به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

یادگیری ماشین

گزارش اولیه پروژه

نام و نام خانوادگی	امیرحسین قاسمی- حدیثه مصباح- امیرپویا کارخانه یوسفی- محمد مهدی سلمانی زارچی
شماره دانشجویی	۸۱۰۱۰۲۱۷۴-۸۱۰۱۰۰۴۴۰-۸۱۰۱۰۲۲۵۳-۸۱۰۱۰۱۰۳۲
تاریخ ارسال گزارش	۱۴۰۲/۰۹/۳۰

فهرست

۱-توضیحی مختصر راجع به نحوه کار با داده های صوتی در الگوریتم های یادگیری ماشین

۱-۱-مقدمه

فناوری تشخیص گفتار یکی از شگفتی‌های محاسبات مدرن است که امکان تبدیل صدا به زبان نوشتاری را فراهم می‌کند. این فرآیند پیچیده در چهار مرحله اصلی کار می‌کند: تجزیه و تحلیل، جدا کردن، دیجیتالی کردن و تطبیق صدا با مناسب ترین نمایش متن با استفاده از یک الگوریتم. به این ترتیب، کلمات گفتاری می‌توانند به متن قابل خواندن توسط کامپیوتر تبدیل شوند که هم ماشین‌ها و هم انسان‌ها آن را درک می‌کنند.

برای رمزگشایی دقیق گفتار انسان، نرم‌افزار تشخیص گفتار باید بتواند خود را با محیط‌های در حال تغییر وفق دهد. الگوریتم‌هایی که ضبط‌های صوتی را به بازنمایی‌های متنی تجزیه و تحلیل می‌کنند، بر روی مدولاسیون‌های صوتی مختلف مانند لهجه‌ها، گویش‌ها، سبک‌های گفتاری، جمله‌بندی‌ها و الگوهای گفتاری آموزش داده می‌شوند. علاوه بر این، این فناوری با قابلیت حذف نویز طراحی شده است تا کلمات گفتاری را از هر صدای پس زمینه حواس پرتی متمایز کند.

تبدیل سیگنال‌های صوتی به داده‌هایی که رایانه می‌تواند درک کند، فناوری‌های صوتی تشخیص خودکار گفتار اغلب با استفاده از یک مدل آکوستیک آغاز می‌شوند. همانطور که یک دماسنج دیجیتال خوانش دمای آنالوگ را به اعداد تبدیل می‌کند، مدل آکوستیک امواج صوتی را به کد باینری تبدیل می‌کند. سپس، مدل‌های زبان و تلفظ با استفاده از زبان‌شناسی محاسباتی برای تشکیل کلمات و جملات از هر صدا در بافت و توالی استفاده می‌کنند.

با این حال، پیشرفت‌های اخیر در فناوری تشخیص خودکار گفتار، رویکرد جدیدی را برای این فرآیند اتخاذ می‌کند و از مدل شبکه عصبی انتها به انتها به جای تکیه بر الگوریتم‌های متعدد استفاده می‌کند. مدل‌های انتها به انتها دقیق‌تر و مؤثرتر بوده‌اند، اما مدل‌های هیبریدی همچنان بیشترین استفاده را در سیستم‌های ASR تجاری دارند.

برای تهیه یک گزارش اولیه در مورد کار با داده‌های صوتی در الگوریتم‌های یادگیری ماشین، باید این مسئله را از چندین جنبه کلیدی مورد بررسی قرار دهیم. بدین منظور بخش‌های بعدی را ارائه خواهیم داد.

۲-۱-مقدمه‌ای بر داده‌های صوتی در یادگیری ماشین

۱-۲-۱-مروری بر داده های صوتی

داده‌های صوتی به شکل دیجیتالی، منبعی پیچیده و غنی از اطلاعات است. طیف وسیعی از فعالیت‌های انسانی و صداهای محیطی را در بر می‌گیرد و یک رسانه منحصر به فرد برای کاربردهای یادگیری ماشین ارائه می‌دهد. برخلاف داده‌های بصری، داده‌های صوتی موقت هستند و اغلب به تکنیک‌های پردازش متفاوتی نیاز دارند. در شکل خام و اولیه، داده‌های صوتی معمولاً به صورت شکل موج نمایش داده می شوند که نمودار فشار صوت در برابر زمان است.

۲-۲-۱-ویژگی‌های داده‌های صوتی دیجیتال

Sampling Rate: به تعداد نمونه‌های صوتی منتقل شده در هر ثانیه اشاره دارد که بر حسب هرتز یا کیلوهرتز اندازه گیری می‌شود. نرخ نمونه‌برداری بالاتر منجر به کیفیت بهتر اما اندازه داده‌های بزرگتر می‌شود.

Bit Depth: وضوح صدا را تعیین می‌کند. تعداد بیت‌های استفاده شده برای هر نمونه را نشان می‌دهد که بر محدوده دینامیکی و سطح نویز ضبط تاثیر می‌گذارد.

Channels: داده‌های صوتی برای سیستم‌های صدای فراگیر می توانند مونو (تک کانال)، استریو (دو کانال) یا چند کاناله باشند.

۳-۲-۱-اهمیت در یادگیری ماشین

داده‌های صوتی به دلیل کاربرد گسترده، پتانسیل بسیار زیادی در زمینه یادگیری ماشین دارند. الگوریتم‌های یادگیری ماشین می‌توانند الگوها و بینش‌های معنی‌داری را از داده‌های صوتی استخراج کنند و در کاربردهای مختلفی مورد استفاده قرار گیرند که به برخی از آن‌ها در ادامه اشاره خواهیم کرد.

تشخیص گفتار: ترجمه کلمات گفتاری به متن، مورد استفاده در دستیاران مجازی، خدمات رونویسی و ترجمه زبان.

تجزیه و تحلیل موسیقی: شناسایی ژانرها، حالات، یا حتی تولید آهنگ های جدید.

طبقه‌بندی صدا: تشخیص صداهای مختلف محیطی، مفید در سیستم‌های امنیتی یا نظارت بر حیات-وحش.

تشخیص احساسات: تجزیه‌وتحلیل الگوهای صوتی برای ارزیابی حالات عاطفی، قابل استفاده در خدمات مشتری و ارزیابی سلامت روان.

۴-۲-۱-چالش‌های منحصر به فرد

کار با داده های صوتی چالش های منحصر به فردی را ایجاد می‌کند که آن را از انواع دیگر داده‌ها متمایز می کند که به برخی از آن‌ها اشاره خواهیم نمود.

