ECCV'20 ONLINE

23-28 AUGUST 2020

16TH EUROPEAN CONFERENCE ON

COMPUTER VISION

WWW.ECCV2020.EU

ECCV'20
ONLINE
23-28 AUGUST 2020

Online Ensemble Model Compression using Knowledge Distillation

By
Devesh Walawalkar, Zhiqiang Shen and Marios Savvides
Carnegie Mellon University, Pittsburgh PA, USA

E: devwalkar64@gmail.com, {zhiqians, marioss}@andrew.cmu.edu

Overview:

1. We present a novel ensemble model compression framework for simultaneously training multiple compressed models using an online knowledge distillation-based scheme.

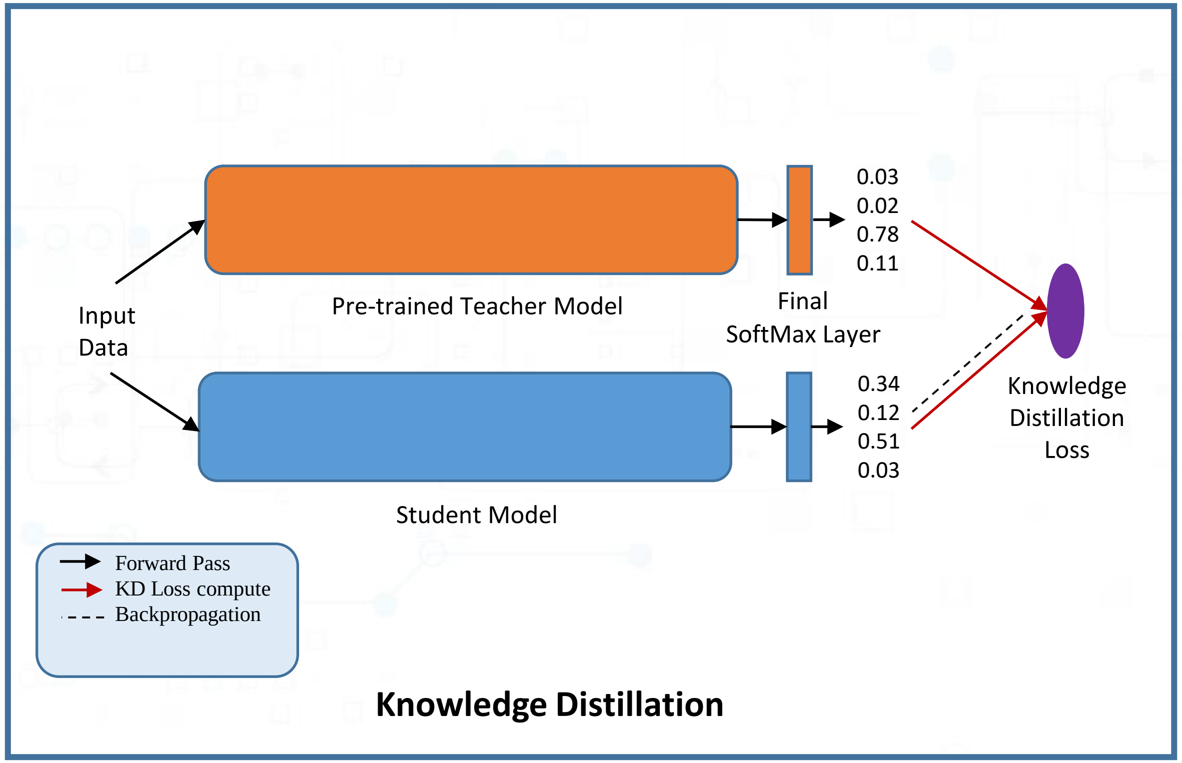
2. Major highlights of our work include:

   a) Pre-trained teacher free and model architecture agnostic approach.
   b) Facilitates simultaneous training of multiple instances of a given architecture compressed to varying degrees.
   c) Significant training time savings over individually training each model.
   d) Consistent performance gains for all compressed models over their baseline individual training.

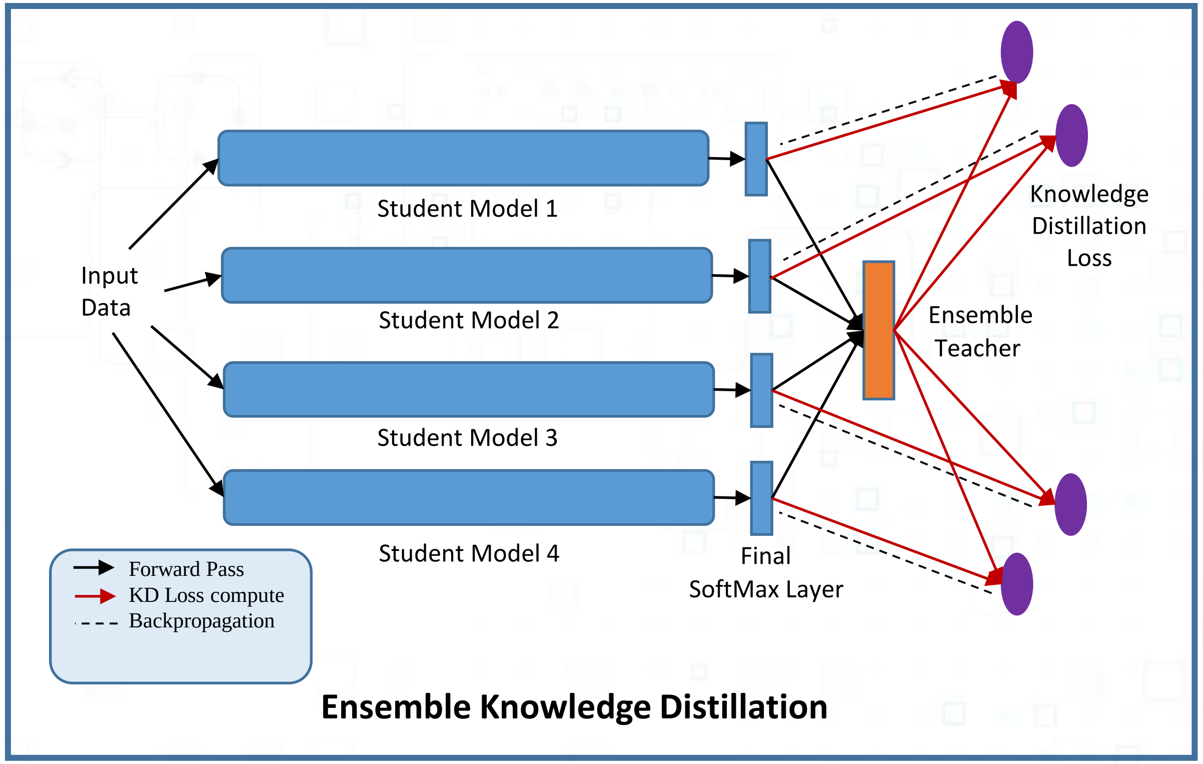
Knowledge Distillation and its Ensemble Model Configuration

