به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس پردازش زبان طبیعی

تمرین چهارم

استاد درس: دکتر هشام فیلی

سرپرست دستیاران آموزشی: سمانه پیمانی‌راد

طراح تمرین:

علی فرتوت، محمد گرجی

علی فرتوت (ایمیل)

محمد گرجی (ایمیل)

آذر ماه ۱۴۰۳

فهرست

مقدمه

موضوع تمرین1:

در سوال اول برای انجام تسک شناسایی موجودیت‌های نام‌دار (NER) از مدل XLM-RoBERTa استفاده خواهیم کرد. برای این منظور، با بهره‌گیری از کتابخانه Hugging Face، فرآیند پیش‌پردازش داده‌ها نیز انجام خواهید داد.

موضوع تمرین2:

در این سوال به آموزش یک مدل llm برای تسک summarization استفاده میکنید. و در ادامه با تکنیک هرس شبکه عصبی آشنا میشوید و به مقایسه نتایج آن میپردازید.

سوال اول: - (50 نمره)

شناسایی موجودیت‌های نام‌دار (NER) یکی از مهم‌ترین وظایف در پردازش زبان طبیعی (NLP) است که هدف آن شناسایی و استخراج موجودیت‌های معنادار مانند اسامی افراد، سازمان‌ها، مکان‌ها، تاریخ‌ها و سایر عناصر مهم از متن می‌باشد. مدل‌های NER نقش کلیدی در کاربردهای مختلفی مانند استخراج اطلاعات، تحلیل احساسات، ترجمه ماشینی و سیستم‌های پاسخ‌گویی به پرسش‌ها ایفا می‌کنند.

وظیفه شما در این سوال این است که ابتدا از دو سورس زبانی مختلف دیتاست فارسی و انگلیسی را بارگزاری کنید. سپس بعد از مراحل پیش‌پردازش؛ مدل مربوطه را بارگزاری کرده و بعد از تغییرات مد نظر مدل، مدل را آموزش میدهید. در این سوال از مدل XLM-RoBERTa استفاده خواهیم کرد. برای این منظور، با بهره‌گیری از کتابخانه Hugging Face، فرآیند پیش‌پردازش داده‌ها را نیز انجام خواهید داد.

بهتر است برای آموزش مدل‌ها زمان مناسبی در نظر بگیرید و کار را به روزهای آخر موکول نکنید.

مجموعه داده (5 نمره)

برای این سوال از دیتاست معروف Conll برای زبان انگلیسی استفاده میکنیم و برای زبان فارسی از ترکیب دو دیتاست PEYAM و ARMAN استفاده میکنیم که در کد زیر در دسترس شما قرار داده شده است. فرمت دیتاست به شکل BIO میباشد که در کلاس با آن آشنا شدید. دیتا به داده های آموزش و تست تقسیم شده است. نیازی به انجام دوباره ندارید.

دیتاست مجموعه‌ای از توکن ها هستند که به همراه آن‌ها لیبل آن وجود دارد. این لیبل‌ها به 2 صورت می‌باشد:

1. لیبل به صورت کلمه
2. لیبل به صورت Id

برای دانلود دیتاست خواسته شده میتوانید از کد زیر استفاده کنید.

```
from datasets import load_dataset
import transformers

persian_dataset = load_dataset("AliFartout/PEYMA-ARMAN-Mixed")
english_dataset = load_dataset("conll2003", trust_remote_code=True)
```

به سوالات زیر پاسخ دهید:

1) مشکلات دیتاست های دانلود شده چیست؟
2) برای حل مشکلات گفته چه راه‌حل‌هایی پیشنهاد میدهید؟

بخش اول: آموزش مدل (15 نمره)

پیش پردازشی با توجه به پاسخ برای قسمت دوم بخش قبلی انجام بدهید. میتوانید از این لینک برای پیش پردازش استفاده کنید. بعد از انجام پیش پردازش برای هر دیتاست باید دو دیتاست را با هم ادغام کنید.

توجه: کتابخانه های مجاز برای این بخش (Pandas, Huggingface Transformers, Dataset etc.) می‌باشد.

بخش دوم: آموزش مدل (25 نمره)

بعد از پیش‌پردازش انجام شده شما باید یک مدل TokenClassifiaction را پیاده سازی کنید.