ماهیت زمانی: داده‌های صوتی ذاتاً داده‌های سری زمانی هستند که به توجه ویژه برای ویژگی‌های زمانی آن نیاز دارند.

تنوع و پیچیدگی: صداها می توانند از نظر زیر و بم، لحن و مدت زمان بسیار متفاوت باشند و به تحلیل پیچیدگی اضافه کنند.

نویز: صداها اغلب شامل نویزهای نامربوط یا حواس‌پرتی است که به تکنیک‌های موثر کاهش نویز نیاز دارند.

۵-۲-۱-نتیجه

درک مبانی داده‌های صوتی برای به کارگیری موثر تکنیک‌های یادگیری ماشین بسیار مهم است. ماهیت غنی و متنوع صدا، چالش‌ها و فرصت‌هایی را برای راه‌حل‌های نوآورانه در کاربردهای متعدد ارائه می‌دهد. همانطور که یادگیری ماشین به تکامل خود ادامه می دهد، نقش داده های صوتی در این حوزه در حال گسترش است و مرزهای جدیدی را برای اکتشاف و توسعه ارائه می‌دهد.

۳-۱-پردازش داده‌های صوتی

پردازش داده‌های صوتی و به نوعی تمیز کردن داده‌های صوتی، گامی حیاتی در آماده‌سازی آن برای مدل‌های یادگیری ماشین است. این مرحله شامل تکنیک‌های مختلفی برای تبدیل صدای خام به فرم قابل استفاده‌تر می‌شود که اثربخشی تجزیه و تحلیل بعدی را افزایش می‌دهد.

۱-۳-۱-تکنیک‌های پیش‌پردازش

یکی از تکنیک‌های پیش پردازش، کاهش نویز می‌باشد. هدف از این کار حذف نویز ناخواسته و تداخلی که می‌تواند سیگنال صوتی را مخدوش کند می‌باشد. از روش‌هایی که برای اینکار وجود دارد می‌توان به گیتینگ طیفی، فیلتر وینر، و حذف نویز مبتنی بر یادگیری ماشین اشاره نمود.

تکنیک دیگر پیش‌پردازش، نرمال‌سازی می‌باشد هدف آن استاندارد کردن سطوح نویز سیگنال صوتی و اطمینان حاصل کردن از دامنه ثابت در نویزهای مختلف می‌باشد. روش‌هایی که برای این کار وجود دارد شامل نرمال‌سازی پیک، نرمال‌سازی RMS یا نرمال‌سازی بزرگی نویز بر اساس بزرگی قابل درک می‌باشد.

همچنین از Mel-frequency Cespstral Coefficients (MFCCs) نیز به‌طور گسترده برای نشان دادن طیف قدرت نویز استفاده می‌شود. در واقع اینگونه پاسخ گوش انسان به نویز را تقلید می‌کنند و به ویژه برای تجزیه و تحلیل گفتار و موسیقی موثر هستند. از طیف‌نگارها نیز جهت نمایش‌های بصری طیف فرکانس‌ها در نویز در جایی که با زمان تغییر می‌کنند استفاده می‌شود که برای شناسایی الگوهای زمانی مفید است. نرخ عبور از صفر نیز از سرعتی که شکل موج صوتی از محور دامنه صفر عبور می‌کند و برای تجزیه و تحلیل محتوای فرکانس سیگنال صوتی مفید است استفاده می‌کند. ویژگی‌های کروما نیز ویژگی‌های هارمونیک و ملودیک موسیقی را که برای تجزیه و تحلیل موسیقی مفید است، استفاده خواهد کرد. موارد اشاره شده مصداق بارز استخراج ویژگی از داده‌های صوتی می‌باشند که به‌صورت مختصر به‌آن‌ها اشاره نمودیم.

۲-۳-۱-افزایش داده‌ها

هدف از افزایش داده‌ها، گسترش مصنوعی مجموعه داده با ایجاد نسخه‌های اصلاح شده از فایل‌های صوتی موجود می‌باشد. این کار به بهبود استحکام مدل‌های یادگیری ماشین کمک می‌کند و از برازش بیش از حد جلوگیری می‌کند. این کار تغییر زیر و بم، سرعت یا اضافه کردن نویز مصنوعی به صدا انجام خواهد شد. از انواع روش‌های مورد استفاده برای این کار می‌توان به موارد زیر اشاره نمود.

Pitch Shifting: جهت تغییر زیر و بم صدا بدون تغییر مدت زمان آن.

کشش زمان: برای کاهش یا افزایش سرعت صدا بدون تأثیر بر سرعت آن.

افزودن نویز مصنوعی: معرفی انواع نویز برای شبیه‌سازی محیط‌های مختلف گوش دادن.

فشرده سازی محدوده دینامیکی: تغییر دامنه بزرگی نویز.

انتخاب ویژگی‌ها به شدت به مسئله موجود بستگی دارد. به عنوان مثال، MFCC ها معمولاً در تشخیص گفتار استفاده می شوند، در حالی که ویژگی های کروما برای تجزیه و تحلیل موسیقی بیشتر مرتبط هستند.

به طور خلاصه، کار با داده‌های صوتی در یادگیری ماشینی شامل تمیز کردن داده‌ها برای حذف نویز و بخش‌های نامربوط، استخراج ویژگی‌های معنی‌دار که به طور مؤثر صدا را نشان می‌دهند، و عادی کردن

این ویژگی‌ها برای آماده‌سازی آن‌ها برای استفاده در مدل‌های یادگیری ماشین است. فرآیندها و ویژگی های دقیق مورد استفاده به نیازهای مسئله بستگی دارد.

۳-۳-۱-تقسیم‌بندی صدا

هدف از این کارتقسیم صداهای ضبط شده طولانی به بخش‌های کوتاه‌تر و قابل کنترل‌تر کردن آن‌ها برای تجزیه و تحلیل می‌باشد. روش‌هایی که برای این کار وجود دارد، تقسیم‌بندی مبتنی بر سکوت، تقسیم‌بندی مبتنی بر انرژی، و رویکردهای مبتنی بر یادگیری ماشین مانند تشخیص نقطه تغییر می‌باشد.

۴-۳-۱-نرمال‌سازی و پاکسازی داده‌ها و ویژگی‌ها

در این روش اطمینان از سازگاری در مجموعه داده با حذف بخش‌های نامربوط از صدا و استاندارد کردن قالب فایل، عمق بیت و نرخ نمونه‌برداری انجام خواهد شد. این کار با استفاده از کاهش سکوت، تبدیل فرمت‌های فایل، نمونه‌برداری مجدد و تبدیل عمق بیت صورت خواهد گرفت.

