

شبکه‌های عصبی و یادگیری عمیق دکتر صفا بخش



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر

رضا آدینه پور ۴۰۲۱۳۱۰۵۵

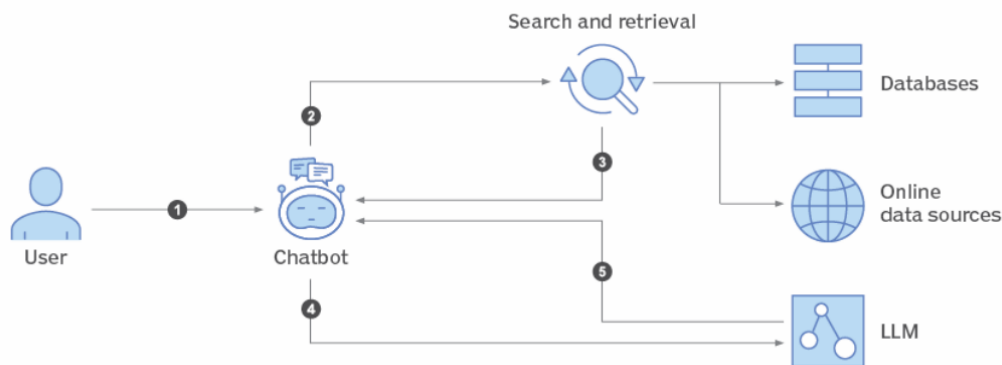
تمرین سوم
شبکه خودسازمانده (SOM)

۲ اردیبهشت ۱۴۰۳

سوال اول - عملی نظری

برای آموزش مدل‌های زبانی بزرگ (Large Language Model) که حاوی میلیون‌ها و میلیارد‌ها پارامتر هستند، از حجم قابل توجهی داده استفاده می‌شود. اما در تمامی این مدل‌ها یک تاریخ قطع آموزش وجود دارد که مدل زبانی هیچ اطلاعاتی در خصوص داده‌های تولید شده پس از این زمان ندارد. به عنوان مثال، تاریخ قطع آموزش مدل GPT-3.5-turbo-instruction سپتامبر ۲۰۲۱ است و از همین رو این مدل ممکن است به سوالات مربوط به رویدادهای سال ۲۰۲۲، ۲۰۲۳ و ۲۰۲۴ پاسخ صحیح ندهد. چنین داده‌هایی که بعد از تاریخ قطع آموزش تولید شده‌اند و یا بخشی از داده‌ی آموزشی اولیه‌ی مدل زبانی نیستند را داده‌ی خارجی می‌گوییم. تکنیک تولید تقویت شده با بازیابی (RAG) رویکردی است که با استخراج داده‌ی خارجی متناسب با فرمان، دریافت شده و افزودن آن به عنوان ورودی به مدل زبانی تلاش می‌کند که فرمان ورودی را تقویت کرده و به مدل زبانی کمک می‌کند تا جواب مرتبط و متناسبی بسازد. به عنوان مثال در پاسخ به یک فرمان متنی مانند «چه کسی شرکت توییتر را در سال ۲۰۲۲ خرید؟» تمامی داده‌های خارجی متناسب با این فرمان را استخراج می‌کند و آن‌ها را به عنوان ورودی به مدل زبانی GPT-3.5-turbo-instruct می‌دهد تا مدل زبانی بتواند با دانش دریافت شده پاسخ متناسبی تولید کند. این رویکرد نیاز به آموزش مجدد و با بازتنظیم (Fine tune) مدل زبانی را برطرف می‌سازد. در این پروژه می‌خواهیم با استفاده از شبکه‌های خودسازمان‌ده این تکنیک را پیاده‌سازی کنیم.

