بسم الله الرحمن الرحیم

محمد عرفان زارع زردینی

۹۸۴۱۱۴۳۲

تمرین سری پنجم درس یادگیری عمیق

مدرس درس: استاد داوود آبادی

پاییز 1402

سوال1)

a)

معماری چند به یک RNN برای کارهایی که ورودی های متوالی (گام های زمانی متعدد) برای تولید یک خروجی واحد باید پردازش شود، مناسب است. حال به بررسی گزینه ها می پردارم:

الف) تشخیص گفتار، تبدیل یک کلیپ صوتی(صوت) به متن می باشد.با توجه به موقعیت هایی که ماهیت متوالی داریم در داده های صوتی، در طول زمان با معماری چند به یک RNN مطابقت دارد. پس برای این تسک **مناسب است**.

ب) این تسک با قطعه متن سروکار داریم. پس لزوما به تحلیل متوالی متن در طول زمان برای تعیین احساسات نیست. برای این تسک بهتر است از CNN استفاده شود. پس برای این تسک RNN **مناسب نیست**.

ج) این تسک همچون تسک قسمت الف بوده و تبدیل صوت به متن است. در این فرمت ، ماهیت پردازش داده های صوتی متوالی برای تعیین جنسیت است. در طول زمان با معماری RNN چند به یک تطابق دارد. پس **مناسب است.**

B)

این سناریو پیرامون پیش بینی رفتار گربه براساس داده های آب و هوای فعلی و داده های چند روز گذشته است. در این مورد ، با توجه به قسمت های مختلف، گزینه ها بررسی شده اند:

الف) این گزینه استفاده از یک RNN دو طرفه را پیشنهاد می کند زیرا هم اطلاعات گذشته و هم اطلاعات آینده را برای پیش بینی در نظر می گیرد. با این حال، در این سناریو، رفتار گربه احتمالا بیشتر به داده های آب و هوای گذشته و آب و هوای فعلی بستگی دارد تا داده های آب و هوای

آینده. بنابراین، در نظر گرفتن داده های آب و هوای آینده ممکن است برای پیش بینی دقیق رفتار گربه ضروری یا مرتبط نباشد. پس این گزینه **مناسب نیست.**

ب)RNN های دو جهته ممکن است زمینه بیشتری را فراهم کنند، اما دقت گرادیان ها در Backpropagation ممکن است لزوماً چالش اصلی در اینجا نباشد. در حالی که RNN های دو طرفه به شبکه اجازه می دهند هم زمینه های گذشته و هم آینده را در نظر بگیرند، این امر دقت بهتری را برای کار خاص پیش بینی رفتار گربه بر اساس داده های آب و هوای تاریخی و فعلی تضمین نمی کند. پس این گزینه **مناسب نیست.**

ج) این گزینه، استفاده از یک RNN یک طرفه را پیشنهاد می‌کند، که در آن پیش‌بینی رفتار گربه در روز t (yt) تنها به داده‌های آب و هوای تاریخی تا روز x1,...,xt متکی است. از آنجایی که رفتار گربه ممکن است لزوماً به داده های آب و هوای آینده بستگی نداشته باشد، یک RNN یک طرفه با بررسی اطلاعات گذشته و فعلی می تواند **انتخاب مناسبی باشد**.

د) مشابه گزینه قبلی، استفاده از RNN یک طرفه (یک طرفه) را پیشنهاد می کند. با این حال، با بیان اینکه رفتار گربه فقط به داده های آب و هوای روز جاری (x) بستگی دارد، مسئله را بیش از حد ساده می کند. این ممکن است وابستگی های زمانی در داده ها را نشان ندهد، زیرا رفتار ممکن است تحت تأثیر الگوهای آب و هوایی تاریخ قبل نیز قرار گیرد.

(C

با توجه به خواسته مسئله ، گزینه هارو بررسی میکنیم:

الف)این گزینه بدان معناست که در مرحله زمانیt در RNN ، احتمال کل توالی توکن هارا از y1 تا yt-1 تخمین میزند. با این حال، در RNN در در مرحله زمانی، معمولا قسمت بعدی دنباله را به جای توزیع احتمال کل دنباله تا آن نقطه پیش بینی می کند.

ب) این گزینه بدان معناست که RNN احتمال اولین توکن که y1 هست را تخمین می زند. با این حال، در یک مدل زبان RNN ، در مرحله زمانی t ، مدل معمولاً توزیع احتمال نشانه بعدی را با توجه به دنباله تا مرحله زمانی t-1 پیش بینی می کند.

ج) این انتخاب نشان دهنده احتمال توکن در مرحله زمانی t با توجه به کل دنباله تا آن نقطه (از y1 تا yt) است. شامل زمینه دنباله تا مرحله زمانی t است، و با پیش‌بینی معمولی که توسط یک مدل زبان RNN در مرحله زمانی t انجام می‌شود، بیشتر همسو می‌شود. با این حال، به اشتباه شرایط را به‌عنوان نشانه در مرحله زمانی t مدنظر میگیرند، که اینطور نیست و مدل معمولاً نشانه بعدی را در دنباله پیش‌بینی می‌کند.

