A VARIATIONAL BAYESIAN APPROACH TO LEARNING LATENT VARIABLES FOR ACOUSTIC KNOWLEDGE TRANSFER

Hu Hu[1], *Sabato Marco Siniscalchi*[1,2], *Chao-Han Huck Yang*[1], *Chin-Hui Lee*[1]

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

ABSTRACT

We propose a variational Bayesian (VB) approach to learning distributions of latent variables in deep neural network (DNN) models for cross-domain knowledge transfer, to address acoustic mismatches between training and testing conditions. Instead of carrying out point estimation in conventional maximum a posteriori estimation with a risk of having a curse of dimensionality in estimating a huge number of model parameters, we focus our attention on estimating a manageable number of latent variables of DNNs via a VB inference framework. To accomplish model transfer, knowledge learnt from a source domain is encoded in prior distributions of latent variables and optimally combined, in a Bayesian sense, with a small set of adaptation data from a target domain to approximate the corresponding posterior distributions. Experimental results on device adaptation in acoustic scene classification show that our proposed VB approach can obtain good improvements on target devices, and consistently outperforms 13 state-of-the-art knowledge transfer algorithms.

Index Terms— Variational inference, Bayesian adaptation, knowledge distillation, latent variable, device mismatch.

1. INTRODUCTION

Recent advances in machine learning are largely due to an evolution of deep learning combined with an availability of massive amounts of data. Deep neural networks (DNNs) have demonstrated state-of-the-art results in building acoustic systems [1, 2, 3]. Nonetheless, audio and speech systems still highly depend on how close the training data used in model building covers the statistical variation of the signals in testing environments. Acoustic mismatches, such as changes in speakers and recording devices, usually cause an unexpected and severe performance degradation [4, 5, 6]. For example, as for acoustic scene classification (ASC), device mismatch is an inevitable problem in real production scenarios [7, 8, 9, 10]. Moreover, the amount of data for the specific target domain is often not sufficient to train a good deep target model to achieve a similar performance to the source model. A key issue is to design an effective adaptation procedure to transfer knowledge from the source to target domains, while avoiding catastrophic forgetting and curse of dimensionality [11, 12, 13, 14, 15] often encountered in deep learning.

Bayesian learning provides a mathematical framework to model uncertainties and incorporate prior knowledge. It usually performs estimation via either maximum a posteriori (MAP) or variational Bayesian (VB) approaches. By leveraging upon target data and prior belief, a posterior belief can be obtained by optimally combining them. In the MAP solution, a point estimate can be obtained, which has been proven effective in handling acoustic mismatches in hidden Markov models (HMMs) [4, 16, 17] and DNNs [12, 18, 13] by assuming a distribution on the model parameters. On the other hand, the VB approach performs an estimation on the entire posterior distribution via a stochastic variational inference method [19, 20, 21, 22, 23, 24]. Bayesian learning can facilitate building an adaptive system for specific target conditions in a particular environment. Thus the mismatches between training and testing can be reduced, and the overall system performance is greatly enhanced.

Traditional Bayesian formulations usually impose uncertainties on model parameters, like Bayesian neural networks [25]. However, for commonly used DNNs, the number of parameters is usually much larger than the available training samples, making an accurate estimation difficult. Moreover, a feature based knowledge transfer framework, namely teacher-student learning (TSL, also called knowledge distillation) [26, 27] has been investigated in recent years. The basic TSL transfers knowledge acquired by the source / teacher model and encoded in its softened outputs (model outputs after softmax), to the target / student model, where the target model directly mimics the final prediction of the source model through a KL divergence loss. The idea is then extended to hidden embedding of intermediate layers [28, 29, 30], where different embedded representations are proposed to encode and transfer knowledge. However, instead of considering the whole distribution of latent variables, they only perform point estimation as in MAP, potentially leading to sub-optimal results and may lose distributional information.