در Feature Scaling نیز درست مانند سایر اشکال داده، مقیاس‌بندی ویژگی در پردازش صدا، محدوده‌های عددی مختلف را به یک محدوده استاندارد می‌آورد. اینکار برای مدل-های حساس به مقیاس داده‌های ورودی مانند SVM یا شبکه‌های عصبی مهم است.

Mean Normalization هم برای تنظیم مقادیر ویژگی‌ها برای داشتن میانگین صفر به‌کار خواهد رفت. اینکار به متمرکز کردن داده‌ها کمک می‌کند.

با استفاده ازVariance Scaling، اطمینان از اینکه همه ویژگی‌ها واریانس یکسانی دارند حاصل خواهد شد. این کار می‌تواند به بهبود عملکرد مدل کمک کند زیرا با همه ویژگی‌ها به طور یکسان رفتار می‌کند.

۵-۳-۱-نتیجه‌گیری

پردازش داده های صوتی یک کار چند وجهی است که نیازمند بررسی دقیق ویژگی‌های صدا و الزامات کار یادگیری ماشین است. پیش‌پردازش موثر، استخراج ویژگی، تقویت داده‌ها و پاکسازی داده‌ها برای اطمینان از قابلیت اطمینان و دقت مدل‌های یادگیری ماشین آموزش داده شده بر روی داده‌های صوتی ضروری است. همانطور که زمینه پردازش صدا در حال پیشرفت است، این تکنیک ها دائماً اصلاح و گسترش می یابند و راه‌های پیچیده‌تری برای استفاده از قدرت داده‌های صوتی در یادگیری ماشین ارائه می‌دهند.

۴-۱- مدل‌های یادگیری ماشین برای داده‌های صوتی

در حوزه یادگیری ماشین، مدل‌ها و الگوریتم‌های مختلف به ویژه برای مدیریت داده‌های صوتی مناسب هستند. این مدل‌ها می توانند الگوها را استخراج کنند، ویژگی‌ها را تشخیص دهند و بر اساس ورودی‌های

صوتی پیش‌بینی کنند. انتخاب مدل اغلب به وظیفه خاصی مانند طبقه بندی، پیش بینی یا تولید بستگی دارد.

۱-۴-۱-مدل‌های یادگیری نظارت‌شده

یکی از این مدل‌ها مبتنی بر طبقه بندی task می‌باشد. هدف این کار دسته‌بندی داده‌های صوتی به کلاس‌های از پیش تعریف شده خواهد بود.

یکی از مدل‌های مورد استفاده قرار گرفته شده، ماشین‌های بردار پشتیبان (SVM) است که برای مجموعه داده‌های کوچک تا متوسط مؤثر است که اغلب برای طبقه‌بندی ژانر، تشخیص احساسات و غیره استفاده می‌شود.

درخت تصمیم و جنگل‌های تصادفی نیز از مدل‌های دیگر هستند که برای مدل‌های قابل تفسیر، که اغلب در طبقه‌بندی نویزمحیطی استفاده می‌شوند، مفید است.

ماشین‌های تقویت گرادیان نیز مانند XGBoost برای کارهای طبقه‌بندی با عملکرد بالا استفاده می‌شود.

taskهای رگرسیون نیز وجود دارد که هدف استفاده از آن‌ها، پیش‌بینی مقادیر پیوسته از داده‌های صوتی، مانند تخمین سن یا خلق و خوی بر اساس صدا خواهد بود. مدل‌های استفاده شده در این حال نیز مدل‌های رگرسیون خطی، رگرسیون چند جمله‌ای برای روابط پیچیده‌تر می‌باشند.

۲-۴-۱-مدل‌های یادگیری بدون نظارت

خوشه بندی از معروف‌ترین کارهایی است که در یادگیری ماشین استفاده می‌شود و هدف از آن‌ها گروه بندی انواع مشابه صداها یا موسیقی بدون برچسب‌های از پیش تعریف شده خواهد بود. مدل‌های معروف در این دسته‌بندی نیز K-means، خوشه‌بندی سلسله مراتبی و مدل‌های مخلوط گاوسی می‌باشند.

در این دسته از مدل‌ها کاهش ابعاد نیز مورد بررسی قرار می‌گیرد که هدف آن ساده‌کردن ویژگی‌ها در عین حفظ اطلاعات ضروری می‌باشد. مدل‌های مورد استفاده نیز تجزیه و تحلیل مؤلفه اصلی (PCA) و t-Distributed Stochastic Neighbor Embedding (t-SNE). می‌باشد.

۳-۴-۱-رویکردهای یادگیری عمیق

در یادگیری عمیق شبکه‌های عصبی وجود دارند که از معروف‌ترین آن‌ها می‌توان به شبکه‌های عصبی کانولوشن (CNN) اشاره نمود که هدف آن استخراج ویژگی‌های سلسله مراتبی از طیف‌نگارها یا MFCC می‌باشد. کاربردهای آن نیز در طبقه بندی صدا، برچسب گذاری صوتی، تشخیص گفتار می‌باشد.

شبکه های عصبی مکرر (RNN) و شبکه‌های حافظه کوتاه مدت (LSTM) نیز دسته‌دیگری از شبکه-های عصبی می‌باشند که هدف آن‌ها مدل‌سازی وابستگی‌های زمانی در داده‌های صوتی است که برای کارهایی که شامل توالی‌هایی مانند گفتار یا موسیقی است، ضروری است. از کاربردهای آن نیز می‌توان به سنتز گفتار، تولید موسیقی، تجزیه و تحلیل توالی زمانی اشاره نمود.

دسته دیگر شبکه‌های عصبی، مبدل‌ها (Transformers) می‌باشد که هدف این شبکه‌ها رسیدگی به وابستگی‌های دوربرد و الگوهای پیچیده در داده‌های صوتی است. از کاربردهای آن می‌توان به سیستم‌های تشخیص گفتار پیشرفته، توصیه موسیقی، تشخیص رویداد صوتی اشاره نمود که استفاده می‌شود.

نوع دیگری از شبکه‌های عصبی، رمزگذارهای خودکار (Autoencoders) می‌باشند که برای یادگیری ویژگی، کاهش ابعاد، و taskهای تولیدی در صدا استفاده خواهد شد. کاربردهای آن‌ها نیز فشرده سازی صدا، حذف نویز و مدل های تولیدی برای موسیقی می‌باشد.

