به نام خدا

درس مبانی یادگیری عمیق

پاسخنامه تمرین سری پنجم

استاد درس : دکتر مرضیه داوودآبادی

دستیاران : مهسا موفق بهروزی، سید محمد موسوی، کمیل فتحی

دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر

نیمسال اول تحصیلی ۱۴۰۲ - ۱۴۰۳

پاسخ های مناسب برای هر سوال، لزوما یکتا نیستند

پاسخ سوالات ۱ تا۳ از پاسخ آقای محمد حسین عباسپور برداشته شده اند

۱. پاسخ صحیح را انتخاب کنید و دلیل انتخاب خود را به‌طور مختصر توضیح دهید. ممکن است سوالی، چند پاسخ صحیح داشته باشد(۱۵ نمره).

a) معماری $many-to-oneRNN$ برای کدام یک از وظایف زیر مناسب است؟

(آ) تشخیص گفتار[1](ورودی: کلیپ صوتی و خروجی: متن)

(ب) دسته‌بندی احساسات (ورودی: یک قطعه متن و خروجی: ۰/۱ برای نشان دادن احساس مثبت یا منفی)

(ج) تشخیص جنسیت از گفتار (ورودی: کلیپ صوتی و خروجی: برچسبی که نشان دهنده جنسیت صحبت کننده است)

در مورد آ ورودی چندتایی است و خروجی که یک متن میباشد شامل چند کلمه (چندتایی) است؛ پس معماری مناسب این حالت $many-to-many$ میباشد. در مورد ب و ج ورودی چندتایی است و خروجی فقط یک عدد (باینری) است (احساس مثبت هست یا نیست. جنسیت یا مرد است یا زن)، پس معماری مناسب این مسائل $many-to-one$ است. (ورودی‌ها هم چون $sequence$ هستند، در همه موارد از شبکه‌های RNN میتوان استفاده کرد)

b) اخلاق گربه جلوی دانشکده (پنبه) به‌شدت به آب‌وهوای فعلی و چند روز گذشته بستگی دارد. فرض کنید داده‌های آب‌وهوایی یک ماه گذشته را به‌صورت $x_1, ..., x_{30}$ و داده‌های مربوط به اخلاق

[1] Speech Recognition

پنبه را به‌صورت $y_1, ..., y_{30}$ جمع‌آوری کرده‌اید. می‌خواهید مدلی بسازید که x را به y نگاشت می‌کند. از کدام یک از RNN یک‌طرفه یا RNN دوطرفه برای این مسئله استفاده می‌کنید؟

(آ) دوطرفه، زیرا پیش‌بینی روز t بر اساس اطلاعات بیشتری انجام می‌شود.

(ب) دوطرفه، زیرا در $backpropagation$ گرادیان‌های دقیق‌تری محاسبه می‌شوند.

(ج) یک‌طرفه، زیرا مقدار y_t تنها به $x_1, ..., x_t$ وابسته است و به $x_{t+1}, ..., x_{30}$ وابسته نیست.

(د) یک طرفه، زیرا مقدار y_t تنها به x وابسته است و به داده‌های آب‌وهوای روزهای دیگر وابسته نیست.

طبق گفته صورت سوال، اخلاق گربه به آب و هوای فعلی و چند روز گذشته بستگی دارد؛ یعنی y_t به $x_1, ..., x_t$ بستگی دارد. پس نیازی به معماری دوطرفه نیست چون نیازی به $x_{t+1}, x_{t+2}, ...$ نداریم. پس گزینه ج صحیح است.

c) فرض کنید در حال آموزش یک مدل زبانی RNN هستید. در مرحله زمانی t ، مدل RNN چه چیزی را تخمین می‌زند؟ بهترین پاسخ را انتخاب کنید.

