دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش کارآموزی

محل کارآموزی: شرکت عصر گویش پرداز

نگارش

امیرمحمد بابائی

شماره دانشجویی: ۹۸۳۱۰۱۱

استاد کارآموزی

دکتر احمد نیک آبادی

تابستان ۱۴۰۱


بسم الله الرحمن الرحيم

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش کارآموزی

محل کارآموزی: شرکت عصر گویش پرداز

نگارش

امیرمحمد بابائی

شماره دانشجویی: ۹۸۳۱۰۱۱

استاد کارآموزی

دکتر احمد نیک آبادی

تابستان ۱۴۰۱

سپاس‌گزاری

اینجانب امیرمحمد بابائی مراتب امتنان و تشکر خود را نسبت به استاد کارآموزی جناب دکتر احمد نیک آبادی که در گذراندن این دوره کارآموزی همواره مرا یاری نموده‌اند، ابراز می‌دارم.

امیرمحمد بابائی
تابستان ۱۴۰۱

چکیده

امروزه، مدل‌های تبدیل گفتار به متن، کاربرد بسیاری در چت‌بات‌ها، تعامل ساده‌تر انسان با رایانه و شبکه‌های پخش آنلاین ویدیو پیدا کرده‌اند. با این حال، مدل‌های فعلی در موقعیت‌های نویزی و دارای کیفیت پایین، عملکرد ضعیفی از خود نشان می‌دهند. یکی از رویکردهای اصلی برای بهبود عملکرد این سیستم‌ها، استفاده از داده‌های کمکی غیر از سیگنال‌های صوتی می‌باشد. از نمونه این نوع داده‌های کمکی، داده‌های تصویری حرکت لب‌های گوینده می‌باشد که توانایی جبران نقص اطلاعات داده‌های صوتی مربوط به گفتار را دارا می‌باشد. رویکرد غالب در این روش، استفاده از روش‌های نظارت‌شده [۱] می باشد اما با این حال، به دلیل محدود بودن داده‌های برچسب گذاری شده صوتی-تصویری، روش مناسبی در حال حاضر نمی‌باشند. رویکرد جدید برای دستیابی به نتایج بهتر، استفاده از مدل‌های خود-نظارتی [۲] می باشد که توانایی رسیدن به عملکرد مناسب با استفاده از حجم داده برچسب‌گذاری شده کمتر را دارا هستند. مدل ای-وی هیوبرت [۳] نمونه ای از مدل‌های مبتنی بر این رویکرد می‌باشد. علاوه بر این، به دلیل نبود دادگان مناسب برای حل این مساله در فارسی، مراحل جمع‌آوری یک دادگان صوتی-تصویری فارسی با استفاده از ویدیوهای آرشیو سایت تلوبیون نیز در این گزارش، ذکر شده است.

واژه‌های کلیدی:

بازشناسی گفتار، بازشناسی گفتار به واسطه صوت و تصویر، دادگان صوتی-تصویری

[1] Supervised

[2] Self-Supervised

[3] AV-HuBERT

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فهرست نمادها

نماد	مفهوم

فصل اول

مقدمه

صوت یکی از مهم‌ترین حالات انرژی در جهان ما می‌باشد و راه ارتباطی اصلی بسیاری از انسان‌ها و دیگر موجودات، از طریق سیگنال های صوتی می‌باشد. به همین دلیل، درک و پردازش این نوع از داده‌ها، اهمیت بسیاری در عصر حاضر برای ما دارا می‌باشد.

علاوه بر این، یکی از بهترین حالات تعامل انسان با رایانه، استفاده از صوت و دستورات گفتاری است. بنابراین، برای دستیابی به چنین قابلیتی، نیاز است که گفتار برای رایانه‌ها قابلیت پردازش و درک پیدا کرده و سپس از آن برای برقراری ارتباط راحت‌تر میان انسان و رایانه استفاده کرد.

برای این کار،امروزه سامانه‌ها و مدل‌هایی وجود دارند که صرفا بر روی قسمت صوتی گفتار متمرکز می‌باشند. این مدل‌ها با اینکه در موقعیت‌های عادی و بدون نویز، به دقت و عملکرد مناسبی دست پیدا کرده‌اند، اما در موقعیت‌های نویزی و با کیفیت پایین، عملکرد نسبتا ضعیفی از خود نشان می‌دهند و به همین دلیل مدل‌های قابل اتکایی نمی‌باشند.

یکی از راهکارها برای قابل‌اتکا کردن این نوع از مدل‌ها، استفاده از داده‌های کمکی می‌باشد. یک نمونه از این داده‌های کمکی، داده های تصویری حرکت لب‌های فرد گوینده می‌باشد. این داده‌ها قابلیت جبران نقص اطلاعات سیگنال‌های صوتی را دارا می‌باشند.

این نوع از عملکرد، معادل عملکرد سیستم شنیداری انسان نیز می‌باشد. در انسان نیز، با اینکه گوش، مهم‌ترین نقش را ایفا می‌کند، اما تنها مولفه نمی‌باشد. این موضوع زمانی واضح‌تر می‌شود که در یک محیط شلوغ، به دنبال درک جملات بیان شده توسط یک گوینده هستیم. در این حالت، حرکت لب‌های فرد در کنار گفتار ضعیفی که از فرد به ما می‌رسد، در کنار هم منجر به درک درست گفتار بیان شده فرد گوینده از سمت ما می‌شود.

علاوه بر این موضوع، برای ساخت و پیاده‌سازی چنین سامانه‌هایی، یکی از مهم‌ترین ارکان، وجود داده‌های آموزشی می‌باشد. این نوع از داده‌ها، در زبان‌هایی نظیر زبان انگلیسی به نسبت، به مقدار بیشتری وجود دارند این در حالی است که در زبان فارسی حجم دادگان‌های موجود به نسبت، کم‌تر می‌باشد. یکی از مواردی که در این گزارش در رابطه با آن صحبت خواهد شد، روش جمع‌آوری و گردآوری یک دادگان صوتی-تصویری برای ارائه و استفاده در حل مساله بازشناسی گفتار به واسطه صوت و تصویر می‌باشد.