۵-۴-۱-یادگیری انتقالی

این رویکرد از مدل‌های از پیش آموزش دیده بهره می‌برد. این کار اجازه می‌دهد تا مدل را در کار جدید با مجموعه داده کوچکتر تنظیم کنید. مبتنی بر این ایده است که دانش آموخته شده در یک کار می‌تواند برای بهبود عملکرد در کار دیگر استفاده شود.

۶-۴-۱-یادگیری چند وظیفه‌ای

این رویکرد به مدل اجازه می‌دهد تا چندین کار را همزمان یاد بگیرد. از اطلاعات به اشتراک گذاشته شده بین آن‌ها برای بهبود عملکرد وظیفه اصلی استفاده می‌کند.

شایان ذکر است که انتخاب روش آموزشی به ویژگی‌های کار و منابع موجود بستگی دارد. روش‌های یادگیری تحت نظارت دقیق‌ترین و در عین حال فشرده‌ترین روش‌های یادگیری هستند. روش های یادگیری بدون نظارت و با نظارت ضعیف به داده‌های کمتری نیاز دارند اما ممکن است عملکرد پایین‌تری داشته باشند. انتقال و یادگیری چند وظیفه‌ای می‌تواند عملکرد را بدون افزایش داده‌های مورد نیاز بهبود بخشد.

۷-۴-۱-چالش‌ها

یکی از چالش‌های موجود در این مسئله، اندازه و کیفیت داده می‌باشد که مسئله‌ای مهم است. زیرا مجموعه داده‌های با کیفیت بالا و بزرگ برای آموزش مدل‌های مؤثر، به‌ویژه مدل‌های یادگیری عمیق، حیاتی هستند.

دیگر چالش موجود، پیچیدگی زمانی داده‌ها می‌باشد. داده‌های صوتی اغلب شامل ساختارهای زمانی پیچیده‌ای هستند که برای ثبت این پویایی‌ها به مدل‌های پیچیده نیاز دارند.

چالش دیگر، منابع محاسباتی موجود می‌باشد که مدل‌های یادگیری عمیق، به ویژه، به قدرت محاسباتی قابل توجهی برای آموزش و استنتاج نیاز دارند.

۸-۴-۱-نتیجه‌گیری

مدل‌های یادگیری ماشین برای داده‌های صوتی متنوع و همه‌کاره هستند، از تکنیک‌های یادگیری ماشین سنتی گرفته تا معماری‌های یادگیری عمیق پیشرفته. هر مدل نقاط قوت خود را دارد و برای انواع مختلف وظایف آنالیز صدا مناسب است. پیشرفت های مستمر در این مدل‌ها، همراه با افزایش قدرت محاسباتی و در دسترس بودن داده ها، باعث پیشرفت چشمگیر در زمینه تجزیه و تحلیل صوتی و یادگیری ماشین شده است.

۵-۱- چالش‌ها و راه‌حل‌ها در کار با داده‌های صوتی در یادگیری ماشین

۱-۵-۱-چالش‌های مسئله

از چالش‌های اساسی این مسئله، مدیریت مجموعه داده‌های بزرگ می‌باشد. مجموعه داده‌های صوتی، به‌ویژه آن‌هایی که برای یادگیری عمیق مورد نیاز هستند، می‌توانند حجیم و پیچیده باشند و به منابع ذخیره‌سازی و محاسباتی قابل توجهی نیاز دارند. تاثیر این موضوع می‌تواند منجر به افزایش هزینه‌ها و چالش‌های محاسباتی، به ویژه برای برنامه‌های کاربردی بلادرنگ شود.

چالش دیگر این است که داده های صوتی می‌توانند متنوع و بدون ساختار باشند. در واقع داده‌های صوتی در فرمت‌های مختلف، سطوح کیفی و شامل طیف گسترده‌ای از صداها و نویزها هستند. این تغییرپذیری ایجاد یک مدل یک‌اندازه برای همه را دشوار می‌کند و بر دقت و تعمیم‌پذیری مدل تأثیر می‌گذارد.

چالش مهم دیگر نیز Overfitting در مدل‌های پیچیده خواهد بود. مدل‌های یادگیری عمیق، با ظرفیت بالای خود، به راحتی می‌توانند به داده‌های آموزشی اضافه شوند، به خصوص زمانی که داده‌ها به اندازه کافی متنوع نباشند. overfitting منجر به مدل‌هایی می‌شود که در داده‌های آموزش عملکرد خوبی دارند اما در داده‌های دیده نشده ضعیف هستند و به‌اصطلاح تعمیم‌پذیری پایینی خواهند داشت.

چالش دیگر وابستگی‌های زمانی می‌باشد. داده‌های صوتی ذاتاً زمانی و متوالی هستند، که در گرفتن وابستگی‌های بلندمدت چالشی ایجاد می‌کند. مشکل اصلی در مدل‌سازی دقیق توالی‌هایی مانند گفتار یا موسیقی، که در آن زمینه و نظم بسیار مهم است به‌وجود خواهد آمد.

مسئله مهم دیگر نویز و تغییرپذیری است که صدای واقعی اغلب حاوی نویز و تغییرات به دلیل شرایط ضبط است. تاثیر نویز و تغییرپذیری می‌توانند ویژگی‌های مهم در داده‌ها را پنهان کنند و عملکرد مدل را کاهش دهند.

۲-۵-۱-راه‌حل‌ها

از راه‌های برطرف ساختن چالش‌ها و مشکلات اشاره شده، می‌توان به ذخیره‌سازی و پردازش کارآمد داده ها اشاره نمود. این کار با استفاده از خدمات ابری صورت می‌گیرد که هدف آن استفاده از منابع ذخیره سازی ابری و محاسباتی برای مقیاس پذیری و کارایی خواهد بود.

از دیگر کارهایی که می‌توان انجام داد، استفاده از تکنیک‌های فشرده‌سازی داده‌ها می‌باشد که استفاده از روش هایی مانند فشرده سازی صدا برای کاهش اندازه داده ها بدون افت کیفیت قابل توجه صورت خواهد گرفت.

تکنیک‌های منظم‌سازی و اعتبار سنجی متقابل نیز از روش‌های دیگر است که از روش‌هایی مانند حذف و عادی سازی دسته ای استفاده می‌شود که از دسته تکنیک هایی است که برای جلوگیری از برازش بیش از حد در شبکه های عصبی استفاده می‌شود.

