سوال یک

A)

ب و ج درست است.

به طور کلی، معماری "many-to-one RNN" برای وظایفی که ورودی آنها یک سری داده‌های زمانی است و خروجی یک مقدار یکتا است، مناسب است. در این نوع معماری، سری ورودی (به عنوان مثال، سری زمانی از واژگان یا ویژگی‌های صوتی) به یک شبکه عصبی بازگشتی (RNN) داده می‌شود و در نهایت یک مقدار یکتا (برچسب یا مقدار خروجی) تولید می‌شود.

در مورد گزینه ب (دسته بندی احساسات)، ورودی شامل یک سری قطعه متنی است و باید برچسب احساس مثبت یا منفی برای آن تولید کنیم. مثلاً با توجه به یک متن ورودی، باید تشخیص دهیم که آیا متن حاوی احساس مثبت یا منفی است. با استفاده از معماری "many-to-one RNN" می‌توانیم با تحلیل سری واژگان در طول زمان، یک برچسب احساسی (مثبت یا منفی) را به عنوان خروجی تولید کنیم.

در مورد گزینه ج (تشخیص جنسیت از گفتار)، ورودی شامل یک سری داده صوتی (مانند کلیپ صوتی) است و ما باید برچسبی که نشان دهنده جنسیت صحبت کننده است را تشخیص دهیم. با استفاده از معماری "many-to-one RNN"می‌توانیم با تحلیل سری ویژگی‌های صوتی در طول زمان، برچسب جنسیت را به عنوان خروجی تولید کنیم.

از طرفی، در مورد گزینه الف (تشخیص گفتار)، ورودی شامل یک کلیپ صوتی است و ما باید متن مربوطه را تولید کنیم. این وظیفه به صورت "one-to-many" است، یعنی با توجه به یک کلیپ صوتی، باید یک سری داده‌های زمانی (متن) تولید کنیم. این مورد با معماری "many-to-one RNN" سازگار نیست زیرا معماری "many-to-one RNN"برای تولید یک مقدار خروجی یکتا استفاده می‌شود و نمی‌تواند یک سری خروجی (متن) تولید کند.

بنابراین، معماری "many-to-one RNN" مناسب برای دسته بندی احساسات (گزینه ب) و تشخیص جنسیت از گفتار (گزینه ج) است، اما برای تشخیص گفتار (گزینه الف) مناسب نیست.

B)

گزینه ج صحیح است.

با توجه به سوال اصلی و فرضیات مطرح شده، ما نیاز داریم تا مدلی بسازیم که بتواند ویژگی‌های آب و هوای فعلی را به ویژگی‌های اخلاقی پنبه نگاشت کند. به عبارت دیگر، ما می‌خواهیم بر اساس داده‌های آب و هوای گذشته(x) ، اخلاق گربه را (y) پیش‌بینی کنیم.

در مسئله مذکور، ورودی ما دنباله‌ای از داده‌های آب و هوای گذشته است(x) ، و ما نیاز داریم که بر اساس این دنباله، خروجی مربوطه یعنی اخلاق گربه (y) را پیش‌بینی کنیم.

وقتی از RNN یک طرفه استفاده می‌کنیم، ما می‌توانیم اطلاعات آب و هوای گذشته را در نظر بگیریم و با توجه به این اطلاعات، اخلاق گربه را پیش‌بینی کنیم. اینجا فرض شده است که مقدار y (اخلاق گربه) تنها به x (داده‌های آب و هوای گذشته) وابسته است و به داده‌های آب و هوای روزهای آینده وابستگی ندارد که طبیعی نیز هست اخلاق گربه طبیعتا به اب و هوای روزهای اینده ربطی ندارد.

بنابراین، استفاده از RNN یک طرفه (گزینه ج) منطقی است زیرا مقدار y تنها به داده‌های آب و هوای گذشته وابسته است و به داده‌های آب و هوای آینده وابستگی ندارد.

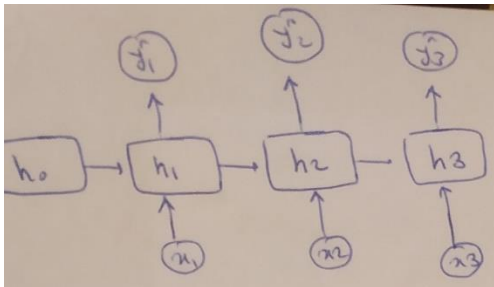
C)

گزینه ج صحیح است.

در مرحله زمانیt ، مدل RNN در حال تخمین زدن P(yt | y1, y2, ..., yt-1) است. به عبارت دیگر، مدل در حال پیش‌بینی احتمال وقوع نشانه (y) در زمان t به شرط داشتن تمام نشانه‌های قبلی (y1, y2, ..., yt-1) است.

بنابراین، گزینه (ج) P(yt | y1, y2, ..., yt-1) بهترین پاسخ است.