در ادامه، در فصل دوم به معرفی شرکت عصرگویش‌پرداز پرداخته و بخشی از مهم‌ترین محصولات و زمینه‌های فعالیت این شرکت بررسی خواهند شد. در فصل سوم، تجربیات کسب شده در این دوره کارآموزی سه‌ماهه، بیان خواهد شد و برخی از چالش‌ها و راه‌حل‌هایی که در این دوره ارائه شدند، بررسی خواهند شد. در نهایت در فصل چهارم، نتیجه‌گیری مربوط به این دوره کارآموزی بیان خواهد شد و پیشنهادهایی در جهت بهبود مدل و دادگان ارائه شده، ذکر خواهد شد.

فصل دوم

معرفی محل کارآموزی

در این قسمت، به طور مختصر، شرکت عصرگویش‌پرداز معرفی شده و در ادامه محصولات اصلی شرکت و همچنین زمینه‌های فعالیت این شرکت ذکر خواهند شد.

۲-۱ معرفی شرکت

عصر گویش پرداز (سهامی خاص) فعال‌ترین شرکت در زمینه هوش مصنوعی و پردازش سیگنال گفتار بوده که فعالیت خود را از ابتدای سال ۱۳۸۲ شروع کرده است. عمده محصولات و خدمات ارائه شده توسط این شرکت برای نخستین بار در کشور و به صورت حرفه‌ای در زمینه‌های پردازش و تشخیص گفتار بوده است. این شرکت با پشتوانه فنی گروهی از متخصصان کشور از دانشگاه صنعتی شریف تأسیس شد که سابقه و تجربه پژوهشی آنها در زمینه‌های مرتبط با پردازش سیگنال به چندین سال قبل از شروع رسمی فعالیت شرکت برمی‌گردد.

۲-۲ محصولات شرکت

عصرگویش پرداز پیشرو در ارائه سیستم های مبتنی بر گفتار برای زبان فارسی، محصولات مختلفی را توسعه داده است که بیشتر آنها برای نخستین بار برای زبان فارسی انجام شده و منحصراً توسط این شرکت تولید می‌شوند. برخی از محصولات این شرکت عبارتند از:

- نویسا: نخستین سامانه تایپ گفتاری فارسی
- نیوشا: نخستین سامانه تلفن گویای هوشمند مبتنی بر گفتار
- آریانا: سامانه متن به گفتار فارسی با صدای طبیعی
- شناسا: تعیین هویت گوینده
- رمزآوا: احراز هویت گوینده
- بینا: تصویر خوان هوشمند
- رومند: چت بات هوشمند
- جویا: سامانه جستجوی عبارات و کلمات در گفتار
- پوشا: سامانه پنهان سازی اطلاعات در تصویر (استگانوگرافی)
- پدیدا: سامانه کشف تصاویر نهان نگاری شده
- پارسیا: اولین نرم‌افزار مترجم گفتار به گفتار فارسی به انگلیسی/ عربی

- نویسیار: اولین نرم‌افزار تایپ هوشمند فارسی
- کارا: نخستین سامانه تشخیص فرمان صوتی برای ویندوز

۲-۳ زمینه‌های فعالیت

این شرکت امروزه دارای گروهی متخصص و منسجم از افرادی با تخصص و تجربه بالا بوده و سابقه طولانی و موفق در زمینه تحقیق و توسعه و کاربردی کردن توانمندی های پژوهشی دارد و علاوه بر ارائه محصولات مختلف در زمینه‌های هوش مصنوعی، پردازش گفتار فارسی و انگلیسی و پردازش تصویر، قادر به انجام پروژه های مختلف و ارائه خدمات در زمینه‌های مختلف نرم‌افزاری می‌باشد. از جمله زمینه‌های فعالیت این شرکت:

- تولید نرم افزارها و سخت افزارهای هوشمند
- هوش مصنوعی و شناسایی الگو
- پردازش سیگنال (گفتار و تصویر)
- تشخیص گفتار و تایپ گفتاری (تبدیل گفتار به متن)
- سنتز گفتار و متن خوان (تبدیل متن به گفتار)
- شناسایی افراد از روی صدا
- پردازش زبان طبیعی
- بهبود کیفیت گفتار
- طراحی دادگان‌های گفتاری و متنی
- طراحی، توسعه و پشتیبانی نرم افزارهای کاربردی مرتبط
- سیستم‌های تلفن گویا (با قابلیت تشخیص گفتار)
- سامانه‌های تلفنی مبتنی بر ویپ (استریسک، الستیکس و ...)
- برنامه نویسی روی ریزکامپیوترها (DSP، تلفن همراه و ...)

با توجه به نوآوری های انجام گرفته در شرکت عصرگویش پرداز، این شرکت علاوه بر انتشار مقاله‌های مختلف در نشریات و کنفرانس‌های علمی ملی و بین‌المللی، دارای افتخارات و تأییدیه‌های متعددی می‌باشد.

فصل سوم

فعالیت‌ها و تجربیات کارآموزی

در این قسمت به تجربیات کسب شده در دوره کارآموزی شرکت عصرگویش‌پرداز پرداخته خواهد شد. در این دوره کارآموزی، در پروژه بازشناسی گفتار به واسطه صوت و تصویر (پرشین ای-وی-اس-آر [1]) فعالیت داشته‌ام. در ادامه، فعالیت‌های انجام شده در این پروژه به تفصیل بیان خواهد شد.

۳-۱ پروژه پرشین ای-وی-اس-آر

در این بخش، در ابتدا به صورت خلاصه مساله و ضرورت حل آن بررسی خواهد شد سپس به بررسی فعالیت‌های انجام شده در جهت حل این مساله و آماده‌سازی یک خدمت [2] برای ارائه آن، پرداخته خواهد شد.

