Equal Accuracy Ratio for Fair CTC Speech Recognition

Heting Gao

University of Illinois at Urbana-Champaign

hgao17@illinois.edu

October 11, 2020

Overview

- Introduction
- Algorithm
- Experiment
- Conclusion

Fairness

- Machine learning algorithms can have bias, reducing opportunities for minority minority group.
    - Credit prediction models (whether to accept a loan application) may favor the old people. [Kamiran and Calders, 2009]
    - Speech recognition products have a higher accuracy over white speakers than black speakers. [Koenecke et al., 2020]
    - Speech recognition models have different accuracy over different dialects [Li et al., 2018]
- If we can identify and formulate the bias on different groups of people (of group attribute A), we may be able to train the model to explicitly reduce it.

Fairness

- Demographic Parity [Kamiran and Calders, 2009]

$$|p_{\hat{Y}|A}(1|0) - p_{\hat{Y}|A}(1|1)| = 0$$

- Equal Odd Gap [Hardt et al., 2016]

$$|p_{\hat{Y}|A,Y}(c|0,y) - p_{\hat{Y}|A,Y}(c|1,y)| = 0$$

- Equal Opportunity Gap [Hardt et al., 2016]

$$|p_{\hat{Y}|A,Y}(y|0,y) - p_{\hat{Y}|A,Y}(y|1,y)| = 0$$

- Predictive rate parity [Zafar et al., 2017]

$$|p_{Y|A,\hat{Y}}(1|0,y) - p_{Y|A,\hat{Y}}(1|1,y)| = 0$$

- These measures assumes binary tabular settings and do not naturally extend to sequence-to-sequence predictions

Algorithm

- Demographic parity is probably not very useful in speech recognition scenario as different groups of people can speak different things (Favorite vs Favourite).
- We adapt equal opportunity gap measure.

$$|p_{\hat{Y}|A,Y}(y|0,y) - p_{\hat{Y}|A,Y}(y|1,y)| = 0$$

1. Matched frames: $p_{\hat{Y}|A,Y}(y|a,y)$ could be measured on each frame. However matched frames would need a ground truth alignment, which are not required for CTC training
2. Matched transcription: $p_{\hat{Y}|A,Y}(y|a,y)$ could be measured using sets of waveforms, with exactly the same transcription. However dataset containing parallel transcriptions are rare.
3. Matched accuracy: $p_{\hat{Y}|A,Y}(y|a,y)$ could be measured using sentence accuracy of an ASR, for user group a, which requires the recognition accuracy is the same for different demographic groups

Equal Accuracy Ratio

- We use matched accuracy to compute accuracy for a user group a,

$$p_{\hat{Y}|A}(Y|a) = \sum_{y} p_{Y|A}(y|a) p_{\hat{Y}|A,Y}(y|a,y).$$

- The equal opportunity measure fairness is defined as

$$|p_{\hat{Y}|A}(Y|0) - p_{\hat{Y}|A}(Y|1)| = 0 \ \forall a, a'$$
$$|\ln p_{\hat{Y}|A}(Y|a) - \ln p_{\hat{Y}|A}(Y|a')| = 0 \ \ \forall a, a'.$$

- We call this measure as equal accuracy ratio
- We then define the equal accuracy ratio loss as

$$\mathcal{L}_{EAR} = \sum_{a,a'} |\ln p_{\hat{Y}|A}(Y|a) - \ln p_{\hat{Y}|A}(Y|a')|.$$

Equal Accuracy Ratio

- We do not have $\ln p_{\hat{Y}|A}(Y|a)$, but we can estimate it as

$$\ln p_{\hat{Y}|A}(Y|a) \approx \frac{1}{|S_a|} \sum_{x^{(i)},y^{(i)} \in S_a} \ln p_{\hat{Y}|X}(y^{(i)}|x^{(i)}),$$

- $\ln p_{\hat{Y}|X}(y^{(i)}|x^{(i)})$ is the CTC loss of ith sample
- We use equal accuracy loss as a regularization to the ordinary CTC loss in the training. The combined loss is defined as

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{EAR},$$

Equal Accuracy Ratio

- In $\mathcal{L}_{EAR}$ we have an absolute difference,

$$\mathcal{L}_{EAR} = \sum_{a,a'} \left| \ln p_{\hat{Y}|A}(Y|a) - \ln p_{\hat{Y}|A}(Y|a') \right|,$$

which can be optimized either increase accuracy of the worse group of decrease accuracy of the better group. The latter is not desirable.

$$\begin{aligned}\mathcal{L}_{WCE} &= \sum_{a,a'} \max \left\{ -\ln p_{\hat{Y}|A}(Y|a), -\ln p_{\hat{Y}|A}(Y|a') \right\}, \\ &= -\sum_{a} N_{\leq a} \ln p_{\hat{Y}|A}(Y|a),\end{aligned}$$

Dataset

Dialect	Abbr	Corpus	# Utts	Len
African American	AA	CORAAL	13908	491
Standard American	SA	Librispeech	28533	6035
Latin American	LA	LDC2014S05	281	28
UK Broadcast News	UK	LDC95S24	10980	1221
Afrikaans Eng	AF	AST Afrikaans	3799	133
Black Eng	XH	AST Black	3323	116
Indian Eng	IN	MaheshChandra	358	16

- Dialect dataset consist of 7 dialects by combine 7 different speech corpus
- "Abbr" column is the abbreviated dialect name used in performance tables.
- "#Utt" column shows the number of utterances in the training set.
- "Len" column shows the total duration of all utterances, in minutes.

