

شبکه‌های عصبی و یادگیری عمیق

دکتر صفابخش



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر

رضا آدینه پور ۴۰۲۱۳۱۰۵۵

تمرین هفتم
شبکه Transformer

۱۹ تیر ۱۴۰۳



سوال اول - تئوری

یکی از دلایل نیاز به مکانیزم توجه، گلوگاهی بود که بین رمزگذار و رمزگشا در مدل های seq2seq به وجود می‌آمد. این مشکل را توضیح دهید و نشان دهید چطور مکانیزم توجه این مشکل را حل کرد. یکی دیگر از مشکلات، عدم توجه مدل به گذشته دور بود. به طور مثال در یک متن به کلمات نزدیک‌تر اهمیت بیشتری داده می‌شد تا کلمات دورتر و وزن کلمات دورتر به صورت نمایی کاهش پیدا می‌کرد. آیا استفاده از lstm و یا lstm دوطرفه می‌تواند این مشکل را به طور کامل رفع کند؟ توضیح دهید.

پاسخ

در مدل‌های Seq2Seq، انکودر دنباله ورودی را پردازش کرده و آن را به یک بردار متنی با طول ثابت تبدیل می‌کند. این بردار متنی باید تمام اطلاعات مربوط به دنباله ورودی را در خود ذخیره کند. سپس دیکودر این بردار با طول ثابت را می‌گیرد و دنباله خروجی را تولید می‌کند. برای دنباله‌های ورودی طولانی، فشرده‌سازی تمام اطلاعات به یک بردار با طول ثابت دشوار است. این منجر به از دست رفتن اطلاعات می‌شود، بردار متنی با طول ثابت ممکن است نتواند تمام جزئیات لازم برای تولید دنباله خروجی منسجم و دقیق را ذخیره کند همین موضوع به عنوان یکی از چالش‌ها م مشکلات مدل‌های RNN مطرح می‌شود.

مکانیزم توجه برای کاهش مشکل گلوگاه در شبکه‌های RNN معرفی شد. مکانیزم توجه بر خلاف روش‌های قبلی، به جای اتکا به یک بردار متنی با طول ثابت، به دیکودر این اجازه را می‌دهد که برای هر خروجی یک بردار متنی پویا ایجاد کند که این بردار متنی پویا یک جمع وزنی از تمام وضعیت‌های پنهان (از گذشته‌های دور تا الان) انکودر است.

مسئله دیگری در مدل‌های Seq2Seq، به‌ویژه با RNN‌ها، دشواری در پردازش وابستگی‌های بلندمدت بود. RNN‌های سنتی و حتی LSTM‌ها تمایل دارند که به ورودی‌های جدید، بیشتر از ورودی‌های دورتر اهمیت دهند.

LSTM‌ها برای کاهش مشکل محو شدن گرادین طراحی شده‌اند که به ضبط وابستگی‌های طولانی‌تر نسبت به RNN‌های معمولی کمک می‌کند. با این حال، تأثیر ورودی‌های دورتر همچنان تمایل دارد که با گذشت زمان کاهش یابد، هرچند نه به اندازه‌ای که در RNN‌های استاندارد دیده می‌شود.

در BiLSTM‌ها دنباله را در هر دو جهت جلو و عقب پردازش می‌کنند و بنابراین اطلاعات را از هر دو زمینه گذشته و آینده فراهم می‌کنند. این رویکرد دوطرفه توانایی مدل را در ضبط وابستگی‌ها در هر دو جهت بهبود می‌بخشد. با این وجود، BiLSTM‌ها همچنان به بردارهای با طول ثابت متکی هستند و با وابستگی‌های بسیار طولانی مشکل دارند.

در عمل این موضوع به عنوان یکی از ضعف‌های این نوع شبکه‌ها محسوب می‌شود و شبکه Transformer و به‌ویژه مکانیزم توجه این مشکل را حل نموده و وابستگی‌های طولانی مدت را در دنباله سیگنال ورودی، بیشتر از سایر شبکه‌ها درک می‌کند.

مکانیزم توجه به دیکودر اجازه می‌دهد تا به هر قسمت از دنباله ورودی به‌طور مستقیم دسترسی داشته باشد، بدون توجه به موقعیت آن. این دسترسی مستقیم به این معنی است که ورودی دور نیز می‌تواند بر ورودی فعلی تأثیرگذار باشد.

سوال دوم - تئوری

در شبکه‌های بازگشتی ما می‌توانستیم از خروجی مرحله قبل در ورودی، تاریخچه و گذشته را مدل کنیم. اما با توجه به اینکه مدل‌های ترنسفورمر از شبکه‌های بازگشتی استفاده نمی‌کنند، چطور می‌توانند بهتر از شبکه‌های بازگشتی گذشته را در نظر بگیرند (نشان دهید). مشکلات ترنسفورمر را در مقایسه با شبکه‌های بازگشتی بیان کنید.