(آ) $P(y_1, y_2, ..., y_{t-1})$

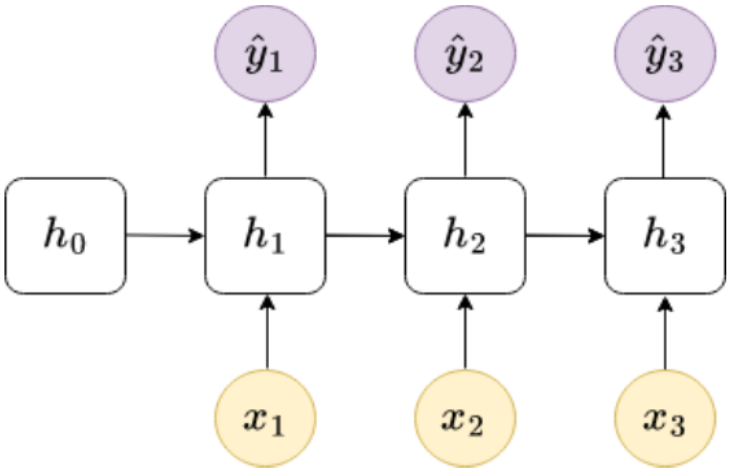
(ب) $P(y_1)$

(ج) $P(y_t|y_1, y_2, ..., y_{t-1})$

(د) $P(y_t|y_1, y_2, ..., y_t)$

شبکه‌های RNN دنباله‌ای هستند؛ یعنی با توجه به مشاهدات تا y_{t-1} برای y_t تصمیم‌گیری می کند. پس گزینه ج صحیح است.

۲. هدف از این تمرین آشنایی با $backpropagation$ در شبکه‌های بازگشتی و به دست آوردن $\frac{dJ}{dW_{hh}}$ است. شبکه بازگشتی زیر را در نظر بگیرید. در روابط زیر σ تابع $softmax$ و ψ تابع فعال‌سازی است(از در نظر گرفتن آن‌ها در محاسبات خود صرف نظر کنید.)(۲۰ نمره)



$$x_t \in \mathbb{R}^3 \qquad W_{hx} \in \mathbb{R}^{4\times 3}$$

$$h_t \in \mathbb{R}^4 \qquad W_{yh} \in \mathbb{R}^{2\times 4}$$

$$y_t, \hat{y}_t \in \mathbb{R}^2 \qquad W_{hh} \in \mathbb{R}^{4\times 4}$$

$$J = -\sum_{t=1}^{3}\sum_{i=1}^{2} y_{t,i} \log\left(\hat{y}_{t,i}\right)$$

$$\hat{y}_t = \sigma\left(o_t\right)$$

$$o_t = W_{yh} h_t$$

$$h_t = \psi\left(z_t\right)$$

$$z_t = W_{hh} h_{t-1} + W_{hx} x_t$$

لطفا پاسخ‌های خود را براساس W_{hh}، W_{yh}، y، $\hat{y}$، h و عبارات مشخص شده در سوال به دست آورید.
(توجه: نیازی نیست همه عبارات در همه پاسخ‌ها ظاهر شوند.)

الف) تابع ضرر $CrossEntropy$ در لحظه t را به صورت:

$$J_t = -\sum_{i=1}^{2} y_{t,i} \log \hat{y}_{t,i}$$

در نظر بگیرید. $\frac{\partial J_t}{\partial o_t}$ را محاسبه کنید.

$$\frac{\partial \log(x)}{\partial x} = \frac{1}{x} \rightarrow \frac{\partial \log\left(\hat{y}_{t,i}\right)}{\partial \hat{y}_{t,i}} = \frac{1}{\hat{y}_{t,i}}$$
$$\frac{\partial J_t}{\partial \hat{y}_t} = \left(-\sum_{i=1}^{2} y_{t,i} \cdot \frac{\partial \log\left(\hat{y}_{t,i}\right)}{\partial \hat{y}_{t,i}}\right) = \left(-\sum_{i=1}^{2} y_{t,i} \cdot \frac{1}{\hat{y}_{t,i}}\right) \quad , \quad \frac{\partial \hat{y}_t}{\partial o_t} = \sigma'\left(o_t\right)$$
$$\frac{\partial J_t}{\partial o_t} = \frac{\partial J_t}{\partial \hat{y}_t} \times \frac{\partial \hat{y}_t}{\partial o_t} = \left(-\sum_{i=1}^{2} y_{t,i} \cdot \frac{1}{\hat{y}_{t,i}}\right) \times \sigma'\left(o_t\right) = g_{o_t}$$

