A Variational Bayesian Approach to Learning Latent Variables for Acoustic Knowledge Transfer

Hu Hu¹, Sabato Marco Siniscalchi¹ ², Chao-Han Huck Yang¹, Chin-Hui Lee¹

¹School of Electrical and Computer Engineering, Georgia Institute of Technology
²Computer Engineering School, University of Enna Kore
Outline

• Introduction
  ➢ Acoustic mismatches and knowledge transfer

• Bayesian Adaptive Learning
  ➢ Bayesian adaptive learning framework
  ➢ Bayesian adaptation for speech processing systems
  ➢ Challenges of Bayesian Adaptation for Deep Models

• Variational Bayesian Knowledge Transfer
  ➢ Bayesian inference of deep latent variables
  ➢ Variational Bayes based adaptive learning
  ➢ Experimental evaluation
Acoustic Variabilities and Mismatches

• In production, acoustic models need to deal with different application scenarios.

• Acoustic variabilities:
  ➢ Speakers: genders, accents, ...
  ➢ Recording devices: handsets, channels, ...
  ➢ Recording environments: scenes, noise types, reverberations, ...
  ➢ ....

• Acoustic mismatches usually cause severe degradation in diverse testing conditions.
• Effective adaptation algorithms are required.
Acoustic Knowledge Transfer

• Acoustic knowledge transfer:
  ➢ Transfer knowledge from the source acoustic domain to the target ones related to testing conditions.
  ➢ It is also referred to as the supervised domain adaptation.

• An example of device adaptation
  ➢ Trained by data from iPhone (Source domain).
  ➢ Adapted to iPad and HomePod (Target domains).
Outline

• Introduction
  ➢ Acoustic mismatches and knowledge transfer

• Bayesian Adaptive Learning
  ➢ Bayesian adaptive learning framework
  ➢ Bayesian adaptation for speech processing systems
  ➢ Challenges of Bayesian Adaptation for Deep Models

• Variational Bayesian Knowledge Transfer
  ➢ Bayesian inference of deep latent variables
  ➢ Variational Bayes based adaptive learning
  ➢ Experimental evaluation
Bayesian Adaptive Learning Framework

• Bayes’ theory:

\[ p(λ|D) = \frac{p(D|λ)p(λ)}{p(D)} \]

- \(λ\): model parameters; \(D\): data; \(S\): source domain; \(T\): target domain.

• For adaptation setups:
  - Prior knowledge learnt from the source domain is encoded in prior distribution:
    \[ p(λ_T) = p(λ_S|D_S) \]
  - The target domain posterior distribution:
    \[ p(λ_T|D_T) = \frac{p(D_T|λ_T)p(λ_S|D_S)}{p(D_T)} \]

• The posterior is usually intractable and difficult to get.
  - An approximation is required: Maximum a posteriori (MAP), Variational Bayes (VB), ...
MAP for GMM-HMM based ASR

- MAP shows good performance for GMM-HMM based ASR system to handle acoustic mismatches [Gauvian, 1994; Lee, 2000].

\[ \lambda^*_T = \arg\max_{\lambda_T} p(\lambda_T|\mathcal{D}_T) = \arg\max_{\lambda_T} p(\mathcal{D}_T|\lambda_T)p(\lambda_T) \]

- Example: GMM and HMM parameters with conjugated prior distributions:
  - HMM parameters: Dirichlet distribution.
  - GMM parameters: Normal-Wishart distribution.

The GMM-HMM system.
MAP for DNN-HMM based ASR

• MAP also shows good performance for DNN-HMM based ASR system for speaker adaptation [Huang, 2015; Huang 2017].

• Linear hidden network (LHN) is used to cast Bayesian assumption.

\[ \text{Loss}_{MAP} = -\log p(D_T|W) - \alpha \log p(W_{lhn}) \]
Challenges of Bayesian Adaptation for Deep Models

- Traditional Bayesian approaches usually focus on model parameters.
  - It works well for traditional statistic models like HMM, GMM, SVM, ...

- For DNN, we have much more parameters than training samples.
  - # of para. >> # of data dimension * # of data [Sebastien, 2021].
  - Especially for the adaptation scenarios.

- Challenges and problems:
  - Difficult to get accurate estimations of model parameters by Bayesian approaches.
  - Curse of dimensionality.
Outline

• Introduction
  ➢ Acoustic mismatches and knowledge transfer

• Bayesian Adaptive Learning
  ➢ Bayesian adaptive learning framework
  ➢ Bayesian adaptation for speech processing systems
  ➢ Challenges of Bayesian Adaptation for Deep Models

• Variational Bayesian Knowledge Transfer
  ➢ Bayesian inference of deep latent variables
  ➢ Variational Bayes based adaptive learning
  ➢ Experimental evaluation

IEEE ICASSP 2022 Paper 2182
We propose to perform Bayesian adaptive learning on **deep latent variables** rather than on DNN weights.

