Variational Domain Adversarial Learning for Speaker Verification

Youzhi Tu[1], Man-Wai Mak[1], Jen-Tzung Chien[2]

[1]Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR of China

[2]Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan

Introduction

- Domain mismatch poses a great challenge to speaker verification (SV). Domain adaptation is often adopted to overcome this problem.
- Domain adversarial neural network (DANN) is a state-of-the-art domain adaptation method for SV. It uses a speaker classifier and a domain discriminator to learn speaker discriminative and domain-invariant features.
- Limitation of DANN: there is no guarantee that the learned features follow a Gaussian distribution, which is an essential requirement for the Gaussian PLDA backend.

Variational Domain-Adversarial Neural Network (VDANN)

- **Methodology**: incorporate a variational auto-encoder (VAE) into the DANN to impose constraint on the distribution of the embedded features.
- **Objective**: produce features that are not only speaker discriminative and domain-invariant but also Gaussian distributed.
- **Architecture**: VDANN comprises a speaker predictor C, a domain classifier D, an encoder E and a decoder G. Their parameters are denoted as θ_c, θ_d, ϕ_e and θ_g, respectively.
- **Optimization**:
  - Keeping θ_c, ϕ_e and θ_g fixed, minimize the domain classification loss with respect to θ_d;
  - Keeping θ_d fixed, maximize the domain classification loss while simultaneously minimizing the speaker classification loss and the VAE loss with respect to θ_c, ϕ_e and θ_g.

References

1. Q. Wang et al., "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in Proc. ICASSP, 2018, pp. 4889–4893.
2. D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in Proc. ICLR, 2014.

VDANN Architecture and Optimization

$$\mathcal{L}_{\text{VDANN}}\left(\theta_c,\theta_d,\phi_e,\theta_g\right)=\mathcal{L}_{\text{C}}\left(\theta_c,\phi_e\right)-\alpha\mathcal{L}_{\text{D}}\left(\theta_d,\phi_e\right)+\beta\mathcal{L}_{\text{VAE}}\left(\phi_e,\theta_g\right)$$

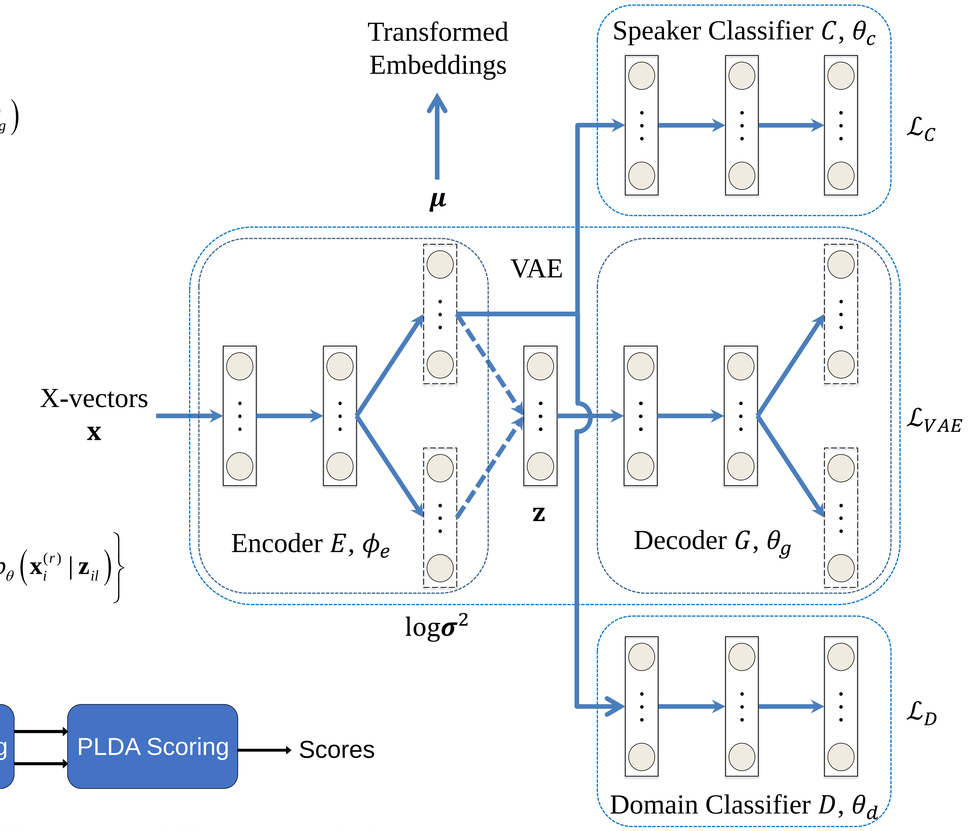
$$\hat{\theta}_d=\underset{\theta_d}{\text{argmax}}\,\mathcal{L}_{\text{VDANN}}\left(\hat{\theta}_c,\theta_d,\hat{\phi}_e,\hat{\theta}_g\right)$$

$$\left(\hat{\theta}_c,\hat{\phi}_e,\hat{\theta}_g\right)=\underset{\theta_c,\phi_e,\theta_g}{\text{argmin}}\,\mathcal{L}_{\text{VDANN}}\left(\theta_c,\hat{\theta}_d,\phi_e,\theta_g\right)$$

