به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس یادگیری عمیق با کاربرد در بینایی ماشین و پردازش صوت

تمرین شماره ۴

خرداد ۱۴۰۱

فهرست

مقدمه

با حل سؤال اول، شما با مفهوم شبکه‌های ترنسفورمری و استفاده از کتابخانه‌ی معروف ترنسفورمرز[1] آشنا خواهید شد. همچنین با به کار بردن این نوع شبکه‌ها در مسائل حوزه‌ی پردازش گفتار، با نمونه کاربردهای احتمالی شبکه‌های عمیق در مسائل روزمره و پر استفاده آشنا خواهید شد.

هدف سوال دوم استفاده از مدل‌های ترنسفورمری بر روی داده‌های متنی می‌باشد. این شبکه‌ها در کاربردهای زیادی مورد استفاده قرار می‌گیرند. از جمله این کاربردها که در این تمرین نیز پوشش داده می‌شود کاربرد سیستم‌های پرسش و پاسخ می‌باشد که در چند سال اخیر مورد توجه بسیاری از پژوهشگران حوزه متن قرار گرفته است

[1] https://huggingface.co/docs/transformers/main/

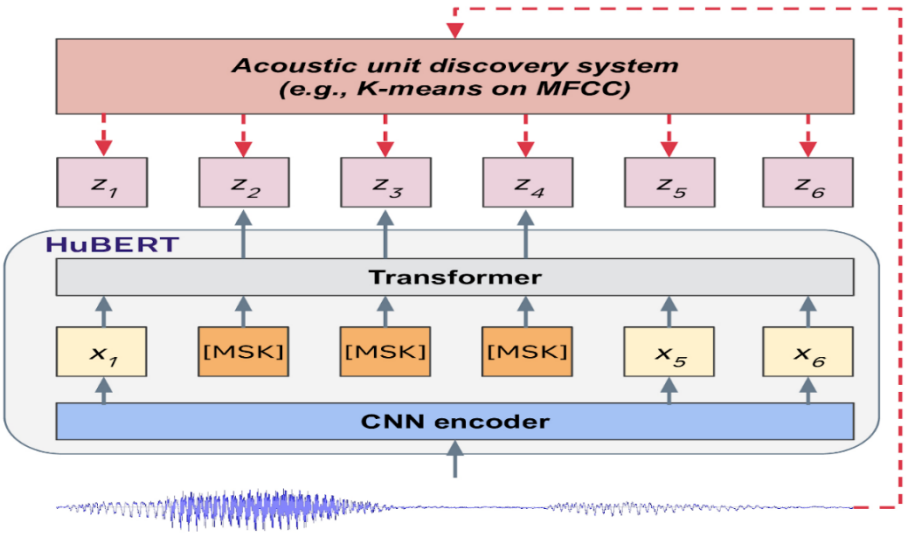
سوال ۱

در این سؤال قصد داریم که با استفاده از کتابخانه‌ی معروف ترنسفورمرز، مدل ترنسفورمری Hubert را برای دو کاربرد مختلف زبانی استفاده نماییم. مدل‌های ترنسفورمری مدل‌های کلانی هستند که اصولاً روی دادگان بسیار انبوه آموزش داده شده‌اند و برای استفاده‌های خرد نیز با روش‌های مرسوم یادگیری انتقالی بسیار کاربرد دارند.

در این سؤال با تمرکز بر روی مسائل تبدیل خودکار صوت به متن و هم‌چنین تشخیص کلید واژه سعی به بررسی توانایی یادگیری انتقالی در مدل‌های ترنسفورمری متداول می‌نماییم.

معرفی مدل HUBERT

مدل Hubert [۱] یک مدل ترنسفورمری صوتی است که بر پایه‌ی مدل Bert [۲] طراحی شده است. ساختار این مدل، مانند مدل Bert تنها یک بخش رمزگذار [۱] دارد و مشابه، با روش یادگیری خود نظارتی [۲] آموزش داده می‌شود.



شکل – ۱مدل ترنسفورمری هیوبرت

[۱] Encoder

[۲] Self supervised

همانطور که در شکل ۱قابل مشاهده است، در ابتدا سیگنال صوتی خام به قسمت‌هایی با اندازه یکسان تقطیع[۱] شده و به یک رمزنگار[۲] کانولوشنی داده می‌شود که مشابه لایه‌ی نهانگر[۳] در دادگان متنی است. پس از عبور از این لایه، یک توصیف نهان تولید می‌شود که به‌عنوان ورودی به یک لایه رمزگذار موقعیت[۴] داده می‌شود.

