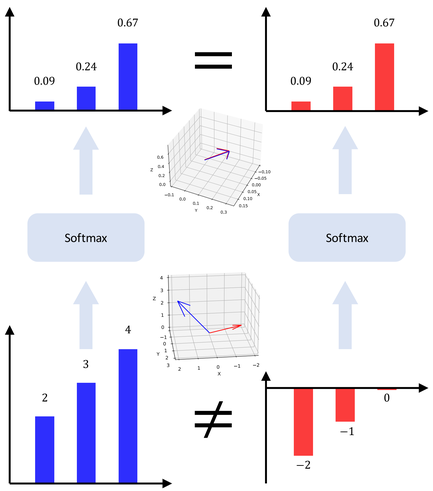
Dual-Head Knowledge Distillation: Enhancing Logits Utilization with an Auxiliary Head

Penghui Yang, Chen-Chen Zong , Sheng-Jun Huang, Lei Feng, Bo An

The softmax function loses certain information



A solution: Replace KL loss with BinaryKL loss

KL loss:

- For multi-class classification
- Use softmax

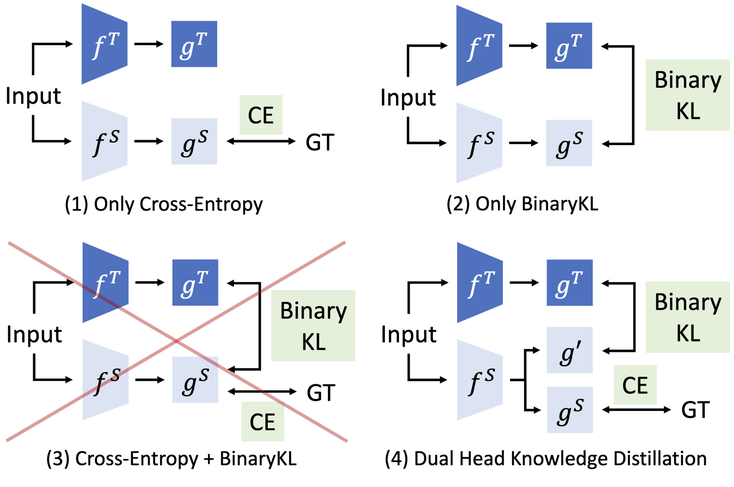
BinaryKL loss:

- Consider each class as a binary classification
- Use sigmoid, which is lossless

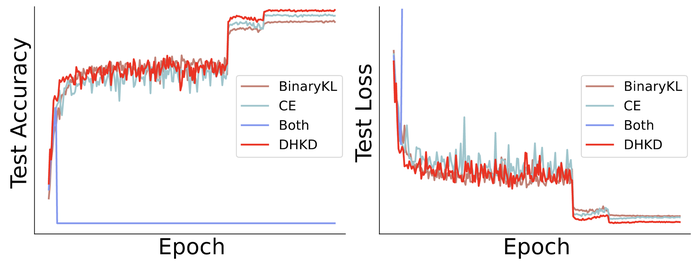
BinaryKL loss:

$$\mathcal{L}_{\text{BinaryKL}} = \tau^2 \sum_{i=1}^{B} \sum_{k=1}^{K} \mathcal{KL}\left(\left[\sigma\left(z_{i,k}^{\mathcal{T}}/\tau\right), 1-\sigma\left(z_{i,k}^{\mathcal{T}}/\tau\right)\right] \| \left[\sigma\left(z_{i,k}^{\mathcal{S}}/\tau\right), 1-\sigma\left(z_{i,k}^{\mathcal{S}}/\tau\right)\right]\right)$$

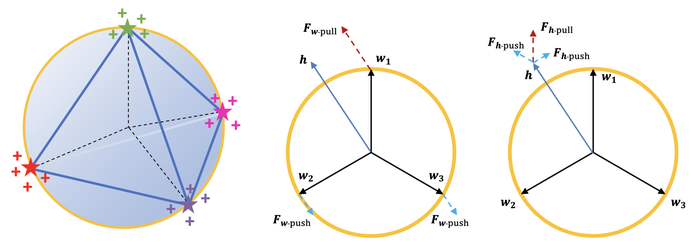
Incompatibility between CE and BinaryKL:



Q: What incompatibility? A: Training collapses.



Q: What causes incompatibility? A: Gradient conflicts in the FC layer.



