MathDNN HW 6

2018-13260 차재현

1 Problem 1

We assume $p < 1$, otherwise all cells are dropped, which is meaningless.

Let $y = (y_i)_i$ be result of linear network and σ be activation function. The condition linear-dropout-activation is equal to linear-activation-dropout is equivalent to

$$\sigma(0) = 0,$$

$$\sigma\left(\frac{y_i}{1-p}\right) = \frac{\sigma(y_i)}{1-p}$$

for all i.

1. $\sigma = \text{ReLU}$. Obviously $\sigma(0) = 0$. When $y_i < 0$, since ReLU vanishes at negative value, both are zero. Lastly, when $y_i \geq 0$, ReLU is just identity, so linear.

2. $\sigma = \text{sigmoid}$. In this case

$$\sigma\left(\frac{y_i}{1-p}\right) < \frac{\sigma(y_i)}{1-p}$$

   for all large y_i. More specifically, for all $\sigma(y_i) > 1 - p$.

3. $\sigma = \text{LeakyReLU}$. Compared to ReLU, only the case $y_i < 0$ is different. In this case, we get

$$\sigma\left(\frac{y_i}{1-p}\right) = \frac{cy_i}{1-p} = \frac{\sigma(y_i)}{1-p}$$

   by $1 - p > 0$.

Therefore ReLU, LeakyReLU guarantee equivalence.

2 Problem 2

Suppose we define a single linear network $y = Ax + b$ without activation for $x \in \mathbb{R}^n, y \in \mathbb{R}^m$. According to Pytorch documentation, every entry of A, b sampled from independent Uniform distribution in $(-1/\sqrt{n}, 1/\sqrt{n})$. Its probability density function is given by

$$p(x) = \frac{\sqrt{n}}{2}\chi_{[-1/\sqrt{n}, 1/\sqrt{n}]},$$

where χ is characteristic function, and its statistical value mean(μ), variance(σ) becomes

$$\mu = \int_{\mathbb{R}} xp(x)dx = 0,$$

$$\sigma = \int_{\mathbb{R}} x^2 p(x)dx = \frac{1}{3n}.$$

Now, assume x have been chosen randomly so that $\mathbb{E}(x) = \mathbf{0}$ and $\mathrm{Cov}(x) = \Sigma$, where Σ is diagonal matrix. Then, from

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} = \begin{pmatrix} \sum_{1\le k\le n} a_{1k}x_k + b_1 \\ \vdots \\ \sum_{1\le k\le n} a_{mk}x_k + b_m \end{pmatrix},$$

for all $1 \le i, j \le m$ we have expectation

$$\mathbb{E}(y_i) = \sum_k \mathbb{E}(a_{ik}x_k + b_i) = 0$$

and covariance

$$\begin{aligned} \mathrm{Cov}(y_i, y_j) &= \mathbb{E}((\sum_k a_{ik}x_k + b_i)\cdot(\sum_k a_{jk}x_k + b_j)) \\ &= \mathbb{E}(\sum_k a_{ik}x_k a_{jk}x_k) + \mathbb{E}(b_j^2), \end{aligned}$$

which is $\mathbb{E}(b_j^2) = \frac{1}{3n}$ unless $i = j$, and if so

$$\begin{aligned} \mathbb{E}(\sum_k a_{ik}x_k a_{jk}x_k) + \mathbb{E}(b_j^2) &= \mathbb{E}(\sum_k (a_{ik}x_k)^2) + \mathbb{E}(b_j^2) \\ &= \sum_k \mathbb{E}(a_{ik}^2)\mathbb{E}(x_k^2) + \mathbb{E}(b_j^2) = \frac{\sum_k \Sigma_{kk} + 1}{3n}. \end{aligned}$$

Applying these facts to problem. Since we are assuming that $x_1, \cdots, x_{n_0}$ are IID with zero-mean and unit variance, we take $\Sigma = I_{n_0}$. Then $\mathbb{E}(y_1) = \mathbf{0}$ and $\mathrm{Cov}(y_1) = \frac{n_0+1}{3n_0} I_{n_1}$. Induction gives $\mathbb{E}(y_L) = 0$ and $\mathrm{Cov}(y_L) = \prod_{0\le l<L} \frac{n_l+1}{3n_l}$.

3 Problem 3

We refer to problem 6 in homework 4 for detailed calculation.

(i) Simply adding derivative of residual term, we get

$$\frac{\partial y_l}{\partial y_{l-1}} = \mathrm{diag}(\sigma'(A_l y_{l-1} + b_l))A_l + I_m$$

when $l < L$. Here I_m is $m \times m$ identity matrix. When $i = L$, since $y_L = A_L y_{L-1} + b_L$ is linear after ignoring b_L, derivative is just A_L.

(ii) First consider the case, differentiating over b_l. Chain rule gives

$$\begin{aligned}\frac{\partial y_L}{\partial b_l} &= \left(\prod_{l<i\leq L} \frac{\partial y_i}{\partial y_{i-1}}\right) \frac{\partial y_l}{\partial b_l} \\ &= A_L \left(\prod_{l<i<L} (\mathrm{diag}(\sigma'(A_i y_{i-1} + b_i))A_i + I_m)\right) \frac{\partial y_l}{\partial b_l}.\end{aligned}$$

Now since such b_l is independent of y_{l-1}, our calculation finished after replacing $\frac{\partial y_l}{\partial b_l}$ to $\mathrm{diag}(\sigma'(A_l y_{l-1} + b_l))$.

To examine the next case, we start from

$$\frac{\partial y_L}{\partial A_l} = \mathrm{diag}\sigma'(A_l y_{l-1} + b_l) \left(\frac{\partial y_L}{\partial y_l}\right)^\top y_{l-1}^\top.$$

As in the previous case, we replace middle term or the right side by

$$\left(\frac{\partial y_L}{\partial y_l}\right)^\top = \left(\prod_{l<i\leq L} \frac{\partial y_i}{\partial y_{i-1}}\right)^\top,$$

which give

$$\begin{aligned}\frac{\partial y_L}{\partial A_l} =& \mathrm{diag}\sigma'(A_l y_{l-1} + b_l) \times \\ & \left\{A_L \left(\prod_{l<i<L} (\mathrm{diag}(\sigma'(A_i y_{i-1} + b_i))A_i + I_m)\right)\right\}^\top y_{l-1}^\top.\end{aligned}$$

(iii) In (ii), terms $A_j, \sigma'(A_j y_{j-1} + b_j)$ appears simultaneously at multiplicands. Even if at least of them becomes zero, since I_m is added at the right side, its diagonal need not be vanished, i.e., $\mathrm{diag}(\sigma'(A_j y_{j-1} + b_j))A_j + I_m \neq \mathbf{0}$.

4 Problem 4

We refer to problem 5 in homework 5.

Write first and second network as concatenated network and splitted network.

(a) When CNN with (input channel, output channel, width, height) $= (n, m, a, b)$, the number of parameter is $nmab$.

- Concatenated network has

  $$256 \times 128 \times 1 \times 1 + 128 \times 128 \times 3 \times 3 + 128 \times 256 \times 1 \times 1 = 212,992$$

  parameters.

- Splitted network has

  $$(256 \times 4 \times 1 \times 1 + 4 \times 4 \times 3 \times 3 + 4 \times 256 \times 1 \times 1) \times 32 = 70,144$$

  parameters.

(b) At Problem_4_b.pdf file