دومین کنفرانس ملی آخرین دستاوردهای مهندسی داده و دانش و محاسبات نرم

کلاس‌بندی مکالمات تلفنی شرکت خدماتی آب و فاضلاب با استفاده از شبکه‌های یادگیری عمیق

الهه بابایی[۱]، عبدالرضا رسولی کناری[۲]، محبوبه شمسی[۳]، سید ابراهیم هزارخانی[۴]

[۱] گروه کامپیوتر، دانشکده برق و کامپیوتر، دانشگاه صنعتی قم ، قم،
babaee.e@qut.ac.ir

[۲] گروه کامپیوتر، دانشکده برق و کامپیوتر، دانشگاه صنعتی قم ، قم،
rasouli@qut.ac.ir

[۳] گروه کامپیوتر، دانشکده برق و کامپیوتر، دانشگاه صنعتی قم ، قم،
Shamsi@qut.ac.ir

[۴] گروه کامپیوتر، دانشکده برق و کامپیوتر، دانشگاه صنعتی قم ، قم،
hezarkhani.se@qut.ac.ir

چکیده

به‌طور حتم یکی از مهم‌ترین موضوعات مطرح و پیچیده در علوم کامپیوتر، تشخیص گفتار است. مهم‌ترین و قوی‌ترین ابزار تشخیص گفتار، استفاده از هوش مصنوعی و الگوریتم‌های تشخیص گفتار است. هنگامی‌که یک صدا به کامپیوتر داده می‌شود، این صدا برای آن قابل‌فهم نیست بلکه باید از الگوریتم‌های تشخیص گفتار استفاده شود تا کامپیوتر به یک فهم خوب از صدای دریافتی برسد. در این پژوهش سعی شده تا با استفاده از الگوریتم‌های کلاس‌بندی، روشی مناسب برای کلاس‌بندی صوت‌های ضبط شده بخش خدمات شرکت آب و فاضلاب به‌عنوان ورودی توسط هوش مصنوعی کامپیوتر ارائه شود. برای این منظور ما جدیدترین شبکه عصبی به نام Yamnet را تغییر داده‌ایم تا گفتار فارسی را به‌عنوان ورودی به‌صورت خام دریافت کند و عملیات کلاس‌بندی گفتار را انجام دهد. ابتدا نمونه‌های آموزشی به‌صورت گفتار تلفنی زبان فارسی تهیه‌شده و به‌صورت خام جهت آموزش داده‌ها به شبکه عصبی داده‌شده‌اند. این مدل بر روی کلاس‌بندی داده‌های آزمایشی به دقتی به‌اندازه ۹۳ درصد رسید. درنهایت، صداهای خام با تبدیل فوریه کوتاه‌مدت و ضریب کپسترال فرکانس مل پردازش شدند و سپس از این ویژگی‌ها به‌عنوان ورودی مدل استفاده شد. مدل با ورودی صوت خام در مقایسه با ورودی پیش‌پردازش صدا با تبدیل فوریه کوتاه‌مدت و ضریب کپسترال فرکانس مل عملکرد بهتر و یا معادل نشان داد.

کلمات کلیدی

پردازش صوتی، یادگیری عمیق، یامنت، تبدیل فوریه کوتاه‌مدت، ضریب کپسترال فرکانس مل

۱- مقدمه

استخراج ویژگی‌های مناسب از صوت و طراحی یک مدل کلاس‌بندی مناسب برای این ویژگی‌ها، به‌عنوان مسئله مهم در تشخیص گفتار خودکار توسط کامپیوتر هستند [۱] و [۲]. اغلب از تبدیل فوریه کوتاه‌مدت یا ضریب کپسترال فرکانس مل استفاده می شود تا ابتدا صوت خام را پردازش کنند و ویژگی‌های مناسب را از آن استخراج کنند و سپس این ویژگی‌ها توسط یک مدل کلاس‌بندی می‌شوند [۱] و [۲]. از اشکالات این روش این است که ویژگی‌های طراحی‌شده ممکن است برای هدف کلاس‌بندی مناسب نباشند.

