به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس پردازش زبان طبیعی

تمرین سوم

استاد درس: دکتر هشام فیلی

سرپرست دستیاران آموزشی: سمانه پیمانی‌راد

طراح تمرین:

پرهام سازدار

علی رمضانی

آبان ماه ۱۴۰۳

فهرست

مقدمه

این تمرین مقدمه ورود به دنیا دیپ لرنینگ در NLP است. سعی کرده ایم روند تاریخی را رعایت کنیم و معماری و تسکی که انجام میدهید طبق روند تاریخی پیشرفت حوزه انتخاب شده است.

اهداف تمرین:

- **یادگیری خوب دیپ لرنینگ در NLP**
- **مهارت لازم برای استفاده در صنعت و پژوهش**
- **آشنایی با معماری و دیتاست های مختلف**

نکات قابل توجه در هنگام پاسخ به سؤالات:

- بهتر است برای آموزش مدل‌ها زمان مناسبی در نظر بگیرید و کار را به روزهای آخر موکول نکنید.
- بدیهی است که باید از colab یا Kaggle استفاده کنید، پیشنهاد جدی ما استفاده از Kaggle است.
- اگرچه سوالات به نظر طولانی میرسند، اما به سادگی از پس همه آن ها برخواهید آمد، به شرطی که با ارامش و به هدف یادگیری و بدون استرس و نگرانی با آن ها رو به رو شوید.
- یادگیری عمیق این تمرین، نقطه عطف ورود به دنیای دیپ لرنینگ در دنیای NLP است و آنچه می آموزید، بسیار ارزشمند است.
- در تمرین هر کتابخانه ای که لازم باشد اشاره شده است، به جز موارد ذکر شده، استفاده از هیچ کتابخانه دیگری مجاز نیست، مگر با هماهنگی با TA های درس.
- حتما می بایست از پایتورچ برای نوشتن کد های خود استفاده کنید و استفاده از تنسورفلو مجاز نیست.
- استفاده از lightning برای نوشتن توابع train و validation در پایتورچ مجاز نیست و توابع باید توسط خودتان نوشته شود.
- فایل قوانین استفاده از ChatGPT را به طور دقیق مطالعه کرده و رعایت نمایید.

سوال اول: SEQUENCE LABELING (50 نمره)

در این تمرین با استفاده از شبکه ی LSTM به پیش بینی لیبل های NER (Named Entity Recognition) میپردازید، در این سوال تنها لازم است از معماری Encoder استفاده کنید.

این تسک، عملا تسکی مهم و در عین حال ساده از منظر NLP سنتی تا پیش از معرفی ترنسفورمر می باشد، در این بخش به خوبی خواهید دانست که قوت مدل ها تا پیش از معماری جدید(ترنسفورمر) چه بود.

مجموعه داده (۲۰ نمره)

دیتاست مورد استفاده در این سوال، دیتاست eriktks/conll2003 می باشد که آن را از Hugging face دریافت کرده و پیش خواهید رفت.

الف) ابتدا dataset را load کرده و به بررسی آن بپردازید، این دیتاست شامل چه ستون هایی(features) است؟(۳ نمره)

ب) میبینند که در این دیتاست به جز tag های ner، دو tag دیگر نیز وجود دارد، هر کدام چه چیزی را نمایش میدهد؟ درباره آن ها توضیح دهید(۴ نمره)

ج) از بخش train دیتاست برای ساخت کلاس vocab استفاده کنید، کلاس vocab شما باید دارای توابع زیر باشد:(۱۰ نمره)

- تابع ساخت vocab
- تابع تبدیل توکن به id
- تابع تبدیل id به توکن
- تابع که اندازه vocab را نمایش دهد

به نکات زیر دقت کنید:

- این دیتاست tokenize شده است و نیازی نیست که مجددا جملات را tokenize کنید.
- توکن های '<pad>','<unk>', '<s>', '</s>' به وکب خود اضافه کنید.
- از یک متغیر cutoff در ساخت vocab استفاده کنید و که توکن های با تکرار کمتر از cutoff را به وکب اضافه نکند، مقدار cutoff را برابر پنج قرار دهید.

د) بعد از نوشتن کلاس vocab و ساختن واژگان،(۳ نمره)

a. سایز واژگان را نمایش دهید
b. مشخص کنید که id = 0 چه توکنی است.
c. توکن "quick" دارای چه id می باشد.

