The University Of Sheffield.

Unsupervised Data Selection For Speech Recognition With Contrastive Loss Ratios

Chanho Park, Rehan Ahmad, Thomas Hain
Speech and Hearing Research Group (SPandH) , University of Sheffield, Sheffield, UK

Paper #6273

Introduction

Semi-supervised learning has become popular

- an increased amount of training data
- negative transfer in multi-domain data sets

This paper aims to

1. select training data for speech recognition matching target data from a data pool
2. maintain or improve the performance of ASR systems while minimising negative transfer

Background

Contrastive representation learning

For representation learning,

- maximises the mutual information of encoded and contextualised embeddings
- comparing density ratios of positive and negative samples for future k steps

$$\mathcal{L}_N = -\mathbb{E}_X\left[log\frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

Submodular function

A function $f : 2^V \rightarrow \mathbb{R}$ is submodular if
$f_A(e) \geq f_B(e)$ for all $A \subseteq B \subseteq V$ and $e \in V \backslash B$
where $f_A(e) = f(A \cap \{e\}) - f(A)$

If the function is monotonically nonincreasing, and given a constraint k,

$$\underset{|S| \leq k}{\arg\max}\{f(S)\}$$

Methods

Loss ratios

f_Ω: loss function trained on the data pool
f_{tgt}: loss function trained on a target data set
α: a number to prevent overflow or underflow
x_t: an observation at time t

$$LR(u) = \frac{1}{T}\sum_{t=1}^{T}\frac{f_\Omega(x_t) + \alpha}{f_{tgt}(x_t) + \alpha}$$

Submodular function

S: a subset of the data pool

$$f_{LR}(S) = \sum_{u \in S}(LR(u))$$

Data sets

Data pool (Ω): 40 hours

AMI	10 hours
Fisher (FS)	10 hours
Tedtalks (TD)	10 hours
Wsjcam0 (WS0)	10 hours

Target data: 1-hour sets for contrastive loss
Test data: 1-hour sets for ASR performance

Experimental setup

Pre-training

wav2vec models for contrastive loss
GMM-HMM models for log-likelihood

Data selection

Calculate $LR(u)$ where $u \in \Omega$
Sort the utterances by $LR(u)$
Select data from Ω on a constraint, e.g. 10h

Hybrid ASR system

GMM-HMM and neural networks

Result - data selection

Data from the same corpus as the target data tend to be selected earlier with CLR than with LL.

Numbers of selected segments. The total numbers for AMI, FS, TD and WS0 were 3526, 3330, 3244 and 3685, respectively.

Contrastive loss ratios (CLR)

target data set	hours of subset			selected data set
	10h	20h	30h	
AMI	3263	3503	3521	AMI
	14	291	1083	FS
	195	1811	2725	TD
	16	1320	3070	WS0
FS	0	669	2209	AMI
	3257	3328	3329	FS
	65	2615	3123	TD
	0	15	1479	WS0
TD	103	1524	2797	AMI
	362	1789	2686	FS
	2773	3181	3219	TD
	0	152	1471	WS0
WS0	104	2166	3299	AMI
	0	4	334	FS
	28	1222	3116	TD
	3527	3684	3685	WS0

Log-likelihood (LL)

target data set	hours of subset			selected data set
	10h	20h	30h	
AMI	2023	2810	3222	AMI
	131	774	1863	FS
	306	1089	2020	TD
	1008	2261	3262	WS0
FS	13	1616	2717	AMI
	3301	3325	3325	FS
	18	1399	2455	TD
	0	349	1646	WS0
TD	1385	2250	2899	AMI
	162	781	1807	FS
	1100	2099	2779	TD
	720	1662	2781	WS0
WS0	845	2492	3208	AMI
	4	337	1699	FS
	57	625	1861	TD
	2680	3653	3685	WS0

References

S. Schneider, A. Baevski, R. Collobert and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech 2019*, Graz, Austria, pp. 3465–3469, [Online]. doi: 10.21437/Interspeech.2019-1873.

A. Krause, and D. Golovin, "Submodular function maximization," in *Tractability: Practical approaches to hard problems*, L. Bordeaux, Y. Hamadi and P. E. Kohli, Eds., p. 71–104. Cambridge University Press, 2014, [Online]. doi: 10.1017/CBO9781139177801.004.

Result - ASR performance

WERs of ASR models trained on data sets selected by CLR were lower than by LL.

WERs(%) on selected data sets.

Method	target	10h	20h	30h	40h
CLR	AMI	31.71	28.62	27.02	**26.69**
	FS	39.57	37.12	**35.49**	35.72
	TD	28.07	25.54	**24.43**	24.58
	WS0	11.14	9.57	**9.32**	9.90
LL	AMI	34.51	29.56	26.95	**26.69**
	FS	40.02	36.80	36.56	**35.72**
	TD	35.19	28.37	26.42	**24.58**
	WS0	11.27	9.90	**9.89**	9.90

Result - negative transfer

ASR models achieved equal to or better performance with less data.

WERs(%) on selected data sets.

Method	target	80%	85%	90%	95%
CLR	AMI	26.98	26.79	**25.91**	26.35
	FS	35.83	36.96	35.83	**35.72**
	TD	24.97	25.25	24.94	**24.34**
	WS0	9.66	9.71	**9.51**	9.66
CL	AMI	27.19	26.55	**25.78**	27.36
	FS	**35.02**	36.11	35.75	35.50
	TD	25.09	24.61	**24.34**	24.59
	WS0	9.56	**9.28**	9.66	9.52

Conclusion

- Using the proposed method, a training set for automatic speech recognition matching the target data set could be selected
- ASR performance can be maintained or improved on the reduced amount of data selected by the method

Acknowledge

This work was conducted at the VoiceBase Research Centre for Speech and Language Technologies at the University of Sheffield, which is funded by VoiceBase Inc.

VoiceBase