Deep Learning: Homework #3

Due on April 28, 2021

Professor Zhen Li

Haoyu Kang
Sno.220041025

Problem 1

Solution

Subproblem (a)

$$
\begin{aligned}
[[L_t f] * w](x) =& \sum_{y\in Z^2}\sum_{k=1}^{K}[L_t f]_k(y) w_k(y-x) \\
=& \sum_{y\in Z^2}\sum_{k=1}^{K} f_k(y-t) w_k(y-x) \\
=& \sum_{y\in Z^2}\sum_{k=1}^{K} f_k(y) w_k(y+t-x) \\
=& \sum_{y\in Z^2}\sum_{k=1}^{K} f_k(y) w_k(y-(x-t) \\
[[L_t f] * w](x) =& [f * w](x-t) \\
=& \sum_{y\in Z^2}\sum_{k=1}^{K} f_k(y) w_k(y-(x-t)
\end{aligned}
\tag{1}
$$

Subproblem (b)

$$
\begin{aligned}
[[L_R f] * w](x) =& \sum_{y\in Z^2}\sum_{k=1}^{K}[L_R f]_k(y) w_k(y-x) \\
=& \sum_{y\in Z^2}\sum_{k=1}^{K} f_k(R^{-1}y) w_k(y-x) \\
=& \sum_{y\in Z^2}\sum_{k=1}^{K} f_k(y) w_k(Ry-x) \\
=& \sum_{y\in Z^2}\sum_{k=1}^{K} f_k(y) w_k(R(y-R^{-1}x) \\
=& \sum_{y\in Z^2}\sum_{k=1}^{K} f_k(y)([L_{R^{-1}} w]_k(y-R^{-1}x)) \\
L_R[f * [L_{R^{-1}} w]](x) =& [f * [L_{R^{-1}} w]](R^{-1}x) \\
=& \sum_{y\in Z^2}\sum_{k=1}^{K} f_k(y)([L_{R^{-1}} w]_k(y-R^{-1}x))
\end{aligned}
\tag{2}
$$

Problem 1 continued on next page...

Subproblem (c)

$$
\begin{aligned}
[[L_u f] * w](g) &= \sum_{h \in G} \sum_{k=1}^{K} [L_u f]_k(h) w_k(g^{-1}h) \\
&= \sum_{h \in G} \sum_{k=1}^{K} f_k(u^{-1}h) w_k(g^{-1}h) \\
&= \sum_{h \in G} \sum_{k=1}^{K} f_k(h) w_k(g^{-1}uh) \\
&= \sum_{h \in G} \sum_{k=1}^{K} f_k(h) w_k((u^{-1}g)^{-1}h) \\
&= [f * w](u^{-1}g) = [L_u[f * w]](g)
\end{aligned}
\tag{3}
$$