۳-۱-۱ مساله بازشناسی گفتار

مهم‌ترین راه ارتباطی انسان، زبان و یکی از ارکان مهم آن، گفتار می‌باشد. بنابراین یکی از مناسب ترین روش‌ها برای ارتباط و تعامل با رایانه‌ها، گفتار می‌باشد. به همین دلیل این مساله، یکی از مهم‌ترین مسائل عصر حاضر می‌باشد.

رویکرد غالب در جهت حل این مساله، ایجاد سامانه‌ای است که با دریافت گفتار به صورت سیگنال‌های صوتی، آن را درک کند و سپس متن متناظر با گفتار را به عنوان خروجی، برگرداند. این رویکرد، عملکرد مناسبی در موقعیت‌های بدون نویز از خود نشان می‌دهد اما در صورت قرارگیری در محیط‌ها و موقعیت‌های نویزی، دچار افت کیفیت شده و عملکرد ضعیفی از خود نشان می‌دهند [۱] .

برای حل این مساله دو رویکرد عمده وجود دارد:

- تقویت گفتار [3]
- بازشناسی گفتار با استفاده از ترکیب داده‌های صوتی و بصری [4]

در این پروژه، برای افزایش پایداری [5] مدل های بازشناسی گفتار در محیط‌های نویزی، از رویکرد دوم استفاده شده است. در این رویکرد، مدل تلاش می‌کند با استفاده از داده‌های بصری - به خصوص حرکت لب‌های فرد گوینده - ضعف قسمت‌های نویزی سیگنال‌های صوتی را جبران کرده و عملکرد بهتری از خود نشان دهد.

[1] PersianAVSR
[2] Service
[3] Speech Enhancement (SE)
[4] Audio-Visual Speech Recognition (AVSR)
[5] Robustness

۳-۱-۲ بررسی پیشینه

اولین گام در این دوره کارآموزی، مرور سوابق پژوهشی در جهت حل این مساله بوده است. برای یافتن مقالات مربوط به این مساله، با استفاده از سایت پیپرزویدکد [۶] ، گوگل اسکولار [۷] و کانکتدپیپرز [۸] فرایند جستجو مقالات را آغاز کرده و در نهایت مقالات مرتبط را با در نظر گرفتن پارامترهای زمان انتشار، وجود پیاده‌سازی در سایت گیت‌هاب [۹] و وجود مدل‌های آماده، جمع‌آوری و در یک برگه گوگل ذخیره کردم. لیست مقالات جمع‌آوری شده، در پیوست قابل مشاهده می‌باشد.

پس از جمع‌آوری تمام مقالات، برای یافتن مقاله مناسب، به بررسی تمام مقالات پرداختم. در کل، یازده مقاله جمع‌آوری شده، دارای پیاده سازی با استفاده از چارچوب‌های [۱۰] تنسورفلو [۱۱] و پایتورچ [۱۲] بودند. از میان این یازده مقاله، به دلیل جدیدتر بودن، وجود پیاده‌سازی در گیت‌هاب و وجود مدل های آماده، مدل ای-وی هیوبرت [۱۳] و مقاله‌های مربوط به آن ([۱]و [۲]) را انتخاب نمودم.

۳-۱-۳ مدل خود-نظارتی ای-وی هیوبرت

مدل ای-وی هیوبرت، یک مدل خود-نظارتی [۱۴] می‌باشد و آموزش آن شامل دو مرحله پیش‌آموزش بر روی داده‌های بدون برچسب و کوک کردن آن با استفاده از داده‌های برچسب‌گذاری شده می‌باشد. به همین دلیل، این مدل با استفاده با حجم کم‌تری از داده‌های برچسب‌گذاری شده، عملکرد بهتری نسبت به مدل های نظارت‌شده [۱۵] از خود نشان می‌دهد [۱] .

ساختار یادگیری این مدل، از رویکرد اصلی آموزش در مدل زبانی معروف برت [۱۶] الهام گرفته شده است. مدل زبانی برت، یک مدل مبتنی بر ترنسفورمر [۱۷] می‌باشد و برای یادگیری سعی می‌کند قسمتی از جمله ورودی - برای مثال تعدادی از کلمات موجود در جمله - را پوشانده و در ادامه با کمک کلمات مجاور و ساختار جمله، کلمات پوشانده شده را حدس بزند. این روش منجر می‌شود با حجم داده برچسب‌گذاری شده کمتر و داده‌های بدون برچسب، مدل درک مناسبی نسبت به ساختار جملات و جایگاه کلمات در جمله به دست آورد [۳] .

[6] Papers With Code (https://paperswithcode.com)
[7] Google Scholar (https://scholar.google.com)
[8] Connected Papers (https://connectedpapers.com)
[9] Github (https://github.com)
[10] Framework
[11] Tensorflow
[12] PyTorch
[13] Audio-Visual HuBERT (AV-HuBERT)
[14] Self-Supervised
[15] Supervised
[16] Bidirectional Encoder Representations from Transformers (BERT)
[17] Transformer

با الهام از این ایده، مدل هیوبرت [18] برای حل مساله بازشناسی گفتار به واسطه صوت [19] پیشنهاد شده است. یکی از تفاوت های اساسی حوزه صوت و متن، ساختار داده ورودی می‌باشد. در حوزه متن، ورودی‌ها به دلیل گسسته بودن، قابل شکستن به ساختارهای کوچک‌تر با معنی به صورت توکن [20] یا کلمات می‌باشند در صورتی که صوت،دارای ماهیت پیوسته بوده و به همین دلیل به طور مستقیم چنین امکانی در این حوزه وجود ندارد. برای حل این مشکل و گسسته سازی صوت و گفتار، پژوهشگران از آواها و هجاها به عنوان کوچکترین ساختارهای معنی‌دار در این حوزه استفاده کرده و گفتار را به صورت ترکیبی از آنها تعریف کردند [۳] .

