به نام آفریننده ای که صدا ها را برای ارتباط و زیبایی آفرید

عنوان تمرین

دسته بندی صداها با استفاده از CNN

عنوان درس

یادگیری ماشین

استاد

دکتر الهام قصرالدشتی

دستیاران آموزشی

مهرداد قصابی

مریم صفوی

گردآورنده

سید حسین حسینی

تابستان ۱۴۰۴

دانشکده مهندسی کامپیوتر

دانشگاه اصفهان

بخش اول: مقدمه و چارچوب مسئله

1.1. تعریف مسئله و اهمیت آن

در دنیای مدرن، حجم داده‌های صوتی تولید شده توسط سیستم‌های مختلف (از مکالمات روزمره و محتوای چندرسانه‌ای گرفته تا داده‌های علمی و صنعتی) به صورت نمایی در حال افزایش است. توانایی تحلیل و درک خودکار این داده‌ها، فرصت‌های بی‌شماری را در حوزه‌های مختلف ایجاد می‌کند. این پروژه بر یکی از مسائل بنیادین در این حوزه، یعنی طبقه‌بندی صدا(Audio Classification) ، متمرکز است.

هدف اصلی پروژه، ساخت یک مدل یادگیری عمیق است که بتواند با دریافت یک قطعه صوتی کوتاه، آن را به یکی از چهار کلاس از پیش تعریف‌شده نسبت دهد:

1. صدای پس‌زمینه(Background) : شامل صداهای محیطی یک جنگل بارانی است که به عنوان خط پایه (Baseline) و نویز طبیعی در نظر گرفته می‌شود.
2. صدای اره‌برقی(Chainsaw) : صدایی مکانیکی و با الگوی فرکانسی مشخص که می‌تواند نشانگر فعالیت‌های انسانی مانند قطع درختان باشد.
3. صدای موتور(Engine) : صدای یک موتور احتراقی که الگوی هارمونیک و پایداری دارد.
4. صدای طوفان(Storm) : شامل صداهای طبیعی مانند باد شدید و رعد و برق که دارای ویژگی‌های پهن‌باند و انرژی بالا در فرکانس‌های پایین است.

اهمیت این مسئله فراتر از یک تمرین آکادمیک است. سیستمی که بتواند این صداها را تشخیص دهد، می‌تواند در کاربردهای زیر به کار گرفته شود:

- پایش محیط زیست :تشخیص خودکار صدای اره‌برقی برای مقابله با جنگل‌زدایی غیرقانونی.
- سیستم‌های هشدار :شناسایی صدای طوفان برای صدور هشدارهای آب‌وهوایی.
- امنیت و نظارت :تشخیص صداهای غیرعادی (مانند صدای موتور در مناطق ممنوعه) در سیستم‌های نظارتی.

1.2. استراتژی حل مسئله: رویکرد بینایی ماشین برای تحلیل صدا

سیگنال‌های صوتی، داده‌هایی پویا و یک‌بعدی هستند. تحلیل مستقیم این سیگنال‌ها با مدل‌های سنتی می‌تواند پیچیده و نیازمند دانش تخصصی در حوزه پردازش سیگنال باشد. از سوی دیگر، شبکه‌های عصبی کانولوشنی (CNN) در سال‌های اخیر توانایی خارق‌العاده‌ای در استخراج ویژگی‌های سلسله‌مراتبی از داده‌های دوبعدی (تصاویر) از خود نشان داده‌اند.

استراتژی کلیدی این پروژه، تبدیل هوشمندانه مسئله تحلیل صدا به یک مسئله بینایی ماشین است. این کار از طریق تبدیل سیگنال‌های صوتی به طیف‌نگاره (Spectrogram) انجام می‌شود.

- طیف‌نگاره چیست؟ یک نمایش بصری از محتوای فرکانسی یک سیگنال صوتی در طول زمان. در این نمودار، محور افقی زمان، محور عمودی فرکانس و شدت رنگ (یا روشنایی) در هر نقطه، دامنه (انرژی) آن فرکانس در آن لحظه زمانی را نشان می‌دهد.
- چرا این تبدیل قدرتمند است؟ زیرا الگوهای فرکانس-زمانی که مشخصه هر صدا هستند (مانند هارمونیک‌های یک نت موسیقی، نویز پهن‌باند یک صدای صنعتی یا الگوهای متغیر صدای انسان) به صورت الگوهای بصری قابل تشخیص در طیف‌نگاره ظاهر می‌شوند. به این ترتیب، یک CNN می‌تواند با یادگیری این الگوهای بصری، صداها را با دقت بالایی طبقه‌بندی کند. ما در واقع به مدل یاد می‌دهیم که به جای "شنیدن"، "ببیند."