یادگیری عمیق[۱] شاخه‌ای از هوش مصنوعی است. ایده مدل‌های یادگیری عمیق این است که شبکه‌های عصبی بزرگ را با مقادیر فزاینده داده تغذیه کنیم، این شبکه‌ها به‌طور خودکار ویژگی‌ها را از داده‌های ورودی استخراج می‌کنند و مدل آموزش دیده شده را می‌توان برای پیش‌بینی بر روی داده‌های دیده نشده در زمان آموزش استفاده کرد [۳]. در کلاس‌بندی تصاویر این شبکه‌های عمیق به‌طور موفقیت‌آمیزی برای استخراج ویژگی از عکس‌ها مورداستفاده قرارگرفته‌اند [۳]. یکی از مسائلی که در کلاس‌بندی صوت‌ها وجود دارد این است که آیا می‌توان از این شبکه‌ها به‌عنوان استخراج ویژگی از صوت‌های خام استفاده کرد. در [۱] و [۲]. نشان داده شده است که لایه‌های پایین‌تر شبکه عصبی عمیق می‌توانند به‌عنوان استخراج ویژگی‌های مناسب از صوت‌های خام انگلیسی قرار بگیرند و لایه‌های بالایی را می‌توان به‌عنوان کلاس‌بندی استفاده کرد. در [۲] نشان داده شده است که یک شبکه عصبی عمیق کانولوشن می‌تواند در استخراج ویژگی و کلاس‌بندی واج‌های انگلیسی بهتر از استفاده از ضریب کپسترال فرکانس مل برای استخراج ویژگی‌ها و سپس کلاس‌بندی این ویژگی‌ها با یک شبکه عمیق عصبی عمل می‌کند. در [۴] آخرین شبکه عصبی با نام یامنت را اصلاح کرده‌ایم تا بتوانیم مکالمات فارسی مربوط به واحد خدمات شرکت آب و فاضلاب را کلاس‌بندی کنیم. مدل یامنت برای کلاس‌بندی ۵۱۲ کلاس از صداهایی مانند صدای حیوانات، طبیعت و ... استفاده شده است. به‌طور پیش‌فرض در مدل یامنت ابتدا با استفاده از تبدیل فوریه کوتاه‌مدت ویژگی‌های اصوات استخراج می‌شود و سپس این ویژگی‌ها توسط شبکه عصبی موبایل نت ورژن یک کلاس‌بندی می‌شوند. ما ورودی مدل را تغییر دادیم تا مدل بتواند فایل‌های صوتی خام فارسی را کلاس‌بندی کند. در این مقاله به دلایل زیر از شبکه یامنت استفاده شده است:

- این شبکه جدیدترین شبکه کلاس‌بندی صوت است.
- به دلیل استفاده از شبکه موبایل نت زمان آموزش این مدل کوتاه‌تر از شبکه‌های معمولی است.
- محدودیت در قدرت محاسباتی کامپیوترهای در دسترس. آموزش مدل‌های یادگیری عمیق به دستگاهی با قدرت محاسباتی (GPU) و حافظه بسیار بالا نیاز دارد. هر چه داده‌های ورودی بیشتر باشد، زمان آموزش بیشتری مورد نیاز است.
- ساختن یک مدل یادگیری عمیق از ابتدا به مقدار زیادی داده و منابع محاسباتی نیاز دارد. یکی از راه‌های حل این مشکل استفاده از آموزش انتقالی[۲] است. با استفاده از یامنت می‌توانیم وزن اولیه را با استفاده از مدل آموزش‌دیده در مجموعه داده Audioset مقداردهی کنیم و سپس مدل را آموزش دهیم.

در مقاله [۴] همچنین ورودهایی شامل ورودی پیش‌فرض یامنت یعنی تبدیل فوریه کوتاه‌مدت و ضریب کپسترال فرکانس مل به مدل داده شده است و دقت مدل با ورودی‌های مختلف با هم مقایسه شده‌اند.

