Adversarial Training of Invariant Features for Speech Recognition

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

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

serdyuk@iro.umontreal.ca

Abstract

Recent advances in domain adaptation allow us to construct networks that are invariant to certain labeled factors. The reverse gradient algorithm uses an adversarial learning approach to train a network to produce the desired label as well as to deceive a second classifier that is trained on adaptation labels. This is tightly connected to the ideas behind generative adversarial networks (GANs). Therefore, some insights for training GANs are needed for the reverse gradient's successful application. We adapt this approach to train an acoustic system that is invariant to undesirable factors such as the recording environment noise type. We conduct experiments on a popular dataset.

Index Terms: speech recognition, adversarial training, reverse gradient

1. Introduction

One of the most challenging aspects of automatic speech recognition (ASR) is the mismatch between the training and testing acoustic conditions. During testing, a system may encounter new recording conditions, microphone types, speakers, accents and types of background noises. Furthermore, even if the test scenarios are seen during training, there can be significant variability in their statistics. Thus, it's important to develop ASR systems that are invariant to unseen acoustic conditions.

Several model and feature based adaptation methods such as Maximum Likelihood Linear Regression (MLLR), feature-based MLLR and i-vectors (Saon et al., 2013) have been proposed to handle speaker variability; and Noise Adaptive Training (NAT; Kalinli et al., 2010) and Vector Taylor Series (VTS; Un et al., 1998) to handle environment variability. With the increasing success of Deep Neural Network (DNN) acoustic models for ASR investigated in (Hinton et al., 2012; Seide et al., 2011; Sainath et al., 2011), and in more recent works (Miao et al., 2015; Sainath et al., 2015) complicated structure of acoustic conditions is modeled within a single network. This allows us to take advantage of the network's ability to learn highly non-linear feature transformations, with greater flexibility in constructing training objective functions that promote learning of noise invariant representations. The main idea of this work is to force the acoustic model to learn a representation which is invariant to noise conditions, instead of explicitly using noise robust acoustic features (Section 4). This type of noise-invariant training requires noise-condition labels during training only. It is related to the idea of generative adversarial networks (GAN) and the gradient reverse method proposed by Goodfellow et al. (2014) and Ganin and Lempitsky (2014) respectively (Section 2). Then we discuss previous work in Section 3. We present results on the Aurora-4 speech recognition task in Section 5 and summarize our findings in Section 6.

2. Background

Generative Adversarial Networks consist of two networks: the generator and the discriminator. 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 "counterfeit" generated image and the "genuine" image taken from the dataset. Thus, the discriminator is just a classifier network with a sigmoid output layer and can be trained with gradient back-propagation. This gradient can be propagated further to the generator network (assuming that the output of the generator is continuous).

The two networks in the GAN setup are competing with each other: the generator is trying to deceive the discriminator network, while the discriminator tries to do its best to recognize if there was a deception, similar to adversarial game-theoretic settings. Formally, the objective function of GAN training is

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

$$\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \tag{2}$$

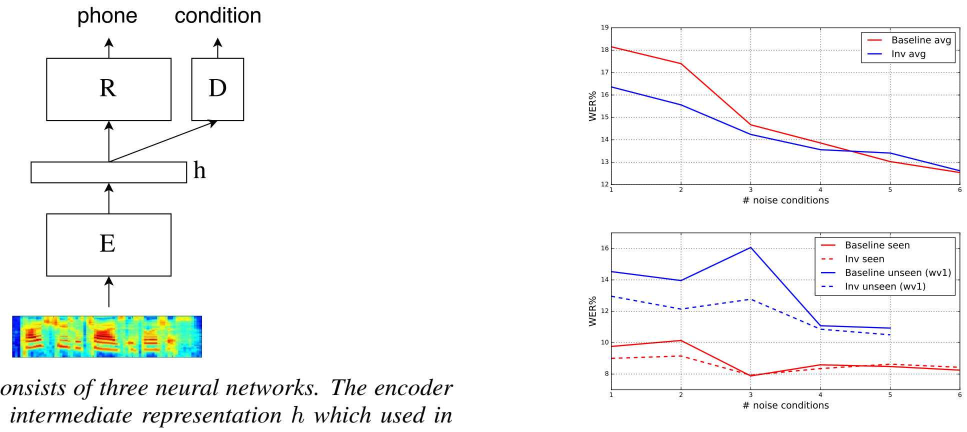
The maximization over the discriminator D forms a usual cross-entropy objective, the gradients are computed with respect to the parameters of D. An important property of this objective is that the gradient of the composition $D(G(\cdot))$ is well defined. Therefore the generator G can be trained with the back-propagation algorithm using the gradient signal from the discriminator D. The generator G is minimizing the classification objective using the gradients propagated through the second term. The minimization over G makes it produce samples which D classifies as originating from the train data.