In this work, we aim at establishing a Bayesian adaptation framework based on latent variables of DNNs, where the knowledge is transferred in the form of distributions of deep latent variables. Thus, a novel variational Bayesian knowl-

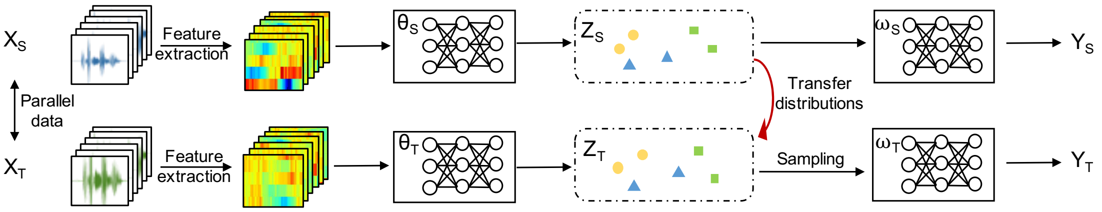


Fig. 1: Illustration of the proposed knowledge transfer framework.

edge transfer (VBKT) approach is proposed. We take into account of the model uncertainties and perform distribution estimation on latent variables. In particular, by leveraging upon variational inference, the distributions of the source latent variables (prior) are combined with the knowledge learned from target data (likelihood) to yield the distributions of the target latent variables (posterior). Prior knowledge from the source domain is thus encoded and transferred to the target domain, by approximating the posterior distributions of latent variables. An extensive and thorough experimental comparison against 13 recent cut-edging knowledge transfer methods is carried out. Experimental evidence demonstrates that our proposed VBKT approach outperforms all competing algorithms on device adaptation tasks of ASC.

2. BAYESIAN INFERENCE OF LATENT VARIABLES

2.1. Knowledge Transfer of Latent Variables

Suppose we are given some data observations $\mathcal{D}$, and let $\mathcal{D}_S = \{x_S^{(i)}, y_S^{(i)}\}_{i=1}^{N_S}$ and $\mathcal{D}_T = \{x_T^{(i)}, y_T^{(i)}\}_{i=1}^{N_T}$ indicate the source and target domain data, respectively. Our framework requires parallel data, e.g., for each target data sample $x_T^{(i)}$, there exists a paired data sample $x_S^{(j)}$ from the source data, where $x_T^{(i)}$ and $x_S^{(j)}$ share the same audio content but recorded by different devices. Consider a DNN based discriminative model with parameters λ to be estimated, where λ usually represents network weights. Starting from the classical Bayesian approach, a prior distribution $p(\lambda)$ is defined over λ, and the posterior distribution after seeing the observations $\mathcal{D}$ can be obtained by the Bayes Rule as follows,

$$p(\lambda|\mathcal{D}) = \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})}. \tag{1}$$

Figure 1 illustrates the overall framework of our proposed knowledge transfer approach. Parallel input features of X_S and X_T are firstly extracted and then fed into neural networks. In addition to network weights, we introduce the latent variables Z to model the intermediate hidden embedding of DNNs. Here Z refers to the unobserved intermediate representations, encoding transferable distributional information. We then decouple the network weights into two independent subsets, θ and ω, as illustrated by the subnets in the 4 squares in Figure 1, to represent weights before and after Z is generated, respectively. Thus we have

$$p(\lambda) = p(Z, \theta, \omega) = p(Z|\theta)p(\theta)p(\omega). \tag{2}$$

Note that the relationship in Eq. (2) holds for both prior $p(\lambda)$ and posterior $p(\lambda|\mathcal{D})$. Here we focus on transferring knowledge in a distribution sense via the latent variables Z. With parallel data, we thus assume that there exists Z retaining the same distributions across the source and target domains. Specifically, for the target model we have $p(Z_T|\theta_T) = p(Z_S|\theta_S, \mathcal{D}_S)$, as the prior knowledge learnt from the source encoded in $p(Z_S|\theta_S, \mathcal{D}_S)$.

2.2. Variational Bayesian Knowledge Transfer

Denoting λ in the target model as λ_T, typically, the posterior $p(\lambda_T|\mathcal{D}_T)$ is often intractable, and an approximation is required. In this work, we propose a variational Bayesian approach to approximate the posterior; therefore, a variational distribution $q(\lambda_T|\mathcal{D}_T)$ is introduced. For the target domain model, the optimal $q^*(\lambda_T|\mathcal{D}_T)$ is obtained by minimizing KL divergence between the variational distribution and the real one, over a family of allowed approximate distributions $\mathcal{Q}$:

$$q^*(\lambda_T|\mathcal{D}_T) = \underset{q \in \mathcal{Q}}{argmin}\, \mathrm{KL}(q(\lambda_T|\mathcal{D}_T) \parallel p(\lambda_T|\mathcal{D}_T)). \tag{3}$$