۲- کارهای مرتبط

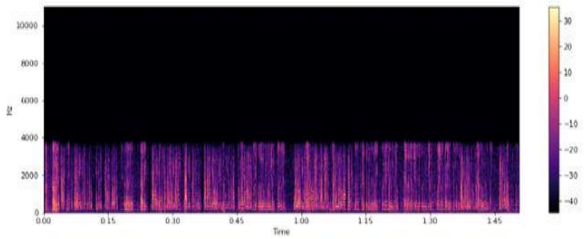
در پژوهش‌های انجام شده، تلاش‌هایی برای مدل سازی سیگنال گفتار خام با پیش‌پردازش کم یا بدون پیش‌پردازش انجام شده است. در مقاله [۵] یک رویکرد جدید برای مدل‌سازی امواج صوتی گفتار با استفاده از یک ماشین بولتزمن محدود[۳] ارائه کرده‌اند. نتایج اولیه نشان داده که عملکرد تشخیص واج با این روش بهتر از روش‌هایی بر اساس ضرایب کپسترال فرکانس مل است. در مقاله [۶] یک شبکه‌های عصبی کانولوشن را برای کلاس‌بندی واج‌ها پیشنهاد داده‌اند. ورودی شبکه‌های عصبی کانولوشن گفتار خام است. معماری شبکه عصبی از دو مرحله تشکیل شده است: یک مرحله از یک لایه کانولوشن و به دنبال آن یک مرحله کلاس‌بندی کننده پرسپترون چندلایه. مطالعات تشخیص واج بر روی مجموعه داده TIMIT نشان داد که رویکرد پیشنهادی قادر به دستیابی به عملکردی قابل مقایسه یا بهتر از رویکرد استاندارد استخراج ویژگی‌های کپسترال به دنبال کلاس‌بندی با شبکه‌های عصبی مصنوعی[۴] است. در مقاله [۷] به بررسی این موضوع که چگونه می‌توان مدل‌های صوتی مبتنی بر شبکه عصبی عمیق برای تشخیص خودکار گفتار را بدون استخراج ویژگی‌ها آموزش داد. در این پژوهش مشخص شد که افزودن لایه‌های کانولوشن در ورودی به بهبود عملکرد سیستم کمک می‌کند. در مقاله [۸]، یک معماری ترکیبی به نام CLDNN مورد بررسی قرار گرفت که در آن سیگنال گفتار خام به‌عنوان ورودی به سیستم عصبی کانولوشن تغذیه می‌شود، خروجی سیستم عصبی کانولوشن متعاقباً توسط یک مدل حافظه طولانی کوتاه-مدت دو جهته (BLSTM) پردازش شده و به یک شبکه عمیق تغذیه می شود. مشخص شد که این رویکرد عملکردی قابل مقایسه با حالتی است که ورودی مدل CLDNN ویژگی‌های ضرایب کپسترال مل به سیستم است. در [۲] یک رویکرد مدل‌سازی صوتی سرتاسر را با استفاده از شبکه‌های عصبی کانولوشن بررسی کرده است که در آن شبکه‌های عصبی کانولوشن سیگنال گفتار خام ورودی را دریافت می‌کند و احتمالات شرطی کلاس مدل پنهان مارکوف را در خروجی تخمین می‌زند. نتایج این تحقیق نشان می‌دهد استخراج ویژگی‌های از لایه‌های کانولوشن در مقایسه با ویژگی‌های استخراج شده از کپسترال استاندارد، بهتر هستند. کارهای مرتبط ذکر شده بر روی واج‌های انگلیسی انجام شده است و هنوز تا یافتن بهترین مدل و بهترین ویژگی‌هایی که می‌توان برای استخراج ویژگی‌ها از صوت‌های فارسی و کلاس‌بندی آن‌ها استفاده کرد، فاصله زیادی هست. در این مقاله به بررسی توانایی مدل یامنت در کلاس‌بندی فایل‌های صوتی فارسی با استفاده از داده‌های خام به‌عنوان ورودی مدل می‌پردازیم. به‌جای تعیین تصادفی وزن‌های مدل در ابتدای آموزش، از وزن یامنت آموزش دیده روی مجموعه داده Audioset استفاده شد. با این روش آموزش مدل سریع‌تر بوده و نیازی به داده‌های زیادی برای آموزش ندارد.

۳- پیش‌زمینه

در این بخش مرور مختصری بر روی الگوریتم‌های موردنیاز روش پیشنهادی مقاله خود خواهیم داشت.

۳-۱- تبدیل فوریه زمان کوتاه

تبدیل فوریه کوتاه‌مدت[۹]، یک تبدیل مرتبط با فوریه است که از فواصل زمانی متفاوت سیگنال، از سیگنال تبدیل فوریه می‌گیرد. در این تبدیل ابتدا سیگنال به بازه‌های زمانی کوتاه با مقداری ا شتراک تق سیم می شود و تبدیل فوریه هر بازه گرفته می‌شود. خروجی این تبدیل یک ماتریس با مقادیر زمان و فرکانس می‌با شد که می‌توان شدت هر فرکانس را در بازه‌های زمانی مختلف به دست آورد.

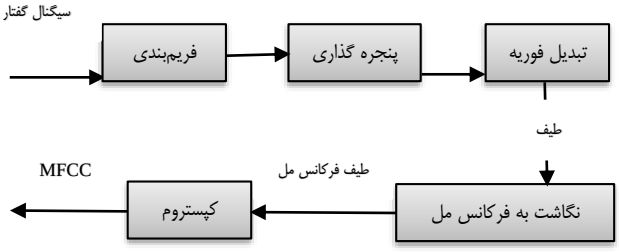


شکل (۱): نمونه ای‌ی از نمودار STFT

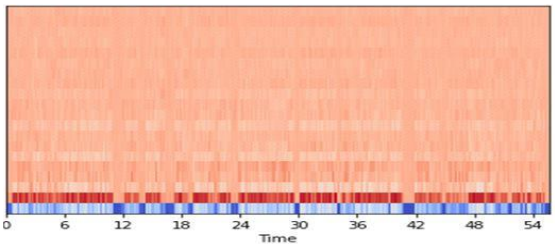
۳-۲- ضریب کپسترال فرکانس مل

ضریب کپسترال فرکانس مل MFCC [۱۰]، تبدیلی است از ترکیب دو تبدیل فوریه و کسینوس که نتیجه آن سیگنال را از حوزه زمانی به حوزه زمان-فرکانس تبدیل می‌کند. مراحل انجام عملیات مرتبط به شرح زیر است:

