به نام خالق رنگین کمان

ستاره باباجانی – 99521109

سوال 1: a) این معماری برای موارد (ب) و (ج) استفاده میشود.

این معماری برای وظایف دسته بندی احساسات و تشخیص جنسیت ازگفتار مناسب است ولی برای تشخیص گفتار مناسب نیست زیرا آنها برای وظایفی طراحی شده اند که در آنها دنباله ای از ورودیها داریم و خروجی ما واحد است. در موارد (ب) و (ج) خروجی واحد ولی در مورد (آ) خروجی چندتا است.

حال به بررسی هر گزینه میپردازیم:

1. طبقه بندی احساسات:

- ورودی دنباله ای از کلمات است.
- خروجی یک برچسب احساسی (مثبت یا منفی) است.
- RNN می تواند با در نظر گرفتن ترتیب کلمات و روابط آنها نحوه آشکار شدن احساس در طول متن را تصویر کند.

2. تشخیص جنسیت از گفتار:

- ورودی دنباله ای از فریم های صوتی (کلیپ گفتار) است.
- خروجی یک برچسب تک جنسیتی (مرد یا زن) است.

- RNN می تواند چگونگی تکامل ویژگی های صوتی مانند آهنگ های زیر و بمی صدا و آهنگ را ردیابی کند.

3. چرا تشخیص گفتار مناسب نیست:

- تشخیص گفتار مستلزم نگاشت دنباله ای از فریم های صوتی به دنباله ای از متن است.(پس به معماری many-to-many احتیاج داریم.)
- در اصل، RNN های چند به یک در وظایفی که درک معنی یا الگوی کلی در یک دنباله برای تولید یک طبقه بندی دقیق و واحد ضروری است، برتری می یابند.

b) گزینه (ج) یعنی یک طرفه (مقدار yt تنها به x1,...,xt وابسته است و به xt+1,...,x30 وابسته نیست) درست است زیرا اخلاق پنبه به آب و هوای فعلی و چند روز گذشته بستگی دارد و به آب و هوای آینده که هنوز اتفاق نیوفتاده است، بستگی ندارد.