بخش دوم: آماده‌سازی و پیش‌پردازش داده‌ها

2.1. جمع‌آوری و سازماندهی مجموعه داده

مرحله ابتدایی هر پروژه یادگیری ماشین، آماده‌سازی داده است. در این پروژه:

1. دانلود داده‌ها :چهار مجموعه فایل صوتی با فرمت wav.از یک منبع آنلاین (گوگل درایو) دانلود می‌شوند. هر مجموعه مربوط به یکی از کلاس‌های تعریف‌شده است.

2. ایجاد ساختار پوشه‌ای :یک ساختار دایرکتوری منطقی و استاندارد (dataset/audio/[class_name])ایجاد می‌شود. این کار نه تنها برای نظم‌دهی ضروری است، بلکه فرآیند خودکارسازی مراحل بعدی را بسیار ساده‌تر می‌کند. این ساختار به ما اجازه می‌دهد تا با یک حلقه for، عملیات تولید طیف‌نگاره و برچسب‌گذاری را برای تمام کلاس‌ها اجرا کنیم.

2.2. مهندسی ویژگی: هنر تبدیل صدا به تصویر معنادار

این مرحله، حیاتی‌ترین بخش پیش‌پردازش است. ما با استفاده از کتابخانه Librosa، هر فایل صوتی را به یک تصویر طیف‌نگاره تبدیل می‌کنیم. انتخاب‌های فنی در این مرحله تأثیر مستقیمی بر کیفیت ویژگی‌های ورودی به مدل دارد.

- بارگذاری سیگنال صوتی :هر فایل wav.با تابع librosa.loadبه یک آرایه عددی (سری زمانی) تبدیل می‌شود. این آرایه، دامنه سیگنال را در نقاط زمانی مختلف نشان می‌دهد.
- ایجاد طیف‌نگاره مِل(Mel Spectrogram) :
  - انتخاب مقیاس مِل :به جای استفاده از مقیاس فرکانسی خطی(Hz) ، از مقیاس مل استفاده شده است. این انتخاب بر اساس یافته‌های روان‌شناسی آکوستیک انجام شده و نحوه درک فرکانس توسط گوش انسان را شبیه‌سازی می‌کند. انسان‌ها به تفاوت‌های بین فرکانس‌های پایین (مثلاً ۱۰۰ هرتز و ۲۰۰ هرتز) بسیار حساس‌تر از تفاوت‌های مشابه در فرکانس‌های بالا (مثلاً ۱۰۰۰۰ هرتز و ۱۰۰۱۰ هرتز) هستند. مقیاس مل این حساسیت غیرخطی را مدل می‌کند و در نتیجه، ویژگی‌های صوتی مهم‌تر را برای مدل ما برجسته‌تر می‌سازد.
  - تبدیل به دسی‌بل(dB) : توان طیف‌نگاره به مقیاس لگاریتمی دسی‌بل تبدیل می‌شود. این کار دامنه دینامیکی گسترده سیگنال را فشرده کرده و باعث می‌شود تفاوت‌های جزئی در انرژی فرکانس‌ها، که ممکن است برای چشم یا مدل قابل تشخیص نباشند، بهتر دیده شوند. این نیز با نحوه درک بلندی صدا توسط انسان مطابقت دارد.

- ذخیره‌سازی تصویر :تصویر طیف‌نگاره تولید شده با حذف تمام حاشیه‌ها، محورها و برچسب‌های اضافی ذخیره می‌شود. این کار تضمین می‌کند که ورودی CNN یک ماتریس خالص از مقادیر شدت فرکانس باشد و هیچ اطلاعات نامرتبطی (نویز) وارد فرآیند یادگیری نشود.

