عنوان پروژه (کارشناسی ارشد)

تشخیص احساسات گفتاری با استفاده از شبکه عصبی کانولوشن

پروژه/ پایان‌نامه/ رساله برای دریافت درجه کارشناسی/ ارشد/ دکتری

در رشته مهندسی / گرایش

نام دانشجو:

استاد (اساتید) راهنما:

جنابت آقای

استاد (اساتید) مشاور:

مهر ماه 1400

تصویر صورتجلسه دفاع از پروژه/ پایان‌نامه/ رساله

تأییدیه هیئت داوران جلسه دفاع از پروژه/ پایان‌نامه/ رساله

دانشکده مهندسی

نام دانشجو:

عنوان پروژه/ پایان‌نامه/ رساله:

تاریخ دفاع:

رشته مهندسی گرایش:

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
1	استاد راهنمای اول				
2	استاد راهنمای دوم				
3	استاد مشاور اول				
4	استاد مشاور دوم				
5	استاد مدعو خارجی اول				
6	استاد مدعو خارجی دوم				
7	استاد مدعو داخلی اول				
8	استاد مدعو داخلی دوم				

مجوز بهره‌برداری از پروژه/ پایان‌نامه/ رساله

بهره‌برداری از این پروژه/ پایان‌نامه/ رساله در چهارچوب مقررات کتابخانه و باتوجه به محدودیتی که توسط استاد راهنما به‌شرح زیر تعیین می‌شود، بلامانع است:

☐ بهره‌برداری از این پروژه/ پایان‌نامه/ رساله برای همگان بلامانع است.

☐ بهره‌برداری از این پروژه/ پایان‌نامه/ رساله با اخذ مجوز از استاد راهنما، بلامانع است.

☐ بهره‌برداری از این پروژه/ پایان‌نامه/ رساله تا تاریخ ..... ممنوع است.

نام استاد (اساتید) راهنما:

تاریخ:

امضا:

تقدیم به:

تشکر و قدردانی:

با تشکر از زحمات استاد گرانقدر جناب

چکیده

موضوع پایان نامه پیش‌بینی و تشخیص احساسات گفتاری با استفاده از شبکه عصبی عمیق کانولوشنال است . تشخیص احساسات گفتاری (SER) یکی از چالش برانگیزترین کارها در حوزه تجزیه و تحلیل سیگنال گفتار است ، این یک چالش حوزه تحقیقاتی است که سعی می کند احساسات را از سیگنال های گفتاری استنباط کند. دقت در تشخیص احساساتی اعم از شادی، غم، ترس، عصبانیت و آرامش بین زنان و مردان بویژه در زمانی که سیگنالهای صدا دارای سرو صدا و نویز پس زمینه است و یا احساسات در گفتار بدرستی و با شفافیت منعکس نمی شود، کار پیچیده ای می باشد. یک شبکه عصبی پیچیده (کانولوشنال) تحت نظارت عمیق برای طبقه بندی هر احساس در صداهای جمع آوری شده آموزش دیده است. با بهره گیری از کتابخانه تنسورفلو Tensorflow و کراسKeras لایه های متعدد در محیط Google Colab برای مدل شکل گرفته است. مجموعه با تعداد 25 Actor متشکل از صداهای زنان و مردان با احساسات مختلف جمع آوری شده از بانک Ravdess مورد آزمایش قرار می گیرد. در آخر بخش مهم اجرای مدل شبکه عصبی پیچیده (کانولوشنال) و تست مدل است که با اندازه گیری متریک دقت میتوان به میزان دقت در مدل واقف شد.

واژه‌های کلیدی: شبکه عصبی عمیق کانولوشنال ، تشخیص احساسات، Tensorflow، Keras، احساسات گفتاری.

