of ResNet-20 (student) with KD (left), SHAKE (middle) and the teacher (right) on CIFAR-100.

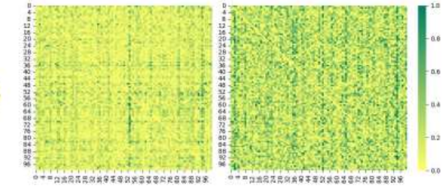
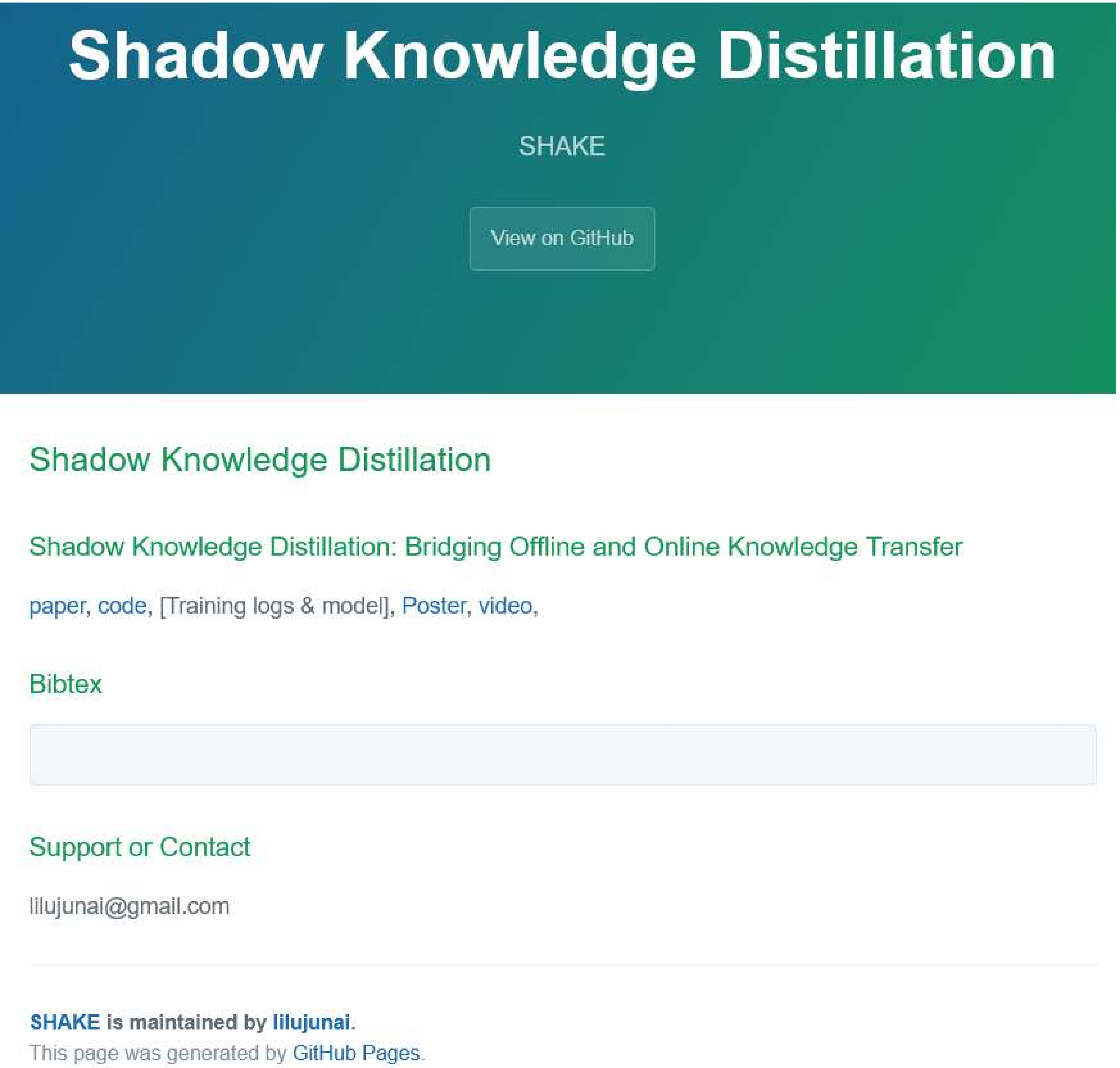


Figure 7: Logits correlation visualization of teacher-student for the student (ResNet-20) with KD (right) and SHAKE (left) on CIFAR-100.



