به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس یادگیری ماشین

گزارش اولیه

نام و نام خانوادگی	امیرحسین پورداود – امیرغرقابی – محمد ابوذری – مهدی سلیمانی زادگان
شماره دانشجویی	۸۱۰۱۰۱۱۲۰ – ۸۱۰۱۰۲۲۱۷ – ۸۱۰۱۰۱۰۸۷ – ۸۱۰۱۰۲۱۷۶
تاریخ ارسال گزارش	۱۴۰۲.۱۰.۰۳

فهرست

شکل‌ها

جدول‌ها

قسمت ۱. نحوه کار با داده ها

در الگوریتم‌های یادگیری ماشین، کار با داده‌های صوتی نیازمند مراحلی برای پیش‌پردازش و استخراج ویژگی‌ها است. در زیر توضیحی مختصر ارائه می‌شود:

پیش‌پردازش داده‌های صوتی: در این مرحله، ممکن است نیاز باشد داده‌های صوتی را پیش‌پردازش کنید. به عنوان مثال، می‌توانید از فیلترها برای حذف نویزهای موجود در سیگنال استفاده کنید یا داده‌ها را نرمال‌سازی کنید تا مقیاس واحدهای آنها یکسان شود.

استخراج ویژگی‌ها: در این مرحله، ویژگی‌های معنادار و مفید از داده‌های صوتی استخراج می‌شوند. این ویژگی‌ها ممکن است شامل مقادیر مربوط به طول، انرژی، فرکانس و زمان استفاده شده در سیگنال صوتی باشند. به عنوان مثال، می‌توانید از ویژگی‌های مانند طول سیگنال، میانگین فرکانس، طیف فرکانسی، ضرایب اندازه و غیره استفاده کنید.

تمیز کردن داده‌ها: در برخی موارد، ممکن است نیاز باشد داده‌های صوتی را تمیز کنید. این شامل حذف نویزهای غیرضروری، حذف قسمت‌های بی‌معنی یا ناخواسته از سیگنال صوتی و حذف اشکال دیگر است که ممکن است تأثیر منفی بر عملکرد الگوریتم‌های یادگیری ماشین داشته باشد.

نرمالایز کردن: ای روش در داده ها که باعث میشود داده های مسله از نظر یکسان نسبت به افراد مختلف و همچنین جلوگیری از افزایش ضرایب روش تخمین ما قابل استفاده باشند و قبل از سوار کردن داده ها در مسله لازم میباشد.

در زیر هر کدام از روش ها بصورت جزیی تری توضیح داده شده اند :

۱- Cleaning Data

در این قسمت سعی میشود با استفاده از روش های آماری و محاسباتی سعی کنیم داده ها صوتی را پاکسازی و اماده ی تحلیل و استخراج ویژگی نماییم

روش های مرسوم در زیر ذکر شده اند :

۱. حذف نویز: اگر داده‌های صوتی شامل نویزهای غیرضروری هستند، می‌توانید از روش‌های حذف نویز مانند فیلترهای کاهش نویز (noise reduction filters) استفاده کنید. به عنوان مثال، فیلترهای کاهش نویز مانند فیلتر میانگین متحرک (Moving Average filter) یا فیلتر میانه (Median filter) را می‌توانید استفاده کنید.

۲. حذف تداخل و بازتاب: برای حذف تداخل و بازتاب صداها می‌توان از روش‌هایی مانند استفاده از فیلترهای آنتی‌تداخل (Anti-Interference Filters)، فیلترهای آنتی‌بازتاب (Anti-Reverberation Filters)، یا الگوریتم‌های آنتی‌بازتاب (Anti-Reverberation Algorithms) استفاده کرد. این روش‌ها برای کاهش تداخل‌ها و بازتاب‌های غیرمطلوب در سیگنال صوتی استفاده می‌شوند.

۳. تمیز کردن با استفاده از یادگیری ماشینی: روش‌های یادگیری ماشینی مانند شبکه‌های عصبی مصنوعی و الگوریتم‌های یادگیری ماشینی می‌توانند برای تشخیص و حذف نویزها و خرابی‌ها در داده‌های صوتی استفاده شوند. با استفاده از مجموعه‌های آموزشی که داده‌های صوتی تمیز و نویزدار را شامل می‌شوند، مدل‌های یادگیری ماشینی می‌توانند به صورت خودکار نویزها و خرابی‌ها را تشخیص دادهو از سیگنال صوتی تمیز شده بازسازی نمایند.

۴. تقسیم برازش (Resampling): در برخی موارد، ممکن است نیاز به تغییر فرکانس نمونه‌برداری داده‌های صوتی باشد. با تقسیم برازش داده‌ها، می‌توانید نمونه‌برداری با فرکانس مطلوب را انجام دهید.

