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

تقدیم به پدر و مادر مهربانم که در تاریکی‌های زندگی، چراغ راهم بوده‌اند.

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فهرست جداول

جدول صفحه

فصل اول

مقدمه

فصل دوم

معرفی محل کارآموزی

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.

۲-۱ 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 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

۲-۲ 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 سه پروژه بازشناسی گفتار به واسطه صوت و تصویر (پرشین ای-وی-اس-آر [۱])، اصلاح گرامری و نحوی متن (کاتب [۲]) و اصلاح نشانه‌گذاری متن [۳] خروجی مدل بازشناسی گفتار (کاتب ای-اس-آر [۴]) فعالیت داشته‌ام. در ادامه، فعالیت‌های انجام شده در هر یک از پروژه‌ها 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 [۵] 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.

برای حل این مساله دو رویکرد عمده وجود دارد:

- تقویت گفتار [۶]
- بازشناسی گفتار با استفاده از ترکیب داده‌های صوتی و بصری [۷]

در این پروژه، برای افزایش پایداری [۸] 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

[1] PersianAVSR
[2] Kateb
[3] Punctuation
[4] Kateb-ASR
[5] Service
[6] Speech Enhancement (SE)
[7] Audio-Visual Speech Recognition (AVSR)
[8] Robustness

لب‌های فرد گوینده - ضعف قسمت‌های نویزی سیگنال‌های صوتی را جبران کرده و عملکرد بهتری از خود نشان دهد.

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 [۹] ، گوگل اسکولار [۱۰] و کانکتدپیپرز [۱۱] فرایند جستجو مقالات را آغاز کرده و در نهایت مقالات مرتبط را با در نظر گرفتن پارامترهای زمان انتشار، وجود پیاده‌سازی در سایت گیت‌هاب [۱۲] 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 [۱۳] تنسورفلو [۱۴] و پایتورچ [۱۵] بودند. از میان این یازده مقاله، به دلیل جدیدتر بودن، وجود پیاده‌سازی در گیت‌هاب و وجود مدل های آماده، مدل ای-وی هیوبرت [۱۶] و مقاله مربوط به آن [۱۷] را انتخاب نمودم.

۳-۱-۳ مدل خود-نظارتی ای-وی هیوبرت

مدل ای-وی هیوبرت، یک مدل خود-نظارتی [۱۸] 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 [۱۹] از خود نشان می‌دهد.

ساختار یادگیری این مدل، از رویکرد اصلی آموزش در مدل زبانی معروف برت [۲۰] الهام گرفته شده است. مدل زبانی برت، یک مدل مبتنی بر ترنسفورمر [۲۱] 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

[9] Papers With Code (https://paperswithcode.com)
[10] Google Scholar (https://scholar.google.com)
[11] Connected Papers (https://connectedpapers.com)
[12] Github (https://github.com)
[13] Framework
[14] Tensorflow
[15] PyTorch
[16] Audio-Visual HuBERT (AV-HuBERT)
[17] Robust Self-Supervised Audio-Visual Speech Recognition (https://arxiv.org/abs/2201.01763)
[18] Self-Supervised
[19] Supervised
[20] Bidirectional Encoder Representations from Transformers (BERT)
[21] Transformer

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.

با الهام از این ایده، مدل هیوبرت [۲۲] برای حل مساله بازشناسی گفتار به واسطه صوت [۲۳] 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 [۲۴] 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.

