On Compressing U-net Using Knowledge Distillation

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

We study the use of knowledge distillation to compress the U-net architecture. We show that, while standard distillation is not sufficient to reliably train a compressed U-net, introducing other regularization methods, such as batch normalization and class re-weighting, in knowledge distillation significantly improves the training process. This allows us to compress a U-net by over 1000x, i.e., to 0.1% of its original number of parameters, at a negligible decrease in performance.

1 Introduction

Deep learning has recently achieved ground-breaking results on several computer vision tasks, yielding highly effective architectures for many common vision tasks, such as image recognition [13] (VGG16, VGG19), human pose estimation [10] (Stacked Hourglass Networks) and image segmentation [12] (U-net). However, the increasing number of parameters of these models restricts their deployability in resource-constrained environments, such as on mobile devices. As a consequence, several lines of research have emerged to train more compact networks that achieve a performance similar to the popular, large ones.

In particular, most neural network compression approaches fall in three broad categories: weight quantization [11, 16], architecture pruning [7, 14, 1] and knowledge distillation [2, 5, 4]. While the first two aim to reduce the size of a given large model by either limiting the number of bits used to represent each parameter, or by removing some of its units or layers, the third, which we investigate here, seeks to train a compact model (student) using knowledge acquired by the larger one (teacher). The first distillation method was introduced by Caruna et al. [2] who focused on the multi-layer perceptron case with an RMSE-based error metric. The concept was then popularized by Hinton et al. [5] with an approach exploiting the teacher's predicted probabilities to train the student and by Romero et al. [4] who proposed to further exploit the teacher's intermediate representations to guide the student. Since then, several application-driven knowledge distillation strategies have been developed, e.g., for face model identification [9], object detection [3] and face verification [15].

In this paper, we study the use of the knowledge distillation technique of [5] to compress a U-net architecture for biomedical image segmentation. We first show that, without performing any distillation, the number of parameters of the U-net can be reduced drastically at virtually no loss of accuracy. We then observe that a direct application of knowledge distillation is insufficient to further compress the U-net, and propose to complement distillation with batch normalization and class re-weighting. As evidenced by our experiments, this allows us to reduce the U-net size to 0.1% of its original number of parameters at only a negligible loss of segmentation accuracy.

2 Methodology

U-net Architecture. The U-net architecture, initially introduced in [12] and depicted by Fig. 1, is a fully convolutional network with skip connections, comprising a contracting path and an expansive path. It relies on a channel depth of 64 at the first level and doubles it in 4 consecutive stages, reaching 1024 at the bottom level. This is then reduced back to 64 by the expansive part.
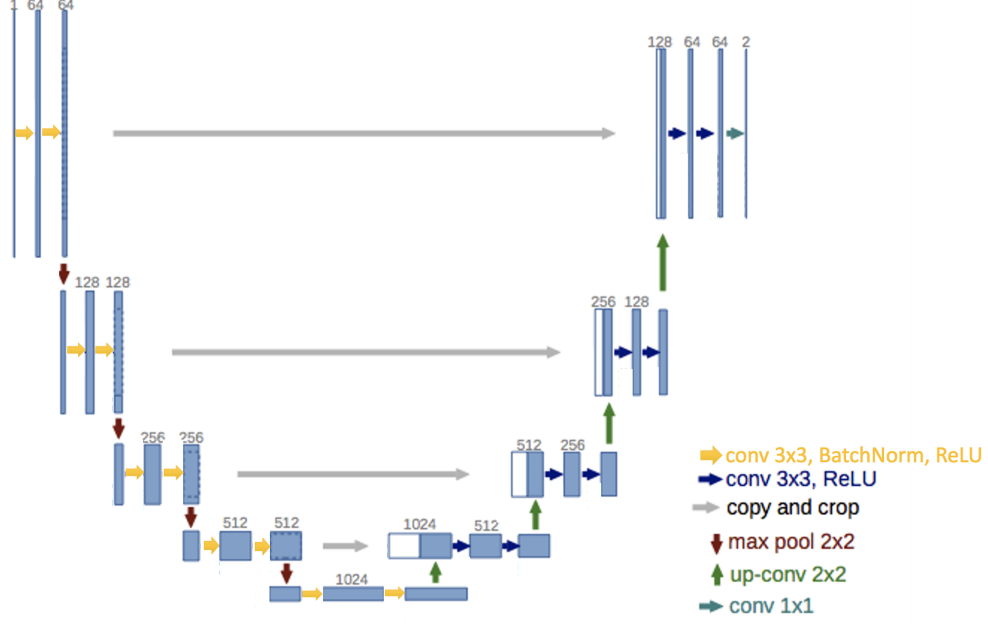
Figure 1: U-net architecture. Each blue box corresponds to a multi-channel feature map, with the number of channels indicated on top of the box. White boxes represent copied feature maps. Arrows of different colors denote different operations. The golden arrows correspond to our modified version of the original U-net of [12], where we added batch normalization.