In this work, we focus on latent variables, Z, and we assume a non-informative prior over θ_T and ω_T. Next, by substituting Eqs. (1), (2), and the prior distribution into Eq. (3), we arrive, after re-arranging the terms, to the following variational lower bound $\mathcal{L}(\lambda_T; \mathcal{D}_T)$ as

$$\begin{aligned} \mathcal{L}(\lambda_T; \mathcal{D}_T) = \mathbb{E}_{Z_T \sim q(Z_T|\theta_T, \mathcal{D}_T)} \log p(\mathcal{D}_T|Z_T, \theta_T, \omega_T) \\ - \mathrm{KL}(q(Z_T|\theta_T, \mathcal{D}_T) \parallel p(Z_S|\theta_S, \mathcal{D}_S)). \end{aligned} \tag{4}$$

Simply put, a Gaussian mean-field approximation is used to specify the distribution forms for both the prior and posterior over Z. Specifically, each latent variable z in Z follows an M-dimension isotropic Gaussian, where M is the hidden embedding size. Given a parallel data set, we can approximate the KL divergence term in Eq. (4) by establishing a mapping of each pair of Gaussian distributions across domains via sample pairs. We denote the Gaussian mean and variance for the source and target domains as $\mu_S; \sigma_S^2$ and $\mu_T; \sigma_T^2$, respectively. A stochastic gradient variational Bayesian (SGVB) estimator [21] is then used to approximate the posterior, with the network hidden outputs being regarded as the mean of the Gaussian. Moreover, we assign a fixed value σ^2 to both σ_S^2 and σ_T^2, as the variance for all individual Gaussian components. We can now obtain a close-form solution for the KLD term in Eq. (4). Furthermore, by adopting Monte Carlo to generate N_T-pairs of sample, the lower bound in Eq. (4) can be approximated empirically as:

$$\mathcal{L}(\lambda_T;\mathcal{D}_T) = \sum_i^{N_T} \mathbb{E}_{z_T^{(i)}\sim\mathcal{N}(\mu_T^{(i)},\sigma^2)} \log p(y_T^{(i)}|x_T^{(i)}, z_T^{(i)}, \theta_T, \omega_T) - \frac{1}{2\sigma^2}\sum_i^{N_T} \|\mu_T^{(i)} - \mu_S^{(i)}\|_2^2, \qquad (5)$$

where the first term is the likelihood, and the second term is deduced from the KL divergence between prior and posterior of the latent variables. Each instance of $z_T^{(i)}$, is sampled from the posterior distribution as $z_T^{(i)}|\theta_T, \mathcal{D}_T \sim \mathcal{N}(\mu_T^{(i)}, \sigma^2)$, thus the expectation form in the first term can be reduced. To flow the gradients of sampling operation through deep neural nets, a reparameterization trick [31, 21] is adopted during the network training. In the inference stage, as it's a classification task, we directly take $z_T^{(i)} = \mu_T^{(i)}$ to simplify the computation.

3. EXPERIMENTS

3.1. Experimental Setup

We evaluate our proposed VBKT approach on the acoustic scene classification (ASC) task of DCASE2020 challenge task1a [32]. The training set contains ~10K scene audio clips recorded by the source device (device A), and 750 clips for each of the 8 target devices (Device B, C, s1-s6). Each target audio is paired with a source audio, and the only difference between the two audios is the recording device. The goal is to solve the device mismatch issue for one specific target device at a time, i.e., device adaptation, which is a common scenario in real applications. For each audio clip, log-mel filter bank (LMFB) features are extracted, and scaled to [0,1] before feeding into the classifier.

Two state-of-the-art models, namely: a dual-path resnet (RESNET) and a fully convolutional neural network with channel attention (FCNN), are tested according to the challenge results [32, 33]. We use the same models for both the source and target devices. Mix-up [34] and SpecAugment [35] are used in the training stage. Stochastic gradient descent (SGD) with a cosine-decay restart learning rate scheduler is used to train all models. Maximum and minimum learning rates are 0.1, and 1e-5, respectively. The latent variables are based on the hidden outputs before the last layer. Specifically, the hidden embedding after batch-normalization but before ReLU activation of the second last convolutional layer is utilized. As stated in Section 2, a deterministic value is set for σ. In our experiments, we generate extra data [36] and compute the average standard deviation over each audio clip, where we finally set $\sigma = 0.2$. For the other 13 tested cut-edging TSL based methods, we mostly follow the recommended setups and hyper-parameter settings in their original papers. The temperature parameter is set to 1.0 for all when computing KL divergence with soft labels. [1]

Table 1: Comparison of average evaluation accuracies (in %) on recordings of the DCASE2020 ASC data set. Each method is tested with and without the combination of the basic TSL method. Each cell represents the average value over 32 experimental results for 8 target devices × 4 repeated trails.

