Problem 11.1

$$\begin{aligned}
y &= h_3 + f_4[h_3(x)] \\
&= h_2 + f_3[h_2(x)] + f_4[h_2 + f_3[h_2(x)]] \\
&= h_1 + f_2[h_1(x)] + f_3[h_1 + f_2[h_1(x)]] \\
&\quad + f_4[h_1 + f_2[h_1(x)] + f_3[h_1 + f_2[h_1(x)]]] \\
&= x + f_1(x) \\
&\quad + f_2(x + f_1(x)) \\
&\quad + f_3(x + f_1(x) + f_2(x + f_1(x))) \\
&\quad + f_4(x + f_1(x) + f_2(x + f_1(x)) + f_3(x + f_1(x) + f_2(x + f_1(x))))
\end{aligned}$$

Problem 11.2

(i) # of path with length 0/1/2/3 is $\binom{3}{0}$, $\binom{3}{1}$, $\binom{3}{2}$, $\binom{3}{3}$ respectively

(ii) $\binom{5}{0}$, $\binom{5}{1}$, $\binom{5}{2}$, $\binom{5}{3}$, $\binom{5}{4}$, $\binom{5}{5}$ respectively.

(iii) # of path with length l of k residual block is $\binom{k}{l}$

Problem 11.3

$$\begin{aligned}
\frac{\partial y}{\partial f_1} &= 1 + \frac{\partial f_2}{\partial f_1} + \frac{\partial f_3}{\partial f_1}\left(1 + \frac{\partial f_2}{\partial f_1}\right) + \frac{\partial f_4}{\partial f_1}\left(1 + \frac{\partial f_3}{\partial f_1} + \frac{\partial f_3}{\partial f_1}\left(1 + \frac{\partial f_2}{\partial f_1}\right)\right) \\
&= 1 + \frac{\partial f_2}{\partial f_1} + \left(\frac{\partial f_3}{\partial f_1} + \frac{\partial f_3}{\partial f_1}\frac{\partial f_2}{\partial f_1}\right) + \left(\frac{\partial f_4}{\partial f_1} + \frac{\partial f_4}{\partial f_1}\frac{\partial f_2}{\partial f_1} + \frac{\partial f_4}{\partial f_1}\frac{\partial f_3}{\partial f_1} + \frac{\partial f_4}{\partial f_1}\frac{\partial f_3}{\partial f_1}\frac{\partial f_2}{\partial f_1}\right)
\end{aligned}$$

Problem 11.4
We consider the first residual layer first.
The first path is $p_1 = f_1[\phi x + \beta]$
The second path is $p_2 = x$

$$\begin{aligned}
\mathbb{E}[p_1 p_2] &= \mathbb{E}[x f_1[\phi x + \beta]] = x\mathbb{E}[f_1[\phi x + \beta]] \\
&= x f_1[\mathbb{E}(\phi)x + \mathbb{E}(\beta)] = 0
\end{aligned}$$

where we use $\phi \sim \mathcal{N}(0, \sigma^2), \beta \sim \mathcal{N}(0, \sigma_\beta^2), f(\cdot) is RELU$
Similarly we get $\mathbb{E}[p_1] = 0$ and $\mathbb{E}[p_2] = x$
$\text{Cov}[p_1, p_2] = \mathbb{E}[p_1 p_2] - \mathbb{E}[p_1]\mathbb{E}[p_2] = 0 \Rightarrow p_1, p_2$ is uncorrelated.
(we can use similar idea to prove other path by induction)

The variance of XxY of variable x, y is given by:

$$\begin{aligned}
\text{Var}[xy] &= \mathbb{E}[(xy - \mathbb{E}(xy))^2] \\
&= \mathbb{E}[((x - \mathbb{E}(x)) + (y - \mathbb{E}(y)))^2] \\
&= \mathbb{E}[(x - \mathbb{E}(x))^2] + \mathbb{E}[(y - \mathbb{E}(y))^2] + 2\mathbb{E}[(x - \mathbb{E}(x))(y - \mathbb{E}(y))] \\
&= \text{Var}[x] + \text{Var}[y] + \text{Cov}[x, y]
\end{aligned}$$

if x, y are uncorrelated, $\mathrm{Cov}[x, y] = 0$
Problem 11.5

$$\begin{aligned}
\frac{\partial z_i^{'}}{\partial z_i} &= \gamma \frac{\partial f_{7i}}{\partial z_i}, \\
\frac{\partial f_{7i}}{\partial z_i} &= \frac{\partial f_{2i}}{\partial z_i} f_6 + f_{2i} \frac{\partial f_6}{\partial z_i}, \\
\frac{\partial f_{2i}}{\partial z_i} &= 1 \\
\frac{\partial f_6}{\partial z_i} &= -\frac{1}{f_5^2} \frac{\partial f_5}{\partial z_i}, \\
\frac{\partial f_5}{\partial z_i} &= \frac{1}{2\sqrt{f_4 + \epsilon}} \frac{\partial f_4}{\partial z_i} \\
f_4 = E[f_{3i}] = \frac{1}{I} \sum_1^I f_{2j}^2 \Rightarrow \frac{\partial f_4}{\partial z_i} &= \frac{1}{I} \sum_1^I \frac{\partial f_{2i}^2}{\partial z_i} = \frac{2}{I} f_{2i} \frac{\partial f_{2i}}{\partial z_i} \\
\frac{\partial f_{2i}}{\partial z_i} &= 1 - \frac{\partial f_1}{\partial z_i}, \\
\frac{\partial f_1}{\partial z_i} &= \frac{1}{I} \sum_1^I \frac{\partial z_j}{\partial z_i} = \frac{1}{I} \\
\Rightarrow \frac{\partial z_i'}{\partial z_i} &= \gamma \left\{ \frac{1}{\sqrt{\sigma^2 + \epsilon}} - \frac{(z_i - \mu)^2}{(\sigma^2 + \epsilon)^{\frac{3}{2}}} \frac{I(I-1)}{I^2} \right\}
\end{aligned}$$

Problem 11.6

# of parameter = 20 * (1+1) + 20 * (20+1) * 9 + (20+1) = 3841
input → hidden hidden hidden → output
# of parameter due to BN = (1+1) * 9 = 18
# of parameter in total = 3841 + 18 = 3859

Problem 11.7

L2 Regularization tends to make weight matrix W to be small by adding penalty $(\lambda \|W^T W\|)$, too small W will cause input of activation function to be small, but with BN, the input of activation function is always $r\tilde{z} + \beta$ with $\tilde{z} \sim N(0, e^2)$ this will cause W to be extremely small therefore causing numerical problem.

Problem 11.8

(1)

- 1. BN: $2 \times 512 = 1024$

- 2. ReLU: 0 since ReLU is non-parameter
- 3. 3x3 Conv: $(3 \times 3 \times 512 + 1) \times 512$
- In total: 2360832

(2)

- Block1:
  - BN: $2 \times 512 = 1024$
  - 1x1 Conv: $(1 \times 1 \times 512 + 1) \times 128$
  - In total: 66688
- Block 2:
  - BN: $2 \times 128 = 256$
  - 3x3 Conv: $(3 \times 3 \times 128 + 1) \times 128$
  - In total: 147840
- Block 3:
  - BN: $2 \times 128 = 256$
  - 1x1 Conv: $(1 \times 1 \times 128 + 1) \times 512$
  - In total: 66304
- In total: 280832 (much less than question(1))

Problem 11.9

Picture with very large size will need too much parameter while very small picture may not good for training. Besides suitable size is easier to maintain consistent layer shapes.