2.3. آماده‌سازی نهایی داده‌های تصویری برای مدل

پس از تولید تصاویر طیف‌نگاره، آن‌ها باید برای ورود به مدل Keras آماده شوند:

1. بارگذاری و تغییر اندازه :تمام تصاویر با تابع image.load_imgاز Keras بارگذاری شده و به اندازه ثابت 224x224پیکسل تبدیل می‌شوند. این استانداردسازی ابعاد برای ورودی CNN الزامی است.
2. تبدیل به آرایه(Tensor) : هر تصویر به یک آرایه NumPy سه‌بعدی (224x224x3) تبدیل می‌شود که ابعاد آن به ترتیب ارتفاع، عرض و کانال‌های رنگی (RGB) هستند.
3. تقسیم‌بندی داده‌ها بهینه :داده‌ها با نسبت ۲۰-۸۰ به مجموعه‌های آموزش و آزمون تقسیم می‌شوند. استفاده از stratify=yدر تابع train_test_splitتضمین می‌کند که توزیع کلاس‌ها در هر دو مجموعه یکسان باشد. این یک روش استاندارد برای جلوگیری از سوگیری نمونه‌گیری (Sampling Bias) و کسب نتایج ارزیابی قابل اعتماد است.
4. نرمال‌سازی :مقادیر پیکسل‌ها از بازه [0, 255]به بازه [0, 1]نگاشت می‌شوند. این کار به الگوریتم بهینه‌سازی (مانندAdam) کمک می‌کند تا سریع‌تر و پایدارتر به نقطه بهینه همگرا شود.
5. رمزگذاری وان-هات :برچسب‌های عددی کلاس‌ها (0, 1, 2, 3) به بردارهای باینری تبدیل می‌شوند. این فرمت برای تابع هزینه categorical_crossentropy، که در طبقه‌بندی چندکلاسه استفاده می‌شود، ضروری است و به مدل اجازه می‌دهد خطای خود را برای هر کلاس به طور مستقل محاسبه کند.

بخش سوم: معماری، آموزش و ارزیابی مدل

3.1. معماری شبکه عصبی کانولوشنی (CNN)

معماری طراحی‌شده یک CNN کلاسیک و مؤثر است که از دو بخش اصلی تشکیل شده است:

1. بخش استخراج ویژگی (Feature Extraction) : شامل چهار بلوک متوالی از لایه‌های Conv2D و MaxPooling2D.
   - لایه‌های کانولوشن (Conv2D) : این لایه‌ها با فیلترهای کوچک (در اینجا 3*3)، الگوهای محلی را در طیف‌نگاره شناسایی می‌کنند. لایه‌های اولیه الگوهای ساده مانند خطوط افقی، عمودی یا قطری را یاد می‌گیرند. در لایه‌های عمیق‌تر، این ویژگی‌های ساده با هم ترکیب شده و الگوهای پیچیده‌تر و معنادارتری مانند هارمونیک‌های پایدار (مشخصه موتور) یا نویزهای پهن‌باند (مشخصه طوفان) را شکل می‌دهند.
   - لایه‌های تجمعی (MaxPooling2D) : پس از هر لایه کانولوشن، یک لایه تجمعی قرار دارد که ابعاد نقشه ویژگی را نصف می‌کند. این کار دو مزیت کلیدی دارد: کاهش چشمگیر حجم محاسبات و پارامترها، و ایجاد ناوردایی نسبت به جابجایی (Translation Invariance) ، یعنی مدل می‌تواند یک الگو را حتی اگر کمی در تصویر جابجا شده باشد، تشخیص دهد.
2. بخش طبقه‌بندی (Classification) :
   - لایه Flatten: خروجی چندبعدی بخش کانولوشنی را به یک بردار یک‌بعدی تبدیل می‌کند.
   - لایه‌های Dense: این لایه‌ها یک شبکه عصبی کاملاً متصل را تشکیل می‌دهند که وظیفه طبقه‌بندی نهایی را بر عهده دارند. لایه پنهان با ۱۰۲۴ نورون، ترکیب‌های غیرخطی از ویژگی‌های استخراج‌شده را یاد می‌گیرد.
   - لایه خروجی: با ۴ نورون (به تعداد کلاس‌ها) و تابع فعال‌سازی softmax، یک توزیع احتمال روی کلاس‌ها تولید می‌کند.

3.2. فرآیند آموزش مدل

- پیکربندی(Compilation) : مدل با بهینه‌ساز Adam(یک الگوریتم بهینه‌سازی کارآمد و محبوب)، تابع هزینه categorical_crossentropy(مناسب برای طبقه‌بندی چندکلاسه) و معیار ارزیابی accuracy (دقت) پیکربندی می‌شود.
- آموزش(Fitting) : مدل برای ۱۲ دور (epoch) روی داده‌های آموزشی آموزش داده می‌شود. انتخاب ۱۲ دور یک نقطه شروع مناسب است که به مدل فرصت کافی برای یادگیری می‌دهد بدون اینکه لزوماً منجر به بیش‌برازش شدید شود.