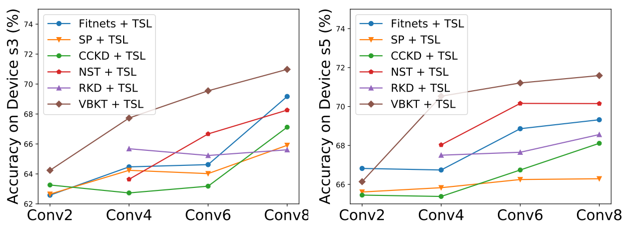
Method	RESNET avg. (%)	RESNET w/ TSL avg. (%)	FCNN avg. (%)	FCNN w/ TSL avg. (%)
Source model	37.70	-	37.13	-
No transfer	54.29	-	49.97	-
One-hot	63.76	-	64.45	-
TSL [27]	68.04	68.04	66.27	66.27
NLE [37]	65.64	67.76	64.47	64.53
Fitnet [28]	66.73	69.89	67.29	69.06
AT [29]	63.73	68.06	64.16	66.35
AB [38]	65.34	68.69	66.21	66.91
VID [39]	63.90	68.56	63.79	65.75
FSP [40]	64.44	68.94	65.33	66.01
COFD [30]	64.92	68.57	66.69	68.63
SP [41]	64.57	68.45	65.74	67.36
CCKD [42]	65.59	69.47	66.52	68.29
PKT [43]	64.65	65.43	64.84	67.25
NST [44]	68.35	68.51	67.13	68.84
RKD [45]	65.28	68.46	65.63	67.27
VBKT	**69.58**	**69.90**	**69.96**	**70.50**

3.2. Evaluation Results on Device Adaptation

Evaluation results of device adaptation on the DCASE2020 ASC task are shown in Table 1. The source models are trained on data recorded by Device A, where we can get a classification accuracy of 79.09% for RESNET and 79.70% for FCNN, on the source test set, respectively. There are 8 target devices, i.e., Device B, C, s1-s6. The accuracy reported in each cell of Table 1 is obtained by averaging among 32 experimental results, from 8 target devices and 4 trails for each. The first and third columns represent results by directly using the knowledge transfer methods; whereas the second and fourth columns list accuracies obtained when further combined with the basic TSL method.

We look at the results without the combination of TSL at first. The 1st row gives results by directly testing the source model on target devices. We can observe a huge degradation (from ~79% to ~37%) when compared with the results on source test set. That shows the device mismatch is indeed a critical aspect in acoustic scene classification, as the device changing causing a serve performance drop. The 2nd and 3rd rows in Table 1 give results of target model trained by target data either from the scratch or fine-tuned on source model. By comparing them we can argue the importance of knowledge transfer when building a target model.

[1] Code available: `https://github.com/MihawkHu/ASC_Knowledge_Transfer`



(a)

(b)

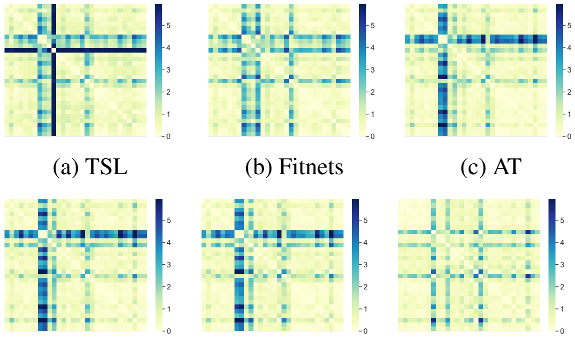
Fig. 2: Evaluation results for using different hidden layers of FCNN model, on two target devices: (a) Device s3 and (b) Device s5. The basic TSL is combined with all methods.

The 4th to 16th rows in Table 1 show the evaluated results of 13 recent top TSL based methods. The result of basic TSL [26, 27], which minimizes KL divergence between model outputs and soft labels, is shown in the 4th row. We can observe a gain obtained by TSL when compared with one-hot fine-tuning. If we compare the other methods (5th to 16th rows) with basic TSL, they only show small advantages on this task. The bottom row of Table 1 shows results of our proposed VBKT approach. It not only outperforms one-hot fine-tuning by a large margin (69.58% vs. 63.76% for RESNET, and 69.12% vs. 61.54% for FCNN), but also attains superior classification results to those obtained with other algorithms.

