UNSUPERVISED DATA SELECTION FOR SPEECH RECOGNITION WITH CONTRASTIVE LOSS RATIOS

Chanho Park, Rehan Ahmad, Thomas Hain

Department of Computer Science, University of Sheffield,
cpark12, r.ahmad, t.hain@sheffield.ac.uk

ABSTRACT

This paper proposes an unsupervised data selection method by using a submodular function based on contrastive loss ratios of target and training data sets. A model using a contrastive loss function is trained on both sets. Then the ratio of frame-level losses for each model is used by a submodular function. By using the submodular function, a training set for automatic speech recognition matching the target data set is selected. Experiments show that models trained on the data sets selected by the proposed method outperform the selection method based on log-likelihoods produced by GMM-HMM models, in terms of word error rate (WER). When selecting a fixed amount, e.g. 10 hours of data, the difference between the results of two methods on Tedtalks was 20.23% WER relative. The method can also be used to select data with the aim of minimising negative transfer, while maintaining or improving on performance of models trained on the whole training set. Results show that the WER on the WSJCAM0 data set was reduced by 6.26% relative when selecting 85% from the whole data set.

Index Terms— data selection, unsupervised, contrastive loss, submodular, speech recognition

1. INTRODUCTION

The performance of an automatic speech recognition (ASR) system is affected by the amount of training data used. It would be ideal to use more data with transcriptions. Unfortunately, since labelling is time-consuming, unlabelled data are used as supplementary to the labelled data set [1]. The above method using both labelled and unlabelled data sets is called semi-supervised learning. Furthermore, an unsupervised speech recognition method that does not require labelled data [2] has been investigated to gain more information from unlabelled data. As a result, the training time has increased as the amount of unlabelled data required for these methods has increased.

The semi-supervised learning method has shown improvement in WER in both cases of using additional data in a single domain [3] and multi-domain data consisting of various data sets [4]. In [4], the model trained on a multi-domain training data set outperformed the model trained on a single-domain data set. However, in the case of multi-domain data, the degradation of WER performance between different data sets, called negative transfer, was observed in [5]. For example, the WER of the model trained on both Fisher and WSJCAM0 data sets was worse than that of the model trained on one of the data sets. Moreover, a similar observation exists in [4]. More subsets in pre-training showed mixed results in terms of WER performance in the scenario when target data were included in the training set.

So far, two observations have been discussed: increased training time and negative transfer. Given a budget of the amount of training data, they need to be addressed by selecting an optimal data subset without performance degradation. Confidence scores can be used to select only high confidence data. However, it is time-consuming because they are built upon the top of ASR systems [6]. Second, to address the computational cost of confidence scores, a proxy function which is smaller but faster has been proposed [7]. However, the performance of a proxy model is less accurate than that of the original as it is proposed to reduce the computational cost instead of the accuracy. There has been another approach to data selection using a submodular function in a budget [8, 9]. A submodular function converts a set into a measurable value so that the optimal subset in the budget can be found by the value.

In this paper, an unsupervised submodular data selection method by using contrastive loss ratios is proposed. The proposal is based on a contrastive loss function learned by predicting future latent representations conditioned on past latent representations [10, 11]. After pre-training, the loss of the model trained on a target data set indicates how likely the frame is in the context of the data set. The loss is compared to that of the model trained on a training data set. Whether the frame is close to the target data or to the training data is decided by the ratios between them. By the utterance-level mean value of the frame-level ratios, all the utterances can be ordered. A function for a contrastive loss ratio is monotonic and non-decreasing by adding utterances. With the characteristics of the function, it can be used as a submodular function for selecting an optimal subset.

The remaining part of the paper proceeds as follows: in Section 2, background information is introduced. In Section 3, the proposed approach is described in detail. After the experimental setup in Section 4, the results are shown in Section 5. Then, the conclusion is in Section 6.

2. BACKGROUND

2.1. Contrastive representation learning

A contrastive loss function maximises the similarity between data representations in a category while minimising it in different categories.

A representation learning method by using the contrastive loss function has been proposed in [10, 11]. First, an audio frame x_t at time t is embedded by convolutional neural networks. Next, an embedding z_t is contextualised by using an autoregressive model to exploit high-level latent information between different parts. Then, the context representation c_t is used to predict future embeddings z_{t+k} where k is the number of future steps.

Instead of predicting x_{t+k} directly, the mutual information of x_{t+k} and c_t is maximised. Based on the model, InfoNCE is suggested as an objective function in [10]. Both the encoder and autoregressive model are trained to jointly optimise the InfoNCE loss:

$$\mathcal{L}_N = -\mathbb{E}_X \left[log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right] \quad (1)$$

where f is a density ratio and can be unnormalised, and X is a set of random samples including a positive sample. To score the mutual information, a simple log-bilinear model was used in [10]. Similarly, the wav2vec model in [11] used $\sigma(z_{t+k}^\top h_k(x_t))$ as f where σ is a sigmoid function, and h_k is an affine transformation.