- ابتدا تبدیل فوریه روی داده‌ها اعمال می‌شود.
- سپس توان‌های نمودار طیف را با استفاده از پنجره‌های مثلثی روی هم قرارگرفته‌اند به مقیاس Mel تبدیل می‌شود.
- لگاریتم هر توان در مقیاس فرکانس مل گرفته می‌شود.
- تبدیل کسینوس به نتیجه مورد ۳ اعمال می‌شود.
- نتیجه استفاده از این تبدیل، ماتریس و درنتیجه یک نمودار سه‌بعدی می‌باشد که می‌توان از آن برای استخراج ویژگی‌های صوتی استفاده شود.



شکل (۲): تبدیل سی‌گنال گفتار به ضری‌ب کپسترال فرکانس مل



شکل (۳): نمونه‌ای از نمودار MFCC

۳-۳- دادهای خام به‌عنوان وردی شبکه عصبی

در این روش صوت‌ها با نرخ نمونه‌برداری ۱۶ کیلوهرتز نمونه‌برداری می شوند. هر صوت به پنجره‌های ۰,۴۸ ثانیه‌ای بافا صله ۰,۲۴ ثانیه‌ای بین آن‌ها تق سیم شد؛ مانند روش تبدیل فوریه زمان کوتاه هرکدام از این پنجره‌ها به پنجره‌های ۰,۰۲۵ ثانیه‌ای بافاصله ۰,۰۱ ثانیه تقسیم شد. ورودی مدل یک دسته از فریم‌های ۰,۴۸ ثانیه‌ای که هر کدام به پنجره های ۰,۰۲۵ ثانیه‌ای تقسیم شده‌اند، می‌باشد.

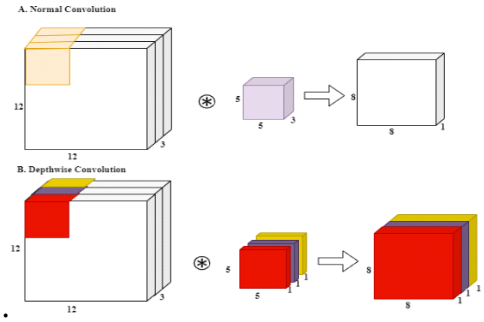
۳-٤- شبکه عصبی کانولوشن

مدل عصبی کانولوشن یک مدل یادگیری نظارت شده است که برای تحلیل‌های تصویری یا گفتاری در یادگیری ماشین استفاده می‌شوند. یک شبکه عصبی کانولوشن از لایه‌های کانولوشن، توابع غیرخطی، لایه‌های ادغام و لایه‌های کاملاً متصل تشکیل شده است. یک لایه کانولوشن شامل مجموعه‌ای از هسته‌ها است که پارامترهای آن‌ها باید آموخته شود و برای ا ستخراج ویژگی‌ها از داده‌ها ا ستفاده می شود. هر ه سته روی ارتفاع و عرض ورودی می‌لغزد و حا صل ضرب نقطه‌ای بین ورودی‌های ه سته و هر موقعیت ورودی محاسبه می‌شود. پس از هر لایه کانولوشن، یک تابع فعال‌سازی غیرخطی برای معرفی ویژگی‌های غیرخطی به مدل استفاده می‌شود. تابع غیرخطی Relu متداول‌ترین تابع فعال‌سازی در مدل‌های یادگیری عمیق هستند. یک لایه تلفیقی معمولاً بین دو لایه کانولوشن قرار می‌گیرد و برای کاهش تعداد پارامترها ا ستفاده می شود. لایه‌های نهایی شامل لایه‌های کاملاً مت صل ه ستند که ویژگی‌های ا ستخراج شده تو سط لایه قبلی را می‌گیرند و احتمالات یا امتیازات کلاس را ایجاد می‌کنند. این لایه‌ها به‌طور کامل به تمام نورون‌های لایه قبلی متصل هستند [۳] و [۱۱].