We further investigate the combination of the proposed approach and the basic TSL. The experimental results are shown in the 2nd and 4th columns of Table 1, for RESNET and FCNN models, respectively. Specifically, the original cross entropy (CE) loss is replaced by an addition of 0.9 × KL loss with soft labels and 0.1 × CE loss with hard labels. There is no change to basic TSL so results remain the same in the 4th row. For the other tests (from 5th to 16th rows), when combining with basic TSL, most tested methods can attain further gains. Indeed, such a combination is recommended by some studies [28, 29, 43]. Finally, we compare our proposed VBKT approach with others, when combined with TSL, the accuracy can be further boosted, and it still outperforms other tested methods under the same setup.

3.3. Effects of Hidden Embedding Depth

In our basic setup, the hidden embedding before the last convolutional layer is utilized for modeling latent variables. Ablation experiments are further carried out to investigate the effects of using different-layer hidden embedding. Experiments are performed on FCNN since it has a sequential architecture with stacked convolutional layers, namely 8 convolutional layers and 1×1 convolutional layer for outputs. Results are shown in Figure 2. Methods use all hidden layers, like COFD, are not covered here. We don't have results for RKD and NST on Conv2 due to they exceeds the available memory on our machine. From the results we can see that the best performance is obtained from last layer for most of the assessed approaches. Moreover, the hidden embedding closer to the model output allows for a higher accuracy than that closer to the input. Therefore we can argue that late features are better than early features in transferring knowledge across domains. That is in line with what is observed in [28, 44]. Finally, VBKT attains a very competitive ASC accuracy and consistently outperforms other methods independently of the selected hidden layers.



(a) TSL

(b) Fitnets

(c) AT

(d) SP

(e) CCKD

(f) VBKT

Fig. 3: Visualized heatmaps of the intra-class discrepancy between target outputs. FCNN on target device s5 is used.

3.4. Visualization of Intra-class Discrepancy

To better understand the effectiveness of the proposed VBKT approach, we compare the intra-class discrepancy between target model outputs (before softmax). The visualized heatmap results are shown in (a)-(f) of Figure 3. Here we randomly select 30 samples from the same class and compute L_2 distance between model outputs of each two. Thus each cell in subnets of Figure 3 represents the discrepancy between two outputs, as the darker color means bigger intra-class discrepancy. From these visualization results we can argue that the one obtained by our proposed VBKT approach in Figure 3f has consistently smaller intra-class discrepancy than those produced by others, implying that VBKT brings up more discriminative information and results in a better cohesion of instances from the same class.

4. CONCLUSION

In this study, we propose a variational Bayesian approach to address the cross-domain knowledge transfer issues when deep models are used. Different from previous solutions, we propose to transfer knowledge via prior distributions of deep latent variables from the source domain. We cast the problem into learning distributions of latent variables in deep neural networks. In contrast to conventional maximum a posteriori estimation, a variational Bayesian inference algorithm is then formulated to approximate the posterior distribution in the target domains. We assess the effectiveness of our proposed VB approach on the device adaptation tasks for the DCASE2020 ASC data set. Experimental evidence clearly demonstrate that the target model obtained with our proposed approach outperforms all other tested methods in all tested conditions.

5. REFERENCES

[1] Geoffrey Hinton, Li Deng, Dong Yu, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Interspeech*, 2011.

[3] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui 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, 2014.

[4] Chin-Hui Lee and Qiang Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1241–1269, 2000.

[5] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[6] Peter Bell, Joachim Fainberg, Ondrej Klejch, Jinyu Li, Steve Renals, and Pawel Swietojanski, "Adaptation algorithms for speech recognition: An overview," *arXiv preprint arXiv:2008.06580*, 2020.

[7] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "A multi-device dataset for urban acoustic scene classification," *arXiv preprint arXiv:1807.09840*, 2018.

[8] Shayan Gharib, Konstantinos Drossos, Emre Cakir, Dmitriy Serdyuk, and Tuomas Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," *arXiv preprint arXiv:1808.05777*, 2018.

[9] Khaled Koutini, Florian Henkel, Hamid Eghbal-zadeh, and Gerhard Widmer, "Low-complexity models for acoustic scene classification based on receptive field regularization and frequency damping," *arXiv preprint arXiv:2011.02955*, 2020.

[10] Hu Hu, Sabato Marco Siniscalchi, Yannan Wang, and Chin-Hui Lee, "Relational teacher student learning with neural label embedding for device adaptation in acoustic scene classification," *Interspeech*, 2020.

