

# In the name of GOD

Sina Alinejad

NLP HW4

Student id: 99521469

سوال ۱-

سوال ۱-

$$P(\text{امروز کتاب خوانیم}) = P(\text{ما امروز} | \text{کتاب خوانیم})$$

$$P(\text{کتاب خوانیم}) = \frac{452}{1872} \times \frac{241}{1923} \times \frac{320}{1245} \approx 0.0073$$

$$P(\text{دیروز کتاب خوانیم}) = P(\text{ما دیروز} | \text{کتاب خوانیم})$$

$$P(\text{کتاب خوانیم}) = \frac{411}{1872} \times \frac{48}{201} \times \frac{345}{945} \approx 0.0024$$

سوال ۲-

سوال ۲-

$$P(w_1^n) = P(w_1 w_2 \dots w_n)$$

$$P(Y|X) = \frac{P(X, Y)}{P(X)} \rightarrow P(X, Y) = P(X)P(Y|X)$$

$$P(w_1 w_2 \dots w_n) = P(w_1 w_2 \dots w_{n-1}) P(w_n | w_1 w_2 \dots w_{n-1})$$

با باز کردن عبارت اول داریم:

$$= P(w_1 \dots w_{n-2}) P(w_{n-1} | w_1 \dots w_{n-2}) P(w_n | w_1 \dots w_{n-1})$$

با ادامه این کار داریم:

$$= P(w_1) P(w_2 | w_1) P(w_3 | w_1 w_2) \dots P(w_n | w_1^{n-1})$$

$$= \prod_{k=1}^n P(w_k | w_1^{k-1})$$

سوال ۳-

الف) از آنجا که  $k=2$  است، در این مرحله فقط دو جمله داریم و هردو انتخاب میشوند.

ب) از آنجا که لگاریتم احتمالات را داریم، به جای ضرب، جمع میکنیم.

$$P(\text{neural neural}) = p(\text{neural}) + p(\text{neural} \mid \text{neural}) = -0.65 - 0.8 = -1.45$$

$$P(\text{neural network}) = p(\text{neural}) + p(\text{neural} \mid \text{network}) = -0.65 - 0.6 = -1.25$$

$$P(\text{network neural}) = p(\text{network}) + p(\text{network} \mid \text{neural}) = -0.73 - 0.6 = -1.33$$

$$P(\text{network network}) = p(\text{network}) + p(\text{network} \mid \text{network}) = -0.73 - 0.8 = -1.53$$

در نتیجه، جمله های neural network و network neural انتخاب میشوند.

ج)

$$P(\text{neural network neural}) = p(\text{neural network}) + p(\text{neural} \mid \text{neural network}) = -1.25 - 0.8 = -2.05$$

$$P(\text{neural network network}) = p(\text{neural network}) + p(\text{network} \mid \text{neural network}) = -1.25 - 0.6 = -1.85$$

$$P(\text{network neural neural}) = p(\text{network neural}) + p(\text{neural} \mid \text{network neural}) = -1.33 - 0.8 = -2.13$$

$$P(\text{network neural network}) = p(\text{network neural}) + p(\text{network} \mid \text{network neural}) = -1.33 - 0.6 = -1.93$$

پس در این مرحله جمله های neural network network و network neural network انتخاب میشوند.

د) در این مسئله، جستجوی beam توانست محتمل ترین جمله را پیدا کند. البته همیشه اینگونه نیست، زیرا در هر مرحله به اندازه سائز beam search محتمل ترین ها را انتخاب میکنیم ولی ممکن است از مواردی که حذف شده بودند، بعدا احتمال بالاتری نسبت به بقیه بدست آورند.

ث) اردر آن  $O(M k T \log(m*k))$  است. در کل  $T$  بار باید فرایند سورت کردن و انتخاب بهترین ها را انجام دهیم و در هر مرحله، از  $k$  گره،  $M$  گره دیگر بیرون میزنند. در نتیجه  $M k O(\text{sort})$  را خواهیم داشت. اگر الگوریتم سورت را  $O(n*\log n)$  بگیریم در نهایت به همان  $O(M k T \log(m*k))$  میرسیم.

سوال ۴-

(الف)

- اگر گیت های ورودی و خروجی را حذف کنیم، LSTM توانایی به روز رسانی حالت پنهان و کنترل خروجی خود را از دست می دهد.

- دروازه فراموشی به تنهایی برای مدیریت تمام جنبه های رفتار LSTM کافی نیست.

- این مدل کمتر انعطاف پذیر و تطبیق پذیر می شود و به طور بالقوه منجر به عملکرد کمتر از حد مطلوب می شود.

- در عمل، گیت های ورودی و خروجی نقش مهمی در این امکان ایفا می کنند که به LSTM اجازه می دهد یاد بگیرد که چه زمانی حالت پنهان خود را به روز کند، چه زمانی آن را بازنشانی کند، و چگونه اطلاعات جدید را ترکیب کند.

