CYCLEGAN BANDWIDTH EXTENSION ACOUSTIC MODELING FOR AUTOMATIC SPEECH RECOGNITION

David Haws, Xiaodong Cui

IBM Research AI
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

ABSTRACT

Although narrowband (NB) and wideband (WB) speech data primarily differ in sampling rate, these two common input sources are difficult to simultaneously model for automatic speech recognition (ASR). Meanwhile, cycle consistent generative adversarial networks (CycleGANs) have been shown value in a number of acoustic tasks such as mapping between domains due to their powerful generators. We apply CycleGAN to the task of bandwidth extension (BWE) and test a variety of architectures. The CycleGANs produce encouraging losses and reconstructed spectrograms. In order to further reduce word error rates (WER) we add an additional discriminative loss to the CycleGAN BWE architecture. This more closely matches our ASR goal and we show gains in WER compared to a standard BWE model discriminatively trained only to map from upsampled narrowband (UNB) to WB data.

Index Terms— speech recognition, deep neural networks, bandwidth extension, cycle consistent generative adversarial networks, acoustic modeling

1. INTRODUCTION

Narrowband (NB) and wideband (WB) data speech data differ primarily in the sampling rate. However, their spectral characteristics are distinct enough that it is difficult to train an automatic speech recognition (ASR) model that handles both these domains simultaneously. Training in only one domain and then upsampling or downsampling input data accordingly does not yield sufficient accuracy. That is, training a WB ASR model and then upsampling NB data, or conversely, training a NB ASR model and then downsampling WB data is insufficient. The goal is to have one model that is capable of handling both domains.

In this paper we propose a bandwidth extension (BWE) model that is trained using a cycle consistent generative adversarial network (CycleGAN). CycleGANs have increased in popularity recently due to their gains in a variety of tasks in vision and speech. CycleGANs jointly train two powerful generators one mapping from domain A to domain B and the other mapping from domain B to domain A. Moreover, they impose a *cycle loss* such that the composition of each map is close to the original input data. CycleGANS are thus able to better take advantage of both domains and the joint training of the generators often leads to better results than simply training a single mapping one direction between two domains. We train CycleGAN for the BWE task and show that the reconstruction losses and spectrograms indeed look promising.

Typical BWE techniques rely on minimizing a reconstruction error such as minimum mean square error, L1, or L2 loss. However, since the goal is to improve ASR word error rates (WER), we also explore using a discriminative loss[1]. We train a standard BWE model using this discriminative loss. We compared this to a CycleGAN trained using standard losses in addition to the discriminative loss. This helps guarantee the performance of the BWE more closely matches the goal of improving WER.

The paper is organized as follows. Section 2 discusses related work on CycleGANS and BWE. Section 3 gives the mathematical formulation of CycleGANs and BWE. Section 4 provides details on the model architecture and system configurations. Experimental results are provided in section 5. Lastly a summary is given in section 6.

2. RELATED WORK

CycleGANS first found use in vision mapping between two domains of images, such as horses and zebras, and pictures and paintings [2]. Since then they have been used in a variety of vision and speech tasks. For example [3] explored mapping noisy to clean data as well as accented data. In [4] [5] the authors explored using CycleGANs for gender mapping. CycleGANs have even been used for dereverberation [6].

BWE is a still evolving field of inquiry in speech and signal processing. NB speech signals, i.e. telephony speech signals, suffer from degraded quality due to the lack of high frequency spectral information eliminated by the low-pass band limitation of communication channels. Many researchers have studied BWE in order to improve quality and intelligibility[7, 8, 9, 10, 11, 12]. The goal of BWE is to estimate the missing high frequency spectral components and thus "extend" the bandwidth of the signal. We do acknowledge there are additional spectral changes because of channel

Model	Data	WER
CNN WB	WB	14.5%
—	UNB	19.8%
—	UNB + BWE	16.4%
CNN NB	NB	15.9%
CNN UNB	WB	19.2%
—	UNB	16.5%

Table 1. WER for acoustic models trained on WB, UNB, and NB data. Model name "CNN WB" means a CNN acoustic model trained on WB data, etc. Second column "Data" specifies the test data where "UNB + BWE" means UNB passed through a discriminatively trained BWE model.

effects, but these will not be addressed in this paper and we leave to future work.