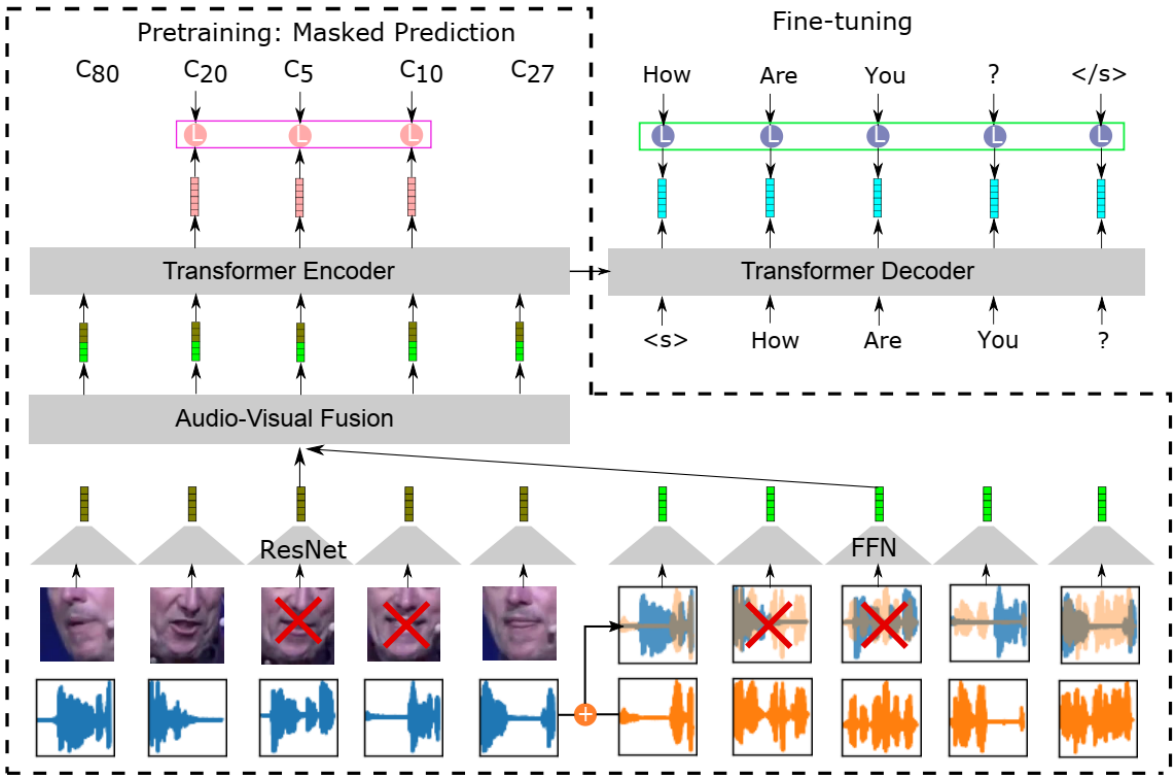
در این مدل، برای استخراج آواها، از استخراج‌کننده ویژگی [21] ام-اف-سی-سی [22] استفاده شده است. این استخراج‌کننده ویژگی با دریافت گفتار، ویژگی‌های با ابعاد ۳۹ را در هر لحظه تولید می‌کند. در نهایت با یک الگوریتم خوشه‌بندی نظیر کا-مینز [23] آواهای اصلی مشخص شده و در فرایند آموزش به عنوان واحدهای سازنده گفتار، شرکت می‌کنند. فرایند استخراج ویژگی، تنها در دور [24] اول به واسطه استخراج‌کننده ویژگی ام-اف-سی-سی انجام شده و در مراحل بعدی به واسطه بازنمایی موجود در لایه‌های میانی شبکه کدکننده [25] ترنسفورمر انجام می‌شود [۳] .

در ادامه برای یادگیری مدل، بخشی از آواها و هجاهای اصلی که در فرایند خوشه‌بندی مشخص شده‌اند، در گفتار ورودی پوشانده شده و مدل تلاش می‌کند تا با توجه به ارتباط میان آواها و یادگیری ساختار آنها، بخش پوشانده را حدس بزند. در این روش، از تابع خطا آنتروپی متقاطع [26] و الگوریتم‌های بهینه‌سازی نظیر الگوریتم آدام [27] استفاده شده است [۳] .

مدل ای-وی هیوبرت، بر پایه مدل هیوبرت ارائه شده است و رویکردی مشابه را این‌بار برای حل مساله بازشناسی گفتار به واسطه صوت و تصویر در پی می‌گیرد. همانطور که در تصویر ۳-۱ مشاهده می‌شود، در این مدل فریم‌های صوتی و بصری ویدیو، به ترتیب به واسطه کدکننده صوتی و کدکننده بصری به یک بازنمایی متراکم [28] تبدیل می‌شوند [۲] .

در کدکننده بصری، از یک مدل رزنت-هجده [29] شده است. مدل های رزنت، به جای ساختار ترتیبی لایه‌ها، دارای اتصالات خارج از ترتیب بوده که موجب کاهش مشکل محوشدن گرادیان [30] و به تبع آن، افزایش تعداد لایه‌های مدل می‌شود. این نوع از مدل‌ها، پیش از ارائه مدل‌های مبتنی بر ویژن‌ترنسفورمر

[18] HuBERT
[19] Automatic Speech Recognition
[20] token
[21] Feature Extractor
[22] Mel-Frequency cepstrum coefficients (MFCC)
[23] K-means
[24] Epoch
[25] Encoder
[26] Cross Entropy
[27] Adam
[28] Dense Representation
[29] ResNet-18
[30] Vanishing Gradient



شکل ۳-۱: معماری مدل ای-وی هیوبرت [۲]

[۳۱] ، دارای بهترین عملکرد در حوزه تصویر بوده‌اند. پیش از ورودی دادن فریم‌های بصری ویدیو به این مدل رزنت-هجده، تغییرات زیر بر روی تصویر اعمال می شود [۱] .