فهرست مطالب

فهرست علائم اختصاری

شبکه عصبی کانولوشنال Convolutional Neural Network ..... *CNN*

فصل 1:

مقدمه

1-1- معرفی

تشخیص احساسات مبتنی بر گفتار دارای مزایای کاربردی عملی فراوانی است. در واقع تشخیص احساسات فرایند شناسایی احساسات انسانی است. دقت افراد در تشخیص احساسات دیگران بسیار متفاوت است. استفاده از این فناوری برای کمک به افراد در تشخیص احساسات یک حوزه تحقیقاتی نسبتاً نوپا است. به طور کلی ، این فناوری در صورتی که از چند روش در زمینه استفاده کند ، بهترین عملکرد را دارد. تا به امروز ، بیشترین کار بر روی تشخیص خودکار حالات چهره از طریق ویدئو ، عبارات گفتاری از طریق صدا ، عبارات نوشتاری از متن اندازه گیری می شود.

تشخیص احساسات یکی از مهمترین استراتژی های بازاریابی در دنیای امروز است. شما می توانید موارد مختلف را برای یک فرد به طور خاص متناسب با علاقه خود شخصی سازی کنید.به همین دلیل ، بنای این تحقیقاین است که بتوان احساسات افراد را فقط با صدای آنها تشخیص داد که اجازه می دهد بسیاری از برنامه های مرتبط با هوش مصنوعی مدیریت شود. برخی از مثالها می تواند شامل مراکز تماس برای پخش موسیقی هنگام عصبانی شدن فرد در تماس باشد. یکی دیگر می تواند یک ماشین هوشمند باشد که هنگام عصبانیت یا ترس سرعت خود را کاهش می دهد.در نتیجه این نوع برنامه دارای پتانسیل زیادی در جهان است که می تواند به نفع شرکت ها و حتی ایمنی مصرف کنندگان باشد.

رابط کاربری صوتی (VUI) تعامل گفتاری انسان با رایانه ها را ممکن می سازد ، از تشخیص گفتار برای درک دستورات گفتاری و پاسخ به سوالات استفاده می کند و معمولاً برای پخش پاسخ از متن به گفتار استفاده می کند. دستگاه فرمان صوتی (VCD) دستگاهی است که با رابط کاربری صوتی کنترل می شود.

SER اگرچه چندان محبوب نیست ، SER در این سالها حوزه های زیادی را وارد این عرصه کرده است ، از جمله:

حوزه پزشکی: در دنیای پزشکی از راه دور که بیماران در بسترهای تلفن همراه مورد ارزیابی قرار می گیرند ، توانایی یک متخصص پزشکی در تشخیص احساس بیمار در واقع می تواند در روند بهبود مفید باشد.

خدمات به مشتریان: در مرکز تماس از مکالمه برای تجزیه و تحلیل مطالعه رفتاری همراهان تماس با مشتریان استفاده می شود که به بهبود کیفیت خدمات کمک می کند.

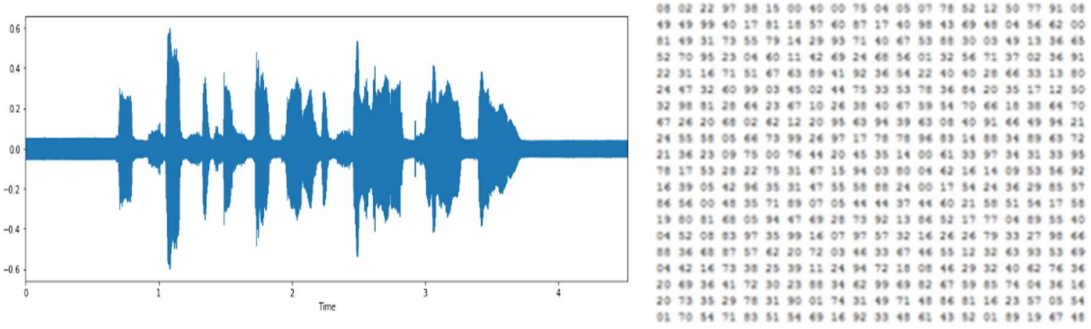
سیستم های توصیه گر: توصیه می شود محصولات را بر اساس احساسات مشتریان نسبت به آن محصول به مشتریان توصیه شود.

