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

برای قسمت داده های برچسب دار از مجموعه داده ShEMO[5] استفاده شده است. اندازه داده ها در حدود ۳ ساعت و ۲۵ دقیقه است که از نمایشنامه های رادیویی برخط[6] استخراج شده است. این داده ها صحبتهای ۸۷ نفر فارسی زبان می باشد.

[1] Automatic Speech Recognition
[2] Self-Supervised Learning
[3] Supervised Learning
[4] https://commonvoice.mozilla.org/en
[5] https://www.kaggle.com/datasets/mansourehk/shemo-persian-speech-emotion-detection-database
[6] Online

۳. پیاده سازی

تنظیم دقیق[۷] مدل یادگیری در زبان فارسی بر روی مجموعه داده هدف توسط افراد مختلف انجام گرفته است. برای تسریع در اجرا و ارزیابی نتایج در این تحقیق از یکی از پیاده سازی‌ها[۸] استفاده و اجرا شده است. پیاده سازی در دو بخش ذیل مدنظر قرار گرفته است.

۳,۱ آموزش و تنظیم دقیق مدل

برای قسمت تنظیم دقیق مدل از محیط کلب[۹] و زبان برنامه نویسی پایتون استفاده شده است. جزئیات پیاده سازی در فایل مورد نظر[۱۰] وجود دارد که در ادامه مراحل اصلی پیاده‌سازی تشریح می شود.

۳,۱,۱ تفکیک داده های آموزش از تست

داده های برچسب دار آموزش و تست به نسبت ۱ به ۱۰ و به صورت تصادفی تعیین می شوند. تعداد داده های آموزش ۲۵۵۴ و تعداد داده های تست ۲۸۴ می باشد. داده های با فرمت csv ساخته و ذخیره می‌شوند. برای آگاهی از صحت اطلاعات برخی از داده ها به صورت تصادفی نمایش داده می‌شوند.

۳,۱,۲ پیش پردازش داده ها

عملیات پیش پردازش متن به شرح ذیل بر روی داده های آموزش و تست به صورت جداگانه انجام می‌گیرد. تعداد حروف مجموعه داده‌های آموزش و تست جهت بررسی و مقایسه اطلاعات نمایش داده می‌شود.

- **حذف کاراکترهای خاص** : برخی از کاراکترهای خاص نظیر ? از داده های متنی حذف می‌شود. لیست مجموعه این کاراکترها در برنامه مشخص شده است.
- **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 شود. به عنوان مثال فاصله ها در برخی از کلمات به نیم فاصله تبدیل می‌شود.

---

[7] Fine-Tuned
[8] https://huggingface.co/m3hrdadfi/wav2vec2-large-xlsr-persian
[9] https://colab.research.google.com/
[10] Fine_Tune_XLSR_Wav2Vec2_on_Persian_ShEMO_ASR_with_Transformers_ipynb.ipynb

۳,۱,۳ پیش پردازش فایل های صوتی

از آنجاییکه فایل های صـــوتی مجموعه داده با نرخ نمونه گیری ۴۸ کیلوهرتز نمونه گیری شـــده اند در این گام نمونه گیری کلیه فایل های صوتی با نرخ ۱۶ کیلوهرتز انجام می‌شود.

۳,۱,٤ یادگیری

در این مرحله مدل بر ا ساس اطلاعات برچ سب دار آموزش داده می شود. شاخص ارزیابی در طول مدت زمان آموزش میزان خطای کلمات می‌باشد که تابع compute_metrics این محاسبه را انجام می‌دهد.

۳,۲ ارزیابی

هدف از پیاده سازی این بخش تعیین میزان پیش بینی خطا بر اساس مدل آموزش داده شده در مرحله قبل می‌باشد. این بخش از برنامه با زبان برنامه نوی سی پایتون[11] تحت محیط تو سعه نرم افزار پایچارم[12] بر روی سی ستم عامل ویندوز انجام شـده اسـت. میزان خطا بر اسـاس شـاخص WER[13] تعیین می‌گردد. بر این اسـاس خطا بر روی مجموعه داده مورد نظر **٪۳٤,۸۷** می‌باشـد که فاصـله زیادی با مقاله اصـلی[14] دارد. با عنایت به چالش نحوه بارگزاری فایل های mp3 از تابع کتابخانه ی audio2numpy استفاده شده است.

٤. سایر مدل ها

افراد متعددی مدلهای تنظیم شده مختلفی را طراحی و اجرا نموده اند که از منظر کارایی و بر ا ساس نتایج اعلامی به شرح 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

[11] Python
[12] PyCharm
[13] Word Error Rate
[14] https://arxiv.org/abs/2006.11477