2.2. Submodular function maximisation

Selecting data from a data pool is to find discrete sets of feasible solutions. This problem can be solved by adopting a submodular function [12] for the set.

$$f : 2^V \to \mathbb{R} \quad (2)$$

where V is a finite set, 2^V is the set of all subsets of V and $f(\emptyset) = 0$. f is submodular if $f_A(e) \geq f_B(e)$ for all $A \subseteq B \subseteq V$ and $e \in V \setminus B$ where $f_A(e) = f(A \cap \{e\}) - f(A)$. If the function is concave, the optimal solution S is a set which maximises the value of the submodular function.

$$\max\{f(S) : S \subseteq V\} \quad (3)$$

To solve this problem, a greedy solution can be used. However, it is NP-hard and requires the non-deterministic polynomial time. One possible idea to avoid it is for the function to be monotonic.

$$f(A) \leq f(B), \text{ if } A \subseteq B \subseteq V \quad (4)$$

With monotonic functions, it is clear that f is maximised at V. Now the optimal solution can be found by considering a constraint k for data selection.

$$\underset{|S| \leq k}{\arg \max}\{f(S)\} \quad (5)$$

In other words, each element of a set can be selected by the value of the function in order.

3. PROPOSED METHOD

3.1. Loss ratios

As a submodular function, the mean of the ratios of all the frames in an utterance is used. First, ratios between the values of a contrastive loss function f_Ω trained on the whole training data set Θ_Ω and those of another contrastive loss function f_{tgt} on a target data set Θ_{tgt} can be calculated as below:

$$LR(u) = \frac{1}{T} \sum_{t=1}^{T} \frac{f_\Omega(x_t) + \alpha}{f_{tgt}(x_t) + \alpha} \quad (6)$$

where α is a number to prevent overflow or underflow of the loss ratio and x_t is an observation at time t. Then, the accumulated $LR(u)$ of all utterances in a subset S which is included in Ω is defined as a submodular function like below:

$$f_{LR}(S) = \sum_{u \in S} \big(LR(u) \big) \quad (7)$$

This modular function f_{LR} is non-negative as it is a sum of means of ratios between non-negative losses in Equation 1. Moreover, it is normalised, which means that the value of the function is zero when the input is null.

3.2. Negative transfer minimisation

When a budget is not given, the amount of data can be reduced without degradation of WER performance by minimising the negative transfer. Negative transfer is performance degradation when another domain data set is added to a training set. To maximise the performance of an ASR model, negative transfer should be minimised. As the optimal data subset is selected by the value of the submodular function in Equation 7, a data set of negative transfer can be filtered by a threshold. The threshold is investigated by a grid search method. Furthermore, a value of $f_{tgt}(x_t)$ represents how well the model is fitted into the target data set. If an utterance is affected by negative transfer, $f_{tgt}(x_t)$ would be high. Thus, another threshold of loss is also explored in the same method.

4. EXPERIMENTAL SETUP

Similar to the previous study [5], a data pool (DP) was generated to select from. It consisted of 4 data sets: AMI [13], Fisher [14] (FS), Tedtalks [15] (TD) and Wsjcam0 [16] (WS0). The training data pool was 40 hours in total: 10 hours of each data set. For pre-training, two 1-hour sets of each data set were target and test data sets. There are three parts of the experiment: pre-training; data selection; building ASR systems.

4.1. Pre-training

A wav2vec [11] was modified for contrastive loss ratios. To fit into the small size of the target data set, kernel sizes and strides of encoder layers were changed from (10,8,4,4,4) and (5,4,2,2,2) to (16, 16) and (2,3,4), respectively. The number of prediction steps was set to 6 instead of 12. Each model was trained to 200 epochs, then stopped early with patience of 10 epochs. Each model was trained on individual target and training data sets.

A GMM-HMM system was trained for log-likelihood. As the amount of training data for this purpose was an hour, the accuracy of the system was not as high as the hybrid ASR system described in Section 4.3. Then, the training data were decoded by using the models, and the log-likelihood of the utterances in the decoding was applied for data selection.

4.2. Data selection

After pre-training, frame-level losses on a target data set and the training data set were calculated. The losses were added to α to prevent ratios of them from overflowing. Then, the mean values of the losses of an utterance were used as a score for data selection. The data were selected by a greedy method based on the score. The utterances were sorted by the score, then data in a budget was selected in order. When a budget was not given, contrastive loss ratios and losses were used to reduce the amount of the selected data by minimising negative transfer. The thresholds of contrastive loss ratios and losses were investigated empirically.

