

گزارش تمرین هفتم درس رایانش عصبی و یادگیری عمیق

تمرین هشتم:

آشنایی با شبکه‌های ترنسفورمر

نام دانشجو: مجید ادیبیان

شماره دانشجویی: ۴۰۰۱۳۱۰۷۸

نام استاد درس: دکتر صفابخش

زمستان ۱۴۰۱

بخش اول (سوالات تئوری):

الف) در هنگامی که یک پیش‌بینی به ازای یک دنباله ورودی انجام می‌دهیم ممکن است قسمت‌های مختلفی از آن دنباله برای ما اهمیت‌های متفاوتی داشته باشند که این موضوع به این معناست که نباید برای پیش‌بینی مناسب به تمام دنباله موجود به یک مقدار اهمیت داد. مکانیزم توجه این امکان را فراهم می‌کند که قسمت‌های پراهمیت‌تر در هر لحظه از پیش‌بینی مشخص شود و از اطلاعات آن‌ها در پیش‌بینی مورد نظر بیشتر استفاده شود. برای پیاده‌سازی این موضوع میزان شباهت بردار مورد نظر در پیش‌بینی با بردار هر یک از قسمت‌های دنباله ورودی (مثلاً هر کلمه از متن ورودی) بررسی می‌شود هرچه میزان این شباهت بیشتر بود ضریب بزرگتری برای آن بردار در نظر گرفته می‌شود. در نهایت بر اساس این ضرایب بردارهای موجود جمع وزن‌دار شده و در پیش‌بینی استفاده می‌شوند.

ب) در self-attention تنها یک دنباله از بردارها داریم و برای یافتن بردار تعبیه‌ای مناسب برای هر یک از مکانیزم توجه آن با سایر بردارهای اطراف در همان دنباله استفاده می‌کنیم و در نتیجه بردار کوثری بردار مورد نظر است و بردارهای value و key از بردارهای اطراف آن در همان دنباله به دست می‌آیند. در حال که در cross-attention هر بار میزان توجه یک بردار از یک دنباله بردار را با هر یک از بردارهای دنباله بردار دیگر می‌یابیم که در نتیجه بردار query همان بردار مورد نظر از دنباله اول است و بردارهای key و value بردارهای حاصل از دنباله بردار دنباله دوم است.

ج) در مدل BERT هدف آن است که با استفاده از تنها قسمت کدگذار از مدل ترنسفورمر مدلی ساخته شود که بتواند با استفاده از حجم زیادی داده متنی بدون برچسب بردارهای تعبیه مناسب برای کلمه و همچنین کل جمله را تولید کند. برای انجام این کار معماری کلی استفاده شده همان معماری قسمت کدگذار در ترنسفورمر است که البته تعداد لایه‌های کدگذار در آن چند حالت مختلف در نظر گرفته شده تا مدل‌های از تعداد پارامترهای کم تا زیاد ساخته شود. یکی از ایده‌های مدل BERT قسمت توکن‌بندی در آن است که از Word Piece استفاده شده است. این روش باعث می‌شود که هر کلمه‌ای را توکن‌بندی کرده و اگر کلمه شناخته شده نیست آن را به قسمت‌های کوچک‌تر می‌شکند تا به قسمت‌های شناخته‌شده برسد. ایده دیگر این مدل استفاده از حجم زیاد داده متنی بدون برچسب است. روش آموزش بر روی داده‌های بدون برچسب در آن به دو روش پوشاندن کلمه و پیش‌بینی جمله بعدی است. به این صورت که دو جمله از پیکره متنی مورد استفاده انتخاب می‌شود و در کنار هم قرار می‌گیرند و برخی از کلمات این جملات mask می‌شود و سپس سعی می‌شود در خروجی مدل کلمات mask شده پیش‌بینی شود و تشخیص داده شود که دو جمله استفاده شده دو جمله متوالی هستند یا خیر که

برای اولی از بردار خروجی مربوط به کلمه mask شده استفاده می‌شود و برای دومی از خروجی بردار مربوط به توکن CLS که در ابتدای کل دنباله اضافه شده است و اطلاعات کل دو جمله را در خود دارد. ایده دیگر این مدل استفاده از کدگذاری مکانی در کنار کدگذاری جمله است. همان‌طور که گفته شد دو جمله در کنار هم به مدل داده می‌شود و برای هر جمله یک عدد اختصاص می‌یابد تا با بردار آن جمع زده شود و مشخص کند که هر بردار مربوط به کدام جمله است.