توجه شود که برای پیاده سازی از کلاس RobertaForTokenClassification نمی‌توانید استفاده کنید. برای اینکار یک کلاس تعریف کنید و از مدل xlmRoberta به عنوان Backbone استفاده کرده و ازlogit های خروجی آن استفاده کنید تا یک مدل Token Classification انجام بدهید. آموزش مدل های Token Classification با روش‌های دیگر اندکی فرق دارد. برای این کار نیاز به Align کردن دارید. میتوانید از این لینک استفاده کنید.

به سوالات زیر پاسخ دهید.

1) چرا باید Alignment انجام شود؟
2) راه حل پیشنهادی چیست؟ این کار چگونه مشکل به وجود آمده را حل میکند؟

حال بعد از مراحل انجام شده مدل را به اندازه 1 ایپاک آموزش دهید.

برای نتایج متریک ها به علاوه loss validation کافی میباشد. (نیازی به کشیدن نمودار نیست فقط log آموزش در notebook وجود داشته باشد.)

برای آموزش ممکن است به مشکل Out of Memory بخورید. برای رفع مشکل میتوانید به این لینک مراجعه کنید. همچنین برای راحتی کار شما میتوانید از مقادیر زیر برای TrainingArguments استفاده بکنید.

```
training_args = TrainingArguments(
    output_dir="Roberta-fa-en-ner", log_level="error",
num_train_epochs=1,
    gradient_checkpointing=True,
    eval_accumulation_steps=10,
    per_device_train_batch_size=24,
    per_device_eval_batch_size=24,
    seed=42,
    logging_strategy="steps", eval_strategy="steps",
```

```
    save_steps=1e6, weight_decay=0.01, disable_tqdm=False,
    logging_steps=1e6, eval_steps=400, push_to_hub=False)
```

بخش سوم: متریک‌ها (5 نمره)

برای متریک ها شما باید علاوه بر accuracy، f1-score، recall و precision نیز استفاده بکنید.

علت تاکید بر کافی نبودن accuracy را بیان کنید.

خروجی مدل آموزش داده شده را به ازای یک جمله دلخواه نشان دهید.

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

در این تمرین عملی، شما با یکی از موضوعات پیشرفته و جذاب در حوزه هوش مصنوعی و پردازش زبان طبیعی، یعنی مدل‌های زبانی بزرگ (Large Language Models - LLMs) کار خواهید کرد. هدف این تمرین، یادگیری و تجربه عملی استفاده از مدل‌های زبانی در سناریوهای مختلف، بهبود عملکرد آن‌ها از طریق تنظیم دقیق، و بهینه‌سازی مدل‌ها برای کاربردهای واقعی است.

تمرین به سه بخش اصلی تقسیم شده است که در هر بخش، با یکی از مراحل کلیدی کار با مدل‌های زبانی آشنا می‌شوید:

1. در بخش اول، از یک مدل قدرتمند و از پیش‌آموزش‌دیده (LLAMA-3.2-1B) استفاده خواهید کرد تا وظیفه خلاصه‌سازی متنی را بدون هیچ تنظیم اضافی انجام دهید. شما دو سناریو مهم Zero-Shot و Few-Shot را تجربه کرده و تاثیر استفاده از مثال‌های کمکی بر عملکرد مدل را تحلیل می‌کنید.
2. در بخش دوم، به سراغ تنظیم دقیق (Fine-Tuning) یک مدل کوچک‌تر (SMOLLM2-360M) می‌روید. هدف این مرحله، بهینه‌سازی مدل برای وظیفه خاص خلاصه‌سازی است. در اینجا، اهمیت تنظیم دقیق، انتخاب پارامترهای مناسب و استفاده از داده‌های آموزشی برای تقویت مدل را درک خواهید کرد.
3. در بخش سوم، با یکی از تکنیک‌های کاربردی در بهینه‌سازی مدل‌های بزرگ، یعنی هرس مدل (Pruning)، آشنا می‌شوید. شما وزن‌های مدل تنظیم‌شده را کاهش داده و تاثیر این هرس را بر عملکرد و کارایی مدل بررسی می‌کنید. سپس با تنظیم دقیق مجدد (Re-fine-tuning)، تلاش می‌کنید تا افت عملکرد ناشی از هرس را جبران کنید.

