به نام خدا

**درس مبانی یادگیری عمیق**

**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

---

# مهلت تحویل : ۱۴۰۲/۱۰/۰۱

**لطفا به نکات موجود در سند قوانین انجام و تحویل تمرین ها دقت فرمایید.**

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).

https://marcossilva.github.io/en/2019/08/12/coursera-deep-learning-module-5-week-1.html

$a$) معماری $many - to - one RNN$ 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)

$b$) 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 پنبه را به‌صورت $y_1, ..., y_{30}$ جمع‌آوری کرده‌اید. می‌خواهید مدلی بسازید که $x$ را به $y$ نگاشت می‌کند. از کدام $RNN$ 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.

[1] Speech Recognition

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

$c$) 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.

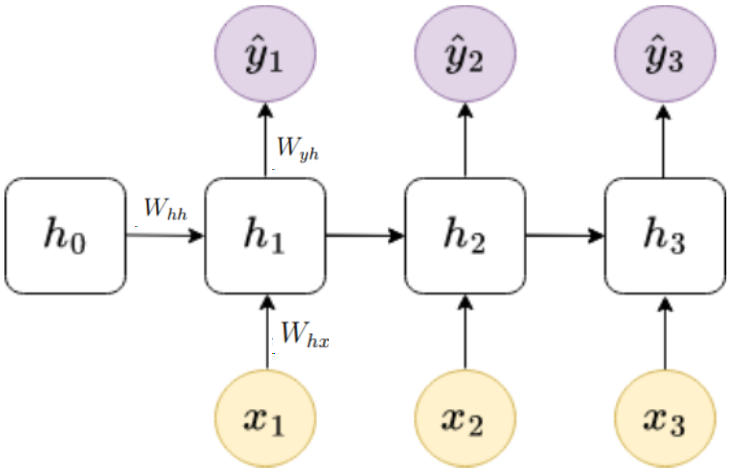
(آ) $P(y_1, y_2, ..., y_{t-1})$

(ب) $P(y_1)$

(ج) $P(y_t|y_1, y_2, ..., y_{t-1})$

(د) $P(y_t|y_1, y_2, ..., y_t)$

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 $\sigma$ تابع $softmax$ و $\psi$ 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.

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، $i \in [1,3]$ محاسبه کنید. پاسخ خود را بر حسب $g_{o_t}$ و متغیرهای ذکر شده بنویسید.

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.

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.

۳. یک نسخه فرضی از $attention$ به نام "$argmax$" را تصور کنید که دقیقاً مقدار[2] متناظر با کلیدی[3] که بیشترین شباهت به پرس‌وجو[4] 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؟

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

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

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

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

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

آموزش بهبود بخشیم؟ )

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