Appendix

Caoyun Fan[1,†], Wenqing Chen[2,†], Jidong Tian[1], Yitian Li[1], Hao He[1,*], and Yaohui Jin[1]

[1] Shanghai Jiao Tong University, Shanghai, China
{fcy3649, frank92, yitian_li, hehao, jinyh}@sjtu.edu.cn
[2] Sun Yat-sen University, Guangzhou, China
chenwq95@mail.sysu.edu.cn

1 Validity of SGD Algorithm

The SGD process is expressed as follows:

$$
\begin{aligned}
\widehat{\mathcal{R}}(\theta_t) &= \frac{1}{|S|} \sum_{(X_i,Y_i)\in S} l(F_{\theta_t}(X_i), Y_i) \\
\widehat{g_S}(\theta_t) &= \nabla_{\theta_t} \widehat{\mathcal{R}}(\theta_t) \\
\theta_{t+1} &= \theta_t - \eta \cdot \widehat{g_S}(\theta_t)
\end{aligned}
\tag{1}
$$

The validity of SGD algorithm is guaranteed:

$$
\begin{aligned}
\mathcal{R}(\theta) &= \frac{1}{|S|} \sum_{i=1}^{|S|} l_i(\theta) = \mathbb{E}\left[\widehat{\mathcal{R}}(\theta)\right] \\
\nabla_\theta \mathcal{R}(\theta) &= \frac{1}{|S|} \sum_{i=1}^{|S|} \nabla_\theta l_i(\theta) = \mathbb{E}\left[\nabla_\theta \widehat{\mathcal{R}}(\theta)\right]
\end{aligned}
\tag{2}
$$

Eq. 2 shows $\widehat{\mathcal{R}}(\theta)$ and $\nabla_\theta \widehat{\mathcal{R}}(\theta)$ are the un-biased estimations of $\mathcal{R}(\theta)$ and $\nabla_\theta \mathcal{R}(\theta)$.

2 Dataset Introduction

2.1 NYUv2

The NYUv2 dataset is comprised of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. It is a challenging indoor scene dataset in various room types (bathrooms, living rooms, studies, etc.), and this dataset has three tasks: 13-class semantic segmentation, depth estimation, and surface normal prediction.

[†] These authors contributed equally.

[⋆] Corresponding author.

2.2 CityScapes

Cityscapes is a large-scale database that focuses on semantic understanding of urban street scenes. It provides semantic, instance-wise, and dense pixel annotations for 30 classes grouped into 8 categories (flat surfaces, humans, vehicles, constructions, objects, nature, sky, and void). The dataset consists of around 5000 fine annotated images and 20000 coarse annotated ones. Data was captured in 50 cities during several months, daytimes, and good weather conditions. It was originally recorded as video so the frames were manually selected to have the following features: a large number of dynamic objects, varying scene layout, and varying background.

3 Universality of Weight Design

For the multiple losses of multi-task learning, the most popular method is to design an appropriate weight coefficient for each task and get the final loss by weighted sum as $L = \sum_{k=1}^{T} \omega_k L_k$. In fact, for any design of loss function:

$$L = f(L_1, L_2, \ldots, L_n) \tag{3}$$

As long as the function satisfies strict monotonicity and continuity for each loss $\{L_1, L_2, \ldots, L_n\}$, it can meet the optimization requirements of BP algorithm. When calculating the gradient for parameter, because of the chain rule, the gradient is expressed as Eq. 4:

$$\frac{dL}{d\theta} = \sum_{i=1}^{n} \frac{dL}{dL_i} * \frac{dL_i}{d\theta} \tag{4}$$

So no matter what form of loss function, it will degenerate into a specific weight design in BP algorithm, as shown in Eq. 5:

$$\begin{aligned} \frac{dL}{d\theta} &= \sum_{i=1}^{n} \omega_i * \frac{dL_i}{d\theta} \\ \text{where} \quad \omega_i &= \frac{dL}{dL_i} \end{aligned} \tag{5}$$

4 Momentum Estimation

The gradient $\widehat{g_S}(\theta)$ obtained by SGD is stochastic, so the gradient can be decomposed into expected gradient and gradient noise:

$$\begin{aligned} &\widehat{g_S}(\theta) = g(\theta) + n_{g(\theta)} \\ &\text{where} \quad n_{g(\theta)} \sim \mathcal{N}(0, \frac{C}{|S|}) \end{aligned} \tag{6}$$

According to Eq. 6, the momentum gradient with noise can be expressed as:

$$\begin{aligned} m_t &= \gamma m_{t-1} + (1-\gamma) \cdot \widehat{g_S}(\theta_t) \\ &\approx (1-\gamma) \sum_{i=1}^{t} \gamma^{t-i} \cdot \widehat{g_S}(\theta_i) \\ &= (1-\gamma) \sum_{i=1}^{t} \gamma^{t-i} \cdot g(\theta_i) + n_{m_t} \end{aligned} \tag{7}$$

Since we assume that the gradient noise obeys the same distribution, and n_{m_t} can be simplified as:

$$\begin{aligned} n_{m_t} &= (1-\gamma) \sum_{i=1}^{t} \gamma^{t-i} \cdot n_{g(\theta_i)} \\ &= (1-\gamma)\, \mathcal{N}(0, \frac{\sum_{i=1}^{t} \gamma^{2(t-i)} C}{|S|}) \\ &\approx (1-\gamma)\, \mathcal{N}(0, \frac{C}{(1-\gamma^2)\,|S|}) \\ \therefore n_{m_t} &\sim \sqrt{\frac{1-\gamma}{1+\gamma}} \mathcal{N}(0, \frac{C}{|S|}) \end{aligned} \tag{8}$$

5 Gradient Norm

The 2-norm. The gradient noise is represented as $n \sim \mathcal{N}(0, C/|S|)$, where $C/|S|$ as the covariance matrix is positive semi-definite. Thus, $C/|S|$ can be orthogonally diagonalized as $C/|S| = Q^T \mathit{\Lambda} Q$, where $\mathit{\Lambda}$ is a non-negative diagonal matrix, and $tr(C/|S|) = tr(C)/|S| = tr(\mathit{\Lambda})$. This means that any Gaussian noise can be rotated to another Gaussian noise with each component orthogonal. Therefore, the high-dimensional noise N is decomposed into multiple orthogonal one-dimensional noises $N = \{n_1, \ldots, n_n\}$, and $\mathbb{E}[N^2] = \sum_{k=0}^{n} \mathbb{E}[n_k^2]$. According to the nature of Gaussian distribution (This is the reason why we select the 2-norm.): When $x \sim \mathcal{N}(0, \sigma^2)$, $\mathbb{E}[x^2] = \sigma^2$, we get $\mathbb{E}[n_k^2] = \sigma_k^2$. Since $\sigma_k^2 = \mathit{\Lambda}_{kk}$ in the high-dimensional Gaussian distribution, $\mathbb{E}[N^2] = \sum_{k=0}^{n} \mathit{\Lambda}_{kk} = tr(\mathit{\Lambda}) = tr(C)/|S|$. We will add the explanation to the supplementary material in the next version.