[11] Ian J Goodfellow, Mehdi Mirza, et al., "An empirical investigation of catastrophic forgetting in gradient-based neural networks," *arXiv preprint arXiv:1312.6211*, 2013.

[12] Zhen Huang, Sabato Marco Siniscalchi, I-Fan Chen, Jiadong Wu, and Chin-Hui Lee, "Maximum a posteriori adaptation of network parameters in deep models," *Interspeech*, 2015.

[13] James Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[14] Warren B Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality*, vol. 703, John Wiley & Sons, 2007.

[15] Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, et al., "Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review," *International Journal of Automation and Computing*, vol. 14, no. 5, pp. 503–519, 2017.

[16] J-L Gauvain and Chin-Hui Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on speech and audio processing*, vol. 2, no. 2, pp. 291–298, 1994.

[17] Olivier Siohan, Cristina Chesta, and Chin-Hui Lee, "Joint maximum a posteriori adaptation of transformation and hmm parameters," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 417–428, 2001.

[18] Zhen Huang, Sabato Marco Siniscalchi, and Chin-Hui Lee, "Bayesian unsupervised batch and online speaker adaptation of activation function parameters in deep models for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 64–75, 2017.

[19] Shinji Watanabe, Yasuhiro Minami, Atsushi Nakamura, and Naonori Ueda, "Variational bayesian estimation and clustering for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 365–381, 2004.

[20] Alex Graves, "Practical variational inference for neural networks," in *Advances in neural information processing systems*. Citeseer, 2011, pp. 2348–2356.

[21] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[22] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner, "Variational continual learning," *arXiv preprint arXiv:1710.10628*, 2017.

[23] Wei-Ning Hsu and James Glass, "Scalable factorized hierarchical variational autoencoder training," *arXiv preprint arXiv:1804.03201*, 2018.

[24] Shijing Si, Jianzong Wang, Huiming Sun, Jianhan Wu, et al., "Variational information bottleneck for effective low-resource audio classification," *arXiv preprint arXiv:2107.04803*, 2021.

[25] Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun, "Hands-on bayesian neural networks–a tutorial for deep learning users," *arXiv preprint arXiv:2007.06823*, 2020.

[26] Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong, "Learning small-size dnn with output-distribution-based criteria," in *Interspeech*, 2014.

[27] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[28] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[29] 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.

[30] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi, "A comprehensive overhaul of feature distillation," in *CVPR*, 2019, pp. 1921–1930.

[31] Tim Salimans, David A Knowles, et al., "Fixed-form variational posterior approximation through stochastic linear regression," *Bayesian Analysis*, vol. 8, no. 4, pp. 837–882, 2013.

[32] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," *arXiv preprint arXiv:2005.14623*, 2020.

[33] Hu Hu, Chao-Han Huck Yang, Xianjun Xia, Xue Bai, et al., "Device-robust acoustic scene classification based on two-stage categorization and data augmentation," *arXiv preprint arXiv:2007.08389*, 2020.

[34] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[35] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[36] Hu Hu, Chao-Han Huck Yang, Xianjun Xia, Xue Bai, et al., "A two-stage approach to device-robust acoustic scene classification," in *ICASSP*. IEEE, 2021, pp. 845–849.

[37] Zhong Meng, Hu Hu, Jinyu Li, Changliang Liu, Yan Huang, Yifan Gong, and Chin-Hui Lee, "L-vector: Neural label embedding for domain adaptation," in *ICASSP*. IEEE, 2020, pp. 7389–7393.

[38] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *AAAI*, 2019, vol. 33, pp. 3779–3787.

[39] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai, "Variational information distillation for knowledge transfer," in *CVPR*, 2019, pp. 9163–9171.

[40] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *CVPR*, 2017, pp. 4133–4141.

[41] Frederick Tung and Greg Mori, "Similarity-preserving knowledge distillation," in *CVPR*, 2019, pp. 1365–1374.

[42] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang, "Correlation congruence for knowledge distillation," in *CVPR*, 2019, pp. 5007–5016.

[43] Nikolaos Passalis and Anastasios Tefas, "Learning deep representations with probabilistic knowledge transfer," in *ECCV*, 2018, pp. 268–284.

[44] Zehao Huang and Naiyan Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," *arXiv preprint arXiv:1707.01219*, 2017.

[45] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho, "Relational knowledge distillation," in *CVPR*, 2019, pp. 3967–3976.