

دانشگاه تهران پردیس دانشکدههای فنی دانشکده برق و کامپیوتر



Deep Generative Models

تمرین شماره ۳

طراحان: محمد جواد مؤمنی حسین آزاد ملکی نیکو نقویان

زمان تحویل: ۱۴۰۲/۱۰/۱۱

آذر ۱۴۰۲

فهرست

٢	سوال ۱– LLM
۵	سوال Prompt engineering -۲
۶	سوال ۳– Speech synthesis سوال ۳
۶	مقدمه
۶	مدل SpeechT5
٨	تولید گفتار به زبان فارسی
٨	مجموعه داده
٩	شبکههای کمکی
٩	پيادەسازى و آموزش
١	سیاست نمرهدهی
١	نمونه کد بارگذاری مدل ها و ابزار های لازم برای پردازش ورودی
١,	نكات پياده سازى و تحويل

سوال ۱- LLM

در این تمرین بر روی حوزه جذاب مدلهای زبانی بزرگ (LLMs) تمرکز میکنیم. وظیفه اصلی در نظر گرفته شده خلاصهسازی مکالمات میباشد بدین جهت از مجموعه داده ی TweetSumm" استفاده میباشد. هدف این که یک مجموعه داده خلاصهسازی گفتگو، شامل حدود ۸۷۹ گفتگو با خلاصههای مربوطه میباشد. هدف این تمرین آشنا شدن شما با تجربه عملی استفاده و بکارگیری مدلهای زبانی بزرگ (LLMs) است.

بررسي وظايف:

برای این تمرین از مجموعه داده "TweetSumm" که مجموعهای از گفتگوها میباشد استفاده خواهیم کرد. شما باید این گفتگوها را پردازش و سپس با استفاده از رویکردهای بیان شده، خلاصه نمایید. هدف آن است که این مکالمات را به خلاصههایی سازگار و آگاهانه تبدیل کنید. برای این کار، شما از مدل قدرتمند زبانی" ده این مکالمات را به خلاصههایی بزرگ با ۳ میلیارد پارامتر استفاده خواهید کرد.

گام اول in-context learning اول

گام اول شما، آزمایش و ارزیابی استراتژیهای مختلف in-context learning می باشد. بدین منظور لازم است روش های ذیل را آزمایش و ارزیابی نمایید:

• zero-shot : از طریق ساخت یک prompt مناسب، کارایی مدل در انجام وظیفه خلاصه سازی گفتگو را بصورت zero-shot ارزیابی نمایید.

• one-shot : در هنگام ساخت دستورالعمل و prompt از رویکرد one-shot استفاده نموده و کارایی مدل را در این رویکرد ارزیابی نمایید، آیا بهبودی ایجاد شده است؟

گام دوم (۳۰ نمره)

در این قسمت مدل را با استفاده از روش LoRA تنظیم دقیق (fine-tune) مینماییم. رویکرد LoRA بر این پایه استوار است که تفاوت بین وزنهای مدل پیش آموزش و تنظیم دقیق شده اغلب دارای رتبه ذاتی پایین (low intrinsic rank) میباشند ازینرو لازم نیست تمام وزنهای مدل آموزش ببینند و به عبارت دیگر، می توان وزنهای مدل تنظیم دقیق شده را به طور مؤثر با یک ماتریس با رتبه پایین تقریب زد. آز آنجایی که

¹ https://huggingface.co/datasets/Salesforce/dialogstudio/viewer/TweetSumm

² https://huggingface.co/stabilityai/stablelm-3b-4e1t

³ Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. (2021) 'LoRA: Low-Rank Adaptation of Large Language Models', arXiv. Available at: https://doi.org/10.48550/arXiv.2106.09685

مدلهای زبانی بزرگ میلیاردها پارامتر دارند و تنظیم دقیق و آموزش تمام این پارامترها نیاز به زیرساخت و هزینهی بالا دارد، بدین جهت از رویکرد LoRA میتوان در آموزش مدلها بهره جست.

