A TRANSFER LEARNING AND PROGRESSIVE STACKING APPROACH TO REDUCING DEEP MODEL SIZES WITH AN APPLICATION TO SPEECH ENHANCEMENT

Sicheng Wang[1], Kehuang Li[1], Zhen Huang[1], Sabato Marco Siniscalchi[1,2], Chin-Hui Lee[1]

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. USA
[2]University of Enna Kore, 94100 Enna, Italy

{sichengwang, cl182}@gatech.edu, {kehlekernel, huangzhenee, siniscalchi77.19}@gmail.com

ABSTRACT

Leveraging upon transfer learning, we distill the knowledge in a conventional wide and deep neural network (DNN) into a narrower yet deeper model with fewer parameters and comparable system performance for speech enhancement. We present three transfer-learning solutions to accomplish our goal. First, the knowledge embedded in the form of the output values of a high-performance DNN is used to guide the training of a smaller DNN model in *sequential transfer learning*. In the second *multi-task transfer learning* solution, the smaller DNN is trained to learn the output value of the larger DNN, and the speech enhancement task in parallel. Finally, a *progressive stacking transfer learning* is accomplished through multi-task learning, and DNN stacking. Our experimental evidences demonstrate 5 times parameter reduction while maintaining similar enhancement performance with the proposed framework.

***Index Terms*—** Transfer learning, model compression, model stacking, multi-task training, speech enhancement

1. INTRODUCTION

While deep neural networks (DNNs) have achieved the state-of-the-art performances in many tasks, including automatic speech recognition (ASR) [1][2] speech enhancement [3][4], image classification [5] and object detection [6], the large-sized parameters in the models take up considerable memory for storage. The complex models also require a lot of time and power to perform matrix multiplication during prediction. These two factors pose key challenges to deploy such models on embedded systems, on which there is often limited memory, power, and bandwidth. Yet the interests in transferring the learned models to low-resource platforms is keen [7, 8, 9] due to the ubiquity of smart mobile devices. Most studies on this topic aim to reduce the size of the DNN while maintaining a similar accuracy or quality of prediction. For example, network analysis and transformation are attempted in [7, 8] to remove redundant parameters in image classification and handwritten digit recognition. A small-footprint ASR system is implemented on a cell phone in [9]. It is natural to extend these ideas to speech enhancement that could be embedded in local processing in mobile devices.

Transfer learning, or learning with “knowledge transfer” [10], is a machine learning paradigm that can play a key role in *model compression* [11]. Indeed, it focuses on sharing/transfer knowledge among/across different domains/tasks, and often uses knowledge learned previously to solve new problems faster or with better solutions. In [12], for example, a multi-task learning approach [13] exploits the knowledge acquired in a more general setting (phone classification) to address a more specific, and difficult task (senone classification). In maximum a posteriori (MAP) speaker adaptation [14], the key idea is to transfer knowledge from a source to a target model by condensing all information about the source domain into a prior density of the model parameters. This allows us to find the most probable model with respect to the target data under MAP. Other successful transfer learning applications are available in the literature, and the readers are referred to [10] for details.

In this paper, we use transfer learning as a viable and effective vehicle to compress the function that is learned by a large DNN into a much smaller neural architecture that has comparable performance and is faster to execute at run-time. In particular, we devise three different model compression solution based on transfer learning with the goal to reduce the size of a highly accurate, wide DNN into a narrower yet deeper DNN with a comparable accuracy. In the first approach, we sequentially transfer the generalisation ability of a cumbersome model (teacher) to a smaller model (student) by using the output values generated by the cumbersome model for the training set as “targets” for the smaller model. This approach is inspired by [11] and [15]. In [11], the original (often small) training set is used to train an ensemble of neural networks, which are in turn employed to label a large unlabeled data set. A single neural network is then trained on this much larger, ensemble labeled, data set. In our approach, we do not use an ensemble of neural networks as primary source of knowledge, and teacher and student DNNs are built on the same training data. In contrast to [15], we address a regression rather than a classification task. In the second solution, the student's architecture is modified by adding an auxiliary output layer to the original one. The auxiliary output layer is latched to the output of the teacher, and knowledge distillation is accomplished through a multi-task transfer learning approach [13]. A multi-stage transfer learning approach is finally devised by combining a progressive approach with multi-task learning. In doing so, we progressively increase the deepness of the student while reducing its complexity, in terms of size, and attaining a similar accuracy with the teacher.

