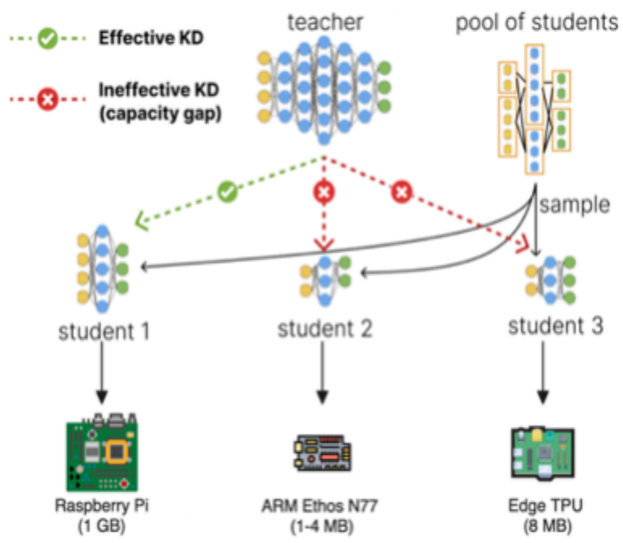
Generalizing Teacher Networks for Effective Knowledge Distillation Across Student Architectures

Kuluhan Binici*, Weiming Wu*, Tulika Mitra*
*School of Computing, National University of Singapore, Singapore

Introduction

Knowledge Distillation (KD) transfers knowledge from a large teacher model to a smaller student model.

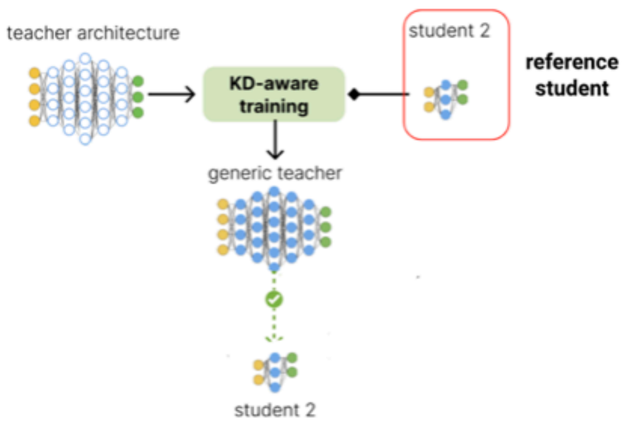
The capacity gap between teacher and student model architectures reduces KD efficiency



Specializing Teacher Models for Students

KD-aware training can be used to specialize teacher models for optimized transfer of knowledge to a reference student

- Effective
- Has to be repeated for each student, undesirable when distilling multiple different student models

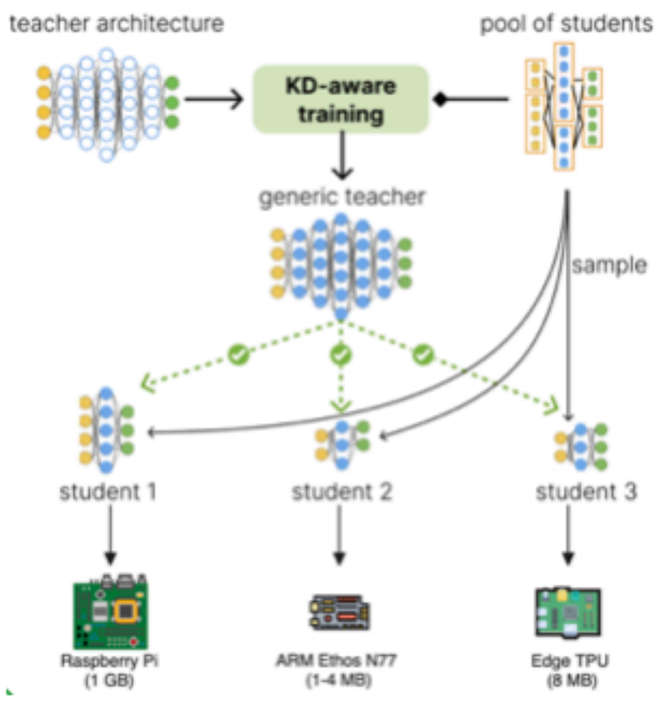


$$L_{CT} = \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{y_{gt}\log(\hat{y}_{s_i})}_{L_{CE_S}} + \underbrace{\alpha T^2 D_{KL}\left(\frac{z_t}{T}\middle\|\frac{z_{s_i}}{T}\right)}_{L_{KL}}\right) + \underbrace{y_{gt}\log(\hat{y}_t)}_{L_{CE_T}}$$

Generic Teacher Networks (GTN)

Objectve: Train a single teacher to perform well across multiple student architectures.

- Saves training time by creating a one-off teacher suitable for KD to all student models contained in a finite pool.
- Enables effective KD across various student models and deployment platforms without increasing training cost.



