A Multi-Discriminator CycleGAN for Unsupervised Non-Parallel Speech Domain Adaptation

Ehsan Hosseini-Asl, Yingbo Zhou, Caiming Xiong, Richard Socher

Overview

Motivation

- Training a robust ASR system is highly dependent on factorizing linguistic features (text) from non-related variations, or adapting the inter-domain variations of source and target.
- Traditional domain adaptation methods (voice conversion) require parallel data of source and target that is difficult to obtain in practice.
- An unsupervised domain adaptation is desirable for building a robust ASR system.

Proposal

- A Cyclic adversarial learning model to adapt spectrogram representation across speech domains using frequency-dependent multi discriminators

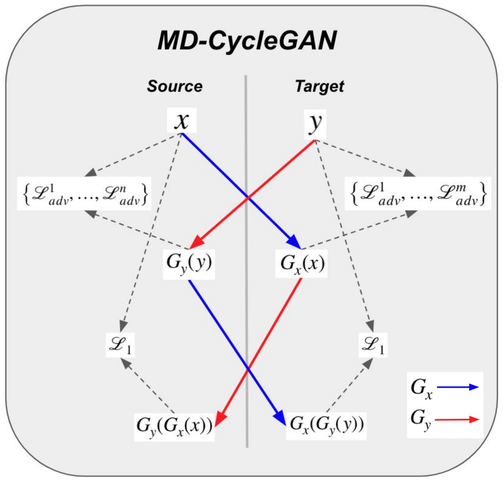
Result

- **Quantitatively:** relative 7.41% improvement of phoneme error rate on TIMIT, and 11.10% word error rate on WSJ, compare to baselines.
- **Qualitatively:** our model generates more natural sounding speech, when conditioned on data from the other domain.

Multi-Discriminators (MD)-CycleGAN

Model:

- Learning the frequency-based mapping functions (generators) $\{G_X, G_Y\}$
- Define multiple frequency-based discriminators $\left\{D_X^{f_{j\in n}}, D_Y^{f_{i\in m}}\right\}$
- $f_{j\in n}$ represents the j-th frequency-band of X domain with n frequency bands
- The frequency band selection in each domain can share a portion of frequency spectrum, or be exclusive, based on the domain spectrogram distribution.



Adversarial loss for pair of generator and discriminators $\left(G_X, D_Y^{f_{i\in m}}\right)$:

$$\mathcal{L}_{MD-CGAN}\left(G_X, D_Y^{f_{i\in m}}\right) = \mathbb{E}_{(y)\sim p_{data}(y)}\left[\sum_{i=1}^{m} \log D_Y^{f_i}(y)\right] + \mathbb{E}_{x\sim p_{data}(x), z\sim p_z(z)}\left[\sum_{i=0}^{m} \log\left(D_Y^{f_i}(x, G_X(z, x))\right)\right]$$

Total loss for MD-CycleGAN:

$$\mathcal{L}_{MD-CycleGAN} = \mathcal{L}_{MD-CGAN}(G_X, D_Y^{f_{i\in m}}) + \mathcal{L}_{MD-CGAN}(G_Y, D_X^{f_{j\in n}}) + \mathcal{L}_{cycle}(G_X, G_Y)$$

TIMIT Train → Test set domain adaptation

		Male (PER)			Female (PER)	
Model	Train	Val	Test	Train	Val	Test
-	Female (Baseline)	40.70	42.79	Male (Baseline)	35.70	30.69
CycleGAN	Female → Male	40.10	42.38	Male → Female	32.94	30.07
	Female& → Male	39.20	42.21	Male & → Female	31.29	29.03
MD-CycleGAN	Female → Male	**29.83**	33.46	Male → Female	28.80	25.44
	Female & → Male	30.01	**33.27**	Male & → Female	**25.98**	**24.13**
FHVAE [2]	-	-	-	Male + Z_1	-	26.20
	Male (baseline)	20.06	22.51	Female (baseline)	24.51	23.21