We evaluate our proposed solutions on speech enhancement, which has attracted a large amount of research attention in recent years because of the growing challenges in many real-world applications, including mobile speech communication, hearing aids and robust ASR [16]. In these tasks, DNNs have been proven to significantly outperform other conventional techniques [4]. Our experimental evidence demonstrates that a 5 times parameter reduction can be achieved while maintaining a similar enhancement performance with the proposed framework.

2. RELATED WORK

2.1. Speech Enhancement with Deep Neural Networks

DNN-based regression to reconstruct clean speech magnitude spectrum is introduced in [4]. In the training phase, a large training set

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

of synthesized noisy speech is trained to minimize the mean square error between the log-power spectral (LPS) features of target clean speech and that of enhanced speech. During enhancement, the LPS features of noisy speech pass through the trained DNN to obtain the enhanced LPS features. Assuming that phase is not as sensitive to human perception, enhanced speech is synthesized from the predicted magnitude and the original noisy phase. DNNs are also stacked [17] to create multi-context networks in order to leverage the contextual information more thoroughly to predict ideal time-frequency masks and clean speech with good quality.

2.2. Transfer Learning

Transfer learning is a machine learning paradigm that relays the knowledge acquired in one task to another instead of training the second system independently. As an emerging field of active research, transfer learning could be further divided into many topics depending on the similarity between the source and target domains, the closeness of the source and target tasks, or number of target tasks [10]. When both the source and target domains are in a common feature space sharing similar marginal probability distributions, and output spaces are alike, it is known as homogeneous transfer learning [18]. In the case where the domain features are different (either in mismatched feature space or dissimilar probability distribution), or the task have changed, knowledge needs to be transferred across domain/tasks, hence heterogeneous transfer [10]. Depending on the number of tasks to be learned simultaneously during the transfer, one could categorize it into sequential or multi-task learning [13].

2.2.1. Teacher-student Networks

Prior efforts in transferring knowledge from a larger teacher network to a smaller student network include [15] in ASR and [19] in image classification. In both tasks, the teacher's knowledge is embedded in the posterior distribution at the output layer of the DNN. Supervised by the teacher's output, the student network adapts to the teacher and becomes imparted with the teacher's knowledge. In [15], the KL divergence between the two networks is minimized. The teacher could be further used to generate more labels of unseen data for the student. The authors were able to achieve 5.08% relative word error rate reduction with this student network than with a small network trained directly. The authors of [19] included additional hints from the teacher, which are inserted into the middle of the student network as intermediate supervision, resulting in a network of 10 times smaller. Both studies have demonstrated the effectiveness of homogeneous knowledge transfer when using learned target from more capable networks (in the sense of more parameters) to smaller ones for classification tasks. We wish to explore mechanisms to transfer knowledge from larger to smaller DNNs for speech enhancement in a regression task.

2.2.2. Network Transformation

The different architecture of the student network from its teacher naturally calls for the idea of network morphism. *Net2Net* is a concept developed in [20] that primarily transfer the knowledge of a learned network to a deeper or wider network. As a result, the student network is argued to be more potent with learning thanks to its larger size. The idea is further extended to *MorphNet* [21] to train a child network with knowledge transferred from its parent at a much faster rate. However, transferring knowledge from larger networks to smaller ones was not addressed in either work.

2.3. Compression of Neural Networks

During training, the redundancy of parameters in deep networks facilitates the searching for good local minimum of the loss function, provided that reasonable regularization are adopted [22]. Multitude of work have been attempted to compress these trained models. Authors of [11] trained compact models to mimic large ensembles of networks with negligible loss in performance by generating pseudo data and labeling these data with the ensembles. The idea of network surgery by removing less important weights was explored in [23] and [24]. More recently, the authors in [7, 25] extended the idea and achieved an overall of 30 to 50 times model compression on the ImageNet data set via pruning, quantization, and encoding. Matrix decomposition with low-rank estimation is employed in [26] generating about 1 to 7 times model compression on the same data set. However, since knowledge is condensed into a smaller model instead of being transferred, such techniques may be sensitive to variations in data distributions. Nonetheless, we consider these compression techniques as post-processing steps that could be applied to many networks, large or small, to achieve further reduction in model sizes.

