Hypersphere Embedding and Additive Margin for Query-by-example Keyword Spotting

Haoxin Ma[1,2]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China

mahaoxin2019@ia.ac.cn

Abstract

Query-by-example (QbE) keyword spotting is convenient for users to define their own keywords and useful in device control. However, conventional regular softmax, which is commonly used for training QbE models, has two limitations. First, the learned features are not discriminative enough. Second, norm variations of the unnormalized features affect computing cosine similarities. To address these issues, this paper introduces normalization and additive margin into residual networks for QbE keyword spotting. Our methods achieve improvements in the experiments.

1. Introduction

Query-by-example (QbE) keyword spotting is the task of detecting the predefined keyword in a series of speech recordings. The most typical application is to activate a device by a customized wake-up word. A QbE keyword spotting system detects audio segments directly without the need to build a robust automatic speech recognition system and can easily deal with the out-of-vocabulary (OOV) and low-resource situations.

Previous studies on QbE methods can be divided into two categories. One comprises the dynamic time warping (DTW) based methods [1, 2, 3] which calculate similarity of the frame-level feature sequences. However, DTW costs much time and computation. The other category comprises the embedding learning methods which project the acoustic features into a fixed-dimensional space and evaluate the similarity of embedding vectors, which is simple but efficient. Several works [4, 5, 6, 7] demonstrate the success of embedding learning and it outperforms the DTW in QbE keyword spotting methods. However, norm variations of the unnormalized features leave a gap between training and testing. Only after normalization, the vectors are on a unit hypersphere, and the inner product operation of softmax can become the cosine calculation, compatible with the criterion (cosine similarity) in testing. Besides, regular softmax loss used for training can only separate different classes apart without making features of the same class compact, which leads to a limitation in discriminative feature learning [8, 9].

Normalization is a helpful method to eliminate the gap between training and testing. Feature normalization can boost the performance in facial recognition [10, 11]. Additionally, additive margin softmax (AM-softmax) [9] is proposed to improve the softmax loss and has recently proven to work well in facial verification. It introduces an additive margin via subtracting a hyper-parameter m in the cosine space [12], which can minimize the intra-class variation.

2. Qbe framework

A user can predefine a specific keyword, and its bottleneck feature (BNF) is extracted by the neural network from the user's speech recordings. When testing, the same neural network extracts the BNF of the input audio. Then, the system calculates the cosine similarity to make a decision whether the audio is the predefined keyword.

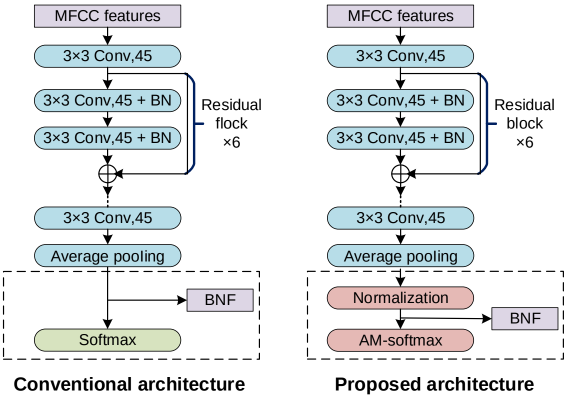
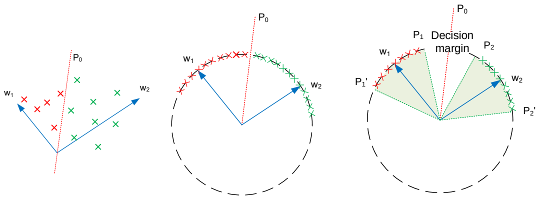


Figure 1: *The left conventional architecture is trained with regular softmax loss, and the right one introduces normalization and AM-softmax loss. The blue parts are the networks: residual block × 6 means that six residual blocks are stacked. The parts with dashed lines show the training methods.*



