بسمه تعالی

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه اموزشی هوش مصنوعی

نام دانشجو: فاطمه وهابی

شماره دانشجویی: 4013614052

عنوان گزارش: تمرین سوم

نام درس: پردازش زبان طبیعی

نام استاد: دکتر حمیدرضا برادران کاشانی

نام حل تمرین: آقای امیرمسعود سلطانی

بهار 1402

فهرست

پرسش ها

برای نگاشت کلمات به فضای تعبیه به منظور اجرای بازی مبتنی بر واژگان، من پیشنهاد می‌کنم از مپینگ word2vec (E2) که در بخش 3.2.2 توضیح داده شده است استفاده کنیم.

نگاشت word2vec انتخاب مناسبی برای این کار است زیرا روابط معنایی بین کلمات را به تصویر می‌کشد و آن‌ها را به عنوان بردارهای عددی نشان می دهد. نزدیکی کلمات در اسناد مختلف را در نظر می‌گیرد و کلمات مشابه را با هم گروه بندی می‌کند که با الزامات بازی که در آن کلمات نیاز به فاصله معینی از یکدیگر دارند، همسو می‌شود.

برای محاسبه وزن‌ها و به دست آوردن بردارهای معادل برای واژگان هر کلمه با استفاده از word2vec می‌توانیم این مراحل را دنبال کنیم:

1. استفاده از یک شبکه عصبی با یک لایه پنهان خطی. تعداد نورون های ورودی باید برابر با تعداد ابعاد در بردارهای نمایش منفرد (E1) باشد.

2. تعداد نورون های لایه پنهان باید برابر با بعد بردار مپینگ (تتا) باشد.

3. شبکه عصبی را با استفاده از مجموعه داده خود آموزش می‌دهیم. در طول آموزش، شبکه دارای یک لایه خروجی با تابع فعال سازی Softmax خواهد بود.

4. فرآیند آموزش شامل بهینه‌سازی وزن شبکه بر اساس نزدیکی کلمات در یک پنجره در مقایسه با کلمه ورودی است. این به شبکه امکان می‌دهد تا احتمال وقوع یک کلمه خاص را در مجاورت سایر کلمات در سند یاد بگیرد.

5. پس از آموزش شبکه، می‌توانیم بردار مپینگ E2(Wi) را برای هر کلمه Wi در فرهنگ لغت بدست آوریم. این بردار از مقادیر تتا $[\theta_1, \ldots, \theta_h]$) تشکیل شده است که نشان دهنده کلمه در فضای تعبیه است.

برای محاسبه شباهت کسینوس بین دو بردار کلمه E2(Wi) و E2(Wj) ، می‌توانیم از فرمول زیر استفاده کنیم:

$$\text{Cosine similarity} = \frac{E2(Wi)E2(Wj)}{||E2(Wi)||\,|E2(Wj)|}$$

این فرمول حاصل ضرب نقطه‌ای را بین دو بردار محاسبه می‌کند و آن را بر حاصل ضرب نرم آنها تقسیم می‌کند. مقدار به دست آمده نشان‌دهنده شباهت کسینوس بین دو کلمه است که درجه شباهت یا نزدیکی آنها را در فضای تعبیه شده نشان می دهد.

ویژگی‌هایی که این روش مپینگ باید برای بازی مبتنی بر واژگان داشته باشد، عبارتند از:

1. بازنمایی معنایی: روش نگاشت باید معنای کلمات را به تصویر بکشد و آن‌ها را به عنوان بردار در فضای تعبیه نمایش دهد. این امکان مقایسه معنی‌دار و محاسبه فاصله بین کلمات را فراهم می‌کند.

2. آگاهی از Proximity: روش باید مجاورت کلمات در اسناد مختلف را در نظر بگیرد و کلمات مشابه را با هم گروه‌بندی کند. این تضمین می‌کند که کلمات ویژه پیشنهادی در بازی سطح مشخصی از رابطه معنایی داشته و معیارهای فاصله مورد نیاز را برآورده می‌کنند.

3. محاسبه فاصله: روش نگاشت باید امکان محاسبه کارآمد فواصل بین بردارهای کلمه در فضای تعبیه را فراهم کند. این امکان بررسی اینکه آیا فاصله بین کلمات خاص حداقل A است یا خیر و بررسی اینکه آیا کلمه پیشنهادی بازیکن با کلمات موجود فاصله کمتری از B دارد یا خیر را ممکن می‌سازد.

4. مقیاس پذیری: روش نگاشت باید مقیاس‌پذیر باشد تا حجم وسیعی از واژگان را مدیریت کند. با پیشرفت بازی و معرفی کلمات بیشتر، روش باید بتواند کلمات جدید را در خود جای دهد و یکپارچگی فضای تعبیه را حفظ کند.

