DS-GA 1008 Deep Learning Question 1

Shuang Gao

October 2020

1 Problem 1

1.1 (a)

Convolutional networks make two assumptions about the structure of the input data. Say (1) what these assumptions are, (2) what advantages we gain by using a convolution when they are true. Give an example of a data set where these assumptions would not hold.

(1) The assumptions are (i) the features are localized; (ii) the features are stationary.

(2) It can largely reduce the number of weights need to train in the neural network, leading to less computational complexity. It also has advantages such as faster convergence, better generalization, higher parallelization, and lower constraint of input size.

Example: In a data set with graph structures (e.g. knowledge graph, social networks), these assumptions would not hold.

1.2 (b)

Let $x[\cdot]$ and $y[\cdot]$ be two 1-D signals of length n. Consider a module which computes the cross-correlation between these two input signals with circular boundary conditions:

$$z[k] = \sum_{i=0}^{n-1} x[i] \cdot y[(i+k) \ mod \ n], \quad k \in \{0, \dots, n-1\}$$

Give the coefficients of the Jacobian of $z[\cdot]$ with respect to $x[\cdot]$ and $y[\cdot]$, i.e. give (1) $\frac{\partial z[k]}{\partial x[i]}$ and (2) $\frac{\partial z[k]}{\partial y[i]}$ for $k, i \in \{0, \dots, n-1\}$.

$$\frac{\partial z[k]}{\partial x[i]} = y[(i+k) \ mod \ n], \quad for \ i, k \in \{0, \dots, n-1\}$$

For $y[i]$, if $i \geq k$, the term $y[i]$ would appear is $x[i-k] \cdot y[i \ mod \ n] = x[i-k] \cdot y[i]$; if $i < k$, the term $y[i]$ would appear is $x[n+i-k] \cdot y[n+i \ mod \ n] = x[n+i-k] \cdot y[i]$. So

$$\frac{\partial z[k]}{\partial y[i]} = x[(n+i-k) \ mod \ n] = x[(i-k) \ mod \ n], \quad for \ i, k \in \{0, \dots, n-1\}$$

1.3 (c)

For $f(\cdot)$, i) the dimentionality of output space is $\mathbb{R}^{1000}$; ii) the number of trainable parameters in the layer is 100,000 totally; iii) the number of operations is 199,000 ($= 1000 \times (100 + 99) \times 1$).

For $g(\cdot)$, i) the dimentionality of output space is $\mathbb{R}^{100 \times 10}$; ii) the number of trainable parameters in the layer is 30 totally; iii) the number of operations is 5000 ($= 10 \times (3 + 2) \times 100$).