۵. استفاده از فیلترهای ترکیبی: با استفاده از فیلترهای ترکیبی مثل فیلترهای پایین‌گذر (Low-pass filters) و بالاگذر (High-pass filters)، می‌توانید برخی اجزای غیرضروری را از سیگنال صوتی حذف کنید. این فیلترها می‌توانند در حذف نویزهای بالا و پایین فرکانس، سیگنال‌های تداخلی یا اجزای غیرضروری کمک کنند.

۲- Normalization

همچنین برای نرمال سازی داده های صوتی راه های مختلفی مطرح شده است که در زیر هریک را به اختصار توضیح میدهیم:

۱. مقیاس‌بندی مین-مکس (Min-Max Scaling): در این روش، مقادیر داده‌های صوتی را بین یک محدوده‌ی مشخص (معمولاً ۰ تا ۱ یا -۱ تا ۱) نرمال می‌کنید. برای هر نمونه، فرمول زیر را می‌توانید استفاده کنید:

$$X_{\text{نرمال}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

۲. استانداردسازی زنجیره‌ای Z-score Normalization در این روش، مقادیر داده‌های صوتی را بر اساس میانگین و انحراف معیار آنها استانداردسازی می‌کنید. برای هر نمونه، فرمول زیر را می‌توانید استفاده کنید:

$$X_{\text{نرمال}} = \frac{X - \mu}{\sigma}$$

۳. لگاریتم نرمال: در برخی موارد، ممکن است داده‌های صوتی شامل دامنه پهنی باشند که باعث مشکل در آموزش مدل‌ها می‌شود. با استفاده از لگاریتم نرمال، می‌توانید داده‌های صوتی را بازنمایی کنید و دامنه پهن آنها را کاهش دهید. که در اینجا اپسیلون یک عدد بسیار کوچک در حد e-10 میباشد که برای صفر نشدن لگاریتم استفاده میشود.

$$X_{\text{نرمال}} = \log(X + \epsilon)$$

۴. نرمال‌سازی مکمل واحدهای خطی L2 Normalization در این روش، بردارهای داده‌های صوتی را بر اساس مجموع مربعات عناصر آنها نرمال می‌کنید. برای هر بردار داده X، فرمول زیر را می‌توانید استفاده کنید:

$$X_{\text{نرمال}} = \frac{X}{\|X\|_2}$$

۳- استخراج ویژگی ها در داده ای صوتی :

استخراج ویژگی‌ها در داده‌های صوتی یک مرحله مهم در پردازش سیگنال صوتی و تجزیه و تحلیل آنهاست. این ویژگی‌ها معمولاً برای استفاده در مدل‌های یادگیری ماشینی، تشخیص سیگنال صوتی، تشخیص سخنرانی، تشخیص حالت و احساس و دیگر برنامه‌های مرتبط با صوت استفاده می‌شوند. در زیر چند روش رایج برای استخراج ویژگی‌ها در داده‌های صوتی را بررسی می‌کنیم:

۱. تبدیل فوریه: تبدیل فوریه (Fourier Transform) یک روش قدرتمند برای تبدیل سیگنال صوتی از دامنه زمان به دامنه فرکانس است. با استفاده از تبدیل فوریه، می‌توانید اطلاعات فرکانسی سیگنال صوتی را استخراج کنید. تبدیل فوریه معمولاً با استفاده از الگوریتم‌هایی مانند تبدیل فوریه سریع (FFT) انجام می‌شود.

۲. ملودی‌گرام MFCCs (Mel-frequency Cepstral Coefficients - MFCCs): یکی از روش‌های معروف برای استخراج ویژگی‌ها در داده‌های صوتی است. با استفاده از MFCCs، اطلاعات مربوط به ملودی و فرکانس سیگنال صوتی استخراج می‌شود. این ویژگی‌ها عموماً شامل استخراج طیف ملودیک، لگاریتم طیف ملودیک، و استفاده از تبدیل کوشی و ضرایب سفت‌سازی هستند.

۳. طیف‌های متعدد (Spectrograms): یک طیف‌نما (Spectrogram) نمایشی است که بر اساس تبدیل فوریه زمانی (Short-time Fourier Transform) از سیگنال صوتی استخراج می‌شود. طیف‌نما نشان می‌دهد که چه میزان از هر فرکانس در طول زمان برای سیگنال صوتی استفاده شده است. با استفاده از طیف‌های متعدد، می‌توانید اطلاعات مربوط به فرکانس و زمان سیگنال صوتی را دریافت کنید.