با استفاده از نقشه (E2) word2vec و در نظر گرفتن این ویژگی‌ها، می‌توانیم یک بازی مبتنی بر واژگان را پیاده‌سازی کنیم که جفت کلمات را پیشنهاد می‌کند و در عین حال از فواصل و روابط مناسب بین آنها اطمینان می‌یابد.

کدزنی

کدها و تحلیل‌های این بخش در فایل ژوپیتر وجود دارد.

به دلیل به هم نریختن ترتیب گزارشات به انگلیسی نوشته شده است.

پرسش هایی از گذشته

تمرین صفرم

در ارزیابی مدل های زبانی، معیارهای مختلفی مانند دقت(accuracy) ، فراخوانی(recall) ، دقت پیش بینی (precision)، ماتریس درهم‌ریختگی(confusion matrix) ، و معیارهای مشتق شده از آن‌ها مانند F1 score مورد استفاده قرار می‌گیرند. اما به دلیل وجود تعادل بین دقت و فراخوانی در معیار F1 score ، این معیار به نسبت به سایر معیارها بهتر است.

در ادامه برخی از دلایل استفاده از معیار F1 score برای ارزیابی مدل های زبانی را بیان می‌کنم:

- **تعادل بین دقت و فراخوانی: معیار F1 score ، به دلیل استفاده از میانگین هارمونیک دقت و فراخوانی، به مدل ها کمک می کند تا در پیش بینی های خود، تعادلی مناسب بین دقت و فراخوانی داشته باشند. این معیار به خصوص برای مواردی که تعداد نمونه های هر کلاس متفاوت است، مناسب است.**
- **تمرکز بر روی کلاس‌های کمیتی: در مواردی که تعداد نمونه های هر کلاس متفاوت است، معیارهایی مانند دقت و فراخوانی می‌تواند به نمونه های کلاس‌های کمیتی بیشتر توجه نکند. اما با استفاده از معیار F1 score، تأثیر کلاس‌های کمیتی در معیار ارزیابی کاهش پیدا می‌کند.**
- **قابلیت تفسیر: معیار F1 score به راحتی قابل تفسیر است و برای ارزیابی عملکرد مدل های مختلف در مسائل مختلف، می‌توان از آن استفاده کرد.**

به طور کلی معیار F1 score به دلیل مزیت هایی که در مقایسه با سایر معیارها دارد، به عنوان یکی از معیارهای مهم در ارزیابی مدل های زبانی مورد استفاده قرار می‌گیرد.

تمرین اول

محدودیت‌های مدل‌های زبانی شامل چالش‌های مختلفی هستند که می‌توان به برخی از آن‌ها اشاره کرد. این محدودیت‌ها عبارتند از:

وابستگی به لغات قبلی: مدل‌های زبانی برای تولید هر کلمه بعدی به لغات قبلی وابستگی دارند و برای تولید هر جمله باید لغات قبلی را با دقت کامل وارد کنیم. این محدودیت باعث می‌شود که نمی‌توانیم به صورت همزمان چند جمله مستقل را تولید کنیم.

مشکل در پردازش متون بلند: با افزایش تعداد کلمات و طول متن، مدل‌های زبانی با مشکلاتی مواجه می‌شوند. همچنین، تعداد پارامترهای مدل و حافظه مورد نیاز برای آموزش و استفاده از مدل نیز با طول متن افزایش می‌یابد.

مشکل در تفسیر معنای جملات: مدل‌های زبانی فقط بر پایه آمار و احتمالات متن، بدون درک واقعیت‌های جهانی، متن را تولید می‌کنند. بنابراین، در برخی موارد، تولیدات مدل‌های زبانی به نظر غیرمعقول یا نامناسب برسد.

مشکل در حفظ پیوستگی و همسانی: مدل‌های زبانی به صورت کلمه به کلمه جملات را تولید می‌کنند و این می‌تواند منجر به عدم پیوستگی و همسانی جملات شود. در برخی مدل‌های مبتنی بر Transformer سعی در رفع مشکلات حفظ پیوستگی و همسانی داشته باشند، اما همچنان ممکن است در برخی موارد جملات تولید شده ناهمسان باشند یا به نظر غیرطبیعی برسند.

وابستگی به داده آموزش: مدل‌های زبانی وابستگی قوی به داده آموزشی دارند و اگر داده آموزشی ناکافی، ناهمگن یا تراکم کمتری داشته باشد، مدل‌های زبانی ممکن است در تولید متن دچار مشکل شوند.

عدم درک مفهوم و منطق: مدل‌های زبانی بر اساس آمار و احتمالات کار می‌کنند و نیاز به درک واقعیت‌های جهانی ندارند. بنابراین، ممکن است در تولید متن‌های پرتکرار یا نامعقول، خطاهایی داشتباهی نوشته شوند.