- ۶۸ نقطه کلیدی چهره تشخیص داده شده و سپس به واسطه یک تبدیل خطی، این نقاط کلیدی به یک دستگاه مختصات متمرکز بر چهره انتقال پیدا می‌کند.
- یک منطقه مورد علاقه [۳۲] با ابعاد ۹۶×۹۶ حول دهان فرد در چهره بریده می‌شود.
- کانال‌های رنگی تصویر به سطح خاکستری منتقل می‌شوند.
- در جهت داده‌افزایی [۳۳] یک کادر با ابعاد ۸۸×۸۸ به صورت تصادفی از منطقه مورد علاقه بریده شده و به صورت تصادفی به صورت افقی قرینه [۳۴] می شود.

به دلیل تاثیر بیشتر داده‌های صوتی نسبت به داده‌های بصری در این مساله، برای کاهش تاثیر داده‌های صوتی و افزایش تاثیر داده‌های بصری در یادگیری مدل، از یک شبکه تماما متصل عصبی استفاده شده است. فریم‌های صوتی خام ورودی، پیش از ورودی داده شدن به شبکه عصبی، به واسطه

[31] Vision Transformer (ViT)
[32] Region of Interest (ROI)
[33] Data Augmentation
[34] Horizontal Flip

یک استخراج کننده ویژگی خاص [۳۵] به یک بردار ۲۶ بعدی با فاصله ۱۰ میلی‌ثانیه تبدیل می‌شود. علاوه بر این، به دلیل تفاوت نرخ برداشت فریم های صوتی و بصری - فریم‌های صوتی با فرکانس صد هرتز و فریم‌های بصری با فرکانس ۲۵ هرتز برداشت می‌شود - به ازای هر فریم بصری،چهار فریم صوتی برداشت می‌شود تا هماهنگی زمانی میان دو نوع داده حفظ شود [۱] .

پس از ایجاد یک بازنمایی متراکم از فریم‌های صوتی و بصری ویدیو، این بازنمایی به عنوان ورودی به لایه‌های کدکننده ترنسفورمر داده می‌شود. همانطور که در مدل زبانی برت و مدل بازشناسی گفتار هیوبرت توضیح داده شد، در اینجا نیز شبکه کدکننده به دنبال حدس آواهای پوشانده شده است و تلاش می‌کند با این کار، ماهیت و ارتباط میان آواها را به صورت بدون نظارت درک کند. در این مرحله، با استفاده از داده‌های بدون برچسب، می‌توان پیش‌آموزش [۳۶] مدل را انجام داد و در نهایت برای تکمیل فرایند یادگیری مدل و انتقال دانش کسب شده در فرایند پیش‌آموزش به مساله اصلی، مدل با استفاده از داده‌های برچسب‌خورده، کوک می‌شود. برای این انتقال دانش، از یک شبکه کدبرگردان [۳۷] ترنسفورمری به همراه تابع زیان سی-تی-سی [۳۸] استفاده می‌شود [۱] .

۳-۱-۴ دادگان‌های صوتی-تصویری موجود

پیش از ارائه مدل‌های خود-نظارتی یا نیمه-نظارت‌شده، رویکرد غالب مدل‌ها در جهت حل مساله بازشناسی گفتار به واسطه صوت و تصویر، مبتنی بر مدل‌های نظارت‌شده بوده است. به همین دلیل، عملکرد و دقت این مدل‌ها به حجم دادگان وابسته بوده و در صورت محدود بودن آن، کیفیت و عملکرد پایینی از خود نشان می‌دادند.

با توجه به این توضیحات، بررسی دادگان‌های موجود ضروری به نظر می‌رسد. در ادامه، به صورت مختصر شرحی در رابطه با مشخصات دادگان‌های ال-آر-اس دو و سه [۳۹] و وکس-سلب [۴۰] بیان خواهد شد.

دادگان ال-آر-اس دو و سه

دادگان ال-آر-اس دو، یکی از بزرگ‌ترین دادگان‌های موجود برای مساله لب‌خوانی می‌باشد. این دادگان با استفاده از برنامه‌های تلویزیونی - به ویژه اخبار و برنامه‌های گفتگومحور [۴۱] - شبکه انگلیسی زبان بی-بی-سی [۴۲] تشکیل شده است. پس از ارائه این دادگان، دادگان دیگری با نام ال-آر-اس سه،

[35] Log Filterbank Energy
[36] Pre-Train
[37] Decoder
[38] Connectionist Temporal Classification Loss (CTC Loss)
[39] LRS2-BBC, LRS3-TED
[40] Vox-Celeb
[41] Talk Show
[42] BBC

این‌بار با استفاده از ویدیوهای سخنرانی‌های برنامه‌های تد [۴۳] و تد-ایکس [۴۴] توسط محققان دانشگاه آکسفورد [۴۵] منتشر شد. این دو دادگان، از بزرگ‌ترین دادگان‌های برچسب‌گذاری شده به زبان انگلیسی می‌باشند و معیار ارزیابی [۴۶] در مساله بازشناسی گفتار به واسطه صوت و تصویر محسوب می‌شوند [۱] .

دادگان وکس-سلب

این دادگان، یک دادگان چندزبانه است که در ابتدا برای مساله تشخیص گوینده چندزبانه با استفاده از داده‌های صوتی و بصری ارائه شده است. بر روی هم، این دادگان شامل بیش از دو هزار و ۴۴۲ ساعت گفتار از بیش از شش هزار گوینده که از سایت اشتراک‌گذاری ویدیو یوتیوب [۴۷] استخراج شده است، می‌باشد. همچنین، این دادگان شامل زیرنویس و متن اصلی که در ویدیوها بیان می‌شود، نمی‌باشد [۱] . مزیت این دادگان نسبت به دادگان ال-آر-اس دو و سه، تنوع بیشتر موقعیت‌ها و صحنه‌هایی است که وجود دارد.

۳-۱-۵ ایجاد دادگان فارسی

با توجه به توضیحات داده شده در بخش قبل، دادگان‌های موجود برای حل مساله بازشناسی گفتار به واسطه صوت و تصویر، غالبا به زبان انگلیسی بوده و در زبان فارسی قابل استفاده نمی‌باشند. به همین دلیل، در این دوره تصمیم به ایجاد و جمع‌آوری یک دادگان فارسی گرفتیم.

