A Refined DARTS-based Method for Speaker Recognition

Minghao Duan Hiroshi Watanabe

Graduate School of Fundamental Science and Engineering, Waseda University

1 Introduction

Speaker recognition identifies or verifies individuals based on their unique acoustic features and is widely used in voice assistants and network security. Traditional models often use CNN backbones such as VGG-Net or ResNet. However, based on the studies from AutoSpeech [1], VGG-Net and ResNet are not optimally suited for speech recognition tasks. Manually searching for more effective architectures is time-consuming. Differentiable Architecture Search (DARTS) [2] automates this process and has achieved remarkable success in image classification by discovering optimal architectures. Autospeech is the first to apply a pure DARTS method to speaker recognition, demonstrating its effectiveness.

In this paper, we introduce self-attention pooling [3] and LSTM to a DARTS-optimized CNN to capture speaker features and temporal dependencies. Compared to the baseline AutoSpeech, our method improves accuracy in both speaker identification and verification tasks.

2 Related Work

Neural Architecture Search (NAS) aims to automate the design process of neural networks. Differential Architecture Search (DARTS) builds on traditional NAS methods by transforming the discrete architectural search space into a continuous search space, thus enabling gradient-based optimization. This approach allows joint optimization of architecture parameters and network weights, reducing computational cost and improving the search efficiency.

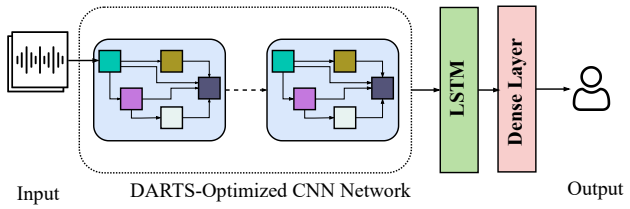

Figure 1 Overview of proposed method

3 Proposed Method

In this work, we apply a DARTS-optimized CNN to speaker recognition, combining self-attention pooling and LSTM. As Figure 1 shows, the process consists of three steps.

First, we use DARTS to search for the optimal neural network architecture.

Second, after determining the optimal architecture, we replace the max pooling layer with self-attention pooling to improve search efficiency by avoiding high computational costs during the architecture search phase. Self-attention pooling captures long-term variations more effectively, as supported by previous studies.

Finally, during the training phase, we apply an LSTM after the CNN backbone, which is widely used to better capture long-term temporal dependencies in speech.

4 Experiment

We train and evaluate our proposed model on the VoxCeleb1 [4] dataset, which contains more than 100,000 utterances from 1251 celebrities, extracted from YouTube.

We select AutoSpeech, a model that applies a pure DARTS-optimized CNN for speaker recognition tasks, for comparative evaluation. We evaluate two models under the same settings, with the number of cells set to 8 and the initial channel size set to 128. As shown in Table 1, our proposed method exhibits better performance in all three metrics of Top-1 accuracy (%), Top-5 accuracy (%) for speaker identification, and Equal Error Rate (EER, %) for speaker verification.

Table 1 Speaker identification and verification performance on the VoxCeleb1 dataset.

	Top-1(%)	Top-5(%)	EER(%)	Parameters
AutoSpeech	87.57	95.98	8.96	18 million
Ours	**88.13**	**96.71**	**8.91**	25 million

5 Conclusion

In this work, we applied Differentiable Architecture Search (DARTS) to speaker recognition tasks and optimized its performance by integrating self-attention pooling and LSTM. Our method demonstrated significant effectiveness in speaker recognition tasks, leading to improved performance in both speaker identification and verification compared to AutoSpeech.

References

[1] S. Ding *et al.*, "AutoSpeech: Neural Architecture Search for Speaker Recognition," arXiv:2005.03215, May 2020.

[2] H. Liu *et al.*, "DARTS: Differentiable Architecture Search," arXiv:1806.09055, June 2018.

[3] P. Safari *et al.*, "Self-Attention Encoding and Pooling for Speaker Recognition," arXiv:2008.01077, 2020.

[4] A. Nagrani *et al.*, "VoxCeleb: A Large-Scale Speaker Identification Dataset," INTERSPEECH, Aug. 2017.