Shadow Knowledge Distillation

SHAKE

View on GitHub

Shadow Knowledge Distillation

Shadow Knowledge Distillation: Bridging Offline and Online Knowledge Transfer

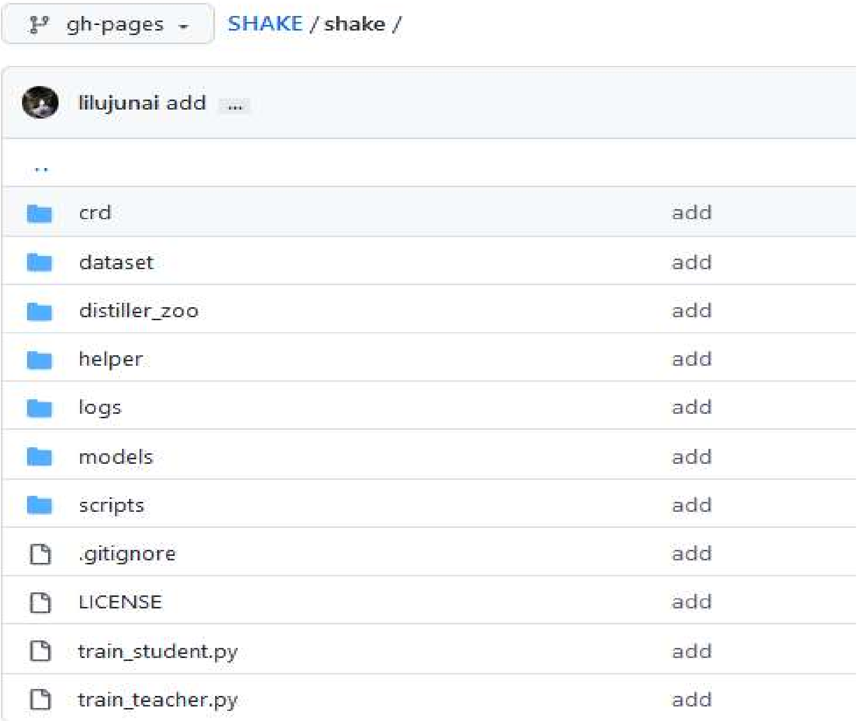
paper, code, [Training logs & model], Poster, video,

Bibtex

Support or Contact

lilujunai@gmail.com

SHAKE is maintained by lilujunai.
This page was generated by GitHub Pages.

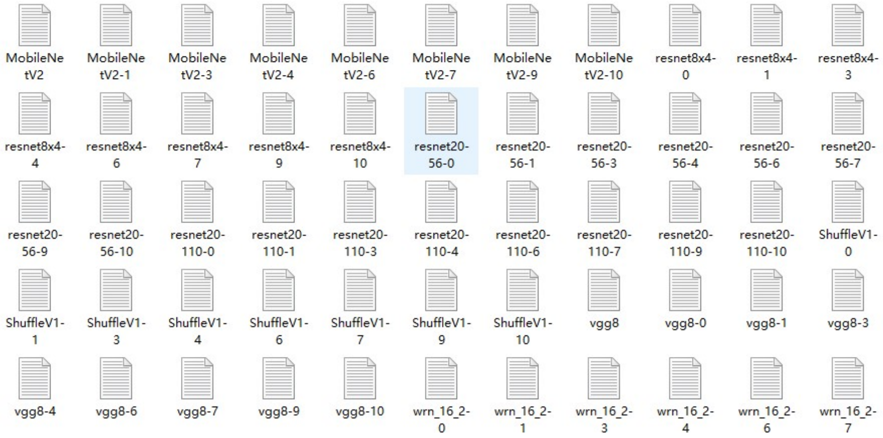


gh-pages SHAKE / shake /

lilujunai add ...

..

crd	add
dataset	add
distiller_zoo	add
helper	add
logs	add
models	add
scripts	add
.gitignore	add
LICENSE	add
train_student.py	add
train_teacher.py	add



MobileNetV2, MobileNetV2-1, MobileNetV2-3, MobileNetV2-4, MobileNetV2-6, MobileNetV2-7, MobileNetV2-9, MobileNetV2-10, resnet8x4-0, resnet8x4-1, resnet8x4-3

resnet8x4-4, resnet8x4-6, resnet8x4-7, resnet8x4-9, resnet8x4-10, resnet20-56-0, resnet20-56-1, resnet20-56-3, resnet20-56-4, resnet20-56-6, resnet20-56-7

resnet20-56-9, resnet20-56-10, resnet20-110-0, resnet20-110-1, resnet20-110-3, resnet20-110-4, resnet20-110-6, resnet20-110-7, resnet20-110-9, resnet20-110-10, ShuffleV1-0

ShuffleV1-1, ShuffleV1-3, ShuffleV1-4, ShuffleV1-6, ShuffleV1-7, ShuffleV1-9, ShuffleV1-10, vgg8, vgg8-0, vgg8-1, vgg8-3

vgg8-4, vgg8-6, vgg8-7, vgg8-9, vgg8-10, wrn_16_2-0, wrn_16_2-1, wrn_16_2-3, wrn_16_2-4, wrn_16_2-6, wrn_16_2-7

NEURAL INFORMATION
PROCESSING SYSTEMS

Thanks for listening!

Lujun Li, Zhe Jin

https://lilujunai.github.io/SHAKE/