پیش از شروع فرایند جمع‌آوری ویدیوها، مقالات متناظر با دادگان‌های ال-آر-اس سه و وکس-سلب را مطالعه کردم. با توجه به نکات ذکر شده در این مقالات، می‌بایست به سوالات زیر جواب داده می‌شد:

- منبع جمع‌آوری ویدیوها
- چگونگی استخراج ویدیوها
- چگونگی پردازش ویدیوها
- چگونگی فرایند برچسب‌زنی داده‌های استخراج‌شده

منبع جمع‌آوری ویدیوها

با توجه به این موضوع که داده‌های دادگان‌های مطرح انگلیسی، با استفاده از برنامه‌های تلویزیونی و ویدیوهای اشتراک گذاشته شده در سایت اشتراک‌گذاری ویدیو یوتیوب به دست آمده بودند، گزینه‌های زیر از گزینه‌های مطرح برای جمع‌آوری ویدیوهای فارسی بودند:

[43] TED
[44] TEDx
[45] Oxford University
[46] Benchmark
[47] YouTube

- سایت تلوبیون
- سایت آپارات
- سایت یوتیوب
- شبکه اجتماعی توییتر [۴۸]
- شبکه اجتماعی اینستاگرام [۴۹]
- سایت‌های آموزش ویدیویی نظیر فرادرس و مکتب‌خونه

در نهایت، از میان این گزینه‌ها، سایت تلوبیون به دلیل برخورداری از ویدیوهای برنامه‌های تلویزیونی و نیازمندی به پالایش کمتر داده‌های این سایت، انتخاب شد. بنا به الگوگیری از دادگان ال-آر-اس دو و سه، در این مرحله تصمیم نهایی بر استخراج ویدیوهای آرشیوی شبکه خبر صدا و سیما جمهوری اسلامی ایران شد.

در ادامه برای ارتقای این دادگان، می‌توان از داده‌های ویدیویی دیگر برنامه‌های تلویزیون به خصوص برنامه‌های گفتگومحور دیگر شبکه‌های صدا و سیما جمهوری اسلامی ایران نیز استفاده نمود، اما در این مرحله به دلیل محدود بودن زمان این عمل به نسخه‌های بعدی این دادگان موکول شده است.

چگونگی استخراج ویدیوها

برای استخراج ویدیوهای آرشیو تلویزیون از سایت تلوبیون، یک اسکریپت به زبان پایتون [۵۰] و با استفاده از کتابخانه سلنیوم [۵۱] توسعه داده شده است. زبان پایتون یکی از زبان‌های مفسری [۵۲] مطرح می‌باشد. علاوه بر این، کتابخانه سلنیوم، اجازه استفاده به صورت خودکار از مرورگر را به توسعه‌دهندگان داده و توانایی خودکار سازی فرایندهای مرورگر را با استفاده از زبان‌های برنامه‌نویسی دیگر می‌دهد.

از دیگر گزینه‌های مطرح برای انجام فرایند استخراج ویدیوها، استفاده از کتابخانه ریکوئستز [۵۳] و بیوتیفول سوپ [۵۴] بوده است. با این حال، به دلیل بارگذاری کند [۵۵] سایت تلوبیون، امکان استفاده از این کتابخانه‌ها وجود نداشت و به همین دلیل از کتابخانه سلنیوم برا انجام این کار استفاده شده است.

این اسکریپت با استفاده از یک برنامه راه‌اندازی [۵۶] مربوط به مرورگر، با مرورگر متصل شده و سپس فرایند اتوماسیون و خودکار سازی را شروع می‌کند. این اسکریپت در ابتدا، با توجه به ورودی تعیین شده

[48] Twitter
[49] Instagram
[50] Python
[51] Selenium
[52] Interpreter
[53] requests
[54] BeautifulSoup
[55] Lazy Loading
[56] Driver

به تعداد روزهای تعیین شده، از آرشیو ویدیوهای امروز شروع کرده و به تدریج به سراغ ویدیوهای روزهای قبل رفته و لینک‌های دانلود مربوط به هر ویدیو را استخراج کرده و در یک فایل متنی، ذخیره می‌کند. علاوه بر این، پس از استخراج لینک‌های دانلود ویدیوها، توانایی شروع دانلود ویدیوها را در اختیار دارد.

چگونگی پردازش ویدیوها

ویدیوها پس از دانلود، می‌بایست پردازش شده تا آماده ورودی داده شدن به مدل شوند. در این قسمت با بررسی اسکریپت‌های موجود به صورت عمومی، متوجه شدیم که خط لوله [۵۷] پردازشی مربوط به دادگان وکس-سلب به صورت عمومی در سایت گیت‌هاب منتشر شده است و بدون تغییر قابل استفاده می‌باشد. این خط لوله، مربوط به مدل صوتی-تصویری سینک‌نت [۵۸] می‌باشد که برای هماهنگ کردن حرکت لب‌های فرد گوینده و صوت استفاده می‌شود.

در این مدل، با استفاده از چارچوب پایتورچ و مدل‌های تشخیص چهره آماده با نام اس-سه-اف-دی [۵۹] چهره‌های موجود در تصویر تشخیص داده شده و در طول ویدیو دنبال می‌شوند. و در نهایت، تمامی چهره‌های موجود در ویدیو، به صورت ویدیوهایی از ویدیو اصلی جدا شده و برای استفاده در حل مسائل دیگر نظیر مساله بازشناسی گفتار به واسطه صوت و تصویر قابل استفاده می‌باشد. در تصویر ۳-۲ چهار نمونه از خروجی‌های خط لوله پردازشی مدل سینک‌نت را مشاهده می‌کنید.

شکل ۳-۲: چهار نمونه از خروجی‌های خط لوله پردازشی مدل سینک‌نت

پیش از استفاده از این خط لوله پردازشی، به دنبال توسعه یک خط لوله از ابتدا بوده‌ام. برای این کار،

[57] Pipeline
[58] SyncNet
[59] S3FD

ابتدا از مدل ام-تی-سی-ان-ان [60] که یک مدل تشخیص چهره با قابلیت آشکارسازی در یک صحنه [61] می‌باشد، استفاده کردم. این مدل قابلیت تشخیص چندین چهره در یک تصویر را داشته است. علاوه بر تشخیص کادر چهره فرد، قادر به تشخیص نقاط کلیدی چهره از جمله محل چشمان، محل لب و محل بینی فرد نیز بوده است.