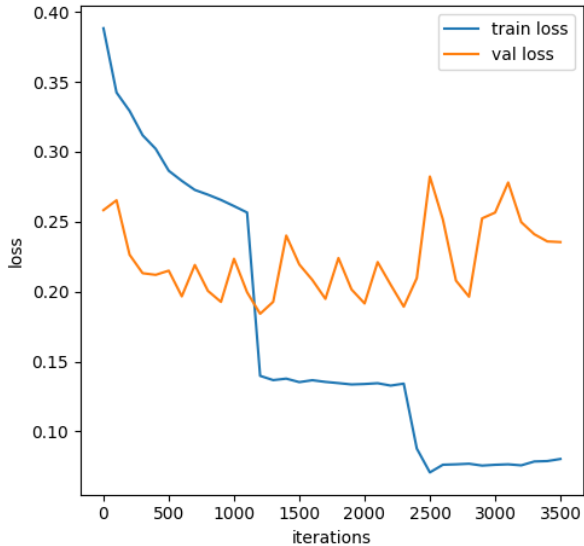
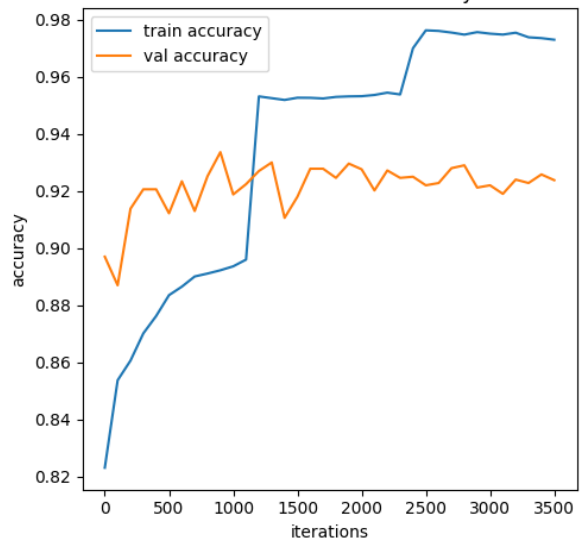
بخش دوم (پیاده‌سازی):

د) در ابتدا مانند فرایند انجام شده در تمرین ۶ ابتدا داده‌های IMDB دانلود شده و پیش‌پردازش می‌شود که همان‌طور که در تمرین ۶ توضیح داده شد این پیش‌پردازش‌ها شامل حذف لینک‌ها و علائم نگارشی و کوچک کردن تمام حروف متن‌ها است. سپس متن پیش‌پردازش شده باید با استفاده از tokenizer مدل BERT توکن‌بندی شود که برای این کار ابتدا tokenizer این مدل دانلود شده و سپس مجموعه متن‌های پیش‌پردازش شده به آن داده شده است تا دنباله آیدی مربوط به هر یک از این متن‌ها را تولید کند. در این فرایند حداکثر طول جمله‌ها ۵۰۰ در نظر گرفته شده است چرا که همانند تمرین ۶ با بررسی که شده بود مشخص شد که طول‌های بیشتر از آن، تعداد بسیار کمی دارند و از طرفی مدل BERT می‌تواند تا ۵۱۲ توکن ورودی را بگیرد.

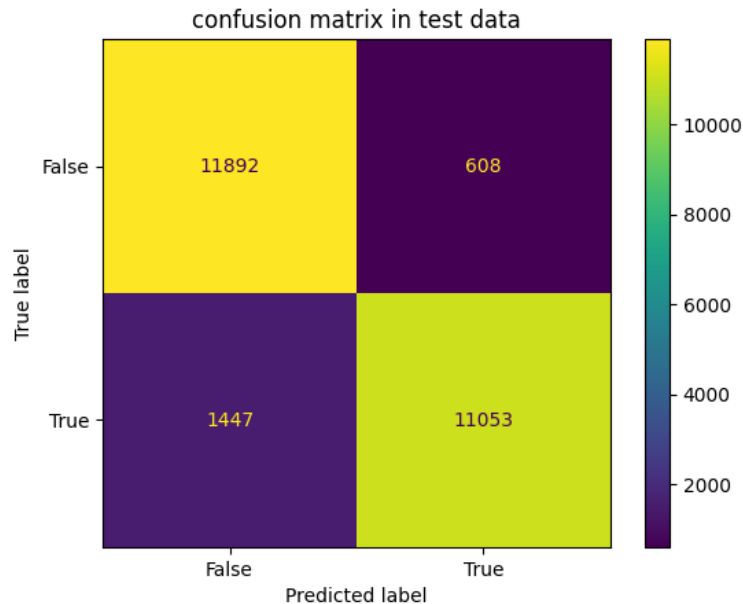
در ادامه داده‌های آموزش به دو قسمت آموزشی و ارزیابی تقسیم شده‌اند که در نتیجه این کار تعداد ۲۰۰۰۰ نظر برای آموزش و ۵۰۰۰ نظر برای ارزیابی و ۲۵۰۰۰ نظر برای تست وجود خواهد داشت. که پس از فرایند گفته شده در بالا در batch ها ۱۶ تایی برای آموزش مدل قرار می‌گیرند.