در این قسمت ابتدا توضیح مختصری از نحوه عملکرد LoRA داشته و سپس این رویکرد را بر روی مدل rank بیاده نمایید.) فراپارامتر rank بیاده نمایید.) فراپارامتر rank بیاده نمایید.) فراپارامترهای منظور میتوانید از کتابخانه PEFT استفاده نمایید.) فراپارامترهای مدل پس از اعمال بر روی مقادیر ۱۶ و ۳۲ گذاشته و برای هر یک ابتدا گزارش دهید چند درصد از پارامترهای مدل پس از اعمال LoRA قابل آموزش هستند و چه تعداد پارامتر به مدلها افزوده شده، سپس به آموزش مدل با استفاده از رویکرد بیان شده بپردازید و نتایج را باهم مقایسه کنید. تمام فراپارامترهای بکار گرفته شده در هنگام آموزش بهمراه دلیل استفاده از آنها را ذکر کرده و توابع خطا و نحوه آموزش مدلها را بصورت مختصر توضیح دهید. (به عنوان مثال توضیح دهید LoRA را بر روی کدام یک از لایههای مدل اعمال کردهاید، همچنین توجه داشته باشید هنگام مواجه با محدودیتهای حاصله از حافظه کارت گرافیک میتوانید از رویکردهای rank کنید. در صورت بکارگیری، بهره ببرید، بدین جهت میتوانید از کتابخانههایی چون bitsandbytes استفاده کنید. در صورت بکارگیری، رویکرد مورد استفاده را بهمراه میزان کاهش حجم مدل گزارش دهید و بطور خلاصه بیان کنید این رویکردها چه مزیتها و پیامدهایی خواهند داشت.)

ارزیابی و مقایسه نتایج: در نهایت، شما نتایج رویکردهای مختلف از جمله in-context learning با روشهای مختلف از جمله Lora و one-shot و one-shot و مدلهای اموزش دیده شده با تنظیم دقیق Lora را مقایسه خواهید کرد. بدین جهت از معیار ROUGE بصورت (rouge1 , rouge2 , rougeL) برای ارزیابی کیفیت و کارایی استفاده کنید.

توجه: مراحل آماده سازی دادههای آموزش و دستورالعملها را بصورت خلاصه ذکر نمایید. در صورت استفاده از کتابخانههای آماده در هریک از مراحل بصورت انتزاعی و خلاصه نحوه عملکرد آنها را شرح دهید.

سوال ۲- Prompt engineering

در این بخش قصد داریم تاثیر رویکرد instruction tuning را بر روی تسک BBH زیر BBH زیر از بنچمارک BBH براسی کرده و نتایج را تحلیل کنیم. BBH زیر مجموعه ۲۳ تسک از big-bench است که برای ارزیابی قابلیت های استدلال چند مرحله ای مدل های زبانی بزرگ طراحی شدهاست. همچنین در انتها با دو مقاله درباره ی ایدههای بهبود روش (Cot thought (COT) آشنا می شویم.

- اً. ابتدا نیاز است تا بخش های ۴٬۳٬۲۰۱و۷ از مقاله "۴٬۳٬۲۰۱و از مقاله "<u>Models</u>" را به صورت خلاصه بیان کنید.(۵ نمره)
- ۲. از API مدل flan-t5-large استفاده کرده و با سه روش 3-shot، answer-only مجموعه داده تست را به مدل بدهید. دقت (accuracy) هر روش را حساب کرده، نمودار barplot آن را نمایش داده و نتایج به دست آمده را تحلیل کنید.(برای طراحی پرامپ ورودی از مقاله قسمت قبل کمک بگیرید و همچنین برای مثالهای روش few-shot از دیتاست آموزش می توانید استفاده کنید)(۱۰ نمره)
- Active Prompting with Chain-of- "در مقاله" few-shot ور مثالهای روش few-shot یک راه برای انتخاب مثالهای روش Thought for Large Language Models" معرفی شده است. متد و متریکهای مقاله را توضیح "Thought for Large Language Models" دهید. (۳ نمره)
- \$. در مقاله " <u>LANGUAGE MODELS</u> . ومقاله " <u>LANGUAGE MODELS</u> . ومقاله " <u>LANGUAGE MODELS</u> . ومقاله " <u>COT</u> آورده شده است، آن را توضیح دهید. چرا این روش نسبت به COT بهتر عمل می *کند*؟(۲ نمره)

توجه : لطفا فایل csv خروجی هر روش قسمت ۲ را نیز آپلود کنید.