با این حال، به دلیل دقت بالاتر و سادگی پیاده‌سازی خط لوله مربوط به مدل سینک‌نت، استفاده از این مدل آماده را برای پردازش نهایی ویدیوها برگزیدم.

البته در فرایند پردازش ویدیوها، دسته‌ای از خروجی‌های این مدل به اشتباه استخراج شده و نیازمند حذف در مرحله بعدی می‌باشند. حالت کلی این ویدیوهای به اشتباه استخراج شده به شکل زیر می‌باشد:

- ویدیوهایی که فرد صحبت نمی‌کند ولی به دلیل وجود گفتار پس‌زمینه به اشتباه استخراج شده است.
- ویدیوهایی که فرد گوینده، ماسک به صورت داشته است.
- ویدیوهایی که دارای صدای دوبله‌شده می‌باشد. برای نمونه ترجمه مترجم بر روی صحبت‌های یک فرد غیر فارسی زبان.

از آنجایی که این نوع از ویدیوهای خروجی، از احتمال رخداد پایینی برخوردارند، می‌توان یکی از دو رویکرد زیر را در مواجهه با آنها در پیش گرفت:

- تشخیص به صورت دستی و حذف آنها
- کوک کردن یک مدل به واسطه تمامی خروجی‌های مدل و تشخیص ویدیوهای با ضریب اعتماد پایین [62]

با توجه به مشورت‌های انجام شده با منتورهای این دوره، تصمیم نهایی بر این شد که رویکرد دوم اتخاذ شود. به گونه‌ای که در ابتدا تمام خروجی‌ها استخراج شده و پس از کوک کردن مدل نهایی، با یافتن ویدیوهای با ضریب اعتماد پایین، این ویدیوها بررسی شده و در صورتی که یکی از حالات ذکر شده در بالا باشند، حذف شوند.

علاوه بر این، به دلیل حجم بالای ویدیوهای دانلود شده (حدود ۲۳۲ گیگابایت)، فرایند پردازش ویدیو به سادگی میسر نبوده است. با توجه به امکانات شرکت، سرور [63] دارای کارت گرافیک [64] شرکت، دارای حافظه کافی برای انتقال کامل داده‌ها به این سرور و سپس پردازش تمامی آن‌ها وجود نداشت. به همین دلیل برای رفع این مشکل، از ارتباط میان سرورهای شبکه به صورت محلی استفاده کرده و در هر مرحله، یک ویدیو را با استفاده از دستور اس-سی-پی [65] از سرور دارای ویدیوها دانلودشده به سرور

[60] MTCNN
[61] Single Shot Detector (SSD)
[62] Low Confident
[63] Server
[64] GPU
[65] scp

فعلی منتقل کرده و پس از اعمال پردازش مربوطه بر روی این ویدیو، خروجی‌های میانی مربوط به این داده را در سرور پردازشی پاک کرده و خروجی نهایی را به سرور اولیه که دارای حجم ذخیره‌سازی کافی بوده است، ارسال می‌کردم.

از آنجایی که فرایند پردازش ویدیوهای دانلود شده فرایندی زمانبر می‌باشد و همچنین ارتباط شبکه محلی میان سرورهای شرکت، دارای سرعت بالایی - حدود ۱ گیگابیت به ازای هر ثانیه - دارا می‌باشد، این عامل منجر به کندی تاثیرگذاری در سیستم نهایی نشده و قابل چشم‌پوشی می‌باشد.

چگونگی فرایند برچسب‌زنی داده‌های استخراج‌شده

داده‌های آموزشی مرتبط با مساله بازشناسی گفتار به واسطه صوت و تصویر، علاوه بر فریم‌های صوتی و بصری، نیازمند متن بیان شده در فریم‌های ویدیو نیز می‌باشند. به همین دلیل، نیاز است که پس از دانلود ویدیوها و پردازش آنها، متن مرتبط با هر یک از این ویدیوها استخراج شده و به عنوان برچسب این ویدیو مشخص شود. برای انجام فرایند برچسب‌زنی داده‌ها در دسترس، با توجه به مشاوره‌های انجام شده با منتورهای دوره کارآموزی، راهکارهای زیر مطرح شده است:

- جمع‌سپاری [۶۶] داده‌های ویدیویی
- برچسب‌گذاری به صورت دستی توسط عوامل دوره کارآموزی
- استفاده از یک مدل آماده تبدیل گفتار به متن [۶۷] و بررسی خروجی آن به ازای هر یک از ویدیوها

با توجه به زمان محدود دوره کارآموزی، گزینه اول و دوم به دلیل زمانبر بودن، کنار گذاشته شده و تصمیم نهایی بر این شد که با استفاده از یک مدل تبدیل گفتار به متن و سپس بررسی خروجی این مدل به ازای هر ویدیو، این فرایند را سرعت بخشیده و دادگان را سریع‌تر آماده کرد.

در حال حاضر، حدود دو هزار و سیصد و نود ویدیو دانلود شده است و این تعداد ویدیو، در حال پردازش توسط خط لوله پردازشی سینک‌نت می‌باشد. به دلیل زمانبر بودن فرایند پردازشی، فعالیت فعلی در این پروژه تا به اینجا محدود شده است. پس از آماده سازی دادگان، امکان کوک کردن مدل ای-وی هیوبرت و دیگر مدل‌های آماده نیز وجود دارد و می‌توان میزان مفید بودن این دادگان فارسی جمع‌آوری شده را به طور بهتری ارزیابی کرد.

علاوه بر این، این دادگان قابل استفاده در مسائل دیگری نظیر تشخیص لب‌خوانی به زبان فارسی نیز می‌باشد و توانایی استفاده برای کوک کردن مدل‌های آماده موجود را نیز دارا می‌باشد.