در لایه‌ی رمزگذار موقعیت، یک بردار ثابت یاد گرفته می‌شود و این بردار ثابت با ورودی جمع گشته تا تعبیری از موقعیت قطعه ورودی در دنباله خروجی از این لایه به وجود بیاید. سپس خروجی این لایه به رمزگشای ترنسفورمری داده می‌شود تا یک توصیف غنی با استفاده از رویکرد «توجه‌محور»[۵] به‌دست آید.

در این مسئله به خاطر اینکه از مدل سبک‌تری استفاده کنیم که در محیط گوگل کولب قابل استفاده باشد، از مدل DistilHubert [۳] استفاده می‌کنیم که تعداد پارامترهای آن ۲۵٪ مدل اصلی است و با وجود اینکه نسبت به مدل اصلی ضعیف‌تر است اما امکان آزمایش‌های مختلف را به ما می‌دهد. ما از نسخه‌ای از پیش آموزش داده شده‌ی انگلیسی این مدل استفاده خواهیم کرد که لینک آن پاورقی قابل رؤیت است.[۶]

کتابخانه‌ی TRANSFORMERS

این کتابخانه یکی از معروف‌ترین و پرکاربردترین کتابخانه‌های حال حاضر در صنعت جهت حل مسائل صوت و متن می‌باشد که توسط کمپانی Huggingface به‌صورت متن‌باز در حال توسعه می‌باشد. مدل‌های ترنسفرموری بسیاری در این کتابخانه پیاده‌سازی شده‌اند که به همراه مخازن داده‌ی موجود در کتاب‌خانه‌ی datasets[۷] اکوسیستم کاملی را برای محققین و مهندسین فعال در حوزه‌ی یادگیری عمیق در زمینه‌ی صوت و متن فراهم می‌کند.

---

[۱] Segment
[۲] Encoder
[۳] Embedder
[۴] Positional Encoding Layer
[۵] Attention-based
[۶] https://huggingface.co/ntu-spml/distilhubert
[۷] https://huggingface.co/docs/datasets/index

همچنین در آخرین سیر افزونه‌های اکوسیستم Huggingface، محیط spaces [۱] امکان بارگذاری مدل‌های یادگیری ماشین را به علاقه‌مندان جهت نشان دادن خروجی‌های ابتدایی می‌دهد.

پایگاه‌های داده

در این سؤال از مجموعه داده superb [۲] که تجمیعی از چند مجموعه داده پرکاربرد در حوزه‌ی پردازش صوت می‌باشد استفاده خواهیم نمود. با توجه به آن‌که این مجموعه داده مرتب در حال تغییر است، ما از نسخه‌ی ۲.۲.۲ این مجموعه داده استفاده خواهیم نمود.

برای مسئله‌ی تبدیل صوت به متن، از بخش "asr" این مجموعه داده که مبتنی بر مجموعه‌ی داده‌ی Librispeech [۳] می‌باشد، استفاده خواهیم نمود. لازم به ذکر است این مجموعه داده انگلیسی است.

برای مسئله‌ی تشخیص کلیدواژه‌ها، از بخش "ks" این مجموعه داده استفاده خواهیم کرد که مبتنی بر دادگان SpeechCommands [۴] می‌باشد که یکی از پرکاربردترین مجموعه دادگان برای ارزیابی روش‌های مختلف در این حوزه می‌باشد. این مجموعه داده نیز انگلیسی می‌باشد.

تابع هزینه و معیار ارزیابی

در بخش‌های عملی این سؤال، تابع هزینه را برای دادگان آزمایشی، اعتبارسنجی (هر دو در طول آموزش) و ارزیابی (پس از اتمام آموزش) گزارش نمایید. همچنین معیارهای ارزیابی نیز باید برای دادگان اعتبارسنجی در طول آموزش، و برای دادگان ارزیابی پس از اتمام آموزش گزارش شوند.

تابع هزینه‌ی بخش تبدیل صوت به متن، تابع هزینه‌ی CTC می‌باشد و معیار ارزیابی مورد استفاده در این سؤال معیار WER می‌باشد، که توضیحات هر دو در اسلایدهای مربوط به درس آورده شده‌اند. همچنین تابع هزینه بخش تشخیص کلید واژه، تابع هزینه‌ی نام‌آشنای آنتروپی متقاطع [۵] است که برای طبقه‌بندی

---

۱ https://huggingface.co/spaces
۲ https://huggingface.co/datasets/superb
۳ http://www.openslr.org/۱۲
۴ https://www.tensorflow.org/datasets/catalog/speech_commands
۵ Cross Entropy

چند دسته به کار می‌رود. معیارهای ارزیابی نیز دقت[۱]، صحت[۲] و بازیابی[۳] به‌صورت میکرو و ماکرو می‌باشند. همچنین برای دادگان ارزیابی در این بخش ماتریس درهم‌ریختگی [۴] نیز باید گزارش شود.

