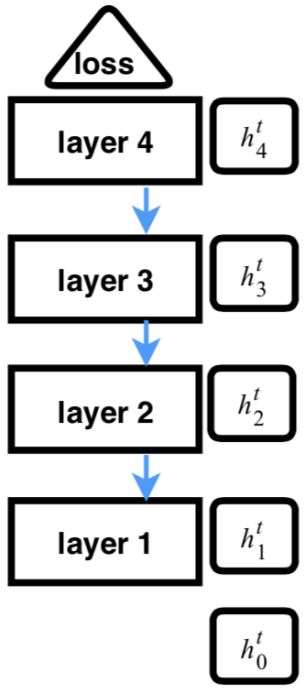
Training Neural Networks Using Features Replay

Zhouyuan Huo[1], Bin Gu[1,2], Heng Huang[1,2]

[1]Department of Electrical and Computer Engineering, University of Pittsburgh
[2] JD.com

November 28, 2018



Method BP

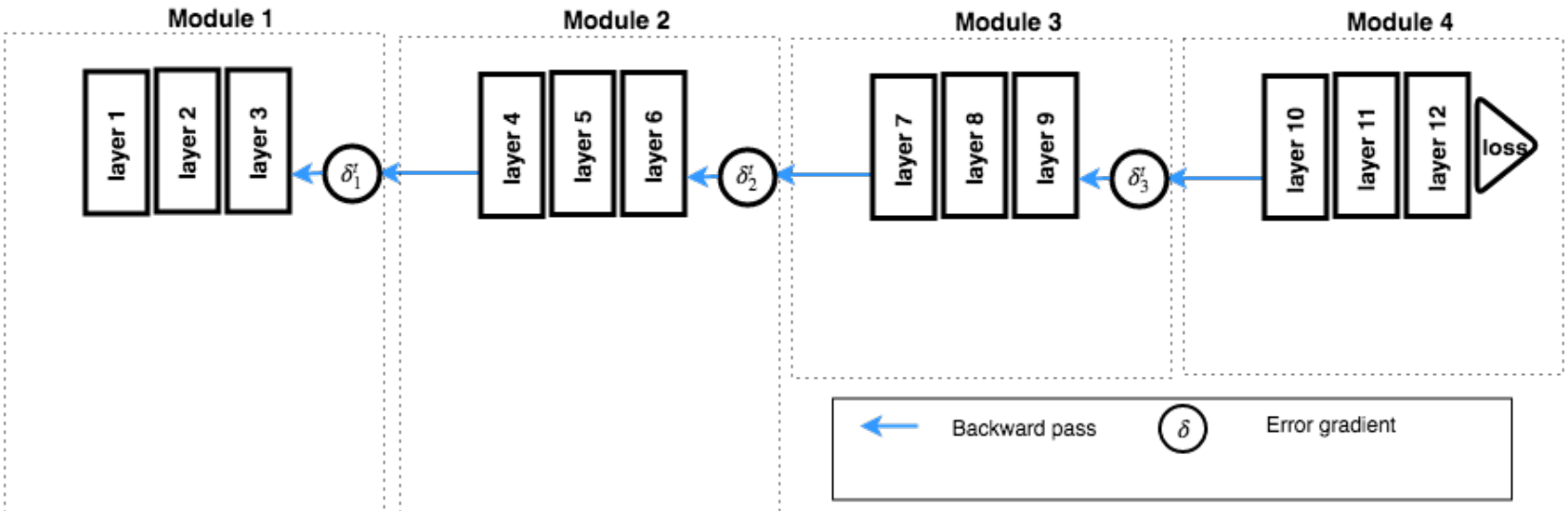
Backpropagation algorithm:

- step 1: Forward pass.
- step 2: Backward pass.

Problem:

- Backward time is about **2** times of forward time.
- Backward locking.
- Backward cannot be parallellized.