In this work we explore using a variety of neural networks structures for CycleGANs for BWE. Specifically we study CNNs, LSTMs, and Residual Networks for CycleGAN BWE as well as exploring the parameter space. Moreover, we analyze adding a discriminative loss based on a trained acoustic model to further improve the WER.

3. BANDWIDTH EXTENSION & CYCLEGAN

3.1. CycleGAN

CycleGANs are a set of models which learns a mapping between two domains A and B, in our case upsampled narrowband (UNB) and WB data [2]. There are two generators $G_{A\to B}$ and $G_{B\to A}$ which map from A to B and B to A respectively. Additionally there are two discriminators D_A and D_B whose role is to distinguish a true sample from A versus a fake and a true sample from B versus a fake respectively. The generator $G_{A\to B}$ is trained such that the distribution of features $p_b(\mathbf{b})$ is indistinguishable from the "fake" features $G_{A\to B}(\mathbf{a})$, and similarly for the generator $G_{B\to A}$ and the distribution $p_a(\mathbf{a})$. The adversarial relationship of the generator to the discriminator is given by

$$\min_{G_{A\to B}} \max_{D_B} \mathbb{E}_{\mathbf{a}\sim p_a(\mathbf{a})}[\log D_B(\mathbf{t})] + \mathbb{E}_{\mathbf{b}\sim p_b(\mathbf{b})}[\log(1 - D_B(G_{A\to B}(\mathbf{a})))].$$

In their original form [2], the discriminators were trained to output a probability that the input data was from the true domain. However, others found that such discriminators lead to stability problems during training and suggested to use the *Earth-Mover* or Wasserstein-1 distance, which is roughly the minimum cost of transporting mass in order to transform two distributions[13, 14]. The discriminators are constrained to be 1-Lipshitz by enforcing that their gradients close to 1. Recall that a continuous function f is L-Lipshitz if $|f(x_1)-f(x_2)| \leq L|x_1 - x_2| \ \ \forall x_1, x_2$ [15]. Thus the WGAN loss is defined as

$$\begin{aligned}\mathcal{L}_{WGAN}(G_{A\to B}, D_B) = \mathbb{E}_{\mathbf{b}\sim p_b(\mathbf{b})}[D_B(\mathbf{b})] \\ - \mathbb{E}_{\mathbf{a}\sim p_a(\mathbf{a})}[D_B(G_{A\to B}(\mathbf{a})] \\ - \beta\mathbb{E}_{\tilde{\mathbf{b}}\sim p_b(\hat{\mathbf{b}})}[(||\Delta_{\hat{\mathbf{b}}} D_B(\hat{\mathbf{b}})||_2 - 1)^2] \\ - \beta\mathbb{E}_{\tilde{\mathbf{a}}\sim p_a(\hat{\mathbf{a}})}[(||\Delta_{\hat{\mathbf{a}}} D_A(\hat{\mathbf{a}})||_2 - 1)^2],\end{aligned}$$

where the last two terms are the gradient penalty and $\tilde{\mathbf{b}} := \alpha\mathbf{t} + (1-\alpha)G_{A\to B}(\mathbf{a})$ such that $\alpha \sim U(0,1)$, $\mathbf{a} \sim p_A(\mathbf{a})$, and $\mathbf{t} \sim p_B(\mathbf{t})$. We define $\tilde{\mathbf{a}}$ similarly.

The key to CycleGAN's success is the cycle consistency loss which ensures that the composition of the generators yields a feature close to the original feature. Specifically we have

$$\begin{aligned}\mathcal{L}_{cyc} := \mathbb{E}_{\mathbf{a}\sim p_a(\mathbf{a})}[||G_{A\to B}(G_{B\to A}(\mathbf{a})) - \mathbf{a}||_1] \\ + \mathbb{E}_{\mathbf{b}\sim p_b(\mathbf{b})}[||G_{B\to A}(G_{A\to B}(\mathbf{b})) - \mathbf{b}||_1].\end{aligned}$$

Thus, summing $\mathcal{L}_{WGAN}(G_{A\to B}, D_B)$, $\mathcal{L}_{WGAN}(G_{B\to A}, D_A)$, and $\mathcal{L}_{cyc}$, we can train the CycleGAN by alternating between updating the discriminators and then the generators.

