SDKD: Saliency Detection with Knowledge Distillation

Lin Yang[1,2,3]

1.School of Computer Science and Technology, Shandong University, Qingdao, China
2.Advanced innovation Center For Future Visual Entertainment (AICFVE), Beijing, China
3. Peking University, Beijing, China

ermu0628@gmail.com

***Abstract*—Lifelong learning is challenging for deep neural networks due to their susceptibility to catastrophic forgetting. Catastrophic forgetting occurs when a trained network is not able to maintain its ability to accomplish previously learned tasks when it is trained to perform new tasks. We study the problem of lifelong object recognition, extending a trained network to new task with a slightly forgetting of previous tasks. In contrast to traditional methods, we disentangle this problem with two aspects: background remove problem and classification problem. Qualitative and quantitaive experimental results on datasets show that the effectiveness of the proposed approach.**

I. Introduction

It is well-known that neural networks (NNs) suffer from catastrophic forgetting (CF) (McCloskey & Cohen, 1989), which refers to the phenomenon that when learning a sequence of tasks, the learning of each new task may cause the NN to forget the models learned for the previous tasks. Without solving this problem, an NN is hard to adapt to lifelong or continual learning, which is important for AI.

Problem Statement: Given a sequence of tasks $T = (T_1, T_2, \ldots, T_N)$*, we want to learn them one by one in the given sequence such that the learning of each new task will not forget the models learned for the previous tasks.*

In recent years, many approaches (often called continual learning) have been proposed to lessen the effect of CF, e.g., elastic weight consolidation (EWC) (Kirkpatrick et al., 2017), gradient episodic memory (GEM) (Lopez-Paz et al., 2017), generative replay (GR) (Shin et al., 2017), etc.

Lifelong object recognition task need to consider many metrics: average accuracy, model size, inference time, replay size and so on (She et al., 2019, Feng et al.,2019). Although these existing studies devote to solving CF problem, they are not suitable for this task. E.g., EWC is very effective for CF problem, but it need too much time to compute every parameter's gradient. So, use suitable model and constraints to solve this problem can get better result.

In this paper, we introduce a model to solve this task. At first, we need to get image saliency map, and then use Knowledge Distillation (KD) to address catastrophic forgetting for traditional supervised method, the result on the validation and test datasets show our model's effectiveness.

Figure 1.

II. Method

Rethink this diffculty-incremental task, what is "diffculty" in this task ? It means environmental factors in this task can become variant. For figure 1, due to the messy background, the "main" object's category shown on the left image can be hardly distinguished , but in contrast to the right image. So, firstly, the model need to know which object int the image need to be classified. Due to memory size , model size , and inference time constraints, we choose MobileNet-V2 as our network backbone. Figure 2 show the architecture of the network.

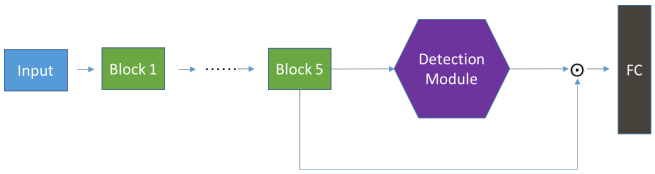


Figure 2. Overview of network architecture.

A. Saliency Detection

We use cascaded partial decoder framework which contains two branches to get image saliency map. In each branch, we use a fast and effective partial decoder. The first branch generates an initial saliency map which is utilized to refine the features of the second branch.

Since that the features of shallower layers contribute less to performance but have large resolution, we construct a partial decoder that only integrates features of deeper layers. Figure 3 show the architecture of detection module. We set the Block 5 as an optimization layer, and use the last two convolutional blocks to construct two branches (an attention one and a detection one). And then feed the detection result to refined layer to filter other "minor" objects and noise, only

retain the "main" object, and then feed it into classification layer, get the final result.

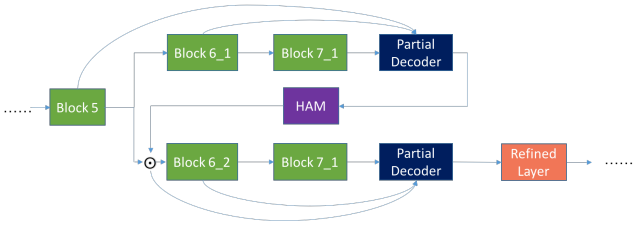


Figure 3 . The architecture of detection module.

