بسمه تعالی

تمرین کامپیوتری شماره ی ۵ پردازش زبان طبیعی

مهلت تحویل تمرین: ۲۴ خرداد

فاطمه ایمانی پور(fatemeh.imanipour@ut.ac.ir)

امیرحسین فقیهی فر (a.faghihifar@gmail.com)

مقدمه

در این تمرین قصد داریم با روند آموزش یک مدل ترجمه ماشینی مبتنی بر شبکه های عصبی و پارامتر های موثر در آن آشنا شویم. برای این کار ابزار های قدرتمند متنوعی وجود دارد که در این تمرین قصد داریم به 3 مورد از آنها بپردازیم:

۱) ابزار MarianNMT

این ابزار بر پایه‌ی زبان C توسعه داده شده و از سریعترین ابزارهای آموزش مدل‌های ترجمه‌ی ماشینی می باشد. از لینک زیر می توانید با این ابزار و نحوه ی آموزش مدل‌های ترجمه ماشینی با آن بیشتر آشنا شوید:

https://marian-nmt.github.io/

۲) ابزار OpenNMT

این ابزار متن باز در دو نسخه ی مبتنی بر Tensorflow و PyTorch قابل استفاده است. در این پروژه شما باید از نسخه ی PyTorch این ابزار استفاده کنید. از لینک زیر می‌توانید تمام اطلاعات لازم برای استفاده از این ابزار را بدست آورید:

https://opennmt.net/

همینطور برای آشنایی پایه ای با این ابزار می توانید از لینک زیر در Google Colab با برخی ویژگی های پایه ای این ابزار آشنا شوید:

https://colab.research.google.com/drive/1Nkd9UFlDX4NhX_gVQwDS-77s2jV7zTqE#scrollTo=ZdTjS0bTSVLk

۳) ابزار FairSeq

این ابزار متن باز مبتنی بر PyTorch که توسط شرکت Facebook ارائه شده است، قابلیت‌های مختلفی دارد و می توان از آن در آموزش مدل‌های مختلف در حوزه‌های متنوعی در NLP استفاده کرد. یکی از ماژول‌های این ابزار مربوط به آموزش مدل‌های ترجمه‌ی ماشینی است. لینک آشنایی با این ابزار در زیر قرار داده شده است:

https://github.com/pytorch/fairseq

شرح پروژه

در این پروژه می بایست دو ابزار از سه ابزار ارایه شده را انتخاب کرده و دو مدل ترجمه‌ی ماشینی با معماری Transformer آموزش دهید. در پروسه‌ی آموزش این دو مدل می بایست علاوه بر توضیح فرایند اجرا شده، نتایج را به صورت کامل مستند کنید. در گزارش شما می بایست به سوالات زیر پاسخ داده شود:

۱) نتایج بدست آمده از دو ابزاری که انتخاب کرده اید را مقایسه کنید. همینطور در صورت وجود، تفاوت بین نتایج دو ابزار را تحلیل کنید.

۲) شرح دهید که در پروسه‌ی آموزش مدل، دو ابزار انتخاب شده چه مزایا و معایبی نسبت به هم داشته اند.

۳) پارامترهایی که در پروسه‌ی آموزش مدل تنظیم (tune) کرده اید را به طور کامل شرح دهید. ابتدا شرح دهید که پارامترها چه کاری انجام می دهند و سپس توضیح دهید چه پارامترهایی را نسبت به حالت پیش فرض تغییر داده‌اید.

۴) در هر ابزار انتخابی ۵ پارامتری که فکر میکنید در کیفیت مدل خروجی تاثیر گذار هستند را نام برده و هر یک را مختصرا توضیح دهید.

۵) توضیح دهید برای آموزش یک ماشین ترجمه‌ی مناسب چه پیش پردازش‌هایی لازم است انجام گیرد و این پیش‌پردازش‌ها را روی دادگان اعمال کرده و توضیح دهید برای پیش پردازش از چه ابزاری استفاده کرده اید. فایل حاوی داده های پیش پردازش شده را در گزارش بیاورید.

۶) روند تغییرات Bleu مدل آموزش دیده را در یکی از ابزار های انتخابی خود بر روی مجموعه dev با افزایش تعداد epoch ها نشان دهید. برای این کار می‌توانید از دستور ذخیره مدل های میانی در ابزار مورد نظر خود استفاده کنید.

نکات مهم

تذکر۱: توصیه اکید میشود که برای استفاده از هر ۳ ابزار از Google Colab و GPU تعبیه شده روی آن استفاده کنید .

تذکر۲: معیار ارزیابی مدل‌ها BLEU می باشد. برای محاسبه‌ی امتیاز BLEU می توانید از توابع موجود در هر یک از ابزار های معرفی شده (در صورت وجود) و یا از ابزار Moses یا کتابخانه‌ی NLTK یا استفاده نمایید. لینک هر دو در زیر آمده است:

https://github.com/moses-smt/mosesdecoder

https://www.nltk.org/

تذکر ۳: در مورد تعداد epoch ها یا طول زمان آموزش مدل محدودیت خاصی قائل نمی شویم اما در نظر داشته باشید که :

- ✓ در این تمرین قصد داریم با روند آموزش یک مدل ماشین ترجمه و الزامات آن آشنا شویم . انتظار ما این است که مدل نهایی که ارائه می دهید در میانه مسیرآموزش باشد. یکی از راه های بررسی این موضوع کنترل دستی خروجی مدل بر روی داده های تست است برای مثال معمولا خروجی یک مدل ترجمه که تازه شروع به آموزش کرده است، تکرار تنها چند کلمه خاص است و بدیهی است که این مدل به عنوان مدل نهایی پذیرفته نیست.
- ✓ با توجه به کم بودن حجم مجموعه داده اولیه و محدودیت های منابع انتظار تولید یک ماشین باکیفیت را نداریم. نگران نتایج ضعیف احتمالی نباشید. :)
- ✓ انتظار نداریم که مجموع زمان آموزش در هر ابزار بیش از ۶ ساعت به طول انجامد و الزاما صرف زمان بیشتر برای آموزش مدل امتیاز محسوب نمیشود.
- ✓ مجددا تاکید میکنیم که در این تمرین هدف اصلی آن است که مسیر آموزش یک ماشین ترجمه مبتنی بر شبکه های عصبی به درستی طی شود و بتوانید تحلیل درستی از نتایج وشرایط پیش امده داشته باشید.

- ✓ در مسیر آموزش مدل اگر متوجه مشکلی شدید و امکان رفع آن را نداشتید، مشکل احتمالی و راه حل پیشنهادی خود را در گزارش عنوان کنید. کیفیت گزارش شما و تحلیل و بررسی درست و دقیق مسائل بخش قابل توجهی از نمره نهایی این تمرین را به خود اختصاص خواهد داد.

تذکر ۴: میتوانید تمرین را در قالب گروه های حداکثر دو نفره انجام دهید.

لطفا علاوه بر فایل گزارش، فایل اسکریپت دستورات اجرا شده و یا اگر در Google Colab اجرا کرده اید فایل notebook آن به همراه خروجی سیستم های آموزش داده شده برای فایل های تست را نیز ارسال کنید.

لطفا به قوانین انجام تمرین ها که پیش از این عنوان شده توجه داشته باشید.

موفق باشید.