Anti-Retroactive Interference for Lifelong Learning

Runqi Wang[1], Yuxiang Bao[1], Baochang Zhang[1⋆], Jianzhuang Liu[2], Wentao Zhu[3], and Guodong Guo[4]

[1] Beihang University
[2] Huawei Noah's Ark Lab
[3] Kuaishou Technology
[4] Institute of Deep Learning, Baidu Research
{runqiwang,bczhang}@buaa.edu.cn

1 Convergence of the Task-Specific Models

Property *Given the set of models (parameters)* $\{\phi_1^t, ..., \phi_n^t, \phi_b^{t-1}\}$ *during training task n with Alg. 2 of the paper and the distance matrix* $\mathbf{dif}^*$ *in Eq. 8 of the paper, if every model can be optimized to a global optimum, then all these models converge to the same optimum.*

Analysis We fuse the information of each task-specific model parameters $\{\phi_1^t, ..., \phi_n^t, \phi_b^{t-1}\}$ through the distance matrix $\mathbf{dif}^*$:

$$\begin{bmatrix} \phi_1^{t\prime} \\ \cdot \\ \cdot \\ \cdot \\ \phi_n^{t\prime} \\ \phi_b^{t-1\prime} \end{bmatrix} = \mathbf{dif}^* \cdot \begin{bmatrix} \phi_1^t \\ \cdot \\ \cdot \\ \cdot \\ \phi_n^t \\ \phi_b^{t-1} \end{bmatrix}, \tag{11}$$

where $\phi_b^{t-1\prime} = \sum_{i=1}^{n+1}(d^*_{n+1,i} \cdot \phi_i^t) = \phi_f^t$. Thus, $\phi_b^{t-1\prime}$ and ϕ_f^t are the same.

Because the sum of the elements in each row of the matrix $\mathbf{dif}^*$ is equal to 1, there exists an eigenvector $\mu = [1, ..., 1]^T$ and an eigenvalue $\lambda = 1$ associated with μ [1], *i.e.*,

$$\mu = \mathbf{dif}^* \cdot \mu. \tag{12}$$

Due to the regularization (in Eq.7 of the paper) of $\mathbf{dif}$, the elements of $\mathbf{dif}^*$ tend to be the same ($\approx \frac{1}{n+1}$) with training progress. After a sufficiently large number of epochs, in the sense that t is large enough, the vector $[\phi_1^{t\prime}, ..., \phi_n^{t\prime}, \phi_b^{t-1\prime}]^T$ consisting of the task-specific and base models would converge to the same optimum according to Eq. 12. Moreover, it is a vector in the eigenspace associated with the eigenvalue $\lambda = 1$ [2].

$$\Phi \in E(\lambda), \lambda = 1, \tag{13}$$

⋆ Corresponding author.

where $E(\lambda)$ denotes the eigenspace associated with the eigenvalue λ.

When all the models, $\phi_1, ..., \phi_n, \phi_b$ are ideally optimized to the same model, they share the same knowledge, thus eliminating information loss and retroactive interference in the task-specific model fusion.

2 Visualization of A and B

In Fig. 9 below, the background mask $\mathbf{B}$ is selected as $(1-\mathbf{A}\circ\mathbf{A})$. We can clearly see that the spatial attention of the input $\mathbf{x}$ covers the target object regions. The background mask $\mathbf{B}$ focuses on the background regions of the input $\mathbf{x}$, which can guide the background attack.

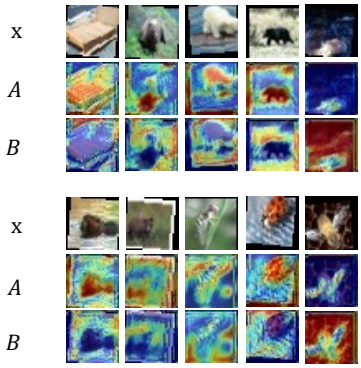

Fig. 9. Visualization results of the spatial attention and background mask. The areas with high values are shown in red, and those with low values are shown in blue.

References

1. Greub, W.H.: Linear algebra. Springer Science & Business Media (2012)
2. Kincaid, D.R., Cheney, E.W.: Numerical analysis : mathematics of scientific computing. Numerical analysis : mathematics of scientific computing (2002)