یکی از مولفه های مهم این فرآیند تشخیص گفتار با استفاده از شبکه عصبی کانولوشنال است ، که کار زیادی را طلب می کند و نیاز به تخصص در این حوزه دارد. به تازگی ، تکنیک های تشخیص گفتاری رایانه ای و یادگیری ماشینی با موفقیت برای بررسی خودکار احساسات انسانی بر روی صداهای ضبط شده استفاده شده است. تعداد زیادی از مقالات و تحقیقات اخیر در تشخیص احساسات و توصیف حالت های انسانی به وضوح نشان دهنده علاقه روزافزون به این حیطه تحقیقاتی است.

2-1- شبکه عصبی کانولوشن

شبکه عصبی پیچشی یا شبکه عصبی کانولوشن، از مهم‌ترین نوآوری‌ها در حوزه‌ی بینایی کامپیوتر به حساب می‌آیند. لغت شبکه عصبی در سال ۲۰۱۲، معروفیت فراوانی کسب کرد؛ در این سال الکس چریشفسکی، با استفاده از شبکه عصبی توانست برنده جایزه ImageNet (المپیک سالیانه بینایی کامپیوتر) شود. ریشفسکی توانست خطای دسته‌بندی (classification) را از ۲۶ درصد به ۱۵ درصد کاهش دهد. این کاهش در آن زمان بسیار چشم‌گیر بود. از آن زمان، شرکت‌های متعددی از یادگیری عمیق به عنوان هسته اصلی محصولات خود استفاده کرده‌اند. گوگل، فیسبوک، آمازون، اینستاگرام و پینترست از شبکه عصبی استفاده می‌کند تا تصاویر را به صورت خودکار تگ‌گذاری نماید؛ با این حال بیش‌ترین استفاده‌ی شبکه عصبی در پردازش تصویر است. در این بخش از معرفی متد تحقیق، به چیستی شبکه عصبی کانولوشن پرداخته خواهد شد و اینکه چگونه از آن در دسته‌بندی تصاویر استفاده می‌شود.

دسته‌بندی تصاویر یا اصوات در واقع پروسه ای است که در آن تعدادی تصویر یا صوت را از ورودی می‌گیریم و در خروجی، کلاس آن‌ها (نوع احساسات صوت، مثلا در اینجا کلاس به احساساتی از قبیل شادی، غم، ترس، عصبانیت و آرامش بین زنان و مردان اطلاق می شود) یا درصد احتمال تعلق به هر کلاس را مشخص می‌کنیم. انجام چنین عملی، یعنی تشخیص و نام‌گذاری (labeling) اصوات، کلاسیفیکیشن نامیده می شود.



شکل (1-1) تصویر سمت چپ تصویر واقعی است که چشمان ما در هنگام پخش فایل صوتی در پلیرها می بیند و تصویر سمت راست چیزی است که کامپیوتر برای آنالیز می بیند.

فصل 2:

پیشینهٔ تحقیق

2-1- مقدمه

2-2- متد شبکه عصبی کانولوشن CNN [7]

ورودی صوتی و تصویری و خروجی کلاسه شده در یک شبکه عصبی کانولوشن

وقتی یک کامپیوتر صوت و یا تصویری را به عنوان ورودی دریافت می‌کند، آن را به صورت آرایه‌ای از اعداد می‌بیند. تعداد آرایه‌ها به سیگنال صوتی (بر اساس داده های متوالی و سری زمانی) و یا سایز تصویر (بر اساس پیکسل) بستگی دارد. برای مثال اگر یک صوت از پیش ضبط شده فرمت wav را به کامپیوتر دهیم، آرایه جانشین آن دارای m*n خانه خواهد بود. هر کدام از خانه‌ها یا المنت ها نیز عددی بین 1- تا 1 را می‌گیرند. این عدد شدت آوا را نشان می‌دهد. این اعداد هر چند در وهله اول بی‌معنی به نظر می‌رسند، اما در پردازش صدا با استفاده از الگوریتم ها، ابزار مناسب، همین اعداد هستند. ایده اصلی آن است که به کامپیوتر یا مدل پردازش صدا، آرایه‌ای از اعداد، شبیه آن چه توضیح داده شد، داده و کامپیوتر نیز در خروجی چنین چیزی را مشخص می‌کند: این صدا با احتمال ۸۰ درصد دارای احساسات خوشحالی است، و یا با احتمال ۱۵ درصد دارای احساسات ترس است.