4.3. Hybrid ASR system

Hybrid systems of GMM-HMM and neural networks were built for speech recognition. They were based on the system described in [17]. For alignment, monophone, triphone, discriminant analysis (LDA) and maximum likelihood linear transform (MLLT), and Speaker Adaptive Training (SAT) models were trained in sequence. Then, neural networks (Nnet2) were trained by using the labelled frames generated by the GMM-HMM models. Moreover, the language model was trained on merged text of all data sets.

5. RESULTS

5.1. Baseline

An ASR system with the data pool was built as a baseline system. It was trained on all 40 hours of the data pool, and each test data set and the combined data set (DP) were scored separately as seen in Table 1.

Table 1: WERs(%) of baseline systems

Feature	AMI	FS	TD	WS0	DP
MFCC	26.69	35.72	24.58	9.90	25.04

5.2. Data selection

The results of data selection are shown in Table 2. The first column is a target data set which a pre-training model was trained on. When segments were selected by the contrastive loss ratios between a pre-training model on AMI and another model on the data pool, 3263, 3503 and 3521 segments of AMI were selected for sets of 10h, 20h and 30h from the pool, respectively. At the same time, 2023, 2810 and 3222 segments were selected by log-likelihood, respectively. Data from the same corpus with the target data tended to be selected by contrastive loss ratios rather than by log-likelihood.

Table 2: Numbers of selected segments by contrastive loss ratios and log-likelihood. The total numbers for AMI, FS, TD and WS0 were 3526, 3330, 3244 and 3685, respectively. CLR and LL stand for contrastive loss ratio and log-likelihood, respectively.

target data set	hours of subset (CLR/LL)			selected data set
	10h	20h	30h	
AMI	3263/2023	3503/2810	3521/3222	AMI
	14/131	291/774	1083/1863	FS
	195/306	1811/1089	2725/2020	TD
	16/1008	1320/2261	3070/3262	WS0
FS	0/13	669/1616	2209/2717	AMI
	3257/3301	3328/3325	3329/3325	FS
	65/18	2615/1399	3123/2455	TD
	0/0	15/349	1479/1646	WS0
TD	103/1385	1524/2250	2797/2899	AMI
	362/162	1789/781	2686/1807	FS
	2773/1100	3181/2099	3219/2779	TD
	0/720	152/1662	1471/2781	WS0
WS0	104/845	2166/2492	3299/3208	AMI
	0/4	4/337	334/1699	FS
	28/57	1222/625	3116/1861	TD
	3527/2680	3684/3653	3685/3685	WS0

Given 10, 20, 30 hours budgets, most of the selected data for the first 10 hours were from the same data set as the target data. As the budget increases, data from the different data sets were selected more because the amount of each data set was limited to 10 hours. However, one of the data sets was less frequently selected until 30 hours. For instance, when a target data set was WS0, only 334 utterances of FS were selected for the 30-hour data set. This was relatively lower than the other data selection results. The negative effect of the remaining data was observed in the following section by using the WER performance.

5.3. Loss ratios vs log-likelihood

Data selected sets in Table 2 were used for training ASR models. The performance of each system was measured in terms of WER. The results of ASR systems trained on data selected by contrastive loss ratios outperformed those by log-likelihood of the hybrid system. For example, in Table 3, the WER performances of models trained on data sets by contrastive loss ratios outperformed the others except 20 hours and 30 hours when the target sets were FS and AMI, respectively.

Table 3: WERs(%) on selected data sets. For example, the WER of an ASR system trained on the subset of 10h when a target data set was AMI was 31.71%.

Method	target	10h	20h	30h	40h
	AMI	31.71	28.62	27.02	**26.69**
CLR	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
	AMI	34.51	29.56	26.95	**26.69**
LL	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

5.4. Negative transfer minimisation

For a grid search of thresholds, different amounts of data sets were selected by contrastive loss ratios and losses. Based on the result in Table 3, from 80% to 95% of the data pool were selected as training data sets.

As shown in Table 4, the optimal subsets for WER performance were between 80% and 95% of the whole data set. The best performances for AMI, FS, TD and WS0 on data sets selected by CLR were 25.91%, 35.72%, 24.34% and 9.51%, respectively. When data sets were selected by CL, they were 25.78%, 35.02%, 24.34% and 9.28% for AMI, FS, TD and WS0, respectively. These WERs were competitive to the baseline with the CLR method when budgets were given

Table 4: WERs(%) on selected data sets for negative transfer minimisation by contrastive loss ratios and losses. CL stands for contrastive loss.

Method	target data set	80%	85%	90%	95%
	AMI	26.98	26.79	**25.91**	26.35
CLR	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
	AMI	27.19	26.55	**25.78**	27.36
CL	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

in Table 3. For example, when data sets selected by CL, the WER of a model on the 80% data set for FS were 35.02%, while the baseline performance was 35.72%. In other words, the WER performance on the data pool can be achieved on a smaller data set by eliminating data that might cause negative transfer.

