

Supplementary: Deep Knowledge Distillation using Trainable Dense Attention

Bharat Bhushan Sau*¹
sau.bharatbhusan@gmail.com

Soumya Roy*²
meetsoumyaroy@gmail.com

Vinay P. Namboodiri³
vpn22@bath.ac.uk

Raghu Seshu Iyengar¹
bm15resch11003@iith.ac.in

¹ Indian Institute of Technology
Hyderabad, India

² Indian Institute of Technology
Kanpur, India

³ University of Bath,
England

In this supplementary report, we provide additional details of attention module construction and training hyperparameters along with the results of the experiments conducted to analyse the effects of different components of dense attention module.

1 Design choices of Attention module

We have used the following configurations for designing the attention module:

1. **More importance to higher level attention:** Reduction rate defines the number of output channels in the encoder part of a multi-channel spatial attention module. In a convolutional network, initial/lower layers learn local features (like edges and corners) and later/higher layers learn global features (like contours). In order to give more importance to the later layers (i.e., higher level attention), we keep a higher reduction rate (i.e., higher number of encoder output channels) for the higher attention modules than for the lower ones. For example, if we divide the network into 4 blocks (where block₁ is at a initial-level than block₄), block₁, block₂, block₃ and block₄ have reduction rates of $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$ and $\frac{1}{2}$ respectively.
2. **Kernel size in encoder:** Kernel size decides the complexity of attention encoder. To learn better attentional information, we use a kernel of size 3×3 , which captures inter-dependency of neighboring pixels better than a 1×1 kernel. However, we can increase kernel size and complexity of the encoder further and thus get even more effective positional information.
3. **Activation function:** We apply activation on the output of the encoder before providing it as input to the decoder. We observe that different activation functions have different effects on the learning of transferable attentional information. Due to its smooth gating functionality, Swish [8] is more effective than other commonly used activations like ReLU [9] and sigmoid etc.

*Bharat Bhushan Sau and Soumya Roy have contributed equally.

4. **Number of layers in encoder and decoder:** We use only one convolutional layer in the encoder and decoder modules to demonstrate that, even at minimal complexity, it is possible to learn very effective and transferable attentional features.

2 Additional Experiments

2.1 Effect of increase in accuracy of teacher on training student:

A teacher network, equipped with dense attention module, has slightly higher accuracy than that without dense attention. In section 4.2 of the main paper, we have claimed that this slight difference in teacher’s accuracy have very little effect on student improvement.

In order to validate this, we use ResNet50 with dense attention (ResNet50-DATN) as teacher and MobileNet as student for methods like AT [10] and Margin-ReLU [10]. Results are shown in Table 1. Here, we see that the accuracy of MobileNet improves very little even when we use ResNet50-DATN as teacher. We have other results in support of our claim:

1. In Table 7 (Deep teacher to ResNet18) in main paper, we see that the absolute difference in Top-1 accuracy between ResNet152 (Top-1 error: 21.69) and ResNet34 (Top-1 error: 26.69) is 5, yet student (ResNet18) trained by them has lower improvement in accuracy. For AT [10], there is degradation of accuracy. For Margin-ReLU [10], improvement is an absolute 0.91. For Dense-ATN, it is an absolute 1.02. This shows that the rate of improvement in student accuracy is much lower compared to the rate of improvement in teacher accuracy. This is because the student model has much lower capacity compared to teacher, hence its rate of improvement is much lower compared to teacher.
2. In Table 3 (Results on Places365) and Table 4 (Results on CUB) in main paper, difference in accuracy of teacher with and without dense attention is negligible. Yet, our method gives significant improvement over previous state-of-the-art methods.

Method	Teacher				Student	
	ResNet50		ResNet50-DATN		MobileNet	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Baseline(no transfer)	-	-	-	-	30.78	11.08
AT[10]	23.84	7.14	-	-	30.44	10.67
AT[10]	-	-	22.73	6.40	30.26	10.48
Margin-ReLU[10]	23.84	7.14	-	-	28.75	9.66
Margin-ReLU[10]	-	-	22.73	6.40	28.42	9.51
Dense-ATN	-	-	22.73	6.40	25.93	8.14

Table 1: Results of MobileNet (student) when using ResNet50 with dense attention (ResNet50-DATN) as teacher: Slight increment in accuracy of the same teacher network has little effect in increasing accuracy of student network.

2.2 Multiplicative Attention is more transferable than Additive Attention

Output of the attention-decoder module can be multiplied with feature maps (like in Eqn. 3 in main paper), or it can be added up like in Double Attention [10]. We term the first type of

attention as *multiplicative* attention, and the later type as *additive* attention.

Here, we experimentally show that multiplicative attention (like Dense-ATN) is more transferable than additive attention. We have tried the following additive variants:

1. Direct Addition (No sigmoid): Output of decoder is added to feature maps directly, i.e., output of Eqn. 2 (in main paper) without sigmoid is added to x_i .
2. Sigmoid-Addition: Sigmoid function is applied on output of decoder and then added to feature maps, i.e., output of Eqn. 2 (in main paper) is added to x_i .
3. Double Attention [10]: It is an additive attention and is useful for increasing accuracy of a network in stand-alone fashion.