بخش اول: ساخت DATASET و DATA LOADER (۱۰ نمره)

در این بخش به آماده سازی dataset و data loader بپردازید، همچنین دقت کنید که به دلیل تفاوت در طول جملات و استفاده از batch در آموزش شبکه، باید طول تمامی جملات هم اندازه باشند. در نتیجه می بایست طول جملات کوتاه تر با استفاده از توکن <pad> افزایش یابند تا هم اندازه شوند، برای این کار شما لازم است از متغیر collate_fn در ساخت data loader استفاده کنید.(نوشتن توابع ۷ نمره)

الف) توضیح دهید که استفاده از collate_fn چه برتری نسبت به پد کردن تمام جملات به اندازه بزرگترین جمله دیتاست دارد؟(۱ نمره)

ب) بعد از ساختن دیتالودر، ابعاد(shape) اولین batch دیتالودر در مجموعه train و validation را نمایش دهید.(۲ نمره)

بخش دوم: آموزش شبکه (۲۰ نمره)

الف) شبکه Encoder با ویژگی های زیر تعریف کنید:(۵ نمره)

- شامل لایه Embedding
- شامل یک لایه RNN دو طرفه(bidirection)
- شامل یک لایه Linear

بعد از آموزش شبکه برای ۱۰ ایپاک موارد زیر را نمایش دهید:

- نمودارهای های تابع هزینه و دقت روی دیتای train و validation را رسم کنید
- با استفاده از classification_report از کتابخانه sklearn، دقت بدست آمده روی هر لیبل را نمایش دهید

نکته مهم: دقت کنید که لیبل هایی که برای pad کردن استفاده کردید باید حین آموزش شبکه mask شوند(یعنی در محاسبه loss و دقت مورد استفاده قرار نگیرند) تا دقت بدست آمده واقعی باشد.

پارامتر های شبکه:

embedding_size = 64

hidden_size = 64

batch_size = 32

num_epochs = 10

learning_rate = 1e-3

ب) بخش الف را تکرار کنید، اما این بار از biLSTM برای شبکه استفاده کنید و نمودارها و classification report را نمایش دهید. :(۵ نمره)

ج) بخش ب را تکرار کنید، اما این بار از biGRU برای شبکه استفاده کنید و نمودارها و classification report را نمایش دهید. :(۵ نمره)

ج) تفاوت نتایج بخش الف و ب و ج را تحلیل کنید؟ آیا تفاوتی که میبیند قابل توجه است؟تحلیل خود را ذکر کنید.(۵ نمره)

سوال دوم: SUMMERIZATION (50 نمره + 5 نمره امتیازی)

در این بخش به تعریف یک شبکه Encoder-Decoder با استفاده از معماری GRU میپردازید، همچنین از یک دیتاست خلاصه سازی استفاده خواهید کرد.

معماری این بخش، شامل مکانیزم Attention نیز خواهد بود. اگر به شکل تاریخی نگاه کنیم، این معماری و این تسک(خلاصه سازی) دقیقا آخرین فعالیت هایی که است که پس از آن ها، معماری ترنسفورمر معرفی شد.

یادگیری مفاهیم این بخش، در یادگیری و فهم معماری ترنسفورمر اهمیت بسیاری دارد، همچنین تسک خلاصه کردن، از جمله تسک هایی است که تفاوت اساسی با تسک های طبقه بندی دارد(مثل سوال یک) و برای مدل های دیپ لرنینگ چالش جدی به شمار میرفت.

از سوی دیگر، معیار ارزیابی تسک های خلاصه سازی با سایر تسک های طبقه بندی متفاوت است.

با یادگیری خوب این بخش با توانایی و محدودیت های معماری های GRU آشنا شده و با معیارهای ارزیابی در تسک های Text Generation آشنا خواهید شد.

مجموعه داده (۱۵ نمره)

دیتاست مورد استفاده در این سوال، دیتاست EdinburghNLP/xsum می باشد که آن را از Hugging face دریافت کرده و پیش خواهید رفت.

الف) مانند سوال یک dataset را load کرده و با توجه به بزرگی dataset ، تنها از ۳۰ درصد دیتای آموزش و ۱۰ درصد دیتای ولیدیشن استفاده کنید.(۲ نمره)

ب) این دیتاست برخلاف سوال ۱، tokenize نشده است. با استفاده از regex این دیتا را tokenize کنید به طوری که :(۳ نمره)

- همه کلمات را lower کنید
- همه punctuation را حذف کنید
- از white-space به عنوان توکنایزر استفاده کنید.