ب) مقدار $\frac{\partial J_t}{\partial o_t}$ را در متغیر g_{o_t} ذخیره می‌کنید. $\frac{\partial J_t}{\partial h_i}$ را برای یک i دلخواه، $i \in [۱,۳]$ محاسبه کنید. پاسخ خود را بر حسب g_{o_t} و متغیرهای ذکر شده بنویسید.
مقدار $\frac{\partial J_t}{\partial h_i}$ در صورتی تعریف شده است که $i <= t$ باشد(چون به مقادیر آینده دسترسی نداریم)

$$h_t = W_{hh} \cdot h_{t-1} + W_{hx} \cdot x_t \rightarrow \frac{\partial h_t}{\partial h_{t-1}} = W_{hh}$$
$$\frac{\partial o_t}{\partial h_t} = W_{yh}$$
$$\frac{\partial J_t}{\partial h_i} = \frac{\partial J_t}{\partial \hat{y}_t} \times \frac{\partial \hat{y}_t}{\partial o_t} \times \frac{\partial o_t}{\partial h_t} \times \left(\prod_{k=0}^{t-i-1} \frac{\partial h_{t-k}}{\partial h_{t-k-1}}\right) = g_{o_t} \times W_{yh} \times W_{hh}^{t-i} = g_{h_t}$$
$$t = 3, i = 3 \rightarrow \frac{\partial J_t}{\partial h_i} = g_{o_t} \times W_{yh} \times 1 = g_{o_t} \times W_{yh}$$
$$t = 3, i = 2 \rightarrow \frac{\partial J_t}{\partial h_i} = g_{o_t} \times W_{yh} \times W_{hh}$$

ج) مقدار $\frac{\partial J_t}{\partial h_i}$ را در متغیر g_{h_t} ذخیره می‌کنید. $\frac{\partial J_t}{\partial w_{hh}}$ را برحسب g_{h_t} و متغیرهای ذکر شده به‌دست آورید.

$$\frac{\partial h_i}{\partial W_{hh}} = h_{i-1}$$
$$\frac{\partial J_t}{\partial W_{hh}} = \sum_{i=1}^{t} \frac{\partial J_t}{\partial h_i} \cdot \frac{\partial h_i}{\partial w_{hh}} = \sum_{i=1}^{t} \frac{\partial J_t}{\partial h_i} \cdot h_{i-1} = \sum_{i=1}^{t} g_{h_t}(i) \cdot h_{i-1} = gW_{hh,t}$$

د) مقدار $\frac{\partial J_t}{\partial w_{hh}}$ را در متغیر $g_{W_{hh},t}$ ذخیره می‌کنید. $\frac{\partial J}{\partial w_{hh}}$ را برحسب $g_{W_{hh},t}$ و متغیرهای ذکر شده به دست آورید.

$$\frac{\partial J}{\partial W_{hh}} = \sum_{t=1}^{t=3} \frac{\partial J_t}{\partial W_{hh}} = \sum_{t=1}^{t=3} gW_{hh,t}$$

۳. یک نسخه فرضی از $attention$ به نام "$argmax$" را تصور کنید که دقیقاً مقدار[۲] متناظر با کلیدی[۳] که بیشترین شباهت به پرس‌وجو[۴] را دارد، برمی‌گرداند؛ شباهت با استفاده از ضرب داخلی اندازه‌گیری می‌شود(۲۰ نمره).