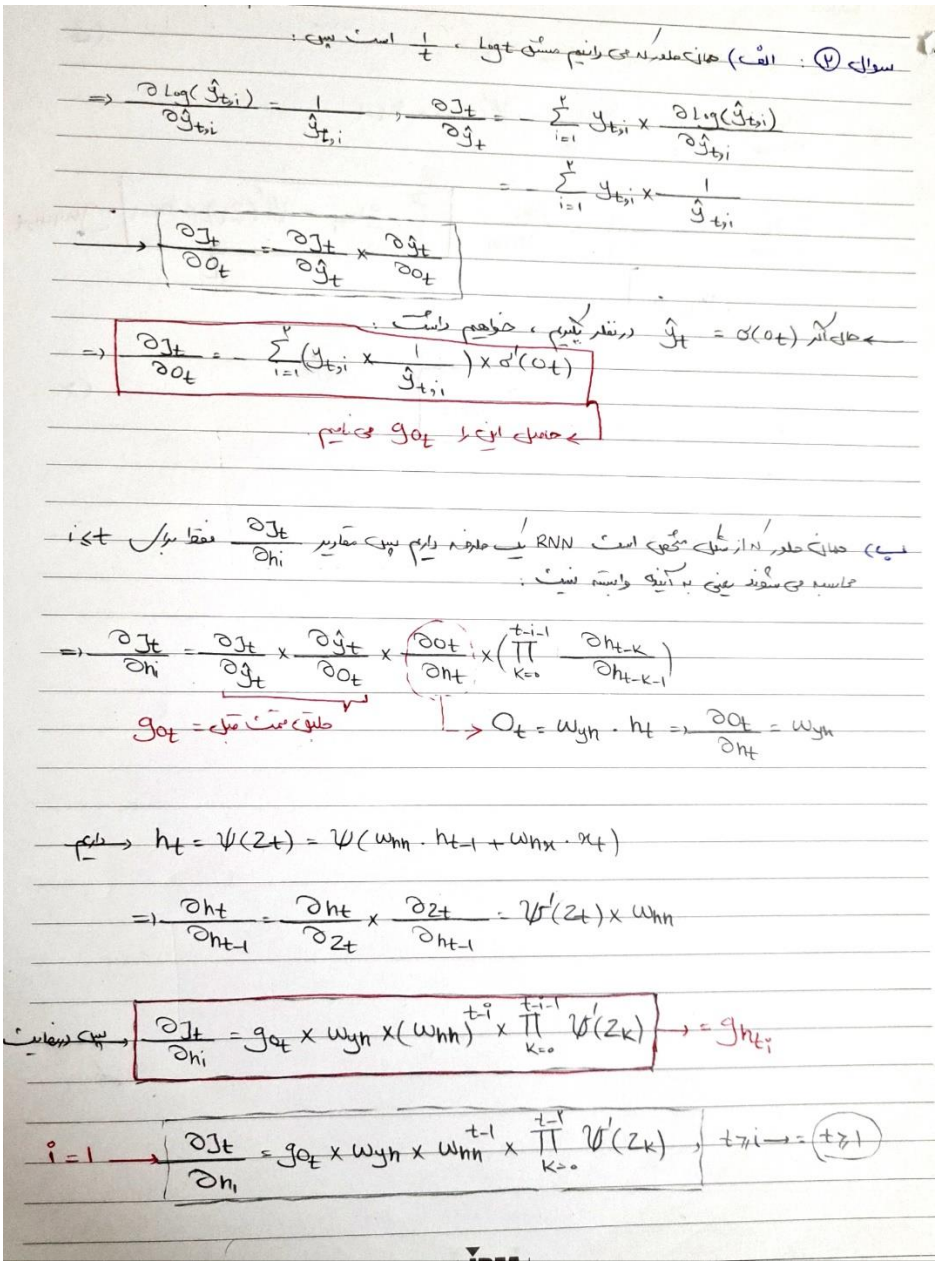
یک RNN یک طرفه، که توالی آب و هوا را به صورت زمانی پردازش می کند، می تواند به طور موثر این رابطه را مدل کند. استفاده از یک RNN دو جهته غیر ضروری خواهد بود و با در نظر گرفتن اطلاعات نامربوط آینده، به طور بالقوه نویز ایجاد می کند.

بنابراین، یک RNN یک‌طرفه مناسب‌ترین انتخاب برای این مشکل است، که به طور دقیق از وضعیت خلق و خوی پنبه بر اساس تاریخچه آب و هوای مربوطه عکس می‌گیرد.

c) گزینه (ج) درست است زیرا در یک مدل زبان سعی میشود که گام بعدی بر اساس دانش تمامی مراحل قبلی پیش بینی شود.

- تخمین P(y, y1,...,yt-1) شامل پیش‌بینی کل دنباله تا t-1 است که تمرکز RNN در مرحله زمانی t نیست.
- تخمین P(y1) زمینه کلمات قبلی را نادیده می گیرد که برای مدل سازی زبان بسیار مهم است.
- تخمین P(yt | y, y1,...,yt) ممکن نیست زیرا RNN هنوز yt را در مرحله زمانی t مشاهده نکرده است.