- An unobservable representation of data, corresponding to intermediate hidden embedding from a specific layer of DNN.

An example of deep latent variables.

- Z indicates the deep latent variables.
  - Prior: \( p(Z) \); Posterior: \( p(Z|X) \).
  - We decouple DNN weights to \( \theta \) and \( \omega \).

\[
X \sim p_{\text{data}}(X) \quad \rightarrow \quad \theta \quad \rightarrow \quad Z \sim p(Z) \quad \rightarrow \quad \omega \quad \rightarrow \quad Y \sim p(Y)
\]
Deep Latent Variables (Cont’d)

• Acoustic scene model embedding.
  - 10 different scene classes:
    - Airport, metro, ...
  - 3 general classes C1-C3:
    - Indoor, outdoor, transportation.
  - Hidden embedding is generated by a DNN model and reduced to 2 dimensions.

• Deep latent variable has its own distribution form.
• Deep latent variable encodes structural relationships.

A visualization of deep latent variables [Hu, 2020].
Bayesian Inference of Deep Latent Variables

- Latent variables are introduced in addition to DNN weights.

\[ p(\lambda) = p(Z, \theta, \omega) = p(Z|\theta)p(\theta)p(\omega) \]
Bayesian Inference of Deep Latent Variables (Cont’d)

• Prior knowledge for target model is learnt from the source domain

\[ p(Z_T | \theta_T) = p(Z_S | \theta_S, \mathcal{D}_S) \]

• Posterior with latent variables:

\[
p(\lambda_T | \mathcal{D}_T) = \frac{p(\mathcal{D}_T | \lambda_T)p(\theta_T)p(\omega_T)p(Z_S | \theta_S, \mathcal{D}_S)}{p(\mathcal{D}_T)}
\]

• Variational Bayes (VB) based estimation way
  - Perform a distribution estimation to obtain the full posterior.
Variational Bayes based Adaptive Learning

• Set a variational distribution to approximate the real distribution.

• Minimize the KLD between them, by

\[ q^*(\lambda_T|\mathcal{D}_T) = \arg\min_{q \in \mathcal{Q}} \text{KL}(q(\lambda_T|\mathcal{D}_T) \parallel p(\lambda_T|\mathcal{D}_T)) \]

• Get a full VB expression with \( Z, \theta \) and \( \omega \).

  ➢ By taking a non-informative prior over \( \theta \) and \( \omega \), we can arrive at the variational lower bound:

\[ \mathcal{L}(\lambda_T; \mathcal{D}_T) = \mathbb{E}_{Z_T \sim q(Z_T|\theta_T, \mathcal{D}_T)} \log p(\mathcal{D}_T|Z_T, \theta_T, \omega_T) - \text{KL}(q(Z_T|\theta_T, \mathcal{D}_T) \parallel p(Z_T|\theta_T)) \]
Variational Bayes based Adaptive Learning (Cont’d)

• Gaussian mean-field variational inference (GMFVI) estimation is used:

  ➢ Each hidden embedding is assumed to be sampled from individual Gaussians:

  \[
  q(Z|\theta, D) = \prod_{i}^{N_T} \mathcal{N}(Z^{(i)}; \mu^{(i)}, (\sigma^{(i)})^2 I)
  \]

  ➢ Final learning objective:

  \[
  \mathcal{L}(\lambda_T; D_T) = \sum_{i}^{N_T} \mathbb{E}_{z_T^{(i)} \sim \mathcal{N}(\mu_T^{(i)}, \sigma^2)} \log p(y_T^{(i)}|x_T^{(i)}, z_T^{(i)}, \theta_T, \omega_T) - \frac{1}{2\sigma^2} \sum_{i}^{N_T} \|\mu_T^{(i)} - \mu_S^{(i)}\|_2^2
  \]
Experimental Setup of Acoustic Scene Classification

• Data set: DCASE 2020 ASC data set.
  • Code available: https://github.com/MihawkHu/ASC_Knowledge_Transfer

• Source domain data:
  ➢ Recorded by a Zoom F8 audio recorder.
  ➢ ~10K training audio clips.

• Target domain data:
  ➢ Recorded by 8 different devices:
    o iPhone SE, Samsung Galaxy S7, ...
  ➢ Each has 750 training audio clips.

• Two state-of-the-art models [Hu, 2020] are used: RESNET and FCNN.
Teacher-Student Learning Family

• Teacher-student learning (TSL) is used as a comparison.
  ➢ Transfers knowledge from the teacher network to the student network.
  ➢ The basic approach is to minimize the KLD between outputs of teacher model and student model.