Test → Train set domain adaptation

		Train			
		Dataset		Dataset	
		TIMIT (PER)		WSJ (CER / WER)	
Test	Model	Male	Female	Male	Female
Male (baseline)	-	22.516	42.79	3.19 / 8.16	14.31 / 27.66
Male → Female	CycleGAN	-	43.427	-	-
	MD-CycleGAN	-	**37.00**	-	**6.82 / 15.68**
Female (baseline)	-	32.08	-	7.57 / 16.38	2.80 / 6.71
Female → Male	CycleGAN	**32.60**	23.215	-	-
	MD-CycleGAN	25.76	-	**5.93 / 13.18**	-

Analysis

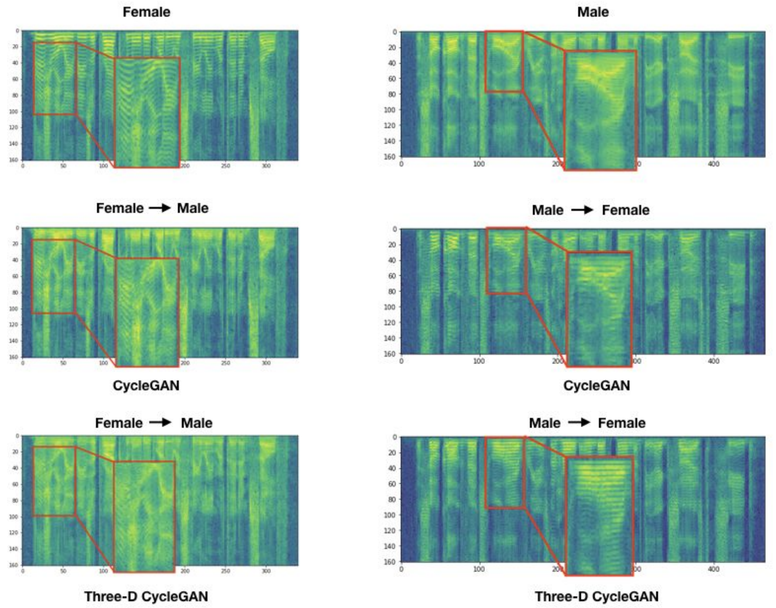
- Male/Female ratio: TIMIT=0.7/0.3 WSJ=0.5/0.5
- Train -> Test: ASR model is retrained on the adapted train set, to match the test distribution,
- Test -> Train: ASR model is fixed and evaluated on new test sets
- To evaluate the generalization of the trained generator: using TIMIT-trained generators on WSJ without retraining.
- ASR performance is significantly improved by reducing the gap to male and female baselines.
- checkerboard problem is mitigated, by learning the characteristic of target domain using multiple independent discriminators.

WSJ Train → Test set domain adaptation

	Male		Female	
Train	CER	WER	CER	WER
Female (baseline)	14.31	27.66	2.80	6.71
Female & → Male	**5.20**	**12.39**	-	-
Male (baseline)	3.19	8.16	7.57	16.38
Male & → Female	-	-	**4.22**	**9.46**

Spectrogram Adaptation

CycleGAN generator converges only by pretraining the discriminator first, unless the generator will learn identity mapping function. Three-D CycleGAN results are achieved without pretraining.



ASR Prediction

ASR prediction mismatch when train/test on different genders, and when adapting using Multi-Discriminator CycleGAN, on WSJ (eval92) dataset