سوال Speech synthesis -۳

مقدمه

در این سوال قصد داریم تا از مدل از پیش آموزش دیده SpeechT5 استفاده کنیم و یک مینی پروژه تبدیل متن به گفتار را برای زبان فارسی پیاده سازی نماییم. این فرایند به ما این امکان را می دهد تا عملکرد مدل ذکر شده را در شناسایی متن و تولید گفتار به زبان فارسی (که با توجه به داده های مورد استفاده برای پیش آموزش، یک زبان جدید به شمار می آید)، مورد ارزیابی قرار دهیم. از این طریق، شما با چالشهای آموزش مدلهای پیشرفته گفتاری، شامل در ک نحوه کار کرد مدل، تهیه و پیش پردازش داده های مورد نیاز و اعمال تغییرات لازم برای آموزش مدل روی یک زبان جدید آشنا خواهید شد.

مدل SpeechT5

در سال ۲۰۲۱ مایکروسافت مقاله مربوط به یک سیستم پیشرفته پردازش گفتار بنام SpeechT5 را منتشر نمود. این مدل بر پایه معماری مدل T5 میباشد، اما با توسعه و افزودن قابلیتهای مرتبط با پردازش گفتار روی معماری پیشین، اکنون قادر است که وظایف متنوعی از جمله تشخیص گفتار، تبدیل صدا، و تولید گفتار و... را انجام دهد. اگر بخواهیم دقیق تر صحبت کنیم، باید بگوییم که این مدل همزمان سه مدل گفتاری در یک معماری ثابت است:

- گفتار به متن: وظایفی چون تشخیص گفتار [†] و یا شناسایی گوینده ^۵ و...
 - متن به گفتار: تولید گفتار ۶
- گفتار به گفتار: وظایفی چون تبدیل صدا V ، از بین بردن نویز $^{\Lambda}$ ، تقویت گفتار P و...

ایده اصلی که این مدل را قادر میسازد تا برای همه کارکردهایی که بالا ذکر کردیم، قابل استفاده باشد، این است که یک مدل واحد را روی ترکیبی از دادههای تبدیل متن به گفتار، گفتار به متن، متن به متن و گفتار به گفتار آموزش دهیم! به این ترتیب، مدل به طور همزمان از متن و گفتار یاد میگیرد. نتیجه این رویکرد

Speech Recognition⁵

Speaker Identification ^a

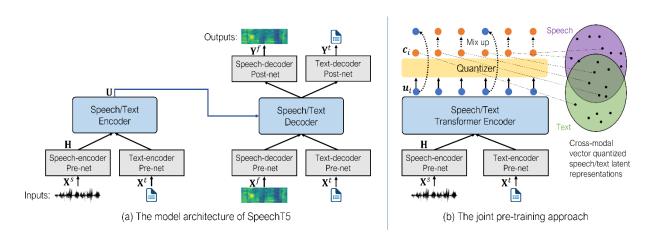
Speech synthesis^{*}

Voice conversion^v

Audio denoising^A

speech enhancement

پیش آموزش مدلی است که دارای فضایی یکپارچه از بازنماییهای پنهان ۱۰ است که هم در متن و هم گفتار مشترک است.