این تمرین نه‌تنها شما را با ابزارها و تکنیک‌های رایج در کار با مدل‌های زبانی آشنا می‌کند، بلکه شما را به تحلیل نتایج و ارائه بینش‌های عمیق‌تر درباره نحوه استفاده از مدل‌ها در کاربردهای واقعی تشویق می‌کند.

مجموعه داده (5 نمره)

1. مجموعه داده CNN/DailyMail برای خلاصه‌سازی متنی استفاده می‌شود. با مراجعه به مستندات، نحوه بارگذاری این مجموعه داده با استفاده از کتابخانه HuggingFace Datasets را توضیح داده و پیاده‌سازی کنید.
    - داده‌ها را به سه بخش: آموزشی (5000 نمونه)، اعتبارسنجی (500 نمونه)، و آزمایشی (100 نمونه) تقسیم کنید.
    - ساختار مجموعه داده را بررسی کرده و یک نمونه را به همراه "متن" و "خلاصه" چاپ کنید. همچنین کمترین، بیشترین و میانگین طول متن و خلاصه را بدست آورید.

نکات:

- به نسخه 3.0.0 از مجموعه داده دسترسی داشته باشید.

مفاهیم:

- پردازش داده‌ها: آیا نیاز به پیش پردازش دادگان برای این سوال وجود دارد؟ توضیح دهید.

بخش اول: استفاده از LLAMA-3.2-1B (15 نمره)

1. مدل LLAMA-3.2-1B را با استفاده از کتابخانه Transformers بارگذاری کنید.
    - این مدل را در حالت کوانتایزر 4 بیتی (4-bit Quantization) پیکربندی کنید.
    - توکنایزر مناسب را بارگذاری کرده و مطمئن شوید که pad token تعریف شده است.
2. سناریوی Zero-Shot Summarization را پیاده‌سازی کنید.
    - یک مقاله نمونه از داده‌های آزمایشی انتخاب کنید و خلاصه‌ای از آن با مدل تولید کنید.
    - عملکرد مدل را با استفاده از متریک‌های ROUGE و BERTScore ارزیابی کنید.
3. سناریوی Few-Shot Summarization را پیاده‌سازی کنید.

- سه نمونه از داده‌های آموزشی را به صورت تصادفی به عنوان مثال برای مدل انتخاب کنید.
- مدل را به کمک این مثال‌ها هدایت کرده و خلاصه‌ای برای یک مقاله آزمایشی تولید کنید.
- عملکرد مدل را با استفاده از متریک‌های ROUGE و BERTScore ارزیابی کنید، همچنین نتایج را تحلیل کرده و تفاوت بین سناریوهای Zero-Shot و Few-Shot را شرح دهید.

نکات:

- برای تنظیم 4-bit Quantization، از کلاس BitsAndBytesConfig در Transformers استفاده کنید.
- برای ارزیابی ROUGE و BERTScore، از بسته‌های rouge_score و bert_score استفاده کنید.
- میتوانید در Few-Shot Summarization، از مقادیر بیشتری از داده‌های آموزشی به‌عنوان ورودی مدل استفاده کنید.

مفاهیم:

- در حوزه تولید متن ارزیابی خروجی مدل های زبانی با چه چالش هایی مواجه است؟
- معیار های ROUGE و BERTScore را به تفصیل توضیح دهید و مزایا و معایب هر کدام را مخصوصا برای خلاصه سازی متن توضیح دهید.
- چرا در Quantization از کوانتایزر 4 بیتی استفاده می‌کنیم؟ تاثیر آن بر دقت و سرعت چیست؟
- چه معیارهایی برای انتخاب بهترین نمونه های ورودی در سناریوی Few-Shot وجود دارد؟
- به نظر شما در این سوال استفاده از نسخه Instruct-Tune شده مدل بهتر است یا نسخه ساده؟ این دو نسخه چه تفاوت‌هایی با یکدیگر دارند

بخش دوم: تنظیم دقیق شبکه SMOLLM2-360M (20 نمره)

1. مدل SMOLLM2-360M را بارگذاری کرده و برای خلاصه‌سازی متنی تنظیم دقیق کنید.
   - طول ورودی‌ها را با توجه به اعداد بدست آمده در سوال بخش مجموعه داده محدود کنید.

