Domain Adversarial Training for Accented Speech Recognition

Sining Sun [1-3] , Ching-Feng Yeh [2] , Mei-Yuh Hwang [2] ,
Mari Ostendorf [3] , Lei Xie [1]

Northwestern Polytechnical University [1]
Mobvoi AI Lab, Seattle, USA [2]
University of Washington, Seattle , USA [3]

Outline

- Introduction
- Domain Adaptation
- Domain Adversarial Training (DAT)
- DAT for ASR
- Experimental Results
- Conclusion

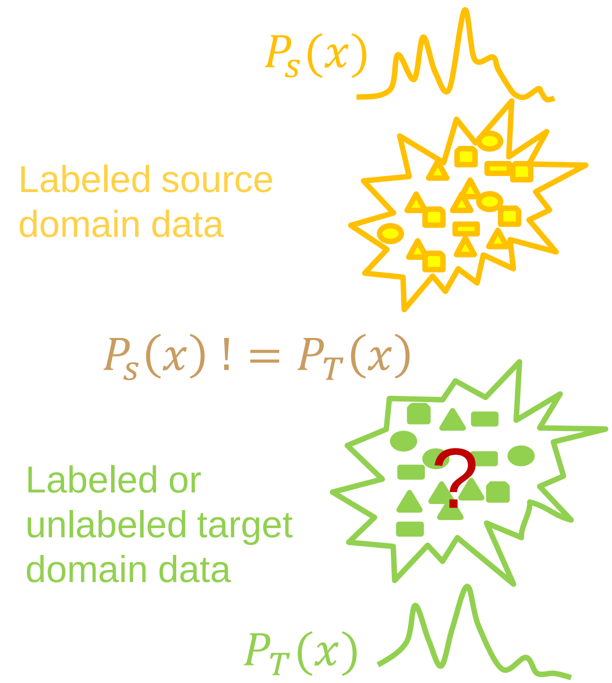


Introduction

- Challenges in ASR
  - Noise, reverberation, accents……
  - Mismatch between training and test data
  - Lack of supervised training data
- Our work
  - Improve ASR performance for accented speech, using unsupervised domain adaptation
  - Learn accent-invariant features using DAT
  - Explore how semi-supervised learning can influence the performance of DAT

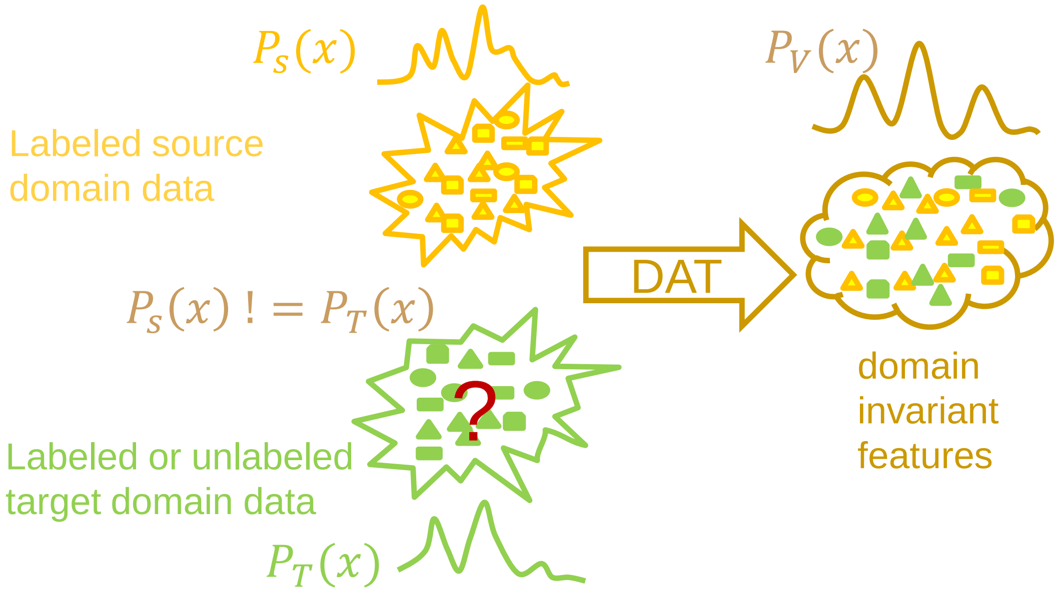
Domain Adaptation

- Domain adaptation
  - Training data
    - Labeled source domain data
    - Labeled or unlabeled target domain data
  - Test data
    - Data with the similar distribution of target domain
  - Task
    - Improve performance on test set using limited target domain data



