Deep Elastic Networks with Model Selection for Multi-Task Learning: Supplementary Material

Chanho Ahn*
Dept. of ECE and ASRI
Seoul National University
mychahn@snu.ac.kr

Eunwoo Kim*
Department of Engineering Science
University of Oxford
ekim@robots.ox.ac.uk

Songhwai Oh
Dept. of ECE and ASRI
Seoul National University
songhwai@snu.ac.kr

1. Details of Hierarchical Structure

The estimator of the proposed method can produce multiple network models of different sizes based on the hierarchical structure in a block. To control the actual speed-up for inference, each hierarchy accesses a different number of channels in each convolution layer. The ratio of the required number of channels for each level can be adjusted. As shown in Figure 1, the lowest level of hierarchy is represented and it accesses only a few channels. The highest level of hierarchy contains all channels in the figure. If the block is based on a residual block [2], the lowest level does not include any channels.

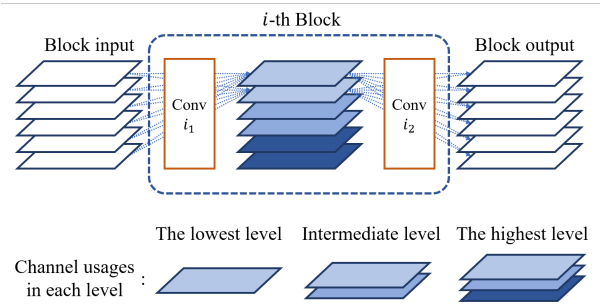

Figure 1. An example of the hierarchical structure in a block (the lowest level of hierarchy is shown as by dotted lines). Each hierarchical structure of a block contains different numbers of channels in the layer such that the lower level of hierarchy uses less channels and higher level uses more channels. The number of convolution filters used at each level depends on the channel usage.

2. Ablation Studies

We evaluated the performance depending on the number of levels in each block or depending on the initial model distribution. We used WRN-32-4 [4] as a backbone network and the CIFAR-100 dataset [3].

*Indicates equal contribution

First, we tested the performance on varying numbers of levels. The number of candidate models increases greatly as the number of levels increases, while the size of the selector is held fixed (the number of candidate models is h^n, where h and n are the number of levels and blocks, respectively). Figure 2 shows that the larger the number of levels, the smaller the network size can be found as exploring a larger model space. The performance also improved incrementally until the number of levels is four. However, when the number of levels is five, the performance is degraded due to the failure on dealing with a number of candidate models.

Second, we verified the effect of the initial model distribution. We applied two other distributions to compare with the proposed model distribution as described in Section 3.2 in the main paper: uniform distribution (Uniform) and random distribution which is obtained from the untrained initial selector (Random). The initial model distribution was used for training the estimator in the initial stage. As observed in Figure 2, we can verify that learning the network with the proposed initial distribution shows the best performance. Using the other distributions in the initial stage, the accuracy of the initial stage converged to the 2 to 3 % lower value compared to our method. Our approach reveals high performance in the initial stage and this affects the overall performance in Figure 2-(b).

3. Model Distribution for Test Set

We describe the model distribution for the test set to verify that diverse models can be selected depending on given input instances. The proposed framework was trained on three datasets, CIFAR-100 [3], Tiny-ImageNet, and STL-10 [1], based on a backbone network, WRN-32-4 [4]. We designed the estimator to have 15 blocks each of which contains four levels of hierarchy. Figure 3 shows the histogram of different models which are used for instances in the test sets. We can observe variability of selected models and the distribution of chosen models is neither deterministic nor

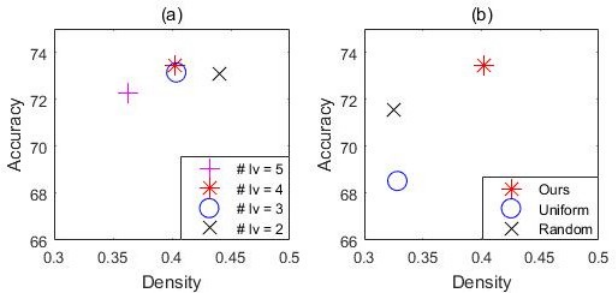

Figure 2. Two ablation studies: (a) performance on varying numbers of levels, and (b) performance with different model distributions. "# lv" is the number of levels in each block. "Uniform" and "Random" denote that the corresponding methods learn the estimator with a uniform model distribution and the random model distribution from the untrained selector in an initial stage, respectively.

uniform. We also calculated the average of probabilities that each level is selected over the test set. As shown in Figure 3-(b), the high values represent that the corresponding levels of hierarchy are frequently selected over the test set and there are common filters which are used for the most instances.

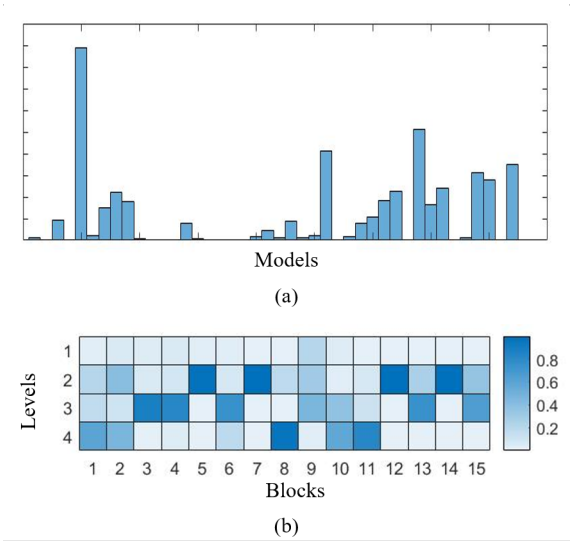

Figure 3. Evaluation of selected models using a network with 15 blocks and four hierarchical levels. (a) A histogram of models selected by the proposed algorithm on the test set. (b) The mean of probabilities that each level is selected by the proposed algorithm on the test set.