موارد قابل توجه عملیاتی

در این بخش موارد عملیاتی مربوط به این سؤال را مرور می‌کنیم. همان طور که در جدول ۱ قابل مشاهده است، ابر پارامترهای اصلی این سؤال به‌صورت زیر خواهند بود.

جدول ۱ – ابر پارامترهای مهم در این سؤال

Parameter	Value
Learning Rate	۱e-۳
Weight Decay	۵e-۳
Num Warmup Steps	۲۵۰
Sampling Rate	۱۶khz

```
from datasets import load_dataset
## ASR
dataset = load_dataset("superb", "asr", revision="2.2.2")
## KS
dataset = load_dataset("superb", "ks", revision="2.2.2")
```

کد ۱– قطعه دستور بارگذاری مجموعه داده‌های سؤال ۱

---

[۱] Accuracy
[۲] precision
[۳] Recall
[۴] Confusion Matrix

برای بارگذاری مجموعه داده‌ها لطفاً از دستورهای کد ۱ استفاده نمایید تا نسخه‌های مجموعه داده‌های استفاده شده در طول این سؤال یکی باشد.

برای یادگیری پیشنهاد می‌شود که از کلاس Trainer خود کتابخانه ترنسفورمرز استفاده کنید، که گزینه‌های زیادی جهت سهولت و تکرارپذیری روند آموزش در اختیار کاربران قرار می‌دهد. از مهم‌ترین این گزینه‌ها، موارد مربوط به کنترل روند ثبت معیارهای ارزیابی و تابع هزینه در طول آموزش، و هم‌چنین دقت محاسبات انجام شده (fp۱۶، fp۳۲ یا دقت مخلوط[۱]) می‌باشد. برای اطلاعات بیشتر به مستندات کتابخانه رجوع نمایید.

هم‌چنین در طول یادگیری وزن‌های پشته‌ی[۲] اصلی مدل ترنسفورمری که در کتابخانه‌ی ترنسفورمرز به‌عنوان استخراجگر ویژگی از آن یاد می‌شود را قفل کنید تا ثابت بماند. در این حالت شما یادگیری بسیار سریع‌تری خواهید داشت. برای این کار می‌توانید تابع freeze_feature_extractor را بر روی مدل خود قبل یادگیری صدا نمایید. در صورتی که از کلاس Trainer استفاده نمی‌کنید، باید به‌صورت دستی وزن‌های مربوط به لایه‌ی استخراجگر ویژگی را تثبیت نمایید.

بخش اول

در این بخش در مورد مفاهیم اصلی مورد بحث در ساختار مدل و مقاله و کد سؤال می‌شود.

۱. چرا در آموزش Hubert یک روند جدای آموزشی برای خوشه‌بندی صوت داریم؟
۲. چرا در آموزش Hubert از تابع هزینه‌ی Cross Entropy استفاده می‌شود و هدف از این طبقه‌بندی چند کلاسه چیست؟
۳. فرآیند تقطیر[۳] را مطالعه کنید و توضیح دهید که چگونه این فرآیند در DistilHubert پیاده‌سازی شده است. (خیلی خلاصه)

---

[۱] Mixed precision

[۲] Backbone

[۳] Distillation

۴. با مطالعه‌ی اسناد کتابخانه‌ی Huggingface، ورودی‌های اصلی مدل HubertForCTC را برای انجام یک پیش‌بینی توضیح دهید و بگویید که هر کدام چه نقشی ایفا می‌کنند و در کجای ساختار مدل به‌صورت ورودی مورد استفاده قرار می‌گیرند؟ (به‌طور مثال نقش input_ids چیست؟)

بخش دوم

مدل DistilHubert را با کلاس HubertForCTC بارگذاری کنید، پشته‌ی مدل (وزن‌های بخش از پیش تمرین داده شده) را تثبیت[۱] کنید تا در طول آموزش تغییر نکنند و بر روی مجموعه داده supreme بخش asr با مشخصاتی که قبلاً توضیح داده شده‌اند آموزش دهید و موارد قابل گزارش را ثبت و گزارش کنید.