Knowledge Distillation is the process of distilling knowledge (specifically output softmax distribution) of a pre-trained teacher model onto another model called a student.

Ensemble Knowledge Distillation represents distillation of knowledge from an ensemble teacher to each of its ensemble students.



Knowledge Distillation

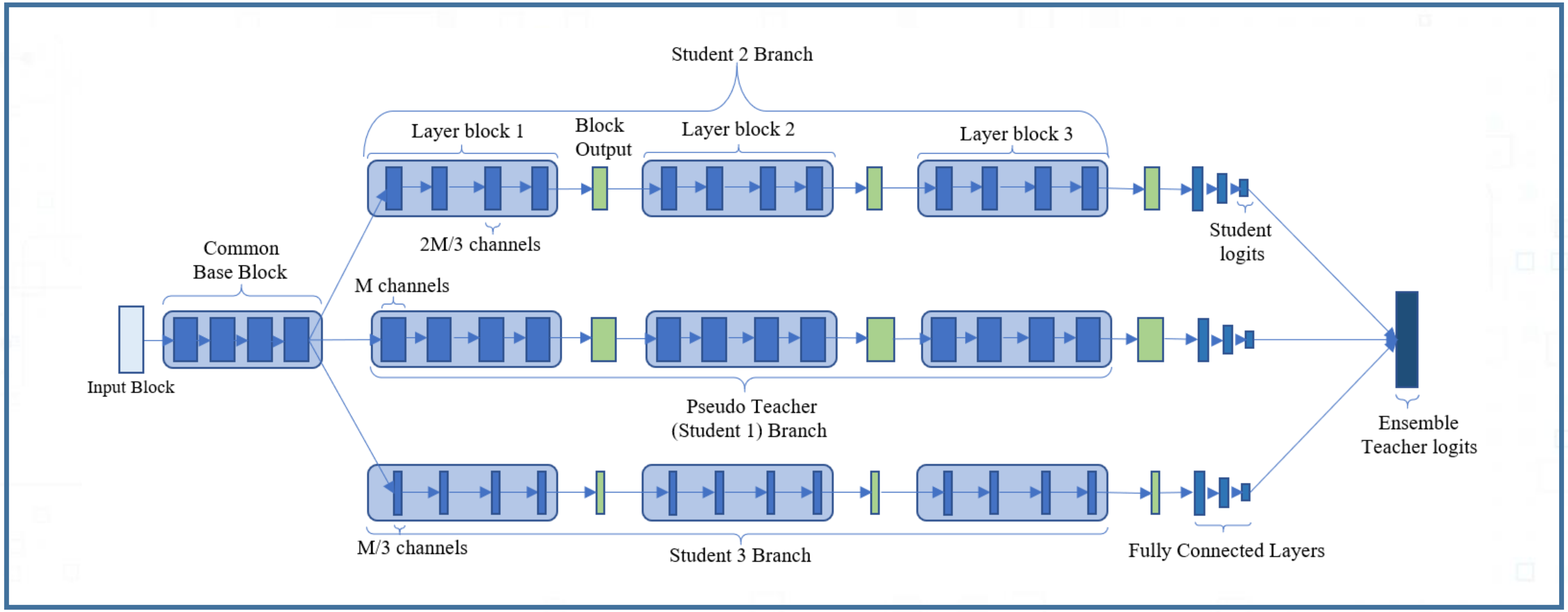


Ensemble Knowledge Distillation

Pre-trained teacher free and Model architecture agnostic approach

Overview of our model compression framework for a 3-student ensemble:

1. Each student is composed of the base block and one of the network branches on top of it.
2. The original model is the first student in the ensemble, termed as pseudo teacher. The ensemble teacher is a weighted combination of all student's output logits.
3. The layer channels for the compressed student branches are reduced by a specific ratio (here M, 2M/3, M/3) with respect to the original model's M channels.



Simultaneous multi-compressed ensemble model training

We incorporate a combination of three separate losses which facilitates final output distribution and intermediate layer knowledge Distillation from pseudo teacher to each ensemble student