الف) با استفاده از توجه $argmax$، خروجی لایه توجه برای این پرس و جو چه خواهد بود؟

$$keys = \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ -2 \\ -4 \end{bmatrix} \right\}$$

$$q = \begin{bmatrix} 3 \\ -1 \\ -1 \end{bmatrix}$$

$$values = \left\{ \begin{bmatrix} 6 \\ 1 \\ -2 \end{bmatrix}, \begin{bmatrix} 6 \\ -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 6 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 \\ 1 \\ 2 \end{bmatrix} \right\}$$

inner product:

$$\text{keys } [0] \cdot q = 3 \times 1 + (-1) \times 2 + (-1) \times 3 = -2$$

$$\text{keys } [1] \cdot q = 3 \times 2 + (-1) \times 2 + (-1) \times 1 = 3$$

$$\text{keys } [2] \cdot q = 3 \times 0 + (-1) \times 1 + (-1) \times (-1) = 0$$

$$\text{keys } [3] \cdot q = 3 \times 0 + (-1) \times (-2) + (-1) \times (-4) = 6$$

$$\text{argmax}([-2, 3, 0, 6]) = 3 \rightarrow \text{ output } = \text{ values } [3] = \begin{bmatrix} 6 \\ 1 \\ 2 \end{bmatrix}$$

ب) این انتخاب طراحی (استفاده از $argmax$) چه تاثیری بر توانایی ما در آموزش مدل‌هایی که از مکانیزم توجه استفاده می‌کنند، دارد؟ (راهنمایی: به این فکر کنید که چگونه گرادیان‌ها از لایه آخر به سمت لایه اول شبکه منتقل می‌شوند. آیا می‌توانیم پرس‌وجوها یا کلیدهای خود را طی فرایند آموزش بهبود بخشیم؟)

[2] Value
[3] Key
[4] Query

در مکانیسم‌های توجه، استفاده از تابع «$argmax$» می‌تواند پیامدهای قابل‌توجهی برای توانایی مدل در یادگیری و سازگاری داشته باشد. در اینجا دلیل آن است:

- $Non-differentiability$:

تابع '$argmax$' غیر قابل مشتق‌پذیر است، به این معنی که گرادیان ندارد. این مسئله در طول $back-propagation$، فرآیندی که شبکه‌های عصبی توسط آن یاد می‌گیرند، مشکل ایجاد می‌کند. در طول انتشار $back-propagation$، گرادیان ها از لایه خروجی شبکه به لایه ورودی بازگردانده می شوند. اگر از «$argmax$» برای انتخاب مشابه ترین عنصر استفاده شود، گرادیان در آن نقطه تعریف نشده است و به طور موثر جریان گرادیان ها را متوقف می کند و از یادگیری شبکه جلوگیری می کند.

- $Sparse\ gradients$:

حتی اگر بتوانیم یک $subgradient$ برای "$argmax$" تعریف کنیم، آن مقدار کم خواهد بود (عمدتاً صفر) زیرا "$argmax$" فقط حداکثر مقدار را انتخاب می کند و بقیه را نادیده می گیرد. این می تواند منجر به همگرایی کندتر در طول آموزش شود، زیرا تنها بخش کوچکی از شبکه به روز می شود.

- $Lack\ of\ softness$:

$argmax$ تصمیم سختی می گیرد، به این معنی که فقط یک چیز (حداکثر مقدار) را انتخاب می کند. در زمینه مکانیسم‌های توجه، این می‌تواند مشکل‌ساز باشد، زیرا ممکن است برای مدل مفید باشد که تمرکز خود را بر روی بخش‌های متعدد ورودی به جای تنها یک قسمت توزیع کند.

۴. به نوتبوک $Question4.ipynb$ رفته و با مطالعه آن، موارد خواسته شده را تکمیل کنید(۴۵ نمره). به نوتبوک $Q4.ipynb$ مراجعه کنید.