Self-teaching for machine translation

Tianyi Zhou

November 14, 2021

$$\begin{array}{ll} \min_A & L(D_{val}; V^*(W^*(A))) \\ s.t & V^*(W^*(A)) = \text{argmin}_V \sum_{i=1}^{M} l(u_i, g(u_i; W^*(A)); V) \\ & W^*(A) = \text{argmin}_W \sum_{i=1}^{N} a_i l(x_i, y_i; W) \end{array}$$

1 Optimization Algorithm

One-step gradient decent of $\widetilde{W}(A)$:

$$\widetilde{W}(A) \doteq \overline{W}(A) = W - \xi_W \nabla_W \sum_{i=1}^{N} a_i l(x_i, y_i, W)$$

Approximation of $\widetilde{V}(\widetilde{W}(A))$:

$$\widetilde{V}(\widetilde{W}(A)) \doteq \overline{V}(\overline{W}(A))$$

$$= V - \xi_v \nabla_v \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}(A); V) \tag{1}$$

For the validation stage:

$$min_A L(D_{val}, \overline{V}(\overline{W}(A)))$$

$$A = A - \xi_A \nabla_A \overline{V}(\overline{W}(A)) \cdot \nabla_{\overline{V}} L(D_{val}, \overline{V}) \tag{2}$$

$$\nabla_A \overline{V}(\overline{W}(A)) = \nabla_A (V - \xi_V \nabla_V \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}(A)); V))$$

$$= \xi_V \nabla_V \nabla_A \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}(A)); V)$$

$$= \xi_V \cdot [\nabla_V \nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V) \cdot \nabla_A \overline{W}(A)]$$

now equation (2) becomes:

$$A = A - \xi_A \xi_V \cdot \nabla_V \nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V) \cdot \nabla_A \overline{W}(A) \cdot \nabla_{\overline{V}} L(D_{val}, \overline{V}) \quad (3)$$

Using Finite difference method to approximate

$$\nabla_V \nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V) \cdot \nabla_{\overline{V}} L(D_{val}, \overline{V})$$

$$= \frac{1}{2\alpha_W} (\nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V^+) - \nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V^-))$$

where:

$$V^{\pm} = V \pm \alpha_W \nabla_{\overline{V}} L(D_{val}, \overline{V})$$

$$\alpha_W = \frac{0.01}{||\nabla_{\overline{V}} l(D_{val}, \overline{V})||_2}$$

Second term in (3):

$$\nabla_A \overline{W}(A) = \nabla_A (W - \xi \nabla_W \sum_{i=1}^{N} (\alpha_i l(x_i, y_; W))$$

$$= \xi_W \nabla_W \nabla_A \sum_{i=1}^{n} \alpha_i l(x_i, y_i; W)$$

now equation (3) becomes:

$$A = A - \xi_A \xi_V [\frac{1}{2\alpha_W} (\nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V^+) - \nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V^-))] \cdot [\xi_W \nabla_W \nabla_A \sum_{i=1}^{n} \alpha_i l(x_i, y_i; W)]$$

$$A = A - \xi_W \xi_A \xi_V \frac{1}{2\alpha_W} [(\nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V^+) - \nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V^-))] \cdot [\nabla_W \nabla_A \sum_{i=1}^{n} \alpha_i l(x_i, y_i; W)]$$

Using Finite difference method to approximate

$$[\nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V^+)] \cdot [\nabla_W \nabla_A \sum_{i=1}^{n} \alpha_i l(x_i, y_i; W)]$$

$$= \frac{1}{2\alpha_A}(\nabla_A \sum_{i=1}^{n} \alpha_i l(x_i, y_i; W^+) - \nabla_A \sum_{i=1}^{n} \alpha_i l(x_i, y_i; W^-))$$

where:

$$W^{\pm} = W \pm \alpha_A(\nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V^+))$$

$$\alpha_A = \frac{0.01}{||\nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V^+)||_2}$$

Using Finite difference method to approximate

$$[\nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V^-)] \cdot [\nabla_W \nabla_A \sum_{i=1}^{n} \alpha_i l(x_i, y_i; W)]$$

$$= \frac{1}{2\alpha_A}(\nabla_A \sum_{i=1}^{n} \alpha_i l(x_i, y_i; W^+) - \nabla_A \sum_{i=1}^{n} \alpha_i l(x_i, y_i; W^-))$$

where:

$$W^{\pm} = W \pm \alpha_A(\nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V^-))$$

$$\alpha_A = \frac{0.01}{||\nabla_{\overline{W}} \sum_{i=1}^{M} l(u_i, g(u_i, \overline{W}); V^-)||_2}$$