بخش سوم

این بار مسئله‌ی تشخیص «کلید واژه» را بررسی می‌کنیم. بخش KS مجموعه داده را بارگذاری کنید. نموداری از فراوانی دسته‌های مختلف داده رسم کنید و در صورت لزوم کلاس خاصی از داده را حذف نمایید. حال می‌خواهیم مدل را آموزش دهیم:

۱- با توجه به اینکه این دادگان دچار مشکل عدم توازن فراوانی هستند، چه راهکارهایی را می‌توان برای آموزش بهتر مدل اتخاذ کرد؟

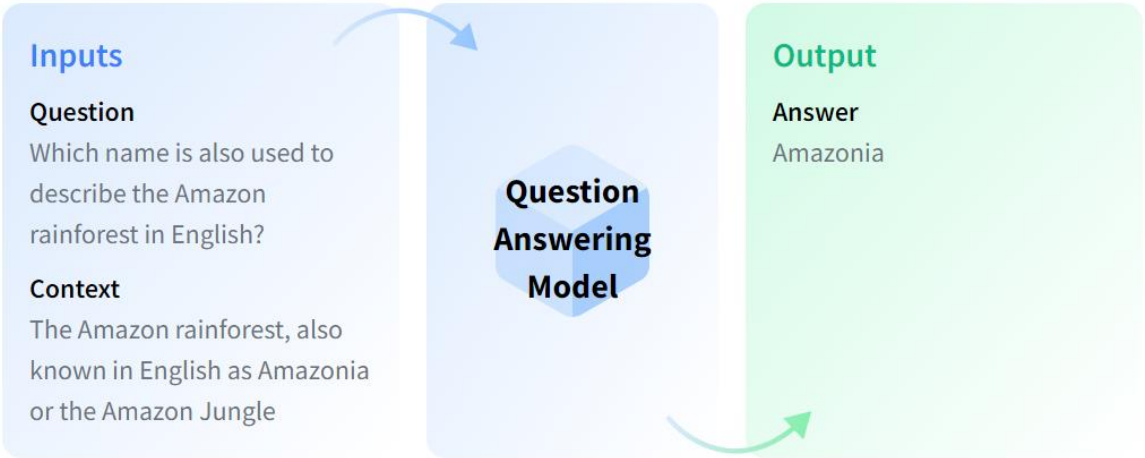
۲- در این مسئله با پیاده‌سازی یک تابع هزینه‌ی وزن‌دار، مشکل عدم توازن را بهبود ببخشید. (در صورتی که از Trainer استفاده می‌کنید باید یک زیر کلاس جدید درست کنید.)

نهایتاً مدل را آموزش دهید و موارد قابل گزارش را ثبت و گزارش کنید.

[۱] Freeze

سوال ۲ (امتیازی- ۲۰ درصد نمره امتیازی)

یکی از مسائلی که در پردازش زبان طبیعی مورد توجه قرار گرفته است، مسئله پرسش و پاسخ[۱] می‌باشد. در این مسئله یک متن داده می‌شود و با توجه به آن باید پاسخ سوال مورد نظر را یافت کرد. برای مثال شکل ۲نمونه‌ایی از پاسخ تولید شده از سوال مورد نظر با استفاده از متن داده شده می‌باشد.



شکل – ۲نمونه‌ای از عملکرد مدل پرسش و پاسخ

مجموعه دادگان

در زبان انگلیسی انواع مختلفی از مجموعه دادگان برای مسئله پرسش و پاسخ وجود دارد. یکی از معروف‌ترین آن‌ها SQuAD[۲] می‌باشد که ما در این تمرین از نسخه دوم این مجموعه داده استفاده می‌کنیم. SQuAD در واقع مجموعه دادگان درک مطلب می‌باشد که شامل سؤال‌هایی است که افراد آن‌ها را از مجموعه متون سایت ویکی‌پدیا جمع‌آوری کرده‌اند. دارای سه بخش سوال، متن و پاسخ می‌باشد. در این مجموعه داده پاسخ به هر سوال بخشی از متن یا گستره‌ای از قسمت خواندنی مربوطه است، یا ممکن است سوال بی پاسخ باشد.

---

[۱] Question-Answering

[۲] **S**tanford **Qu**estion **A**nswering **D**ataset

برای دسترسی به دادگان آموزش و اعتبار سنجی و ارزیابی می‌توانید به این لینک زیر مراجعه کنید:

https://rajpurkar.github.io/SQuAD-explorer/

بخش اول

با استفاده از مدل زبانی Bert[1] که از جمله مدل‌های زبانی ترنسفورمری[2] می‌باشد. مدلی طراحی و پیاده‌سازی کنید. ساختار مدل خود را توضیح دهید. (توجه کنید که می‌بایست آن‌را بر روی داده‌های داده شده FineTune کنید). برای آموزش مدل خود دو تا از از مدل‌های DistilBERT, Bert, XLM, XLNet را انتخاب کنید و خروجی را با هم مقایسه کنید.

