REPLACEMENT LEARNING: TRAINING VISION TASKS WITH FEWER LEARNABLE PARAMETERS

Anonymous authors
Paper under double-blind review

A SUPPLEMENTARY MATERIAL

A.1 EXPERIMENT IMPLEMENT DETAILS

In our experiments on CIFAR-10 Krizhevsky et al. (2009), SVHN Netzer et al. (2011), and STL-10 Coates et al. (2011) datasets, we utilize the AdamW optimizer Loshchilov & Hutter (2017) with a weight decay factor of 1e-4 for ViT-B, ViT-L Dosovitskiy et al. (2021), ResNet-32, and ResNet-110 He et al. (2016). We employ batch sizes of 1024 for CIFAR-10 Krizhevsky et al. (2009), SVHN Netzer et al. (2011), and STL-10 Coates et al. (2011). The training duration spans 250 epochs, starting with initial learning rates of 0.01, following a cosine annealing scheduler Loshchilov & Hutter (2016).

For ImageNet Deng et al. (2009), We use the AdamW optimizer Loshchilov & Hutter (2017) with a weight decay factor of 1e-4. Different hyperparameters are used for each architecture: batch size is 128 for ViT-B Dosovitskiy et al. (2021) and ResNet-34 He et al. (2016), and batch size is 32 for ResNet-101 and ResNet-152 He et al. (2016). Training lasts 100 epochs with initial learning rates of 0.04 for ViT-B Dosovitskiy et al. (2021) and ResNet-34 He et al. (2016), and 0.01 for ResNet-101 and ResNet-152 He et al. (2016).

We recognize that in the Transformer Encoder of the ViT Dosovitskiy et al. (2021) architecture, one layer consists of an MLP and a Multi-Head Attention. When freezing layers, we freeze only the gradients of the Multi-Head Attention, without altering the gradient descent of the MLP during forward propagation. For the ResNet He et al. (2016) architecture, we refer to each residual block as a layer, where each layer is composed of two convolutions. The entire layer is frozen during gradient freezing, with the parameters derived from the parameter integration mechanism entering the next layer via the residual connection.

A.2 GENERALIZATION STUDY

In this section, we aim to investigate the generalization performance of our proposed Replacement Learning. To evaluate its effectiveness, we utilize the checkpoints trained on the CIFAR-10 Krizhevsky et al. (2009) and test them on the STL-10 Coates et al. (2011), taking inspiration from previous work Qu et al. (2021).

As shown in Table 1, with the usage of our Replacement Learning, we witness a significant improvement in test accuracy, surpassing all backbones' end-to-end training Rumelhart et al. (1985). These findings emphasize the efficacy of our Replacement Learning in improving the generalization

Table 1: Generalization study. Checkpoints are trained on the CIFAR-10 and tested on the STL-10. The data in the table represents the test accuracy.

Backbone	Test Accuracy	Backbone	Test Accuracy
ResNet-32	36.88	ViT-B	28.31
ResNet-32*	**37.95 (↑ 1.07)**	**ViT-B***	**30.14 (↑ 1.83)**
ResNet-110	39.19	ViT-L	26.25
ResNet-110*	**39.76 (↑ 0.57)**	**ViT-L***	**28.02 (↑ 1.77)**

capabilities of supervised learning, ultimately leading to enhanced overall performance in the image classification task.

A.3 Algotithm

Algorithm 1 Replace Learning

Initialize θ_l for all layers $l = 1$ to n
Set k as the interval for freezing layers
Define frozen layer indices $\mathcal{F} = \{\, l \mid l \bmod k = 0 \,\}$
Initialize learnable parameters a_l and b_l for $l \in \mathcal{F}$
for each mini-batch (x, y) **do**
    $h_0 \leftarrow x$
    for $l = 1$ to n **do**
        if $l \in \mathcal{F}$ **then**
            $\theta_l \leftarrow a_l \times \theta_{l-1} + b_l \times \theta_{l+1}$
            $h_l \leftarrow f_l(h_{l-1}; \theta_l)$
        else
            $h_l \leftarrow f_l(h_{l-1}; \theta_l)$
        end if
    end for
    Compute loss $\mathcal{L} \leftarrow \mathcal{L}(h_n, y)$
    Backpropagate to compute gradients
    for $l = n$ down to 1 **do**
        if $l \in \mathcal{F}$ **then**
            Compute gradients $\frac{\partial \mathcal{L}}{\partial a_l}$ and $\frac{\partial \mathcal{L}}{\partial b_l}$
            Update $a_l \leftarrow a_l - \eta \times \frac{\partial \mathcal{L}}{\partial a_l}$
            Update $b_l \leftarrow b_l - \eta \times \frac{\partial \mathcal{L}}{\partial b_l}$
        else
            Compute gradient $\frac{\partial \mathcal{L}}{\partial \theta_l}$
            Update $\theta_l \leftarrow \theta_l - \eta \times \frac{\partial \mathcal{L}}{\partial \theta_l}$
        end if
    end for
end for

References

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Zhan Qu, Huan Jin, Yang Zhou, Zhen Yang, and Wei Zhang. Focus on local: Detecting lane marker from bottom up via key point. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14122–14130, 2021.

David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985.