سوال دو



$x_t \in R^3$

$h_t \in R^4$

$y_t, \hat{y}_t \in R^2$

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

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

$o_t = W_{yh} h_t$

$h_t = \psi(z_t)$

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

الف)

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

$$\frac{\partial J_t}{\partial o_t} = -\sum_{i=1}^{2} \frac{\partial (y_{t,i} \log \hat{y}_{t,i})}{\partial o_{t,i}} \quad , \hat{y}_t = \sigma(o_t)$$

$$\Rightarrow -\sum_{i=1}^{2} y_{t,i} \frac{\partial (\log \hat{y}_{t,i})}{\partial o_{t,i}} = -\sum_{i=1}^{2} \frac{y_{t,i}}{\hat{y}_{t,i}} \frac{\partial (\hat{y}_{t,i})}{\partial o_{t,i}}$$

$\sigma(x) = \hat{y}(1-\hat{y})$

$$= -\sum_{i=1}^{2} \frac{y_{t,i}}{\hat{y}_{t,i}} (\hat{y}_{t,i})(1 - \hat{y}_{t,i})$$

$$\boxed{\frac{\partial J_t}{\partial o_t} = -\sum_{i=1}^{2} (y_{t,i})(1 - \hat{y}_{t,i})}$$

ب)

$o_t = W_{yh} h_t$

$\frac{\partial J_t}{\partial o_t} = g_{o_t}$, $\frac{\partial J_t}{\partial h_i}$ $i \in [1,3]$

$$\frac{\partial J_t}{\partial h_i} = \sum_{j=1}^{3} \frac{\partial J_t}{\partial o_{t,j}} \cdot \frac{\partial o_{t,j}}{\partial h_i} = \sum_{j=1}^{3} g_{o_{t,j}} \frac{\partial o_{t,j}}{\partial h_i} = \sum_{j=1}^{3} g_{o_{t,j}} W_{yh_{ij}}$$

ج)

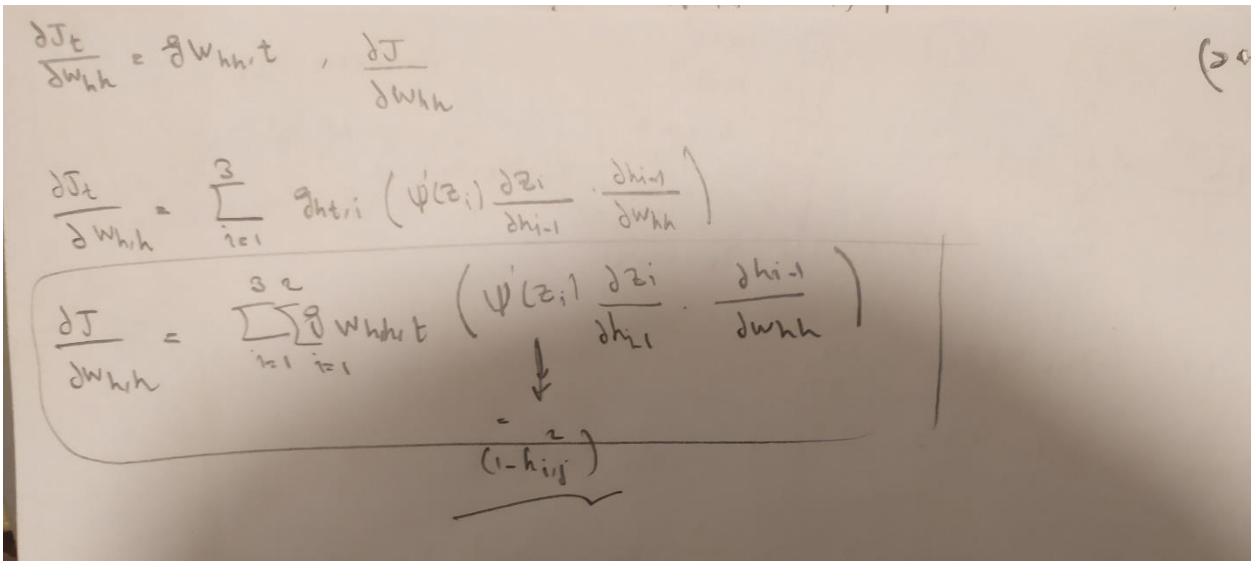
$\frac{\partial J_t}{\partial W_{hh_i}} = g_{h_t}$ $\frac{\partial J_t}{\partial W_{hh}} = \sum_{j=1}^{3} \frac{\partial J_t}{\partial h_{i,j}} \frac{\partial h_{i,j}}{\partial W_{hh}} = \sum_{j=1}^{3} g_{h_{t,j}} \frac{\partial h_{i,j}}{\partial W_{hh}}$