شیوه عملکرد شبکه عصبی کانولوشن چیست؟

متد حل مسئله در شبکه عصبی کانولوشن دریافت صدا و یافتن ویژگی‌های منحصر به فرد آوا با استفاده از کتابخانه librosa مانند قدرت ، صدای پیکربندی و مجرای صوتی از سیگنال گفتار است. سپس تشخیص دهد در صدا، احساسی موجود است یا نه. مثلا برای تشخیصش احساسات، ابتدا به مولفه های جزئی‌تر آن مانند قدرت، طول موج توجه می کند و ضمن تطبیق با الگوهای موجود در داده های آموزشیTrain Set ، در می یابد که در چه احساسی در صدا وجود دارد. در نتیجه مدل برای درک و تشخیص احساسات در صداهای پیچیده‌ای مثل صدا همراه با نویز پس زمینه ، ابتدا ویژگی‌های (feature) ساده‌تر آن صدا مانند قدرت و تغییر طول موج را تشخیص می‌دهد. در یک شبکه عصبی، لایه‌های متعددی وجود دارند؛ در هر یک از این لایه‌ها، ویژگی‌های خاصی تشخیص داده می‌شوند و در نهایت، در لایه‌ی آخر، صدا به طور کامل شناسایی می‌شود. روندی که توضیح داده شد، فرایند کلی نحوه کار یک شبکه عصبی کانولوشن بود؛ حال به جزئیات بیش‌تری پرداخته می شود.

ارتباط شبکه عصبی پیچشی با بیولوژی

در این قسمت مفاهیم پایه‌ای‌تر مورد بررسی قرار می گیرد. عبارت شبکه عصبی کانولوشن قرابت زیادی با زیست شناسی و نوروساینس دارد. ساختار شبکه عصبی پیچشی (CNN) در حقیقت از قشر بینایی مغز الهام گرفته شده است. در سال ۱۹۶۲، دو دانشمند با نام‌های هابل و ویزل، آزمایش جالبی انجام دادند. آن‌ها نشان دادند که با دیدن لبه‌ها با اشکال مختلف، سلول‌های خاصی در قشر بینایی مغز تحریک می‌شوند. برای مثال با دیدن خطوط افقی، سلول‌های خاصی تحریک می‌شوند و با دیدن خطوط عمود بر هم سلول‌های متفاوتی حساسیت نشان می‌دهند. هابل و ویزل دریافتند که این سلول‌ها به شکل ستونی و خیلی منظم در کنار همدیگر قرار گرفته‌اند و حاصل همکاری آن‌ها با هم این است که انسانها می‌توانند ادراک تصویری خوبی از محیط پیرامون داشته باشند. اساس کار شبکه عصبی کانولوشن نیز مانند قشر بینایی مغز است. در حقیقت در یک CNN، لایه‌های مختلفی وجود دارند که هر یک لایه مخصوص شناسایی موارد خاصی است. در نهایت نیز خروجی مدل ادراک تصویری کامل است.