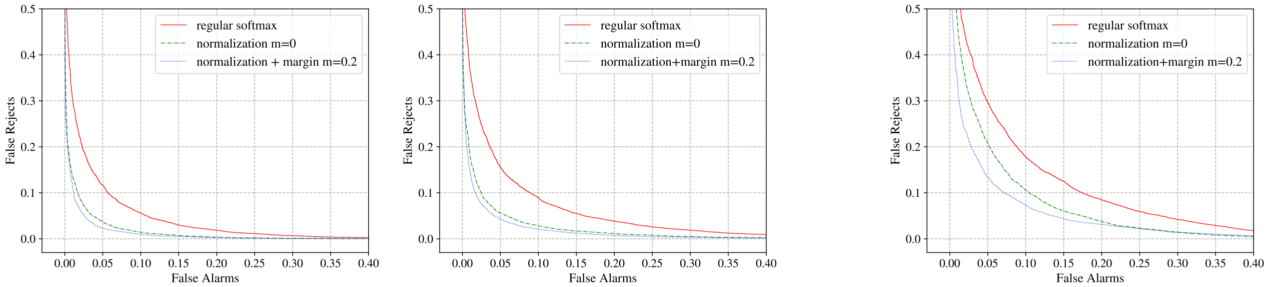
(a) a (b) b (c) c

Figure 2: *A comparison of regular softmax (a), normalization operation (b), and AM-softmax(c): the "×" marks represent the embedding vectors of samples. Different colors correspond to different labels.*

3. Hypersphere embedding and additive margin methods

As demonstrated in Figure 1, compared with the conventional architecture, our proposed architecture replaces the regular softmax with AM-softmax and adds a normalization operation before the softmax layer in ResNet. After normalization, both the feature vectors and the weight vectors are normalized on a unit hypersphere, forming a hypersphere embedding.

The function of AM-softmax is as follows:



(a) DET curves on the in-vocabulary set

(b) DET curves on the OOV set

(c) DET curves on the cross-corpus set

Figure 3: *DET curves of the experiment results.*

$$\begin{aligned} L_{AM-s} &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s(\cos\theta_{y_i}-m)}}{e^{s(\cos\theta_{y_i}-m)}+\sum_{j=1,j\neq i}^{k}e^{s\cos\theta_i}} \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{w_{y_i}^T y_i}}{e^{s(w_{y_i}^T y_i-m)}+\sum_{j=1,j\neq i}^{k}e^{s w_j^T y_i}}. \end{aligned} \tag{1}$$

where y is the BNF. Vector w is the weight which stands for the prototype of the class. Subscripts i and j denote the i-th or the j-th sample. k and n denote the sizes of classes and samples, respectively. Since the vectors y and w are normalized, $w^T y = \cos\theta$. Cosine values can be calculated directly. m is the additive margin. s is the scaling factor. Because an additive margin m is subtracted from $\cos\theta$, the value of AM-softmax is less than the corresponding regular softmax one. If the value of $\cos\theta - m$ wants to be the same as the regular softmax, a larger cosine value is needed. Thereby the distance between the sample of the same label will be more compact and the intra-class differences will be reduced.

Figure 2 shows the improvement of each step. After normalization, radial variations of the vectors are removed. Thus, the calculation can just focus on the angular similarity. After AM-softmax, the decision boundary becomes a decision margin rather than one simple vector boundary P_0 [9].

4. Experiments

Since there are few publicly available datasets used for QbE keyword spotting, we develop our new dataset based on the existing AISHELL-1 dataset [13] and HelloNPU corpus [14]. According to the data preparation idea in the work of A. Jansen *et al.* [15], we select speech segments from the forced alignments of transcripts. The duration of the segments are at least 0.5 seconds and not exceeding 1 second. The labels contain at least 2 characters as texts. We select the segments with the frequency in the top 5,000. We divide the segments into disjointed sets named as training set and development set, including 200,095 utterances and 25,604 utterances, respectively. For evaluation, we design 3 types of sets including the in-vocabulary set, the OOV test set, and the cross-corpus test set. Cross-corpus set is an entirely different dataset from HelloNPU corpus. Therefore, we can use it for a further evaluation of our methods. Each test size is made up of 20,000 speech segment pairs. Half of the label are positive (which means the keyword is ”Hello Xiaogua”) and half of the label are negative (which means the keyword is not ”Hello Xiaogua”).

