Anti-Retroactive Interference for Lifelong Learning

Runqi Wang¹, Yuxiang Bao¹, Baochang Zhang¹*, Jianzhuang Liu², Wentao Zhu³, and Guodong Guo⁴

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) \{φᵀ¹, ..., φᵀⁿ, φᵀ⁻¹\} during training task \(n\) with Alg. 2 of the paper and the distance matrix \(\text{dif}^*\) 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 \{φᵀ¹, ..., φᵀⁿ, φᵀ⁻¹\} through the distance matrix \(\text{dif}^*\):

\[
\begin{bmatrix}
\phiᵀ¹ \\
\vdots \\
\phiᵀⁿ \\
\phiᵀ⁻¹
\end{bmatrix} = \text{dif}^* \cdot 
\begin{bmatrix}
\phiᵀ¹ \\
\vdots \\
\phiᵀⁿ \\
\phiᵀ⁻¹
\end{bmatrix},
\]

where \(\phiᵀ⁻¹ = \sum_{i=1}^{n+1} (d_{n+1,i}^{*} \cdot \phiᵀ_i) = \phiᵀ_f\). Thus, \(\phiᵀ⁻¹\) and \(\phiᵀ_f\) are the same.

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

\[
\mu = \text{dif}^* \cdot \mu.
\]

Due to the regularization (in Eq. 7 of the paper) of \(\text{dif}\), the elements of \(\text{dif}^*\) tend to be the same (≈ \(\frac{1}{n+1}\)) with training progress. After a sufficiently large number of epochs, in the sense that \(t\) is large enough, the vector \([\phiᵀ¹, ..., \phiᵀⁿ, φᵀ⁻¹]^T\) 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].

\[
Φ \in E(\lambda), \lambda = 1,
\]

* Corresponding author.
where $E(\lambda)$ denotes the eigenspace associated with the eigenvalue $\lambda$.

When all the models, $\phi_1, \ldots, \phi_n, \phi_b$, 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 $B$ is selected as $(1 - A \circ A)$. We can clearly see that the spatial attention of the input $x$ covers the target object regions. The background mask $B$ focuses on the background regions of the input $x$, 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.](image)

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)