Several practical guidelines were proposed for optimizing GANs by Radford et al. (2015) and further explored by Salimans et al. (2016). The second term of Eq. 2 is frequently exchanged with

$$-\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(D(G(\boldsymbol{z})))]. \tag{3}$$

The gradient of this term has the same sign but it has better properties. We experimented with several losses for our discriminator and chose one similar to Eq. 2 but we maintain the negative sign in our notation to demonstrate the general direction of the gradient of the discriminator loss.

Optimization of the GAN objective corresponds to a minimization of the so-called Jenson-Shannon divergence between the data distribution and the distribution represented by the generator. This divergence is zero if and only if the two distributions are identical. From this perspective, we can say that the discriminator provides a training signal which allows one to match



(a) *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.*

(b) *Top: 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).*

Figure 1: *The model and summarized results.*

distributions in general. In this work we are interested in matching feature distributions of different noise conditions.

Prior work by Ganin and Lempitsky (2014) proposed a method for 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 a hidden representation that is invariant to the domain labels and still 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 specific distributions similar to each other.

The network is trained with three goals: the hidden representation h should be helpful for the target classifier, harmful for the domain classifier, and the domain classifier should have a good classification accuracy. More formally, the authors define the loss function as

$$L = L_1(\hat{y}, y; \theta_R, \theta_E) + \alpha L_2(\hat{d}, d; \theta_D) - \beta L_3(\hat{d}, d; \theta_E), \quad (4)$$

where y is the ground truth class, d is the domain label, corresponding hat variables are the network predictions, and θ_E, θ_R and θ_D are the subsets of parameters for the encoder, recognizer and the domain classifier networks respectively. The hyper-parameters α and β denote the relative influence of the loss functions terms.

3. Related Work

Neural networks display some robustness towards different recording conditions and speaker by themselves and the effectiveness of representations produced by a neural network for internal noise reduction is discussed by Yu et al. (2013a). This work sets a baseline for experiments on the Aurora-4 dataset.

To be even more robust with respect to different recording conditions and speakers, one can either aim to adapt the model parameters to these new situations or to learn representations which are invariant to them. Most approaches so far, are based on adaptation. Unfortunately, many of the adaptation methods which have been designed for GMM-HMM systems cannot be applied to DNN-based systems. For this reason, a large body of recent work has investigated new adaptation methods for neural networks.

Some of the linear and affine transformations used for GMM adaptation can still be applied to neural networks when one limits these adaptations to the very last layer, as was shown in Yao et al. (2012). The improvements seemed to rely mostly on the adaptation of the bias parameters. A clear downside of this approach is that it doesn't utilize the representational power provided by the non-linear multi-layer structure of deep neural networks.

Many neural network speaker adaptation methods are based on i-vectors. The most common way to exploit i-vectors is by providing them as additional inputs (Senior and Lopez-Moreno, 2014; Saon et al., 2013). This means that i-vectors also need to be available during testing and this may not always be practical.

