به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس پردازش زبان‌های طبیعی

تمرین ۵

خرداد ماه ۱۴۰۲

*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 GPU ، استفاده 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 ترجمه‌ی انگلیسی به فارسی توسعه دهیم و تاثیر استفاده از شبکه های از پیش آموزش داده شده چند زبانه مانند mBERT و ... را نیز مورد مطالعه قرار دهیم. معماری استفاده شده در این تمرین یک شبکه encoder-decoder شامل یک لایه encoder و یک لایه decoder با معماری LSTM با مکانیزم attention است(معماری ذکر شده به صورت پیش ساخته در ابزار Fairseq موجود است).

نکاتی درباره‌ی پیاده سازی، موارد تحویلی و استفاده از Fairseq

نکته: تمامی مدل های مورد مطالعه در این تمرین از معماری یکسان استفاده می‌کنند. مشخصات مدل مذکور به شرح زیر است:

- معماری encoder-decoder شامل یک لایه encoder و یک لایه decoder با معماری LSTM با مکانیزم attention
- تابع هزینه label smoothed cross entropy به مقدار label smoothing برابر ۰,۲
- نرخ یادگیری تطبیقی inverse sqrt با مقدار اولیه ۰,۰۰۲۵
- بهینه ساز adam با پارامتر های beta1=0.9 و beta2=0.98
- استفاده از dropout با مقدار ۰,۲۵

نکته: برای تمام مدل های ساخته شده در این تمرین می بایست موارد زیر را گزارش کنید:

- نمودار loss با استفاده از tensorboard
- فایل log مربوط به نمودار loss با فرمت csv
- مقدار BLEU برای دادگان ارزیابی

برای آشنایی بیشتر شما با ابزار Fairseq ، Hands-on ای در رابطه با کار با این کتابخانه و در کل آموزش و ارزیابی مدل‌های ترجمه‌ی ماشینی تهیه شده است. توجه کنید این Hands-on صرفا برای آشنایی بیشتر شما با Fairseq تهیه شده است و برای آزمایش‌های مختلف، ممکن است نیاز به تغییر یا حذف بعضی از بخش‌های آن داشته باشید. (لینک کولب Hands-on)

مجموعه داده

برای آموزش مدل‌های ترجمه‌ی ماشینی از مجموعه دادگان موازی استفاده می‌شود. مجموعه دادگان موازی شامل جملاتی به زبان مبدا و مقصد است که به صورت متناظر ترجمه‌ی همدیگر هستند. در این تمرین ازمجموعه داده‌ی AFEC برای آموزش مدل‌های ترجمه‌ی ماشینی استفاده خواهد شد. این مجموعه داده شامل جملات فارسی و انگلیسی هم تراز شده و جملات ترجمه شده توسط انسان است. برای اطلاعات بیشتر درباره‌ی دیتاست AFEC می‌توانید مقاله‌ی آن را بخوانید (لینک مقاله).

در این تمرین ما ۶۸۴۰۰۰ جمله از این دیتاست را انتخاب کردیم و از این تعداد، ۶۷۰۰۰۰ جمله را به دیتاست آموزش، ۱۰۰۰۰ جمله را به دیتاست ارزیابی و ۴۰۰۰ جمله را هم به دیتاست اعتبارسنجی اختصاص دادیم.

سوال ۱

در بخش اول این تمرین به دنبال مطالعه تاثیر BPE در سیستم ترجمه ماشینی هستیم. به این منظور ابتدا با طی مراحل زیر یک مدل بدون استفاده از BPE می سازیم:

پیش پردازش دادگان با استفاده از fariseq-preprocess

برای پیش پردازش داده ها با استفاده از دستور fairseq-preprocess از گزینه های nwordssrc و nwordstgt به منظور تعیین اندازه vocab مربوط به زبان مبدا و مقصد با مقدار ۴۰k برای هر یک استفاده نمایید.

آموزش مدل با استفاده از fairseq-train

برای آموزش مدل، پارامترهای fairseq-train را مطابق مقادیر گفته شده مقداردهی نمایید و مدل را برای حداقل ۵ ایپاک آموزش دهید.

استفاده از fairseq-generate برای ترجمه دادگان ارزیابی

پس از اتمام آموزش مدل با استفاده از دستور fairseq-generate دادگان ارزیابی را ترجمه کرده و در فایل مجزا ذخیره نمایید.

استفاده از Tensorboard برای بررسی فرایند آموزش

پس از اتمام ارزیابی مدل، به کمک Tensorboard می‌توانید نمودار پارامترهای مختلف (مانند loss) در طول فرایند آموزش را مشاهده کنید. برای رسم نمودار loss و به دست آوردن فایل log مقادیر آن (با فرمت csv) از این ابزار استفاده کنید.

حال با طی مراحل زیر اقدام به ساخت مدل با استفاده از داده های پیش پردازش شده با استفاده از BPE می‌پردازیم:

آموزش مدل BPE برای دادگان انگلیسی و فارسی

با استفاده از کتابخانه sentencepeice برای مجموعه دادگان آموزش انگلیسی و فارسی به طور مجزا مدلی با اندازه vocab برابر ۵k آموزش دهید.

پردازش دادگان با مدل های BPE

با استفاده از مدل های BPE تمامی دادگان AFEC شامل دادگان انگلیسی و فارسی در هر سه دسته آموزشی، اعتبار سنجی و ارزیابی را پردازش کرده و در فایل جدید ذخیره نمایید.