پاسخ

مدل‌های ارنسفرمر برخلاف شبکه‌های (RNN)، تمام توکن‌ها را به صورت موازی پردازش می‌کنند که منجر به افزایش قابل توجهی در کارایی محاسباتی، به ویژه در سخت‌افزارهای مدرن مانند GPU ها می‌شود. این پردازش موازی به ارنسفرمرها اجازه می‌دهد تا دنباله‌های طولانی را به طور موثرتری مدیریت کنند، زیرا از مشکل گلوگاه پردازش مرحله به مرحله در RNN اجتناب می‌شود. به طور دقیق این مکانیزم توجه است که این امکان را به ما می‌دهد تا برخلاف شبکه‌های RNN بدون هیچ کامپوننت بازگشتی و فیدبک‌ای اطلاعات و وابستگی‌های گذشته را درک کنیم و آن را مدل کنیم. اما در برابر همه این مزایا، ترنسفرمرها معایبی نیز دارند که در ادامه به معرفی چند مورد از آنها می‌پردازیم:

۱. حجم محاسبات و حافظه زیاد:

مکانیزم توجه خود دارای پیچیدگی زمانی و حافظه‌ای زیاد نسبت به طول دنباله است. این امر باعث می‌شود ترنسفرمرها برای دنباله‌های بسیار طولانی منابع زیادی مصرف کنند. در مقابل، RNN ها دارای پیچیدگی محاسباتی کمتر نسبت به ترنسفرمرها هستند که می‌تواند برای کاربردها کارآمدتر باشد.

۲. مدیریت دنباله‌های بسیار طولانی:

با اینکه ترنسفرمرها در درک وابستگی‌های طولانی بهتر هستند، عملکرد آنها با دنباله‌های بسیار طولانی می‌تواند به دلیل پیچیدگی بیش از حد به صورت توان دوم کاهش یابد.

۳. نیاز به داده‌های زیاد برای آموزش:

ترانسفورمرها به طور کلی به مقادیر زیادی از داده‌های آموزشی و منابع محاسباتی قابل توجهی برای آموزش موثر نیاز دارند. در مقابل، RNN ها می‌توانند از نظر داده‌ای کارآمدتر و ارزان‌تر برای آموزش باشند، و با مجموعه داده کوچکتری در مقایسه با ترنسفرمرها آموزش ببینند.

سوال سوم - تئوری

ترنسفورمرها نسبت به شبکه‌های seq2seq قابلیت موازی‌سازی بیشتری دارند. با ذکر جزئیات توضیح دهید.

پاسخ

مدل‌های Seq2Seq معمولاً از RNN یا LSTM استفاده می‌کنند که پردازش ترتیبی دارند. اما در مقابل ترنسفورمرها به صورت موازی داده‌ها را پردازش می‌کنند.

۱. در مدل‌های seq2seq

اگر فرض شود دنباله ای با طول n داریم، هر گام زمانی در RNN یا LSTM باید به ترتیب پردازش شود و برای هر گام زمانی، یک پردازش زمانی $O(1)$ انجام می‌شود. بنابراین، کل پیچیدگی زمانی برای پردازش توالی ورودی $O(n)$ است.

هر گام زمانی باید منتظر تکمیل گام قبلی باشد، بنابراین پردازش‌ها نمی‌توانند به‌طور موازی انجام شوند. در بهترین حالت، هر گام زمانی می‌تواند به‌طور موازی با پردازش داخلی خود (مثل محاسبات داخل سلول‌های LSTM) انجام شود، اما این قابلیت موازی‌سازی به‌طور کلی محدود است.

در مقابل در

۲. ترنسفورمرها

اگر مجدداً فرض شود توالی ای ورودی با طول n داریم، در لایه Self Attention، هر نشانه می‌تواند به تمام نشانه‌های دیگر در توالی توجه کند. این کار با محاسبه ماتریس توجه انجام می‌شود که پیچیدگی زمانی $O(n^2)$ دارد. محاسبات ماتریسی برای محاسبه توجه به‌طور موازی قابل انجام است. بنابراین، پیچیدگی زمانی یک لایه خود توجهی $O(n^2)$ است.

محاسبات ماتریسی در لایه Self Attention (مثل محاسبه ماتریس KQ^T برای توجه) به‌طور کامل به‌صورت موازی انجام می‌شوند.

سوال چهارم - تئوری

یکی از مشکلات ترنسفورمرها مرتبه هزینه محاسباتی و هزینه ذخیره‌سازی عملیات self-attention است که از مرتبه $O(N^2)$ می‌باشد. تلاش‌هایی برای کاهش این مشکل انجام شد. مقالاتی نشان دادند که عملکرد softmax باعث می‌شود تا بتوانیم این پیچیدگی را کاهش دهیم. توضیح دهید چرا عملکرد softmax باعث وجود این مسئله می‌شود. همچنین یکی از پیشنهادها برای حل این مشکل استفاده از مکانیزم‌های توجه کرنلی است. در مورد این مکانیزم تحقیق کنید و نشان دهید چطور این روش منجر به کاهش پیچیدگی می‌شود. یک کرنل به دلخواه انتخاب کنید و عبارت «۱» را بازنویسی کنید و مرتبه زمانی و حافظه مورد نیاز برای عملکرد self-attention را محاسبه کنید. لطفاً به مقاله که برای انتخاب کرنل مراجعه کردید، ارجاع دهید.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

سوال پنجم - عملی