Selecting Useful Knowledge from Previous Tasks for Future Learning in a Single Network

Feifei Shi, Peng Wang, Zhongchao Shi, Yong Rui
Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
AI Lab, Lenovo Research
Lenovo Research

Introduction

Continual learning can learn new tasks incrementally while avoiding catastrophic forgetting. PackNet freezes and transfers all the weights of the previous tasks. Based this idea, We propose a novel gradient-based threshold method to reuse the knowledge of the previous tasks selectively when learning new tasks.

We creatively introduce a knowledge-selective mask, which can guide the network to make great use of previous weights selected based on current data. Our method can actively utilize the previous knowledge, and alleviate the negative effects that previous knowledge may bring.

Experiments

We evaluate our method on four classification datasets, and our method obtains competitive results when compared to PackNet. The results are shown in TABLE II.

TABLE II
CLASSIFICATION ACCURACY ON VGG16.

Dataset	PackNet	Ours	Individual network
CUBS	78.44	78.58	79.78
Stanford Cars	83.09	83.14	86.99
Flowers	89.06	89.04	91.78
Average	83.53	83.59	86.18

To further verify the effectiveness of the algorithm, we divide CUBS dataset into five disjoint sub-datasets. And the results in TABLE III and TABLE IV indicate that the gradient values in the network indeed show whether the previous parameters is useful for current data.

TABLE III
CLASSIFICATION ACCURACY ON SUB DATASETS OF CUBS.

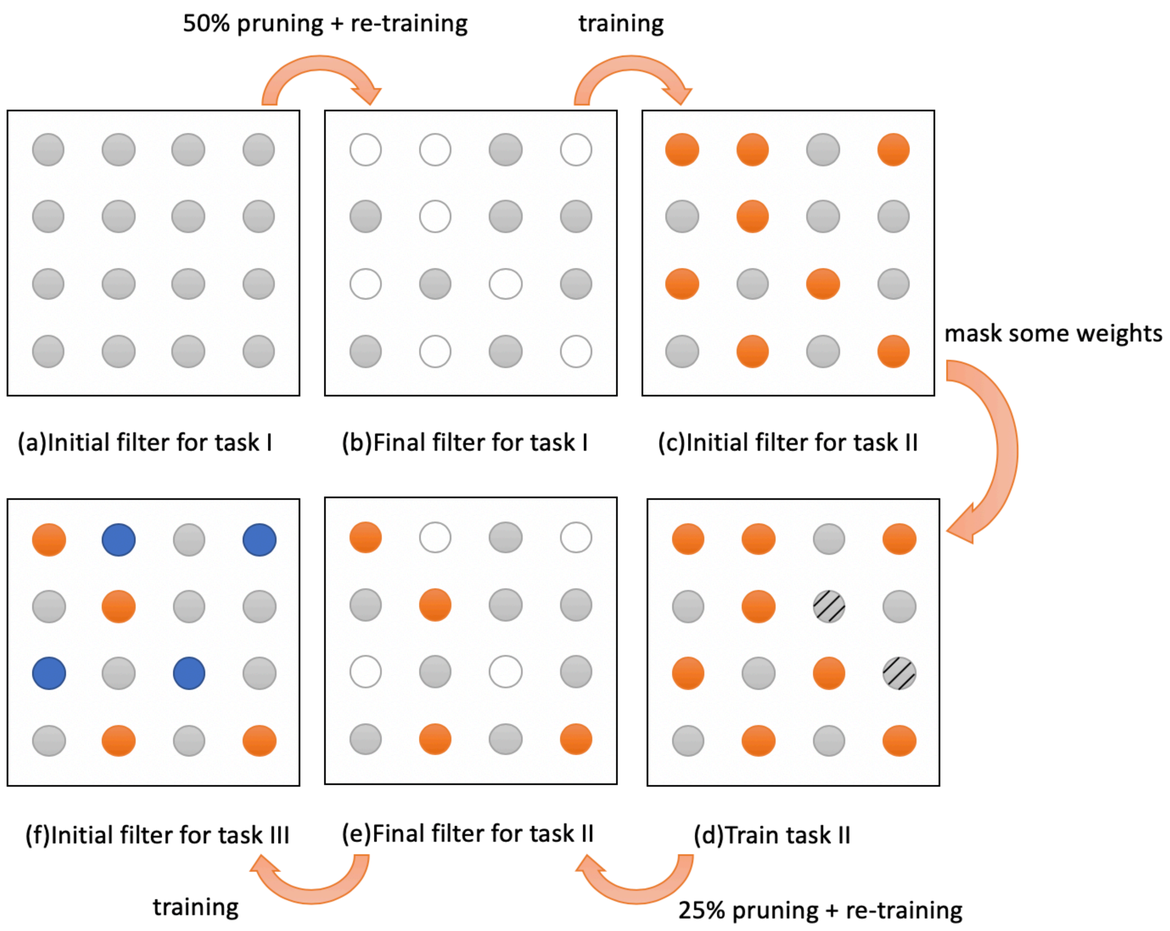
Dataset	PackNet	Ours	Individual network
CUB2	84.01	84.2	84.67
CUB3	83.54	83.64	84.29
CUB4	83.66	83.85	84.95
CUB5	82.8	82.84	83.42
Average	83.5	83.63	84.32

TABLE IV
CLASSIFICATION ACCURACY ON REVERSED SUB DATASETS OF CUBS.

Dataset	PackNet	Ours	Individual network
CUB4	84.97	85.26	84.95
CUB3	83.48	83.48	84.29
CUB2	83.55	83.89	84.67
CUB1	79.94	80	82.53
Average	82.98	83.16	84.11

Methods

The main process of our algorithm is packing multiple tasks into a single network incrementally by iterative pruning and re-training network. We freeze the weights of the previous tasks to avoid catastrophic forgetting. When learning the current task, the knowledge-selective mask will be generated according to the gradient in the network. The previous weights that may prevent current learning will be masked, and the remaining previous weights will be reused by the current learning.



The gradient-based threshold method can generate the knowledge-selective mask to pick suitable previous weights for the current task. We filter out previous weights that may prevent current learning according to the range of absolute value of the gradient, in addition, we set a new threshold to control the number of the masked weights for ensuring the transfer ability of the model.

$$W_0 = k * (|g|_{max} - |g|_{min}) + b$$

The previous weights that are not helpful for the current learning will be masked by the knowledge-selective mask. And our method can reuse the knowledge of the previous tasks selectively.

Conclusions

We have presented a novel method of picking suitable previous knowledge for current learning while avoiding catastrophic forgetting. The proposed method can make great use of the gradient values in the current network, the important weights learned in the past can be selected by knowledge-selective mask to learn the current data.