B. *Knowledge Distillation*

For catastrophic forgetting problem, we use Knowledge Distillation to prevent it. Because our model can learn saliency detection to get the "main" object in the image, so the CF problem in our model perform in two aspects: saliency detection and classification. For i-th task, we regard (i-1)-th model as teacher network, and i-th model as student network.

For saliency detection, we use "Factor Transform" to constrain network forget previous tasks, it is said that if one fully understand a thing, he / she should be able to explain it by himself / herself. We use an autoencoder as a teacher translator, and an encoder as student translator, which has same architecture with teacher translator's encoder. Aim to project saliency maps from teacher network and student network to same space.

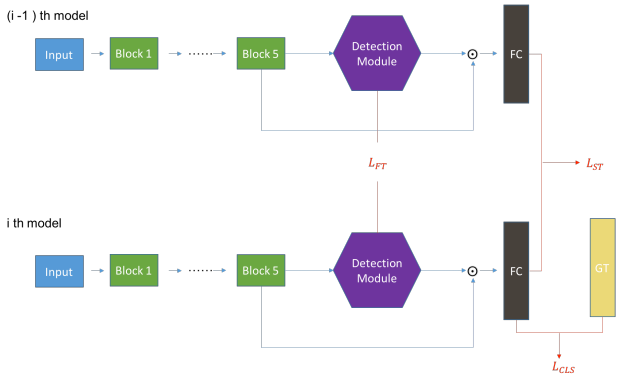


Figure 4. Knowledge distillation with the network.

In order to extract the factor from the teacher network, we train the teacher translator in an unsupervised way by assigning the reconstruction loss at the beginning of every task training preocess. Then use student translator to translate student network's saliency map output, compute L1 loss between teacher network output and student network. So, we can get loss:

$$L_{ft} = || \frac{F_T}{||F_T||_2} - \frac{F_S}{||F_S||_2} || \qquad (1)$$

For classification, we use soft target loss (Hinton, 2015),

$$L_{st} = KL(o_s, o_t)*(T*T*2*\alpha) \qquad (2)$$

where T = 20, α = 0.7.

Combine these loss, we get our total loss:

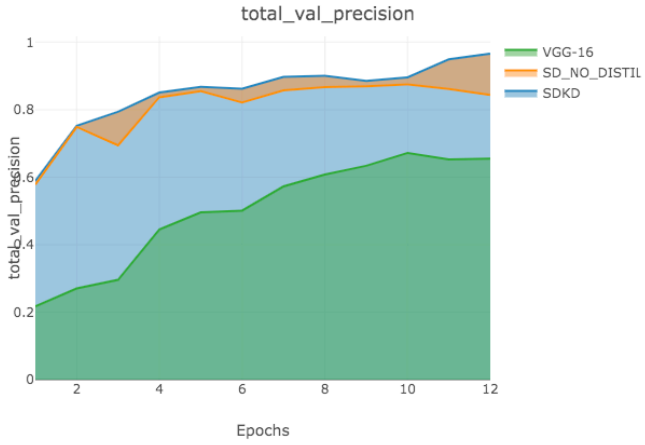
$$L = L_{cls} + \lambda_1 L_{ft} + \lambda_2 L_{st} \qquad (3)$$

where $\lambda_1 = 1$, $\lambda_2 = 1$.

Figure 4 shows the illustration of total loss.

III. Ablation Study

We use vgg-16 model as our baseline model, and use mean accuracy on validation dataset as metric.



References

[1] Wu, Zhe, Li Su, and Qingming Huang. "Cascaded Partial Decoder for Fast and Accurate Salient Object Detection." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[2] Kim, Jangho, SeongUk Park, and Nojun Kwak. "Paraphrasing complex network: Network compression via factor transfer." *Advances in Neural Information Processing Systems*. 2018.

[3] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

[4] Feng, F., Chan, R. H., Shi, X., Zhang, Y., & She, Q. (2019). Challenges in Task Incremental Learning for Assistive Robotics. IEEE Access.

[5] She, Q., Feng, F., Hao, X., Yang, Q., Lan, C., Lomonaco, V., ... & Qiao, F. (2019). OpenLORIS-Object: A Robotic Vision Dataset and Benchmark for Lifelong Deep Learning. arXiv preprint arXiv:1911.06487.