شکل ۱- نمای کلی از معماری مدل SpeechT5

این مدل از یک ساختار ترنسفورمری کدگذار-رمزگشا۱۱ بهره می برد و به عبارت بهتر یک مدل تبدیل دنباله به دنباله ۲۱ به شمار می آید که این امکان را فراهم می کند تا هم متن و هم گفتار را به عنوان ورودی قبول کرده و خروجی های مناسب را در هر دو فرمت تولید کند. برای این منظور از یک سری به اصطلاح پیششبکهها ۲۱ پسشبکهها ۱۲ ستفاده می شود. وظیفه پیششبکهها تبدیل متن یا گفتار ورودی به بازنماییهای پنهانی است که توسط ترنسفورمر استفاده می شود. پسشبکهها خروجی ها را از ترنسفورمر می گیرند و دوباره به متن یا گفتار تبدیل می کنند. یکی دیگر از ویژگی های برجسته SpeechT5 توانایی آن در حفظ هویت گویندگان است. در واقع این مدل با دریافت یک بردار تعبیه ۱۵ که نوعی بازنمایی از هویت و ویژگی های گوینده است، می تواند گفتاری با صدای گوینده مورد نظر تولید کند. این قابلیت به ویژه در مواردی مانند وظیفه تبدیل صوت، که می خواهیم گفتار ورودی با صدای یک گوینده مشخص باز تولید شود و یا در ترجمه گفتاری، که نیاز است صدای گوینده در خورجی حفظ شود، اهمیت دارد.

Hidden representation\.

Transformer encoder-decoder"

Sequence to sequence¹⁷

Pre-net¹⁷

Post-net^{\f}

Embedding vector¹⁰

تولید گفتار به زبان فارسی

اگرچه همانطور که در مباحث مقدماتی این سوال مشاهده کردیم، SpeechT5 یک مدل از پیش آموزشدیده همه فن حریف است، اما یک محدودیت قابل توجه دارد، مانند مدلهای گفتاری چون MMS قابلیت تولید گفتار (یا سایر وظایف) برای زبانهای غیر انگلیسی را ندارد. لذا در این سوال ما قصد داریم تا این مدل را بر روی زبان فارسی فاین تیون کنیم و عملکرد مدل در تولید گفتار فارسی بیازماییم.

مجموعه داده

اولین چالشی که برای آموزش یک مدل تولید گفتار وجود دارد، یافتن یا جمعآوری یک مجموعه داده غنی میباشد. یعنی دیتاست نه تنها باید حجم مناسبی داشته باشد بلکه باید شامل موضوعات مختلف و همچنین گوینده های متفاوتی باشد تا مدل توانایی تولید گفتار با صداها و هویتهای متفاوت را نیز پیدا کند. از این رو ما برای این تمرین از یک مجموعه داده فارسی از مجموعه Common Voice که توسط موزیلا ارائه شده است استفاده خواهیم نمود که اگرچه برای یک پروژه واقعی احتمالا کافی نخواهد بود، اما برای آنچه در این تمرین از شما انتظار داریم، از حجم و البته تنوع گوینده کافی برخوردار است.

برای دانلود این دیتاست از طریق این لینک به صفحه دانلود رفته، زبان فارسی را انتخاب کرده و سپس نسخه ۱۳ دیتاست را انتخاب نمایید، درنهایت با وارد کردن ایمیل و تایید موارد خواسته شده لینک دانلود برای شما فعال خواهد شد (در صورتی که از کولب استفاده می کنید، لینک دانلود را داخل "" قرار دهید تا هنگام دانلود با خطا مواجه نشوید). در نهایت نیز لازم است از جداول train و test که با فرمت tsv در دیتاست ذخیره شدهاند استفاده نمایید. این جداول شامل فهرستی از جملات به زبان فارسی و نیز آدرس فایلهای صوتی مرتبط با هر جمله میباشند. توجه کنید که لزومی بر استفاده از تمام دیتاست وجود ندارد، حتی با حدود یک سوم از مجموعه دادهای که در فایل train.tsv فهرست شده نیز میتوانید مدل را آموزش دهید و نتیجه بگیرید. البته در این صورت تنظیم هایپرپارامترها به حساسیت و دقت بیشتری نیاز دارد. اما لازم است هر حجمی از دیتاست (مجموع زمان فایلهای صوتی) را که برای آموزش استفاده می کنید در گزارش خود ذکر نمایید. (برای این کار لازم نیست زمان همه فایلها محاسبه شود، می توانید به طور میانگین مدت زمان هر فایل صوتی را ۴ ثانیه در نظر بگیرید)

