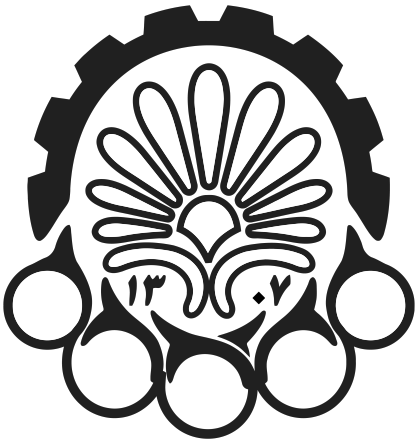
شبکه‌های عصبی و یادگیری عمیق
دکتر صفابخش



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر

رضا آدینه پور ۴۰۲۱۳۱۰۵۵

تمرین سوم
شبکه خودسازمان‌ده (SOM)

۳۱ فروردین ۱۴۰۳

دانشکده مهندسی کامپیوتر

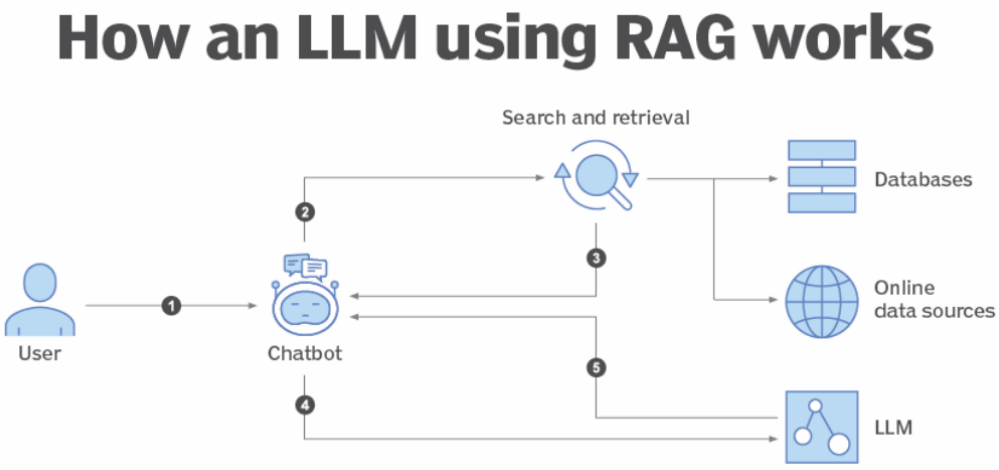
شبکه‌های عصبی و یادگیری عمیق

تمرین سوم

رضا آدینه پور ۴۰۲۱۳۱۰۵۵

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 شده‌ی پس از این زمان ندارد. به‌عنوان مثال، تاریخ قطع آموزش مدل GPT-3.5-turbu-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 به‌عنوان ورودی به مدل زبانی تلاش می‌کند که فرمان ورودی را تقویت کرده و به مدل زبانی کمک می‌کند تا جواب مرتبط و متناسبی بسازد. به‌عنوان مثال در پاسخ به یک فرمان متنی مانند «چه‌کسی شرکت توییتر را درسال ۲۰۲۲ خرید؟» تمامی داده‌های خارجی متناسب با این فرمان را استخراج می‌کند و آن‌ها را به‌عنوان 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.



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» دیاگرام 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 درنظر 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.

۲. مجموعه داده ارائه شده در این پروژه شامل رویدادهای سه‌سال 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 آ ن انجام دهید.