سوال 2:



سوال ۴ : الف) همانطور که در کلاس مشتق $\log t$ شد $\frac{1}{t}$ است پس :

$$\frac{\partial \text{Log}(\hat{y}_{t,i})}{\partial \hat{y}_{t,i}} = \frac{1}{\hat{y}_{t,i}} \quad , \quad \frac{\partial J_t}{\partial \hat{y}_t} = -\sum_{i=1}^{2} y_{t,i} \times \frac{\partial \text{Log}(\hat{y}_{t,i})}{\partial \hat{y}_{t,i}} = -\sum_{i=1}^{2} y_{t,i} \times \frac{1}{\hat{y}_{t,i}}$$

$$\rightarrow \boxed{\frac{\partial J_t}{\partial o_t} = \frac{\partial J_t}{\partial \hat{y}_t} \times \frac{\partial \hat{y}_t}{\partial o_t}}$$

$\hat{y}_t = \sigma(o_t)$ حال اگر در نظر بگیریم ، خواهیم داشت :

$$\Rightarrow \boxed{\frac{\partial J_t}{\partial o_t} = -\sum_{i=1}^{2} \left(y_{t,i} \times \frac{1}{\hat{y}_{t,i}} \right) \times \sigma'(o_t)}$$

حاصل این را g_{o_t} می‌نامیم

ب) همانطور که از شکل مشخص است RNN ما دارای چند مقادیر پنهان است فقط برای $t=3$ محاسبه می‌شوند یعنی به اینکه وابسته نیست :

$$\Rightarrow \frac{\partial J_t}{\partial h_i} = \frac{\partial J_t}{\partial \hat{y}_t} \times \frac{\partial \hat{y}_t}{\partial o_t} \times \frac{\partial o_t}{\partial h_t} \times \left(\prod_{k=0}^{t-i-1} \frac{\partial h_{t-k}}{\partial h_{t-k-1}} \right)$$

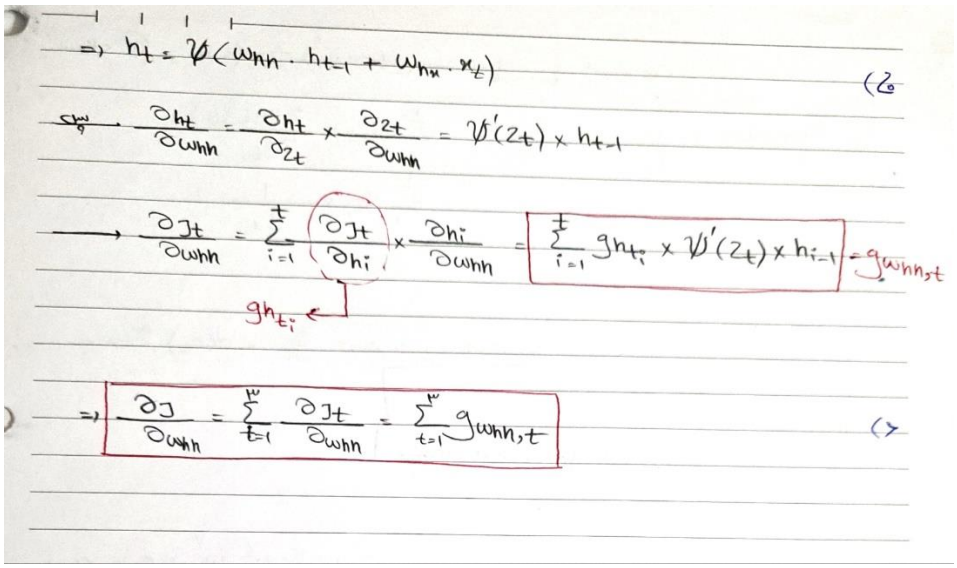
g_{o_t} = طبق مثبت قبل $\qquad o_t = w_{yh} \cdot h_t \Rightarrow \frac{\partial o_t}{\partial h_t} = w_{yh}$

$$h_t = \psi(z_t) = \psi(w_{hh} \cdot h_{t-1} + w_{hx} \cdot x_t)$$

$$\Rightarrow \frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial h_t}{\partial z_t} \times \frac{\partial z_t}{\partial h_{t-1}} = \psi'(z_t) \times w_{hh}$$

پس سمت راست $\rightarrow \boxed{\frac{\partial J_t}{\partial h_i} = g_{o_t} \times w_{yh} \times (w_{hh})^{t-i} \times \prod_{k=0}^{t-i-1} \psi'(z_k)} \rightarrow = g_{h_{t,i}}$

$i=1 \rightarrow \boxed{\frac{\partial J_t}{\partial h_1} = g_{o_t} \times w_{yh} \times w_{hh}^{t-1} \times \prod_{k=0}^{t-2} \psi'(z_k)}$, $t \neq i \rightarrow = (t \neq 1)$



سوال 3: الف) همان طور که میدانیم این مکانیزم مقداری را انتخاب میکند که بیشترین شباهت را به پرس و جو با استفاده از ضرب داخلی دارد.

