Notes on Backpropagation of NN models

Zhenyu Liao

Huazhong University of Science & Technology
School of Electronic Information and Communications

September 29, 2021

Backpropagation: basic setting

Consider a single-hidden-layer neural network with N neurons and tanh activation, the input $\mathbf{x} \in \mathbb{R}^p$ is a (column) data vector of dimension p,

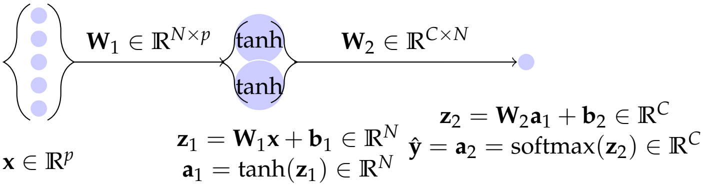

Figure: Illustration of a single-hidden-layer (FC) neural network model.

We have the following system of equations (all vectors are considered **column** vectors):

$$\mathbf{z}_1 = \mathbf{W}_1\mathbf{x} + \mathbf{b}_1 \in \mathbb{R}^N \tag{1}$$

$$\mathbf{a}_1 = \tanh(\mathbf{z}_1) \in \mathbb{R}^N \tag{2}$$

$$\mathbf{z}_2 = \mathbf{W}_2\mathbf{a}_1 + \mathbf{b}_2 \in \mathbb{R}^C \tag{3}$$

$$\hat{\mathbf{y}} = \mathbf{a}_2 = \operatorname{softmax}(\mathbf{z}_2) \in \mathbb{R}^C \tag{4}$$

Backpropagation: part 1

$$
\begin{aligned}
\mathbf{z}_1 &= \mathbf{W}_1\mathbf{x} + \mathbf{b}_1 \in \mathbb{R}^N \\
\mathbf{a}_1 &= \tanh(\mathbf{z}_1) \in \mathbb{R}^N \\
\mathbf{z}_2 &= \mathbf{W}_2\mathbf{a}_1 + \mathbf{b}_2 \in \mathbb{R}^C \\
\hat{\mathbf{y}} &= \mathbf{a}_2 = \operatorname{softmax}(\mathbf{z}_2) \in \mathbb{R}^C
\end{aligned}
$$

with the i-th entry of $\hat{\mathbf{y}}$ given by

$$
[\hat{\mathbf{y}}]_i = e^{[\mathbf{z}_2]_i} / \Big(\sum_{k=1}^C e^{[\mathbf{z}_2]_k}\Big) = e^{[\mathbf{z}_2]_i} / (\mathbf{1}_C^\mathsf{T} e^{\mathbf{z}_2}) \tag{5}
$$

for $\mathbf{1}_C \in \mathbb{R}^C$ the column vector of all ones.

Consider the cross-entropy loss $L : \mathbb{R}^C \times \mathbb{R}^C \mapsto \mathbb{R}$ defined as

$$
L = L(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{k=1}^C [\mathbf{y}]_k \log[\hat{\mathbf{y}}]_k = -\sum_{k=1}^C [\mathbf{y}]_k \log \frac{e^{[\mathbf{z}_2]_k}}{\mathbf{1}_C^\mathsf{T} e^{\mathbf{z}_2}} \tag{6}
$$

for *one-hot* label vector $\mathbf{y} \in \{0,1\}^C$, we have

$$
\begin{aligned}
\frac{\partial L}{\partial [\mathbf{z}_2]_i} &= \sum_{k=1}^C \frac{\partial L}{\partial [\hat{\mathbf{y}}]_k} \frac{\partial [\hat{\mathbf{y}}]_k}{\partial [\mathbf{z}_2]_i} = -\sum_{k=1}^C \frac{[\mathbf{y}]_k}{[\hat{\mathbf{y}}]_k} \frac{\partial [\hat{\mathbf{y}}]_k}{\partial [\mathbf{z}_2]_i} &(7)\\
&= -\frac{[\mathbf{y}]_i}{[\hat{\mathbf{y}}]_i} \frac{\partial [\hat{\mathbf{y}}]_i}{\partial [\mathbf{z}_2]_i} - \sum_{k\neq i}^C \frac{[\mathbf{y}]_k}{[\hat{\mathbf{y}}]_k} \frac{\partial [\hat{\mathbf{y}}]_k}{\partial [\mathbf{z}_2]_i} &(8)\\
&= \frac{[\mathbf{y}]_i}{[\hat{\mathbf{y}}]_i} [\hat{\mathbf{y}}]_i([\hat{\mathbf{y}}]_i - 1) + \sum_{k\neq i}^C \frac{[\mathbf{y}]_k}{[\hat{\mathbf{y}}]_k} [\hat{\mathbf{y}}]_i [\hat{\mathbf{y}}]_k &(9)\\
&= [\hat{\mathbf{y}}]_i - [\mathbf{y}]_i &(10)
\end{aligned}
$$