ساختار شبکه عصبی پیچشی

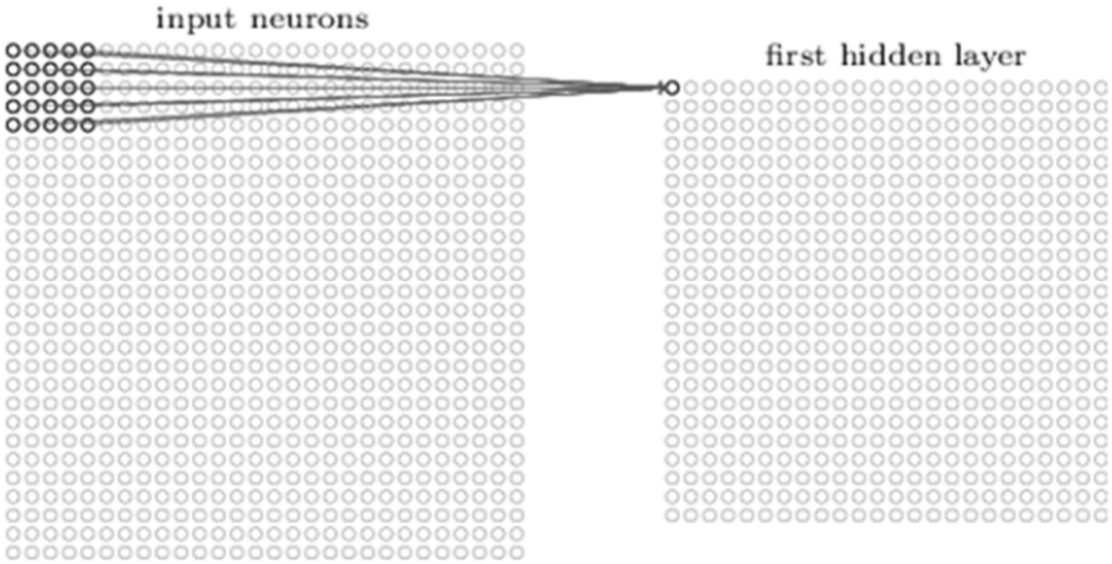
همان‌طور که اشاره شد، در یک شبکه عصبی پیچشی، کامپیوتر یک تصویر را به عنوان ورودی می‌گیرد؛ سپس این تصویر وارد یک شبکه‌ی پیچیده با چندین لایه‌ی پیچشی و غیر خطی می‌شود. در هر یک از این لایه‌ها، عملیات‌هایی انجام می‌شود و در انتها بر روی خروجی، یک کلاس یا درصد وقوع چند کلاس مختلف نشان داده می‌شود. قسمت سخت ماجرا، لایه‌های میانی و نحوه عملکرد آن‌هاست. در ادامه به بررسی مهم‌ترین لایه‌ها پرداخته می شود.

لایه اول در شبکه عصبی کانولوشن

لایه اول در یک شبکه عصبی پیچشی، همیشه یک لایه‌ی کانولوشنال است. همان‌طور که قبلا اشاره کردیم، ورودی این لایه یک آرایه از اعداد است. لایه اول در شبکه عصبی مانند یک چراغ قوه کار می‌کند. در یک اتاق تاریک، چراغ قوه‌ای را تصور کنید که بر گوشه‌ی بالا و سمت چپ تصویر انداخته می شود و محدوده‌ای از تصویر روشن می نمایاند و آن قسمت دیده می شود. سپس چراغ قوه بر روی قسمت‌های دیگر تصویر تابانده می شود تا کم‌کم کل تصویر را روشن نمایاند. همین روند دریادگیری ماشین، رخ می دهد (گرچه این روند در تشخیص احساسات در صدا نیز رخ می دهد و هر پارت از صدا آرام آرم با استفاده از فیلترینگ، پرداطش می شود).

در شبکه عصبی کانولوشن، به این چراغ قوه، فیلتر (filter) (یا نورون یا کرنل) می‌گوییم. آن قسمتی از تصویر یا صوت که چراغ قوه به آن نور می‌تاباند، محدود پذیرش (receptive field) نام دارد. لازم به ذکر است، فیلترها نیز خود آرایه‌هایی از اعداد هستند. به اعداد موجود در فیلتر، وزن (weight) یا پارامتر

(parameter) گفته می‌شود. لازم به ذکر است که عمق این فیلتر باید با عمق صوت و یا تصویر برابر باشد. فیلتر در هر نگاه، یک قسمت از صوت و یا تصویر را می‌بیند. سپس بر روی صوت و یا تصویر حرکت می‌کند تا قسمت‌های دیگر را هم اسکن کند. به این حرکت فیلتر بر روی تصویر، پیچیدن (convolve) گفته می شود. همین طور که فیلتر از تصویر عبور می‌کند، اعداد موجود در فیلتر با آرایه عددی پیکسل‌های صوت و یا تصویر ضرب می‌شود. در نهایت نیز تمام حاصل ضرب‌ها با یکدیگر جمع می‌شوند و به یک عدد می‌رسیم.

