Unsupervised data selection for Speech Recognition with contrastive loss ratios

Chanho Park, Rehan Ahmad and Thomas Hain

Speech and Hearing Research Group (SPandH)
The University of Sheffield

2022 IEEE International Conference
on Acoustics, Speech and Signal Processing

Table of Contents

Motivation

Semi-supervised learning

- an iterative process of labelling and training
- selects confident data for better performance

Current methods

- confidence score: scoring the whole training data set
- proxy function: smaller and faster, but less accurate

Issues

- increased amount of unlabelled training data
- negative transfer when training and test data are in different domains

Aims

- to avoid iterative computations
- to select reduced amount of data while minimising negative transfer

Contrastive representation learning

A contrastive loss function

- maximises the similarity between data representations in a category
- minimises it between data representations in different categories

For representation learning,

- maximises the mutual information of encoded and contextualised embeddings
- predicts the encoded embedding of future k-step based on the context embeddings
- comparing density ratios of positive and negative samples

In this paper, wav2vec[1] model was adopted as a representation learning model

[1] 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.

Submodular function

Selecting data from a data pool is to find discrete sets of feasible solutions

$$f : 2^V \rightarrow \mathbb{R}$$

A function is submodular if

$$f_A(e) \geq f_B(e) \text{ for all } A \subseteq B \subseteq V \text{ and } e \in V \backslash B$$
$$\text{where } f_A(e) = f(A \cap \{e\}) - f(A)$$

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

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

VoiceBase

Proposed method

Contrastive 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

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

Experimental setup

corpus	hours		
	target	data pool	test
AMI	1	10	1
Fisheer (FS)	1	10	1
Tedtalks (TD)	1	10	1
Wsjcam0 (WS0)	1	10	1

Data pool: 40 hours of training data sets for ASR models
Target data: 1-hour sets of training data for contrastive loss
Test data: 1-hour sets of evaluation data for ASR performance

Experimental setup

Contrastive representation learning

 $\longrightarrow$ f_{Ω}

data pool

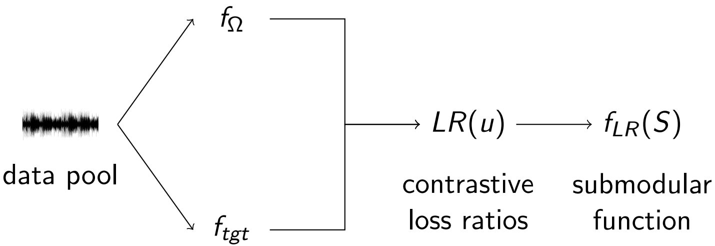
 $\longrightarrow$ f_{tgt}

target data set

wav2vec models

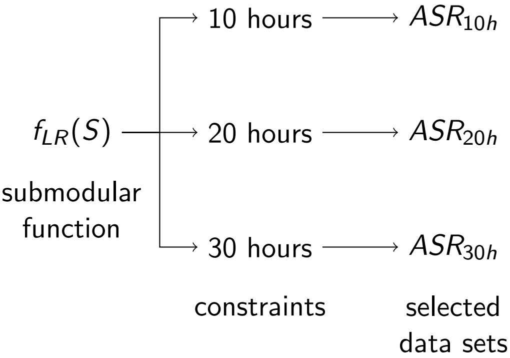
Experimental setup

Contrastive loss ratios



Experimental setup

Data selection



Results

The numbers of segments selected by the proposed method:

Contrastive loss ratios

target data set	hours of subset 10h	20h	30h	selected data set
AMI	3263	3503	3521	AMI
	14	291	1083	FS
	195	1811	2725	TD
	16	1320	3070	WS0
WS0	104	2166	3299	AMI
	0	4	334	FS
	28	1222	3116	TD
	3527	3684	3685	WS0

Log-likelihood

target data set	hours of subset 10h	20h	30h	selected data set
AMI	2023	2810	3222	AMI
	131	774	1863	FS
	306	1089	2020	TD
	1008	2261	3262	WS0
WS0	845	2492	3208	AMI
	4	337	1699	FS
	57	625	1861	TD
	2680	3653	3685	WS0

Results

Given a 10 hours of constraint:

Data selection

target/ selected	segments		total
	CLR	LL	
AMI	3263	2023	3526
FS	3257	3301	3330
TD	2773	1110	3244
WS0	3527	2680	3685

Results

Given a 10 hours of constraint:

Data selection

target/ selected	segments CLR	segments LL	total
AMI	3263	2023	3526
FS	3257	3301	3330
TD	2773	1110	3244
WS0	3527	2680	3685

ASR performance

target/ selected	WER(%) CLR	WER(%) LL
AMI	31.71	34.51
FS	39.54	40.02
TD	28.07	35.19
WS0	11.14	11.27

Results

ASR performance on selected data sets

target	10h	20h	30h	40h
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

Results

Negative transfer

Method	selected	80%	85%	90%	95%	100%
CLR	AMI	26.98	26.79	**25.91**	26.35	26.69
	FS	35.83	36.96	35.83	**35.72**	35.72
	TD	24.97	25.25	24.94	**24.34**	24.58
	WS0	9.66	9.71	**9.51**	9.66	9.90
CL	AMI	27.19	26.55	**25.78**	27.36	26.69
	FS	**35.02**	36.11	35.75	35.50	35.72
	TD	25.09	24.61	**24.34**	24.59	24.58
	WS0	9.56	**9.28**	9.66	9.52	9.52

Conclusion

- By using the proposed method, a training set for automatic speech recognition matching the target data set could be selected.
- ASR models trained on the data sets selected by the proposed method outperformed the model trained on the data pool
- ASR performance could be maintained or improved on the reduced amount of data selected by the method

QnA