۳-۵- شبکه موبایل نت

موبایل نت یک کلاس از شبکه‌های عصبی هستند که توسط محققان شرکت گوگل در سال ۲۰۱۷ توسعه یافتند [۱۲]. موبایل نت یک شبکه عصبی کانولو شن سبک و سریع ا ست که برای پیاده سازی روی د ستگاه‌های موبایل طراحی شده‌اند. در این شبکه عصبی نوعی لایه کانولوشن جدید به نام

کانولوشن عمق-جداپذیر[۷] معرفی شد که به‌طور قابل‌توجهی تعداد پارامترها را در مقایسه با شبکه کانولوشن استاندارد کاهش می‌دهد [۱۲]. تفاوت اصلی بین لایه‌های کانولوشن‌های معمولی و کانولوشن‌های عمق-جداپذیر در این است که در لایه‌های کانولوشن‌های معمولی عملیات کانولوشن برای همه کانال‌های ورودی اعمال می‌شود، درحالی‌که در لایه‌های کانولوشن‌های عمق-جداپذیر عملیات کانولوشن به‌طور مجزا روی هر کانال انجام می‌شود. پس از اعمال عملیات کانولوشن بر هر کانال به‌طور جداگانه، یک لایه کانولوشن یک‌دریک اعمال می‌شود تا نتایج عملیات کانولوشن عمقی بر هر کانال مجزا را ترکیب کند. ازآنجایی‌که لایه‌های عمق-جداپذیر به محاسبات کمتری نسبت به لایه‌های کانولوشن معمولی نیاز دارد، موبایل نت سریع‌تر از مدل‌های کانولوشن معمولی است و انرژی کمتری مصرف می‌کند، بنابراین می‌تواند روی دستگاه‌های تلفن همراه بدون پردازنده‌های گرافیکی قدرتمند اجرا شوند [۱۲].



شکل (۴): پی‌چی‌دگی عمیق و کانولوشن معمولی[۱۳]

۶-۳- شبکه عصبی یامنت

یامنت یک شبکه عصبی عمیق است که ۵۱۲ کلاس را در مجموعه داده‌های صوتی پیش‌بینی می‌کند [۴]. یامنت از معماری موبایل نت ورژن یک استفاده می‌کند [۱۲]. در این شبکه ابتدا پیش‌پردازش زیر بر روی داده‌ها انجام می‌شود:

- تمام صوت‌ها به‌صورت مونو و ۱۶ کیلوهرتز نمونه‌برداری می‌شوند.
- یک طیف‌نگار با استفاده از تبدیل فوریه زمان-کوتاه با اندازه پنجره ۲۵ میلی‌ثانیه، جهش پنجره ۱۰ میلی‌ثانیه و یک پنجره هان دوره‌ای محاسبه می‌شود.
- در مرحله بعد یک طیف‌نگار مل با نگاشت طیف‌نگار ۶۴ مل باند محاسبه می‌شود. سپس این ویژگی‌ها در نمونه‌های ۵۰ درصد همپوشانی ۰٫۹۶ ثانیه‌ای قرار می‌گیرند که در آن هر نمونه ۶۴ باند مل و ۹۶ فریم هرکدام ۱۰ میلی‌ثانیه را پوشش می‌دهد.
- این بردارهای ۹۶ در ۶۴ تایی به مدل موبایل نت وارد می‌شوند.

خروجی لایه‌های کانولوشن این مدل یک بردار با بعد ۱۰۲۴ است و این خروجی به لایه‌های کلاس‌بندی برای کلاس‌بندی صوت‌ها داده می‌شود.

۴- روش پیشنهادی

مدل یامنت آموزش‌دیده برای استفاده و آموزش مدل بر روی داده‌های قابل‌دسترسی می‌باشد [۴]. این مدل فقط قابل‌استفاده برای یادگیری انتقالی است. در یادگیری انتقالی، مدل با مجموعه داده‌های بزرگی آموزش داده می‌شود و سپس از همان مدل مجدداً استفاده می‌شود و برای یک کار مشابه دوباره آموزش داده می‌شود [۱۴]. در این روش وزن‌های لایه‌های کانولوشن میانی ثابت و بدون تغییر می‌ماند و وزن‌های لایه‌های کلاس‌بندی با داده‌های جدید آموزش داده می شوند [۴]. برای آموزش مدل از ابتدا و آموزش وزن‌های لایه‌های میانی، باید تغییراتی در برنامه نوشته‌شده انجام شود. ما تغییرات لازم را بر روی کد انجام دادیم تا بتوانیم وزن تمام لایه‌های مدل را از ابتدا آموزش دهیم.

۱-۴- نحوه آموزش مدل‌ها

در یادگیری ماشینی، هایپرپارامتر پارامتری است که مقدار آن برای کنترل فرآیند یادگیری استفاده می‌شود. در مقابل، مقادیر سایر پارامترها (معمولاً وزن گره‌ها) از طریق آموزش به دست می‌آیند. در شبکه عصبی یامنت، هایپرپارامتر شامل اندازه پنجره STFT، فاصله بین پنجره‌های STFT، اندازه پنجره وردی به STFT، فاصله بین پنجره‌های ورودی STFT، نرخ یادگیری و کاهش نرخ یادگیری است. در این مقاله، هایپرپارامترها به صورت دستی انتخاب شده‌اند.