۴. ویژگی‌های مبتنی بر زمان (Time-based Features): ویژگی‌های مبتنی بر زمان مانند میانگین مقدار مطلق (Average Amplitude), انرژی صوتی (Energy), کوتاه مدت تغییر لحظه‌ای (Short-term Temporal Variation) و غیره، معمولاً برای تشخیص سیگنال صوتی و تشخیص سخنرانی استفاده می‌شوند. این ویژگی‌ها اطلاعات مربوط به شکل موج صوتی در طول زمان را ارائه می‌دهند.

۵. ویژگی‌های مبتنی بر فرکانس (Frequency-based Features): ویژگی‌های مبتنی بر فرکانس مانند باندهای موازی فرکانسی (Mel Filter Banks) و طیفی از توان فرکانسی (Power Spectrum) استفاده می‌شوند. این ویژگی‌ها اطلاعات مربوط به توان و شدت فرکانس‌های موجود در سیگنال صوتی را نشان می‌دهند.

۶. ویژگی‌های مبتنی بر زمان-فرکانس (Time-Frequency-based Features): ویژگی‌های مبتنی بر زمان-فرکانس مانند تبدیل مومانتوم زمان-فرکانس (Time-Frequency Moment Transform) و تبدیل ویولت (Wavelet Transform) استفاده می‌شوند. این ویژگی‌ها اطلاعات مربوط به تغییرات زمانی و فرکانسی سیگنال صوتی را در طول زمان استخراج می‌کنند.

۷. ویژگی‌های مبتنی بر حالت و احساس (Emotion-based Features): در برخی برنامه‌ها مانند تشخیص حالت و احساس در صدا، ویژگی‌های مبتنی بر حالت و احساس مانند تغییرات صوتی (Pitch Variation)، شدت صوتی (Intensity) و فرکانس بندی احساسی (Emotional Mel-frequency Cepstral Coefficients) استفاده می‌شوند.

قسمت ۲. روش های پیاده سازی تسک ASR

نام ASR به تکنولوژی ای گفته می شود که گفتار را به متن تبدیل می کند. به طور کلی دو دسته روش برای انجام این کار وجود دارد.

روش های مختلفی سنتی مانند روش های آماری (statistical) و روش های مدرن مانند انتها به انتها (End-to-End) وجود دارد. در این بخش به این دو روش می پردازیم.

۱. روش سنتی – آماری

رویکردهای آماری سنتی اغل بر مدل های احتمالاتی و متد های آماری برای مدل سازی روابط بین ویژگی های صوتی و واحد های زبانی متکی هستند. برخی از روش های مرسوم و کلیدی عبارتند از:

- مدل های مارکف مخفی

۱- این مدل ها آماری هستند که دنباله ای از حالت های قابل مشاهده (ویژگی های صوتی) را از طریق یک سری حالت های پنهان (واحدهای آوایی یا واحد های شبه کلمه ای) نشان می دهد. انتقال بین این لایه ها به صورت احتمالاتی مدل می شوند.

- GMMها

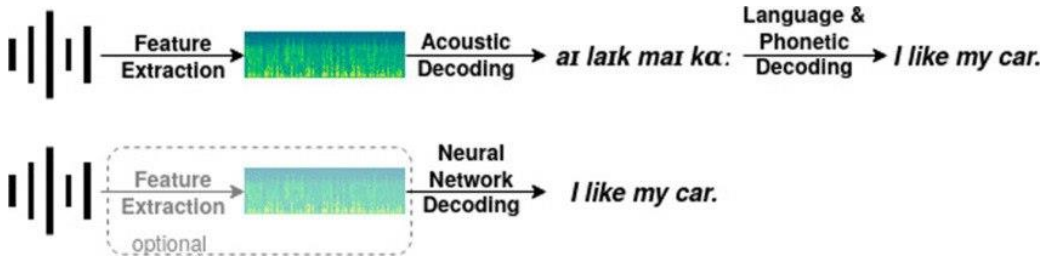
۲- این مدل ها، توزیع احتمال را به صورت مجموع چند توزیع گوسی مدل می کند. در ASR، از این مدل ها برای مدل سازی توزیع ویژکی های صوتی مرتبط با واحدهای آوایی مختلف استفاده می شود.

- Vector Quantization

۳- VQ یک تکنیک کوانتزیزه کردن است که برای نمایش بردارهایی با مقادیر پیوسته با مجموعه ای محدود از نمادهای گسسته به کار می رود. در ASR از این روش برای کوانتیزه کردن بردارهای ویژگی صوتی به مجموعه کوچکتری از بردار ها استفاده می شود. از این روش برای خوشه بندی feature space و رویکردهای مبتنی بر codebook برای کاهش ابعاد بردارهای ویژگی صوتی استفاده می شود.