From the results shown in Table 2, we hypothesize that, multiplication of attention-maps with feature maps is more helpful to learn transferable attention maps than doing addition.

Attention variant	Teacher (ResNet50)		Student (MobileNet)	
	Top-1	Top-5	Top-1	Top-5
Direct Addition (No sigmoid)	23.31	6.89	26.85	8.63
Sigmoid-Addition	23.55	6.72	26.75	8.43
Double Attention [10]	23.12	6.59	28.02	9.33
Sigmoid-Multiplication (Ours)	22.73	6.40	25.93	8.14

Table 2: Multiplicative vs Additive Attention: Multiplication of sigmoid attention is more transferable than additive attention.

2.3 Improvement in Teacher-Student Similarity:

In knowledge distillation, knowledge learnt by teacher network helps the student network to improve its accuracy. In other words, knowledge transfer method is meant to increase similarity between teacher and student output distributions, thus increasing its accuracy as by-product. Hence, we also report the similarity score between teacher and student in order to measure the effectiveness of the knowledge transfer method. One standard way of measuring similarity between two output distributions is KL-divergence. Lower the value of KL-divergence, better the similarity. In Table 3, we report KL-divergence between MobileNet (student) and ResNet50 (teacher). We compare our method against Margin-ReLU [10], which gives best result among existing methods on ResNet50-MobileNet pair.

Method	KL-Div with Teacher	Top-1 Error(%)
Baseline	0.491	30.78
Margin-ReLU[10]	0.362	28.75
Dense-ATN	0.321	25.93

Table 3: KL-divergence between MobileNet(student) and ResNet50(teacher) probability outputs.

2.4 Determining reduction rate of attention blocks:

We keep a higher reduction rate for final attention block compared to the internal blocks, in order to reduce overfitting in internal layers. For example, if reduction rate of the final block

is $\frac{1}{4}$, then for internal layers it will be $\frac{1}{8}$. Here, we experiment with different reduction rates. Results are shown in Table 4. We see that, when reduction rate of the final block is $\frac{1}{2}$ (and for internal blocks it is $\frac{1}{4}$), the network is able to learn the best transferable attention maps.

Reduction Rate	Teacher (ResNet50)		Student (MobileNet)	
	Top-1	Top-5	Top-1	Top-5
1/4	22.79	6.42	26.20	8.36
1/2	22.73	6.40	25.93	8.14
1.0	23.63	6.30	26.04	8.16

Table 4: Effect of reduction rate: first column represents reduction rate of final attention module. Hence, we fix reduction rate = $\frac{1}{2}$ for all other experiments.

3 Training Hyperparameters

3.1 Teacher training:

In most knowledge distillation methods, the teacher network is not modified. In this work, we modify the original network with dense multi-channel spatial attention modules. The modified network is trained using the standard hyperparameters for that dataset. For example, in case of ImageNet, the teacher network is trained from scratch for 100 epochs with the standard learning hyperparameters. In case of CUB [8], the teacher network is initialized with ImageNet weights and trained for 100 epochs.

It should be noted that the addition of attention blocks does not significantly improve the accuracy of teacher network, rather they act as better knowledge extractors for training the student network.

3.2 Student training:

For ImageNet, Places365 and CUB dataset, we train the network for 100 epochs, use a batch size of 256 and employ Stochastic Gradient Descent (SGD) as the optimizer. The initial learning rate (lr) is 0.1 and the learning rate is decayed by 0.1 after every 30 epochs. Initial value of β is 1000 and decayed by 0.25 at every 30 epochs.

For smaller subset of ImageNet, i.e. 10% of ImageNet dataset, we follow the hyperparameters used in [8]. The network is trained for 200 epochs, with initial learning rate of 0.1, and lr_decay of 0.1 at 140, 160 and 180 epochs. β value is also decayed by 0.25 whenever learning rate is decayed.

For CIFAR100 [9] dataset, the network is trained for 200 epochs, with initial learning rate of 0.1, and lr_decay of 0.1 at 100 and 150 epochs. Initial value of β is 1000 and decayed by 0.30 whenever learning rate is decayed.

For data augmentation we use horizontal flips and random crops. In all cases, attention module is placed at the end of a residual block. Reduction rate for these modules are: 1/4, 1/4, 1/4 and 1/2 respectively.

3.3 Choice of β value:

β is loss multiplier for attention (see Eqn. 7 in main paper). It is important to mention that the β value needs to be decayed. We start with a β value of 1000 and use a β decay rate of 0.25 for all experiments. The β value is decayed whenever there is a learning rate decay. If we don't use β decay, then at later stages of training, the student network overfits to the teacher attention maps, thus increasing the generalization error. This strategy is also followed in [10].

References

- [1] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A²-nets: Double attention networks. In *Advances in Neural Information Processing Systems*, pages 352–361, 2018.
- [2] Byeongho Heo, Jeessoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. *arXiv preprint arXiv:1904.01866*, 2019.
- [3] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.
- [4] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.
- [5] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.
- [6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
- [7] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.
- [8] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485, 2019.