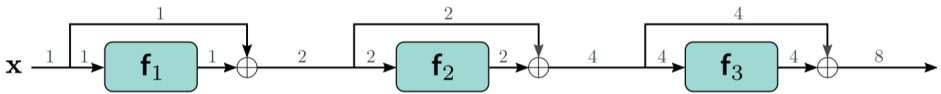
Residual Networks (Part 3)

Dr. Alireza Aghamohammadi

Motivation

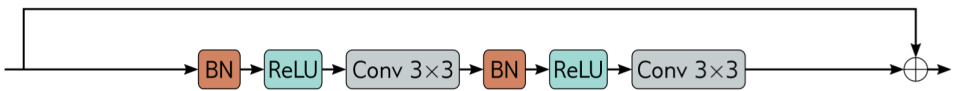
- Despite utilizing He initialization in the residual block, the values tend to grow exponentially during the forward pass through the network[1].



[1] Reference: "Understanding Deep Learning"

Batch Normalization

- *Batch normalization*, also known as *BatchNorm*, adjusts each activation h by shifting and rescaling it.
- This process ensures that the mean and variance across the batch $\mathcal{B}$ are learned values during the training phase.
- Typically, batch normalization is applied prior to the activation function[2].



[2]Reference: "Understanding Deep Learning"

Algorithm

- **Input:** The algorithm takes as input the values of h_i for a mini-batch $\mathcal{B}$, along with the parameters γ and δ that are to be learned during the process.
- **Output:** The output of the algorithm is the transformed values of h_i.

$$
\begin{aligned}
&1. \quad m_h = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} h_i \\
&2. \quad s_h = \sqrt{\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} (h_i - m_h)^2} \\
&3. \quad h_i = \frac{h_i - m_h}{s_h + \epsilon} \quad \forall i \in \mathcal{B} \\
&4. \quad h_i = \gamma h_i + \delta \quad \forall i \in \mathcal{B}
\end{aligned}
$$

- After performing this operation, the activations for each member of the batch have a mean of δ and a standard deviation of γ. Both these values are learned during the training process.
- Each hidden unit undergoes batch normalization independently.