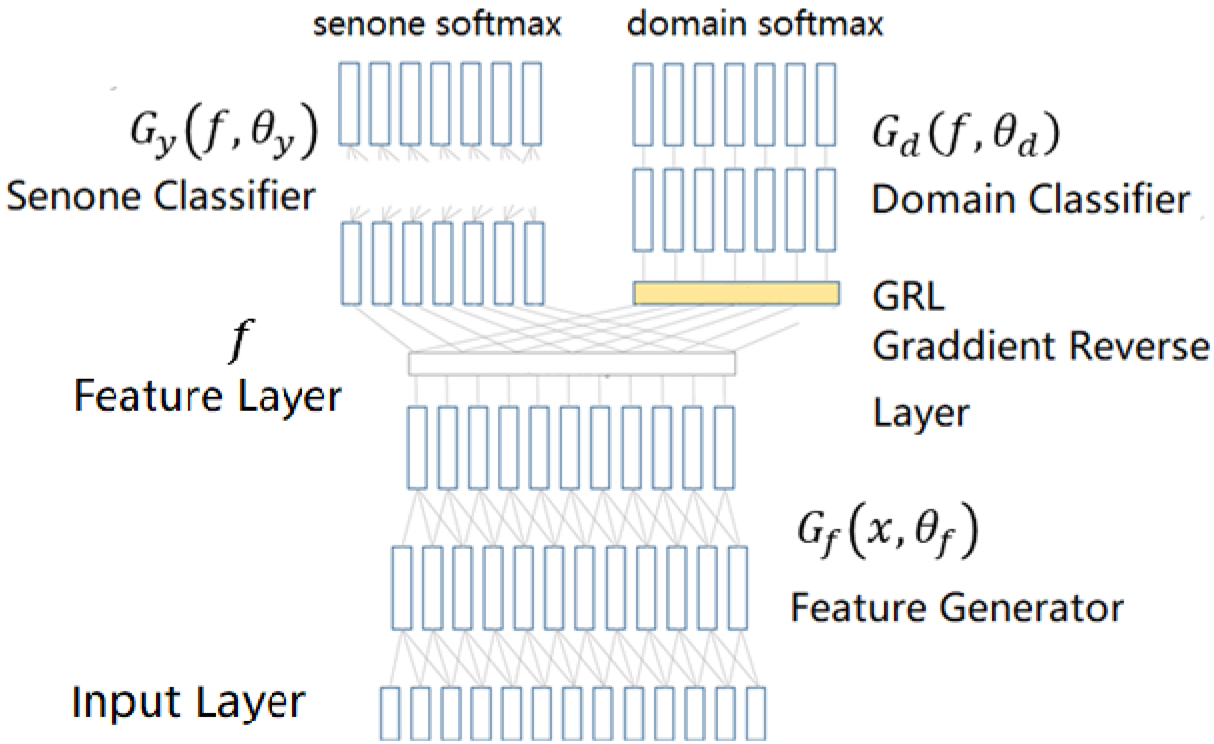
Domain Adversarial Training

- Given labeled or unlabeled target domain data
  - DAT tries to learn features that are
    - Domain-invariant
    - Classification-discriminative



DAT for Speech Recognition

- Gradient reverse layer (GRL) based adversarial training



  - GRL: multiply a constant **negative** factor to gradients generated by $G_d(f, \theta_d)$

DAT for Speech Recognition

- Training
  - Loss function

Indicator for labeled or not

Domain cross entropy

$$E(\theta_f, \theta_y, \theta_d) = \frac{1}{N}\sum_{i=1}^{N}\left(I_d(i)L_y^i(\theta_f, \theta_y) - \lambda I_{vad}(i)L_d^i(\theta_f, \theta_d)\right)$$