اعتبار سنجی متقابل نیز راهکار دیگری است که استفاده از تکنیک‌هایی مانند اعتبار سنجی متقاطع k-fold برای اطمینان از تعمیم مدل به خوبی به داده های دیده نشده و به‌صورت کلی generalization را صورت می‌دهد.

پیش پردازش پیشرفته و استخراج ویژگی نیز موضوع مهم دیگری است که در آن‌ها از الگوریتم‌های کاهش نویز استفاده می‌شود. استفاده از تکنیک‌های پیشرفته حذف نویز برای تمیز کردن داده های صوتی از خروجی‌های استفاده از این روش است.

استخراج ویژگی قوی و مهم نیز روش دیگری است که با استفاده از ویژگی‌هایی مانند MFCC یا طیف نگاره‌هایی که کمتر در معرض نویز و تغییرپذیری هستند این کار را انجام می‌دهد.

استفاده نمودن از مدل‌های مناسب نیز موضوعی مهم است. به‌طور مثال RNN و LSTMs برای مدل‌سازی وابستگی‌های زمانی به طور موثر در داده‌های متوالی مورد استفاده قرار می‌گیرند.

مکانیسم‌های توجه و ترانسفورماتورها نیز مدل‌های دیگری هستند که همانطور که اشاره نمودیم، برای ثبت وابستگی‌های دوربرد و الگوهای پیچیده در صدا مورد استفاده قرار می‌گیرند. با توجه به کاربرد و استفاده‌ای که می‌خواهیم داشته باشیم، باید به انتخاب مدل توجه ویژه‌ای داشته باشیم.

افزایش و تنوع داده‌ها نیز موضوعی قابل توجه است که افزایش داده‌های آموزشی منجر به ایجاد تغییرات افزایش تنوع و استحکام خواهد شد.

جمع آوری نمونه‌های داده‌های متنوع نیز اهمیت زیادی دارد. اطمینان از مجموعه داده های آموزشی شامل طیف گسترده ای از شرایط ضبط و انواع صدا خواهد بود.

۳-۵-۱-نتیجه‌گیری

چالش‌های کار با داده‌های صوتی برای یادگیری ماشین بسیار مهم هستند، اما غیرقابل حل نیستند. با به کارگیری استراتژی‌های مدیریت داده کارآمد، مدل‌های پیشرفته یادگیری ماشین و تکنیک‌های پیش پردازش قوی، می‌توان به این چالش‌ها به طور موثر رسیدگی کرد. علاوه بر این، پیشرفت‌های مداوم در فناوری و روش شناسی به طور پیوسته بر این موانع غلبه می‌کند و راه را برای سیستم‌های تجزیه و تحلیل صوتی پیچیده‌تر و دقیق‌تر هموار می‌کند.

۶-۱- مطالعات موردی و کاربردهای داده های صوتی در یادگیری ماشین

۱-۶-۱-مطالعات موردی

از جمله کاربردهایی که وجود دارد، دستیارهای صوتی (به عنوان مثال، سیری، الکسا، دستیار گوگل) می‌باشند که چالش اصلی آن‌ها درک و پاسخگویی دقیق به گفتار انسان در حالت بلادرنگ می‌باشد. برای تحقق این خواسته استفاده از الگوریتم‌های پیشرفته تشخیص گفتار و پردازش زبان طبیعی برای تفسیر و پاسخ به پرسش های کاربر مورد استفاده قرار می‌گیرد. با این کار تعامل انسان و رایانه دچار تحول شده و فناوری در دسترس‌تر و شهودی‌تر خواهد شد.

استفاده دیگر در سیستم‌های توصیه موسیقی (به عنوان مثال، Spotify، Apple Music) می‌باشد که چالش این کار ارائه توصیه‌های موسیقی شخصی به کاربران است که کاری دشوار است. راه‌حل برای این قضیه این است که مدل‌های یادگیری ماشین ترجیحات موسیقی، سابقه گوش دادن و ویژگی‌های صوتی آهنگ‌ها را تجزیه و تحلیل کنند تا فهرست‌های پخش شخصی‌سازی شده را پیشنهاد دهند. با تنظیم موسیقی که با سلیقه‌های فردی همسو می‌شود، تجربه کاربر را بهبود می‌بخشد.

طبقه بندی صدای محیطی برای شهرهای هوشمند نیز از کاربردهای مهم دیگر است که نظارت و طبقه‌بندی صداهای شهری برای ایمنی عمومی و نظارت بر محیط زیست را انجام می‌دهد. برای انجام این کار استقرار شبکه‌های حسگر که از یادگیری ماشین برای شناسایی و دسته‌بندی صداهای شهری (مانند ترافیک، آژیرها و غیره) استفاده می‌کنند، امری ضروری است. این کار موجب بهبود مدیریت شهری و ایمنی عمومی از طریق تجزیه و تحلیل صدا در حالت بلادرنگ خواهد شد.

کاربردهای مراقبت های بهداشتی (به عنوان مثال، تجزیه و تحلیل ضربان قلب و سرفه) نیز از موارد دیگر است که چالش آن تشخیص شرایط سلامت از طریق تجزیه و تحلیل صوتی غیر تهاجمی می‌باشد. برای این کار مدل‌های یادگیری ماشین آموزش دیده برای تشخیص الگوهای موجود در داده‌های صوتی که با شرایط پزشکی خاص مرتبط است مورد استفاده باید قرار گیرند که به موجب آن، تشخیص از راه دور و غیر تهاجمی را فعال می‌کند که این کار به ویژه در پزشکی از راه دور و نظارت بر بیمار مفید است.

۲-۶-۱-کاربردها

یکی از کاربردها، تبدیل گفتار به متن می‌باشد که در صنایع حقوقی، پزشکی و رسانه‌ای برای تبدیل گفتار به متن نوشتاری کارآمد استفاده می‌شود.

کاربرد دیگر تولید خودکار موسیقی است که تولید آهنگ‌های جدید با استفاده از الگوریتم‌ها، موجب فراتر رفتن از مرزهای خلاقیت و آهنگسازی خواهد شد.

در تشخیص احساسات از طریق صدا نیز با تجزیه و تحلیل الگوهای صوتی برای ارزیابی احساسات، قابل استفاده در خدمات مشتری برای بهبود کیفیت تعامل مورد استفاده قرار خواهد گرفت.

پایش بیوآکوستیک برای حفاظت از حیات وحش نیز موردی دیگر است که در آن با استفاده از داده‌های صوتی برای نظارت بر جمعیت حیات وحش و تنوع زیستی، و ضرورت برای تحقیقات زیست محیطی و تلاش‌های حفاظتی تلاش می‌شود.