پیش پردازش دادگان با استفاده از fariseq-preprocess

با استفاده از fairseq-preprocess پیش پردازش دادگان را انجام دهید.

آموزش مدل با استفاده از fairseq-train

برای آموزش مدل، پارامترهای fairseq-train را مطابق مقادیر گفته شده مقداردهی نمایید و مدل را برای حداقل ۵ ایپاک آموزش دهید.

استفاده از fairseq-generate برای ترجمه دادگان ارزیابی

پس از اتمام آموزش مدل با استفاده از دستور fairseq-generate دادگان ارزیابی را ترجمه کرده و در فایل مجزا ذخیره نمایید.

استفاده از Tensorboard برای بررسی فرایند آموزش

پس از اتمام ارزیابی مدل، به کمک Tensorboard می‌توانید نمودار پارامترهای مختلف (مانند loss) در طول فرایند آموزش را مشاهده کنید. برای رسم نمودار loss و به دست آوردن فایل log مقادیر آن (با فرمت csv) از این ابزار استفاده کنید.

سوال ۲

در این بخش از تمرین به دنبال استفاده از مدل های چند زبانه از پیش آموزش داده شده به طور خاص mBERT در سیستم ترجمه ماشینی هستیم. یکی از ساده ترین روش های استفاده از این مدل ها بکار گیری وزن های لایه embedding بعنوان مقدار اولیه وزن های شبکه خودمان می باشد. برای این منظور می بایست گام های زیر را به ترتیب طی نمایید:

دانلود Bert Tokenizer و Bert Model

برای دانلود tokenizer و مدل مورد نیاز می توانیم از کتابخانه huggingface استفاده کنیم. به این منظور به لینک زیر مراجعه نمایید.

https://huggingface.co/bert-base-multilingual-cased

پردازش دادگان AFEC با استفاده از Bert Tokenizer

با استفاده از tokenizer دانلود شده تمامی دادگان AFEC شامل دادگان انگلیسی و فارسی در هر سه دسته آموزشی، اعتبار سنجی و ارزیابی را tokenize کرده و در فایل جدید ذخیره نمایید.

پیش پردازش دادگان با استفاده از fariseq-preprocess

با توجه به یکسان بودن tokenizer برای هر دو زبان مبدا و مقصد در هنگام پیش پردازش، هنگام اجرای fairseq-preprocess از گزینه ی joined-dictionary استفاده نمایید.

ذخیره وزن های لایه embedding شبکه Bert با فرمت مناسب

به منظور استفاده از وزن های لایه embedding شبکه Bert در ابزار Fairseq می بایست وزن های مربوط به این لایه را با فرمت زیر در یک فایل متنی ذخیره نماییم. به طور مثال:

```
2 3
hello 0.14234 0.13423 0.23412
world  0.54497 0.68455 0.21342
```

در این مثال ساده vocab_size برابر ۲ و embedding_size برابر ۳ می باشد (خط اول فایل). سایر خطوط شامل یکی از عناصر vocab و بردار embedding مربوط به آن می باشد که با کاراکتر فاصله از یکدیگر جدا شده اند.

آموزش مدل با استفاده از fairseq-train

برای آموزش مدل با مقدار دهی اولیه وزن ها می بایست از گزینه‌ی encoder-embed-path و decoder-embed-path در کنار گزینه‌ی share-all-embeddings استفاده نمایید. یک مدل با حالت ثابت بودن وزن های مربوط به embedding و یک مدل با بروز رسانی وزن های embedding آموزش دهید. (برای ثابت نگه داشتن وزن های embedding می توانید از گزینه های encoder-freeze-embed و decoder-freeze-embed استفاده نمایید.)

استفاده از fairseq-generate برای ترجمه دادگان ارزیابی

پس از اتمام آموزش مدل با استفاده از دستور fairseq-generate دادگان ارزیابی را ترجمه کرده و در فایل مجزا ذخیره نمایید.

استفاده از Tensorboard برای بررسی فرایند آموزش

پس از اتمام ارزیابی مدل، به کمک Tensorboard می‌توانید نمودار پارامترهای مختلف (مانند loss) در طول فرایند آموزش را مشاهده کنید. برای رسم نمودار loss و به دست آوردن فایل log مقادیر آن (با فرمت csv) از این ابزار استفاده 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.
- توجه کنید آموزش مدل‌های ترجمه‌ی ماشینی وظیفه‌ای GPU intensive است و در صورتی که لپتاپ شما GPU ی مناسبی نداشته باشد، مجبور به استفاده ازGoogle Colab هستید. آموزش هر مدل ترجمه (برای حداقل ۵ ایپاک) ممکن است بین ۳۰ تا ۶۰ دقیقه طول بکشد و در این تمرین شما باید ۴ مدل آموزش دهید که ممکن است در مجموع از محدودیت استفاده‌ی روزانه‌ی GPU های Colab بیشتر باشد. در نتیجه مدیریت کنید که انجام تمرین را به روزهای آخر موکول نکنید تا این محدودیت باعث ارسال با تاخیر تمرین‌تان نشود.
- 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:

Mersad.esalati@gmail.com

Pedram.rostami9@gmail.com

مهلت تحویل: ۱۴۰۲/۳/۱۳