		Train on Male
Test on Female	True	ASSOCIATED INNS KNOWN AS AIRCOA IS THE GENERAL PARTNER OF AIRCOA HOTEL PARTNERS AND HAS A ONE PERCENT GENERAL PARTNER INTEREST IN THE PARTNERSHIP
	Female	AFFECTED ENDS NONE IS AIR COLA IS THE GENERAL PARTNER OF ALOHA TELL PRINTERS AND HAS A ONE PERCENT GENERAL PARTNER INTEREST IN THE PARTNERSHIP
	Female→Male	ASSOCIATED INNS NONE HAS AIRCOA IS THE GENERAL PARTNER OF THE ARCO HOTEL PARTNERS AND AS A ONE PERCENT GENERAL PARTNER INTEREST IN THE PARTNERSHIP
	True	ALTHOUGH THOSE GAINS ERODED DURING THE AFTERNOON STOCK PRICES STAYED WITHIN A NARROW RANGE UNTIL THE LAST HALF HOUR OF TRADING
	Female	ALL THE THIS GAINS A ROLE DURING THE AFTERNOON STOCK PRICES STAYED WITHIN A NARROW RANGE UNTIL THE LAST THATCHER OF TRADING
	Female→Male	ALTHOUGH THOSE GAINS A ROLE DURING THE AFTERNOON STOCK PRICES STAYED WITHIN A NARROW RANGE UNTIL THE LAST TOUGH HOUR OF TRADING
	True	LA GUARDIA HAS ONLY FIFTY SEVEN GATES BUT AT PEAK HOURS DOZENS OF MORE PLANES MAY BE ON THE GROUND
	Female	THE GUIDE A HAS ONLY FIFTY SEVEN GATES THAT AT PEAK HOURS DOZENS OF MORE PLANES NAVY ON A GROUND
	Female→Male	THE GUARD A HAS ONLY FIFTY SEVEN GATES BUT AT PEAK HOURS DOZENS OF MORE PLANES MAY BE ON A GROUND
	True	HE SAID THE SALES HAD HAD A MAJOR PSYCHOLOGICAL IMPACT ON IRAN AND A NEGATIVE MILITARY IMPACT ON IRAQ
	Female	HE FED THE FAILS AND HAD A MAJOR PSYCHOLOGICAL INTACT ON IRAN AND A NEGATIVE MILITARY INTACT UNDER
	Female→Male	HE SAID THE SALES HAD HAD A MAJOR PSYCHOLOGICAL IMPACT ON IRAN AND A NEGATIVE MILITARY IMPACT ON IRAQ
	True	HE SAYS IT CAN GET CORNY TO SAY THAT MUSIC IS A UNIVERSAL LANGUAGE BUT IT REALLY IS
	Female	HE SAYS IT CAN GET CORN TO SANTA MUSIC IS A UNIVERSAL LANGUAGE THAT IT REALLY IS
	Female→Male	HE SAYS IT CAN GET CORNER TO SAY THE MUSIC IS A UNIVERSAL LANGUAGE BUT A REALLY IS
	True	GLASNOST HAS ALSO BEEN GOOD TO LAWRENCE LEIGHTON SMITH
	Female	FLAT NEST HAS ALSO BEING GOOD TO LAWRENCE LADEN SMITH
	Female→Male	GLASNOST HAS ALSO BIG GOOD TO LAWRENCE LADEN SMITH
	True	BIDS TOTALING SIX HUNDRED FIFTY ONE MILLION DOLLARS WERE SUBMITTED
	Female	BIDS TUMBLING SIX HUNDRED FIFTY ONE MILLION DOLLARS WERE SUBMITTED
	Female→Male	BIDS TOTALING SIX HUNDRED FIFTY ONE MILLION DOLLARS WERE SUBMITTED