کاربرد دیگر ابزارهای ترجمه و یادگیری زبان هستند که در واقع ابزارهایی هستند که زبان گفتاری را به‌صورت بلادرنگ ترجمه می‌کنند و با تجزیه و تحلیل تلفظ و روان به یادگیری زبان کمک خواهند کرد.

نظارت و امنیت مبتنی بر صدا نیز از جمله کاربردهای دیگر است که در آن با استفاده از تشخیص صدا در سیستم‌های امنیتی، تشخیص ناهنجاری‌هایی مانند شکستن شیشه یا ورود غیرمنتظره را ثبت خواهند کرد.

۳-۶-۱-نتیجه‌گیری

این مطالعات موردی و کاربردها پتانسیل گسترده داده‌های صوتی را در یادگیری ماشین در بخش‌های مختلف نشان می‌دهد. از افزایش تجربیات کاربر در پلتفرم‌های دیجیتال گرفته تا کمک به ایمنی عمومی و مراقبت‌های بهداشتی، برنامه‌ها متنوع و تاثیرگذار هستند. همانطور که فناوری همچنان به پیشرفت خود ادامه می دهد، دامنه استفاده‌های نوآورانه از داده‌های صوتی در یادگیری ماشین احتمالاً حتی بیشتر گسترش خواهد یافت و راه‌حل‌های جدیدی برای مشکلات پیچیده ارائه می‌دهد.

۷-۱- نتیجه‌گیری و جمع‌بندی نهایی

ما ماهیت منحصر به فرد داده‌های صوتی، ویژگی‌های دیجیتالی آن و اهمیت آن در حوزه یادگیری ماشین را بررسی کرده‌ایم. داده‌های صوتی، با ساختار زمانی و پیچیده خود، چالش‌ها و فرصت‌هایی را برای تجزیه و تحلیل و کاربرد ارائه می‌دهند.

تکنیک‌های پردازش را نیز بررسی نمودیم که در واقع نقش حیاتی پیش‌پردازش، استخراج ویژگی، و افزایش داده‌ها در تهیه داده‌های صوتی برای مدل‌های یادگیری ماشین برجسته شد. تکنیک‌هایی مانند کاهش نویز، نرمال‌سازی و استخراج ویژگی هایی مانند MFCCها و طیف نگارها برای تجزیه و تحلیل موثر ضروری هستند.

همچنین مدل‌های مختلف یادگیری ماشین، از روش‌های سنتی مانند SVM و درخت تصمیم گرفته تا معماری‌های یادگیری عمیق پیشرفته مانند CNN، RNN و LSTM، برای کارهای مختلف پردازش صدا مناسب هستند را بررسی نمودیم. انتخاب مدل به نیازهای خاصی که مربوط به کار انجام شده می‌باشد، بستگی دارد.

در این بین ما چالش‌هایی را که در پردازش داده‌های صوتی با آن مواجه می‌شویم، مانند مدیریت مجموعه داده‌های بزرگ، برخورد با داده‌های متنوع و بدون ساختار، و غلبه بر خطر overfitting در مدل‌های پیچیده را بررسی کردیم. راه‌حل‌هایی مانند ذخیره‌سازی کارآمد داده، تکنیک‌های منظم‌سازی و انتخاب مدل مناسب مورد بحث قرار گرفتند.

کاربردهای متنوع داده‌های صوتی در یادگیری ماشین از طریق مطالعات موردی مختلف، از جمله دستیارهای صوتی، سیستم‌های توصیه موسیقی، مراقبت‌های بهداشتی و نظارت بر محیط زیست نشان داده شد. این برنامه‌ها تأثیر تحول‌آفرین تجزیه و تحلیل داده‌های صوتی را در بخش‌های مختلف نشان می‌دهند.

پتانسیل داده‌های صوتی در یادگیری ماشین بسیار زیاد است. با پیشرفت‌های مداوم در فناوری و روش‌ها، ظرفیت استخراج بینش‌های معنادار و ایجاد برنامه‌های کاربردی نوآورانه از داده‌های صوتی به سرعت در حال گسترش است. این تکامل باعث پیشرفت در زمینه هایی مانند مراقبت های بهداشتی، سرگرمی، نظارت بر محیط زیست و فراتر از آن می‌شود.

با وجود پیشرفت، محدودیت‌هایی وجود دارد، به ویژه در مورد کیفیت داده‌ها، نگرانی‌های حفظ حریم خصوصی، و نیاز به مجموعه داده‌های بزرگ و متنوع برای آموزش مدل‌های قوی وجود دارد. علاوه بر این، پیچیدگی داده‌های صوتی به منابع محاسباتی قابل توجهی نیاز دارد و نیاز به تحقیقات مداوم برای رسیدگی موثر به این چالش‌ها وجود دارد.

روندهای آینده در تجزیه و تحلیل داده‌های صوتی ممکن است شامل مدل‌های یادگیری عمیق پیچیده-تر، قابلیت‌های پردازش در زمان واقعی و ادغام تجزیه و تحلیل صوتی با انواع دیگر داده‌ها برای بینش غنی‌تر باشد. همگرایی تجزیه و تحلیل داده‌های صوتی با فناوری‌های نوظهور مانند واقعیت افزوده و اینترنت اشیا نیز امکانات هیجان انگیزی را ارائه می‌دهد.

حوزه داده‌های صوتی در یادگیری ماشین یکی از زمینه‌های رشد و نوآوری پویا است. همانطور که ما به توسعه روش‌های پیشرفته‌تر و کارآمدتر برای پردازش و تجزیه و تحلیل داده‌های صوتی ادامه می‌دهیم، احتمال استفاده از آن احتمالاً افزایش می‌یابد و راه‌حل‌های جدید و بهبود یافته‌ای را برای چالش‌های موجود و نوظهور ارائه می‌دهد. در نتیجه، اکتشاف داده‌های صوتی در یادگیری ماشین نشان‌دهنده یک حوزه مطالعه پر جنب و جوش و در حال تحول است که نویدبخش پیشرفت‌های مداوم در سال‌های آینده خواهد بود.

۲-توضیحی مختصر راجع به روش های پیاده سازی تسک ASR (تبدیل اتوماتیک صوت به متن)

۲-۱-مقدمه

ASRیک فناوری است که به رایانه‌ها اجازه می‌دهد زبان گفتاری را تشخیص داده و به متن تبدیل کنند. دو روش اصلی برای اجرای ASR وجود دارد که شامل روش‌های آماری سنتی و روش‌های جدیدتر end-to-end. هر دو روش دارای ویژگی‌های متمایزکننده‌ای از نظر نیازهای داده، منابع پردازش و دقت هستند.