1] Traditional Cross Entropy Loss

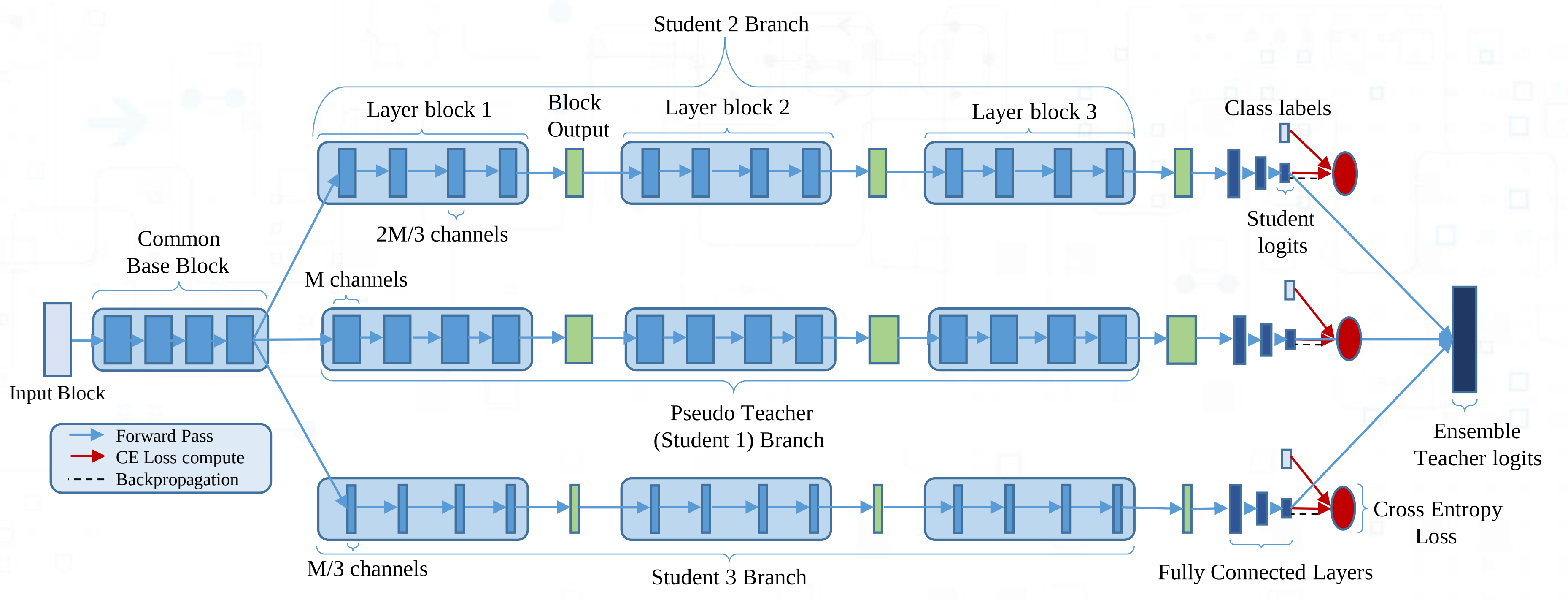
$$X_{ijk} = \frac{exp(x_{ijk})}{\sum_{k=1}^{C} exp(x_{ijk})}$$

$$L^{Normal} = \sum_{i=1}^{S}\sum_{j=1}^{N}\sum_{k=1}^{C} -\mathbb{1}_{jk}\log(X_{ijk})$$

i, j, k : Student, Batch, Class Index

X_{ijk} : Respective Student SoftMax Output

$\mathbb{1}_{jk}$: One hot label Indicator



Traditional Cross Entropy Loss Computation

Simultaneous multi-compressed ensemble model training

2] Knowledge Distillation Loss

$$X_{ijk} = \frac{exp(\frac{x_{ijk}}{T})}{\sum_{k=1}^{C} exp(\frac{x_{ijk}}{T})}$$

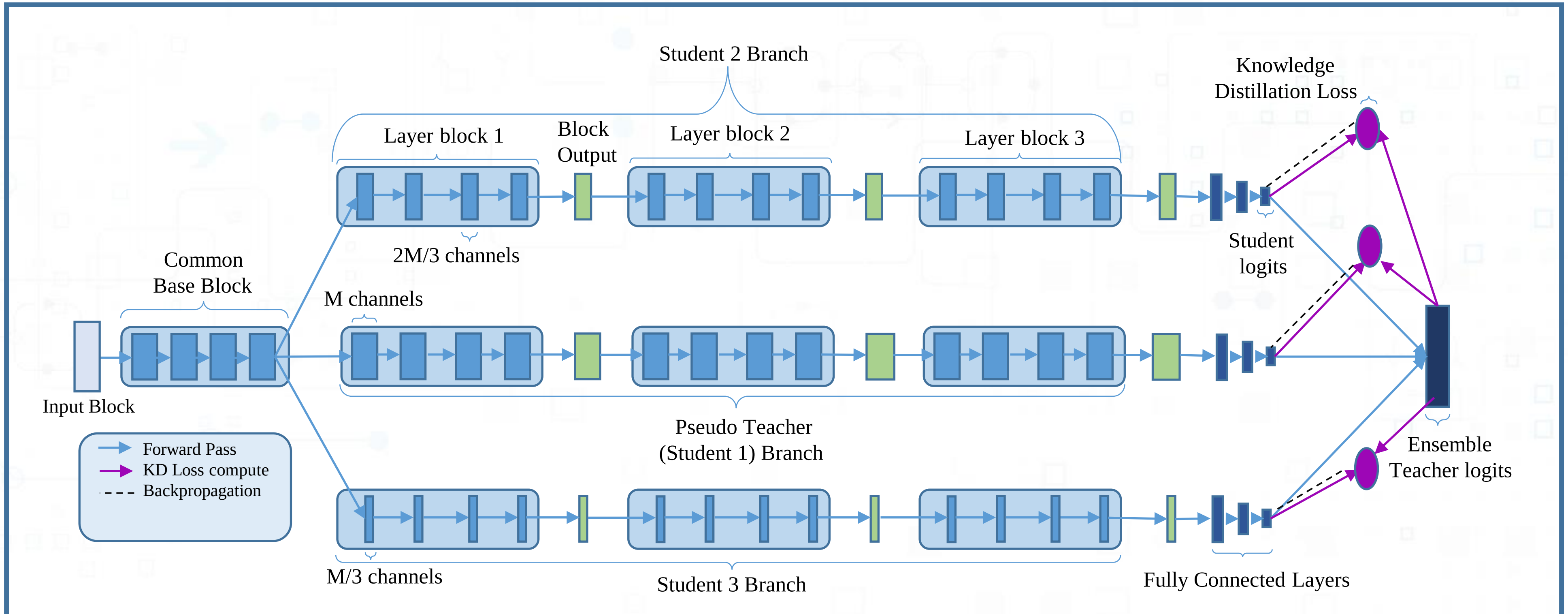
$$L^{KD} = \sum_{i=1}^{S}\sum_{j=1}^{N}\sum_{k=1}^{C} X^T_{jk} \log\left(\frac{X^T_{jk}}{X_{ijk}}\right)$$

i, j, k : Student, Batch, Class Index

X_{ijk} : Respective Student SoftMax Output

X^T_{jk} : Ensemble Teacher SoftMax Output

T : SoftMax Softening Temperature



Knowledge Distillation Loss Computation

Simultaneous multi-compressed ensemble model training

3] Intermediate Loss

$$l_{block}^{intermediate} = \sum_{l=2}^{S}\left(\sum_{m=1}^{N}(|x_m^{PT} - x_m^{l}|)^2\right)$$

