Quiz 3 **Name:**
Week 3 Sept/23/2020 **On your left:**
CS 280: Fall 2020 **On your right:**
Instructor: Xuming He

Instructions:
Please answer the questions below. Show all your work. This is an open-book test. NO discussion/collaboration is allowed.

Problem 1. Convolution Kernel (10 points)
We have a video sequence and we would like to design a 3D convolutional neural network to recognize events in the video. The frame size is 32x32 and each video has 30 frames. Let's consider the first convolutional layer.

- We use a set of 3x3x3 convolutional kernels. Assume we have 64 kernels and apply stride 2 in spatial domain and temporal domain, what is the size of output feature map? Use proper padding if needed.

$T \times H \times W \quad 30 \times 32 \times 32 = m_0 \times m_1 \times m_2$

$k = 3 \times 3 \times 3 \quad S = 2 \times 2 \times 2$

$$O_i = \left\lfloor \frac{m_i + 2 \times P_i - d_i \times (k_i - 1) - 1}{S_i} + 1 \right\rfloor$$

the shape of output is $k \times O_0 \times O_1 \times O_2$

~~Take the zero padding~~ P_i: 2 ← clearly tell the padding.

$1 \times 30 \times 32 \times 32 \xrightarrow{\text{conv3d}} 64 \times 14 \times 15 \times 15$

$C_1 \times T_1 \times H_1 \times W_1 \qquad \underset{2}{\underline{C_2}} \times \underset{3}{\underline{T_2}} \times \underset{3}{\underline{H_2 \times W_2}}$

Any reasonable padding is OK.

Quiz 3
Week 3 Sept/23/2020
CS 280: Fall 2020
Instructor: Xuming He

Name:
On your left:
On your right:

Problem 2. Adam (10 points)
Explain why the bias correction is needed in the Adam update. Hint: You should derive the update rule for t=1.

① what is bias correction

w/o bias correction

$m_0, v_0 = 0, 0$

$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$

$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$

$\theta_t = \theta_{t-1} - \alpha \frac{m_t}{\sqrt{v_t} + \varepsilon}$

bias correction

$m_t \leftarrow \frac{m_t}{1-\beta_1^t}$

$v_t \leftarrow \frac{v_t}{1-\beta_2^t}$

5'

② for t=1

$m_1 = \beta_1 \cdot 0 + (1-\beta_1) g_1$ $\quad \hat{m}_1 = \frac{m_1}{1-\beta_1} = g_1$

$v_1 = \beta_1 \cdot 0 + (1-\beta_2) g_1^2$ $\quad \hat{v}_1 = \frac{v_1}{1-\beta_2} = g_1^2$

$\theta_1 = \theta_0 - \alpha \frac{\hat{m}_1}{\sqrt{v_1} + \varepsilon}$

Core idea: $t \to 0$, the estimate m_t (for g_t) and v_t (for g_t^2) is biased towards m_0 (and v_0). Bias correction corrects the bias by a factor of $\frac{1}{1-\beta^t}$

③ Explanation

5'

$m_t = \beta \cancel{m_{t-1}} + (1-\beta) g_t$

$\cancel{\beta m_{t-1}} = \beta^2 \cancel{m_{t-2}} + \beta(1-\beta) g_{t-1}$

$\vdots$

$\cancel{\beta^{t-1} m_1} = \beta^t m_0 + \beta^{t-1}(1-\beta) g_1$

$\therefore m_t = \beta^t m_0 + (1-\beta) \sum_{i=0}^{t-1} \beta^i g_{t-i}$

$E[m_t] = (1-\beta) \cdot E[g_t] \frac{1}{1-\beta} (1-\beta^t) + \beta^t \cdot E[m_0]$

$= (1-\beta^t) E[g_t] + \beta^t \cdot E[m_0]$

when $t \to \infty$,

$\because \beta \in (0,1)$, $\beta^t \to 0$

$\therefore E[m_t] \to 1 \cdot E[g_t] + 0$.

m_t is an unbiased estimate for g_t.

when $t \to 0$

$\beta^t \to 1$, $1-\beta^t \to 0$

$\therefore E[m_t] \to E[m_0]$

m_t is biased towards m_0.

However, let $\hat{m}_t = \frac{m_t}{1-\beta^t}$

$E[\hat{m}_t]$

$= E[g_t] + \frac{\beta^t}{1-\beta^t} \cdot E[m_0]$

If initialize with $m_0 = 0$.

$E[\hat{m}_t] = E[g_t]$ unbiased!