Reinterpretation of 'Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks'

Akihisa Watanabe
Waseda University
Tokyo, Japan
akihisa@ruri.waseda.jp

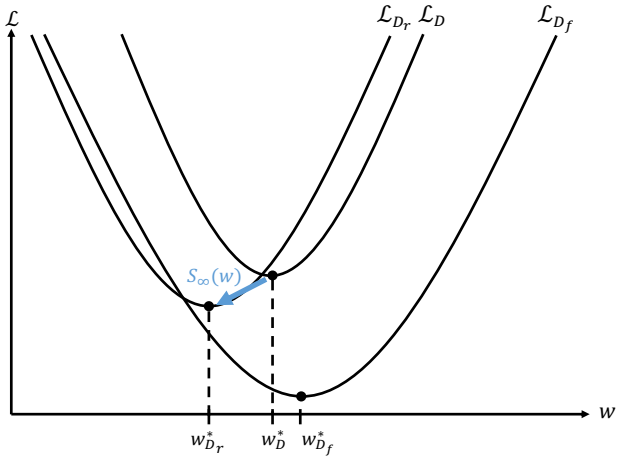

Figure 1. Shifting from the optimal value w^*_D, obtained from training on the entire dataset D, to the optimal value $w^*_{D_r}$ on the dataset D_r, that excludes the data D_f to be forgotten. The shift is guided by eq.(9).

Abstract

I report an alternative interpretation of 'Eternal Sunshine of the Spotless Net' [1], suggesting that the process can be viewed as a shift of the optimal value. Additionally, I identify and rectify a minor error in the proof of a proposition within the same paper.

1. Forgetting via Shifting optimal values

The objective of selective forgetting is to induce the model to forget specific data. More specifically, the goal is to emulate the state of the model as if it were retrained on the entire dataset excluding a specific subset, without actually having to retrain it from scratch. This is achieved by shifting the model's parameters from the optimal value obtained from training on the entire dataset to the optimal value on the dataset that excludes the data to be forgotten. In the following, we provide a mathematical formulation of how this is accomplished.

First, we break down the entire dataset D into the dataset D_f to be forgotten and its complement D_r, where set $D = D_f \cup D_r$. Here, we make the following assumption,

Assumption 1.1.

$$\mathcal{L}_D = \mathcal{L}_{D_r} + \mathcal{L}_{D_f}.$$

We assume that the loss function $\mathcal{L}_D$ in D can be expressed as the sum of the losses when learning with each dataset D_r and D_f individually.

Furthermore, we assume that $\mathcal{L}_{D_r}$ and $\mathcal{L}_{D_f}$ can be expressed as quadratic functions:

Assumption 1.2.

$$\mathcal{L}_{D_r}(w) = \frac{1}{2}b(w - w^*_{D_r})^2 \tag{1}$$

$$\mathcal{L}_{D_f}(w) = \frac{1}{2}c(w - w^*_{D_f})^2. \tag{2}$$

Where b and c are constants, w is the model parameter (a set of weights) and $w^*_{D_r}, w^*_{D_f}$ are the optimal values of each loss function(see Figure.1).

Here, the loss function $\mathcal{L}_D$ when learning with the entire dataset D is expressed as follows,

$$\begin{aligned}
\mathcal{L}_D(w) &= \mathcal{L}_{D_r}(w) + \mathcal{L}_{D_f}(w) \\
&= \frac{1}{2}b(w - w^*_{D_r})^2 + \frac{1}{2}c(w - w^*_{D_f})^2 \\
&= \frac{1}{2}b(w^2 - 2ww^*_{D_r} + w^{*^2}_{D_r}) + \frac{1}{2}c(w^2 - 2ww^*_{D_f} + w^{*^2}_{D_f}) \\
&= \frac{1}{2}\left\{(b+c)w^2 - 2w(bw^*_{D_r} + cw^*_{D_f}) - w^{*^2}_{D_r} + w^{*^2}_{D_f}\right\} \\
&= \frac{1}{2}(b+c)\left(w - \frac{bw^*_{D_r} + cw^*_{D_f}}{b+c}\right)^2 + C. \qquad (3)
\end{aligned}$$

For simplicity sake, we let C be a constant. From this, it can be seen that the optimal value of $\mathcal{L}_D$ is

$$w_D^* = \frac{bw_{D_r}^* + cw_{D_f}^*}{b+c}. \tag{4}$$

We assert that forgetting is accomplished by shifting from the optimal value w_D^* of the trained model parameters to the optimal value $w_{D_r}^*$ of the loss function when learning only with D_r(see Figure.1). In other words, we want to calculate the value of $w_{D_r}^*$, which can be derived from eq. (4) as follows,

$$\begin{aligned} w_D^*(b+c) &= bw_{D_r}^* + cw_{D_f}^* \\ w_{D_r}^* &= \frac{b+c}{b} w_D^* - \frac{c}{b} w_{D_f}^* \\ w_{D_r}^* &= w_D^* - \frac{c}{b} w_{D_f}^* \left(w_{D_f}^* - w_D^* \right). \end{aligned} \tag{5}$$

However, this formula requires $w_{D_f}^*$ to be known, so we need to devise a way.