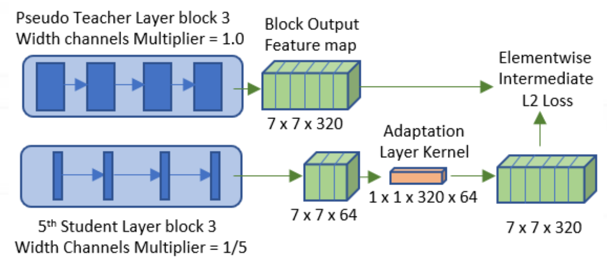
$$L^{intermediate} = \sum_{b=1}^{B} l_b^{intermediate}$$

m, l : Batch, Student Index

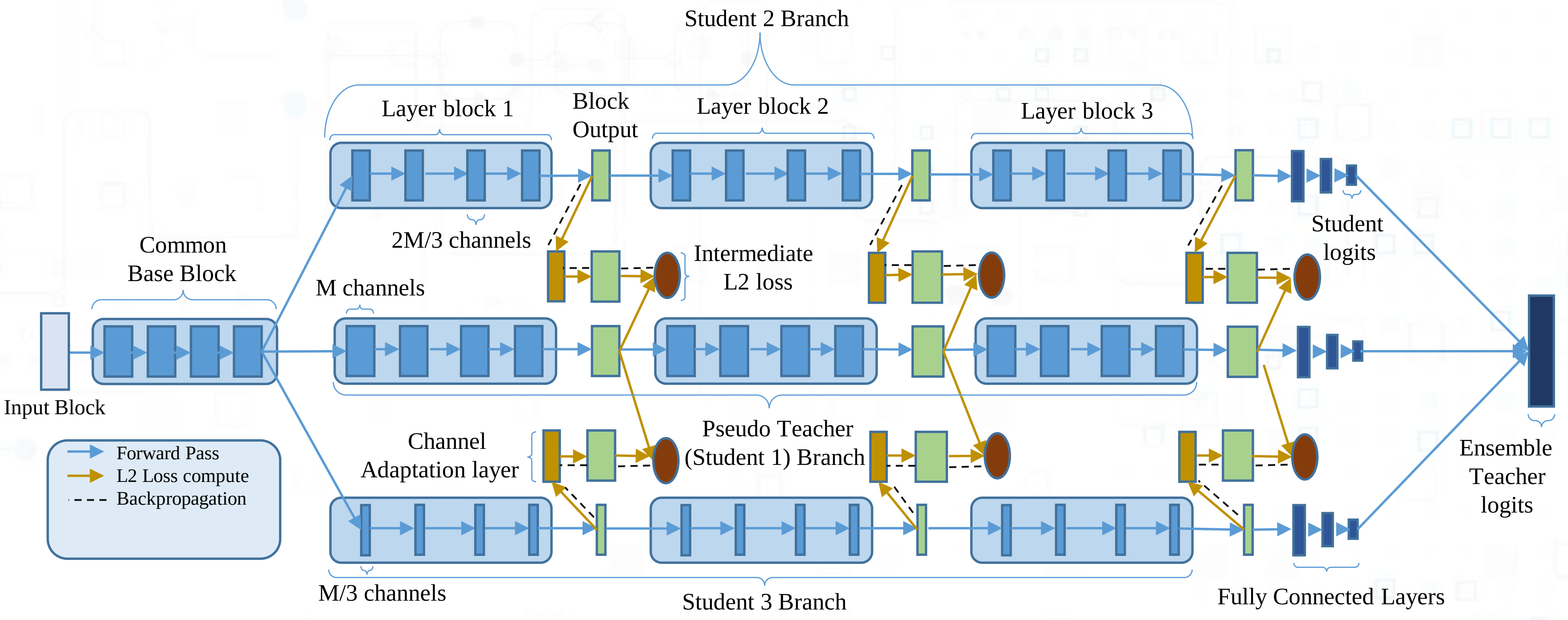
x_m^l : Feature map of size H x W x C corresponding to student 1 and batch sample m