به طور خلاصه، حذف گیت های ورودی و خروجی از یک LSTM به طور قابل توجهی قابلیت های آن را محدود می کند. دروازه فراموشی به تنهایی نمی تواند به طور کامل از دست دادن این گیت های دیگر را جبران کند و احتمالاً عملکرد مدل آسیب خواهد دید. LSTM ها به عنوان یک سیستم کامل طراحی شده اند و هر سه گیت با هم کار می کنند تا چالش های یادگیری وابستگی های طولانی مدت و مدیریت شیب های در حال از بین رفتن را برطرف کنند.

(ب)

- با تنظیم دروازه فراموشی روی صفر، LSTM مانند یک سیستم حافظه ثابت رفتار می کند.

- از ورودی های جدید یاد نمی گیرد یا وضعیت داخلی خود را تنظیم نمی کند.

- توانایی شبکه برای یادگیری الگوها، توالی ها و وابستگی های پیچیده به شدت محدود خواهد بود.

- پیش بینی ها احتمالاً نادرست خواهند بود، زیرا مدل نمی تواند خود را با زمینه های متغیر تطبیق دهد یا وضعیت حافظه خود را به روزرسانی کند

به طور خلاصه، اگر دروازه فراموشی همیشه ۱ باشد (هرگز فراموش نشود) و گیت ورودی همیشه ۰ باشد (هرگز به روز نمی شود)، وضعیت داخلی سلول حافظه LSTM ثابت می ماند و مانع از توانایی های یادگیری و پیش بینی آن می شود. انعطاف پذیری ارائه شده توسط گیت های ورودی و فراموشی برای عملکرد مؤثر LSTM بسیار مهم است.

(ج)

مزایای بیشتر کردن لایه های LSTM:

- این کار باعث میشود مدل ما قدرت بیشتری برای بازنمایی دیتا پیدا کند. شبکه های عمیقتر میتوانند ویژگیهای پیچیده تر و دقیقتری را یاد بگیرند.
- یادگیری ویژگیهای سلسله مراتبی نیز یکی از مزایای افزایش تعداد لایه های LSTM میباشد. در این حالت، هر لایه میتواند روی سطح خاصی از Abstraction تمرکز کند.
- تعداد لایه های بیشتر باعث میشود مدل بتواند sequence های طولانی تری را بازنمایی کند.

معایب:

- تعداد زیاد لایه ها ممکن است باعث شود مدل دیتای آموزشی را حفظ کند. به این پدیده بیش برازش گفته میشود. این مورد باید کنترل شود.
- تعداد بیشتر لایه ها به معنای منابع مصرفی بیشتر است. زمان آموزش و inference و همچنین فضای حافظه مورد نیاز بیشتر میشود.
- مشکل دیگر که ممکن است پیش بیاید، مسئله vanishing gradient است که وقتی مدل عمیقتر میشود، معمولاً پیش می آید و با روشهایی مثل skip connection جلوی آن را تا حدودی گرفت.
- شبکه های عمیقتر نیازمند دیتاست بزرگتری هستند که در بعضی مواقع امکان تهیه چنین دیتاستی را نداریم.

سوال ۵-

-i

## ۱. مزایای پنجره بازگشت بزرگتر:

- وابستگی‌های بلندمدت: پنجره بازگشت بزرگتر به مدل امکان می‌دهد تا وابستگی‌های بلندمدت در داده‌ها را درک کند. به عنوان مثال:
  - در پردازش زبان طبیعی، درک متن یک جمله اغلب نیاز به در نظر گرفتن کلماتی از چند جمله قبلی دارد.
  - در پیش‌بینی سری‌های زمانی مالی، قیمت‌های سهام ممکن است به وقوع پیوسته‌هایی که هفته‌ها یا ماه‌ها قبل رخ داده‌اند، وابسته باشند.
- بهبود حافظه: LSTM ها در حفظ حافظه در دنباله‌های طولانی عالی عمل می‌کنند. پنجره بزرگتر توانایی آن‌ها در به خاطر سپردن اطلاعات مرتبط از زمان‌های دورتر را افزایش می‌دهد.
- عمومی‌سازی بهتر: پنجره بازگشت بزرگتر به مدل کمک می‌کند تا با در نظر گرفتن متناسب بازه‌های زمانی، بهتر عمومی‌سازی کند.

## ۲. معایب یک پنجره بازگشت بزرگتر:

- پیچیدگی محاسباتی بیشتر:
  - پنجره‌های بازگشت بلندتر نیاز به پردازش مراحل زمانی بیشتری در طول آموزش و پیش‌بینی دارند.
  - زمان آموزش و استفاده از حافظه افزایش می‌یابد، به ویژه برای شبکه‌های عمیق.
- خطر بیش‌برازش:
  - اگر پنجره بیش‌تر از حد باشد، مدل ممکن است نویز یا الگوهای غیرمرتبط از داده‌های آموزش را حفظ کند.
  - بیش‌برازش در زمانی رخ می‌دهد که مدل در داده‌های آموزش عملکرد خوبی دارد، اما در داده‌های ناشناخته عملکرد ضعیفی دارد.
- تأخیر در تطبیق:
  - پنجره بازگشت بزرگتر تأخیر در تطبیق مدل با تغییرات اخیر در داده‌ها ایجاد می‌کند.