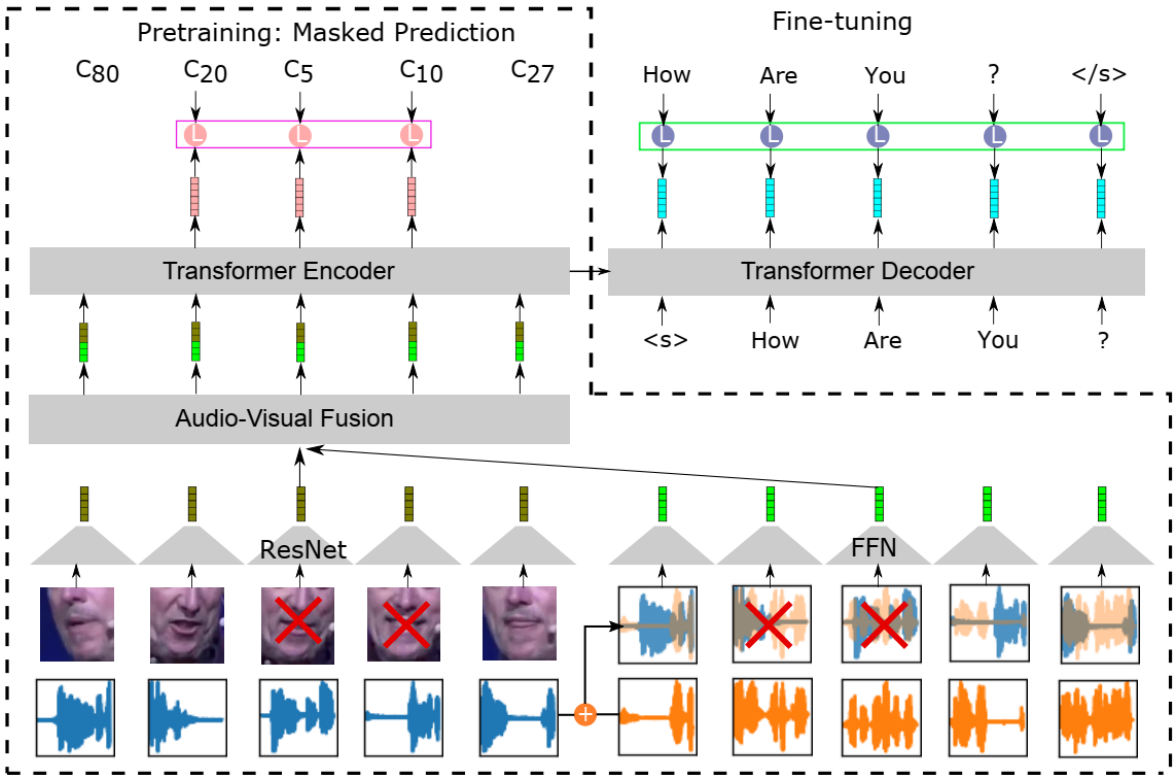
در این مدل، برای استخراج آواها، از استخراج‌کننده ویژگی [۲۵] ام-اف-سی-سی [۲۶] استفاده شده است. این استخراج‌کننده ویژگی با دریافت گفتار، ویژگی‌های با ابعاد ۳۹ را در هر لحظه تولید می‌کند. در نهایت با یک الگوریتم خوشه‌بندی نظیر کا-مینز [۲۷] آواهای اصلی مشخص شده و در فرایند آموزش به عنوان واحدهای سازنده گفتار، شرکت می‌کنند. فرایند استخراج ویژگی، تنها در دور [۲۸] اول به واسطه استخراج‌کننده ویژگی ام-اف-سی-سی انجام شده و در مراحل بعدی به واسطه بازنمایی موجود در لایه‌های میانی شبکه کدکننده [۲۹] ترنسفورمر انجام می‌شود.

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 [۳۰] و الگوریتم‌های بهینه‌سازی نظیر الگوریتم آدام [۳۱] استفاده شده است.

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 [۳۲] تبدیل می‌شوند.

---

[22] HuBERT
[23] Automatic Speech Recognition
[24] token
[25] Feature Extractor
[26] Mel-Frequency cepstrum coefficients (MFCC)
[27] K-means
[28] Epoch
[29] Encoder
[30] Cross Entropy
[31] Adam
[32] Dense Representation



شکل ۳-۱: معماری مدل ای-وی هیوبرت [۱]

در کدکننده بصری، از یک مدل رزنت-هجده [۳۳] شده است. مدل های رزنت، به جای ساختار ترتیبی لایه‌ها، دارای اتصالات خارج از ترتیب بوده که موجب کاهش مشکل محوشدن گرادیان [۳۴] و به تبع آن، افزایش تعداد لایه‌های مدل می‌شود. این نوع از مدل‌ها، پیش از ارائه مدل‌های مبتنی بر ویژن‌ترنسفورمر [۳۵] ، 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.

- 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 [۳۶] با ابعاد ۹۶×۹۶ حول دهان فرد در چهره بریده می‌شود.
- کانال‌های رنگی تصویر به سطح خاکستری منتقل می‌شوند.
- در جهت داده‌افزایی [۳۷] یک کادر با ابعاد ۸۸×۸۸ به صورت تصادفی از منطقه مورد علاقه بریده

[33] ResNet-18
[34] Vanishing Gradient
[35] Vision Transformer (ViT)
[36] Region of Interest (ROI)
[37] Data Augmentation

شده و به صورت تصادفی به صورت افقی قرینه [۳۸] 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 یک استخراج کننده ویژگی خاص [۳۹] 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.

۳-۲ پروژه کاتب

۳-۳ پروژه کاتب ای-اس-آر

[38] Horizontal Flip

[39] Log Filterbank Energy

منابع و مراجع

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