2. مدل تنظیم‌شده را ذخیره کرده و چند نمونه تصادفی از مجموعه داده آزمایشی را خلاصه کنید.
    - خلاصه‌های تولیدشده را با خلاصه‌های مرجع مقایسه کنید.
    - عملکرد مدل را با استفاده از متریک‌های ROUGE و BERTScore ارزیابی کنید.
3. تحلیل کنید که تنظیم دقیق چه تاثیری بر عملکرد مدل داشته است. آیا تنظیم دقیق باعث بهبود خلاصه‌ها شده است؟ دلایل خود را ارائه دهید.

نکات:

- از کلاس SFTTrainer برای تنظیم دقیق استفاده کنید.
- مدل را پس از تنظیم دقیق ذخیره کرده و برای تولید خلاصه از آن استفاده کنید.

مفاهیم:

- توضیح دهید در چه سناریویی استفاده از مدل های عام منظوره بزرگ (بدون آموزش) و در چه شرایطی تنظیم دقیق مدل های کوچکتر منطقی و یا ضروری است؟
- استفاده از دقت عددی نصفه (FP16) چه تاثیری بر کارایی مدل دارد و چگونه صورت میگیرد؟
- در مورد مدل های زبانی سری SMOLLM2 توضیح دهید و معماری آن را به صورت کلی شرح دهید؟
- توضیح دهید چه تکنیک هایی برای بدست آمدن و آموزش مدل های زبانی کوچک با دقت قابل قبول وجود دارد؟ به نظر شما در مورد مدل SMOLLM2 کدام یک از این تکنیک ها استفاده شده است.
- به نظر شما برای تنظیم دقیق مدل استفاده از نسخه Instruct-Tune شده بهتر است یا نسخه ساده؟ چرا؟

بخش سوم: هرس شبکه (20 نمره)

1. مدل تنظیم‌شده را بارگذاری کنید و هرس را اعمال کنید.
   - با استفاده از روش L1 Unstructured Pruning، 30 درصد از وزن‌های مدل را هرس کنید.
   - پراکندگی (Sparsity) هر لایه و پراکندگی کلی مدل را محاسبه کنید.
2. عملکرد مدل هرس‌شده را با مجموعه داده آزمایشی ارزیابی کنید.
   - مقادیرROUGE و BERTScore را با مقادیر مدل تنظیم‌شده مقایسه کنید.
   - تحلیل کنید که هرس چه تاثیری بر عملکرد مدل داشته است.
3. مدل هرس‌شده را دوباره تنظیم دقیق کنید.
   - تنظیم دقیق مدل هرس‌شده را انجام داده و تاثیر آن بر عملکرد را بررسی کنید.
   - آیا تنظیم دقیق مجدد توانسته افت عملکرد ناشی از هرس را جبران کند؟

نکات:

- از ماژول torch.nn.utils.prune برای اعمال هرس استفاده کنید.
- برای تحلیل پراکندگی، تعداد وزن‌های صفر و کل وزن‌ها را محاسبه کنید.
- پس از هرس، از SFTTrainer برای تنظیم دقیق مجدد استفاده کنید.

مفاهیم:

- 3 مورد از تکنیک های مختلف هرس مدل های زبانی را توضیح دهید. تکنیک های هرس را از نظر ساختارمندی یا بدون ساختار بودن و همچنین محلی یا فراگیر بودن بررسی کنید و مزایا و معایب هر کدام را نام ببرید.

- هرس مدل در مقابل دیگر روش های فشرده سازی مدل های زبانی چه مزایا و معایبی دارد؟ (تکنیک های دیگر اعم است از Quantization, Distillation, low-rank factorization)
- تاثیر هرس بر عملکرد: چگونه هرس باعث کاهش حافظه و سرعت پردازش می‌شود؟ چه تاثیری بر دقت دارد؟
- تنظیم دقیق مجدد پس از هرس: چرا تنظیم دقیق مجدد می‌تواند عملکرد مدل هرس‌شده را بهبود دهد؟

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان NLP-CA4-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، 14 روز زمان تحویل بدون جریمه دارید که تنها از 7 روز آن برای هر تمرین می‌توانید استفاده کنید، در صورتی که این 14 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	18 آذر 1403
مهلت تحویل بدون جریمه	28 آذر 1403
مهلت تحویل با تأخیر، با جریمه **10** درصد	5دی 1403