د)

این گزینه احتمال توکن در مرحله زمانی t با توجه به دنباله تا مرحله زمانی t-1 است. در یک مدل زبان RNN، مدل توزیع احتمال نشانه بعدی را بر اساس دنباله تا مرحله زمانی قبلی t-1 پیش بینی می کند.

حال با توجه به بررسی های انجام شده، **گزینه د) مناسب است.** در مرحله زمانی t، یک مدل زبانی RNN توزیع احتمال نشانه بعدی که Yt است را با توجه به دنباله تا مرحله زمانی قبلی t-1 تخمین میزند. این پیش‌بینی به تولید توکن بعدی در دنباله بر اساس زمینه‌ای که از توکن‌های قبلی آموخته شده است کمک می‌کند.

سوال 2)

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

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

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

$h_t = \psi(z_t)$ ، $z_t = w_{hh} \cdot h_{t-1} + w_{hx} \cdot x_t$ ، $J_i = -\sum_{i=1}^{2} y_{t,i} \log(\hat{y}_{t,i})$ (ب

$o_t = w_{yh} \cdot h_t \Rightarrow \frac{\partial J_t}{\partial h_i} = \sum_{j=1}^{3} \frac{\partial J_t}{\partial o_{t,j}} \cdot \frac{\partial o_{t,j}}{\partial h_i} \Rightarrow \frac{\partial J_t}{\partial h_i} = \sum_{i=1}^{3} g_{o_{t,j}} \cdot \frac{\partial o_{t,j}}{\partial h_i}$

$\Rightarrow O_t \cdot w_{yh} \cdot h_t \Rightarrow \frac{\partial J_t}{\partial h_i} = \sum_{j=1}^{3} g_{o_{t,j}} \cdot w_{yh_{i,j}}$

$$\frac{\partial J_t}{\partial w_{hh}} = \sum_{i=1}^{3} g_{h_{t,i}} \cdot \frac{\partial h_i}{\partial w_{hh}} \Rightarrow h_i = \psi(z_i) \rightarrow z_i = w_{hh} \cdot h_{i-1} + w_{h_x} x_i$$ (ج

$$\frac{\partial h_i}{\partial w_{hh}} = \frac{\partial \psi(z_i)}{\partial w_{hh}} + \frac{\partial \psi(z_i)}{\partial h_{i-1}} \cdot \frac{\partial h_{i-1}}{\partial w_{hh}} \Rightarrow$$

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

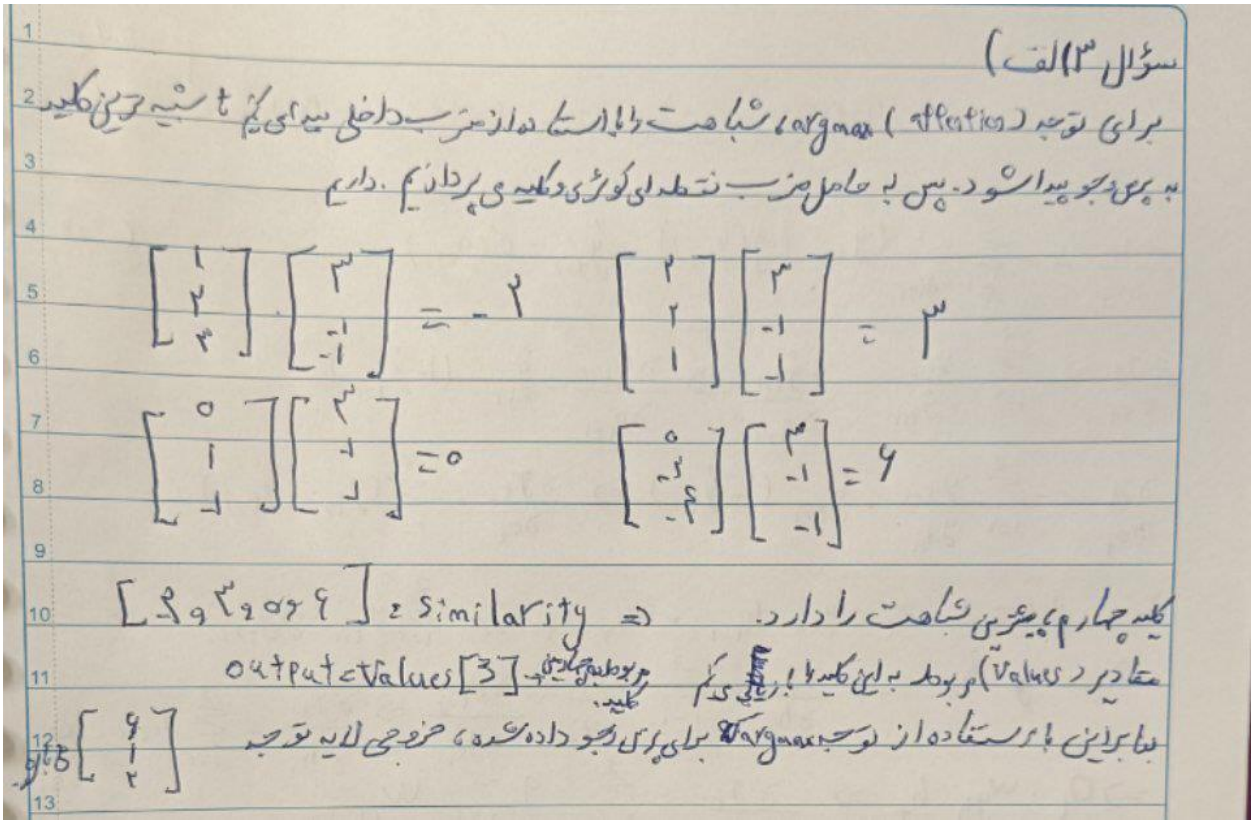
$$\psi'(z_i) = \psi(z_i) \cdot (1 - \psi(z_i)) \Rightarrow \frac{\partial \psi(z_i)}{\partial h_{i-1}} = \psi'(z_i) \cdot w_{hh}$$ (د