Methodology

GTN training conditions the teacher model to conform with the architectural capacities of various student architectures.

To avoid excessive time cost, a weight-sharing reconfigurable supernet architecture is used to represent different student architectures at each iteration of teacher conitioning process

Experiment Results

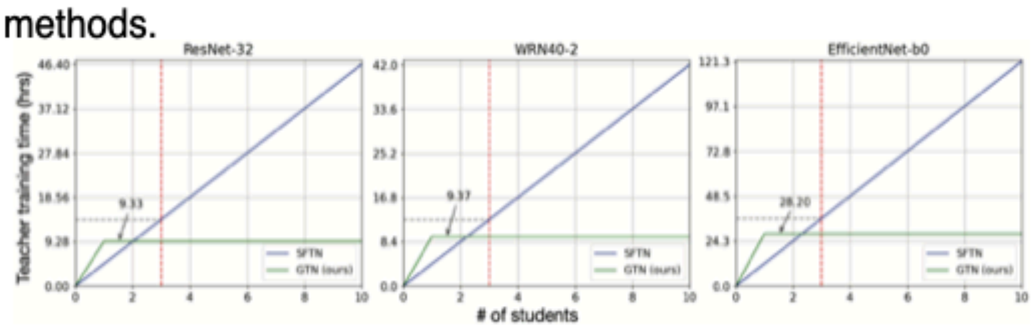
Experimented with 7 different student models sampled from the finite architecture pool, each having different resource footprint

	$s1$		$s2$		$s3$		$s4$		s^s_{nas}		s^m_{nas}		s^l_{nas}	
Bit-width	32b	8b	32b	8b	32b	8b	32b	8b	32b	8b	32b	8b	32b	8b
Memory Size (MB)	22.2	5.5	29.9	7.5	50.6	12.6	25.4	6.4	6.7	1.7	28.6	7.2	77.1	19.3
Arm Ethos N77 (1-4 MB)	X		X		X		X		✓		X		X	
Edge TPU (8 MB)	✓		✓		X		✓		✓		✓		X	
Raspberry Pi (1 GB)	✓		✓		✓		✓		✓		✓		✓	

GTN consistently outperforms baseline methods that specialize teacher models for individual student architectures

Method	no KD	Vanilla KD	SCKD	SFTN				GTN (ours)
CIFAR-100								
Reference Student	N/A	N/A	N/A	s1	s2	s3	s4	supernet
Teacher acc.	N/A	78.04	78.04	81.54	81.46	81.49	80.97	80.71
s1 acc. (Δ)	73.22	75.40	75.34 (−0.06)	75.82 (+0.42)	76.09 (+0.69)	75.34 (−0.06)	76.03 (+0.63)	**76.70** (+1.30)
s2 acc. (Δ)	77.92	77.89	77.47 (−0.42)	78.23 (+0.34)	78.52 (+0.63)	77.90 (+0.01)	77.76 (−0.13)	**78.85** (+0.96)
s3 acc. (Δ)	76.87	77.21	77.08 (−0.13)	77.66 (+0.45)	78.05 (+0.84)	78.05 (+0.84)	77.94 (+0.73)	**78.22** (+1.01)
s4 acc. (Δ)	75.60	76.42	75.77 (−0.65)	77.20 (+0.78)	77.71 (+1.29)	77.23 (+0.81)	77.66 (+1.24)	**77.79** (+1.37)
ImageNet-200								
Reference Student	N/A	N/A	N/A	s5	s6	s7	s8	supernet
Teacher acc.	N/A	65.28	65.28	69.71	69.08	68.96	69.00	70.10
s5 acc. (Δ)	62.63	64.86	65.36 (+0.50)	65.54 (+0.68)	64.99 (+0.13)	65.08 (+0.22)	65.30 (+0.44)	**65.63** (+0.77)
s6 acc. (Δ)	64.34	64.79	65.68 (+0.45)	66.00 (+1.21)	65.71 (+0.92)	65.88 (+1.09)	65.62 (+0.83)	**66.08** (+1.29)
s7 acc. (Δ)	62.62	64.37	63.97 (−0.40)	64.69 (+0.32)	64.73 (+0.36)	64.71 (+0.34)	65.13 (+0.76)	**65.74** (+1.37)
s8 acc. (Δ)	62.15	63.53	63.95 (+0.42)	63.48 (−0.05)	64.13 (+0.60)	63.88 (+0.35)	64.13 (+0.60)	**64.29** (+0.76)

GTN's one-off training saves time as the number of deployment scenarios increases. When more than two student models are to be distilled and deployed, our approach takes over the time advantage against specialization-based methods.



Conclusion

This study presents a generic teacher model that effectively transfers knowledge across diverse student architectures, addressing the capacity gap in Knowledge Distillation (KD). Our approach improves KD performance with a constant, low training cost, making it practical for multi-platform deployments.