3.2. Bandwidth Extension

BWE techniques typically are based on minimizing a reconstruction error such as the minimum mean square error, L1, or L2 loss. For example

$$\theta^* = \arg\min \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{y}_i - f_\theta(\mathbf{x})||_2^2$$

where $\mathbf{x}_i$ is the WB data, $\mathbf{y}_i$ is the NB data, and f_θ is the BWE mapping. However, since our task is ASR classification we can further specialize the BWE loss. Given a trained WB acoustic model, we can discriminatively train a BWE network to minimize the cross entropy of mapped NB data with respect to the trained WB acoustic model. That is, we pass the NB data through the BWE model to be trained, then pass that data through the fixed WB acoustic model. From there we take the cross-entropy between this and the label. Discriminative BWE is discussed in detail in [1].

4. IMPLEMENTATION

4.1. Features

The sample rate of the WB speech and NB speech is 16KHz and 8KHz respectively. From the speech data[16], 40-dimensional logmel features are extracted and then a global cepstral mean normalization (CMN) followed by an utterance-based CMN is applied. We compute the static logmel features, as well as their delta and double deltas and input those

Model	WER	Discriminative Loss UNB	Discriminative Loss WB	Cycle Loss UNB	Cycle Loss WB	Identity Loss UNB	Identity Loss WB
Resnet	18.1%			0.120	0.074	0.038	0.060
Resnet(disc)	16.1%	0.85	1.06	0.110	0.068	0.035	0.054
LSTM(disc)	16.3%	0.87	1.08	0.108	0.118	0.056	0.042
CNN(disc)	20.7%	1.28	1.53	0.290	0.261	0.127	0.137

Table 2. Word Error Rate (WER) and losses for the BWE task using CycleGAN trained networks. Reported are the discriminative loss, cycle loss, and identify loss. Cycle Loss UNB is the L_1 loss of the a UNB sample mapped to WB then mapped back to UNB using the generators versus its input data. Identify Loss UNB is the L_1 loss of a UNB sample UNB passed through the map from WB to UNB versus its input data.

into the CMNs. Lastly we provide a temporal context of 11 frames. The UNB speech signals are passed through the WB Mel filter banks after upsampling in the time domain, which gives rise to zeros in the outputs of the upper Mel filter bins (the zero-padding effect).

4.2. GAN models

Multiple network structures were explored, but we only present three here. First, a Resnet similar to [2] and [3] was trained. For convention we will write a convolutional layer with kernel k, stride s, and input channels i, and output channels o as `conv_kxk_sxs_ixo`. The model Resnet9 consists of `conv_3x3_1x1_3x32`, leaky relu with slope 0.2 (LRelu), `conv_3x3_2x2_32x64`, LRelu, instance normalization (IN), `conv_3x3_2x2_64x128`, LRelu, IN. Next is a series of 9 residual blocks. Each residual block is composed of `conv_3x3_1x2_128x128`, LRelu, IN, `conv_3x3_1x1_128x128`, LRelu, IN, with a residual connection. Following, the 9 residual blocks is a 3x3 deconvoluational layer with stride 2x2 input channels 128 and output channels 64 `deconv_3x3_2x2_128x64`, LRelu, IN, `deconv_3x3_2x2_63x32`, LRelu, IN, `conv_3x3_1x2_32x3`. Additionally, the Resnet network was modified to only map to the upper 9 bands of the feature space. [1]

The second generator network explored was a CNN model. This CNN consists of 4 convolutional layers, 2 max-pooling layers and 3 fully connected (FC) layers. Every 2 convolutional layers are followed by one max-pooling layer. The first 2 convolutional layers use 3x3 kernels with a stride 1x1 and padding 1x1. The second 2 convolutional layers again use 3x3 kernels with a stride 1x1 and padding 1x1. The 2 max-pooling layers use 2x2 kernels with a stride 1x1. The 3 FC layers have $1,024$ hidden units. All activation functions are Relu except the last FC layer which uses tanh. The first two convolutional layers have 128 feature maps, and the second two convolutional layers have 256 feature maps.

Thirdly, we explored as generators a bidirectional LSTM with 4 layers and 512 hidden units followed by a FC layer to reconstruct a feature of the correct input dimension. As in the case of the Resnet above, we design the LSTM architecture to only map to upper 9 bands.