Senone cross entropy

Indicator for speech or not

  - Optimization

$$\theta_y^*, \theta_f^* = \min_{\theta_y, \theta_f} E(\theta_y, \theta_f, \theta_d)$$

$$\theta_d^* = \max_{\theta_d} E(\theta_y, \theta_f, \theta_d)$$

$$\Rightarrow$$

$$\theta_f \leftarrow \theta_f - \alpha\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial L_y^i}{\partial \theta_f} I_d(i) - \lambda\frac{\partial L_d^i}{\partial \theta_f} I_{vad}(i)\right)$$

$$\theta_y \leftarrow \theta_y - \alpha\frac{1}{N}\sum_{i=1}^{N}\frac{\partial L_y^i}{\partial \theta_y} I_d(i)$$

$$\theta_d \leftarrow \theta_d - \alpha\frac{1}{N}\sum_{i=1}^{N}\lambda\frac{\partial L_d^i}{\partial \theta_d} I_{vad}(i)$$

DAT for Speech Recognition

- DAT versus Multi-Task Learning (MTL)

$$E(\theta_f, \theta_y, \theta_d) = \frac{1}{N}\sum_{i=1}^{N}\left(I_d(i)L_y^i(\theta_f, \theta_y) - \lambda I_{vad}(i)L_d^i(\theta_f, \theta_d)\right)$$

  - If $\lambda < 0$, it is the regular MTL
  - If $\lambda = 0$, no domain information is used
  - If $\lambda > 0$, it becomes DAT



Experimental Results

- Dataset
  - Source domain data
    - 360 hours standard accent Mandarin training data with transcriptions (Std)
  - Target domain data
    - 600 hours accented Mandarin speech from 6 different provinces (HN, SC, GD, JX JS and FJ) of China
    - 100 hours per accent
    - These data were transcribed



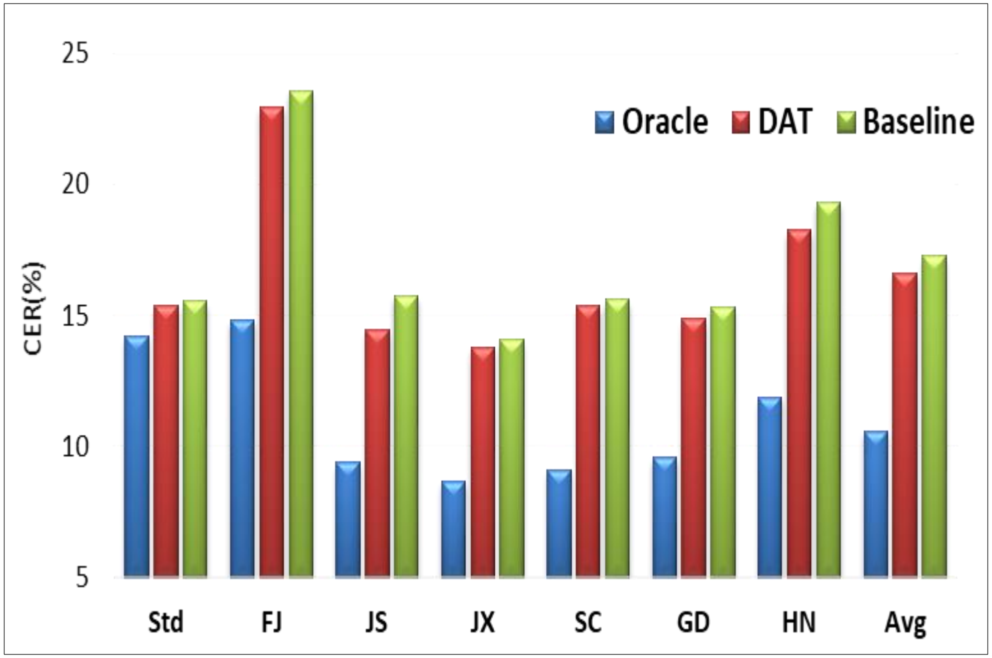
Experimental Results

- Acoustic feature
  - 23-dimensional filterbanks with 3-dimensional pitch
- Acoustic model
  - TDNN with LF-MMI
  - 7 layers and each layer has 625 hidden units with ReLU
  - 5998 output units
  - Trained by Kaldi



Experimental Results

- Accent-invariant feature extraction across all accents using unsupervised DAT



Baseline:

Trained using 360 hours Std data

Oracle:

Trained using all 960 hours training data with transcriptions

DAT:

Trained with unsupervised DAT, all accented data were used.