Starting Channel Depth	64	16	4	2
Test Loss	0.1021	0.0822	0.0974	**0.8227**
Training Loss	0.0256	0.0220	0.0286	**0.3423**
Number of Iterations	80,000	150,000	300,000	290,000

Table 1: Performance of original U-net architectures (without batch normalization) when varying the channel depth in the first layer. The number of iterations corresponds to the minimum test loss.

Knowledge Distillation. Let us consider a binary image segmentation problem, where the input is a $W \times H$ image $\mathbf{x}$ and the output a binary $W \times H$ label map $\mathbf{y}$. Given training pairs $(\mathbf{x}^t, \mathbf{y}^t)$, the training procedure of a U-net makes use of the cross-entropy loss

$$\mathcal{H}(\mathbf{y}^t, \hat{\mathbf{y}}^t) = -\sum_{(i,j)} \mathbf{y}^t_{ij} \log(\hat{\mathbf{y}}^t_{ij}) + (1 - \mathbf{y}^t_{ij}) \log(1 - \hat{\mathbf{y}}^t_{ij}) \;,$$

where a $\mathbf{y}^t_{ij}$ indicates the label of sample t at location (i, j), and $\hat{\mathbf{y}}^t$ is the probability map predicted by the network. In the standard setting, this probability map is obtained via the softmax function. As discussed in [5], however, softer probabilities can be obtained by increasing the temperature T of this softmax. Distillation is then achieved by training a teacher network at a temperature $T > 1$ to generate probabilities $\hat{\mathbf{y}}^{t*}$ for each sample $\mathbf{x}^t$ in a validation set. These probabilities are then employed to train the student network using the cross-entropy $\mathcal{H}(\mathbf{y}^{t*}, \hat{\mathbf{y}}^t)$. This cross-entropy can be used either on its own, for vanilla distillation, or in conjunction with the original cross-entropy $\mathcal{H}(\mathbf{y}^t, \hat{\mathbf{y}}^t)$, for mixed distillation. After being trained at $T > 1$, the softmax temperature of the student network is reduced back to 1.

Improving Distillation. As evidenced by our results in Section 3, using standard distillation, whether vanilla or mixed, did not prove sufficient to distill a standard U-net into a very small one. To overcome this, we therefore propose two modifications of the original strategy. First, as indicated in Fig. 1, we introduce batch normalization [6] operations in every convolution layer of the contracting path. Second, we re-weight the classes according to their proportions in the training set. Specifically, the contribution of each foreground pixel in the loss is multiplied by a weight w_f equal to the ratio of the number of background pixels over the number of foreground ones. In practice, because there

2

T_{transfer}	Test Loss	Training Loss (Hard)	Training Loss (Soft)
2	0.490	0.268	0.167
5	0.505	0.267	0.239
10	0.506	0.268	0.305
15	0.507	0.270	0.326
20	0.507	0.271	0.334

Table 2: Performance of the 2-Unet using mixed distillation with class re-weighting (without batch normalization). The training soft losses are reported after multiplying by $T_{transfer}^2$ to adjust for scale.

Network	# trainable parameters	IoU Score	Cross Entropy Loss
Original 64-Unet	31,042,434	0.804	0.102
Original 4-Unet	122,394	**0.807**	**0.097**
Our 2-Unet (soft loss only)	**30,902**	0.752	<u>0.134</u>
Our 2-Unet (mixed distillation)	30,902	<u>0.759</u>	0.135

Table 3: Comparison of our method with a 64-Unet. Note that we achieve a competitive Intersection over Union Score to the original 64-Unet with only **0.1%** of its capacity.

are many more background pixels, this weight is larger than 1, i.e., $w_f = 17.8$. The contribution of the background pixels is kept unchanged, i.e., $w_b = 1.0$. These two modifications, i.e., batch normalization and class re-weighting, are performed on the student network only, to further help overcoming the general difficulty in training shallow networks. The former reduces the internal covariate shift and the latter combats the inherent class imbalance. As shown below, these two modifications allowed us to significantly reduce the necessary U-net size for accurate segmentation.