شبكههاي كمكي

مدل SpeechT5 برای آموزش سه نوع ورودی اصلی دریافت می کند: متن خام (سطح کاراکتر)، گفتار خام (شکل موج) و بردار تعبیه گوینده X-Vector که در بخش معرفی مدل به آن اشاره کردیم. دو مورد اول را از طریق دیتاست در اختیار دارید، اما برای تولید بردار تعبیه گوینده لازم است از یک مدل اضافی استفاده نمایید. جهت مشاهده توضیحات مربوط به این مدل، نحوه استخراج بردارهای تعبیه گوینده و سایر موارد به این لینک مراجعه نمایید. توجه به این نکته مهم است که مدل مذکور بر روی گفتار انگلیسی آموزش داده شده است، اما از آنجایی که چنین مدلی برای زبان فارسی به راحتی قابل دسترس نیست، به استفاده از همین مدل بسنده می کنیم. مدل دومی که برای استفاده از SpeechT5 نیاز است، یک Vocoder میباشد که البته ربطی به فاز آموزش مدل ندارد، اما برای تولید گفتار به آن نیاز خواهید داشت. از آنجایی که مدل به عنوان خروجی نوعی اسپکتروگرام X تولید می کند، برای تبدیل آن به سیگنال صوتی X لازم است تا از مدل X HiFi-GAN استفاده نمایید.

پیادهسازی و آموزش

از آنجایی که برخی از مدلهای از پیش آموزشدیدهای که پیشتر راجع به آنها صحبت کردیم، فقط در transformers وجود دارند، لازم است تا برای فاین تیون کردن SpeechT5 از کتابخانه Huggingface ابزارها و مدلهای ارائه شده توسط آن استفاده نماییم. اما توجه کنید که برای این سوال، در استفاده از این ابزارها یک سری محدودیتها وجود دارد. بطور کلی شما تنها برای بارگذاری و استفاده از مدلهای از پیش آموزش دیده و ابزارهای پیش پردازش دادههای متنی (Tokenizer) و گفتاری (FeatureExtractor) می توانید از کتابخانه transformers استفاده نمایید. لذا استفاده از کلاسهایی مانند Processor و Processor مجاز نمی باشد. همچنین توجه داشته باشید که loss در خود مدل SpeechT5 محاسبه شده و در خروجی بازمی گردد، و لازم نیست تابع خطا را پیاده سازی نمایید.

^{۱۶} تعبیههای گوینده روشی برای نمایش هویت گوینده به صورت فشرده، به عنوان یک بردار با اندازه ثابت، بدون توجه به طول گفتار است. این تعبیهها اطلاعات ضروری را در مورد صدای گوینده، لهجه، لحن و دیگر ویژگیهای منحصربه فرد که یک گوینده را از دیگری متمایز می کند را حفظ میکنند. متداول ترین تکنیک ها برای ایجاد بردار های تعبیه گوینده عبارتند از:

[■] I-Vectors: بر اساس مدل مخلوط گاوسی (GMM) هستند و به روشی بدون نظارت (Unsupervised) تولید میشوند.

[■] X-Vectors: با استفاده از شبکههای عصبی عمیق بدست می آیند و عملکرد بالاتری نسبت به بردار های I-Vectors دارند. در این روش یک شبکه عصبی برای تمایز بین گویندگان آموزش میبیند و گفتارهای با طول متغیر را به بردار های تعبیه ای با ابعاد ثابت تبدیل میکند.

Log-mel-spectrogram'

میدانیم که برای وظیفه تولید گفتار معیار قابل محاسبه و بیطرفی 1 وجود ندارد، لذا لازم است تا ارزیابی نحوه آموزش توسط خودتان صورت گیرد. بدین ترتیب باید بعد از هر ایپاک، با استفاده از متنهای موجود در مجموعه ارزیابی، تعدادی گفتار تولید کنید و عملکرد مدل را بررسی نمایید.

