Invariant Representations for Noisy Speech Recognition

Dmitriy Serdyuk[12], Kartik Audhkhasi[2], Philémon Brakel[1], Bhuvana Ramabhadran[2], Samuel Thomas[2], Yoshua Bengio[1†]

[1] MILA, Université de Montréal, [2] IBM Watson, † CIFAR Fellow

Introduction

- Modern automatic speech recognition (ASR) systems need to be robust under acoustic variability arising from environmental, speaker, channel, and recording conditions. Ensuring such robustness to variability is a challenge in modern day neural network-based ASR systems, especially when all types of variability are not seen during training.
- We attempt to address this problem by encouraging the neural network acoustic model to learn invariant feature representations.
- We use ideas from recent research on image generation using Generative Adversarial Networks and domain adaptation ideas extending adversarial gradient-based training. A recent work from Ganin et al. proposes to use adversarial training for image domain adaptation by using an intermediate representation from the main target classification network to deteriorate the domain classifier performance through a separate neural network. Our work focuses on investigating neural architectures which produce representations invariant to noise conditions for ASR.
- We evaluate the proposed architecture on the Aurora-4 task, a popular benchmark for noise robust ASR. We show that our method generalizes better than the standard multi-condition training especially when only a few noise categories are seen during training.

Generative Adversarial Networks

- The generator network G has an input of randomly-generated feature vectors and is asked to produce a sample, e.g. an image, similar to the images in the training set. The discriminator network D can either receive a generated image from the generator G or an image from the training set. Its task is to distinguish between the "fake" generated image and the "real" image taken from the dataset [GPAM+14].

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

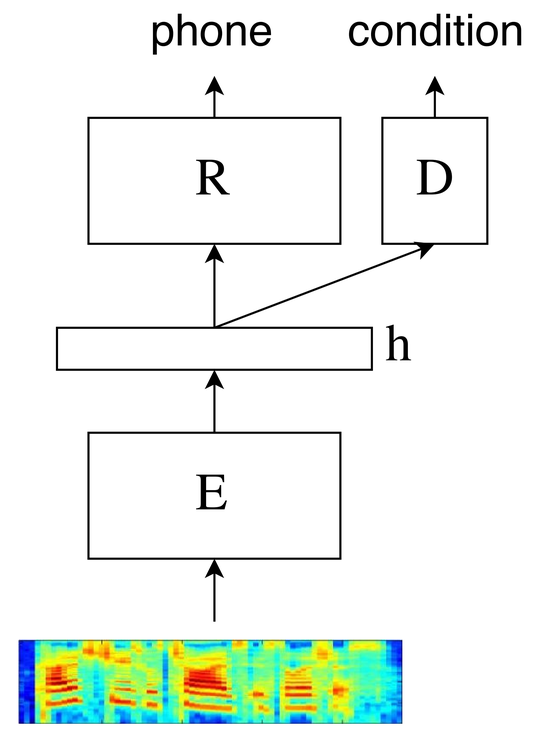
Domain Adaptation with Reverse Gradient

- Prior work by [GL14] proposed a method of training a network which can be adapted to new domains. The training data consists of the images labeled with classes of interest and separate domain (image background) labels. The network has a Y-like structure: the image is fed to the first network which produces a hidden representation h. Then this representation h is input to two separate networks: a domain classifier network (D) and a target classifier network (R). The goal of training is to learn the hidden representation that is invariant to the domain labels and performs well on the target classification task, so that the domain information doesn't interfere with the target classifier at test time. Similar to the GAN objective, which forces the generation distribution be close to the data distribution, the *gradient reverse method* makes domain distributions similar to each other.

Invariant Representation Network

- The model consists of three neural networks. The encoder E produces the intermediate representation h which used in the recognizer R and in the domain discriminator D. The hidden representation h is trained to improve the recognition and minimize the domain discriminator accuracy. The domain discriminator is a classifier trained to maximize its accuracy on the noise type classification task.

$$L = L_1(\hat{y}, y; \theta_R, \theta_E) + \alpha L_2(\hat{d}, d; \theta_D) - \beta L_3(\hat{d}, d; \theta_E)$$