بخش دوم

برای این تسک از چه معیارهایی استفاده می‌شود؟ آن‌ها را نام ببرید و هر کدام را توضیح دهید و عملکرد مدلتان را بر روی آن بیان کنید.

[1] Bidirectional Encoder Representations from Transformers

[2] transformer

منابع

[۱] W.-N. Hsu *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," vol. ۲۹, pp. ۳۴۵۱-۳۴۶۰, ۲۰۲۱.

[۲] J. Devlin, M.-W. Chang, K. Lee, and K. J. a. p. a. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," ۲۰۱۸.

[۳] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit BERT," in *ICASSP ۲۰۲۲-۲۰۲۲ IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ۲۰۲۲, pp. ۷۰۸۷-۷۰۹۱: IEEE.

نکات:

- مهلت تحویل این تمرین تا ۳۰ خرداد است.
- انجام این تمرین به‌صورت یک نفره می‌باشد.
- شما قادر نیستید هیچ تمرینی را با بیش از ۷ روز تاخیر بارگذاری کنید (دقیقاً ۷ روز پس از مهلت آپلود، سامانه بسته خواهد شد).
- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است. لطفاً تمامی نکات و فرض هایی که برای پیاده سازی ها و محاسبات خود در نظر می‌گیرید را در گزارش ذکر کنید. **دقت داشته باشید** ۵۰ درصد از نمره تمرین شما مربوط به گزارش است.
- کدهای خود را به‌صورت عکس در داخل گزارش کپی نکنید و با فرمتی مناسب آن را در گزارش قرار دهید.
- داخل کدها کامنت‌های لازم را قرار دهید و تمامی موارد مورد نیاز برای اجرای صحیح کد را ارسال کنید.
- الزامی به ارائه توضیح جزئیات کد در گزارش نیست. اما باید نتایج به‌دست آمده را گزارش و تحلیل کنید.
- گزارش را در قالب تهیه شده که روی صفحه درس در سامانه eLearn بارگذاری شده، بنویسید. در صورت تمایل می‌توانید از Latex نیز برای نوشتن گزارش استفاده نمائید.
- در گزارش خود برای تصاویر زیرنویس و برای جداول هم بالانویس اضافه کنید.
- برای انجام این تمرین فقط مجاز به استفاده از زبان برنامه نویسی Python هستید و امکان استفاده از کتابخانه‌های یادگیری عمیق نظیر Tensorflow و PyTorch را ندارید.
- از آدرس دهی مطلق در کدهای خود استفاده نکنید و به‌جای آن از آدرس دهی نسبی استفاده نمایید.
- فایل‌های ارسال شده باید به فرمت py. باشد و از ارسال فایل تمرین‌ها به‌صورت ipynb خودداری نمائید. همچنین ساختار کلی کدهای شما باید حداقل شامل فایل‌های زیر باشد. همچنین پیشنهاد می‌گردد قسمت‌های بارگذاری داده، توابع مورد استفاده در فایل‌هایی مانند dataloader و utils قرار داده شود.

توضیح	نام فایل
ساختار مدل	Model
کد آموزش و اجرای مدل	Main

- کد شما باید قابلیت اجرا بر روی قسمت کوچکی از داده‌ها را داشته باشد تا دستیار آموزشی مربوطه بتواند با استفاده از کد شما در مدت زمان کوتاهی مدل شما را آموزش دهد.
- در صورت مشاهده‌ی موارد تشابه بین دو یا چند فرد در گزارش کار و یا کد، به طرفین تقلب نمره صفر داده خواهد شد. کپی برداری از کدهای آماده موجود در اینترنت و یا استفاده از کدهای افراد ترم‌های گذشته تفاوت چندانی با تقلب ندارد.
- اگر بخشی از کد را از کدهای آماده اینترنتی استفاده می‌کنید که جزء قسمت‌های اصلی تمرین نمی‌باشد، حتماً باید لینک آن در گزارش و کد ارجاع داده شود، در غیراینصورت تقلب محسوب شده و کل نمره تمرین را از دست می‌دهید
- لطفاً فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در صفحه درس در سامانه eLearn بارگذاری نمائید.

HW۴_[Lastname]_[StudentNumber].zip

- در صورت وجود هرگونه ابهام یا مشکل میتوانید از طریق رایانامه زیر با دستیار آموزشی طراح تمرین فرهود اطاعتی (سوال اول) و رومینا اوجی (سوال دوم) در تماس باشید:

farhoodetaati@gmail.com

Romina.oji@ut.ac.ir