به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

پردازش زبان‌های طبیعی

تمرین شماره ۴

اردیبهشت ۱۴۰۱

فهرست سؤالات

سوال ۱ – ParsiNLU[1] dataset classification

در این سوال قصد داریم از Transformer ها در کاربرد متن استفاده کنیم. یکی از کاربردهایی که در پردازش زبان طبیعی مورد پژوهش قرار گرفته است Textual entailment می‌باشد. Textual entailment در واقع یک Classification است که با کمک آن می‌توان ارتباط بین دو جمله را فهمید. در این تسک ارتباطات به سه دسته تقسیم می‌شود:

۱- Entailment: در این حالت جمله اول اثبات می‌کند که جمله دوم صحیح است.

۲- Contradiction: نوعی تضاد بین دو جمله وجود دارد.

۳- Neutral: جملات ارتباطی به یکدیگر ندارند.

نمونه‌ایی از این سه حالت را می‌توانید در جدول زیر مشاهده کنید:

جمله‌ی اول	جمله‌ی دوم	لیبل
گروهی از مردم در نزدیکی اقیانوس هستند.	جمعیت زیادی نزدیک آب هستند.	Entailment
مردی روی صحنه مشغول نواختن گیتار است.	هیچ مردی روی صحنه گیتار نمی‌زند.	Contradiction
یک دوچرخه سوار تنها در حال پریدن در هوا است.	مردی در حال پریدن در یک استخر پر است.	Neutral

برای انجام این تسک لازم است از Contextualized word embedding ایی همچون BERT استفاده کنید.

[1] https://arxiv.org/abs/2012.06154

یکی از دیتاست‌هایی که توسط مقاله ParsiNLU منتشر شده است مربوط به تسک Textual Entailement می‌باشد. داده‌ها از طریق این لینک زیر در دسترس می‌باشد:

https://huggingface.co/datasets/persiannlp/parsinlu_entailment

۱- در ابتدا تحلیلی بر روی داده‌های train داشته باشید. آیا پیش پردازش خاصی لازم دارد؟ (در صورتی‌که جواب شما منفی است دلیلتان را توضیح دهید و در غیراینصورت موارد مورد نظر را پیاده سازی کرده و دلایلتان را شرح دهید).

۲- شبکه عصبی عمیقی طراحی کنید که به کمک آن بتوانید داده‌ها را طبقه بندی کنید. به‌منظور اعمال Word-embedding از مدل Multiningual BERT به نام XLM-RoBERTa استفاده نمایید. این مدل در Hugging Face به آدرس زیر منتشر شده است:

https://huggingface.co/xlm-roberta-base

۳- بار دیگر شبکه طراحی شده در قسمت قبل را با ParsBERT پیاده سازی کنید و آن‌ها را با هم مقایسه کنید.

(توجه کنید برای انجام این طبقه‌بندی باید مدل‌ها را روی تسک خودتان Fine-tune کنید)

سوال ۲- Multilingual classification

در این سوال به طبقه‌بندی داده‌های انگلیسی و فارسی پرداخته می‌شود که شما بایستی از مدل‌های تک‌زبانه (Mono-lingual) و چندزبانه (Multi-lingual) استفاده کنید. دیتاست train، validation و test در پوشه تمرین قرار داده شده است. داده‌ها از سه ستون source (متن انگلیسی) و targets (متن فارسی ترجمه شده) و category (شامل سه برچسب quran (کتاب قرآن)، bible (کتاب انجیل)، mizan (مجموعه شاهکارهای ادبی ترجمه شده به انگلیسی)) هستند.

در هر کدام از مراحل زیر batch_size=32، max_sequence_length=128، learning_rate=3e-5 و تعداد epochs را برابر 10 قرار دهید. برای آموزش و ارزیابی مدل از داده train و validation استفاده کرده و در نهایت متریک‌های Accuracy، F1-score و AUC را برای داده‌های test گزارش کنید. (می‌توانید از متد classification_report از کتابخانه scikit-learn) استفاده کنید.

۱- شبکه عصبی عمیقی طراحی کنید که به کمک آن بتوانید داده‌های انگلیسی را طبقه‌بندی کنید (می‌توانید از یکی از مدل های BERT، DistilBERT، RoBERTa، DistilRoBERTa و یا ... استفاده کنید).

۲- شبکه عصبی عمیقی طراحی کنید که به کمک آن بتوانید داده‌های فارسی را طبقه‌بندی کنید (از مدل ParsBERT استفاده کنید).

۳- شبکه عصبی عمیقی طراحی کنید که به کمک آن بتوانید داده‌های انگلیسی و فارسی را به همراه هم (راهنمایی: داده های انگلیسی و ترجمه شده فارسی آن را با یک تگ <SEP> به هم بچسبانید) طبقه‌بندی کنید (می‌توانید از یکی از مدل های چندزبانه مثل XLM-RoBERTa و یا ... استفاده کنید).

- ✓ در نهایت تحلیل خود را بر اساس نتایج به دست آمده بر روی داده های test گزارش دهید.
- ✓ آیا مورد سوم (استفاده از مدل های چندزبانی بر روی داده های چندزبانی) باعث بالا بردن دقت شبکه خواهد شد؟

سوال ۳- Cross-lingual zero-shot transfer learning (امتیازی)

شبکه عصبی عمیقی طراحی کنید تا با همان پارامترها و داده های train و validation سوال دوم، مدلی برای زبان انگلیسی به دست آورید و سپس همین مدل را (بدون تغییر وزن ها) روی داده‌های test زبان فارسی سوال دوم اعمال کنید. و به سوالات زیر پاسخ دهید.

(می‌توانید از یکی از مدل‌های چندزبانه مثل XLM-RoBERTa و یا ... استفاده کنید).

۱- انتظار شما از Performance مدل بر روی داده‌های test زبان فارسی، قبل از اجرای این مدل چیست؟

۲- بعد از اجرای این مدل آیا انتظارات پیشین شما برآورده شده است؟ دلیل این Performance ای که گرفته‌اید چیست؟

۳- در چه مواقعی از Cross-lingual zero-shot transfer learning استفاده می‌کنیم در واقع کاربرد آن را توضیح دهید.

نکات:

- پیاده سازی با Tensorflowیا Pytorch باید انجام شود.
- تمامی تصاویر و جداول مورد استفاده در گزارش کار باید به ترتیب دارای توضیح (caption) باشند.
- بخش زیادی از نمره شما مربوط به گزارش کار و روند حل مسئله است. خواهشا از هر گونه اطناب درگزارش کار پرهیز کرده و به موارد خواسته شده به صورت کامل پاسخ دهید.
- تمامی کدها و گزارش مربوطه بایستی در یک فایل فشرده با عنوان NLP_CA4_StudentID تحویل داده شود.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرای مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.
- توجه کنید این تمرین باید به‌صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت وجود هرگونه ابهام یا مشکل می‌توانید از طریق رایانامه‌های زیر با دستیاران آموزشی مربوطه رومینا اوجی (سوال اول) و علی ذوالنور (سوال دوم و سوم) در تماس باشید:

romina.oji@ut.ac.ir

ali.zolnour@ut.ac.ir