CORAAL Dataset

Attr	Group	Abbr	#Utt	Len
Age	-19		7320	250
	20-29		2776	104
	30-50		2590	99
	51+		1122	37
Work	Lower Working Class	LW	3516	125
	Upper Working Class	UW	4359	146
	Lower Middle Class	LM	3647	131
	Upper Middle Class	UM	1159	46
	Upper Class	U	824	28
	Unknown	Unk	403	13
Edu	Elementary School	ES	169	6
	Student in Middle School	StMS	3190	107
	Student in High School	StHS	3510	118
	Some High School.	SHS	1206	41
	High School	HS	3156	108
	Student in College	StCO	192	7
	Some College	SCO	1485	63
	College	CO	847	32
	Graduate School	GS	153	5
Gender.	Male	M	9155	317
	Female	F	4753	174

Dialect Dataset results

Dialect	λ=0	λ=0.001	λ=0.01	λ=0.1	λ=1	λ=10
AA	43.08	**39.07**	42.99	44.28	45.72	46.36
AF	20.88	**18.18**	23.70	22.26	24.81	20.98
AM	14.19	**10.94**	13.73	14.50	18.21	16.12
BR	14.56	**12.21**	17.36	17.09	19.23	16.98
IN	52.80	51.38	**50.95**	51.36	53.67	52.80
LA	38.41	**30.00**	41.70	36.28	32.14	36.46
XH	26.60	**22.11**	29.29	27.58	28.26	26.43
Mean	30.07	**26.27**	31.39	30.48	31.72	30.87
Std	14.97	14.85	14.11	13.97	**13.39**	14.61

Table: Multi-dialect experiments. Refer to Table 9 for the meanings of the abbreviations.

CORAAL Dataset results

Age	λ=0	λ=0.001	λ=0.01	λ=0.1	λ=1	λ=10
-19	55.59	56.60	**53.96**	56.23	55.94	56.72
20-30	55.56	55.99	**53.73**	55.82	56.60	57.13
30-50	56.31	56.99	**54.94**	56.24	56.61	57.04
50+	59.31	59.97	58.59	**58.53**	59.33	59.79
Mean	56.69	57.39	**55.30**	56.70	57.12	57.67
Std	1.78	1.77	2.25	**1.23**	1.50	1.42
Work						
LM	56.16	**54.97**	58.03	55.64	57.05	56.90
LW	55.30	**54.30**	57.44	55.06	56.76	55.60
UW	56.03	**54.68**	58.32	55.55	56.96	56.81
UM	58.01	**55.62**	58.27	55.69	58.15	57.78
U	58.76	**57.25**	59.06	57.33	59.31	57.99
Unk	56.86	**54.71**	57.41	57.46	56.71	56.36
Mean	56.85	**55.26**	58.09	56.12	57.49	56.91
Std	1.31	1.07	**0.62**	1.01	1.04	0.89
Edu						
ES	61.94	61.00	61.54	62.35	**59.24**	60.19
StMS	55.54	**54.86**	55.95	57.03	57.28	56.93
StHS	55.40	**54.55**	56.48	57.31	56.71	55.83
SHS	**55.20**	55.25	56.70	57.73	56.87	55.57
HS	57.27	**56.04**	58.63	59.13	58.06	56.69
StCO	**51.95**	53.25	55.03	59.17	54.79	57.28
SCO	56.12	**55.54**	57.27	57.99	57.48	56.65
CO	54.18	**53.79**	55.70	55.62	55.28	55.04
GS	**54.42**	54.97	54.83	57.04	56.22	55.39
Mean	55.78	**55.47**	56.90	58.15	56.88	56.62
Std	2.74	2.24	2.09	1.92	**1.36**	1.54
Gender						
M	55.74	55.28	55.55	57.32	58.07	**55.21**
F	55.93	56.41	55.56	57.57	57.46	**55.44**
Mean	55.84	55.85	55.55	57.45	57.76	**55.32**
Std	0.13	0.80	**0.01**	0.17	0.43	0.16

Conclusion

- Conclusion
  - There is a trade of between accuracy and variance (fairness)
  - Training with Equal Accuracy Ratio helps reduce variance in accuracy.
  - Training with Equal Accuracy Ratio does not always reduce accuracy.
- Future Work
  - The dialect dataset can be improved by adding more data.
  - Different λ can be tried to see the effect of regularization
  - Different weight can be tried to see if giving more weights on worst performance group brings a more fair model.

The End

Reference I

Hardt, M., Price, E., and Srebro, N. (2016).
Equality of opportunity in supervised learning.
In *Advances in neural information processing systems*, pages 3315–3323.

Kamiran, F. and Calders, T. (2009).
Classifying without discriminating.
In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. (2020).
Racial disparities in automated speech recognition.
Proceedings of the National Academy of Sciences, 117(14):7684–7689.

Reference II

Li, B., Sainath, T. N., Sim, K. C., Bacchiani, M., Weinstein, E., Nguyen, P., Chen, Z., Wu, Y., and Rao, K. (2018).
Multi-dialect speech recognition with a single sequence-to-sequence model.
In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4749–4753. IEEE.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017).
Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment.
In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.