به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس پردازش زبان طبیعی

تمرین۵

خرداد ماه ۱۴۰۳

فهرست

مقدمه

ترجمه ماشینی یکی از زیر شاخه های سنتی و مهم پردازش زبان طبیعی است که در آن نحوه استفاده از نرم افزار رایانه ای در ترجمه متن یا گفتار از یک زبان به زبان دیگر بدون مشارکت انسان مطالعه می شود. ایده استفاده از شبکه های عصبی برای ساخت موتور ترجمه ماشینی اولین بار در سال ۱۹۸۷ مطرح شد. اما به علت کمبود داده آموزشی و توان محاسباتی کافی استفاده از شبکه های عصبی در سیستم های ترجمه ماشینی برای دو دهه به تعویق افتاد. در سال های اخیر با افزایش حجم داده های آموزشی و امکان استفاده از توان پردازشی بالای ، GPUاستفاده از شبکه های عصبی در ترجمه ماشینی منجر به برداشتن گامی چشمگیر در افزایش کیفیت این ماشین ها شده است.

ابزار Fairseq یک ابزار متن باز برای sequence modeling می باشد. با استفاده از Fairseq می توانیم برای وظایف ترجمه، خلاصه سازی، مدل سازی زبان و به طور کلی تمام وظایف تولید متن عملیات ساخت و آموزش مدل را انجام دهیم. این ابزار توسط شرکت Facebook با هدف انعطاف پذیری بالا برای تعریف task و مدل های جدید به روی بستر PyTorch توسعه داده شده است.

در این سری از تمرین ما قصد داریم با استفاده از ابزار fairseq یک سیستم ترجمه ماشینی برای ترجمه ی انگلیسی به فارسی توسعه دهیم. در بخش‌های مختلف به پردازش دادگان، آموزش tokenizer، مقایسه‌ی معماری‌های مختلف برای تسک ترجمه ماشینی و ارزیابی مدل‌ها می‌پردازیم.

همچنین یک هندزآن برای آشنایی با ابزار fairseq، توسط آقای رستمی طراحی شده است که در ضمیمه پیوست می‌شود.

توجه:

ابتدا تمام قسمت‌های تمرین را مطالعه کرده و با تمام مراحل آشنا شوید، سپس شروع به پیاده‌سازی قسمت‌های مختلف، به ترتیبی که مد نظرتان هست بکنید.

سوال اول

این سوال شامل چندین بخش مختلف شامل پردازش داده، آموزش توکنایزر، آموزش مدل با معماری‌های مختلف و ارزیابی مدل‌ها می‌باشد.

برای انجام این تمرین نیاز به نصب کتابخانه‌های sacremoses و sentencepiece از طریق pip دارید. همچنین نیاز به نصب کتابخانه‌ی fairseq دارید که حتما آخرین نسخه‌ی آن از گیت‌هاب را نصب کنید تا موقع گرفتن تست از مدل، به مشکل نخورید (نصب از این طریق کمی بیشتر از حالت نصب از طریق pip طول می‌کشد).

موارد زیر را در تمام مدل‌هایی که آموزش می‌دهید یکسان در نظر بگیرید.

- هر مدل را به اندازه‌ی ۵ ایپاک آموزش دهید. با توجه به حجم داده هر ایپاک حدود ۶ تا ۹ دقیقه طول خواهد کشید. بنابراین بهتر است برای آموزش مدل‌ها زمان مناسبی در نظر بگیرید و کار را به روزهای آخر موکول نکنید. قطعا برای آموزش مدل با دقت قابل قبول به تعداد ایپاک بیشتری نیاز است، اما به دلیل کمبود منابع، همین مقدار کفایت می‌کند.
- از بهینه‌ساز adam با پارامترهای beta به مقدار 0.9 و 0.98 استفاده کنید.
- تابع خطای label_smoothed_cross_entropy با مقدار label-smoothing برابر با 0.2 استفاده کنید.
- به انتخاب مقدار مناسب برای learning rate و warm up steps توجه کنید، این پارامترها در آموزش سریعتر و بهتر مدل تاثیر دارند.
- از ابزار tensorboard استفاده کنید و مقادیر loss در طول فرآیند آموزش برای داده‌ی آموزش و ارزیابی را بدست آورید. عکسی از نمودار آن را گزارش کنید و فایل log مقادیر loss برای آموزش و ارزیابی را با فرمت csv دانلود کرده و در ضمیمه همراه با جواب ارسال کنید.

دادگان (۲۰ نمره)

برای آموزش مدل ترجمه‌ی ماشینی نیاز به داده در دو زبان مورد نظر به صورت موازی داریم. در این تمرین هدف ما ترجمه **از** انگلیسی **به** فارسی می‌باشد. دیتاست مورد استفاده‌ی ما برای آموزش مدل، مجموعه‌داده‌ی میزان (گیت‌هاب، مقاله) می‌باشد.

مجموعه داده را می‌توانید از این لینک دریافت کنید.

ابتدا مجموعه داده را دریافت کرده و موارد گفته شده در پایین را روی این دیتاست گزارش کنید.

