CS2109s - Tutorial 9

Eric Han (TG12-TG15)

Apr 11, 2024

Annoucements

Important admin

- PS6 marking is done!
- PS7 is due **Saturday, April 13 2024, 23:59** (One more to go!!)
- Last tutorial next week! - AMA Next week - https://www.menti.com/alo3igdy3ot9

PS6 Feedback

- Question 3: Task 1.3 - Observations on different model configurations; Plot out and examine the weight values
  - polyfit(x, y, mse, 1, 1e-3, 5000) > Underfitting
  - polyfit(x, y, mse, 6, 1e-3, 5000) > Numerical instability
- Question 6: Task 2.3 - Random seeds, how to compare to ensure fairness?
  - Take a few random seeds and average; Setting a fixed random seed does not work when we are comparing different architecture.

AMA For Last Tutorial

Figure 1: AMA Next week - https://www.menti.com/alo3igdy3ot9

Student Feedback on Teaching (SFT)

NUS Student Feedback https://blue.nus.edu.sg/blue/, due **26 Apr**:

- Don't Mix module/grading/project feedback - **feedback only for teaching**.
- Feedback is confidential to university and anonymous to us.
- Feedback is optional but highly encouraged.
- Past student feedback improves teaching; see https://www.eric-han.com/teaching
  - ie. Telegram access, More interactivity.
- Your feedback is important to me, and will be used to improve my teaching.
  - Good > Positive feedback > Encouragement
    - Teaching Awards (nominate)
    - Steer my career path
  - Bad > Negative feedback (nicely pls) > Learning
    - Improvement
    - Better learning experience

Student Feedback on Teaching (SFT)

Your feedback is important to me; optional, but highly encouraged:

Figure 2: NUS Student Feedback on Teaching - https://blue.nus.edu.sg/blue/

Question 1

Question 1a [G]

Find the cross-correlation (‘convolution’ as per CNN), $\mathbf{x} \bigotimes \mathbf{W}$:

$$\mathbf{x} = \begin{bmatrix} 0.10 & 0.20 & 0.10 & 0.10 & 0.00 \\ 0.80 & 0.90 & 1.00 & 1.00 & 0.90 \\ 1.00 & 1.00 & 1.00 & 1.00 & 1.00 \\ 0.90 & 1.00 & 1.00 & 0.80 & 1.00 \\ 0.00 & 0.10 & 0.10 & 0.20 & 0.00 \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} 1.00 & 1.00 & 1.00 \\ 0.00 & 0.00 & 0.00 \\ -1.00 & -1.00 & -1.00 \end{bmatrix}$$

[@] What is the difference between cross-correlation and convolution? Why are most CNNs implemented as cross-correlation? Find the convolution, $\mathbf{x} * \mathbf{W}$.

Recap

- How to calculate ‘convolution’ as per CNN?

Answer

$$\mathbf{x} \bigotimes \mathbf{W} = \begin{bmatrix} -2.60 & -2.60 & -2.80 \\ -0.20 & 0.10 & 0.10 \\ 2.80 & 2.60 & 2.70 \end{bmatrix}$$

$$\mathbf{x} * \mathbf{W} = \begin{bmatrix} 2.60 & 2.60 & 2.80 \\ 0.20 & -0.10 & -0.10 \\ -2.80 & -2.60 & -2.70 \end{bmatrix}$$

Flip the filter in both dimensions (or rotate 180 degrees) to go between cross-correlation and convolution.

- In this case here the result is the negative of each other, but in general it is **NOT**.

Question 1b

- Image input is $H \times W \times C = 224 \times 224 \times 3$
- First layer is Convolutional Layer with $C_1 = 96$ kernels of size 11×11, stride 4×4 without padding

Recap

How to calculate the output of a convolution?

Question 1b

- Image input is $H \times W \times C = 224 \times 224 \times 3$
- First layer is Convolutional Layer with $C_1 = 96$ kernels of size 11×11, stride 4×4 without padding

Recap

How to calculate the output of a convolution?

- Output height = (Input height + padding height top + padding height bottom - kernel height) / (stride height) + 1
- Output width = (Output width + padding width right + padding width left - kernel width) / (stride width) + 1

Answer

$$H_1 = \left\lfloor \frac{H - K + 2P}{S} \right\rfloor + 1 = 54$$

There are 96 filters so, $54 \times 54 \times 96$

Question 1c [G]

Images are often batched B. B can take values such as $8, 16, 32, 64$.

- Comment on the output shape if we feed the large CNN in part (b) with a batch.
- What are the advantages of using a batch of images rather than a single image?
- [@] Impact of large/small batch sizes and how to determine the optimal size?