Problem Reformulation poster #12

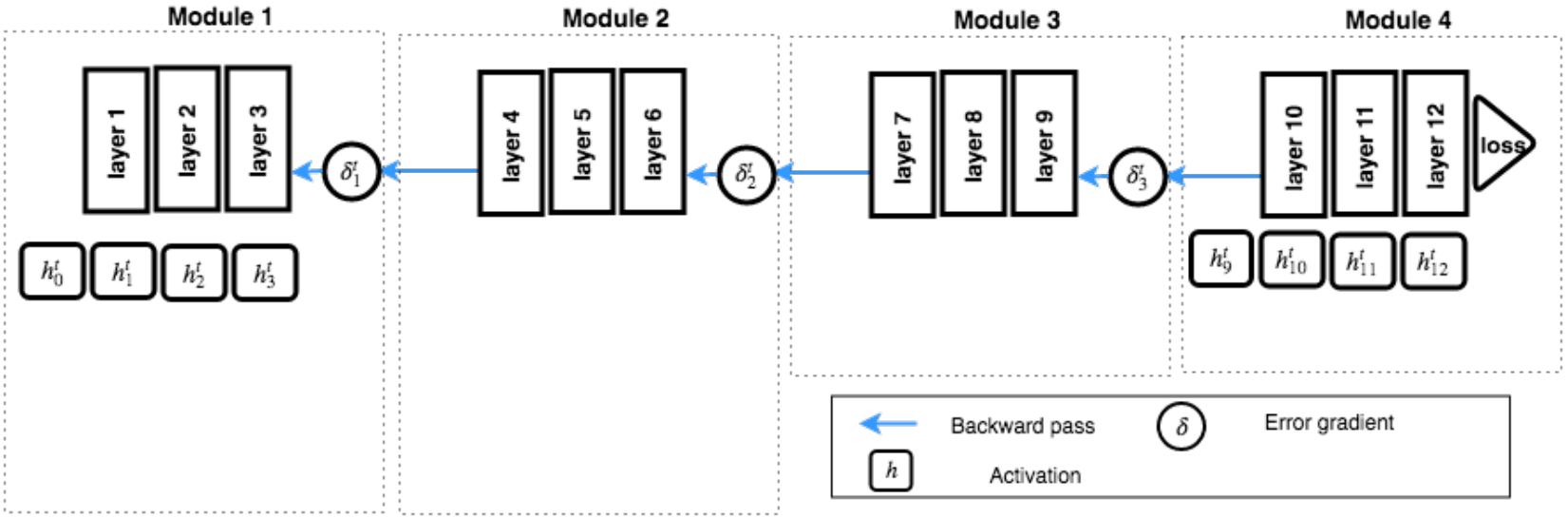


Original formulation:

$$\min_{w} \quad f(h_L, y)$$
$$s.t. \quad h_l = F_l(h_{l-1}; w_l)$$

New formulation:

$$\min_{w,\delta} \quad \sum_{k=1}^{K-1} \left\| \delta_k^t - \frac{\partial f_{h_{L_k}^t}(w^t)}{\partial h_{L_k}^t} \right\|_2^2 + f\left(h_{L_K}^t, y^t\right)$$
$$s.t. \quad h_{L_k}^t = F_{\mathcal{G}(k)}(h_{L_{k-1}}^t; w_{\mathcal{G}(k)}^t)$$



Module 1:

$$\min_{w,\delta} \quad \left\| \delta_1^t - \frac{\partial f_{h_{L_1}^t}(w^t)}{\partial h_{L_1}^t} \right\|_2^2$$

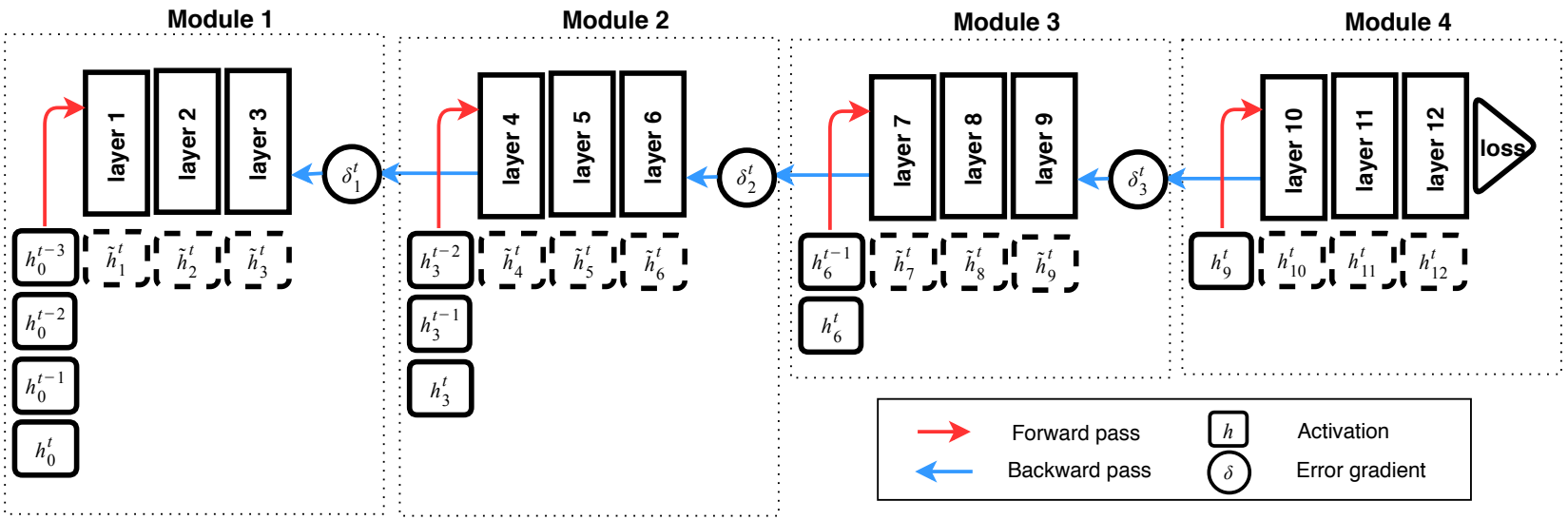
$$s.t. \quad h_{L_1}^t = F_{\mathcal{G}(1)}(h_{L_0}^t; w_{\mathcal{G}(1)}^t)$$

We approximate $\delta_1^t = \frac{\partial f_{h_{L_1}^{t-3}}(w^t)}{\partial h_{L_1}^{t-3}}$.