• Point estimation vs. distribution estimation.
Teacher-Student Learning Family (Cont’d)

• 13 recent cut-edging knowledge transfer methods compared in our experiments:
  - NLE: Neural label embedding [Meng, 2020].
  - Fitnets: Hints for thin nets [Romero, 2014].
  - AT: Attention transfer [Zagoruyko, 2016].
  - AB: Activation boundaries [Heo, 2019].
  - VID: Variational information distillation [Ahn, 2019].
  - FSP: Flow of solution procedure [Yim, 2017].
  - COFD: Comprehensive overhaul feature distillation [Heo, 2019].
  - SP: Similarity preserving [Tung, 2019].
  - CCKD: Correlation congruence knowledge distillation [Peng, 2019].
  - PKT: Probabilistic knowledge transfer [Passalis, 2018].
  - NST: Neuron selectivity transfer [Huang, 2017].
  - RKD: Relational knowledge transfer [Park, 2019].

• All above are implemented and compared. Some are presented in the next few slides.
Experimental Results on Acoustic Scene Classification (1/5)

- Accuracies on source device data:
  - RESNET: 79.09 %, FCNN: 79.70 %.

<table>
<thead>
<tr>
<th>Method</th>
<th>RESNET avg% ± std</th>
<th>FCNN avg% ± std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>37.70</td>
<td>37.13</td>
</tr>
<tr>
<td>No transfer</td>
<td>54.29 ± 0.76</td>
<td>49.97 ± 2.70</td>
</tr>
<tr>
<td>One-hot</td>
<td>63.76 ± 0.59</td>
<td>64.45 ± 0.51</td>
</tr>
<tr>
<td>TSL</td>
<td>68.04 ± 0.34</td>
<td>66.27 ± 0.46</td>
</tr>
<tr>
<td>NLE</td>
<td>65.64 ± 0.53</td>
<td>64.47 ± 0.59</td>
</tr>
<tr>
<td>AT</td>
<td>63.73 ± 0.81</td>
<td>64.16 ± 0.49</td>
</tr>
<tr>
<td>SP</td>
<td>64.57 ± 0.76</td>
<td>65.74 ± 0.37</td>
</tr>
<tr>
<td>RKD</td>
<td>65.28 ± 0.81</td>
<td>65.63 ± 0.22</td>
</tr>
<tr>
<td>V BK T-GMFVI</td>
<td><strong>69.58 ± 0.49</strong></td>
<td><strong>69.96 ± 0.13</strong></td>
</tr>
</tbody>
</table>
### Experimental Results on Acoustic Scene Classification (2/5)

<table>
<thead>
<tr>
<th>Method</th>
<th>RESNET avg% ± std</th>
<th>FCNN avg% ± std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>37.70</td>
<td>37.13</td>
</tr>
<tr>
<td>No transfer</td>
<td>54.29 ± 0.76</td>
<td>49.97 ± 2.70</td>
</tr>
<tr>
<td>One-hot</td>
<td>63.76 ± 0.59</td>
<td>64.45 ± 0.51</td>
</tr>
<tr>
<td>TSL</td>
<td>68.04 ± 0.34</td>
<td>66.27 ± 0.46</td>
</tr>
<tr>
<td>NLE</td>
<td>65.64 ± 0.53</td>
<td>64.47 ± 0.59</td>
</tr>
<tr>
<td>AT</td>
<td>63.73 ± 0.81</td>
<td>64.16 ± 0.49</td>
</tr>
<tr>
<td>SP</td>
<td>64.57 ± 0.76</td>
<td>65.74 ± 0.37</td>
</tr>
<tr>
<td>RKD</td>
<td>65.28 ± 0.81</td>
<td>65.63 ± 0.22</td>
</tr>
<tr>
<td>VBIKT-GMFVI</td>
<td><strong>69.58 ± 0.49</strong></td>
<td><strong>69.96 ± 0.13</strong></td>
</tr>
</tbody>
</table>

- **Accuracies on source device data:**
  - RESNET: 79.09 %, FCNN: 79.70 %.

- Device mismatches causes huge degradations when directly applying the source model.
Experimental Results on Acoustic Scene Classification (3/5)

• Accuracies on source device data:
  - RESNET: 79.09 %, FCNN: 79.70 %.

• Device mismatches causes huge degradation when directly applying the source model.

• Fine-tuning with target data can help ease the mismatch issue.