- Using unsupervised DAT can improve the ASR performance on accented test data

Per-Accent Experiments

- Accent-specific DAT
  - Three accents selected: FJ, SC, HN
  - A different baseline system for each of the following conditions on 100 hours accented speech data

HN
SC
FJ

(1) No Transcripts

Baseline: 300hrs Std data with human transcripts
DAT/MTL: 300hrs Std + 100hrs accented data *without transcripts*

(2) ASR Transcripts

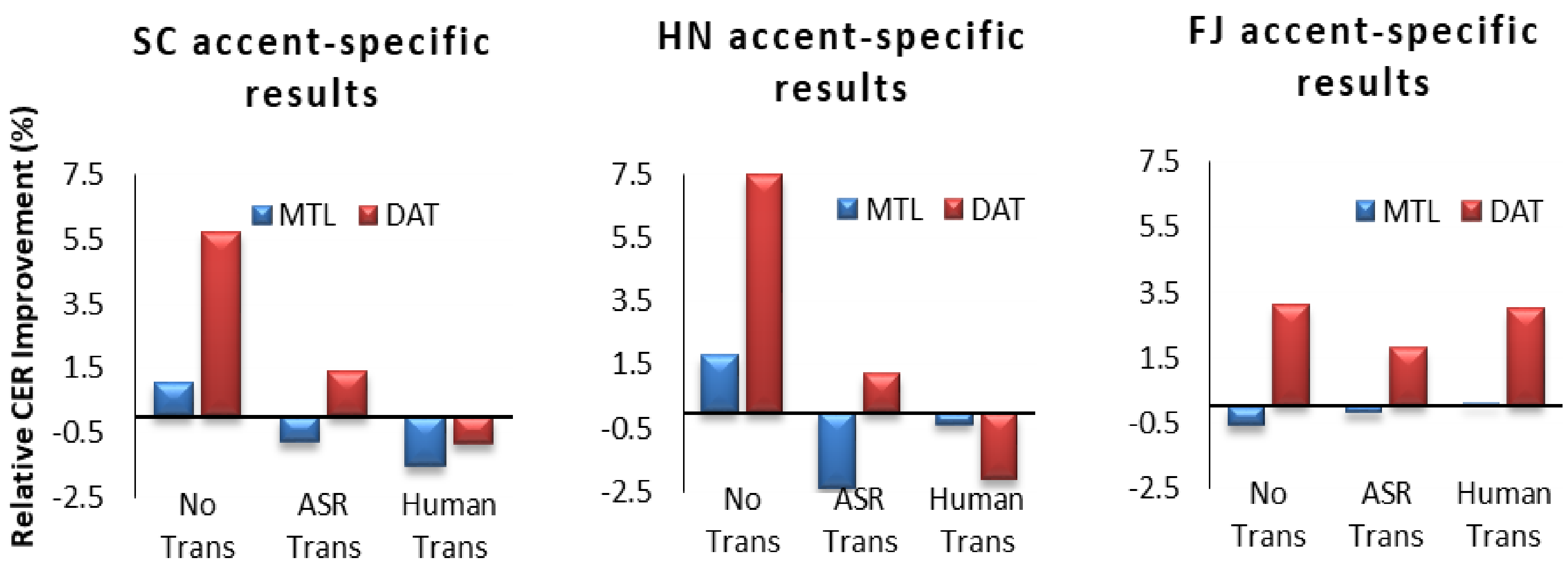
Baseline/DAT/MTL:
300hrs Std data with human transcripts
+ 100hrs accented data with *ASR transcripts*

(3) Human Transcripts

Baseline/DAT/MTL:
300hrs Std data with human transcripts
+ 100hrs accented data with *human transcripts*

Experimental Results

- Relative CER improvement of accent-specific DAT



  - In no transcriptions case, DAT can always help
  - In ASR transcriptions case, DAT contribution shrinks
  - DAT is almost better than MTL on accented test data

Conclusion

- Conclusion
  - Integrated DAT into TDNN AM training for accented speech recognition
  - 7.4% relative CER reduction using unsupervised DAT
  - Explored how automatic transcripts can influence DAT performance
  - 20% relative CER reduction when combining DAT and ASR transcripts
- Future work
  - Compare DAT with other emerging deep domain adaption method
  - Extend DAT to far-field scenario



Thank you!