3.3. تحلیل نتایج و ارزیابی عملکرد

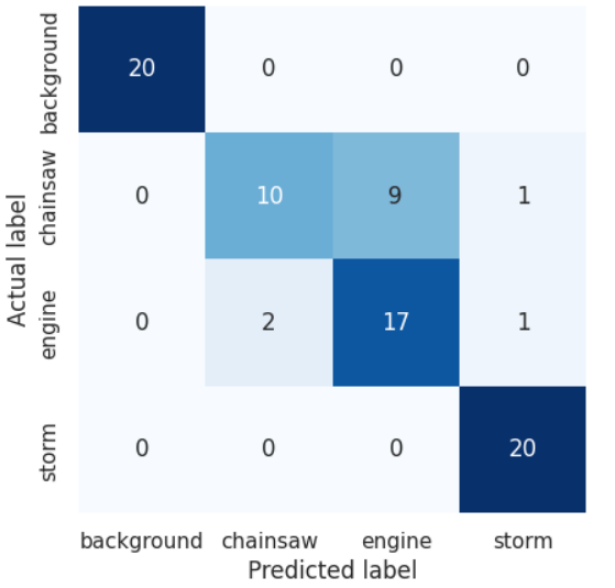
۱ .نمودار دقت آموزش و اعتبارسنجی:

این نمودار، ابزار اصلی ما برای نظارت بر فرآیند یادگیری است. در این پروژه، مشاهده می‌شود که هر دو منحنی دقت آموزش و اعتبارسنجی با روندی تقریباً موازی و نزدیک به هم افزایش می‌یابند. این یک نتیجه ایده‌آل است و نشان می‌دهد که:

- مدل در حال یادگیری الگوهای واقعی و قابل تعمیم است.
- پدیده بیش‌برازش(Overfitting) ، که در آن مدل داده‌های آموزشی را "حفظ" می‌کند ولی روی داده‌های جدید عملکرد ضعیفی دارد، به طور جدی رخ نداده است.

۲ .ماتریس درهم‌ریختگی(Confusion Matrix) :

این ماتریس یک تحلیل عمیق و دقیق از عملکرد مدل برای هر کلاس ارائه می‌دهد:



- اعداد روی قطر اصلی نشان‌دهنده پیش‌بینی‌های صحیح هستند. مقادیر بالای این اعداد (مثلاً ۲۰ برای پس زمینه و ۲۰ برای طوفان) نشان‌دهنده عملکرد عالی مدل در تشخیص این کلاس‌هاست.
- اعداد خارج از قطر اصلی نشان‌دهنده خطاها هستند. تحلیل دقیق‌تر این خطاها، بینش‌های مهمی به ما می‌دهد:
  - خطای اصلی :بیشترین خطا (۹مورد) در طبقه‌بندی اشتباه صدای "اره‌برقی" به عنوان "موتور" رخ داده است. این امر از نظر آکوستیکی قابل توجیه است؛ هر دو صدا دارای انرژی قابل توجهی در فرکانس‌های پایین و میانی هستند و الگوی نویزی دارند. تفکیک آن‌ها نیازمند یادگیری الگوهای ظریف‌تری است.
  - خطای کمتر :موارد بسیار کمی از اشتباه گرفتن "طوفان" یا "پس‌زمینه" با کلاس‌های دیگر وجود دارد. این نشان می‌دهد که الگوهای بصری تولید شده توسط این صداهای مکانیکی (خطوط هارمونیک پایدار و نویزهای پهن‌باند مشخص) برای CNN بسیار متمایز و قابل تشخیص بوده‌اند.

دقت نهایی :دقت کلی مدل روی داده‌های تست حدود ٪۸۴ است که برای یک مدل ساخته‌شده از پایه و با داده‌های نسبتاً محدود، یک نتیجه بسیار خوب و امیدوارکننده محسوب می‌شود.

```
Epoch 12/12
10/10 ━━━━━━━━━━━━━━━━━━━━ 1s 87ms/step - accuracy: 0.8466 - loss: 0.3226 - val_accuracy: 0.8375 - val_loss: 0.4340
```

بخش ششم: ۵ منبع تکمیلی و پیشرفته

این منابع برای گسترش دانش و کاوش در مسیرهای بهبود پروژه بسیار مفید هستند و به جنبه‌های نظری و عملی پیشرفته‌تری اشاره دارند.