$g_{h_{t,j}}$

$h_{i,j} = \psi(z_{i,j})$, $z_{i,j} = W_{hh_i} h_{t-1} + W_{hx} \cdot x_t$

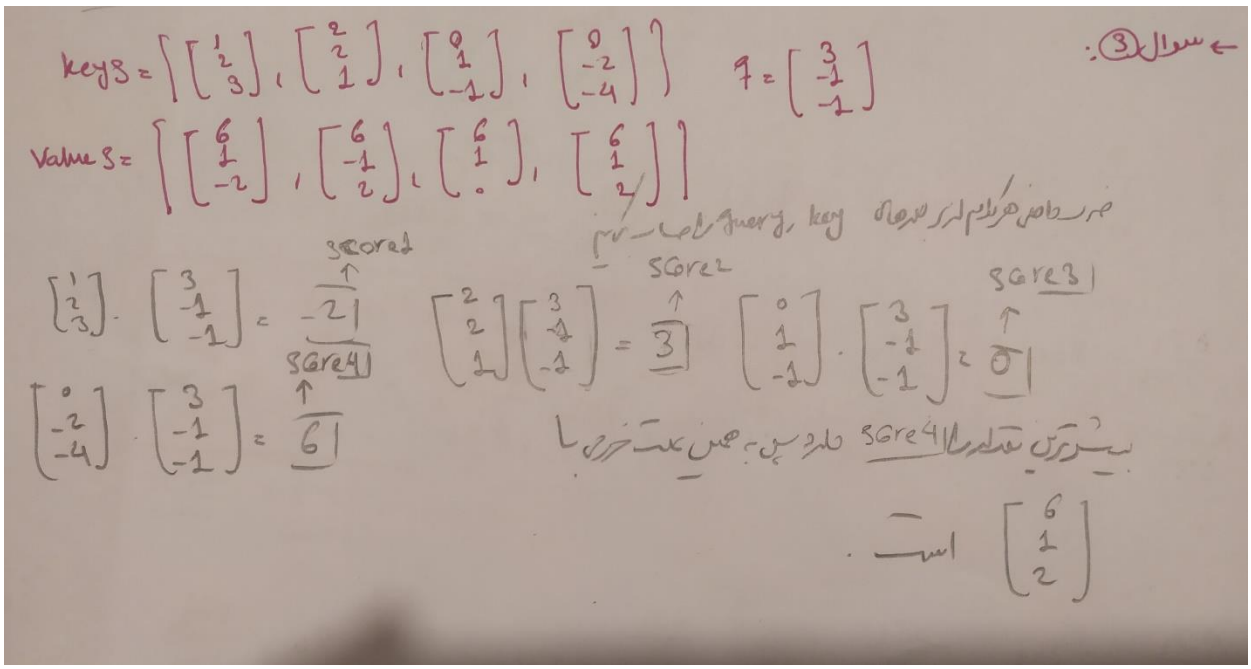
$$\frac{\partial h_{i,j}}{\partial W_{hh}} = \frac{\partial h_i}{\partial z_i} \frac{\partial z_i}{\partial W_{hh}} = \psi'(z_i) \frac{\partial z_i}{\partial W_{hh}} = \psi'(z_i) \left(\frac{\partial z_i}{\partial h_{i-1}} \frac{\partial h_{i-1}}{\partial W_{hh}} \right)$$

$$= \frac{\partial J_t}{\partial W_{hh}} = \sum_{j=1}^{3} g_{h_{t,j}} \left(\psi'(z_i) \frac{\partial z_i}{\partial h_{i,j}} \frac{\partial h_{i-1}}{\partial W_{hh}} \right)$$



سوال سه

الف)



ب)

استفاده از argmax برای مکانیزم توجه در آموزش مدل‌ها تأثیر قابل توجهی بر توانایی ما در آموزش و بهبود مدل‌هایی که از مکانیزم توجه استفاده می‌کنند دارد. در ادامه، تأثیر این انتخاب طراحی را بررسی می‌کنیم:

در مدل‌های با مکانیزم توجه، گرادیان‌ها از لایه آخر به سمت لایه اول شبکه منتقل می‌شوند تا وزن‌ها و پارامترهای شبکه به‌روزرسانی شوند. برای این منظور، از روش پس‌انتشار خطا استفاده می‌شود که بر اساس محاسبه گرادیان است. حالت argmax باعث می‌شود که گرادیان‌ها در طول این فرآیند منتقل نشوند و آموزش مدل مشکلاتی را به وجود آورد.

به‌طور معمول، در آموزش مدل‌های با مکانیزم توجه از روش softmax برای تولید وزن‌های توجه استفاده می‌شود. با استفاده ازsoftmax ، توزیع احتمالاتی از وزن‌های توجه برای هر کلید تولید می‌شود و گرادیان‌ها منتقل می‌شوند. این باعث می‌شود که مدل قادر به یادگیری و بهبود پرس و جوها یا کلیدها در طول فرآیند آموزش شود.

اگر از argmax بجای softmax استفاده شود، گرادیان‌ها در طول فرآیند پس‌انتشار خطا به درستی منتقل نمی‌شوند و آموزش مدل مشکلاتی را به همراه خواهد داشت. این به این معنی است که ما نمی‌توانیم پرس وجوها یا کلیدهای خود را بهبود داده و در فرآیند آموزش تغییراتی را در آنها اعمال کنیم.

بنابراین، استفاده از argmax به جای softmax در مکانیزم توجه می‌تواند توانایی ما در آموزش مدل‌هایی که از مکانیزم توجه استفاده می‌کنند را محدود کند و امکان بهبود پرس وجوها یا کلیدها در طول فرآیند آموزش را به ما ندهد.

سوال چهار

کد ها زده شده است.