۲-۲-روش‌های آماری

در یک و نیم دهه گذشته، تشخیص گفتار تحت سلطه رویکرد ترکیبی سنتی بوده است. بسیاری هنوز به دلیل فراوانی داده‌های تحقیق و آموزشی موجود در ساخت مدل‌های قوی به این روش تکیه می‌کنند.

سیستم‌های آماری ASR سنتی معمولاً مبتنی بر مدل‌های پنهان مارکوف (HMMs) همراه با مدل‌های مخلوط گاوسی (GMMs) هستند. این مدل‌ها برای نشان دادن ویژگی های آماری صداهای گفتاری استفاده می‌شوند. سیستم‌های ASR سنتی از چندین مؤلفه مجزا تشکیل شده‌اند، از جمله یک مدل صوتی، یک مدل زبان و یک مدل تلفظ. مدل آکوستیک برای تشخیص صداهای گفتار آموزش دیده است، مدل زبان توالی کلمات را پیش‌بینی می‌کند، و مدل تلفظ به نگاشت صداهای گفتار به کلمات کمک می‌کند.

مدل آکوستیک (AM) مسئول تشخیص الگوهای آکوستیک گفتار و پیش‌بینی صدا یا واجی است که در هر بخش متوالی بر اساس داده‌های تراز اجباری ادا می‌شود. AM معمولا دارای ساختار GMM یا HMM است.

مدل زبانی (LM) برای مدل‌سازی الگوهای آماری زبان طراحی شده است. می‌توان آن را آموزش داد تا بفهمد کدام عبارات و کلمات به احتمال زیاد با هم صحبت می‌شوند و به آن اجازه می‌دهد تا به طور دقیق احتمال هر کلمه‌ای را پس از مجموعه‌ای از کلمات فعلی پیش‌بینی کند.

مدل واژگانی توضیح می‌دهد که چگونه کلمات به صورت آوایی تلفظ می‌شوند. به طور معمول، یک مجموعه واج جداگانه برای هر زبان مورد نیاز است که توسط آوا شناسان مجرب طراحی شده است.

در حالی که هنوز به طور گسترده روش‌های ذکر شده مورد استفاده قرار می‌گیرد، رویکردهای ترکیبی سنتی برای تشخیص گفتار دارای چند اشکال عمده است. قابل توجه‌ترین میزان دقت پایین آن است. علاوه بر این، هر مدل نیاز به آموزش مستقل دارد که به زمان و کار بیش از حد نیاز دارد. به دلیل حجم قابل توجهی از کار انسانی که در به‌دست آوردن آن‌ها انجام می‌شود، به‌دست‌آوردن داده‌های تراز اجباری نیز

دشوار است. علاوه بر این، دانش تخصصی برای ساخت مجموعه‌های آوایی سفارشی برای افزایش دقت مدل‌ها ضروری است.

از لحاظ نیاز به داده نیز این سیستم‌ها اغلب به مقدار قابل توجهی از داده‌های برچسب‌دار نیاز دارند. داده‌ها باید به واحدهای آوایی تقسیم شوند و هر واحد باید برچسب‌گذاری شود تا سیستم مطابقت بین گفتار و متن را یاد بگیرد.

منابع پردازشی مورد نیاز نیز در این روش‌ها معمولاً در مقایسه با روش‌های end-to-end به توان محاسباتی کمتری نیاز دارند. پردازش شامل مراحل متمایزی مانند استخراج ویژگی، مدل‌سازی صوتی، مدل‌سازی زبان و رمزگشایی است.

از لحاظ دقت نیز در روش‌های سنتی می‌تواند در محیط‌های کنترل‌شده بالا باشد، اما اغلب با گفتار طبیعی، محاوره‌ای یا با نویز زیاد مشکل دارد. این مدل‌ها می‌توانند در برخورد با لهجه‌ها یا تغییرات گفتاری ضعیف عمل کنند.

۳-۲-روش انتها به انتها

رویکرد انتها به انتها در تشخیص گفتار شامل استفاده از شبکه‌های عصبی برای مدل‌سازی مستقیم داده‌های صوتی ورودی، به جای تکیه بر تکنیک‌های سنتی پردازش گفتار مانند استخراج ویژگی‌ها از سیگنال صوتی و سپس اعمال یک مدل جداگانه برای تشخیص است. این رویکرد اغلب به عنوان یک رویکرد انتها به انتها نامیده می‌شود زیرا یک شبکه عصبی واحد، کل فرآیند تشخیص گفتار را بدون نیاز به مراحل میانی انجام می‌دهد.

سیستم‌های یادگیری عمیق انتها به انتها برای تشخیص گفتار معمولاً دارای دو جزء اصلی هستند: یک شبکه رمزگذار که سیگنال صوتی خام را به یک نمایش سطح بالا تبدیل می‌کند و یک شبکه رمزگشا که رونویسی نهایی را ایجاد می‌کند. یکی از متداول‌ترین رویکردها طبقه‌بندی زمانی ارتباطی (CTC) است که به سیستم اجازه می‌دهد یاد بگیرد که ورودی را با خروجی تراز کند.

همانطور که اشاره نمودیم، سیستم‌های ASR انتها به انتها، که اغلب مبتنی بر یادگیری عمیق هستند، از شبکه‌های عصبی برای پردازش مستقیم گفتار به متن استفاده می‌کنند. این سیستم‌ها معمولاً از معماری-هایی مانند شبکه‌های عصبی کانولوشن (CNN)، شبکه های عصبی تکراری (RNN) یا اخیراً از مدل‌های ترانسفورماتور استفاده می‌کنند.

از لحاظ نیاز به داده نیز سیستم‌های انتها به انتها به مقدار زیادی داده نیاز دارند، اما برخلاف روش‌های سنتی، داده‌ها نیازی به برچسب‌گذاری یا بخش‌بندی دقیق ندارند. آن‌ها می‌توانند مستقیماً از گفتار مداوم بیاموزند و آن‌ها را با الگوهای گفتاری متنوع سازگارتر کند.

در بررسی منابع پردازشی نیز باید اشاره نمود که این روش‌ها به دلیل پیچیدگی مدل‌های شبکه عصبی از نظر محاسباتی فشرده هستند. آنها به منابع GPU قابل توجهی هم برای آموزش و هم برای استنباط نیاز دارند، مخصوصاً برای مدل‌های بزرگ مانند Transformers.