۲. روش مدرن – انتها به انتها

رویکرد انتها به انتها از روش های یادگیری عمیق (معمولا شبکه های عصبی) برای مپ کردن ویژی های صوتی به متن و به صورت مستقیم استفاده می کند. در زیر تصویر از این pipeline قابل مشاهده است.



بطور کلی این روش از متدهای یادگیری عمیق بهره میگیرد تا سیگنال صوتی خام را مستقیما به متن تبدیل کند.

مقایسه:

- از نظر پردازش:

روش های انتها به انتها معماری ساده تری دارند و همچنین تعداد مراحل کمتری را نیز نیاز دارند اما محاسبات پیچیده ای دارند. زیرا ترین کردن مدل شبکه های عصبی ممکن است از لحاظ پردازی سنگین و وقت گیر باشند به خصوص اگر از مدل های بزرگ استفاده شود.

از طرفی روش های سنتی مانند GMM ها به علت قدیمی تر بودن، الگوریتم های بهنیه تری دارند و همچنین در مرحله decoding بازدهی بالاتری دارند. اما از طرف دیگر تعداد مراحل بیشتری برای تولید متن خروجی دارند.

- از نظر دقت:

روش های مدرن انتها به انتها به علت استفاده از شبکه های عمیق می تواند دقت بیشتری را فراهم کنند اما از طرفی این اتفاق زمانی می افتد که به اندازه کافی داده ورودی موجود باشد. اگر حجم داده زیاد بوده و محدودیت پردازشی وجود نداشته باشد، روش انتها به انتها می تواند دقت بسیار خوبی را فراهم کند.

- از نظر داده:

روش های انتها به انتها مبتنی بر شبکه عصبی عمیق است کع ورودی را به خروجی به صورت مستقیم مپ می کند. این روش تکیه زیادی بر حجم داده دارد اما نکته قابل توجه این است که این داده ها می توانند بسیار متنوع باشد.

اما روش های مبتنی بر آمار، نیازمند حجم دیتای برچسب خورده هستند اما به نسبت با داده های کمتری می توانند عملکرد خود را داشته باشند همچنین حساسیت زیادی نسبت به کیفیت داده ها دارند.به طور کلی انتخاب میان این روش بستگی به تنوع داده، کیفیت داده و محدودیت توان پردازشی دارد.

قسمت ۳ – fine tuning در شبکه های عصبی

Fine tuning به معنای تنظیم مجدد یک مدل پیش‌آموزش داده شده بر روی داده‌های خاص پروژه است. این کار با هدف بهبود عملکرد مدل دریک تسک خاص پروژه انجام می‌شود.

ممکن است داده‌های مربوط به پروژه ما متفاوت باشند و نیاز به تنظیم مجدد وزن‌ها داشته باشیم تا مدل بهتر با داده‌های ما هماهنگ شود. این اقدام به ما امکان می‌دهد تا عملکرد بهتری در تشخیص گفتار داشته باشیم.

شکل ۱ **Fine tuning** در آموزش شبکه های عصبی

Fine tuning (تنظیم دقیق)

در آموزش شبکه عصبی به فرآیند گرفتن یک مدل از پیش آموزش دیده و آموزش بیشتر آن بر روی یک مجموعه داده جدید کوچکتر یا یک مجموعه داده با توزیع متفاوت، معمولاً برای انطباق مدل با یک وظیفه یا دامنه خاص اشاره دارد.

دلایل استفاده از Fine tuning در شبکه های عصبی :

۱. انتقال یادگیری:

استفاده از Fine tuning به شبکه هایی که بسیار پیچیده و با تعداد زیادی لایه هستند امکان انتقال دانش و یادگیری از مجموعه داده بزرگ پیش آموزش دیده شده را فراهم می کند . این امر می تواند به کاهش نیاز به داده برای آموزش جدید کمک کند.

۲. بهبود عملکرد:

با استفاده از Fine tuning ، می توان بر روی خصوصیات مرتبط با مجموعه داده جدید تمرکز کرده و عملکرد شبکه را بهبود داد.

۳. تطبیق دادن به داده های جدید :

ممکن است ویژگی های موجود در داده های جدید با دوگانه ی پیش تر مورد آموزش شبکه اندکی تفاوت داشته باشد. در این موارد، اجرای Fine tuning امکان بهبود عملکرد شبکه را فراهم می کند . به طور کلی، Fine tuning با کمک داده های موجود در محیطی خارج از داده های پیش آموزش، امکان بهبود و تطبیق شبکه عصبی را با موارد مورد نیاز فراهم می کند.