How an LLM using RAG works



شکل ۱: فرآیند کلی RAG در یک مدل زبانی بزرگ

وظیفه اصلی RAG جست‌وجو معنایی (Semantic search) در پایگاه داده‌های اطلاعاتی و بازیابی اطلاعات خارجی دارای تناسب محتوایی با فرمان داده‌شده به یک مدل زبانی است. برای تسهیل جست‌وجوی معنایی، ابتدا داده‌های خارجی استخراج شده به بازنمایی‌های عددی یا برداری تبدیل می‌شوند که به این بازنمایی، تعبیه‌ی متن (Text embedding) می‌گوییم. در زمان بازیابی نیز ابتدا فرمان متنی به بازنمایی برداری تبدیل می‌شود و سپس نزدیک‌ترین بردارهای داده‌ی خارجی متناسب با آن استخراج می‌شود. شکل «۱» دیگرام کلی این فرآیند را نشان می‌دهد. چالش اصلی این رویکرد این است که جست‌وجوی معنایی ذکر شده به دلیل نیازمندی به محاسبه‌ی فاصله‌ی بردار فرمان با حجم عظیمی از بردارهای داده‌ی خارجی، به منابع پردازشی و

محاسباتی زیاد و زمان قابل توجهی نیاز دارد. بنابر این پیدا کردن رویکردی که جست‌وجوی معنایی را به‌صورت کارا انجام دهد بسیار حائز اهمیت است.

برای افزایش کارایی جست‌وجو معنایی، یک رویکرد رایج این است که بردارهای داده‌های خارجی را خوشه‌بندی کنیم و در زمان جست‌وجو نیز ابتدا خوشه مشابه با بردار فرمان ورودی را پیدا می‌کنیم و سپس شباهت بردارهای داده‌های خارجی متعلق به آن خوشه با بردار فرمان را محاسبه می‌کنیم و اگر شباهت بردارها از یک آستانه بیشتر باشد، آنها را به‌عنوان اطلاعات مرتبط در نظر می‌گیریم.

۱. در این پروژه قصد داریم برای خوشه‌بندی داده‌های خارجی از شبکه خودسازمان‌ده استفاده کنیم. بررسی کنید که در این شبکه‌ها نسبت به سایر روش‌های خوشه‌بندی که در یادگیری ماشین به‌کار گرفته می‌شود، چه مزایا و معایبی دارد؟ به نظر شما، چرا استفاده از شبکه خودسازمان‌ده به صورت با نظارت صورت نمی‌گیرد؟ فرآیند یادگیری این مدل‌ها را توضیح دهید.

پاسخ

قبل از بررسی مزایا و معایب شبکه SOM نیاز است که یک سری پیش‌نیازها را توضیح دهیم. پیش از هر چیزی ابتدا می‌بایست انواع الگوریتم‌های یادگیری ماشین و دلیل استفاده از آنها را توضیح دهیم. الگوریتم‌های یادگیری ماشین به ۳ دسته مختلف تقسیم می‌شوند:

(آ) یادگیری با نظارت (Supervised Learning)

(ب) یادگیری نیمه نظارتی (Semi-supervised Learning)

(ج) یادگیری بدون نظارت (Unsupervised Learning)

در یادگیری با نظارت، داده و لیبل‌های متناظر با آنها را داریم. در یادگیری نیمه نظارتی، صرفاً بخشی از داده‌ها لیبل دارند و لیبل بقیه داده‌ها مشخص نیست. دسته آخر که مورد بحث ماست، یادگیری بدون نظارت است که داده‌های موجود، لیبل ندارند و به ازای داده‌های مختلف، خروجی مناسب را نمی‌دانیم و از الگوهای پنهان در داده‌ها اطلاعی نداریم. در این صورت است که به سمت الگوریتم‌های بدون نظارت می‌آیم تا به الگوریتم این اجازه را بدهیم که هرچه را می‌تواند یاد بگیرد و اطلاعات پنهان در داده‌ها را مشخص کند. الگوریتم‌های خوشه‌بندی در این دسته قرار می‌گیرند و دلیل قرارگیری در این دسته آن است که ما هیچ اطلاعاتی در مورد داده‌های ورودی نداریم و به دنبال ایجاد وابستگی میان آنها هستیم. الگوریتم‌های خوشه‌بندی این امکان را برای ما فراهم می‌سازد تا داده‌های شبیه به هم را در یک دسته قرار دهد. در این باره در صفحه ۱۴۱ [۱] گفته شده است:

« تکنیک‌های خوشه‌بندی زمانی اعمال می‌شوند که کلاسی برای پیش‌بینی وجود نداشته باشد، بلکه زمانی که نمونه‌ها باید به گروه‌های طبیعی تقسیم شوند، اعمال می‌شوند. »

پس اگر با داده‌هایی مواجه بودیم که اطلاعاتی در مورد آنها نمی‌دانیم، خوشه‌یابی بهترین روش برای درک وابستگی‌ها میان داده‌هاست. الگوریتم‌های خوشه‌یابی را می‌توان به‌صورت زیر دسته‌بندی کرد:

- (a) Density-based
- (b) Distribution-based
- (c) Centroid-based
- (d) Hierarchical-based

در الگوریتم‌های خوشه‌یابی مبتنی بر چگالی، داده‌ها بر اساس تراکم و غلظت داده‌ها در نقاط مختلف تقسیم‌بندی می‌شود.

پاسخ

در خوشه‌یابی توزیع شده، اساس خوشه‌یابی به صورت احتمالی است. یعنی برای تمام نقاط یک احتمال تعلق به یک خوشه خاص در نظر گرفته می‌شود که با دور شدن داده از مرکز آن خوشه، احتمال تعلق داده به خوشه مربوطه کاهش پیدا می‌کند.

پراکندگی و سریع‌ترین نوع خوشه‌یابی، خوشه‌یابی Centroid است. این الگوریتم نقطه‌ها را بر اساس چندین مرکز در داده‌ها جدا می‌کند و هر نقطه بر اساس مجذور فاصله‌اش تا مرکز داده به یک خوشه اختصاص می‌یابد. استفاده از خوشه‌بندی سلسله‌مراتبی محدود تر از سایر روش‌هاست. بدین صورت است که برای داده‌هایی که ذاتاً به صورت سلسله‌مراتبی هستند استفاده می‌شود. مانند داده‌های مربوط به یک پایگاه داده.

*

References

- [1] Data Mining: Practical Machine Learning Tools and Techniques, 2016.
- [2] Zeng GL. A deep-network piecewise linear approximation formula. IEEE Access. 2021 Aug 31;9:120665-74.

۲. مجموعه داده ارائه شده در این پروژه شامل رویدادهای سه سال متوالی از ۲۰۲۲ تا ۲۰۲۴ است که از سایت ویکی‌پدیا جمع‌آوری شده است. داده‌ی مربوطه را بارگزاری کنید و پیش‌پردازش‌های متنی شامل حذف کلمات ایست (Stop word)، واحدسازی کلمات (Tokenization) و تبدیل به بردارهای GloVe را روی آن انجام دهید.

۳. پارامترهای ورودی مدل minisom را توضیح دهید. پارامترهای شبکه خودسازمان‌ده خود را تنظیم کنید و شبکه را بر روی داده‌های مربوطه آموزش دهید. (مقادیر تمامی پارامترها را در گزارش خود اضافه کنید). سپس به‌ازای هر داده‌ی ورودی واحد، منطبق (Best matching unit) با آن را به‌دست آورید و به‌عنوان نمایه‌ی داده‌ی مربوطه ذخیره کنید.

۴. برای ۵۰ رویداد که به صورت تصادفی از مجموعه داده انتخاب شده‌اند، نقشه خروجی را رسم کنید. نقشه‌ی به‌دست آمده را تفسیر کنید.

۵. فرآیند جست‌وجو را به صورت زیر برای سه رویداد دلخواه از سه سال گذشته انجام دهید. (می‌توانید از پرسش‌های موجود در فایل sample_questions.txt کمک بگیرید.) و خروجی مربوطه را در گزارش خود اضافه کنید.

- تبدیل پرسش به بردار
- پیدا کردن نمایه‌ی متناسب با پرسش مربوطه
- پیدا کردن تمامی داده‌های خارجی نمایه‌ی مورد نظر
- محاسبه معیار شباهت کسینوسی و خروجی دادن بردارهای داده‌های خارجی با شباهت بیشتر از آستانه. (چرا معیار کسینوسی در این مسئله انتخاب مناسبی است؟)