ج) از بخش train دیتاست برای ساخت کلاس vocab استفاده کنید، کلاس vocab شما باید دارای توابع زیر باشد:(این کلاس vocab دقیقا با سوال یک یکسان است، و میتوانید بدون هیچ تغییری از همان استفاده کنید.)(۵ نمره)

- تابع ساخت vocab
- تابع تبدیل توکن به id
- تابع تبدیل id به توکن
- تابع که اندازه vocab را نمایش دهد

به نکات زیر دقت کنید:

- توکن های '<pad>','<unk>', '<s>', '</s>' به وکب خود اضافه کنید.
- از یک متغیر cutoff در ساخت vocab استفاده کنید و که توکن های با تکرار کمتر از cutoff را به وکب اضافه نکند، مقدار cutoff را برابر ۱۰ قرار دهید.

د) در این سوال میخواهیم از Embedding Glove به عنوان بردارهای از پیش آماده Embedding استفاده کنیم، از سایت زیر، نسخه glove.6B را دانلود کنید:(۵ نمره)

- با استفاده از بردارهای Embedding با ابعاد 100، ماتریس Embedding، برای vocab ساخته شده را شکل دهید، در نهایت ماتریسی با ابعاد زیر خواهید داشت: (Vocab_size, 100)
- دقت کنید که Embedding توکن های خاص(pad و...) را با بردار صفر جایگزین کنید.
- در نهایت برای ۳ کلمه دلخواه در vocab، ۱۰ درایه از بردار Embedding را نمایش دهید.

سایت glove:

http://nlp.stanford.edu/data/glove.6B.zip

بخش اول: ساخت DATASET و DATA LOADER (۵ نمره)

در این بخش به آماده سازی Dataset و Data loader بپردازید، همچنین دقت کنید که به دلیل تفاوت در طول جملات و استفاده از batch در آموزش شبکه، باید تمامی جملات هم اندازه باشند. در نتیجه می بایست طول جملات کوتاه تر با استفاده از توکن <pad> افزایش یابند تا هم اندازه شوند، برای این کار شما لازم است از متغیر collate_fn در ساخت دیتالودر استفاده کنید.

بخش دوم: آموزش شبکه (۳۰ نمره)

در این بخش یک شبکه Encoder-Decoder با استفاده از GRU دو طرفه(bidirectional) خواهید ساخت همچنین از مکانیزم Attention نیز درون شبکه بهره خواهید برد، به ترتیب روند زیر را طی کنید:

الف) مکانیزم Attention را تعریف کنید. (منظور این است که کد مکانیزم Attention را بزنید)(۵ نمره)

ب) شبکه Encoder-Decoder شامل مکانیزم Attention را تعریف کنید.(۱۵ نمره)

دقت کنید شبکه باید شامل لایه های زیر باشد:

- لایه Embedding که از بردارهای Glove استفاده میکند

- معماری GRU یک لایه و دو طرفه(bidirectional)
- یک لایه خطی

پارامتر های شبکه:

embedding_size = 100

hidden_size = 100

batch_size = 32

num_epochs = 10

learning_rate = 1e-3

ج) نمودار loss و ROUGE روی داده train و validation را رسم کنید. برای معیار ROUGE میتوانید از کتابخانه آماده استفاده کنید.(۷ نمره)

د) در مورد معیار ROUGE توضیح دهید، این معیار چطور نتایج را ارزیابی میکند؟(۳ نمره)

ه) سه خلاصه ثابت از دیتای Validation را بعد از هر epoch پرینت کنید، آیا در طول آموزش این جملات بهتر شده اند؟ اگر خیر چرا؟(۵ نمره امتیازی)

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان NLP-CAx-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. دقت کنید که حتما گزارشات خود را در قالب ارائه شده برای تحویل تکالیف که در سامانه برای شما بارگذاری شده است ارسال بفرمایید.
- 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. **دقت کنید که تمامی کدها باید توسط شما اجرا شده باشند و نتایج اجرا در فایل کدهای ارسالی مشخص باشد. به کدهایی که نتایج اجرای آن‌ها در فایل ارسالی مشخص نباشد نمره‌ای تعلق نمی‌گیرد.**
- 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 تأخیر 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 صفر تعلق می‌گیرد و به استاد نیز گزارش می‌گردد.**

تاریخ آپلود تمرین	۲۲ آبان ماه ۱۴۰۳
مهلت تحویل بدون جریمه	۱۰ آذر ماه ۱۴۰۳
مهلت تحویل با تأخیر، با جریمه ۱۰ درصد	۱۷ آذر ماه ۱۴۰۳