با توجه به محدودیتهای سختافزاری، ما سعی کردیم که مدل گفتاری و مجموعه داده را طوری انتخاب کنیم که در صورت پیادهسازی صحیح و تنظیم مناسب هایپرپارامترها، قادر باشید تا مدل را در محیط کولب آموزش دهید و قبل از به پایان رسیدن محدودیت زمانی استفاده از GPU، تولید گفتارهای قابل قبول را مشاهده نمایید (البته فقط برای برخی جملات ساده تر، آموزش کامل مدل به زمان بیشتر و سختافزار بهتری نیاز دارد). لذا هر زمان که مدل شما قادر به تولید حتی یک گفتار صحیح شد، می توانید آموزش را متوقف کنید. پس از آن، گفتار تولید شده را از نظر میزان تطابق صدای گوینده، تلفظ صحیح، لحن و میزان روان بودن گفتار با نمونه اصلی مقایسه کنید و نتیجه را همراه با نمودار خطا مربوط به مجموعه آموزش و ارزیابی گزارش نمایید. همچنین در نهایت، علاوه بر مواردی که تاکنون خواسته شده، لازم است تا گزارش مختصری راجع به ۱) آموزش مدل و تغییراتی که بهجهت فاین تیون کردن مدل روی زبان جدید اعمال کردید و ۲) پیششبکه و پسشبکههایی که در وظیفه تولید گفتار به کار گرفتن آن در مدل SpeechT5 و نحوه عملکرد آن؛ ارائه دهید.

سیاست نمرهدهی

باتوجه به نوع وظیفه خواسته شده از شما در این سوال، امکان شکستن سوال به زیربخشهای مختلف جهت توزیع نمره وجود ندارد. چراکه تمام بخشها بهم وابسته بوده و نمیتوان هیچیک را جدا از دیگری ارزیابی نمود. از این جهت بارم بندی این سوال به شکل زیر خواهد بود:

- پیادهسازی کد ۲۰ نمره
- آموزش صحیح مدل (کاهش معقول خطا در طی فرایند آموزش) Δ نمره
 - ارائه گفتار تولید شده توسط مدل با کیفیت مناسب ۵ نمره
 - گزارش ۱۰ نمره

نمونه کد بارگذاری مدل ها و ابزار های لازم برای پردازش ورودی

نکات پیاده سازی و تحویل

- مهلت ارسال این تمرین تا پایان روز "دوشنبه ۱۱ دی ماه" خواهد بود.
- این زمان قابل تمدید نیست و درصورت نیاز میتوانید از grace time استفاده کنید.
- پیاده سازی با زبان برنامه نویسی پایتون باید باشد و کدهای شما باید قابل اجرا بوده و به همراه گزارش آپلود شوند.
 - انجام این تمرین به صورت یک نفره میباشد.
- در صورت مشاهده هر گونه تشابه در گزارش کار یا کدهای پیادهسازی، این امر به منزله تقلب برای طرفین در نظر گرفته خواهد شد.
- استفاده از کدهای آماده بدون ذکر منبع و بدون تغییر به منزله تقلب خواهد بود و نمره تمرین شما صفر در نظر گرفته می شود
 - در صورت رعایت نکردن فرمت گزارش کار نمره گزارش به شما تعلق نخواهد گرفت.
 - تحویل تمرین به صورت **دستنویس** قابل پذیرش نیست.
 - تمامی تصاویر و جداول مورد استفاده در گزارش کار باید دارای توضیح (caption) و شماره باشند.
 - بخش زیادی از نمره شما مربوط به گزارش کار و روند حل مسئله است.
 - لطفا گزارش ، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه بارگذاری نمائید.

HW3_[Lastname]_[StudentNumber].zip

به طور مثال:

HW3_Zilouchian_12345678.zip

- در صورت وجود سوال و یا ابهام میتوانید از طریق رایانامه زیر با موضوع DGM_HW3 با دستیاران
 آموزشی در ارتباط باشید:
 - 0 سوال اول

محمد جواد مؤمني : momeni.nezhad@mail.ut.ac.ir

0 سوال دوم

نيكو نقويان : <u>nikoo.naghavian@gmail.com</u>

0 سوال سوم

حسین آزاد ملکی : hosein.azadmaleki@gmail.com

شاد و سلامت باشید.