- The model is a DNN-HMM system with a feed-forward neural network trained to predict the CD HMM state (2000 classes).
- Input: 40-dimensional Mel-filterbank features with their deltas and delta-deltas spliced over ±5 frames (total dimension is 1320).
- Baseline: 6-layer DNN with 2048 rectified linear units at every layer.
- The right branch that predicts the domain (noise condition). This branch is discarded in the testing phase.
- The noise condition is the domain label. We are merging all noise types into one label and clean as the other label.
- The invariance term is

$$L_3 = d\log(1-\hat{d}) + (1-d)\log(\hat{d}),$$

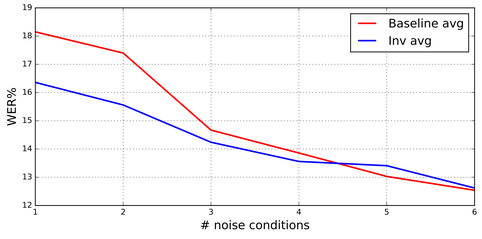
where d is the domain label and $\hat{d}$ is the network prediction.

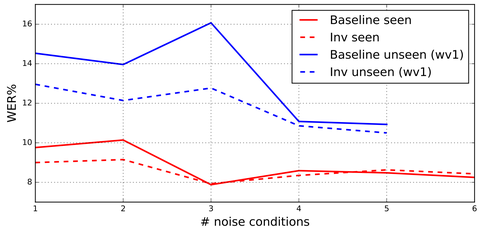
Dataset

- Aurora-4 is based on the Wall Street Journal corpus (WSJ0).
- It contains noises of 6 categories which was added to clean data. Every clean and noisy utterance is filtered to simulate the frequency characteristics.
- The training data:
  - One microphone: Sennheiser.
  - 4400 clean utterances.
  - 446 utterances for each of 6 noise condition.
  - Total of 2676 noisy utterances.
- The testing set:
  - Two microphones.
  - 330 utterances for each condition, each microphone.

Experiments on Aurora-4

In order to evaluate the impact of our method on generalization to unseen noises, we performed 6 experiments with different set of seen noises. The networks are trained on clean data, with each noise condition added one-by-one in the following order: airport, babble, car, restaurant, street, and train.
Average performance of the baseline multi-condition and invariance model varying with the number of noise conditions used for training. Bottom: Average performance on seen versus unseen noise conditions. Testing was performed on all wv1 conditions (Sennheiser microphone).





Results

Average word error rate (WER%) on Aurora-4 dataset on all test conditions, including seen and unseen noise and unseen microphone. First column is the number of noise conditions used for the training. The last row is a preliminary experiment with layer-wise pre-training close to state-of-the-art model and a corresponding invariance training starting with a pretrained model.

Noise cond	Inv	BL	A		B		C		D	
			Inv	BL	Inv	BL	Inv	BL	Inv	BL
1	16.36	18.14	6.54	7.57	12.71	14.09	11.45	13.10	22.47	24.80
2	15.56	17.39	5.90	6.58	11.69	13.28	11.12	13.51	21.79	23.96
3	14.24	14.67	5.45	5.08	10.76	12.44	9.75	9.84	19.93	19.30
4	13.61	13.84	5.08	5.29	9.73	9.97	9.49	9.56	19.49	19.90
5	13.41	13.02	5.12	5.34	9.52	9.42	9.55	8.67	19.33	18.65
6	12.62	12.60	4.80	4.61	9.04	8.86	8.76	8.59	18.16	18.21
6*	11.85	11.99	4.52	4.76	8.76	8.76	7.79	8.57	16.84	16.99

Conclusions

- We show that invariance training helps the ASR system to generalize better to unseen noise conditions and improves word error rate when a small number of noise types are seen during training.
- Related work [Shu16] investigates a similar approach for domain adaptation to noise types and the SNR levels on an in-house dataset based on WSJ.
- Our experiments show that in contrast to the image recognition task, in speech recognition, the gradient of the L_3 term is unreliable and noisy.
- Future research includes enhancements to the domain adaptation network while exploring alternative network architectures and invariance-promoting loss functions.

References

[GL14] Yaroslav Ganin and Victor Lempitsky, *Unsupervised domain adaptation by backpropagation*, ArXiv e-prints (2014).

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, *Generative adversarial nets*, Advances in Neural Information Processing Systems 27 (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), Curran Associates, Inc., 2014, pp. 2672–2680.

[Shu16] Yusuke Shunohara, *Adversarial multi-task learning of deep neural networks for robust speech recognition*, Interspeech 2016 (2016), 2369–2372.