[66] Crowd Sourcing

[67] Automatic Speech Recognition

فصل چهارم

نتیجه‌گیری و پیشنهادها

در این فصل، در ابتدا به مرور نکات ذکر شده و جمع‌بندی آنها پرداخته و سپس پیشنهادهایی در جهت بهبود و ارتقا سامانه و دادگان جمع‌آوری شده ارائه می‌شود.

۴-۱ نتیجه‌گیری و جمع‌بندی

همانطور که در فصل‌های قبل بررسی شد، استفاده از داده‌های تصویری حرکت لب‌های فرد گوینده، داده مناسبی برای جبران نقص در سیگنال‌های صوتی مربوط به گفتار می‌باشد. علاوه بر این، این مکانیزم در سیستم شنیداری انسان نیز وجود دارد و علاوه بر گوش‌ها، دیدن حرکت لب‌های گوینده نیز تاثیر به سزایی در فهم متن بیان شده توسط گوینده دارد.

همچنین، به دلیل محدود بودن داده‌های برچسب‌گذاری شده ویدیویی، استفاده از رویکردهای خود-نظارتی و نیمه-نظارت‌شده نسبت به رویکرد نظارت‌شده، عملکرد بهتری از خود نشان داده و از پایداری بهتری برخوردار خواهد بود. در این مدل‌ها، تلاش می‌شود که دانش کلی نسبت به ماهیت و ارتباط داده‌ها به دست آمده (در فرایند پیش‌آموزش بر روی داده‌های بدون برچسب) و سپس این دانش به طور خاص بر روی حل مساله مورد نظر کوک شود.

این رویکرد، رویکرد مناسبی برای استفاده در زبان‌هایی است که داده ویدیو کافی نداشته باشند؛ چراکه با وجود داده برچسب‌گذاری شده کم نیز، توانایی رسیدن به عملکرد و دقت مناسب را دارا می‌باشند. با این حال، در زبان فارسی داده ویدیویی مناسب برای این مساله موجود نمی‌باشد. به همین دلیل در این مقاله در پی این بر آمدیم تا دادگان ویدیویی فارسی با استفاده از ویدیوهای آرشیو شبکه خبر صدا و سیما جمهوری اسلامی ایران جمع‌آوری کنیم.

۴-۲ پیشنهادها

از جمله پیشنهادهایی که می‌توان در جهت بهبود دادگان فعلی داد، استفاده از ویدیو برنامه‌های گفتگومحور دیگر شبکه‌های صدا و سیما جمهوری اسلامی ایران می‌باشد. علاوه بر این در صورت توسعه یک خط لوله پردازشی دقیق‌تر برای حذف خروجی‌های اشتباه مدل سینک‌نت، می‌توان به دادگان با کیفیت بالاتری دست یافت. این دادگان به دلیل حجم تخمینی کمی که دارد، برای فرایند کوک کردن مدل‌های بازشناسی گفتار به کمک صوت و تصویر، مناسب می‌باشد اما برای اجرای فرایند پیش‌آموزش مناسب نمی‌باشد. یکی از دیگر پیشنهادها در جهت بهبود این دادگان، استفاده از دیگر سایت‌های اشتراک‌گذاری ویدیو آنلاین نظیر آپارات، فرادرس و مکتب‌خونه می‌باشد.

منابع و مراجع

[1] Shi, Bowen, Hsu, Wei-Ning, Lakhotia, Kushal, and Mohamed, Abdelrahman. Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184, 2022.

[2] Shi, Bowen, Hsu, Wei-Ning, and Mohamed, Abdelrahman. Robust self-supervised audio-visual speech recognition. arXiv preprint arXiv:2201.01763, 2022.

[3] Hsu, Wei-Ning, Bolte, Benjamin, Tsai, Yao-Hung Hubert, Lakhotia, Kushal, Salakhutdinov, Ruslan, and Mohamed, Abdelrahman. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. CoRR, abs/2106.07447, 2021.

پیوست

لیست مقالات مرتبط با مساله بازشناسی گفتار به واسطه صوت و تصویر
لینک دسترسی به برگه گوگل مقالات جمع‌آوری شده

واژه‌نامه‌ی انگلیسی به فارسی

A

بازشناسی گفتار به واسطه صوت و تصویر Audio-Visual Speech Recognition

بازشناسی خودکار گفتار Automatic Speech Recognition

B

معیار ارزیابی . . . . . Benchmark

C

آنتروپی متقاطع . . . . . Cross Entropy

جمع‌سپاری . . . . . Crowd Sourcing

ضریب اعتماد . . . . . Confidence

D

داده‌افزایی . . . . . Data Augmentation

کدبرگردان، رمزگشا . . . . . Decoder

بازنمایی متراکم، بازنمایی چگال . . . . Dense Representation

برنامه راه‌اندازی . . . . . Driver

E

دور، مرحله . . . . . Epoch

کدکننده، رمزکننده . . . . . Encoder

F

کوک کردن . . . . . Fine-Tune

استخراج‌کننده ویژگی . . Feature Extractor

چارچوب . . . . . Framework

G

کارت گرافیک . . . . . GPU

H

قرینه به صورت افقی . . . . Horizontal Flip

I

مفسر . . . . . Interpreter

L

بارگذاری کند . . . . . Lazy Loading

P

خط لوله . . . . . Pipeline

پیش‌آموزش . . . . . Pre-Train

R

نظارت‌شده . . . . . Supervised

تقویت گفتار . . . . . Speech Enhancement

T

برنامه گفتگومحور . . . . . Talk Show

توکن . . . . . Token

V

محوشدن گرادیان . . . Vanishing Gradient

پایداری . . . . . Robustness

منطقه مورد علاقه . . . . Region of Interest

S

خود-نظارتی . . . . . Self-Supervised

نیمه-نظارت‌شده . . . . . Semi-Supervised

خدمت . . . . . Service

آشکارسازی در یک صحنه . . . . Single Shot Detector