گزارش فنی

تنظیم دقیق چارچوب Wav2Vec2 در زبان فارسی

Fine-Tuned Wav2vec2 In Persian

حمیدرضا اکبری

خردادماه ۱۴۰۱

۱. هدف

یکی از چارچوب های کارآمد در تشخیص خودکار ‘گفتار[1] چارچوب Wav2Vec می‌باشد. در این چارچوب ابتدا با روش یادگیری خودنظارتی[2] یادگیری اولیه با دادهای بدون برچسب انجام و در گام دوم با کمک یک روش تحت نظارت[3] مدل گام قبل تنظیم دقیق تری خواهد شد. کارایی منا سب این روش حتی با حجم داده های برچ سب دار اندک نکته کلیدی در این چارچوب می با شد. هدف از این گزارش برر سی عملی این چارچوب بر روی یک مجموعه داده و برر سی میزان کارایی این روش در تشخیص گفتار می‌باشد.

۲. مجموعه داده

Common Voice[4] یک مجموعه ی از داده های صوتی در دسترس عموم می باشد. این مجموعه داده توسط افراد داوطلب در سطح جهانی پشتیبانی می‌شود و تولیدکنندگان برنامه های نرم افزار صوتی می‌توانند از این داده ها برای یادگیری مدلهای پیش‌بینی استفاده نمایند. داده های مذکور بیش از ۹۰ زبان مهم جهان را پشتیبانی می‌کند. در زیر به برخی از شاخص های این مجموعه داده در زبان فارسی اشاره شده است.

اندازه داده	**۹گیگابایت**
فرمت فایل صوتی	Mp3
تعداد فایل های صوتی	۴,۰۵۸

۳. مدل یادگیری خودنظارتی

همانطور که اشاره شد در گام اول چارچوب Wav2Vec، یادگیری مدل پایه بر اساس اطلاعات برچسب نشده و با استفاده از روشهای یادگیری خودنظارتی انجام می‌گیرد. در این تحقیق با انتخاب یکی از مدل های خودنظارتی یادگرفته شده بر روی مجموعه داده‌های زبان فارسی، تنظیم دقیق مدل با داده‌های برچسب دار پیگیری و اجرا می‌شود. لذا انتخاب مدل

[1] Automatic Speech Recognition
[2] Self-Supervised Learning
[3] Supervised Learning
[4] https://commonvoice.mozilla.org/en

خودنظارتی یادگیری شده بر روی زبان فارسی یکی از مراحل انجام این تحقیق می‌باشد. در وب سایت Hugging Face[5] لیست انواع مدلهای آموزش داده شده بر روی چارچوب Wav2Vec که توسط کاربران مختلف طراحی و اجرا شده است قابل استفاده می‌باشد. با جستجو می‌توان آنها را شناسایی و برای تنظیم دقیق از آنها استفاده نمود.

٤. پیاده سازی

در این بخش بر اساس مدل پایه انتخاب شده جزئیات پیاده سازی و نتایج حاصل از آن تشریح شده است. کدمنبع[6] به شرح پیوست برای این هدف طراحی و پیاده سازی شده است. این کد با زبان برنامه نویسی پایتون[7] تولید و تحت محیط توسعه نرم افزار پایچارم[8] بر روی سیستم عامل ویندوز پیاده سازی شده است.

در فایل Evaluation میزان خطا پیش‌بینی بر اساس شاخص WER[9] محاسبه و نمایش داده خواهد شد. نحوه یادگیری مدل پایه در آدرس https://colab.research.google.com/github/m3hrdadfi/notebooks/blob/main/Fine_Tune_XLSR_Wav2Vec2_on_Persian_ASR_with_%F0%9F%A4%97_Transformers_ipynb.ipynb قابل دسترسی می‌باشد. ساختار کد به شرح زیر توصیف شده است.

- بارگزاری فایل های صوتی : تابع speech_file_to_array_fn مسئولیت بارگزاری و resample فایل را با اندازه ۱۶۰۰۰ بر عهده دارد. چالشهایی در نحوه بارگزاری فایل های صوتی با فرمت mp3 وجود داشته که برای حل این چالش از توابع کتابخانه ی audio2numpy استفاده شده است.

- پیش پردازش بر روی داده های متنی برچسب دار : بر روی داده های متنی سه پیش پردازش ذیل انجام می‌شود.
  - حذف کاراکترهای خاص : برخی از کاراکترهای خاص نظیر ? از داده های متنی حذف می‌شود. لیست مجموعه این کاراکترها در برنامه مشخص شده است.

[5] https://huggingface.co/
[6] Source Coe
[7] Python
[8] PyCharm
[9] Word Error Rate

- تبدیل برخی از کارکترهای خاص : برخی از حرفهای فارسی نظیر ی و ک به جهت کدینگهای متعدد به یک کد تبدیل می شود. لیست این تبدیلها در برنامه مشخص شده است.

- عملیات نرمالیز کردن متن : با کمک تابع کتابخانه hazm متن نرمالیزه می شـــود. به عنوان مثال فاصله ها در برخی از کلمات به نیم فاصله تبدیل می‌شود.

- تنظیم دقیق : برای انجـام عملیـات تنظیم دقیق از مـدل m3hrdadfi/wav2vec2-large-xlsr-persian استفاده شده است. شاخص ارزیابی استفاده شده برای ارزیابی WER می باشد که بر روی مجموعه داده مورد نظر خطا عدد **٣٤,٨٧٪** می‌باشد که فاصله زیادی با مقاله اصلی[10] دارد.

۵. سایر مدل ها

افراد متعددی مدلهای خودنظارتی مختلفی را طراحی و اجرا نموده اند که از منظر کارایی و بر اسـاس نتایج اعلامی به شـرح جدول ذیل می‌با شد. میزان تمایز برخی از این مدلها ضرورت برر سی بی شتر در نحوه طراحی و پیاده سازی آنها را افزایش خواهد داد.

مدل	کارایی (WER)
jonatasgrosman/wav2vec2-large-xlsr-53-persian	۳۰,۱۲
m3hrdadfi/wav2vec2-large-xlsr-persian	۳۴,۸۷
m3hrdadfi/wav2vec2-large-xlsr-persian-v3	۱۰,۳۶

[10] https://arxiv.org/abs/2006.11477