به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس جستجو و بازیابی اطلاعات در وب

استاد ممتازی

تمرین دوم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

بخش دوم – بازیابی با استفاده از بردارهای معنایی

مطابق با درخواست سوال یک مدل BERT با استفاده از کتابخانه Hugging Face بارگزاری و استفاده شد. برای بدست آوردن یک تعبیه جمله از میانگین بردار تمام توکن‌ها متن استفاده کرده‌ایم. نه تنها میانگین بردارهای لایه آخر، که میانگین تمام بردارهای مخفی چهار لایه آخر مورد استفاده قرار گرفت. علت استفاده از چهار لایه آخر به جای لایه آخر نتایج بهتر گزارش‌شده برای میانگین چهار لایه آخر به جای آخرین لایه بوده است.

در پیاده‌سازی من مطابق با درخواست سوال شبکه BERT تنظیم‌دقیق نمی‌شود. همچنین برای کاهش زمان اجرا، تمام بردارهای مربوط به داده‌های آموزش یک بار تهیه و نگهداری می‌شود.

بخش سوم – بازیابی با آموزش مدل شبکه عصبی

برای این قسمت سه مدل تعبیه کلمه پیش‌آموزش‌یافته GloVe، Word2Vec و FastText از کتابخانه Gensim مورد استفاده قرار گرفت. برای تعبیه متن یک لایه LSTM در دو حالت دوطرفه و یک‌طرفه ارزیابی شد. نهایتا برای ترکیب خروجی دو زیرشبکه از سه حالت مختلف استفاده شد:

1. تفاضل دو بردار حساب می‌شود و سپس با یک لایه متراکم یک خروجی تک عددی بدست می‌آید.
2. دو بردار با هم ادغام می‌شوند و سپس با یک لایه متراکم یک خروجی تک عددی بدست می‌آید.
3. شباهت کسینوسی دو بردار بدست می‌آید.

برای آموزش مدل از خطای کراس‌آنتروپی، بهینه‌ساز آدام و حداکثر ده گام آموش استفاده کردم. نهایتا توجه داشته باشید که برای آموزش مدل بر خلاف تمام مدل‌های تا به الان از GPU استفاده کردم.

شایان ذکر است که برای تسریع زمان اجرا در هنگام آزمون، برای تمام داده‌های آموزش، بردارهای مربوط به زیرشبکه اول را محاسبه و نگهداری کردم. همچنین برای داده تست هم یک بار این بردار را حساب کردم و با تکثیر آن، جفت‌های نهایی را برای لایه ادغام آماده کردم. به این ترتیب زمان آزمون بیشتر مربوط به اعمال لایه ادغام است.

با توجه به توضیحات بیان شده سه سری تنظیم تعبیه کلمه، یک‌طرفه/دوطرفه بودن LSTM و نحوه ادغام دو خروجی را داریم که باید مشخص شود. به صورت مرحله‌ای هر تنظیم را مشخص خواهیم کرد. ابتدا سه حالت تعبیه کلمه را بررسی کردم و برای دو تنظیم دیگر مقادیر پیش‌فرض را لحاظ کردم. نتایج در جدول ۱ آورده شده است. مطابق این جدول بهترین مدل تعبیه کلمه برای کاربرد ما و شرایط آموزش مدل GloVe است.

جدول ۱ - دقت‌های معیار ارزیابی برای شبکه سیامیس با تعبیه‌های کلمه مختلف

	P@5	P@10	MAP	MRR
Word2Vec	۲۳.۹۷٪	۱۶.۷۱٪	۱۷.۳۵٪	۴۷.۶۳٪
FastText	۱۶.۱۶٪	۱۲.۴٪	۱۱.۳۵٪	۳۷.۸۵٪
GloVe	۳۰.۲۷٪	۲۱.۹۹٪	۲۲.۲۲٪	۵۴.۳۳٪

سپس مدل تعبیه کلمه را روی GloVe فیکس کردم و سه استراتژی مختلف برای ادغام بردارهای زیرشبکه را امتحان کردم که نتایج آن در جدول ۲ آورده شده است.