Module 4:

$$\min_{w,\delta} \quad f\left(h_{L_4}^t, y^t\right)$$

$$s.t. \quad h_{L_4}^t = F_{\mathcal{G}(4)}(h_{L_3}^t; w_{\mathcal{G}(4)}^t)$$



Forward pass:

$$h^t_{L_k} = F_{\mathcal{G}(k)}\left(h^t_{L_{k-1}}; w^t_{\mathcal{G}(k)}\right) \quad \text{(Play)}$$

Backward pass:

$$\tilde{h}^t_{L_k} = F_{\mathcal{G}(k)}(h^{t+k-K}_{L_{k-1}}; w^t_{\mathcal{G}(k)}) \quad \text{(Replay)}$$

Apply chain rule using $\tilde{h}^t_{L_k}$ and δ^t_k in each module.

Assumption 1 ***(Sufficient direction)*** *We assume that the expectation of the descent direction* $\mathbb{E}\left[\sum_{k=1}^{K} g^t_{\mathcal{G}(k)}\right]$ *in Algorithm 1 is a sufficient descent direction of the loss* $f(w^t)$ *regarding* w^t. *Let* $\nabla f(w^t)$ *denote the full gradient of the loss, there exists a constant* $\sigma > 0$ *such that,*

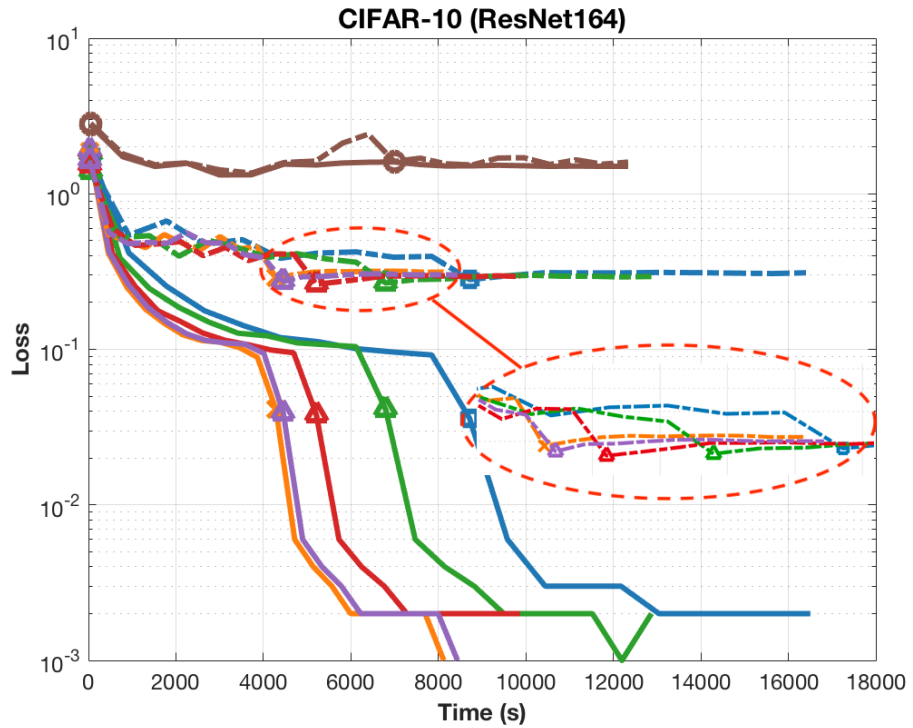
$$\left\langle \nabla f(w^t), \mathbb{E}\left[\sum_{k=1}^{K} g^t_{\mathcal{G}(k)}\right] \right\rangle \geq \sigma \|\nabla f(w^t)\|_2^2. \tag{8}$$

Convergence Guarantee:

$$\frac{1}{\sum_{t=0}^{T-1}\gamma_t}\sum_{t=0}^{T-1}\gamma_t \mathbb{E}\left\|\nabla f(w^t)\right\|_2^2 \leq \frac{f(w^0)-f(w^*)}{\sigma\sum_{t=0}^{T-1}\gamma_t} + \frac{LM}{2\sigma}\frac{\sum_{t=0}^{T-1}\gamma_t^2}{\sum_{t=0}^{T-1}\gamma_t}. \tag{1}$$

Experimental Results poster #12

- Faster Convergence.
- Lower Memory Consumption.
- Better Generalization Error.



Thanks !

Welcome to poster #12
Room 210 & 230 AB