		Train on Female
Test on Male	True	STRONGER PALM OIL PRICES HELPED OIL PRICES FIRM ANALYSTS SAID
	Male	STRONGER PALM ON PRICES HELPED OIL PRICE FROM ANALYSTS SO
	Male→Female	STRONGER PALM OIL PRICES HELPED OIL PRICES FIRM ANALYSTS SAID
	True	CONTACTS STILL INSIDE OWENS CORNING HELP TOO
	Male	CONTACTS STILL INSIDE OWENS CORN AND HELP TO
	Male→Female	CONTACTS STILL INSIDE OWENS CORNING HELP TO
	True	THE WARMING TREND MAY HAVE MELTED THE SNOW COVER ON SOME CROPS
	Male	THE WOMAN TREND MAYOR MELTED THE SNOW COVER ON SOME CROPS
	Male→Female	THE WARMING TREND MAY HAD MELTED TO SNOW COVER ON SOME CROPS
	True	HE MADE A SALES CALL HE SAYS
	Male	HE MADE A SALES CALL HE SERVES
	Male→Female	HE MADE A SALES CALL HE SAYS
	True	IT WASN'T A GIVEAWAY
	Male	IT WASN'T TO GIVE MORE
	Male→Female	IT WASN'T THAT GIVE AWAY
	True	VICE PRESIDENT BUSH MUST BE ESPECIALLY GRATEFUL FOR THE CHANGE OF SUBJECT ANYTHING WAS BETTER THAN THE DRUM-BEAT ABOUT PANAMA AND GENERAL NORIEGA
	Male	VICE PRESIDENT BUSH MUST BE ESPECIALLY GREAT FULL FOR THE CHANGE OF SUBJECT ANYTHING WAS BETTER THAN A DRUM-BEAT ABOUT PANAMA AND TRUMAN MILEAGE
	Male→Female	VICE PRESIDENT BUSH MUST BE ESPECIALLY GREAT FULL FOR THE CHANGE OF SUBJECT ANYTHING WAS BETTER THAN THE DRUM BEAT ABOUT PANAMA AND GENERAL NORIEGA
	True	IN NINETEEN EIGHTY FIVE PENNZOIL WON NEARLY ELEVEN BIL-LION DOLLARS IN DAMAGES AT TRIAL THE BIGGEST JUDGMENT EVER AWARDED A PLAINTIFF
	Male	IN NINETEEN EIGHTY FIVE PENNZOIL ONE NEARLY ELEVEN BIL-LION DOLLARS IN DAMAGES ARE THE BIGGEST GEORGE MEN EVEN ORDERED A PLAINTIFF
	Male→Female	IN NINETEEN EIGHTY FIVE PENNZOIL ONE NEARLY ELEVEN BIL-LION DOLLARS IN DAMAGES AT TRY THE BIGGEST JUDGMENT EVER AWARDED A PLAINTIFF
	True	BUT IT'S DIFFICULT TO SEE WHERE THE COMPANY GOES FROM HERE
	Male	BUT IT'S DIFFICULT TO SEE WHERE THE COMPANY GOT YEAR
	Male→Female	BUT IT'S DIFFICULT TO SEE WHERE THE COMPANY GOES FROM HERE
	True	THEY EXPECT COMPANIES TO GROW OR DISAPPEAR
	Male	THE DEBUT COMPANIES TO GO ON DISAPPEAR
	Male→Female	THEY EXPECT COMPANIES TO GROW OR DISAPPEAR

Contact

- Ehsan Hosseini-Asl, ehosseiniasl@salesforce.com
- Yingbo Zhou, yingbo.zhou@salesforce.com
- Caiming Xiong, cxiong@salesforce.com
- Richard Socher, rsocher@salesforce.com

Check out more of our work at https://einstein.ai

Sound Demos

Selected References

- [1] J.Y. Zhu, T. Park, P. Isola and A. A. Efros, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. ICCV *2017*.
- [2] W.-N. Hsu, Y. Zhang, and J. R. Glass, Unsupervised learning of disentangled and interpretable representations from sequential data, in *NIPS*, 2017.
- [3] S. Pascual, A. Bonafonte, and J. Serra, SEGAN: Speech enhancement generative adversarial network, in *INTERSPEECH*, 2017.
- [4] D. Michelsanti and Z.-H. Tan, Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification, in *INTERSPEECH*, 2017.
- [5] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks, in INTER- SPEECH, 2017.
- [6] T. Kaneko and H. Kameoka, Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks, *arXiv:1711.11293*, 2017.
- [7] C. Donahue, J. McAuley, and M. Puckette, Synthesizing audio with generative adversarial networks, *CoRR*, vol. abs/1802.04208, 2018.