۱. در دیتاست دو فایل وجود دارد، یکی برای فارسی و دیگری برای انگلیسی، تعداد کل خطوط و محتوای سه خط اول این دو فایل را گزارش کنید.
۲. دیتای هر سطر را با whitespace توکنایز کنید و هیستوگرام تعداد توکن‌های هر سطر را برای داده‌های فارسی و انگلیسی به طور جدا گزارش کنید.
۳. در اینجا هدف کاهش حجم دیتاست است، **بر اساس دیتای فارسی**، سطرهایی که تعداد توکن بیشتر از ۵۰ و کمتر از ۱۰ دارند را پیدا کرده و هم از دیتای فارسی و هم از دیتای انگلیسی حذف کنید. سپس تعداد سطرهای جدید دیتاست را گزارش کنید.
۴. پس از فیلتر کردن تعداد سطرهای دیتاست، مجموعه داده را با **یک random seed مشخص و ثابت** shuffle کنید، سپس داده‌ی آموزش، ارزیابی و تست را با تعداد ۵۰۰۰۰۰، ۵۰۰۰ و ۱۰۰۰۰ به ترتیب از ابتدای دیتای shuffle شده جدا کنید.
۵. در نهایت در یک پوشه‌ی raw_data شش فایل جدید با عنوان‌های زیر بسازید.

train.en, train.fa, valid.en, valid.fa, test.en, test.fa

در نهایت دیتای شما به تفکیک آموزش، ارزیابی و تست در یک پوشه آماده‌ی استفاده می‌باشد.

بخش اول: آموزش توکنایزر BPE و پیش پردازش دادگان (۲۰ نمره)

در این بخش از کتابخانه‌ی sentencepiece استفاده کنید و به طور مجزا برای دادگان آموزش فارسی و انگلیسی توکنایزر با روش bpe و اندازه vocab برابر 10K آموزش دهید. سپس دیتای آموزش، ارزیابی و تست را با مدل آموزش دیده توکنایز کرده و در مسیری مشخص ذخیره کنید.

سپس با استفاده از ابزار fairseq-preprocess از دادگان توکنایز شده استفاده کنید و با قرار دادن nwordssrc و nwordstgt به اندازه‌ی 10K پیش پردازش را انجام دهید.

در این قسمت گزارش کنید که فرمان fairseq-preprocess چه کاری انجام می‌دهد و بعد از آن چه فایل‌هایی تولید می‌شود.

بخش دوم: آموزش مدل LSTM ENCODER-DECODER (۲۰ نمره)

در این قسمت از مدل‌های آماده‌ی پیاده‌سازی شده در fairseq استفاده کنید. مدل encoder decoder که با lstm پیاده‌سازی شده است را پیدا کنید. سپس با پارامتر مناسب تعداد لایه‌های encoder و decoder را ۶ قرار دهید. یعنی معماری مدل شما دارای ۱۲ لایه خواهد بود. این مدل را با استفاده از دیتای توکنایز شده (انجام شده با BPE در قسمت قبل)، و با فرمان fairseq-train آموزش دهید.

- حتما از پارامتر مناسب برای ذخیره‌ی بهترین مدل در طی فرآیند آموزش استفاده کنید.
- نحوه‌ی استفاده از دو پارامتر max-tokens-- و batch-size-- را توضیح دهید. مقادیر استفاده شده برای این دو پارامتر را گزارش کنید و بگویید برای کنترل بهتر حجم داده در هر batch چگونه از این دو پارامتر استفاده کرده اید؟

بخش سوم: آموزش مدل TRANSFORMER ENCODER-DECODER (۲۰ نمره)

در این قسمت از مدل‌های آماده‌ی پیاده‌سازی شده در fairseq استفاده کنید. مدل encoder decoder که با transformer پیاده‌سازی شده و در encoder و decoder ۶ لایه وجود دارد را پیدا کرده و انتخاب

کنید. این مدل را با استفاده از دیتای توکنایز شده (انجام شده با BPE در قسمت‌های قبل)، و با فرمان fairseq-train آموزش دهید.

بخش چهارم: معیار ارزیابی و بررسی داده‌ی تست (۲۰ نمره)

در نظر بگیرید تا برای آموزش دو مدل، معیار BLEU در نظر گرفته شود و گزارش شود.

برای هر دو مدل بخش دوم و سوم، بعد از آموزش مدل از فرمان fairseq-generate استفاده کرده و خروجی مدل برای داده‌های تست را در فایل ذخیره کنید. در انتهای این فایل‌ها مقدار BLEU برای داده‌های تست ذخیره می‌شود، آن‌ها را نیز در گزارش خود بیان کنید.

در مورد معیار ارزیابی Comet تحقیق کرده و نحوه‌ی کار آن را توضیح دهید.

فایل خروجی فرمان fairseq-generate را بررسی کنید. در این فایل جملات source, target و machine-translated target قرار دارد. اطلاعات لازم را از فایل بیرون کشیده و از BPE tokenizer هایی که روی داده‌های خود آموزش داده بودیم استفاده کنید و جملات را decode کنید.

از کتابخانه‌ی unbabel-comet استفاده کنید و با استفاده از بهترین مدل comet، نتایج داده‌های تست برای دو مدل آموزش داده شده را ارزیابی کنید و گزارش دهید.

نتایج comet و bleu را با هم مقایسه کنید.

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان NLP_CA5_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.
- تمرین تا یک هفته بعد از مهلت تعیین شده با تاخیر تحویل گرفته می شود. دقت کنید که شما جمعا برای تمام تکالیف، ۱۴ روز زمان تحویل بدون جریمه دارید که تنها از ۷ روز آن برای هر تمرین می‌توانید استفاده کنید، در صورتی که این ۱۴ روز به اتمام رسیده باشد، به ازای هر روز تاخیر در ارسال تمرین، ده درصد جریمه میشوید.
- **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 تشابه 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:
  mstfmasoudii@gmail.com

مهلت تحویل بدون جریمه: ۱۹ خرداد ۱۴۰۳

مهلت تحویل با تأخیر، با جریمه ۱۰ درصد : ۲۶ خرداد ۱۴۰۳