جدول (۱): هایپرپارامترهای استفاده‌شده برای شبکه عصبی یامنت

ردیف	هایپرپارامترهای	مقدار
۱	اندازه پنجره STFT	۰٫۰۲۵ ثانیه
۲	فاصله بین پنجره‌های STFT	۰٫۰۱ ثانیه
۳	اندازه پنجره وردی به STFT	۰٫۴۸ ثانیه
۴	فاصله بین پنجره‌های ورودی STFT	۰٫۲۴ ثانیه
۵	نرخ یادگیری	نرخ یادگیری
۶	کاهش نرخ یادگیری	کاهش نرخ یادگیری

در این کار از الگوریتم بهینه‌سازی آدام[۸] به‌عنوان بهینه‌سازی و از آنتروپی متقاطع به‌عنوان تابع هزینه استفاده شده است. داده‌ها به یک مجموعه آموزشی، مجموعه اعتبار سنجی و مجموعه آزمایشی تقسیم شده‌اند. مجموعه داده‌های آموزشی و آزمایشی به ترتیب برای آموزش مدل و ارزیابی عملکرد شبکه آموزش‌دیده استفاده می‌شود. مجموعه آزمایشی در طول آموزش و انتخاب هایپرپارامتر نادیده گرفته می‌شود و خطاهای آن تعیین می‌کند که چقدر مدل به داده‌های دیده نشده تعمیم می‌یابد. داده های اعتبار سنجی برای مقایسه مدل های مختلف، انتخاب هایپر پارامترها و جلوگیری از برازش بیش‌ازحد در مرحله آموزش استفاده می‌شود. ۱۰ درصد از داده‌ها به‌عنوان مجموعه آزمایشی انتخاب شد و مابقی به ترتیب به‌عنوان مجموعه‌های آموزشی و اعتبارسنجی به ۷۰ و ۲۰ تقسیم شدند. هر مدل حداکثر ۳۰ گام آموزش داده شد و بهینه‌سازی توقف زودهنگام برای جلوگیری از overfit استفاده شد. توقف زودهنگام یک استراتژی بهینه سازی است که هدف آن کاهش بیش‌برازش بدون به خطر انداختن دقت مدل است. بیش‌برازش به پدیده نامطلوبی در یادگیری ماشین گفته می‌شود که اگرچه مدل روی داده استفاده شده برای یادگیری بسیار خوب نتیجه می‌دهد، اما بر روی داده جدید دارای خطای زیاد است. در توقف زودهنگام تابع هزینه مجموعه اعتبار سنجی در طول آموزش ردیابی می شود و اگر پس از تعداد معینی از تکرار بهبود نیابد، آموزش متوقف می شود. ما وزن اولیه مدل‌ها را با استفاده از مدل آموزش‌دیده

در مجموعه داده Audioset مقداردهی کردیم و سپس این وزن‌ها بر روی داده‌های خود آموزش دادیم.

۵- مجموعه داده‌ها[۹]

داده‌های صوتی از اداره خدمات شرکت آب و فاضلاب جمع‌آوری شد. این داده‌ها بر اساس موضوع درخواست به چندین زیرمجموعه تقسیم می شدند. تعداد فایل‌های صوتی بر اساس موضوع درخواست متفاوت بود. برای داشتن کلاس‌هایی با تعداد فایل‌های صوتی متعادل، پنج کلاس انتخاب شد و کلمات کلیدی درخواست با توجه به موضوع درخواست از فایل صوتی انتخاب شده برش داده شد. این فایل‌های برش داده شده به‌عنوان ورودی مدل استفاده شدند. جدول (۲) جزییات این فایل‌ها را نشان می‌دهد.

جدول (۲): مجموعه داده

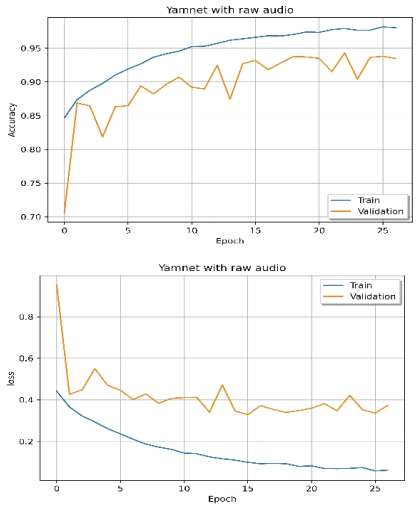
ردیف	عنوان فایل	تعداد فایل صوتی
۱	ترکیدگی لوله در کوچه	۲۵۷
۲	خروج جانوران موذی	۳۴۹
۳	ترکیدگی بین کنتور و محفظه	۳۵۰
۴	قطع آب	۴۱۴
۵	پرداخت	۴۰۰