(a) A simplex equiangular tight frame (b) Gradient directions *w.r.t.* $\boldsymbol{w}$ (c) Gradient directions *w.r.t.* $\boldsymbol{h}$

Proposition 1 *The gradient of* $\mathcal{L}_{\text{overall}}$ *w.r.t. the linear classifier can be formulated as follows:*

$$\frac{\partial \mathcal{L}_{\text{overall}}}{\partial \boldsymbol{w}_k} = -(\boldsymbol{F}_{\boldsymbol{w}\text{-pull}}^{\text{CE}} + \alpha \boldsymbol{F}_{\boldsymbol{w}\text{-pull}}^{\text{BinaryKL}}) - (\boldsymbol{F}_{\boldsymbol{w}\text{-push}}^{\text{CE}} + \alpha \boldsymbol{F}_{\boldsymbol{w}\text{-push}}^{\text{BinaryKL}}),$$

where

$$\boldsymbol{F}_{\boldsymbol{w}\text{-pull}}^{\text{CE}} = \sum_{i=1}^{n_k} (1 - p_k(\boldsymbol{h}_{k,i}^{\mathcal{S}}))\boldsymbol{h}_{k,i}^{\mathcal{S}}, \boldsymbol{F}_{\boldsymbol{w}\text{-pull}}^{\text{BinaryKL}} = \tau \sum_{i=1}^{n_k} (q_k(\boldsymbol{h}_{k,i}^{\mathcal{T}}) - q_k(\boldsymbol{h}_{k,i}^{\mathcal{S}}))\boldsymbol{h}_{k,i}^{\mathcal{S}},$$

$$\boldsymbol{F}_{\boldsymbol{w}\text{-push}}^{\text{CE}} = -\sum_{k' \neq k}^{K} \sum_{j=1}^{n_{k'}} p_k(\boldsymbol{h}_{k',i}^{\mathcal{S}})\boldsymbol{h}_{k',j}^{\mathcal{S}}, \boldsymbol{F}_{\boldsymbol{w}\text{-push}}^{\text{BinaryKL}} = -\tau \sum_{k' \neq k}^{K} \sum_{j=1}^{n_{k'}} (q_k(\boldsymbol{h}_{k,i}^{\mathcal{S}}) - q_k(\boldsymbol{h}_{k,i}^{\mathcal{T}}))\boldsymbol{h}_{k',j}^{\mathcal{S}}.$$

BinaryKL may obstruct near classifier's learning process.

Proposition 2 *The gradient of* $\mathcal{L}_{\text{overall}}$ *w.r.t. the features can be formulated as follows:*

$$\frac{\partial \mathcal{L}_{\text{overall}}}{\partial \boldsymbol{h}} = -(\boldsymbol{F}_{\boldsymbol{h}\text{-pull}}^{\text{CE}} + \alpha \boldsymbol{F}_{\boldsymbol{h}\text{-pull}}^{\text{BinaryKL}}) - (\boldsymbol{F}_{\boldsymbol{h}\text{-push}}^{\text{CE}} + \alpha \boldsymbol{F}_{\boldsymbol{h}\text{-push}}^{\text{BinaryKL}}),$$

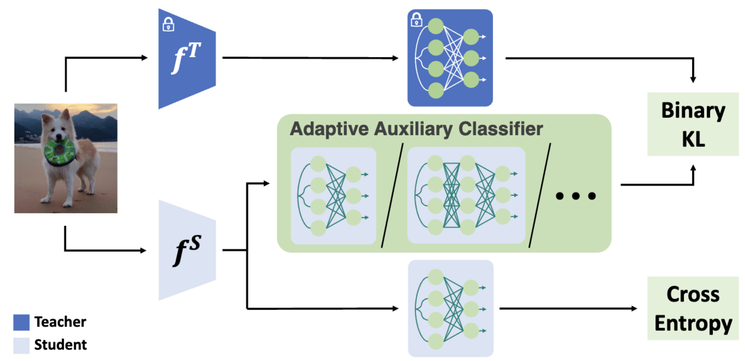
where

$$\boldsymbol{F}_{\boldsymbol{h}\text{-pull}}^{\text{CE}} = (1 - p_c(\boldsymbol{h}^{\mathcal{S}}))\boldsymbol{w}_c^{\mathcal{S}}, \quad \boldsymbol{F}_{\boldsymbol{h}\text{-pull}}^{\text{BinaryKL}} = \tau(q_c(\boldsymbol{h}^{\mathcal{T}}) - q_c(\boldsymbol{h}^{\mathcal{S}}))\boldsymbol{w}_c^{\mathcal{S}},$$

$$\boldsymbol{F}_{\boldsymbol{h}\text{-push}}^{\text{CE}} = -\sum_{k \neq c}^{K} p_k(\boldsymbol{h}^{\mathcal{S}})\boldsymbol{w}_k^{\mathcal{S}}, \quad \boldsymbol{F}_{\boldsymbol{h}\text{-push}}^{\text{BinaryKL}} = -\tau \sum_{k \neq c}^{K} (q_k(\boldsymbol{h}^{\mathcal{S}}) - q_k(\boldsymbol{h}^{\mathcal{T}}))\boldsymbol{w}_k^{\mathcal{S}}.$$

BinaryKL loss provides more detailed information about the teacher model when training the backbone.