x_m^{PT} : Pseudo Teacher feature map for batch sample m

b : Student Layer Block Index



Student Channel Adaptation Technique



Intermediate Loss Computation

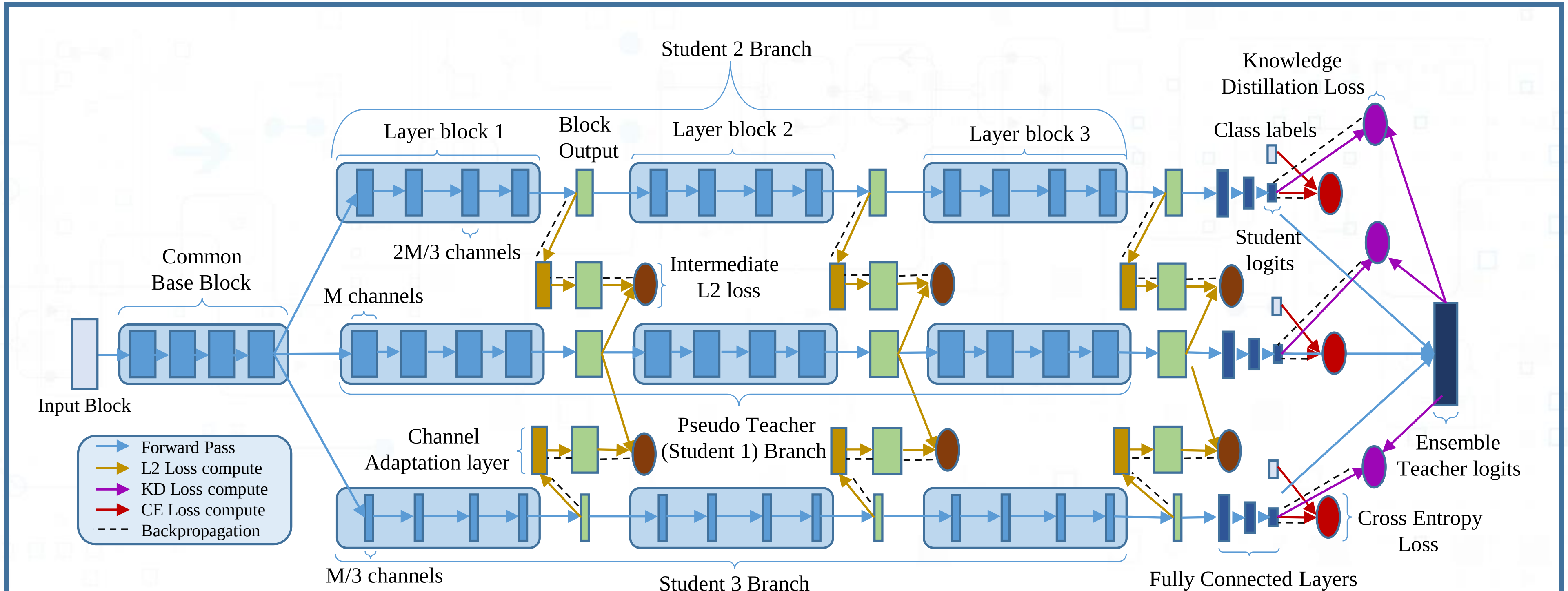
ECCV'20 ONLINE
23-28 AUGUST 2020

Simultaneous multi-compressed ensemble model training

4] Overall Loss

$$L = \alpha L^{Normal} + \beta L^{intermediate} + \gamma L^{KD}$$

Contribution ratio values found through ablation studies
α = 0.7
β = 0.15
γ = 0.15



Overall Loss Computation

Consistent performance gains for all models over their baseline individual training

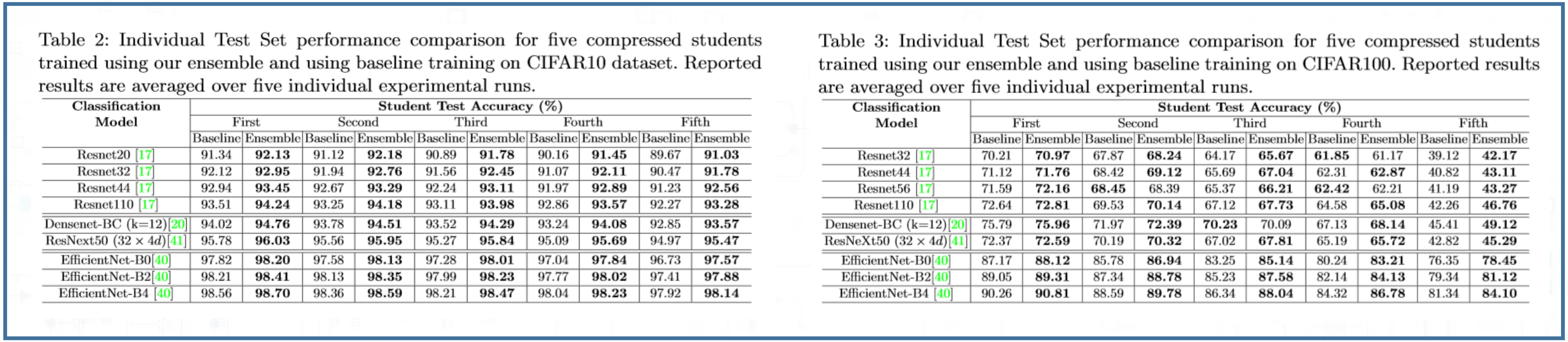


Table 2: Individual Test Set performance comparison for five compressed students trained using our ensemble and using baseline training on CIFAR10 dataset. Reported results are averaged over five individual experimental runs.

Classification Model	Student Test Accuracy (%)									
	First		Second		Third		Fourth		Fifth	
	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble
Resnet20 [17]	91.34	**92.13**	91.12	**92.18**	90.89	**91.78**	90.16	**91.45**	89.67	**91.03**
Resnet32 [17]	92.12	**92.95**	91.94	**92.76**	91.56	**92.45**	91.07	**92.11**	90.47	**91.78**
Resnet44 [17]	92.94	**93.45**	92.67	**93.29**	92.24	**93.11**	91.97	**92.89**	91.23	**92.56**
Resnet110 [17]	93.51	**94.24**	93.25	**94.18**	93.11	**93.98**	92.86	**93.57**	92.27	**93.28**
Densenet-BC (k=12)[20]	94.02	**94.76**	93.78	**94.51**	93.52	**94.29**	93.24	**94.08**	92.85	**93.57**
ResNext50 (32 × $4d$)[41]	95.78	**96.03**	95.56	**95.95**	95.27	**95.84**	95.09	**95.69**	94.97	**95.47**
EfficientNet-B0[40]	97.82	**98.20**	97.58	**98.13**	97.28	**98.01**	97.04	**97.84**	96.73	**97.57**
EfficientNet-B2[40]	98.21	**98.41**	98.13	**98.35**	97.99	**98.23**	97.77	**98.02**	97.41	**97.88**
EfficientNet-B4 [40]	98.56	**98.70**	98.36	**98.59**	98.21	**98.47**	98.04	**98.23**	97.92	**98.14**

Table 3: Individual Test Set performance comparison for five compressed students trained using our ensemble and using baseline training on CIFAR100. Reported results are averaged over five individual experimental runs.