این محدودیت‌ها نشان می‌دهند که هنوز مدل‌های زبانی به تمامی چالش‌های تولید متن پاسخ نداده‌اند و نیاز به توسعه و بهبود دارند.

تمرین دوم

1. برای اثبات عدم دوبخشی یک گراف، باید نشان داد که آن را نمی‌توان به دو مجموعه صحیح تقسیم کرد به‌گونه‌ای که گره‌های هر مجموعه با هم یال نداشته باشند. در صورتی که بتوان یک مورد پیدا کرد که این شرط را نقض کند، می‌توان نتیجه گرفت که گراف دوبخشی نیست.

2. در صورت برداشتن محدودیت دوبخشی، می‌توان برای هر گره، درجه آن را حساب کرد و این اطلاعات را به گراف بیفزاییم. همچنین، اگر دورهای مختلف در گراف را شمارش کرد می‌توان میزان پیچیدگی گراف را بیان کرد و همینطور اینکه بررسی کنیم دارای ساختارهای خاصی مانند دورهای فرد یا زوج است. همچنین، گر اتصالات بین گره‌های مختلف را بررسی کنیم، می‌توانیم با این اطلاعات، روابط بین گره‌ها را تحلیل کنیم. اطلاعاتی مانند وجود یا عدم وجود رابطه وابستگی بین برخی گره‌ها را با این تحلیل به دست می‌آوریم.

4- همان‌طور که در شکل مشاهده می‌کنید معماری های این دو متفاوت است. در ادامه بیش‌تر توضیح می‌دهیم.

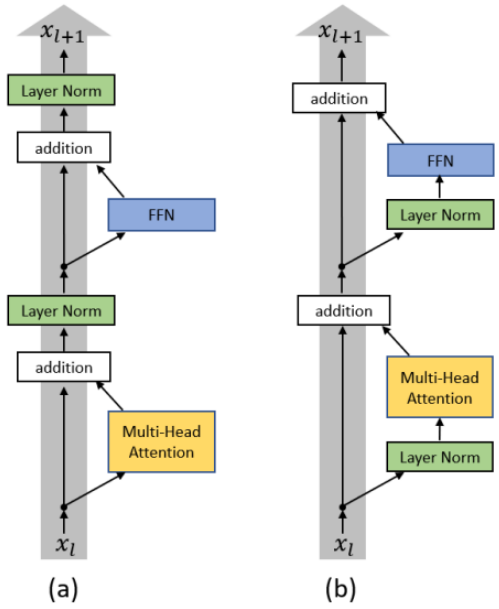

Figure 1 : (a)Post-LN Transformer layer; (b) Pre-LN Transformer layer

در معماری Pre-LN ، لایه نرمال‌سازی قبل از لایه‌های فیدفوروارد قرار می‌گیرد. به عبارت دیگر، هر لایه ابتدا لایه نرمال‌سازی را اعمال کرده و سپس فیدفوروارد را اجرا می‌کند. این روش می‌تواند جریان اطلاعات را در معماری ترنسفورمر بهبود بخشد، زیرا لایه نرمال‌سازی می‌تواند تأثیر بیشتری در جریان اطلاعات بگذارد. به عبارتی این بهبود جریان اطلاعات در معماری Pre-LN ناشی از فرایند نرمال‌سازی است که قبل از لایه‌های فیدفوروارد اجرا می‌شود. با اعمال لایه نرمال‌سازی پیش از فیدفوروارد، تغییرات وارد شده به داده در هر لایه تأثیر کمتری بر جریان اطلاعات خواهد داشت، که این امر می‌تواند به عملکرد بهتر و سریع‌تر شبکه ترنسفورمر منجر شود.

در مقابل، در معماری Post-LN ، لایه نرمال‌سازی پس از لایه‌های فیدفوروارد قرار می‌گیرد. به این ترتیب، هر لایه ابتدا فیدفوروارد را اجرا کرده و سپس لایه نرمال‌سازی را اعمال می‌کند. در این حالت، تأثیر لایه نرمال‌سازی بر جریان اطلاعات کمتر است و احتمالاً عملکرد مدل ضعیف‌تر خواهد بود.

اثربخشی هر روش به مسئله و مجموعه داده مورد استفاده وابسته است. مطالعات مختلف نشان داده‌اند که Pre-LN در برخی موارد عملکرد بهتری نسبت به Post-LN دارد، در حالی که در موارد دیگر، روش Post-LN عملکرد بهتری از خود نشان می‌دهد. بنابراین، برای تحلیل دقیق تفاوت و تأثیر هر روش، نیاز به آزمایش‌های بیشتر و مطالعه دقیق‌تر وجود دارد تا بتوانیم اطلاعات دقیق‌تری را از هر دو بدهیم.