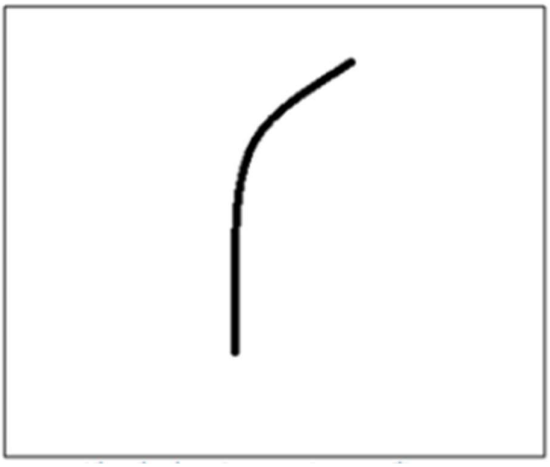


شکل (1-2) صوت و یا تصویر فیلتر چرخشی 5*5 در محدوده داده ورودی و تولید یک نقشه فعال ساز

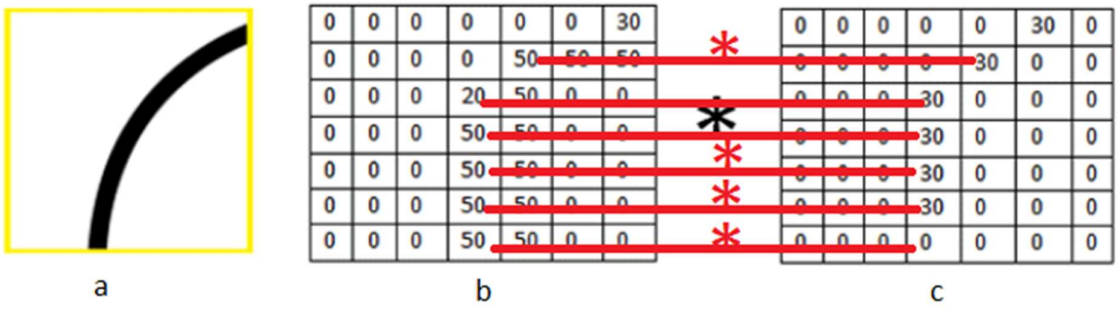
لایه اول در شبکه عصبی کانولوشن / کاربردی

هر یک از فیلترهایی را که در قسمت قبلی به آن‌ها اشاره شد، می‌تواند به عنوان یک شناساگر ویژگی (feature identifier) در نظر گرفته شود. منظور از ویژگی (feature) در این جا، چیزهایی مانند قدرت صدا، طول موج است. فرض بر این است که فیلتر اول، یک فیلتر با ابعاد m*n و یک شناساگر قدرت صدا است. این فیلتر در حقیقت یک ماتریس عددی مانند صوت و یا تصویر زیر است که درایه‌های این ماتریس در محل‌هایی که قدرت صدا در آن وجود دارد، مقادیر عددی بالاتری دارند. حال این فیلتر را بر روی قسمتی از صوت و یا تصویر مد نظرمان قرار می‌دهیم. پس از آن مانند شکل زیر، درایه به درایه اعداد موجود در خانه‌ها را با هم ضرب و حاصل ضرب‌ها را با یکدیگر جمع خواهد شد.

0	0	0	0	0	30	0
0	0	0	0	30	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0



شکل (2-2) تصویر سمت چپ مقادیر درایه های قدرت صوت سمت راست را در درایه های یک ماتریس به نمایش در آورده و در قدرت صوت ، شاهد افزایش از 0 به 30 می باشیم.