Classification Model	Student Test Accuracy (%)									
	First		Second		Third		Fourth		Fifth	
	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble
Resnet32 [17]	70.21	**70.97**	67.87	**68.24**	64.17	**65.67**	**61.85**	61.17	39.12	**42.17**
Resnet44 [17]	71.12	**71.76**	68.42	**69.12**	65.69	**67.04**	62.31	**62.87**	40.82	**43.11**
Resnet56 [17]	71.59	**72.16**	**68.45**	68.39	65.37	**66.21**	**62.42**	62.21	41.19	**43.27**
Resnet110 [17]	72.64	**72.81**	69.53	**70.14**	67.12	**67.73**	64.58	**65.08**	42.26	**46.76**
Densenet-BC (k=12)[20]	75.79	**75.96**	71.97	**72.39**	**70.23**	70.09	67.13	**68.14**	45.41	**49.12**
ResNeXt50 (32 × $4d$)[41]	72.37	**72.59**	70.19	**70.32**	67.02	**67.81**	65.19	**65.72**	42.82	**45.29**
EfficientNet-B0[40]	87.17	**88.12**	85.78	**86.94**	83.25	**85.14**	80.24	**83.21**	76.35	**78.45**
EfficientNet-B2[40]	89.05	**89.31**	87.34	**88.78**	85.23	**87.58**	82.14	**84.13**	79.34	**81.12**
EfficientNet-B4 [40]	90.26	**90.81**	88.59	**89.78**	86.34	**88.04**	84.32	**86.78**	81.34	**84.10**

17 : He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
20 : Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
40 : Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
41 : Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)

Consistent performance gains for all models over their baseline individual training

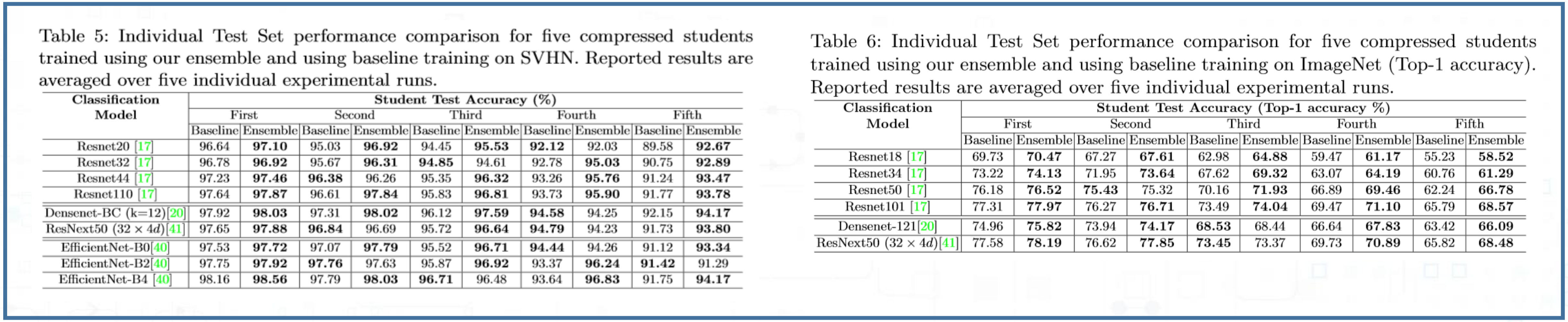


Table 5: Individual Test Set performance comparison for five compressed students trained using our ensemble and using baseline training on SVHN. Reported results are averaged over five individual experimental runs.

Classification Model	Student Test Accuracy (%)									
	First		Second		Third		Fourth		Fifth	
	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble
Resnet20 [17]	96.64	**97.10**	95.03	**96.92**	94.45	**95.53**	**92.12**	92.03	89.58	**92.67**
Resnet32 [17]	96.78	**96.92**	95.67	**96.31**	**94.85**	94.61	92.78	**95.03**	90.75	**92.89**
Resnet44 [17]	97.23	**97.46**	**96.38**	96.26	95.35	**96.32**	93.26	**95.76**	91.24	**93.47**
Resnet110 [17]	97.64	**97.87**	96.61	**97.84**	95.83	**96.81**	93.73	**95.90**	91.77	**93.78**
Densenet-BC (k=12)[20]	97.92	**98.03**	97.31	**98.02**	96.12	**97.59**	**94.58**	94.25	92.15	**94.17**
ResNext50 ($32 \times 4d$)[41]	97.65	**97.88**	**96.84**	96.69	95.72	**96.64**	**94.79**	94.23	91.73	**93.80**
EfficientNet-B0[40]	97.53	**97.72**	97.07	**97.79**	95.52	**96.71**	**94.44**	94.26	91.12	**93.34**
EfficientNet-B2[40]	97.75	**97.92**	**97.76**	97.63	95.87	**96.92**	93.37	**96.24**	**91.42**	91.29
EfficientNet-B4 [40]	98.16	**98.56**	97.79	**98.03**	**96.71**	96.48	93.64	**96.83**	91.75	**94.17**

Table 6: Individual Test Set performance comparison for five compressed students trained using our ensemble and using baseline training on ImageNet (Top-1 accuracy). Reported results are averaged over five individual experimental runs.

Classification Model	Student Test Accuracy (Top-1 accuracy %)									
	First		Second		Third		Fourth		Fifth	
	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble
Resnet18 [17]	69.73	**70.47**	67.27	**67.61**	62.98	**64.88**	59.47	**61.17**	55.23	**58.52**
Resnet34 [17]	73.22	**74.13**	71.95	**73.64**	67.62	**69.32**	63.07	**64.19**	60.76	**61.29**
Resnet50 [17]	76.18	**76.52**	**75.43**	75.32	70.16	**71.93**	66.89	**69.46**	62.24	**66.78**
Resnet101 [17]	77.31	**77.97**	76.27	**76.71**	73.49	**74.04**	69.47	**71.10**	65.79	**68.57**
Densenet-121[20]	74.96	**75.82**	73.94	**74.17**	**68.53**	68.44	66.64	**67.83**	63.42	**66.09**
ResNext50 ($32 \times 4d$)[41]	77.58	**78.19**	76.62	**77.85**	**73.45**	73.37	69.73	**70.89**	65.82	**68.48**

17 : He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
20 : Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
40 : Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
41 : Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)