جدول ۲ - دقت‌های معیار ارزیابی برای شبکه سیامیس با استراتژی‌های مختلف ادغام

	P@5	P@10	MAP	MRR
اختلاف	۰٪	۰٪	۰٪	۰٪
ادغام	۰.۱۴٪	۰.۰۷٪	۰.۰۱٪	۰.۱۴٪
شباهت کسینوسی	۳۰.۲۷٪	۲۱.۹۹٪	۲۲.۲۲٪	۵۴.۳۳٪

با توجه به نتایج جدول ۲ به نظر می‌رسد دو استراتژی اختلاف و ادغام متناسب با پیاده‌سازی فعلی من اصلا مناسب نیستند و همان استراتژی پیش‌فرض شباهت کسینوسی بهترین گزینه است.

نهایتا نتایج LSTM یک‌طرفه را با LSTM دوطرفه مقایسه کردم که نتایج آن در جدول ۳ آورده شده است.

جدول ۳ - دقت‌های معیار ارزیابی برای شبکه سیامیس با تغییر ساختار زیرشبکه

	P@5	P@10	MAP	MRR
LSTM	۲.۴۷٪	۱.۵۱٪	۱.۱۲٪	۶.۷۲٪
BiLSTM	۳۰.۲۷٪	۲۱.۹۹٪	۲۲.۲۲٪	۵۴.۳۳٪

پس شبکه‌ای با تعبیه کلمه GloVe، زیرشبکه‌های BiLSTM و ادغام با شباهت کسینوسی را به عنوان شبکه نهایی سیامیس انتخاب می‌کنم.

بخش چهارم – کار با روش‌های ارزیابی

با بدست آمدن نتایج برای دو مدل جدید می‌توان به صورت خلاصه و بر اساس نتایج این تمرین و تمرین پیشین جدول ۴ را کامل کرد. همچنین جدول زمان اجرا هم مطابق با جدول ۵ تکمیل می‌گردد.

جدول ۴ – دقت‌های معیار ارزیابی تمام مدل‌ها

	P@5	P@10	MAP	MRR
TF-IDF	۴۴.۵۲٪	۳۱.۷۱٪	۳۹.۴۸٪	۶۲.۷۷٪
Unigram	۳۹.۸۶٪	۲۸.۴۲٪	۳۷.۵۵٪	۶۱.۰۵٪
Bigram	۴۰.۲۷	۲۸.۱۵٪	۳۶.۷۲٪	۶۰.۷۵٪
BERT	۴۵.۷۵٪	۳۱.۷۸٪	۴۱.۴٪	۷۶.۶٪
Siamese	۳۰.۲۷٪	۲۱.۹۹٪	۲۲.۲۲٪	۵۴.۳۳٪

جدول ۵ – زمان آموزش و آزمون تمام مدل‌ها

	زمان آموزش (s)	زمان آزمون (s)
TF-IDF	۳.۴۵	۲۰۰.۶۹
Unigram	۰.۵۷	۹۱.۱
Bigram	۰.۹۱	۱۶۵.۴۵
BERT	۲۴۰۵.۲۲	۳۸۵.۱۷
Siamese	۵۶۷۶.۰۳	۲۶۱.۴۹

با بررسی نتایج به نظر می‌رسد که مدل BERT توانسته است به بهترین دقت‌ها از میان تمام مدل‌ها دست پیدا کند. اما زمان آموزش و آزمون آن بسیار بیشتر از رقیب آن یعنی TF-IDF است. به علاوه آنکه در پیاده‌سازی ما مدل BERT را تنظیم دقیق نکردم و زمان آموزش شامل بدست آوردن بردارهای فضای جستجو می‌شود.

شبکه سیامیس مطابق با چیزی که من پیاده‌سازی کردم علی‌رغم تنظیمات مختلف و استفاده از GPU دقت و زمان اجرای بدی دارد و نیازمند بهبود بیشتر است.