Answer

$$H_1 = \left\lfloor \frac{H - K + 2P}{S} \right\rfloor + 1 = 54$$

There are 96 filters so, $54 \times 54 \times 96$

Question 1c [G]

Images are often batched B. B can take values such as $8, 16, 32, 64$.

- Comment on the output shape if we feed the large CNN in part (b) with a batch.
- What are the advantages of using a batch of images rather than a single image?
- [@] Impact of large/small batch sizes and how to determine the optimal size?

Answer

- $B \times H_1 \times W_1 \times C_1$
- Using a batch of images is computationally efficient and more stable in gradient descent convergence.

Question 2

Identify the type of RNN model and the characteristics required for the task:

a. Image Captioning
b. Text Classification
c. Language Translation

Recap

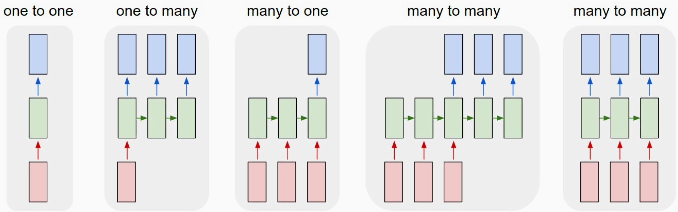

Figure 3: Rectangle is a vector and arrows represent matrix multiply; Input - red, output - blue and green - RNN's state. Taken from https://karpathy.github.io/2015/05/21/rnn-effectiveness/

Answer

a. One-to-many model
  - Input: One image. Output: Multiple words as captions.

b. Many-to-one model
  - Input: Many words. Output: Which class this text belongs to.

c. Many-to-many model
  - Input: Many words / code of language A. Output: Many words / code of language B.

Question 3 [G]

a. Performing sentiment analysis on Covid-19 posts on X. Explain what characteristics of RNN make it a standard model for sentiment analysis and which RNN model you want to use to tackle this problem.
b. Would it be possible to perform sentiment analysis using CNN? Explain why or why not.
c. Image recognition. Suppose we now want to recognize whether the image contains Chihuahua or muffin, briefly explain why CNN is good for image recognition.
d. Examine RNNs for image processing, formulate one.

Recap

- What are CNNs good at?
- What are RNNs good at?

Answer

a. RNN is the method for dealing with sequential input; Many/One RNN - Input: Sentence. Output: Sentiment(+/-).
b. Sentiment analysis strongly relies on context of the whole sentence; CNN convolution need many layers to detect higher level features to capture context.
    - I like durian
    - I do not like durain
    - I do not do not like durain
c. CNNs are very good at capture spatial structure - locality, ie. pixels near to each other are useful together - to recognize eye and layers above to compose the features.
d. Window as a token and we can slide it across to generate the input.
    - https://karpathy.github.io/2015/05/21/rnn-effectiveness/

Question 4 [G]

Dying ReLU Problem - majority of the activations are 0 (meaning the underlying pre-activations are mostly negative), resulting in the network dying midway.

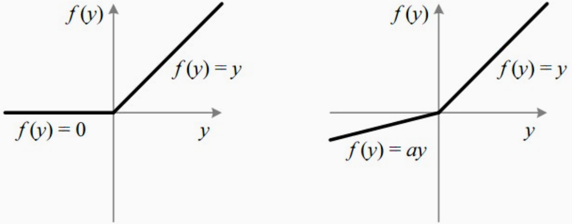

Figure 4: The Rectified Linear Unit (`ReLU`) (left) vs The Leaky Rectified Linear Unit (`Leaky ReLU`) with a as the slope when the values are negative. (right)

- How does `Leaky ReLU` fix this? What happens if we set $a = 1$ in the Leaky ReLU?

Answer

$$\mathrm{ReLU}(x) = \max(0, x), \quad \frac{\partial \mathrm{ReLU}(x)}{\partial x} = \begin{cases} 0, \text{if } x < 0 \\ 1, \text{if } x > 0 \end{cases}$$

- ReLU being stuck at 0 because the gradient is 0[1].
- Leaky ReLU get around this by creating small positive gradient a
- When $a = 1$, the activatation function becomes a linear function (NN loses power)

[1] Segway to last week calculations

Bonus Qn

Tasks

1. Implement `correlate2d(x,W)` and `convolve2d(x,W)` using `numpy`.
2. Calculate the values for question 1.
3. Compare it with `scipy.correlate2d(x,W, mode='valid')` and `scipy.convolve2d(x,W, mode='valid')`.

Buddy Attendance Taking

1. [@] and Bonus declaration is to be done here; You should show bonus to Eric.
2. Attempted tutorial should come with proof (sketches, workings etc. . .)
3. Guest students must inform Eric and also register the attendance.

Figure 5: Buddy Attendance: https://forms.gle/jsGfFyfo9PTgWxib6