6. CONCLUSION

An unsupervised data selection with a submodular function based on contrastive loss ratios has been explored in this paper. For data selection based on budget, this method outperforms a method using log-likelihood produced by the models trained on a target data set. Furthermore, the amount of data selected from the data pool can be minimised without performance degradation by avoiding negative transfer. The performance of the ASR models trained on the training data set selected by contrastive losses and ratios is comparable to the baseline performance on the whole training data set.

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] S. H. Krishnan Parthasarathi, and N. Strom, "Lessons from building acoustic models with a million hours of speech," in *ICASSP 2019 - 2019 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Brighton, UK, pp. 6670–6674, [Online]. doi: 10.1109/ICASSP.2019.8683690.

[2] A. Baevski, W. Hsu, A. Conneau and M. Auli, "Unsupervised speech recognition," presented at 35th Conf. Neural Inf. Process. Syst. (NeurIPS), 2021.

[3] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang et al., "Pushing the limits of semi-supervised learning for automatic speech recognition," presented at Self-Supervised Learn. for Speech and Audio Process. Workshop 34th Conf. Neural Inf. Process. Syst. (NeurIPS), 2020.

[4] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap et al., "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *Proc. Interspeech 2021*, Brno, Czechia, pp. 721–725, [Online]. doi: 10.21437/Interspeech.2021-236.

[5] M. Doulaty, O. Saz and T. Hain, "Data-selective transfer learning for multi-domain speech recognition," in *Proc. Interspeech 2015*, Dresden, Germany, pp. 2897–2901, [Online]. doi: 10.21437/Interspeech.2015-609.

[6] Q. Li, D. Qiu, Y. Zhang, B. Li, Y. He, P. C. Woodland et al., "Confidence estimation for attention-based sequence-to-sequence models for speech recognition," in *ICASSP 2021 - 2021 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Toronto, Ontario, Canada, pp. 6388–6392, [Online]. doi: 10.1109/ICASSP39728.2021.9414920.

[7] C. Coleman, C. Yeh, S. Mussmann, B. Mirzasoleiman, P. Bailis, P. Liang et al., "Selection via proxy: Efficient data selection for deep learning," presented at Int. Conf. Learn. Repr. (ICLR), Addis Ababa, Ethiopia, 2020.

[8] H. Lin, and J. Bilmes, "How to select a good training-data subset for transcription: Submodular active selection for sequences," in *Proc. Interspeech 2009*, Brighton, UK, pp. 2859–2862, [Online]. doi: 10.21437/Interspeech.2009-730.

[9] K. Wei, Y. Liu, K. Kirchhoff and J. Bilmes, "Unsupervised submodular subset selection for speech data," in *2014 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Florence, Italy, pp. 4107–4111, [Online]. doi: 10.1109/ICASSP.2014.6854374.

[10] A. van den Oord, Y. Li and O. Vinyals, "Representation learning with contrastive predictive coding." DeepMind., https://deepmind.com/research/publications/2019/representation-learning-contrastive-predictive-coding (Accessed: Feb 9, 2022).

[11] 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.

[12] 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.

[13] C. Cieri, D. Miller and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *Proc. 4th Int. Conf. Lang. Resour. Eval. (LREC'04)*, Lisbon, Portugal, 2004, vol. 4, pp. 69–71.

[14] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain et al., "The AMI meeting corpus: A pre-announcement," in *Mach. Learn. for Multimodal Interact.*, Berlin, Heidelberg, 2006, vol. 3869 of *Lecture Notes in Computer Science*, pp. 28–39, Springer Berlin Heidelberg, [Online]. doi: 10.1007/11677482_3.

[15] R. W. Ng, M. Doulaty, R. Doddipatla, W. Aziz, K. Shah, O. Saz et al., "The USFD spoken language translation system for IWSLT 2014," in *Proc. 11th Int. Workshop Spoken Lang. Transl.: Eval. Campaign*, Lake Tahoe, CA, USA, 2014, pp. 86–91.

[16] T. Robinson, J. Fransen, D. Pye, J. Foote and S. Renals, "WSJCAMO: A British English speech corpus for large vocabulary continuous speech recognition," in *1995 Int. Conf. Acoust., Speech, Signal Process.*, Detroit, Michigan, vol. 1, pp. 81–84, [Online]. doi: 10.1109/ICASSP.1995.479278.

[17] D. Povey, X. Zhang and S. Khudanpur, "Parallel training of DNNs with natural gradient and parameter averaging," presented at Int. Conf. Learn. Repr. (ICLR), San Diego, CA, USA, 2015.