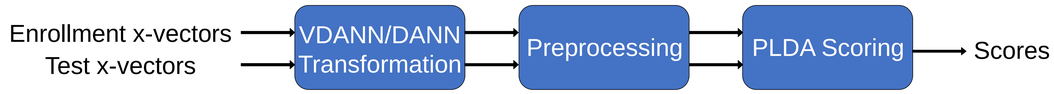
$$\mathcal{L}_{\text{C}}\left(\theta_c,\phi_e\right)=\sum_{r=1}^{R}\mathbb{E}_{p_{\text{data}}\left(\mathbf{x}^{(r)}\right)}\left\{-\sum_{k=1}^{K}y_k^{(r)}\log C\left(E\left(\mathbf{x}^{(r)}\right)\right)_k\right\}$$

$$\mathcal{L}_{\text{D}}\left(\theta_d,\phi_e\right)=\sum_{r=1}^{R}\mathbb{E}_{p_{\text{data}}\left(\mathbf{x}^{(r)}\right)}\left\{-\log D\left(E\left(\mathbf{x}^{(r)}\right)\right)_r\right\}$$

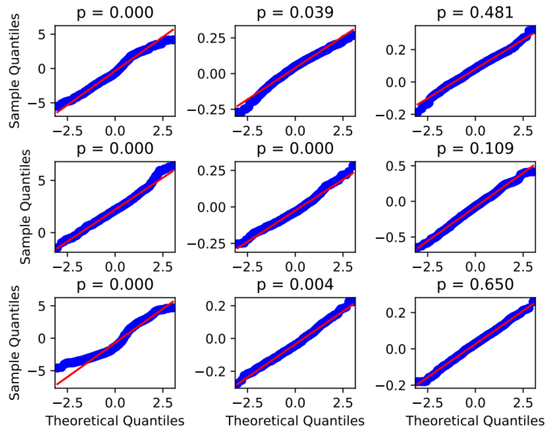
$$\mathcal{L}_{\text{VAE}}(\theta,\phi)=-\sum_{r=1}^{R}\sum_{i=1}^{N_r}\left\{\frac{1}{2}\sum_{j=1}^{J}\left[1+\log\left(\sigma_{ij}^{(r)}\right)^2-\left(\mu_{ij}^{(r)}\right)^2-\left(\sigma_{ij}^{(r)}\right)^2\right]+\frac{1}{L}\sum_{l=1}^{L}\log p_\theta\left(\mathbf{x}_i^{(r)}\mid\mathbf{z}_{il}\right)\right\}$$



Experimental Setup



- X-vectors were extracted using the pre-trained DNN available from the Kaldi repository.
- We trained the VDANN/DANN on SRE04–10, Voxceleb1, Switchboard 2 Phases I–III and SITW datasets. The DANN has the same structure as the VDANN, but without the VAE decoder and the sampling procedure.
- The baseline PLDA model was trained on the SRE04–10 and their augmented x-vectors for SRE16; while for SRE18 the Mixer6 and its augmented x-vectors were also added to the training sets. For VDANN/DANN evaluation, the PLDA model was trained on the transformed x-vectors.
- Pre-processing includes centering and LDA projection (to a 150 dimensional space).



Quantile-quantile (Q–Q) plots of x-vectors and *p*-values from Shapiro–Wilk tests (The larger the *p*, the more Gaussian the distribution.)

SRE16

	No PLDA adaptation		PLDA adaptation	
	EER	minDCF	EER	minDCF
Baseline	11.30	0.890	8.27	0.604
DANN	11.62	0.822	8.43	0.599
VDANN	**11.17**	**0.798**	**8.21**	**0.584**

SRE18-CMN2

	No PLDA adaptation		PLDA adaptation	
	EER	minDCF	EER	minDCF
Baseline	11.21	0.676	9.60	**0.575**
DANN	10.82	0.678	9.28	0.583
VDANN	**10.25**	**0.667**	**9.23**	0.576