For the discriminator networks we used `conv_3x3_1x1_1x64`, LRelu, `conv_3x3_2x2_64x128`, LRelu, followed by a FC layer with 512 nodes, LRelu, a FC layer with 512 nodes, LRelu, and a FC layer with 1 node.

4.3. Acoustic Models

CNN acoustic models are used for WB baseline, and NB baseline, which have the same configuration. There are 2 convolutional layers and each convolutional layer is followed by a max-pooling layer. The first convolutional layer uses 5x5 kernels with a stride is 1x1 and padding 2x2. The second convolutional layer uses the same kernel, stride and padding sizes as those of the first convolutional layer. Both max-pooling layers use a kernel of 2x2 and stride of 2x2. On top of the convolutional and pooling layers are 3 FC layers with $1,024$ hidden units. All activation functions are Relu except the last FC layer which uses sigmoid. The output softmax layer has $9,300$ output units. The first two convolutional layers have 128 layers, while the second two convolutional layers have 256 layers.

4.4. Standard Unidirectional BWE Model

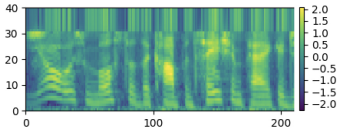
A direct BWE mapping network is constructed which only maps UNB to WB using the discriminative loss described in Subsection 3.2. This BWE is modeled exactly as the CNN described in Subsection 4.2

5. EXPERIMENTAL RESULTS

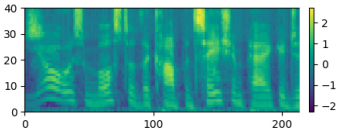
5.1. Baseline

For our experiments we used a 50hr subset of Broadcast News data which is provided at a 16KHz sampling rate[16]. We downsampled the data to 8KHz then upsampled to create corresponding UNB training and test sets. Baseline acoustic models described in 4.3 were trained on the WB, NB, and UNB data. We can see in Table 1 that domain mismatch for

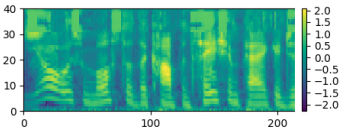
[1] We did explore mapping to the entire feature space, but this led to non-optimal performance.



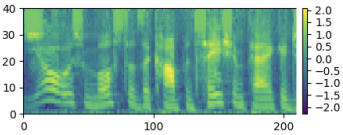
(a) Upsampled Narrowband test data



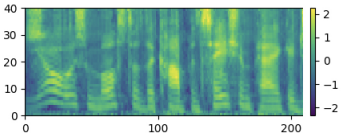
(b) True Wideband test data



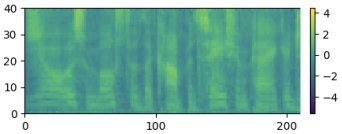
(c) Resnet CycleGAN UNB to WB without disc loss



(d) Resnet CycleGAN UNB to WB with disc loss



(e) LSTM CycleGAN UNB to WB with disc loss



(d) CNN CycleGAN UNB to WB with disc loss

Fig. 1. Logmel spectrograms of UNB data versus CycleGAN BWE models trained with discriminative loss.

the UNB trained acoustic model leads to a 2.7% degradation in WER, and for the WB trained acoustic model leads to a 5.3% degradation.

5.2. GAN

We trained three GAN models, each with a different generator networks: Resnet, LSTM, and CNN. Following recipes in [13, 14] we trained the discriminators D_A and D_B 4 times before updating the generators $G_{A\to B}$ and $G_{B\to A}$. Recall that the discriminators are trained to maximize the loss $\mathcal{L}(G_{A\to B}, G_{B\to A}, D_A, D_B)$, while the generators are trained to minimize said loss. We also added an additional identity constraint $\lambda L_{id_A} = ||G_{A\to B}(\mathbf{b}) - \mathbf{b}||_1$ and $\lambda L_{id_B} = ||G_{B\to A}(\mathbf{a}) - \mathbf{a}||_1$. For loss coefficients we set the cycle loss coefficient α to 10.0, the WGAN gradient coefficient β to 100.0, and the identity loss coefficient λ to 0.5

We report the model type, number of steps and losses for our CycleGAN networks in Table 2. Recall that the Resnet and LSTM only generate the upper 9 bands of the 40 logmel features. We found this performed better than generating the entire band using Resnet or LSTM. We explored many other network structures and parameter selections, but only present a few here. We note a WER of 18.1% for the Resnet model that had no acoustic discriminative training. This improves over simply using the UNB by 1.8% absolute as shown in Table 1. Moreover the cycle losses and identity losses converge. We show the logmel spectrogram of a test example in Figure 4.1. We can see in subfigure (c) that this Resnet is beginning to learn to fill in the upper band.