In the specific case of noise robustness, one can use speech enhancement to obtain more robust features and neural networks have been used for this for decades (Knecht et al., 1995). A further step in this direction is to train systems to recognize speech and perform speech enhancement jointly (Narayanan and Wang, 2014). We argue that speech enhancement is a more difficult task to learn than invariance to certain noise conditions and more likely to suffer from overfitting when the amount of available noisy data is limited. Another downside of enhancement approaches is that there is no reason to expect them to generalize well to noise conditions that were not available during training.

Another adaptation strategy is to retrain the model parameters using a very small number of utterances from the new domain while making sure that the model doesn't stray too far away from the parameter values that were obtained from the train data (Yu et al., 2013b). The adaptation to new domains can also be kept under control by limiting the number of parameters which are adapted to the new data or by limiting the adaptation to a rescaling of the hidden unit activations (Swietojanski and Renals, 2014). The most important difference between these approaches and our work is that we don't adapt the model parameters at all after training. In many situations this may not be possible and retraining of models may require more from the hardware on which the speech recognition system is implemented.

Recently, in a work by Shinohara (2016) a multi-layer sigmoidal network was trained in an adversarial fashion on an in-house transcription task corrupted by noise. This work is very similar to our approach but we evaluate our methods on more challenging benchmark while investigating different numbers of noise conditions. Our work also differs due to the use of more modern rectifier activation based classification networks. Finally, in other very recent work (Saon et al., 2017) a form of adversarial training with a network that predicts an i-vector using the mean square error loss was used.

4. Invariant Representations for Speech Recognition

Most ASR systems are DNN-HMM hybrid systems. The context dependent (CD) HMM states (acoustic model) are the class labels of interest. The recording conditions, speaker identity, or gender represent the domains in GANs. The task is to make the hidden layer representations of the HMM state classifier network invariant with respect to these domains. We hypothesize that this adversarial method of training helps the HMM state classifier to generalize better to unseen domain conditions and requires only a small additional amount of supervision, i.e., the domain labels.

Figure 1a depicts the model, which is same as the model for the gradient reverse method. It is a feed-forward neural network trained to predict the CD-HMM state, with a branch that predicts the domain (noise condition). This branch is discarded in the testing phase. In our experiments we used the noise condition as the domain label, merging all noise types into one label and defining 'clean' as the other label. Our training loss function is Eq. 4 with L_3 set to $d \log(1 - \hat{d}) + (1 - d) \log(\hat{d})$ for stability during training. L_3 term maximizes the probability of an incorrect domain classification in contrast to the gradient reverse where the correct classification is minimized. The terms L_1 and L_2 are regular cross-entropies which are minimized with corresponding parameters θ_E and θ_D. For simplicity, we use only a single hyper-parameter – the weight of the third term.

5. Experiments

We experimentally evaluated our approach on the well-benchmarked Aurora-4 (Parihar and Picone, 2002) noisy speech recognition task. Aurora-4 is based on the Wall Street Journal corpus (WSJ0). It contains noises of six categories which were added to the clean data. Every clean and noisy utterance has been filtered to simulate the phone quality recording using P.341 (Parihar and Picone, 2002). The training data contains 4400 clean utterances and 446 utterances for each noise condition, i.e., a total of 2676 noisy utterances. The test set consists of clean data, data corrupted by 6 noise types, and data recorded with a different microphone for both the clean and noisy conditions.

For both the clean and noisy data, we extracted 40-dimensional Mel-filterbank features with their deltas and delta-deltas spliced over ±5 frames, resulting in 1320 input features that were subsequently mean and variance normalized. The baseline acoustic model was a 6-layer DNN with 2048 rectified linear units at every layer. It was trained using momentum-accelerated stochastic gradient descent for 15 epochs with new-bob annealing (the learning rate is halfed if no improvement on the validation set, as in Morgan and Bourlard, 1995; Sainath et al., 2011).