Dual-Head Knowledge Distillation



Results on CIFAR-100

	Teacher	resnet56	resnet110	resnet32×4	WRN-40-2	WRN-40-2	VGG13
		72.34	74.31	79.42	75.61	75.61	74.64
	Student	resnet20	resnet32	resnet8×4	WRN-16-2	WRN-40-1	VGG8
		69.06	71.14	72.50	73.26	71.98	70.36
features	FitNet	69.21	71.06	73.50	73.58	72.24	71.02
features	RKD	69.61	71.82	71.90	73.35	72.22	71.48
features	CRD	71.16	73.48	75.51	75.48	74.14	73.94
features	OFD	70.98	73.23	74.95	75.24	74.33	73.95
features	ReviewKD	71.89	73.89	75.63	76.12	75.09	74.84
features	SimKD	71.02	73.89	78.04*	75.48	75.21	74.83
features	CAT-KD	71.62	73.62	76.91	75.60	74.82	74.65
logits	KD	70.66	73.08	73.33	74.92	73.54	72.98
logits	DKD	**71.97***	**74.11***	76.32	76.24	74.81	74.68
logits	DHKD	71.19	73.92	**76.54**	**76.36***	**75.25***	**74.84***
	DHKD + ReviewKD	**73.14**	**75.21**	**78.29**	**77.97**	**76.56**	**76.27**

	Teacher	resnet32×4	WRN-40-2	VGG13	ResNet-50	resnet32×4
		79.42	75.61	74.64	79.34	79.42
	Student	ShuffleNet-V1	ShuffleNet-V1	MBN-V2	MBN-V2	ShuffleNet-V2
		70.50	70.50	64.60	64.60	71.82
features	FitNet	73.59	73.73	64.14	63.16	73.54
features	RKD	72.28	72.21	64.52	64.43	73.21
features	CRD	75.11	76.05	69.73	69.11	75.65
features	OFD	75.98	75.85	69.48	69.04	76.82
features	ReviewKD	77.45	77.14	70.37*	69.89	77.78
features	SimKD	77.18	77.23	69.45	71.12	78.39
features	CAT-KD	78.26*	77.35*	69.13	71.36*	78.41*
logits	KD	74.07	74.83	67.37	67.35	74.45
logits	DKD	76.45	76.70	69.71	70.35	77.07
logits	DHKD	**76.78**	**77.25**	**70.09**	**71.08**	**77.99**
	DHKD + ReviewKD	**78.35**	**78.22**	**71.01**	**71.51**	**78.83**

Results on ImageNet

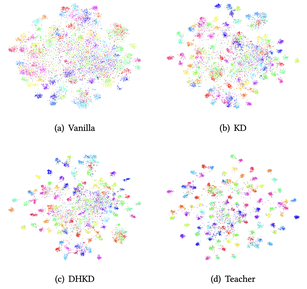
T: ResNet-34
S: ResNet-18

distillation manner			features						logits			
Metric	Teacher	Student	AT	OFD	CRD	ReviewKD	SimKD	CAT-KD	KD	DKD	DIST	DHKD
top-1	73.31	69.75	70.69	70.81	71.17	71.61	71.59	71.26	70.66	71.70	72.07	**72.15**
top-5	91.42	89.07	90.01	89.98	90.13	90.51	90.48	90.45	89.88	90.41	90.42	**90.89**

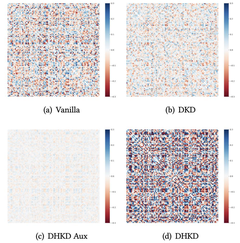
T: ResNet-50
S: MobileNet

distillation manner			features						logits			
Metric	Teacher	Student	AT	OFD	CRD	ReviewKD	SimKD	CAT-KD	KD	DKD	DIST	DHKD
top-1	76.16	68.87	69.56	71.25	71.37	72.56	72.25	72.24	68.58	72.05	**73.24**	72.99
top-5	92.86	88.76	89.33	90.34	90.41	91.00	90.86	91.13	88.98	91.05	91.12	**91.45**

t-SNE Visualization



The difference between the correlation matrices of the teacher's and student's logits



Acknowledgement
This research is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Prepositioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. This research is also supported by the Joint NTU-WeBank Research Centre on Fintech, Nanyang Technological University, Singapore. Sheng-Jun Huang is supported by the NSFC(U2441285, 62222605).