<table>
<thead>
<tr>
<th>Method</th>
<th>RESNET avg% ± std</th>
<th>FCNN avg% ± std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>37.70</td>
<td>37.13</td>
</tr>
<tr>
<td>No transfer</td>
<td>54.29 ± 0.76</td>
<td>49.97 ± 2.70</td>
</tr>
<tr>
<td>One-hot</td>
<td>63.76 ± 0.59</td>
<td>64.45 ± 0.51</td>
</tr>
<tr>
<td>TSL</td>
<td>68.04 ± 0.34</td>
<td>66.27 ± 0.46</td>
</tr>
<tr>
<td>NLE</td>
<td>65.64 ± 0.53</td>
<td>64.47 ± 0.59</td>
</tr>
<tr>
<td>AT</td>
<td>63.73 ± 0.81</td>
<td>64.16 ± 0.49</td>
</tr>
<tr>
<td>SP</td>
<td>64.57 ± 0.76</td>
<td>65.74 ± 0.37</td>
</tr>
<tr>
<td>RKD</td>
<td>65.28 ± 0.81</td>
<td>65.63 ± 0.22</td>
</tr>
<tr>
<td>VBLT-GMFVI</td>
<td>69.58 ± 0.49</td>
<td>69.96 ± 0.13</td>
</tr>
</tbody>
</table>
Experimental Results on Acoustic Scene Classification (4/5)

- Accuracies on source device data:
  - RESNET: 79.09 %, FCNN: 79.70 %.
- Device mismatches causes huge degradation when directly applying the source model.
- Fine-tuning with target data can help ease the mismatch issue.
- Knowledge transfer algorithms show advantages over simply fine-tuning.

<table>
<thead>
<tr>
<th>Method</th>
<th>RESNET avg% ± std</th>
<th>FCNN avg% ± std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source.</td>
<td>37.70</td>
<td>37.13</td>
</tr>
<tr>
<td>No transfer</td>
<td>54.29 ± 0.76</td>
<td>49.97 ± 2.70</td>
</tr>
<tr>
<td>One-hot</td>
<td>63.76 ± 0.59</td>
<td>64.45 ± 0.51</td>
</tr>
<tr>
<td>TSL</td>
<td>68.04 ± 0.34</td>
<td>66.27 ± 0.46</td>
</tr>
<tr>
<td>NLE</td>
<td>65.64 ± 0.53</td>
<td>64.47 ± 0.59</td>
</tr>
<tr>
<td>AT</td>
<td>63.73 ± 0.81</td>
<td>64.16 ± 0.49</td>
</tr>
<tr>
<td>SP</td>
<td>64.57 ± 0.76</td>
<td>65.74 ± 0.37</td>
</tr>
<tr>
<td>RKD</td>
<td>65.28 ± 0.81</td>
<td>65.63 ± 0.22</td>
</tr>
<tr>
<td>VBKT-GMFVI</td>
<td><strong>69.58 ± 0.49</strong></td>
<td><strong>69.96 ± 0.13</strong></td>
</tr>
</tbody>
</table>
Experimental Results on Acoustic Scene Classification (5/5)

<table>
<thead>
<tr>
<th>Method</th>
<th>RESNET avg% ± std</th>
<th>FCNN avg% ± std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source.</td>
<td>37.70</td>
<td>37.13</td>
</tr>
<tr>
<td>No transfer</td>
<td>54.29 ± 0.76</td>
<td>49.97 ± 2.70</td>
</tr>
<tr>
<td>One-hot</td>
<td>63.76 ± 0.59</td>
<td>64.45 ± 0.51</td>
</tr>
<tr>
<td>TSL</td>
<td>68.04 ± 0.34</td>
<td>66.27 ± 0.46</td>
</tr>
<tr>
<td>NLE</td>
<td>65.64 ± 0.53</td>
<td>64.47 ± 0.59</td>
</tr>
<tr>
<td>AT</td>
<td>63.73 ± 0.81</td>
<td>64.16 ± 0.49</td>
</tr>
<tr>
<td>SP</td>
<td>64.57 ± 0.76</td>
<td>65.74 ± 0.37</td>
</tr>
<tr>
<td>RKD</td>
<td>65.28 ± 0.81</td>
<td>65.63 ± 0.22</td>
</tr>
<tr>
<td><strong>VBKT-GMFVI</strong></td>
<td><strong>69.58 ± 0.49</strong></td>
<td><strong>69.96 ± 0.13</strong></td>
</tr>
</tbody>
</table>

- Accuracies on source device data:
  - RESNET: 79.09 %, FCNN: 79.70 %.

- Device mismatches causes huge degradation when directly applying the source model.

- Fine-tuning with target data can help ease the mismatch issue.

- Knowledge transfer algorithms show advantages over simply fine-tuning.

- Our proposed VBKT method improves performance on target devices and outperforms all others.
Appendix: More Results and Analysis

- Effects of Hidden Embedding Depth
  - Methods use only one hidden layer are compared.

- Last layer (Conv8) shows best results than others.

- Layers closer to output show better results.
  - Better transferable properties.

- The proposed method consistently outperforms all others.
• Visualization of intra-class discrepancy
  • 30 samples from the same class are randomly selected.
  • L2 distance between model outputs are computed and visualized.
  • Darker color means bigger intra-class discrepancy.

• The proposed method has consistent smaller intra-class discrepancy than others.
  • It has more discriminative information and better cohesion of instances.
Thank you~