$$\Rightarrow \frac{\partial h_{i-1}}{\partial w_{hh}} = h_{i-2} \Rightarrow \frac{\partial J_t}{\partial w_{hh}} = \sum_{i=1}^{3} g_{w_{hh} t_i} (\psi(z_i) + \psi(z_i) w_{hh} \cdot h_{i-2})$$

سوال3)

منبع: (gpt3)

الف)



ب)

انتخاب طراحی بر مدل های آموزشی با استفاده از مکانیسم های توجه اثر گذار است که در ذیل توضیح دادم. وقتی نحوه انتقال گرادیان ها از آخرین لایه به لایه اول شبکه بررسی شود، مهم است که تأثیر استفاده از توجه argmax را در نظر گرفته شود .
حال به بررسی آن می پردازیم:

جریان شیب (Gradient Flow) : در شبکه های عصبی ، در حین backpropagation، شیب ها از طریق لایه ها به سمت عقب پخش می شوند تا

وزن ها را به روز کنند. البته، با توجه argmax، که شامل عملیات deterministic و غیر قابل تمایز است (از آنجایی که حداکثر شباهت را بدون در نظر گرفتن توابع قابل تمایز مانند softmax انتخاب می‌کند)، گرادیان‌ها ممکن است به آرامی در مکانیسم توجه جریان نداشته باشند.

چالش های آموزشی : ماهیت deterministic ، argmax می تواند منجر به ناپیوستگی گرادیان و مشکلات در انتشار گرادیان های معنی دار از طریق شبکه شود. این می تواند منجر به چالش هایی در یادگیری به طور مؤثر شود ، به خصوص هنگام برخورد با داده های پیچیده یا نویزی.

برای بررسی این مسائل و بهبود روند آموزش داریم:

پارامترهای قابل یادگیری برای پرس و جوها و کلیدها: معرفی پارامترهای قابل یادگیری برای پرس و جوها و کلیدها به مدل اجازه می دهد تا به طور تطبیقی نمایش هایی را تنظیم کند و یاد بگیرد که روابط بین پرس و جوها و کلیدها را بهتر به تصویر بکشد. این پارامترهای قابل یادگیری را می توان در طول آموزش به روز کرد و امکان یادگیری مؤثرتر را فراهم می شود.

- منظم سازی و اصلاحات(Regularization and Modifications) : تکنیک هایی مانند روش های منظم سازی (مانند dropout) یا اصلاح مکانیسم توجه می تواند به کاهش ماهیت جبری argmax کمک کند. به عنوان مثال، معرفی یک مکانیسم نرم‌کننده برای smooth کردن و متمایز کردن توزیع توجه می‌تواند جریان گرادیان بهتری را در طول تمرین تسهیل کند.

بهبود نمایش داده ها یا کلیدها در طول آموزش با استفاده از آنها به عنوان پارامترهای قابل یادگیری امکان پذیر است. با اجازه دادن به این پارامترها برای به روز رسانی بر

اساس داده های آموزشی، مدل به طور بالقوه می تواند بازنمایی های معنی دار بیشتری بیاموزد و روابط بین پرس و جوها و کلیدها را بهتر نمایش دهد.

بدین سان باوجود اینکه روش هایی برای بهبود فرآیند آموزش توجه argmax هست ولی ممکن است چالش هایی را در گرفتن روابط پیچیده ایجاد نماید و نیاز به مدیریت دقیق در بهینه سازی دارد تا یادیگری موثر داشته باشیم.

سوال4) گفته شد نیاز به توضیحات ندارد