Problem 2

Solution

$$
\begin{aligned}
\frac{\partial L}{\partial W_{hz}} &= \sum_t \frac{\partial L_t}{\partial W_{hz}} = \sum_t \frac{\partial L_t}{\partial z_t} \frac{\partial z_t}{\partial net_t} \frac{\partial net_t}{\partial w_{hz}} \\
&= \sum_t \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \frac{-1}{Z_k} \\ \vdots \\ 0 \end{bmatrix}^\top \cdot \begin{bmatrix} \frac{\partial Z_{t,1}}{\partial Net_{t,1}} & \cdots & \frac{\partial Z_{t,1}}{\partial Net_{t,c}} \\ & \vdots & \\ \frac{\partial Z_{t,c}}{\partial Net_{t,1}} & \cdots & \frac{\partial Z_{t,c}}{\partial Net_{t,c}} \end{bmatrix} \cdot \begin{bmatrix} h_{t,1} & h_{t,2} & \cdots & h_{t,c} \end{bmatrix} \\
&= \sum_t \left\{ \begin{bmatrix} 0 & 0 & \cdots & \frac{-1}{Z_k} & \cdots & 0 \end{bmatrix} \cdot \begin{bmatrix} Z_{t,1} - Z_{t,1}^2 & -Z_{t,1}Z_{t,2} & \cdots & -Z_{t,1}Z_{t,c} \\ & & \vdots & \\ -Z_{t,1}Z_{t,c} & -Z_{t,c}Z_{t,2} & \cdots & Z_{t,c} - Z_{t,c}^2 \end{bmatrix} \right\}^\top \cdot \begin{bmatrix} h_{t,1} & h_{t,2} & \cdots & h_{t,c} \end{bmatrix} \\
&= \sum_t (\boldsymbol{Z_t} - \boldsymbol{y_t})^\top h_t
\end{aligned}
\tag{4}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial W_{hh}} =& \sum_t \frac{\partial L_t}{\partial W_{hh}} = \sum_t \left(\frac{\partial L_t}{\partial z_t}\frac{\partial z_t}{\partial net_t}\frac{\partial net_t}{\partial h_t} + \frac{\partial L_{t+1}}{\partial z_{t+1}}\frac{\partial z_{t+1}}{\partial net_t}\frac{\partial net_{t+1}}{\partial h_{t+1}}\frac{\partial net_{t+1}}{\partial h_t}\right)\frac{\partial h_t}{\partial hidden_t}\frac{\partial hidden_t}{\partial W_{hh}} \\
=& \sum_t \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \frac{-1}{Z_k} \\ \vdots \\ 0 \end{bmatrix}^\top \cdot \begin{bmatrix} \frac{\partial Z_{t,1}}{\partial Net_{t,1}} & \cdots & \frac{\partial Z_{t,1}}{\partial Net_{t,c}} \\ & \vdots & \\ \frac{\partial Z_{t,c}}{\partial Net_{t,1}} & \cdots & \frac{\partial Z_{t,c}}{\partial Net_{t,c}} \end{bmatrix} \cdot \begin{bmatrix} W_{hz,1} & W_{hz,2} & \cdots & W_{hz,c} \end{bmatrix} + \\
& \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \frac{-1}{Z_k} \\ \vdots \\ 0 \end{bmatrix}^\top \cdot \begin{bmatrix} \frac{\partial Z_{t+1,1}}{\partial Net_{t+1,1}} & \cdots & \frac{\partial Z_{t+1,1}}{\partial Net_{t+1,c}} \\ & \vdots & \\ \frac{\partial Z_{t+1,c}}{\partial Net_{t+1,1}} & \cdots & \frac{\partial Z_{t+1,c}}{\partial Net_{t+1,c}} \end{bmatrix} \cdot \begin{bmatrix} W_{hz,1} & W_{hz,2} & \cdots & W_{hz,c} \end{bmatrix} \cdot W_{hh} \cdot tanh'(hidden_t) \cdot h_{t-1} \\
=& \sum_t [(Z_t - y_t)^T W_{hz} + (Z_{t+1} - y_{t+1})^T W_{hz}^T W_{hh}] tanh'(hidden_t) \cdot h_{t-1}
\end{aligned} \tag{5}
$$

Problem 3

Solution

Subproblem (a)

○ Given (s, a, r, s'), we use the update equation:

$$
Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \quad max_{a' \in \{-1,1\}} \quad Q(s',a') - Q(s,a)) \tag{6}
$$

○ Using the equation with $\alpha = \frac{1}{2}, \gamma = \frac{1}{3}$, we have:

$$
\begin{aligned}
Q(3,-1) \leftarrow& 0 + \frac{1}{2}(-1 + \frac{1}{3} max_{a'} Q(2,a')) = -\frac{1}{2} \\
Q(2,-1) \leftarrow& 0 + \frac{1}{2}(-1 + \frac{1}{3} max_{a'} Q(3,a')) = -\frac{1}{2} \\
Q(3,1) \leftarrow& 0 + \frac{1}{2}(-1 + \frac{1}{3} max_{a'} Q(4,a')) = -\frac{1}{2}
\end{aligned} \tag{7}
$$

Subproblem (b)

○ We have $\nabla_w J(w)$ as follow:

$$
\begin{aligned}
\nabla_w J(w) =& -2(r + \gamma max_{a'} \hat{q}(s',a';w^-) - \hat{q}(s,a;w)) \nabla_w \hat{q}(s,a;w) \\
=& -2(r + \frac{1}{3} max_{a'} (w^-)^T \begin{bmatrix} s' \\ a' \\ 1 \end{bmatrix} - w^T \begin{bmatrix} s \\ a \\ 1 \end{bmatrix}) \begin{bmatrix} s \\ a \\ 1 \end{bmatrix}
\end{aligned} \tag{8}
$$

◦ using this, the parameter update with a single sample (s, a, r, s') is:

$$\begin{aligned} w' \to & w - \alpha \nabla_w J(w) \\ =& w + \frac{1}{2}(r + \frac{1}{3}max(w^-)^T \begin{bmatrix} s' \\ a' \\ 1 \end{bmatrix} - w^T \begin{bmatrix} s \\ a \\ 1 \end{bmatrix}) \begin{bmatrix} s \\ a \\ 1 \end{bmatrix} \end{aligned} \tag{9}$$