1. .مقاله: شبکه‌های عصبی کانولوشنی-بازگشتی برای طبقه‌بندی صدای محیطی (Convolutional Recurrent Neural Networks for Environmental Sound Classification)

- مرجع : Adavanne, S., Pertilä, P., & Virtanen, T. (2017). *Sound event detection using spatial and harmonic features and convolutional recurrent neural network*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

2. .مقاله: یادگیری انتقال برای طبقه‌بندی صدای محیطی (Transfer Learning for Environmental Sound Classification)

- مرجع : Guzhov, A., et al. (2021). *ESResNet: Environmental Sound Classification Based on Visual-Domain Models*.
- چرا این منبع کلیدی است؟

این پروژه یک CNN را از ابتدا (from scratch) آموزش می‌دهد. اما یک رویکرد بسیار قدرتمند و رایج در بینایی ماشین، یادگیری انتقال (Transfer Learning) است. این مقاله (و مقالات مشابه) نشان می‌دهد که چگونه می‌توان از مدل‌های CNN قدرتمندی مانند ResNet که روی مجموعه داده عظیم ImageNet(شامل میلیون‌ها تصویر از ۱۰۰۰ کلاس مختلف) آموزش دیده‌اند، برای طبقه‌بندی طیف‌نگاره‌های صوتی استفاده کرد.

3 .کتابخانه PyTorch و فریم‌ورک Torchaudio

- مراجع:
  - Paszke, A., et al. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Advances in Neural Information Processing Systems.
  - وب‌سایت Torchaudio: https://pytorch.org/audio/stable/index.html

- چرا این منبع کلیدی است؟

در حالی که این پروژه از TensorFlow/Keras استفاده کرده، PyTorch یک فریم‌ورک یادگیری عمیق بسیار محبوب دیگر است که به خصوص در محیط‌های تحقیقاتی به دلیل انعطاف‌پذیری بالا، طرفداران زیادی دارد .Torchaudio نیز کتابخانه تخصصی PyTorch برای پردازش صداست که ابزارهای مشابه Librosa (مانند تولید طیف‌نگاره مِل و MFCC) را به صورت کاملاً یکپارچه با تنسورهای PyTorch و با قابلیت اجرا روی GPU ارائه می‌دهد. آشنایی با این اکوسیستم، یک جایگزین قدرتمند برای ابزارهای استفاده شده در پروژه فراهم می کند و به توسعه‌دهنده اجازه می‌دهد تا از جدیدترین معماری‌ها و تکنیک‌های منتشر شده در جامعه تحقیقاتی PyTorch به راحتی استفاده کند.

4 .مقاله: افزایش داده برای سیگنال‌های صوتی (Data Augmentation for Audio Signals)

- مرجع: Salamon, J., & Bello, J. P. (2017). *Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification*. IEEE Signal Processing Letters.

- چرا این منبع کلیدی است؟

یکی از بزرگترین چالش‌ها در یادگیری عمیق، کمبود داده‌های آموزشی است. این مقاله به طور خاص به تکنیک‌های افزایش داده (Data Augmentation) برای سیگنال‌های صوتی می‌پردازد

5 .مجموعه دادهAudioSet

- مرجع : Gemmeke, J. F., et al. (2017). *Audio Set: An ontology and human-labeled dataset for audio events*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
    - وب‌سایت مجموعه داده : https://research.google.com/audioset/
- چرا این منبع کلیدی است؟

این مجموعه داده که توسط گوگل منتشر شده، یکی از بزرگترین و جامع‌ترین مجموعه داده‌های صوتی در جهان است AudioSet .شامل بیش از ۲ میلیون قطعه صوتی ۱۰ ثانیه‌ای از ویدیوهای یوتیوب است که با بیش از ۶۰۰ برچسب مختلف (از صدای حیوانات و موسیقی گرفته تا صداهای انسانی و محیطی) حاشیه‌نویسی شده‌اند. این منبع به دو دلیل کلیدی است:

1. محک‌زنی :(Benchmarking) می‌توان از آن برای ارزیابی و مقایسه عملکرد مدل خود با پیشرفته‌ترین مدل‌های جهانی استفاده کرد.
2. یادگیری انتقال :می‌توان یک مدل را روی این مجموعه داده عظیم پیش‌آموزش داد و سپس آن را برای یک کار خاص (مانند پروژه ما) تنظیم دقیق کرد. این رویکرد، مشابه یادگیری انتقال از ImageNet در بینایی ماشین، می‌تواند به نتایج بسیار قدرتمندی منجر شود.

موفق باشید