به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس پردازش زبان‌های طبیعی

تمرین ۳

اردیبهشت ماه ۱۴۰۲

*فهرست

مقدمه و مجموعه داده

در این تمرین هدف بررسی روش‌های تحلیل Sequence با استفاده از شبکه عصبی است. به طور خاص با تحلیل جمله، می‌خواهیم مدلی برای تحلیل احساس آموزش دهیم. مجموعه داده‌هایی که برای این تمرین در نظر گرفته شده است، داده‌های Sentiment۱۴۰ خواهد بود که می‌توانید از آدرس زیر دریافت و با جزئیات آن آشنا شوید.

http://help.sentiment140.com/for-students

کتابخانه مورد استفاده در این تمرین PyTorch خواهد بود. نحوه حل مسئله در این تمرین بسیار مهم بوده و سعی کنید هر قسمت را به طور کامل توضیح دهید.

پیش پردازش

همانطور که می‌دانید، یکی از مراحل آموزش یک مدل هوشمند، پیش پردازش داده‌های اولیه است. مجموعه داده‌ها را ابتدا دریافت کرده و سپس برای استفاده در مدل خود آماده کنید. به عنوان مثال بعضی از توییت‌ها دارای لینک، هشتگ و یا منشن هستند که بایستی برای آن‌ها استراتژی مناسبی در نظر گرفته شود.

بخش ۱

بعد از پیش پردازش دادههای اولیه، یک شبکه بازگشتی ساده طراحی و سپس آن را آموزش داده و ارزیابی کنید.

۱. هر جمله را با استفاده از کاراکتر Space به توکن تبدیل کنید.
۲. سه روش one-hot، Word2vec و Glove را در مدل خود برای تبدیل کلمات به بردار مناسب استفاده کرده و آن‌ها را ارزیابی کنید.
   a. در روش one-hot، بعد از تشکیل بردار اولیه، آن‌ها را از یک لایه خطی با بعد ۱۵۰ عبور داده و در حین آموزش Task، آموزش دهید.
۳. از آنجایی که طول همه جملات یکسان نیستند، Padding مناسبی برای ورودی شبکه در نظر بگیرید. این Padding می‌تواند در سمت راست جمله و یا در سمت چپ آن باشد.
۴. شبکه مناسبی را برای Task طراحی کنید.

. برای این کار می‌توانید قسمت کوچکی از داده‌ها را جدا کرده و سعی کنید مدل‌های خود را با استفاده از آن آموزش دهید.

a. از داده‌ها Development برای بررسی معماری شبکه خود استفاده کنید (بخشی از داده آموزش را جدا کنید).

۵. خروجی آخرین لایه بازگشتی را از یک خطی عبور داده و با اعمال Softmax خروجی نهایی را ایجاد کنید.
۶. ماتریس درهم ریختگی (Confusion Matrix)را ایجاد و آن را تحلیل کنید.

بخش ۲

در این بخش می‌خواهیم LSTM و GRU را بررسی کنیم.

۱. ابتدا مدلی مبتنی بر LSTM طراحی کنید.
۲. بعد مخفی را ۱۵۰ در نظر گرفته و از بهینه ساز Adam استفاده کنید.
۳. از تابع هزینه Cross entropy استفاده کنید.
۴. در لایه Embedding دو روش one-hot و glove را بررسی کنید.
۵. ماتریس درهم ریختگی (Confusion Matrix) را ایجاد و آن را تحلیل کنید.
۶. مدل LSTM را با مدل GRU عوض کرده و نتایج را تحلیل کنید.
۷. مدل GRU و LSTM را با یکدیگر مقایسه کنید (زمان اجرا و ...).

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان NLP_CA3_StudentID تحویل داده شود.

- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto: mollaabbasi.m@ut.ac.ir

مهلت تحویل:۱۴۰۲/۰۲/۱۶