Ensemble Teacher performance comparison with other literature methods

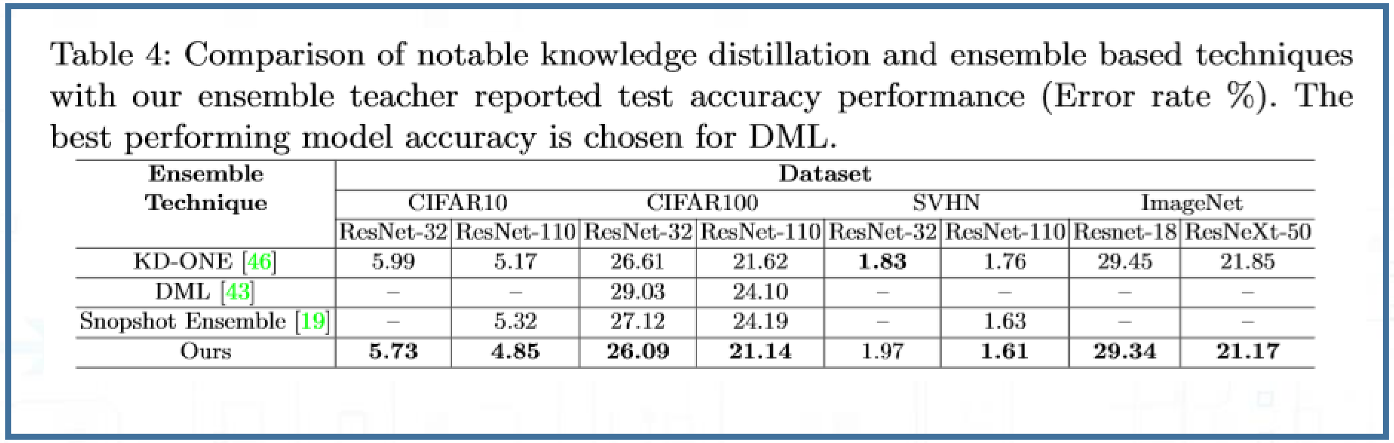


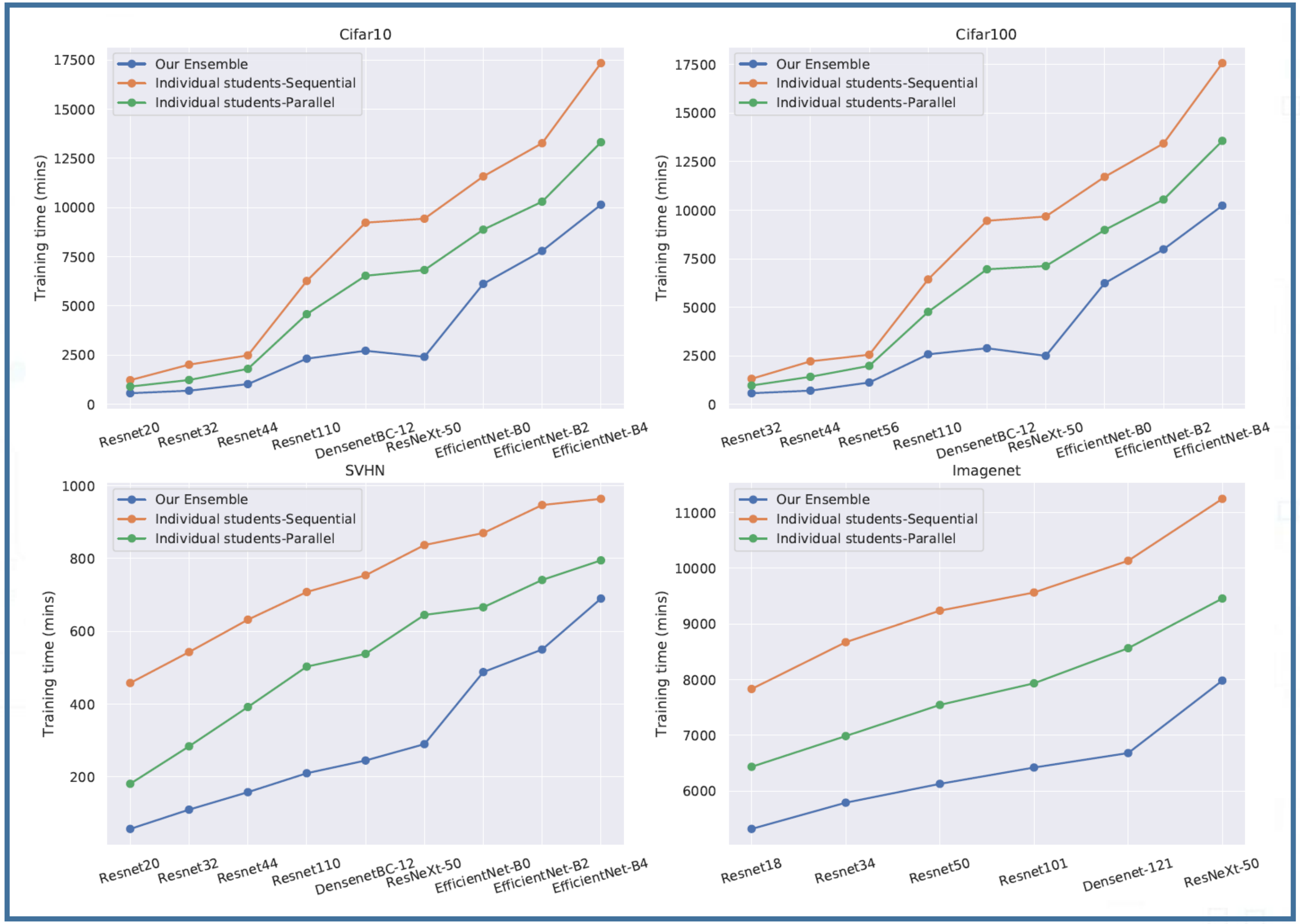
Table 4: Comparison of notable knowledge distillation and ensemble based techniques with our ensemble teacher reported test accuracy performance (Error rate %). The best performing model accuracy is chosen for DML.