حال برای محاسبه بیشترین شباهت(خروجی) مراحل زیر را طی میکنیم:

- محاسبه ضرب داخلی بین پرس و جو و کلیدها:

=> q . Keys[0] = 3 * 1 + (-1 * 2) + (-1 * 3) = -2
=> q · Keys[1] = 3 * 2 + (-1 * 2) + (-1 * 1) = 3
=> q · Keys[2] = 3 * 0 + (-1 * 1) + (-1 * -1) = 0
=> q · Keys[3] = 3 * 0 + (-1 * -2) + (-1 * -4) = 6

- اعمال argmax روی مقادیر بدست آمده برای انتخاب اندیس مشابه ترین:

=> argmax([-2, 3, 0, 6]) = 3

- محاسبه مقدار خروجی از طریق اندیس محاسبه شده:

=> Output = Values[3] = [6, 1, 2]

پس خروجی لایه برای q داده شده با استفاده از این مکانیزم، [6,1,2] است.

ب) استفاده از argmax در زمینه مکانیسم های توجه به دلیل ماهیت غیر قابل تمایز آن با وجود اینکه از نظر محاسباتی ساده است، می تواند چالش هایی را در طول آموزش ایجاد کند:

- تابع argmax شاخص حداکثر مقدار را در مجموعه‌ای از مقادیر انتخاب می‌کند، به این معنی که گسسته است و به راحتی قابل بهینه‌سازی مبتنی بر گرادیان نیست: در مکانیسم‌های توجه، تابع softmax معمولاً برای استخراج وزن‌های توجه بر اساس نمرات شباهت (محاسبه شده با استفاده از پرس و جو و کلیدها) استفاده می‌شود که مجموع نمرات را 1 می کند و امکان توزیع یکنواخت و احتمالی را فراهم می کند که می توان آن را متمایز کرد و در پس انتشار برای آموزش استفاده کرد. وقتی از argmax به جای softmax استفاده می شود، یک ناپیوستگی در گرادیان ها ایجاد می کند. این ناپیوستگی محاسبه مستقیم گرادیان ها با توجه به پارامترهای کلیدها و پرس و جوها را غیرممکن می کند زیرا عملیات argmax خود دارای یک گرادیان کاملاً تعریف شده نیست. در نتیجه، این جریان موثر شیب ها را در طول شبکه در طول انتشار پس از آن مهار می کند.

- توجه argmax ممکن است برای کارهایی که به درک روابط پیچیده بین کلیدها و پرس و جوها دارند، بیش از حد ساده باشد بطوریکه تفاوتهای کوچک یا شباهت در توزیع ها را در بر نگیرد.
- چون خروجی بر اساس شبیه ترین کلید بدون هیچ گونه تبدیل قابل به یادگیری است، مدل نمیتواند در طول فرآیند آموزش بهبود کلید ها را بیاموزد.
- به جای استفاده ازargmax ، استفاده از ترفند-Gumbel Softmax، می تواند مفید باشد. این روش یک عنصر تصادفی را معرفی می کند که عملیات argmax گسسته را با یک تقریب پیوسته و قابل تمایز تقریب می زند. این به گرادیان‌ها اجازه می‌دهد تا در طول آموزش از طریق مکانیسم توجه جریان پیدا کنند و شبکه را قادر می‌سازد تا نمایش‌های بهتری از پرس و جوها و کلیدها را بیاموزد.

پایان