Since our desire is to improve ASR accuracy we next added the acoustic discriminative loss to the cycleGAN generators and assign a coefficient 100.0 to this loss. Additionally, a BWE model which only maps UNB data to WB data was discriminatively trained using the loss described in subsection 3.2. We will refer to this as the *standard BWE model*. For the discriminative loss the two previously trained WB and UNB models were used to generate labels for the training data. WER are given in Table 1. The standard BWE model gives a WER of 16.4% and thus improves 3.4% absolute. We see in Table 2 that the Resnet is able to outperform the standard BWE and achieve a WER of 16.1% on the BWE task. The LSTM is able to outperform the standard BWE and achieve a WER of 16.3% WER on the BWE task. See Figure 4.1.

6. SUMMARY

Creating unified acoustic models able to handle diverse domains is crucial for modern ASR systems. For the NB vs WB mismatch problem previous works have used BWE which maps solely from one domain to another, typically using regression losses such as minimum mean square error. CycleGANs present machinery capable of utilizing mappings from both UNB to WB as well as WB to UNB. Combined with the addition of the discriminative loss, we have shown CycleGANs are better able to handle BWE over a standard model.

We explored dozens of CycleGAN configurations during the process of writing this paper, but have only scratched the surface of possibilities of these models. Specifically, we plan to explore more network configurations and parameter selection. Moreover, the LSTM model showed particular promise visually and with other losses and we hope to reconcile this with improved WER in the future.

7. REFERENCES

[1] Khoi-Nguyen C. Mac1, Xiaodong Cui, Wei Zhang, and Michael Picheny, "Large-scale mixed-bandwidth deep neural network acoustic modeling for automatic speech recognition," Submitted to ICASSP 2019, 2018.

[2] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

[3] Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, "Cross-domain speech recognition using nonparallel corpora with cycle-consistent adversarial networks," in *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*. IEEE, 2017, pp. 134–140.

[4] Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," *arXiv preprint arXiv:1804.00425*, 2018.

[5] Ehsan Hosseini-Asl, Yingbo Zhou, Caiming Xiong, and Richard Socher, "A multi-discriminator cyclegan for unsupervised non-parallel speech domain adaptation," *arXiv preprint arXiv:1804.00522*, 2018.

[6] Ke Wang, Junbo Zhang, Sining Sun, Yujun Wang, Fei Xiang, and Lei Xie, "Investigating generative adversarial networks based speech dereverberation for robust speech recognition," *arXiv preprint arXiv:1803.10132*, 2018.

[7] B. Iser, W. Minker, and G. Schmidt, *Bandwidth extension of speech signals*, Springer, 2008.

[8] N. Prasad and T. Kumar, "Bandwidth extension of speech signals: a comprehensive review," *International Journal of Intelligent Systems Technologies and Applications*, vol. 2, no. 2, pp. 45–52, 2016.

[9] F. Nagel and S. Disch, "A harmonic bandwidth extension method for audio codecs," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 145–148.

[10] H. Pulakka and P. Alku, "Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband Mel spectrum," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2170–2183, 2011.

[11] C. Liu, Q.-J. Fu, and S. Narayanan, "Effect of bandwidth extension to telephone speech recognition in cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 26, no. 5, pp. 77–83, 2018.

[12] Z.-H. Ling, Y. Ai, Y. Gu, and L.-R. Dai, "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 125, no. 2, pp. 883–894, 2009.

[13] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[14] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[15] Houshang H Sohrab, *Basic real analysis*, vol. 231, Springer, 2003.

[16] Tara N. Sainath, Brian Kingsbury, Abdel rahman Mohamed, George E. Dahl, George Saon, Hagen Soltau, Tomas Beran, Aleksandr Y. Aravkin, Bhuvana Ramabhadran, Jonathan G. Fiscus, and et. al, "Improvements to deep convolutional neural networks for lvcsr," `https://arxiv.org/pdf/1309.1501.pdf`, 2013.