Backpropagation: part 2

$$\begin{aligned} \mathbf{z}_1 &= \mathbf{W}_1 \mathbf{x} + \mathbf{b}_1 \in \mathbb{R}^N \\ \mathbf{a}_1 &= \tanh(\mathbf{z}_1) \in \mathbb{R}^N \\ \mathbf{z}_2 &= \mathbf{W}_2 \mathbf{a}_1 + \mathbf{b}_2 \in \mathbb{R}^C \\ \hat{\mathbf{y}} &= \mathbf{a}_2 = \operatorname{softmax}(\mathbf{z}_2) \in \mathbb{R}^C \end{aligned}$$

with the i-th entry of $\hat{\mathbf{y}}$ given by

$$[\hat{\mathbf{y}}]_i = e^{[\mathbf{z}_2]_i} / (\sum_{k=1}^C e^{[\mathbf{z}_2]_k}) = e^{[\mathbf{z}_2]_i} / (\mathbf{1}_C^\mathsf{T} e^{\mathbf{z}_2})$$

and cross-entropy loss $L : \mathbb{R}^C \times \mathbb{R}^C \mapsto \mathbb{R}$

$$L = L(\hat{\mathbf{y}}, \mathbf{y}) = - \sum_{k=1}^C [\mathbf{y}]_k \log [\hat{\mathbf{y}}]_k$$

for *one-hot* label vector $\mathbf{y} \in \{0,1\}^C$.

Backpropagation with the chain rule:

$$\begin{aligned} \frac{\partial L}{\partial [\mathbf{W}_1]_{ij}} &= \sum_{k=1}^C \frac{\partial L}{\partial [\mathbf{z}_2]_k} \frac{\partial [\mathbf{z}_2]_k}{\partial [\mathbf{a}_1]_i} \frac{\partial [\mathbf{a}_1]_i}{\partial [\mathbf{z}_1]_i} \frac{\partial [\mathbf{z}_1]_i}{\partial [\mathbf{W}_1]_{ij}} && (11) \\ &= \sum_{k=1}^C [\hat{\mathbf{y}} - \mathbf{y}]_k \cdot [\mathbf{W}_2]_{ki} \cdot [1 - \tanh^2(\mathbf{z}_1)]_i \cdot [\mathbf{x}]_j && (12) \\ &= [\mathbf{W}_2^\mathsf{T} (\hat{\mathbf{y}} - \mathbf{y})]_i \cdot [1 - \tanh^2(\mathbf{z}_1)]_i \cdot [\mathbf{x}]_j && (13) \end{aligned}$$

where we used the fact that

$$[\mathbf{z}_1]_i = [\mathbf{W}_1 \mathbf{x}]_i + [\mathbf{a}_1]_i = \sum_{j=1}^p [\mathbf{W}_1]_{ij} [\mathbf{x}]_j + [\mathbf{a}_1]_i. \tag{14}$$

Putting in matrix form

$$\frac{\partial L}{\partial \mathbf{W}_1} = \left(\left(\mathbf{W}_2^\mathsf{T} (\hat{\mathbf{y}} - \mathbf{y}) \right) \circ \left(1 - \tanh^2(\mathbf{z}_1) \right) \right) \mathbf{x}^\mathsf{T} \tag{15}$$

$$\mathbf{z}_1 = \mathbf{W}_1\mathbf{x} + \mathbf{b}_1 \in \mathbb{R}^N$$
$$\mathbf{a}_1 = \tanh(\mathbf{z}_1) \in \mathbb{R}^N$$
$$\mathbf{z}_2 = \mathbf{W}_2\mathbf{a}_1 + \mathbf{b}_2 \in \mathbb{R}^C$$
$$\hat{\mathbf{y}} = \mathbf{a}_2 = \mathrm{softmax}(\mathbf{z}_2) \in \mathbb{R}^C$$

with the i-th entry of $\hat{\mathbf{y}}$ given by

$$[\hat{\mathbf{y}}]_i = e^{[\mathbf{z}_2]_i} / (\sum_{k=1}^{C} e^{[\mathbf{z}_2]_k}) = e^{[\mathbf{z}_2]_i} / (\mathbf{1}_C^\mathsf{T} e^{\mathbf{z}_2})$$

and cross-entropy loss $L : \mathbb{R}^C \times \mathbb{R}^C \mapsto \mathbb{R}$

$$L = L(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{k=1}^{C} [\mathbf{y}]_k \log[\hat{\mathbf{y}}]_k$$

for *one-hot* label vector $\mathbf{y} \in \{0,1\}^C$.

Backpropagation with the chain rule:

$$\frac{\partial L}{\partial \mathbf{W}_2} = \delta_1 \mathbf{a}_1^\mathsf{T} \in \mathbb{R}^{C \times N} \tag{16}$$
$$\frac{\partial L}{\partial \mathbf{b}_2} = \delta_1 \in \mathbb{R}^{C} \tag{17}$$
$$\frac{\partial L}{\partial \mathbf{W}_1} = \delta_2 \mathbf{x}^\mathsf{T} \in \mathbb{R}^{N \times p} \tag{18}$$
$$\frac{\partial L}{\partial \mathbf{b}_1} = \delta_2 \in \mathbb{R}^{N} \tag{19}$$

for

$$\delta_1 = \hat{\mathbf{y}} - \mathbf{y} \in \mathbb{R}^C, \tag{20}$$
$$\delta_2 = (1 - \tanh^2(\mathbf{z}_1)) \circ (\mathbf{W}_2^\mathsf{T} \delta_1) \in \mathbb{R}^N \tag{21}$$