As for the experimental setup, we extract the 40-dimensional Mel-frequency cepstrum coefficients (MFCCs) of the input audio and pad them to 99 frames. Then we take the 45-dimensional BNF generated from the ResNet. We use the detection error tradeoff (DET) curve as criterion.

We firstly investigate different values of margin m, $m = [0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35]$. Then we select the best performing margin value $m = 0.2$ for further evaluation. $m = 0$ means that we just employ the normalization for both feature and weight vectors without cosine margin. $m \neq 0$ means that we employ both normalization and additive margin.

We regard regular softmax as the baseline, Figure 3(a), Figure 3(b), and Figure 3(c) demonstrate the improvement of “normalization” and “normalization + margin” on all of the test sets. The performance achieves increasingly better from regular softmax to “normalization”, and from “normalization” to “normalization + margin”. At FAR of 2%, the FRR of “normalization + margin” relatively reduces by 79.86% on the in-vocabulary set, 68.03% on the OOV set, and 46.60% on the cross-corpus set compared to the regular softmax. And at FAR of 2%, the FFR of “normalization + margin” relatively reduces by 79.86% on the in-vocabulary set, 68.03% on the OOV set, and 46.60% on the cross-corpus set compared to the regular softmax.

5. Discussions and Future work

From the above experiments, our proposed methods show improvement over the conventional regular softmax for keyword spotting. These are attributed to two reasons. First, normalization helps neural networks focus on angular optimization that is more compatible to the test metric. Compared to the regular softmax, which implicitly learns features from both Euclidean norm and angle, normalization eliminates the variations in Euclidean norm and constrains the features on hypersphere. Second, the margin m helps to reduce the intra-class distance and, moreover, leads to more discriminative feature learning.

This paper introduces the normalization operation and additive margin into QbE keyword spotting tasks to learn discriminative embedding features. These are simple and have an intuitive geometric interpretation. In the future, we want to train a network extracting embedding features from raw waveform directly. Although MFCC and FBANK are employed in many keyword spotting systems, they are hand-crafted features based on prior knowledge and needed extra processing before being fed into networks. Besides, SincNet [16] is proposed to process raw audio with interpretability and performs well in speaker and speech recognition. Thus, we want to explore its possibility in QbE keyword spotting and make some improvement.

6. References

[1] X. Anguera and M. Ferrarons, “Memory efficient subsequence dtw for query-by-example spoken term detection,” in *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2013, pp. 1–6.

[2] G. Mantena, S. Achanta, and K. Prahallad, “Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 5, pp. 946–955, 2014.

[3] Y. Zhang and J. R. Glass, “Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams,” in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 398–403.

[4] K. Levin, A. Jansen, and B. Van Durme, “Segmental acoustic indexing for zero resource keyword search,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5828–5832.

[5] Y.-A. Chung, C.-C. Wu, C.-H. Shen, H.-Y. Lee, and L. Lee, “Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks,” in *Proc. Interspeech*, 2016, pp. 765–769.

[6] S. Settle and K. Livescu, “Discriminative acoustic word embeddings: Tecurrent neural network-based approaches,” in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 503–510.

[7] H. Lim, Y. Kim, Y. Kim, and H. Kim, “Cnn-based bottleneck feature for noise robust query-by-example spoken term detection,” in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1278–1281.

[8] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[9] F. Wang, J. Cheng, W. Liu, and H. Liu, “Additive margin softmax for face verification,” *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[10] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, “Normface: l 2 hypersphere embedding for face verification,” in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 1041–1049.

[11] Y. Zheng, D. K. Pal, and M. Savvides, “Ring loss: Convex feature normalization for face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5089–5097.

[12] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “Cosface: Large margin cosine loss for deep face recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[13] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[14] S. Wang, J. Hou, L. Xie, and Y. Hao, “Hellonpu: A corpus for small-footprint wake-up word detection research,” in *National Conference on Man-Machine Speech Communication*. Springer, 2017, pp. 70–79.

[15] A. Jansen, S. Thomas, and H. Hermansky, “Weak top-down constraints for unsupervised acoustic model training,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8091–8095.

[16] M. Ravanelli and Y. Bengio, “Interpretable convolutional filters with sincnet,” *arXiv preprint arXiv:1811.09725*, 2018.