٦- ارزیابی و آزمایش

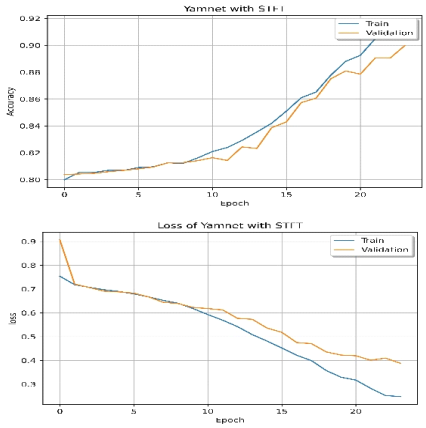
شکل‌های (۵)، (۶) و (۷) به ترتیب تابع هزینه و دقت مدل یامنت را با ورودی صوت خام، تبدیل فوریه کوتاه‌مدت و ضریب کپسترال فرکانس مل در طول آموزش نشان می‌دهند. همان‌طور که در شکل‌ها مشاهده می شود، زمانی که تبدیل فوریه کوتاه‌مدت و ضریب کپسترال فرکانس مل به‌عنوان ورودی استفاده می شوند، در ابتدای آموزش دقت مدل روی داده‌های آموزشی به‌دقت مدل روی داده‌های اعتبار سنجی نزدیک‌تر است از زمانی که صوت خام را به‌عنوان ورودی دریافت می‌کند است. این به این دلیل که در روش تبدیل فوریه کوتاه‌مدت و ضریب کپسترال فرکانس مل بعضی از ویژگی‌های توسط این دو روش استخراج شده و شبکه عصبی برای استخراج دیگر ویژگی‌های استفاده می‌شود ولی در روش ورودی داده خام هیچ‌گونه ویژگی استخراج نشده و مدل از ابتدا شروع به استخراج ویژگی می‌کند. چون از بهینه‌سازی توقف زودهنگام استفاده می‌کنیم، مراحل آموزش مدل‌ها با یکدیگر متفاوت است. در رویکرد مدل یامنت با ورودی تبدیل فوریه کوتاه‌مدت و ضریب کپسترال فرکانس مل، اندازه ورودی مدل پشته‌ای از بردارها به‌اندازه (۱،۶۴،۴۸) و در مورد ورودی خام، پشته‌ای از بردارها به‌اندازه (۱،۴۰۰،۴۶) است. ازاین‌رو، آموزش یامنت با ورودی صدای خام در مقایسه با موارد دیگر زمان بیشتری می‌برد. جدول (۳) نتایج کلاس‌بندی با مدل یامنت و ورودی‌های مختلف را نشان می‌دهد. از جدول، می‌توان دریافت که بهترین عملکرد مدل یامنت روی داده‌های تجربی زمانی حاصل می شود که از صوت خام به‌عنوان ورودی مدل استفاده کنیم. دومین عملکرد برتر با ضریب کپسترال فرکانس مل به‌عنوان ورودی مدل به‌دست‌آمده و بدترین عملکرد توسط مدل با ورودی تبدیل فوریه کوتاه‌مدت به دست آمد. به‌طورکلی تمامی مدل‌ها بر روی مدل کلاس‌بندی فارسی عملکرد معقولی دارند. ازاین‌رو یامنت مدل خوبی برای کلاس‌بندی صوت فارسی است.

جدول (۳): دقت مدل با وردی‌ها متفاوت بر روی داده‌های آزمایشی

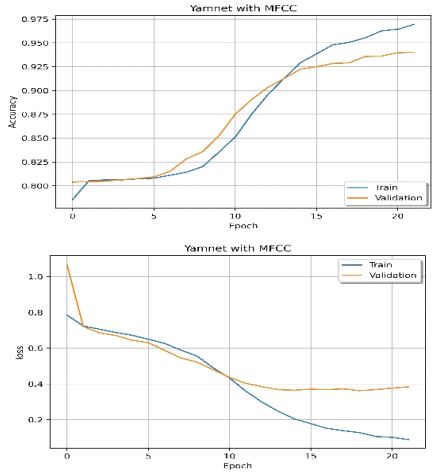
ردیف	مدل	دقت (٪)
۱	وردی صوت خام	۹۳
۲	ورودی MFCC	۹۱,۶
۳	ورودی STFT	۸۸,۲۴



شکل (۵): تغی‌ی‌رات تابع هزی‌نه و دقت در طول آموزش برای ی‌امنت با ورودی داده خام.



