به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

پردازش زبان طبیعی

تمرین پنجم

خرداد 1401

مقدمه

دراین تمرین قصد داریم ابزارهای ترجمه ماشینی مبتنی بر شبکه عصبی را بررسی کنیم و همچنین با روند آموزش یک مدل ترجمه ماشینی مبتنی بر شبکه عصبی آشنا شویم.

در ادامه به معرفی سه تا از ابزاهای معروف این حوزه می‌پردازیم:

۱) ابزار MarianNMT

این ابزار بر پایه‌ی زبان C توسعه داده شده و از سریعترین ابزارهای آموزش مدل‌های ترجمه‌ی ماشینی است. از لینک زیر می‌توانید 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 ویژگی‌های 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 است و توسط شرکت Facebook ارائه شده است. 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

آموزش ترجمه انگلیسی به فارسی

برای آموزش و ارزیابی مدلهای این تمرین بخش کوچکی از پیکره انگلیسی - فارسی AFEC در نظر گرفته شده که دادگان آموزشی، آزمایشی و همچنین validation در پوشه data آورده شده است. می‌خواهیم با کمک این پیکره دو مدل ترجمه ماشینی مبتنی بر شبکه عصبی را آموزش دهیم (با هر ابزار یک مدل).

ابتدا دو تا از ابزارهای معرفی شده در قسمت مقدمه را انتخاب کنید و برای انجام تمرین موارد زیر را در نظر بگیرید:

1. همانطور که می‌دانیم در تسک NMT نیازمند پیش‌پردازش هستیم، یکی از روش‌های پیش‌پردازش اعمال bpe بر پیکره آموزشی است به نظر شما چه پیش‌پردازش دیگری برای ترجمه ماشینی مبتنی بر شبکه عصبی مفید است؟ به دلخواه خود دو روش پیش پردازش دیگر را، با ذکر دلیل، انتخاب کرده و نام ببرید.
2. به نظر شما در زبان فارسی کدام یک از این 3 نوع پیش‌پردازش از اهمیت بیشتری برخوردار است؟ چرا؟ در زبان انگلیسی چطور؟
3. همانطور که می‌دانید معماری‌های متفاوتی برای آموزش مدل‌های sequence to sequence وجود دارد. در این تمرین قصد داریم تا مدلی بر مبنای معماری transformer آموزش دهیم.
4. در این مرحله متناسب با ابزارهایی که انتخاب کرده‌اید، 10 تا از پارامترهای آموزش مدل را شرح دهید (در هر ابزار به شکل جداگانه). ابزار مورد نظر خود را نام برده و ابتدا توضیح دهید که این پارامترها چه کاری انجام می‌دهند و سپس توضیح دهید چه پارامترهایی را لازم است تا نسبت به حالت پیش فرض تغییر دهید.
5. 3 پارامتری که فکر می‌کنید در کیفیت خروجی یک مدل ترجمه ماشینی تاثیر گذار هستند ولی امکان تغییر یا تنظیم آن‌ها باتوجه به منابع موجود (سخت افزاری و محدودیت‌های ابزار) وجود ندارد را نام برده و اهمیت هر یک را مختصرا توضیح دهید. (در هر ابزار به شکل جداگانه)
6. حال می‌خواهیم 2 مدل مختلف را با استفاده از 2 ابزار انتخابی آموزش دهیم که جزییات این بخش به شرح زیر است:

   ➢ ابتدا bpe و دو روش پیش پردازش دیگر را بر روی دادگان آموزش اعمال کنید

   - چگونگی و جزئیات پیاده‌سازی این بخش را در گزارش خود ذکر کنید.

- انتخاب ابزار مناسب در این مرحله برعهده شماست و محدودیتی وجود ندارد.
- خروجی مرحله پیش‌پردازش را به شکل یک فایل جداگانه در گزارش خود بیاورید.

➢ پارامترهای لازم را متناسب با دادگان، منابع موجود و صلاحدید خود مقداردهی کنید.

- پارامترهای آموزش مدل را به شکل کامل در گزارش خود ذکر کنید.

➢ آموزش مدل را تا زمانی که در میانه آموزش قرارگیرد ادامه دهید.

- توصیه میشد تا از google colab برای حل این تمرین استفاده شود.
- 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 محدودیت‌های 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.
- لطفا علاوه برفایل 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.

7. پس از آموزش بایستی روند تغییرات Bleu دو مدل آموزش دیده را بر روی مجموعه dev با افزایش تعداد epochها نشان دهید. برای این کار می‌توانید از دستور ذخیره مدل های میانی در ابزار مورد نظر خود استفاده کنید.(حداقل 5 نقطه را در طول آموزش مدل گزارش دهید)
8. با کمک دادگان موازی test، مقدار Bleu را برای این دو مدل بر روی دادگان آزمایشی گزارش دهید.
9. 3 تا از معیارهایی که فکر میکنید برای ارزیابی یک **ابزار تولید** ماشین ترجمه مناسب هستند را نام ببرید و براساس این معیارها توضیح‌دهید به نظرتان از میان دو ابزار انتخابی کدام انتخاب بهتری بوده است؟

نکات:

- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌اي برخوردار است. لطفاً تمامی نکات و فرض‌هایی که برای پیاده‌سازی‌ها و محاسبات خود در نظر مي‌گیرید را در گزارش خود ذکر کنید.
- در گزارش خود برای تصاویر زیرنویس و برای جداول هم بالانویس اضافه کنید.
- 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.

- میتوانید تمرین را در قالب گروه های حداکثر دو نفره انجام دهید. در ضورت انجام تک نفره تمرین همه موارد باید پاسخ داده شود و تفاوتی در نمره‌دهی وجود نخواهد داشت.
- دستیاران آموزشی ملزم به اجرا کردن کدهای شما نیستند. بنابراین هرگونه نتیجه و یا تحلیلی که در شرح سوال از شما خواسته شده است را به طور واضح و کامل در گزارش بیاورید. در صورت عدم رعایت این مورد، بدیهی است که از نمره تمرین کسر میشود.
- لطفاً گزارش، فایل کدها و سایر ضمایم مورد نیاز را با فرمت زیر در سامانه مدیریت دروس ایلرن بارگذاری نمایید.
- CA5# _[Lastname1]_[StudentNumber1] _[Lastname2]_[StudentNumber2].zip
- در صورت وجود هرگونه ابهام یا مشکل میتوانید از طریق رایانامه‌های زیر با دستیاران آموزشی مربوطه خانم ایمانی پور و آقای طاهرخانی در تماس باشید:

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

طاهرخانی (mahdi.taherkhani@ut.ac.ir)