50.039 Theory and Practice of Deep Learning
Theory Homework 3

Joel Huang 1002530

February 20, 2019

1 Cross Entropy Loss Gradient

Let $h(x_i) = s(w \cdot x_i)$. Then

$$L = (-1) \cdot \sum_{i=1}^{n} y_i \log(s(w \cdot x_i)) + (1 - y_i) \log(1 - s(w \cdot x_i))$$

$$\nabla_w L = \nabla_w \left((-1) \sum_{i=1}^{n} y_i \log(s(w \cdot x_i)) + (1 - y_i) \log(1 - s(w \cdot x_i)) \right)$$

Applying chain rule,

$$\nabla_w L = (-1) \sum_{i=1}^{n} y_i \left(\left(\frac{\partial}{\partial w} (\log(s(w \cdot x_i))) \cdot \frac{\partial}{\partial w} (w \cdot x_i) \right) + \left((1 - y_i) \frac{\partial}{\partial w} (\log(1 - s(w \cdot x_i))) \cdot \frac{\partial}{\partial w} (w \cdot x_i) \right) \right)$$

Using the relationship $\dfrac{\partial \log(s(w \cdot x_i))}{\partial w} = 1 - s(w \cdot x_i)$ and $\dfrac{\partial \log(1 - s(w \cdot x_i))}{\partial w} = -s(w \cdot x_i)$,

$$\nabla_w L = (-1) \sum_{i=1}^{n} (y_i(1 - s(w \cdot x_i))(x_i) + (1 - y_i)(-s(w \cdot x_i))(x_i))$$

$$\nabla_w L = (-1) \sum_{i=1}^{n} ((x_i)(y_i - s(w \cdot x_i)(y_i)) + (x_i)(-s(w \cdot x_i) + s(w \cdot x_i)(y_i)))$$

$$\nabla_w L = (-1) \sum_{i=1}^{n} x_i(y_i - s(w \cdot x_i)(y_i) - s(w \cdot x_i) + s(w \cdot x_i)(y_i))$$

Finally,

$$\nabla_w L = \sum_{i=1}^{n} x_i(s(w \cdot x_i) - y_i) = \sum_{i=1}^{n} x_i(h(x_i) - y_i)$$

2 Einsum notation

Matrix-vector multiplication

$$C_{j,k} = \sum_i A_{ijk} b_i$$

Einsum: $ijk, i \to jk, [A, b]$

$$C_j = \sum_{i,k} A_{ijk} b_{ik}$$

Einsum: $ijk, ik \to j, [A, b]$

Sum over dimensions

$$A_{ik} = \sum_{j,l} A_{ijkl}$$

Einsum: $ijkl \to ik, [A]$

$$A_{ki} = \sum_{j,l} A_{ijkl}$$

Einsum: $ijkl \to ki, [A]$

$$C_i = \sum_{j,k} A_{ijk} A_{ijk}$$

Einsum: $ijk, ijk \to i, [A, A]$

$$C = x^\top A x$$

Einsum: $i, ij, j \to, [x, A, x]$

$$C = AG^\top B, A \in \mathbb{R}^{d \times e}, G \in \mathbb{R}^{f \times e}, B \in \mathbb{R}^{f \times l}$$

Einsum: $ij, kj, kl \to il, [A, G, B]$

$$C_{????} = \sum_{cd} A_{abcd} B_{bcde} E_{cdef}$$

Einsum: $abcd, bcde, cdef \to abef, [A, B, E]$

3 Tensor broadcasting

1. $(3, 1, 2, 3)$ and $(5, 3)$ are not broadcastable.
   - Fill smaller tensor from the left: $(1, 1, 5, 3)$
   - The sizes in the third dimension (2 and 5) are incompatible.
2. $(3, 2, 1, 3, 4)$ and $(5, 3, 4)$ are broadcastable.
   - Fill smaller tensor from the left: $(1, 1, 5, 3, 4)$
   - The sizes in the third dimension (1 and 5) are compatible.
   - $(1, 1, 5, 3, 4)$ is copied till its shape is $(3, 2, 5, 3, 4)$.
3. $(3, 2, 1, 3, 4)$ and $(5, 1, 4)$ are broadcastable.
   - Fill smaller tensor from the left: $(1, 1, 5, 1, 4)$
   - The sizes in the third dimension (1 and 5) are compatible.
   - The sizes in the fourth dimension (3 and 1) are compatible
   - $(1, 1, 5, 1, 4)$ is copied till its shape is $(3, 2, 5, 3, 4)$.
4. $(3, 2, 1, 3, 2)$ and $(5, 3, 1)$ are broadcastable.
   - Fill smaller tensor from the left: $(1, 1, 5, 3, 1)$
   - The sizes in the third dimension (1 and 5) are compatible.
   - The sizes in the fifth dimension (2 and 1) are compatible
   - $(1, 1, 5, 3, 1)$ is copied till its shape is $(3, 2, 5, 3, 2)$.
5. $(3, 2, 1, 3, 2)$ and $(1, 3, 1, 2)$ are broadcastable.
   - Fill smaller tensor from the left: $(1, 1, 3, 1, 2)$
   - The sizes in the third dimension (1 and 3) are compatible.
   - The sizes in the fourth dimension (3 and 1) are compatible.
   - $(1, 1, 3, 1, 2)$ is copied till its shape is $(3, 2, 3, 3, 2)$.
6. $(7, 1)$ and (7) are broadcastable.
   - Fill smaller tensor from the left: $(1, 7)$
   - The sizes in the first dimension (7 and 1) are compatible.
   - The sizes in the second dimension (1 and 7) are compatible.
   - $(1, 7)$ is copied till its shape is $(7, 7)$.