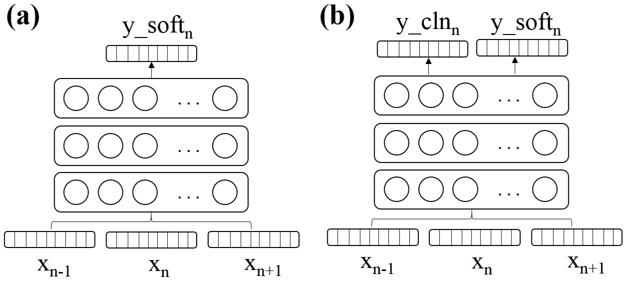

Fig. 1: (a) Supervision with only soft target under sequential learning. (b) Supervision by multi-tasks.

3. TRANSFER LEARNING IN NETWORK COMPRESSION

3.1. Sequential Transfer Learning

We first consider **sequential transfer learning** where both the source and target domains correspond to the same acoustic space [18]. The common task is to remove additive noise to recover the clean speech spectrum. To overcome the challenge that the smaller networks are hampered by reduced model sizes, we assume that the DNN outputs of the teacher network encapsulates the knowledge about the probability distribution of the acoustic features in the feature space, as in [15] yet for a regression task. By using the teacher's outputs as *soft targets* [27] referred to in this paper as opposed to the conventional *clean targets* (with a little abuse of terminology), we distill the knowledge acquired by a more complex model into a smaller one. This is illustrated in Fig.1(a).

3.2. Multi-task Transfer Learning

While the use of soft targets can allow us to transfer the knowledge acquired by a complex neural network into a less complex DNN for the speech enhancement task, the distortion in the teacher output values may hinder a proper deployment of the noise-to-clean mapping of the student. **Multi-task transfer learning** can address such an issue by retaining the access to the clean target speech. In this case, the student is forced to learn two tasks in parallel, namely predicting the clean and learning the teacher's noise-to-clean mapping. We assume that knowledge embedded in the teacher's output could aid learning for the primary task of reconstructing clean target speech, Y_{cln}. The soft targets, Y_{soft}, supplement as secondary targets, weighted by a

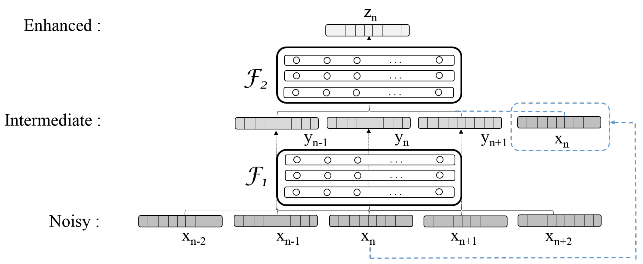

Fig. 2: Example of progressive transfer learning architecture. The noisy speech frame included is indicated with the dashed line.

factor of λ. The loss function L to train the student network is thus:

$$L = \frac{1}{2N}||Y_{cln} - P_{cln}||^2 + \frac{\lambda}{2N}||Y_{soft} - P_{soft}||^2 \qquad (1)$$

where P_{cln} and P_{soft} are the predictions of the clean and soft targets, respectively. N is the number of frames in mini-batch training in which the clean and soft targets are concatenated for the supervision of student network as shown in Fig.1(b).

3.3. Progressive Stacking Transfer Learning

With the trained models from Sec.3.2, we could further transfer knowledge to secondary networks to boost the overall performance. The outputs of the base networks exhibit representations that are very close to the targets. Thus, we select the outputs from the base networks with respect to the clean targets (its primary task) and feed them to the follow-up networks. The secondary networks are also trained under multi-task supervision. Since each stage needs not to finish the whole task, progressive learning warrants the use of simpler architectures (e.g., narrow hidden layers) for each stage. After training the secondary networks, the whole network is combined and fine-tuned to form a deeper yet narrower network. This process could be repeated to stack in more stages. An example with 2-stage networks and 3 frames in each stage is illustrated in Fig.2. Only the primary targets are shown for an ease of understanding.

However, one must be cautioned that the knowledge from the base network may not be entirely accurate and constructive, especially during aggressive compression using very narrow bases. The lesser learning capability of these narrow networks may introduce considerable distortions, usually referred to as *negative knowledge*[10]. There is no universally agreed approach to avoid such negative knowledge. We propose to counteract such artifacts by retaining the access to the original noisy signals for secondary networks during training. This is shown by the dashed line in Fig.2. In the case that only a small context window of 3 is used, the enhanced frame of a two-stage network could be formulated as

$$y_n = \mathcal{F}_1\big(x_{n-1}, x_n, x_{n+1}\big) \qquad (2)$$

$$z_n = \mathcal{F}_2\big(y_{n-1}, y_n, y_{n+1}, x_n\big) \qquad (3)$$

with $\mathcal{F}_i$ represents the network function of a single stage and y_n's represent the intermediate outputs from the first stage.