سپس مدلی ساخته شده است که ابتدای آن BERT است و سپس خروجی نوکن CLS که اطلاعات کل متن ورودی را در خود دارد را به طور متوالی به دو لایه خطی می‌دهیم که در نهایت ابعاد این بردار را از ۷۶۸ به ۱ می‌رساند و بر روی آن یک تابع Sigmoid اعمال می‌کنیم. برای آموزش مدل هم از تابع بهینه‌سازی آدام با نرخ یادگیری $5e-5$ استفاده شده است و مدل تا ۳ اپاک آموزش دیده و پس از هر ۱۰۰ قدم آموزش ارزیابی مدل بر روی داده‌های ارزیابی انجام می‌شود و مقدار خطا و دقت بر روی داده آموزشی و ارزیابی ثبت می‌شود.

خروجی نمودارهای تغییرات خطا و دقت بر روی داده‌های آموزشی و ارزیابی به صورت زیر است:

داده آموزشی و ارزیابی	
<p>train and validation loss</p> 	تغییرات خطا
<p>train and validation accuracy</p> 	تغییرات دقت

پس از آموزش مدل تا ۳ ایپاک، از داده‌های آزمون استفاده می‌کنیم که دیده می‌شود مدل بر روی داده‌های آزمون به دقت ۰.۹۲ رسیده است. نمودار درهم‌ریختگی حاصل از پیش‌بینی مدل بر روی داده‌های آزمون به صورت زیر است:



تحلیل نتایج:

دیده می‌شود که دقت مدل بر روی داده‌های آموزشی حدود ۹۸ درصد و بر روی داده‌های آزمون و ارزیابی برابر ۹۲ درصد است که دقت قابل قبولی می‌باشد. همچنین با بررسی ماتریس درهم‌ریختگی حاصل از داده‌های آزمون دیده می‌شود که در داده‌هایی که به اشتباه پیش‌بینی شده‌اند درصد بیشتری داده‌هایی هستند که نظر مثبت بوده‌اند ولی مدل آن را منفی پیش‌بینی کرده است. همچنین دیده می‌شود که به دلیل پارامترهای زیاد مدل و پیش‌آموزش دیده بودن مدل تنها بعد از ۳ اپاک مدل توانسته به دقت بالایی برسد.

استفاده از ترنسفورمر در تسک‌های مختلف: (امتیازی)

برای نشان دادن کاربر شبکه ترنسفورمر در تسک‌های مختلف سه مدل معروف و پرکاربرد در این حوزه در نظر گرفته شده است و در هر یک تسک متفاوتی اجرا شده است.

(۱) در ابتدا از مدل BERT استفاده شده که برای تسک پرسش و پاسخ در زبان فارسی fine tune شده است. در اینجا تنها مدل آموزش دیده استفاده شده است و یک نمونه پرسش و پاسخ به عنوان خروجی تولید شده است:

محتوا: والیبالیست یک ورزش گروهی و ششمین ورزش پر طرفدار و گسترده‌ترین ورزش در بعضی کشورهای جهان است که در آن بازیکنان در دو تیم شش نفره، در دو سوی توری قرار می‌گیرند و تلاش می‌کنند تا طبق قوانین

بازی، توپ را از روی تور در زمین تیم مقابل فرود آورند. طول زمین والیبال ۱۸ متر و عرض آن ۹ است. هر تیم حداقل باید ۳ ست (یا دست) از ۵ پنج ست بازی را ببرد، تا بتواند پیروز مسابقه شود.

پرسش: هر تیم فوتبال چگونه می‌تواند پیروز مسابقه شود؟

پاسخ: حداقل باید ۳ ست (یا دست) از ۵ پنج ست بازی را ببرد.

۲) سپس از مدل T5 استفاده شده هم قسمت کدگذار و هم کدگشا در ترنسفورمر را دارد و مدل فعلی استفاده شده برای تسک خلاصه‌سازی fine tune شده است. مانند قبل تنها مدل پیش‌آموزش دیده این مدل استفاده شده و یک نمونه خلاصه‌سازی در آن تولید شده است که نتیجه آن را در زیر می‌بینیم (به دلیل بزرگ بودن متنی که از باید خلاصه می‌شد، این متندر این قسمت نیامده):

text: ...

summary: on Tuesday at a private funeral in Houston. Floyd, who was 46, will be buried next to his mother's grave. A Minnesota police officer was caught on video pressing his knee into Mr. Floyd's neck for nearly nine minutes before his death. The officer has been charged with second-degree manslaughter and his bail was set at \$1.25 million. Floyd's last words — "I can't breathe" — were a rallying cry. He was accused of killing himself.

۳) در انتها مدل GPT استفاده شده است که پیاده‌سازی قسمت کدگشا در ترنسفورمر است و مدل فعلی استفاده شده برای تولید متن در زبان فارسی استفاده شده است. به طوری که قسمتی از متن به مدل داده می‌شود و مدل ادامه آن را تولید می‌کند. در زیر نمونه خروجی این مدل را می‌بینیم:

متن ورودی: در یک اتفاق شگفت انگیز، پژوهشگران به این نتیجه رسیدند که در زندگی...

متن خروجی: در یک اتفاق شگفت انگیز، پژوهشگران به این نتیجه رسیدند که در زندگی واقعی نمی‌توانید تنها با کمک کردن به یکدیگر، با انسان‌ها مهربان باشید.