Wujie SUN, Defang CHEN, Can WANG, Deshi YE, Yan FENG, Chun CHEN, 2024. Multi-exit self-distillation with appropriate teachers. *Frontiers of Information Technology & Electronic Engineering*, 25(4):585-599.
https://doi.org/10.1631/FITEE.2200644

Multi-exit self-distillation with appropriate teachers

Key words: Multi-exit architecture; Knowledge distillation; Learning gap

Corresponding author: Can WANG
E-mail: wcan@zju.edu.cn
ORCID: https://orcid.org/0000-0002-5890-4307

Motivation

- Multi-exit architecture has been employed in self-distillation by distilling knowledge from different exits to improve the model capability. It enables adaptive inference with high accuracy without incurring extra computational cost.
- However, existing multi-exit self-distillation methods use mainly the knowledge from deep exits or a single ensemble to guide all exits, and ignore the fact that shallow exit performance may not be significantly improved due to the learning gap between the teacher and the student.

Main idea

- We provide students with an equal number of teachers which are obtained by different weighted combinations of exits' logits. Each student is required to learn mainly from its primary teacher whose knowledge is generally more appropriate for student learning.
- To prevent students from becoming overly focused on their primary teacher and failing to capture the rest valuable knowledge, students are asked to acquire some knowledge from other teachers as well.
- We use a neural network to calculate the weights for composing the teachers, and generate diverse and appropriate knowledge for each student by using a novel loss function.

Framework

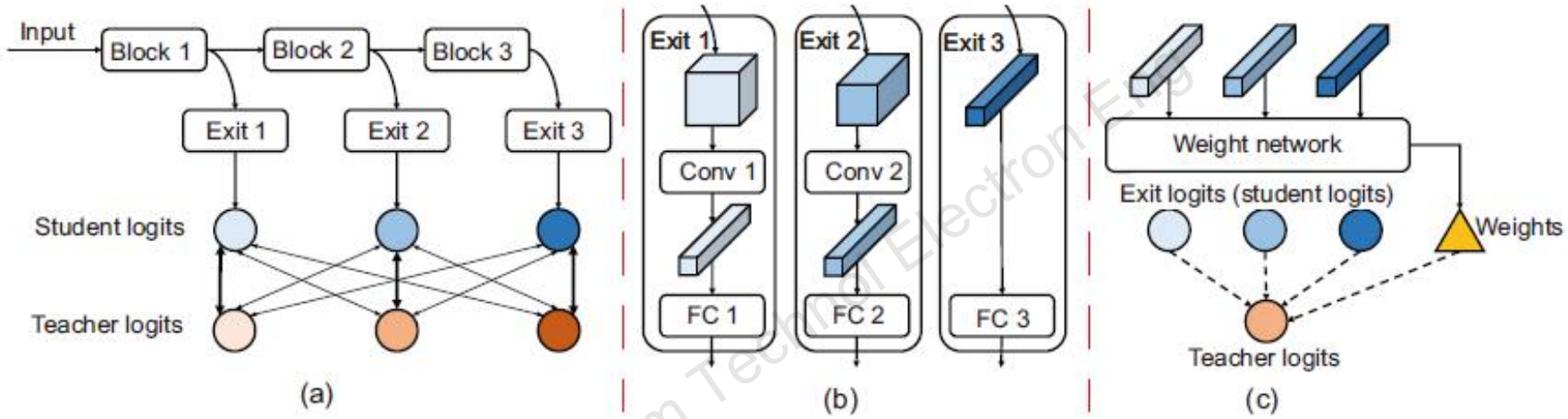
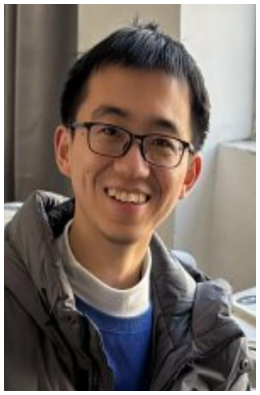


Fig. 4 Framework of MATE: (a) the overall framework, where two-way learning is required between students and teachers; students gain knowledge from teachers, and teachers dynamically adjust ensemble weights based on students' output; (b) exit architecture, where each exit resizes the block's output feature to match the dimensions of the last block's output feature, which is then inputted to the fully connected classifier to generate logits; (c) weight network, where a weight network based on self-attention is used to obtain teacher logits. It takes resized features as the input, and outputs the weights. Teacher logits are computed using weights and logits from all exits. Cuboids indicate the features and circles indicate the logits

Conclusions

In this work, we propose Multi-exit self-distillation with Appropriate TEachers (MATE) to provide diverse and appropriate knowledge for each exit. We highlight the necessity of controlling the learning gap between students and teachers. Experimental results show that our method consistently achieves better performance than state-of-the-art methods with various network architectures on multiple datasets.

Wujie SUN is a PhD candidate in Zhejiang University, China. His main research interests include computer vision and machine learning.



Defang CHEN is a PhD candidate in Zhejiang University, China. His main research interests include diffusion model and knowledge distillation.

Can WANG is a professor in Zhejiang University, China. His main research interests include data mining, machine learning, and information retrieval.

Deshi YE is an associate professor in Zhejiang University, China. His main research interests include online algorithm design and analysis, algorithm game theory, and optimization problems in machine learning.

Yan FENG is an associate professor in Zhejiang University, China. Her main research interests include database and data mining.

Chun CHEN is a professor in Zhejiang University, China.
His main research interests include data mining, computer vision, computer graphics, and embedded technology.