Now, from Assumption(1.1), the gradient is given by,

$$\nabla \mathcal{L}_D = \nabla \mathcal{L}_{D_r} + \nabla \mathcal{L}_{D_f}. \tag{6}$$

Furthermore, using eq.(1) and eq.(2), we have

$$\begin{aligned} (b+c)(w - w_D^*) &= \nabla \mathcal{L}_{D_r} + c(w - w_{D_f}^*) \\ \nabla \mathcal{L}_{D_r} &= (b+c)(w - w_D^*) - c(w - w_{D_f}^*) \\ &= bw - bw_D^* + cw - cw_D^* - cw + cw_{D_f}^* \\ &= b(w - w_D^*) + c(w_{D_f}^* - w_D^*). \end{aligned} \tag{7}$$

Here, substituting $w = w_D^*$, we obtain

$$\nabla \mathcal{L}_{D_r}|_{w=w_D^*} = c(w_{D_f}^* - w_D^*). \tag{8}$$

By substituting eq.(5), we finally get

$$\begin{aligned} w_{D_r}^* &= w_D^* - b^{-1} \nabla \mathcal{L}_{D_r}|_{w=w_D^*} \\ &= w_D^* - \left[\nabla^2 \mathcal{L}_{D_r} \right]^{-1} \nabla \mathcal{L}_{D_r}|_{w=w_D^*}. \end{aligned} \tag{9}$$

This equation suggests that if the trained model has reached the optimal value w_D^*, it can be seen that forgetting can be performed by the simple *Newton Update*.

Also, this optimal value is obtained after the optimization algorithm has converged, $t \to \infty$. By setting $B = \nabla^2 \mathcal{L}_{D_r}$, we arrive at the same equation as stated in Proposition 3 of Golatkar *et al*. [1]:

$$S_\infty(w) = w - B^{-1} \nabla L_{D_r}(w). \tag{10}$$

2. Possible Error in Proof of Proposition 4

I noticed a minor error in Proposition 4 of the version you uploaded on Arxiv. I believe there is an omission of λ in your proof. I have detailed my understanding below for your consideration.

Proposition 2.1. Assume that $h(w)$ is close to w' up to some normally distributed error $h(w) - w' \sim N(0, \Sigma_h)$, and assume that $L_{D_r}(w)$ is (locally) quadratic around $h(w)$. Then the optimal scrubbing procedure in the form $S(w) = h(w) + n, n \sim N(0, \Sigma)$, that minimizes the Forgetting Lagrangian

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{\tilde{w} \sim S(w)} \left[L_{\mathcal{D}_r}(\tilde{w}) \right] \\ &\quad + \lambda \mathbb{E}_\epsilon \left[\mathrm{KL} \left(P(S(w) \mid \mathcal{D}, \epsilon) \| P\left(S_0(w) \mid \mathcal{D}_r, \epsilon\right) \right) \right] \end{aligned}$$

is obtained when $\Sigma B \Sigma = \lambda \Sigma_h$, where $B = \nabla^2 L_{\mathcal{D}_r}(k)$. In particular, if the error is isotropic, that is $\Sigma_h = \sigma_h^2 I$ is a multiple of the identity, we have $\Sigma = \sqrt{\lambda \sigma_h^2} B^{-1/2}$.

Proof. The proof for this proposition follows the same steps as in the paper [1] up to this point, and is therefore omitted for brevity. We continue from the following step:

$$\mathcal{L} = L_{D_r}(h(w)) + \frac{1}{2} \mathrm{tr}(B\Sigma) + \lambda \frac{1}{2} \mathrm{tr}(\Sigma^{-1} \Sigma_h)$$

We now want to find the optimal covariance Σ of the noise to add in order to forget. Setting $\nabla L_\Sigma = 0$ We obtain the condition

$$\begin{aligned} \nabla L_\Sigma &= 0 \\ \frac{1}{2} B - \lambda \frac{1}{2} \Sigma^{-1} \Sigma_h \Sigma^{-1} &= 0 \\ \lambda \Sigma^{-1} \Sigma_h \Sigma^{-1} &= B \\ \lambda \Sigma_h &= \Sigma B \Sigma. \end{aligned}$$

If we further assume the error Σ_h to be isotropic, that is, $\Sigma_h = \sigma_h^2 I$, then this condition simplifies to

$$\begin{aligned} \lambda \sigma_h^2 \Sigma^{-1} I \Sigma^{-1} &= B \\ \lambda \sigma_h^2 \left(\Sigma^{-1} \right)^2 &= B \\ \Sigma^2 &= \lambda \sigma_h^2 B^{-1} \\ \Sigma &= \sqrt{\lambda \sigma_h^2} B^{-\frac{1}{2}}. \end{aligned}$$

□

In summary, it appears that λ is missing from eq.(2) in the paper. I've tried to fill in the missing steps in the proof above. I would appreciate it if you could take a moment to review this.

References

[1] A. Golatkar, A. Achille, and S. Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9301–9309, 2020. 1, 2