4. EXPERIMENTAL SETTINGS

Clean speech was selected from the TIMIT database [28], synthesized with 100 noise types [29] at 6 different SNR levels from -5dB to 20dB at a 5dB interval. The training, development, and *matched* test set contain 15, 1.5, and 0.8 hours of these multi-condition noisy speech respectively. We also created a 1.5-hour *mistmatched* test set using 15 unseen noise types [30] to evaluate the robustness of the DNNs. The optimization of DNN parameters follows the standard recipe described in [3]. The baseline system in [3] is used to as the teacher network to generate soft targets. Preliminary investigation shows that 0.1 to 1 is a reasonable range for the weight parameter λ. Hence, the value of 1 was used for multi-task training. In progressive transfer learning, a context window of 5 is supplied to the base networks. All subsequent stages used 3 context windows from previous stages, together with an original noisy speech frame. Multi-task training is used to supervise all network training. During fine-tuning, the initial learning rate is set to 10 times of that during the last iterations of training the secondary network.

Table 1: AVERAGE PESQ AND LSD BETWEEN VARIOUS NAIVELY TRAINED NETWORK.

Network	Matched		Mismatched	
	PESQ	LSD	PESQ	LSD
Original	2.31	5.60	2.34	6.91
Baseline	3.14	2.52	2.63	3.94
3×800	3.01	2.78	2.56	4.01
6×800	3.04	2.82	2.57	4.12
10×800	2.95	3.21	2.49	4.42

Two objective metrics are used to assess the quality of enhanced speech. Perceptual evaluation of speech quality (PESQ) [31] correlates closely with subjective perception of speech quality. Log spectral distortion (LSD), e.g., [32], which correlates more with the MMSE loss function, measures the distance between the spectra of the original and the corrupted signals.

5. EXPERIMENTS AND RESULT ANALYSIS

5.1. Direct Training without Knowledge Transfer Results

The baseline system [3] uses a 3-layer network with 2048 hidden nodes in each layer. In the effort to establish a thinner but deeper network of the same performance, we are able to train a narrow network of a moderate depth (6 layers, 800 hidden nodes). It performs better than a shallower network of equal width, but it could not be matched against the baseline, as shown in the left part of Table 1 with the match test set, attaining a PESQ score of 3.04 and an LSD of 2.82 dB. Our experiments further confirm that training a thinner and deeper network of 10 layers naively resulted in worse PESQ of 2.95 and an LSD of 3.21 dB shown in the bottom row, caused by the well-known vanishing gradient problems [33]. Similar trend is observed in the test for mismatched noise conditions. In the following discussions, the networks follow the nomenclature that $l \times m$ means a network of l hidden layers of m hidden nodes each.

5.2. Sequential Transfer Learning Results

Student DNNs learn from the soft targets as well as learning from clean targets, if not better. A student of moderate width (800 hidden nodes) of 3, 6, and 10 layers are selected to learn from the clean targets *OR* the soft targets. In Figure 3 we compare four DNN configurations, baseline 3x2048, 3x800, 6x800 and 10x800, as displayed in each set of the bar chart from left to right. The striped bars in all four sets correspond to sequential transfer learning which is comparable with the white bars representing direct learning as explained in Section 5.1. Evident from Fig.3(b), the networks trained under sequential learning with soft targets achieved much lower LSD than those trained with clean targets, while both approach obtain similar PESQ scores in 3(a). The benefit of using soft targets is particularly pronounced for deeper models (10× 800) which could not be trained

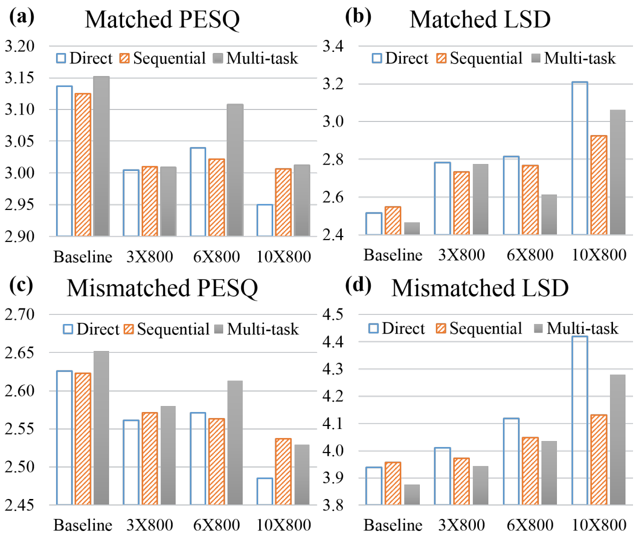

Fig. 3: PESQ and LSD of direct, sequential, and multi-task traininig on (a,b) matched and (c,d) mismatched test set.