3 Experiments

Dataset. We use the Electron Microscopy (EM) Mitochondria Segmentation dataset of [8]. It contains a 5x5x5 micrometer section taken from the CA1 hippocampus region of the brain, corresponding to 1065x2048x1536 voxels of resolution approximately 5x5x5nm. This volume is separated into training and testing sub-volumes, each of which consists of 165 slices. We treat each slice as an image and aim to produce the corresponding binary label map, indicating the presence or absence of a mitochondria at the each location.

Compressing without Distillation. First, we experiment with reducing the number of channels in the first U-Net layer while keeping the doubling trend of the contracting path without any distillation. As shown in Table 1, a U-net with only 4 initial channels (4-Unet) achieves a similar test loss to the original 64-Unet. In the remaining experiments, we use the 4-Unet as teacher network for distillation.

Exploiting Standard Distillation. We then tried to make use of the standard distillation procedure of [5] to train a 2-Unet from a 4-Unet. To this end, we evaluated both vanilla distillation, mixed distillation and sequential distillation, where we started training with soft labels and then finished with hard ones, or vice-versa. All these attempts were unsuccessful, even with class re-weighting. This is illustrated by Table 2 for mixed distillation at different temperatures.

Distilling Our Modified U-net. Finally, we evaluate the use of distillation with our modified U-net that incorporates batch normalization and class re-weighting. Note that for all the distillation experiments, the teacher is the original 4-Unet trained from scratch with hard training loss. As shown in Table 3 for both vanilla distillation and mixed distillation, a 2-Unet achieves segmentation accuracies similar to those of a standard 64-Unet, while requiring only about 0.1% of its capacity. Note that training a 2-Unet without distillation but with our modifications yields a training loss of 0.265 and a test loss of 0.307, which shows that distillation is still required.

4 Conclusion

We have introduced a modified distillation strategy to compress a U-net architecture by over 1000x while retaining an accuracy close to the original U-net. This was achieved by modifying the U-net to incorporate batch normalization and class re-weighting. In the future, we plan to investigate the use of these modifications to perform distillation of other networks and for other application domains.

References

[1] Alvarez, J.M. and Salzmann, M., 2017. Compression-aware training of deep networks. In *Advances in Neural Information Processing Systems* (pp. 856-867).

[2] Buciluǎ, C., Caruana, R. and Niculescu-Mizil, A., 2006, August. Model compression. In Proceedings of the *12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 535-541). ACM.

[3] Chen, G., Choi, W., Yu, X., Han, T. and Chandraker, M., 2017. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems* (pp. 742-751).

[4] Han, S., Pool, J., Tran, J. and Dally, W., 2015. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems* (pp. 1135-1143).

[5] Hinton, G., Vinyals, O. and Dean, J., 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv*:1503.02531.

[6] Ioffe, S. and Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv*:1502.03167.

[7] LeCun, Y., Denker, J.S. and Solla, S.A., 1990. Optimal brain damage. In *Advances in neural information processing systems* (pp. 598-605).

[8] Lucchi, A., Smith, K., Achanta, R., Knott, G. and Fua, P., 2012. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE transactions on medical imaging*, 31(2), pp.474-486.

[9] Luo, P., Zhu, Z., Liu, Z., Wang, X. and Tang, X., 2016, February. Face Model Compression by Distilling Knowledge from Neurons. In proceedings of *Association for the Advancement of Artificial Intelligence* (pp. 3560-3566).

[10] Newell, A., Yang, K. and Deng, J., 2016, October. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision* (pp. 483-499). Springer, Cham.

[11] Rastegari, M., Ordonez, V., Redmon, J. and Farhadi, A., 2016, October. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision* (pp. 525-542). Springer, Cham.

[12] Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

[13] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556.

[14] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015

[15] Wang, C., Lan, X. and Zhang, Y., 2017. Model Distillation with Knowledge Transfer from Face Classification to Alignment and Verification. *arXiv preprint arXiv*:1709.02929.

[16] Zhu, C., Han, S., Mao, H. and Dally, W.J., 2016. Trained ternary quantization. *arXiv preprint arXiv*:1612.01064.