◦ Usiing the sample $(2, -1, -1, 1)$ and the particular values of w and w^- yields:

$$\begin{aligned} w' \to & \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} + \frac{1}{2}(-1 + \frac{1}{3}max_{a'} \begin{bmatrix} 1 \\ -1 \\ -2 \end{bmatrix}^T \begin{bmatrix} 1 \\ a' \\ 1 \end{bmatrix} - \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}^T \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}) \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} \\ = & \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} + \frac{1}{2}(-1 + \frac{1}{3}max_{a'}(1 - a' - 2) - (-2 - 1 + 1)) \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} \\ = & \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} \\ = & \begin{bmatrix} 0 \\ 1/2 \\ 3/2 \end{bmatrix} \end{aligned} \tag{10}$$

Problem 4

Solution

The picture of plotting of training curve and test accuracy under 4 subprobelms is shown as follow:

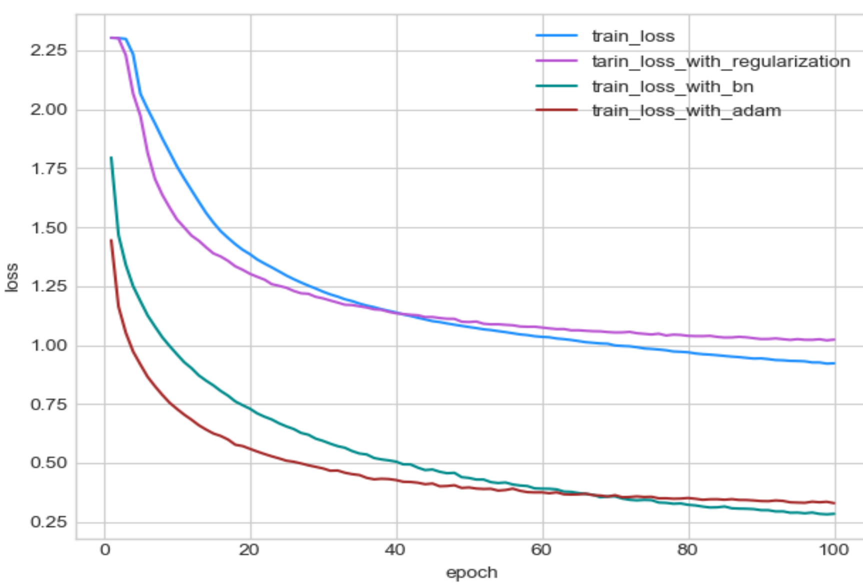

Fig 1: Loss vs Epoch

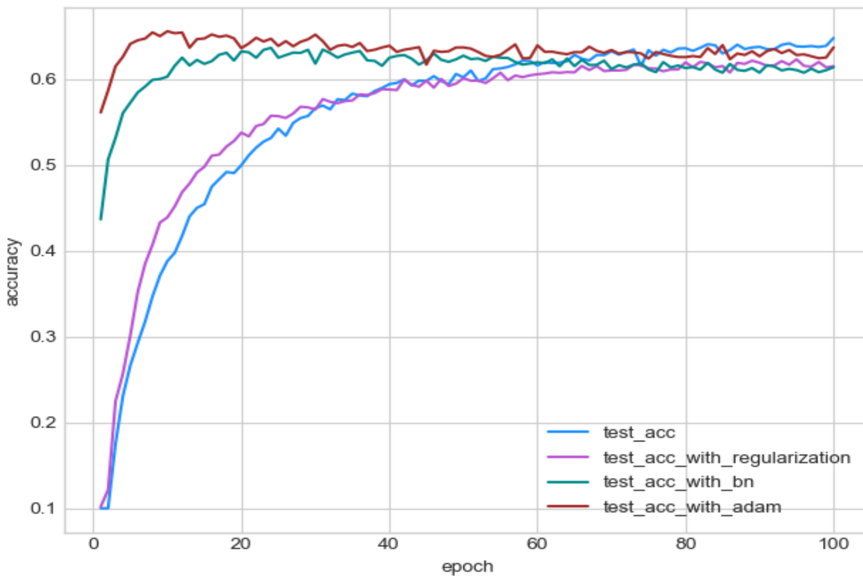

Fig 2: Accuracy vs Epoch

Subproblem (a)

After using softmax loss and regularization, we find that trian loss varies a little compared with original setting. However the test accuracy increase at the early time since it overcomes the overfitting that is brought in original setting. The visulization of first layer filter is as follow:

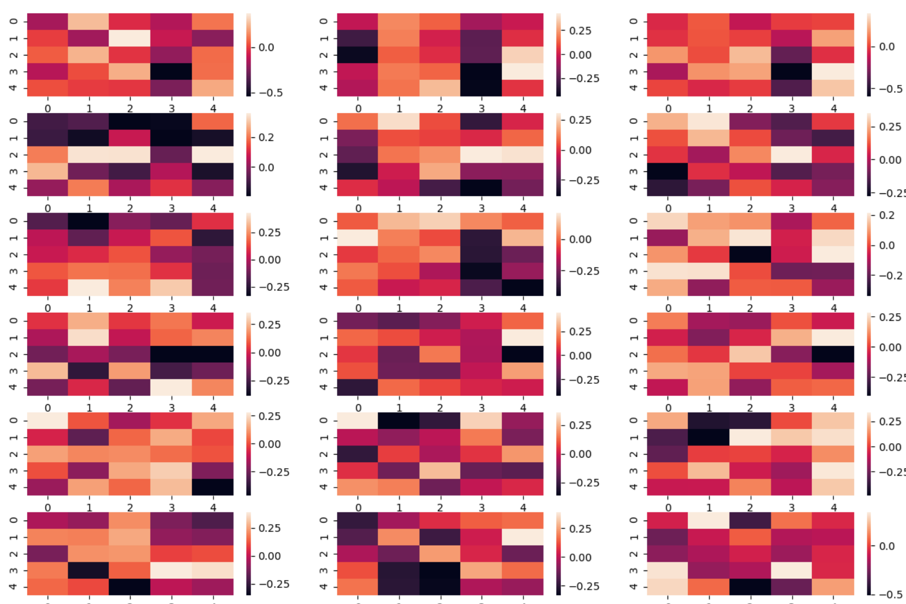

Fig 3: The filters learned in the first convolutional layer

Subproblem (b)

∘ We prove $Var[y_l] = n_l Var[w_l] E[x_l^2]$ as follow

$$
\begin{aligned}
Var[y_l] =& Var[W_{l,i} \cdot x_l] \\
=& n_l \cdot Var[w_l \cdot x_l] \\
=& n_l \cdot (E[(w_l \cdot x_l)^2] - (E[w_l \cdot x_l])^2) \\
=& n_l \cdot (E[(w_l \cdot x_l)^2] - (E[w_l] \cdot E[x_l])^2) \\
=& n_l \cdot E[(w_l \cdot x_l)^2] \\
=& n_l \cdot E[w_l^2] \cdot E[x_l^2] \\
=& n_l \cdot Var[w_l] \cdot E[x_l^2]
\end{aligned} \tag{11}
$$

∘

$$
P(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} & x = 0 \\ Q(x) & x > 0 \end{cases} \tag{12}
$$

$$
p(x) = q(x) \qquad (x > 0)
$$

Then, we can have

$$
\begin{aligned}
E[x_l^2] =& \int_{-\infty}^{\infty} x_l^2 \cdot p(x_l)\, dx \\
=& 0^2 \cdot \frac{1}{2} + \int_{0}^{\infty} x_l^2 \cdot p(x_l)\, dx \\
=& \int_{0}^{\infty} x_l^2 \cdot q(x_l)\, dx
\end{aligned} \tag{13}
$$

We can also have

$$
\begin{aligned}
\frac{1}{2} Var[y_l - 1] =& \frac{1}{2} E(y_{l-1}^2) \\
=& \frac{1}{2} \int_{-\infty}^{\infty} y_{l-1}^2 \cdot q(y)\, dy \\
=& \frac{1}{2} \cdot 2 \int_{0}^{\infty} y_{l-1}^2 \cdot q(y)\, dy \\
=& \int_{0}^{\infty} y_{l-1}^2 \cdot q(y)\, dy
\end{aligned} \tag{14}
$$

Subproblem (c)

From the figure, we see that training loss decrease quickly in the early stage after we add batch nomalization layer. That is becasue BN avoid covariance shift and gradient explosion and disappearance, which accelerate convergece of training loss. In the meanwhile, test accuracy also rise in the begining.

Subproblem (d)

In this part, I use Adam optimizer to investigate if it can solve current trouble. Since adam utilize momentum and adaptive size to upgrade in rach iteration, so it has great speed to decrease that we can see in the figure. At the begining of first 20 epochs, the model with Adam optimizer perform better on both training and testing.