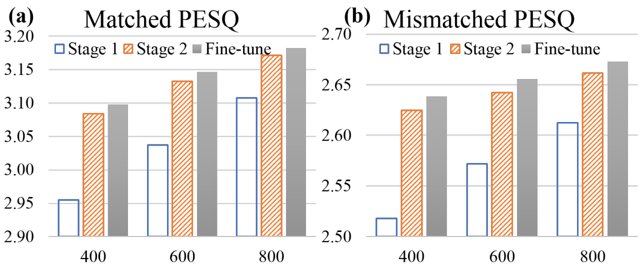

Fig. 4: Average PESQ of networks under progressive learning for (a) matched and (b) mismatched noise conditions.

Table 2: MODEL SIZE AND PERFORMANCE.

Stage	Network	# Params	reduction	PESQ
Baseline	3 × 2048	11.5M	1x	2.63
Stage 1	6 × 800	4.4M	2.6x	2.61
	6 × 600	2.7M	4.2x	2.57
	6 × 400	1.4M	8.1x	2.52
Stage 2	3 × 800	6.7M	1.7x	2.66
	3 × 600	4.2M	2.7x	2.64
	3 × 400	2.3M	5.1x	2.63
Fine tune	3 × 800	6.7M	1.7x	2.67
	3 × 600	4.2M	2.7x	2.66
	3 × 400	2.3M	**5.1x**	**2.64**

well with conventional means. Same observation could be made for the mismatched test set shown in Fig.3 (c) and (d).

5.3. Multi-task Transfer Learning Results

Still in Fig.3, the solid bars represent the PESQ and LSD of networks trained under multi-tasks. When DNN of an equal size as the teacher is trained under multi-task learning, its performance is even superior to the teacher's, manifested as higher PESQ and lower LSD of the solid bars. This suggests the merit of distilled knowledge in the teacher's soft targets. For student DNNs with width of 800, we could observe an increase in PESQ and drop of LSD in all architectures by comparing the solid and the white bars in Fig.3. In fact, the 6-layer network have PESQ score (2.61) very close to the baseline (2.63) for the mismatched noise condition. Nevertheless, the depth of the DNN is still limited to fewer than 10 layers.

5.4. Progressive Transfer Learning Results

Progressive knowledge transfer as shown in Figure 2 allows us to grow the depth of the student DNNs and to potentially boost their performances. In the following experiments, the base networks are all with 6 layers and the subsequent stages are with 3 layers each. Additional contexts together with the noisy speech frames are fed into the subsequent stages, since wider contextual information has proven to be beneficial to DNN-based speech enhancement [4, 17]. However, the narrow networks often could not accommodate the large contexts in the input layer, as substantial reduction in dimension from the input to the hidden layer result in irreversible loss of information. Progressive transfer learning circumvents this problem by using small contexts at each stage. The knowledge from the lower networks is further contextualized with its adjacent frames to be processed by the secondary networks. This way, knowledge is further summarized and relayed to the top layers.

Large gains in PESQ were witnessed after two stages, especially for thinner networks, as shown in the left bar chart set in Fig.4. For example, the PESQ score goes from 2.52 for width 400 in the white bar (indicating Stage 1) to the striped bar of 2.63 (representing Stage 2). This could be attributed to partial knowledge transferred from the base to the top networks. Stage-wise training allows stacking thin networks that matches the performance of a wide network. Fine tuning also gives slight PESQ gains in all cases as indicated by the solid bars.

5.5. Model Reduction Analysis

Finally, we analyze the model compression results, at maintaining a reasonable PESQ score for the mismatched noise tests, when compared with the baseline system. As shown in Table 2 the two-stage network of 400 hidden nodes with fine-tuning in the bottom row has achieved 5 times of parameter reduction from 11.5M to 2.3M parameters at a marginal PESQ gain from 2.52 to 2.64. It has 10 layers (6 hidden, 1 intermediate output, 3 hidden) in total, yet it outperforms that naively trained 10-layer network with 400 hidden nodes (at PESQ = 2.35) not shown in Table 1.