در مسئله دقت این روش‌ها نیز باید گفت که مدل‌های انتها به انتها اغلب به دقت بالاتری دست می‌یابند، به ویژه در شرایط گفتاری طبیعی و متنوع. این روش‌ها در برخورد با لهجه‌ها، گویش‌ها و محیط‌های با نویز بالا بهتر عمل می‌کنند. با این حال، عملکرد آن‌ها به شدت به کمیت و کیفیت داده‌های آموزشی وابسته است.

۴-۲-مقایسه روش‌ها

برای مقایسه این دو روش از لحاظ نیاز به تعداد داده باید گفت که روش‌های سنتی به داده‌های دقیق، تقسیم‌بندی‌شده و برچسب‌گذاری شده و کمتری نسبت به روش‌های انتهابه‌انتها نیاز دارند، در حالی که روش‌های انتها به انتها به حجم زیادی از داده‌های گفتاری پیوسته نیاز دارند.

در منابع پردازشی نیز روش‌های سنتی کم‌مصرف منابع هستند و برای سیستم‌هایی با قدرت محاسباتی محدود مناسب هستند. روش‌های انتها به انتها به منابع محاسباتی قابل توجهی، به‌ویژه پردازنده‌های گرافیکی نیاز دارند.

از لحاظ دقت نیز روش‌های انتها به انتها معمولاً در اکثر سناریوها، به ویژه در مدیریت الگوهای گفتاری متنوع و طبیعی، بهتر از روش‌های سنتی عمل می‌کنند، اما به داده‌ها و قدرت محاسباتی بیشتری نیاز دارند.

در نتیجه، انتخاب بین روش‌های آماری سنتی و روش‌های جدیدتر انتها به انتها در ASR بستگی به الزامات خاص مسئله، از جمله داده‌های موجود، منابع محاسباتی، و نیاز به دقت در شرایط گفتاری متنوع دارد.

۳-مفهوم و دلیل استفاده از fine tuning در آموزش شبکه‌های عصبی عمیق

۳-۱-مقدمه

Fine tuning یک مفهوم مهم در آموزش شبکه‌های عصبی عمیق است، به ویژه در زمینه وظایفی مانند تشخیص خودکار گفتار (ASR). این کار شامل گرفتن یک شبکه عصبی از پیش آموزش دیده و آموزش بیشتر (یا Fine tuning) آن بر روی یک مجموعه داده خاص و اغلب کوچکتر مربوط به یک مسئله خاص است. این تکنیک به چند دلیل مفید است که آن‌ها را به اختصار توضیح خواهیم داد.

۳-۲-Fine tuning

نقطه شروع موضوعی مهم است که Fine tuning با مدلی شروع می‌شود که قبلاً روی یک مجموعه داده بزرگ و کلی آموزش داده شده است. این مدل درک خوبی از ویژگی‌ها و الگوهای موجود در آن مجموعه داده گسترده‌تر ایجاد کرده است.

مدل از پیش آموزش داده شده و سپس بر روی یک مجموعه داده خاص‌تر آموزش داده می‌شود. در مورد ASR، این مجموعه داده‌ای از ضبط‌های گفتار و رونویسی‌های آن‌ها خواهد بود که اصطلاحا به این کار Specialization گفته می‌شود.

در طول Fine tuning، معمولاً تنها بخشی از لایه‌های مدل تنظیم می‌شوند، یا نرخ یادگیری بسیار پایین تنظیم می‌شود تا فقط تنظیمات جزئی در وزن‌ها انجام شود. این کار برای تخصصی کردن درک مدل به مجموعه داده جدید و در عین حال حفظ دانش عمومی که قبلاً کسب کرده بود انجام می‌شود.

دلایل متنوعی برای استفاده از Fine tuning وجود دارد که به برخی از آن‌ها اشاره خواهیم نمود.

استفاده از ویژگی‌های از پیش آموخته شده یکی از دلایل استفاده از Fine tuning می‌باشد. مدل‌های یادگیری عمیق، به‌ویژه در حوزه‌هایی مانند ASR، برای یادگیری مؤثر به مقادیر زیادی داده نیاز دارند. با استفاده از یک مدل از پیش آموزش دیده، از ویژگی‌ها و الگوهایی که قبلاً آموخته است استفاده می کنید، که می‌تواند پایه‌ای قوی برای مسئله ما باشد.

کارایی نیز موضوع دیگر است که از اهمیتی خاص برخوردار است. آموزش شبکه عصبی عمیق از ابتدا به منابع محاسباتی و زمان قابل توجهی نیاز دارد. تنظیم دقیق یک مدل از پیش آموزش دیده بسیار کارآمدتر است زیرا به توان محاسباتی و زمان آموزش کمتری نیاز دارد.

محدودیت داده‌ها نیز مشکلی دیگر است که برای غلبه بر آن از Fine tuning می‌توان استفاده نمود. اغلب، مجموعه داده خاص برای یک مسئله (مانند نوع خاصی از گفتار در ASR) کوچکتر از مجموعه داده‌های مورد استفاده برای آموزش مدل‌های بزرگ است. Fine tuning امکان آموزش موثر حتی با مجموعه داده‌های نسبتا کوچکتر را فراهم می‌کند.

بهبود عملکرد نیز از نتایج مثبت Fine tuning است. Fine tuning می‌تواند منجر به عملکرد بهتر در کارهای تخصصی شود. مدل از پیش آموزش دیده سطحی از درک عمومی را به ارمغان می‌آورد که وقتی با آموزش تخصصی ترکیب می‌شود، می‌تواند منجر به مدلی شود که در کارهای خاص بهتر از مدلی که از ابتدا آموزش دیده است، انجام شود.

این رویکرد امکان تطبیق یک مدل واحد و کلی را با وظایف مختلف خاص فراهم می‌کند که عملی‌تر و کارآمدتر از آموزش مدل‌های فردی برای هر کار است که موجب انطباق پذیری خواهد شد.

به طور خلاصه، Fine tuning یک تکنیک قدرتمند در یادگیری عمیق است که کارایی و اثربخشی مدل‌ها را به ویژه در کارهای تخصصی مانند ASR افزایش می‌دهد. از قدرت مدل‌های از پیش آموزش دیده استفاده می‌کند و آن‌ها را با نیازهای خاص تطبیق می‌دهد، در زمان و منابع صرفه جویی می‌کند و در عین حال اغلب به عملکرد برتر دست می‌یابد.