شکل (۶): تغی‌ی‌رات تابع هزی‌نه و دقت در طول آموزش برای ی‌امنت با ورودی تبدی‌ل فوری‌ه کوتاه مدت



شکل (۷): تغی‌ی‌رات تابع هزی‌نه و دقت در طول آموزش برای ی‌امنت با ورودی تبدی‌ل فوری‌ه کوتاه مدت

۷- نتیجه‌گیری

در این کار به بررسی توانایی مدل یامنت در کلاس‌بندی صدای فارسی پرداختیم. مدل اصلاح شد تا بتوان صدای فارسی خام را به‌عنوان ورودی کلاس‌بندی کرد. این مدل در مجموعه آزمایش به‌دقت ۹۳ در صد د ست‌یافت که نشان می‌دهد این مدل برای این منظور مناسب است. عملکرد مدل با صوت‌های خام به‌عنوان ورودی با مدل‌هایی با ورودی تبدیل فوریه کوتاه‌مدت و ضریب کپسترال فرکانس مل مقایسه شد. مدل با صدای خام به‌عنوان ورودی بهترین عملکرد را در مجموعه آزمایشی به دست آورد که نشان می‌دهد شبکه ع صبی کانولو شن مدل خوبی برای ا ستخراج ویژگی‌ها از فایل صوتی نسبت به ورودی تبدیل فوریه کوتاه‌مدت و ضریب کپسترال فرکانس مل است. یکی از معایب ا ستفاده از صوت خام به‌عنوان ورودی مدل این ا ست که ابعاد ورودی افزایش می‌یابد و زمان بیشتری برای آموزش مدل نیاز است. بااین‌حال، مزیت داده‌های خام این اســت که در زمان اســتفاده از یک مدل آموزش‌دیده برای پیش‌بینی، نیازی به پیش‌پردازش داده‌ها با تبدیل فوریه کوتاه‌مدت یا ضریب کپسترال فرکانس مل نیست و برای پیش‌پردازش داده‌ها به زمان کمتری نیاز داریم.

مراجع

[1] A.-r. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in ۲۰۱۲ □□□□ □□□□□□□□□□□□□□ □□□□□□□□□□□ □□ □□□□□□□□□□, *speech and signal processing (ICASSP)*, 2012: IEEE, pp. ۴۲۷۳-۴۲۷۶.

[2] D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition," *Speech Communication*, vol. 108, pp. 15-3۲, ۲۰۱۹.

[3] G. Nguyen *et al.*, "Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey," *Artificial Intelligence Review*, vol. 52, no. 1, pp. ۷۷-۱۲۴, ۲۰۱۹.

[4] [Online].Available:https://github.com/tensorflow/models/tree/master/research/.

[5] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011: IEEE, pp. 5884-5887.

[6] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *arXiv preprint arXiv:1304.1018*, 2013.

[7] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in LVCSR," in *Sixteenth annual conference of the international speech communication association*, 2015.

[8] T. Sainath, R. J. Weiss, K. Wilson, A. W. Senior, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," 2015.

[9] M. Zeitler, P. Fries, and S. Gielen, "Biased competition through variations in amplitude of γ-oscillations," *Journal of computational neuroscience*, vol. 25, no. 1, pp. 89-1۰۷, ۲۰۰۸.

[10] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, and Q. Tian, "HMM-based audio keyword generation," in *Pacific-Rim Conference on Multimedia*, 2004: Springer, pp. 566-5۷۴.

[11] M. V. Valueva, N. Nagornov, P. A. Lyakhov, G. V. Valuev, and N. I. Chervyakov, "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation," *Mathematics and computers in simulation*, vol. 177, pp. ۲۳۲-۲۴۳, ۲۰۲۰.

[12] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[13] K. Alibabaei, E. Assunção, P. D. Gaspar, V. N. Soares, and J. M. Caldeira, "Real-Time Detection of Vine Trunk for Robot Localization Using Deep Learning Models Developed for Edge TPU Devices," *Future Internet*, vol. ۱۴, □□. ۷, □. ۱۹۹, ۲۰۲۲.

[14] J. West, D. Ventura, and S. Warnick, "Spring research presentation: A theoretical foundation for inductive transfer," *Brigham Young University, College of Physical and Mathematical Sciences*, vol. 1, no. 08, 2007.

زیرنویس‌ها

۱ Deep learning
۲ Transfer larning
۳ Restricted Boltzmann Machine
٤ Artificial Neural Networks (ANN)
٥ Short-Time Fourier Transform (STFT)
٦ mel frequency cepstral coefficient

[۹] Data set

[۷] depth-wise separable convolution

[۸] Adam