CSE 543 - Deep Learning

Lecture 6: Kaiming Initialization

April 26, 2023

Lecturer: Simon Du *Scribe: Ruoqi Shen*

1 Kaiming Initialization

Consider the neural network with ReLU activation:

$$f(x, W^{(0)}, ..., W^{(H-1)}, b^{(0)}, ..., b^{(H-1)}) = W^{(H-1)}\sigma(W^{(H-2)}...\sigma(W^{(0)}x + b^{(0)})... + b^{(H-2)}) + b^{(H-1)}$$

where $x \in \mathbb{R}^{d_0}$, $b^{(h)} \in \mathbb{R}^{d_{h+1}}$ and $W^{(h)} \in \mathbb{R}^{d_{h+1} \times d_h}$. Let $x^{(0)} = x$, and for $h \in \{0, ..., H-1\}$, $z^{(h)} = W^{(h)}x^{(h)} + b^{(h)} \in \mathbb{R}^{d_{h+1}}$ and $x^{(h+1)} = \sigma(z^{(h)})$.

Kaiming Initialization [1] For any $h \in \{0, ..., H_1\}$, initialize $b^{(h)} = 0$ and $W_{ij}^{(h)} \sim \mathcal{N}(0, \frac{2}{d_h})$ for $i \in [d_{h+1}]$ and $j \in [d_h]$.

We can show that under Kaiming initialization, the variance of each layer stays the same.

Proposition 1.1. *For any $h \in \{1, ..., H-1\}$, $i \in [d_{h+1}]$ and $j \in [d_h]$, $\mathbf{Var}(z_i^{(h)}) = \mathbf{Var}(z_j^{(h-1)})$*

Proof. For $W_{ij}^{(h)} \sim \mathcal{N}(0, \frac{2}{d_h})$, $\mathbb{E}[W_{ij}^{(h)}] = 0$. Then, for $z^{(h)} = W^{(h)}x^{(h)}$, $\mathbb{E}[z^{(h)}] = 0$. Since $z_i^{(h)} = \sum_{j'=1}^{d_h} W_{ij'}^{(h)} x_{j'}^{(h)}$, the variance of $z_i^{(h)}$ is given by

$$\begin{aligned}
\mathbf{Var}(z_i^{(h)}) &= d_h \cdot \mathbf{Var}(W_{ij}^{(h)} x_j^{(h)}) \\
&= d_h \cdot \Big(\mathbf{Var}(W_{ij}^{(h)}) \cdot \mathbf{Var}(x_j^{(h)}) + (\mathbb{E}W_{ij}^{(h)})^2 \mathbf{Var}(x_j^{(h)}) + \mathbf{Var}(W_{ij}^{(h)}) \cdot (\mathbb{E}x_j^{(h)})^2\Big) \\
&= d_h \cdot \mathbf{Var}(W_{ij}^{(h)}) \cdot \mathbb{E}[(x_j^{(h)})^2].
\end{aligned}$$

The first steps follows from $W_{ij}^{(h)} x_j^{(h)}$ for $j \in [d_h]$ are i.i.d. The second and the third step follow

from $\mathbb{E}[W_{ij}^{(h)}] = 0$ and $W_{ij}^{(h)}$ and $x_j^{(h)}$ are independent. Next, we can compute $\mathbb{E}[(x_j^{(h)})]^2$,

$$\begin{aligned}
\mathbb{E}[(x_j^{(h)})]^2 &= \int_{-\infty}^{\infty} (x_j^{(h)})^2 \mathcal{P}_x(x_j^{(h)}) dx_j^{(h)} \\
&= \int_{-\infty}^{\infty} (\max\{0, z_j^{(h-1)}\})^2 \mathcal{P}_x(z_j^{(h-1)}) dz_j^{(h-1)} \\
&= \int_{0}^{\infty} (z_j^{(h-1)})^2 \mathcal{P}_z(z_j^{(h-1)}) dz_j^{(h-1)} \\
&= \frac{1}{2} \int_{-\infty}^{\infty} (z_j^{(h-1)})^2 \mathcal{P}_z(z_j^{(h-1)}) dz_j^{(h-1)} \\
&= \frac{1}{2} \mathbf{Var}(z_j^{(h-1)}),
\end{aligned}$$

where $\mathcal{P}_x(\cdot)$ and $\mathcal{P}_z(\cdot)$ are the density functions of $x_j^{(h)}$ and $z_j^{(h-1)}$. The fourth equation follows from the symmetry of the distribution. Then,

$$\begin{aligned}
\mathbf{Var}(z_i^{(h)}) &= d_h \cdot \mathbf{Var}(W_{ij}^{(h)}) \cdot \mathbb{E}[(x_j^{(h)})^2] \\
&= d_h \cdot \mathbf{Var}(W_{ij}^{(h)}) \cdot \frac{1}{2} \mathbf{Var}(z_j^{(h-1)}).
\end{aligned}$$

Since $\mathbf{Var}(W_{ij}^{(h)}) = \frac{2}{d_h}$, $\mathbf{Var}(z_i^{(h)}) = d_h \cdot \frac{2}{d_h} \cdot \frac{1}{2} \cdot \mathbf{Var}(z_j^{(h-1)}) = \mathbf{Var}(z_j^{(h-1)})$. □

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.