6. CONCLUSION AND FUTURE WORK

In this study, we propose a transfer knowledge scheme from a trained large DNN to a smaller network for a regression task. Three transfer approaches are proposed and compared. In multi-task transfer, the use of soft target provides additional guidance that compensates the inadequate learning capability of smaller networks. Progressive stacking provides effective means of incorporating large context of speech frames into narrow networks. Our experiments in speech enhancement find the smaller network could match the performance against wider ones by using much fewer parameters. The tests with both matched and mismatched noise conditions suggest that the proposed transfer learning scheme is effective even when the feature space changes or the underlying probability distribution shifts.

Future effort could include more quantitative evaluation of the quality of teacher's knowledge in an attempt to minimize negative knowledge. This would allow future researchers to set different weights to the teacher's output in multi-task learning. The quantification of knowledge would also allow us to assess the trade-off between knowledge retention and model size under the paradigm of lossy compression, paving ways to methodical selections of optimal model sizes to achieve acceptable system performances.

7. REFERENCES

[1] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[2] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[4] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

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

[6] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Advances in Neural Information Processing Systems*, 2013, pp. 2553–2561.

[7] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Advances in Neural Information Processing Systems*, 2015, pp. 1135–1143.

[8] W. Chen, J. T. Wilson, S. Tyree, K. Q Weinberger, and Y. Chen, "Compressing neural networks with the hashing trick," *CoRR, abs/1504.04788*, 2015.

[9] X. Lei, A. W. Senior, A. Gruenstein, and J. Sorensen, "Accurate and compact large vocabulary speech recognition on mobile devices.," in *INTERSPEECH*, 2013, pp. 662–665.

[10] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[11] C. Bucilă, R. Caruna, and A Niculescu-Mizil, "Model compression," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 535–541.

[12] Z. Huang, J. Li, S. M. Siniscalchi, I-F. Chen, J. Wu, and C.-H. Lee, "Rapid adaptation for deep neural networks through multi-task learning," in *Proc. Interspeech*, 2015.

[13] R. Caruna, "Multitask learning: A knowledge-based source of inductive bias," in *Proc. ICML*, 1993, pp. 41–48.

[14] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[15] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size dnn with output-distribution-based criteria.," in *INTERSPEECH*, 2014, pp. 1910–1914.

[16] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Boca Raton, FL, USA: CRC, 2013.

[17] X.-L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 967–977, 2016.

[18] Z. Huang, S. M. Siniscalchi, and C.-H. Lee, "A unified approach to transfer learning of deep neural networks with applications to speaker adaptation in automatic speech recognition," *in Neurocomputing, doi:10.1016/j.neucom.2016.09.0189*.

[19] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[20] T. Chen, I. Goodfellow, and J. Shlens, "Net2net: Accelerating learning via knowledge transfer," *arXiv preprint arXiv:1511.05641*, 2015.

[21] T. Wei, C. Wang, R. Rui, and C. W. Chen, "Network morphism," *arXiv preprint arXiv:1603.01670*, 2016.

[22] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[23] Y. LeCun, J. S. Denker, S. A. Solla, R. E. Howard, and L. D. Jackel, "Optimal brain damage.," in *NIPs*, 1989, vol. 2, pp. 598–605.

[24] B. Hassibi, D. G. Stork, and G. J. Wolff, "Optimal brain surgeon and general network pruning," in *Neural Networks, 1993., IEEE International Conference on*. IEEE, 1993, pp. 293–299.

[25] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding," *CoRR, abs/1510.00149*, vol. 2, 2015.

[26] Y.-D. Kim, E. Park, S. Yoo, T. Choi, L. Yang, and D. Shin, "Compression of deep convolutional neural networks for fast and low power mobile applications," *arXiv preprint arXiv:1511.06530*, 2015.

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

[28] J. Garofolo, "An acoustic phonetic continuous speech database," *Speech communication*, vol. 30, pp. 95–198, 2000.

[29] G. Hu, "100 nonspeech sounds," `http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html`, 2004.

[30] A. Varga and H. JM Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[31] ITU-T, Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Int. Telecommun. Union-Telecommun. Stand. Sector*, 2001.

[32] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., pp. 873–901. Springer, Berlin, 2008.

[33] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks.," in *Aistats*, 2010, vol. 9, pp. 249–256.