To evaluate the impact of our method on generalization to *unseen* noises (the most typical situation in practice), we performed 6 experiments with different sets of noises seen during training. The networks were trained on clean data, with each noise condition added one-by-one in the following order: airport, babble, car, restaurant, street, and train. The last training group included all noises and therefore matched the standard multi-condition training setup. For every training group, we trained the baseline and the invariance model, where we branched out at the 4^{th} layer to a binary classifier predicting clean versus noisy data. Due to the imbalance between amounts of clean and noisy utterances, we had to oversample noisy frames to ensure that every mini-batch contained equal number of clean and noisy speech frames.

Table 1 summarizes the results. Figure 1b visualizes the word error rate (WER) for the baseline multi-condition training and invariance training as the number of seen noise types varies. We conclude that the best performance gain is achieved when a small number of noise types are available during training. It can be seen that invariance training is able to generalize better to unseen noise types compared with multi-condition training. In practice, it's only possible to train on a small fraction of all the possible noise conditions in the world, so this apparent ability to generalize to unseen conditions is a promising result.

We note that our experiments did not use layer-wise pre-training, commonly used for small datasets. The baseline WERs reported are very close to the state-of-the-art. Our preliminary experiments on a pre-trained network (better overall WER) when using all noise types (last row of Table 1) for training show the same trend as the non-pretrained networks.

6. Discussion

This paper presents the application of generative adversarial networks and invariance training for noise robust speech recognition. We show that invariance training helps the ASR system to generalize better to unseen noise conditions and improves the word error rate when a small number of noise types are seen during training. Our experiments show that in contrast to the image recognition task, in speech recognition, the domain adaptation network suffers from underfitting. Therefore, the gradient of the L_3 term in Eq. 4 is unreliable and noisy. Future research includes enhancements to the domain adaptation network while exploring alternative network architectures and invariance-promoting loss functions.

7. Acknowledgments

We would like to thank Yaroslav Ganin, David Warde-Farley for insightful discussions, developers of Theano (Theano Development Team, 2016), Blocks, and Fuel (van Merriënboer et al., 2015) for great toolkits.

8. References

G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors." in *ASRU*, 2013, pp. 55–59.

O. Kalinli, M. L. Seltzer, J. Droppo, and A. Acero, "Noise adaptive training for robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1889–1901, 2010.

Table 1: *Average word error rates (WER%) on Aurora-4 dataset on all test conditions, including seen and unseen noise and unseen microphone. The first column specifies the number of noise conditions used for the training. The results in the last row are from 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	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

C. K. Un, N. S. Kim *et al.*, "Speech recognition in noisy environments using first-order vector taylor series," *Speech Communication*, vol. 24, no. 1, pp. 39–49, 1998.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *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.

F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks." in *Interspeech*, 2011, pp. 437–440.

T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A.-r. Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 30–35.

Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.

T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Proc. Interspeech*, 2015.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *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. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *ArXiv e-prints*, Sep. 2014.

A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," *arXiv preprint arXiv:1606.03498*, 2016.

D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks – studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.

K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 366–369.

A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 225–229.

W. G. Knecht, M. E. Schenkel, and G. S. Moschytz, "Neural network filters for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 6, pp. 433–438, 1995.

A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2504–2508.

D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7893–7897.

P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 171–176.

Y. Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition," *Interspeech 2016*, pp. 2369–2372, 2016.

G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim *et al.*, "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.

N. Parihar and J. Picone, "Aurora working group: Dsr front end lvcsr evaluation au/385/02," *Inst. for Signal and Information Process, Mississippi State University, Tech. Rep*, vol. 40, p. 94, 2002.

N. Morgan and H. Bourlard, "Continuous speech recognition," *IEEE signal processing magazine*, vol. 12, no. 3, pp. 24–42, 1995.

Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688

B. van Merriënboer, D. Bahdanau, V. Dumoulin, D. Serdyuk, D. Warde-Farley, J. Chorowski, and Y. Bengio, "Blocks and fuel: Frameworks for deep learning," *CoRR*, vol. abs/1506.00619, 2015. [Online]. Available: http://arxiv.org/abs/1506.00619