Ensemble Technique	**Dataset**							
	CIFAR10		CIFAR100		SVHN		ImageNet	
	ResNet-32	ResNet-110	ResNet-32	ResNet-110	ResNet-32	ResNet-110	Resnet-18	ResNeXt-50
KD-ONE [46]	5.99	5.17	26.61	21.62	**1.83**	1.76	29.45	21.85
DML [43]	–	–	29.03	24.10	–	–	–	–
Snopshot Ensemble [19]	–	5.32	27.12	24.19	–	1.63	–	–
Ours	**5.73**	**4.85**	**26.09**	**21.14**	1.97	**1.61**	**29.34**	**21.17**

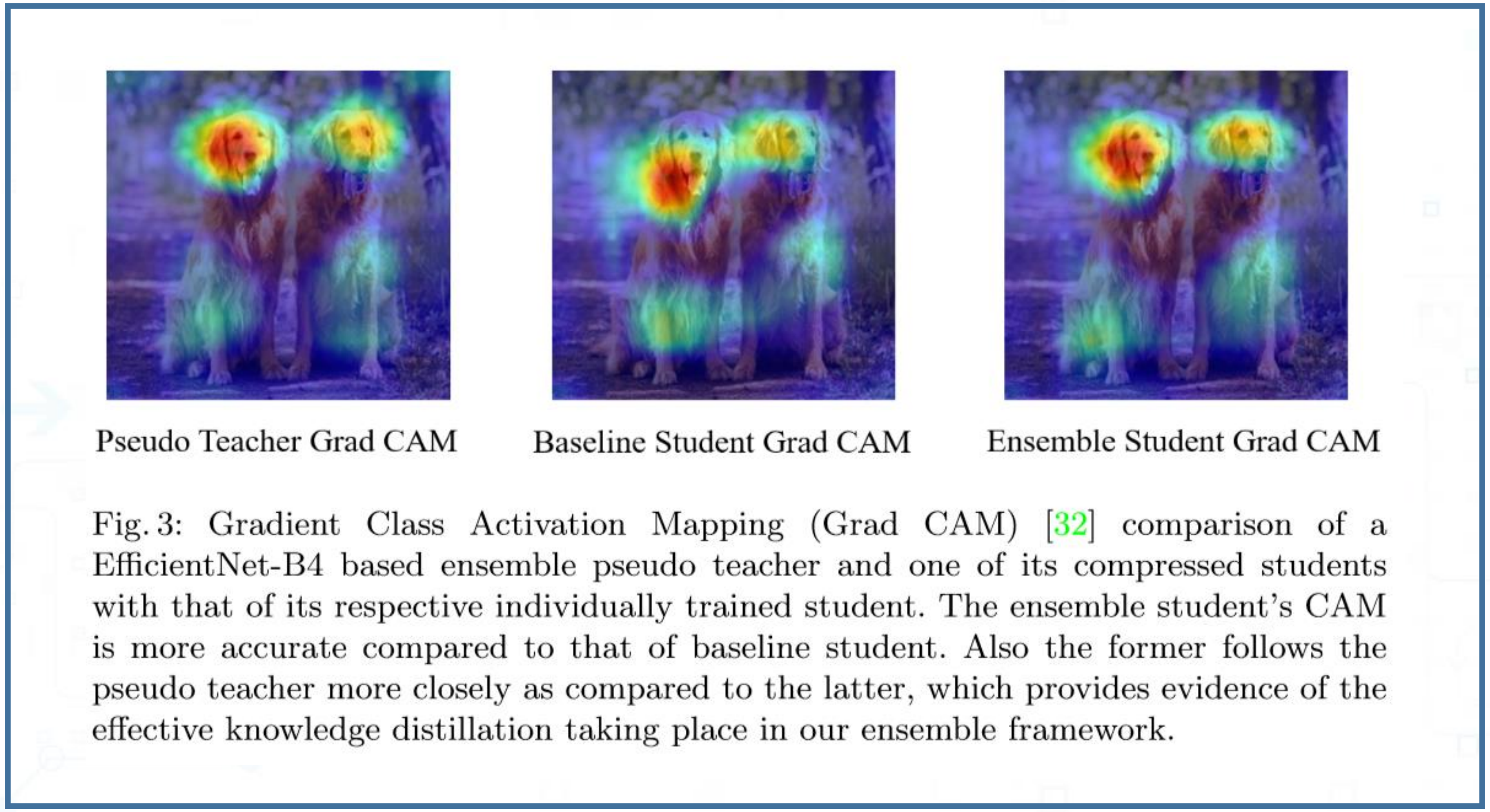
19 : Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., Weinberger, K.Q.: Snapshot ensembles: Train 1, get m for free. arXiv preprint arXiv:1704.00109 (2017)
43 : Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4320–4328 (2018)
46 : Zhu, X., Gong, S., et al.: Knowledge distillation by on-the-fly native ensemble. In: Advances in neural information processing systems. pp. 7517–7527 (2018)

ECCV'20 ONLINE
23-28 AUGUST 2020

Training time savings over training each student model either sequentially or in parallel



Effective Knowledge Distillation property of our ensemble framework



Pseudo Teacher Grad CAM

Baseline Student Grad CAM

Ensemble Student Grad CAM

Fig. 3: Gradient Class Activation Mapping (Grad CAM) [32] comparison of a EfficientNet-B4 based ensemble pseudo teacher and one of its compressed students with that of its respective individually trained student. The ensemble student's CAM is more accurate compared to that of baseline student. Also the former follows the pseudo teacher more closely as compared to the latter, which provides evidence of the effective knowledge distillation taking place in our ensemble framework.

Conclusion:

1. We present a novel model compression technique using an ensemble knowledge distillation learning procedure without requiring the need of any pretrained weights.

2. It manages to provide multiple efficient versions of a given model, compressed to varying degree without making any major manual architecture changes on the user's part.

3. Comprehensive experiments conducted using a variety of current state-of-the-art classification models and academic datasets provide substantial evidence of the framework's effectiveness.

4. Substantial training time gains are achieved using our framework compared to individual model training either sequentially or in parallel.

Thank You