From the experiment, we have found that different models are selected by different groups of images. Examples of selected models and corresponding input images are shown in Figure 4. Three example models are shown in the figure: Model A, B, and C. Model A is selected for images with children and Model B is selected for images with people doing different activities. Note that Model A and B shares the same network architecture. Model C is selected for vessels. Similar groups are selected for Model A and Model B while the selected groups for Model C are different from Model A and B. We can see that each group is learned for specific features and the proposed selector explores appropriate groups for efficient inference.

References

[1] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, 2011.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URL: https://www.cs.toronto.edu/kriz/cifar.html (visited on Mar.1, 2016)*, 2009.

[4] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

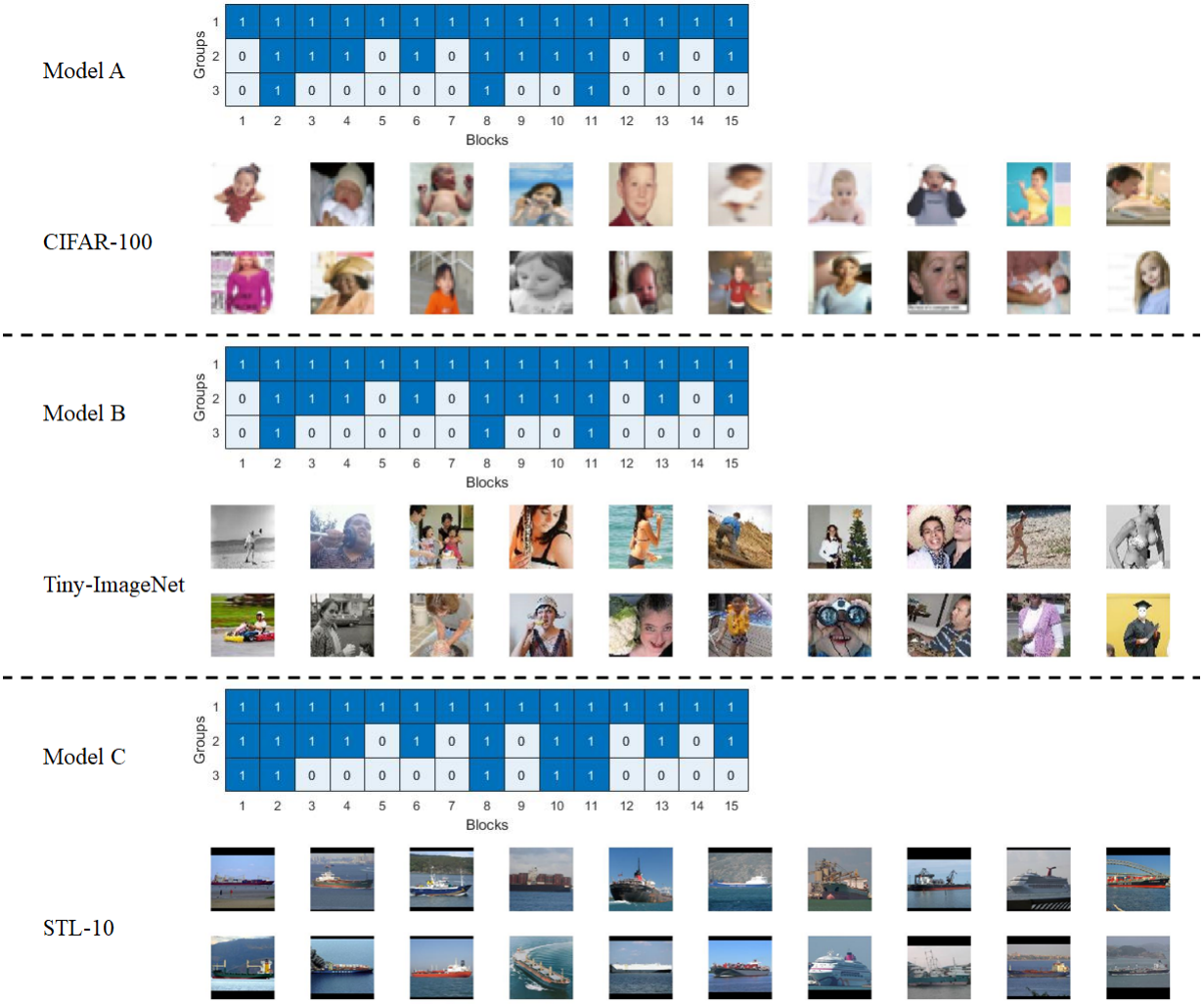

Figure 4. The most selected model structure for each dataset. The result is from the proposed framework jointly trained with three datasets: CIFAR-100, Tiny-ImageNet, and STL-10. "Model" represents convolution groups chosen by the proposed selector. Below the model, examples of input images which selected the model are shown.