Progressive Label Distillation: Learning Input-Efficient Deep Neural Networks

Zhong Qiu Lin
University of Waterloo
Alexander Wong
University of Waterloo
Email: { zhong.q.lin, a28wong}@uwaterloo.ca

Table 1: Accuracy of the student networks produced using direct distillation.

Cropped Length (ms)	500	600	700	800	900
Test Acc (%)	85.82	90.12	92.58	93.81	94.91

Abstract

Much of the focus in the area of knowledge distillation has been on distilling knowledge from a larger teacher network to a smaller student network. However, there has been little research on how the concept of distillation can be leveraged to distill the knowledge encapsulated in the training data itself into a reduced form. In this study, we explore the concept of progressive label distillation, where we leverage a series of teacher-student network pairs to progressively generate distilled training data for learning deep neural networks with greatly reduced input dimensions. To investigate the efficacy of the proposed progressive label distillation approach, we experimented with learning a deep limited vocabulary speech recognition network based on generated 500ms input utterances distilled progressively from 1000ms source training data, and demonstrated a significant increase in test accuracy of almost 78% compared to direct learning.

1 Introduction

Due to the limited computational resources available in such on-device edge scenarios, many recent studies [1, 2, 4] have put greater efforts into designing small, low-footprint deep neural network architectures that are more appropriate for embedded devices.

In this study, we explore a concept we will call *progressive label distillation*, where a series of teacher-student network pairs are leveraged to progressively generate distilled training data. The proposed approach enables the learning of computationally efficient DNNs with greatly reduced input dimensions without the need for collecting and labeling new data. The proposed strategy can be used in conjunction with any efficient deep neural network architecture to further reduce computational costs and memory footprint.

2 Method

An overview of the progressive label distillation strategy can be described as follows. First, we train a teacher network using an original training data with dimensions greater than the target input dimension. Second, a new training data with reduced dimensions is generated from the original training data (e.g., in the case of limited vocabulary speech recognition, one can generate short audio samples by randomly cropping segments from longer audio samples), with the associated labels generated using the prediction results of the teacher network for these dimension-reduced samples. Since the labels are generated based on the knowledge of the teacher network, we will refer to these generated labels as *distilled labels*. Third, a student network with reduced input dimensions is trained with the new input data and the distilled labels generated using the teacher network. This process is repeated in a progressive manner until the desired target input dimension is reached.

3 Experiments and Discussion

To better investigate and explore the efficacy of the introduced notions of label distillation and progressive label distillation, a number of experiments were performed for the task of limited vocabulary speech recognition [2, 3, 5], where the underlying goal is to identify which word from a limited vocabulary was spoken based on an input audio utterance recording. In general, we will first explore direct label distillation for learning student networks with various input dimensions. We will then investigate the effectiveness of progressive label distillation through different teacher-student network pair configurations.

Table 2: Test accuracy of networks learnt using progressive label distillation

# Steps	C900	C800	C700	C600	C500
1					85.82
2				90.12	86.71
2			92.58		87.92
2		93.81			**89.22**
2	94.91				88.94
3			92.58	91.18	84.98
3		93.81		90.97	88.24
3	94.91		93.07		86.12
4		93.81	93.38	90.96	84.90
4	94.91	94.34	92.85		84.25
5	94.91	94.34	92.85	89.15	79.50

Direct Label Distillation In the first experiment, Table 1, we evaluate the efficacy of direct label distillation of learning input-efficient student networks via soft labels and hard labels for a set of student networks with five different input dimensions (i.e., $\{500ms, 600ms, 700ms, 800ms, 900ms\}$) by computing their respective test accuracies.

Progressive Label Distillation In the second experiment, Table 2, we evaluate the efficacy of progressive label distillation of learning input-efficient student networks for different combinations of series of teacher-student networks with progressively smaller input dimensions.

4 Conclusions and Future Work

In this study, we show that progressive label distillation can be leveraged for learning deep neural networks with reduced input dimensions without collecting and labeling new data. This reduction in input dimension results in input-efficient networks with significant reduction in the computation cost. Experiment results for the task of limited vocabulary speech recognition show that the use of progressive label distillation can lead to an input-efficient student network with half the input dimension with a test accuracy of 89.22%, compared to just 12.03% without using label distillation.

References

[1] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[2] Zhong Qiu Lin, Audrey G Chung, and Alexander Wong. Edgespeechnets: Highly efficient deep neural networks for speech recognition on the edge. *arXiv preprint arXiv:1810.08559*, 2018.

[3] Tara N Sainath and Carolina Parada. Convolutional neural networks for small-footprint keyword spotting. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[5] Raphael Tang and Jimmy Lin. Deep residual learning for small-footprint keyword spotting. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5484–5488. IEEE, 2018.