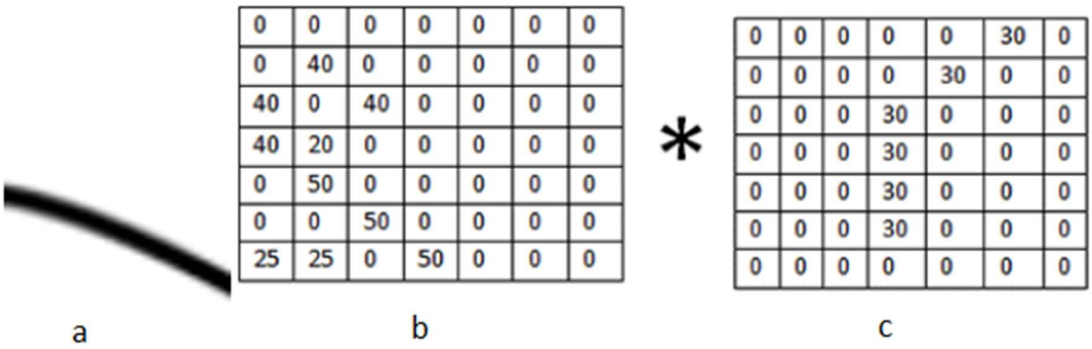
شکل (2-3) تصویر a نشاندهنده قدرت در صدا می باشد، تصویر b محدود پذیرش (receptive field) صدای اصلی، و تصویر c فیلتر یا نورون یا کرنل

$$Filter = (50 * 0) + (50 * 30) + (50 * 30) + (50 * 30) + (20 * 30) + (50 * 30) = 6600$$

همانطور که مشاهده می‌شود، حاصل به‌دست‌آمده، یک عدد بزرگ است. بزرگ بودن این عدد نشانگر آن است که در این ناحیه یک قدرت صدا مانند این فیلتر وجود دارد.

در تصویر زیر، حاصل ضرب عدد کوچکی می‌شود؛ علت آن است که فیلتر با صدای ورودی تطابق ندارد. هدف یافتن یک نقشه فعال‌سازی است؛ قسمت بالا و سمت چپ این نقشه فعال‌سازی، مقدار ۶۶۰۰ را خواهد داشت. این عدد بزرگ نشان‌دهنده‌ی آن است که در ناحیه‌ی خاصی از صدا، با احتمال زیاد یک

قدرت وجود دارد. در این جا تنها از یک فیلتر استفاده شده است. برای آن که اطلاعات بیش‌تری از صدا استخراج شود، نیاز است تا از فیلترهای بیش‌تری استفاده شود؛ استفاده از فیلترهای بیش‌تر یعنی ابعاد بالاتر.



شکل (4-2) تصویر a نشاندهنده قدرت در صدا می باشد، تصویر b محدود پذیرش (receptive field) در صدای اصلی، و تصویر c فیلتر یا نورون یا کرنل

$$Filter\ Result = 0$$

لایه‌های عمیق‌تر شبکه عصبی کانولوشن

در یک شبکه عصبی، علاوه بر لایه‌ی توضیح داده شده، لایه‌های دیگری نیز وجود دارند. این لایه‌ها وظایف و عملکردهای گوناگونی دارند. به طور کلی، لایه‌های داخلی، مسئول نگهداری و حفظ ابعاد و امور غیرخطی هستند. آخرین لایه در شبکه عصبی کانولوشن نیز از اهمیت خاصی برخوردار است.

لایه آخر در شبکه عصبی پیچشی

در لایه آخر یک شبکه عصبی کانولوشن، خروجی سایر لایه‌ها، به عنوان ورودی دریافت می‌شود. خروجی لایه آخر هم یک بردار N بعدی است. N تعداد کلاس‌های موجود است. به عنوان مثال اگر شبکه ایی مانند آنچه در تشخیص احساسات گفتاری ها بر روی صدا ها داریم، احساساتی اعم از شادی، غم، ترس، عصبانیت و آرامش، در نتیجه تعداد کلاس‌ها پنج تاست؛ چون 5 نوع کلاس، شادی، غم، ترس، عصبانیت و آرامش وجود دارد. در بردار N بعدی، هر مولفه، احتمال وقوع یک کلاس را نشان می‌دهد. کاری که لایه‌ی آخر یک شبکه عصبی کانولوشن می‌کند آن است که به ویژگی‌های لایه‌های سطح بالا نگاه می‌کند و میزان مطابقت این ویژگی‌ها را با هر کلاس مقایسه می‌کند؛ هر چه این مطابقت بیش‌تر باشد، احتمال وقوع آن کلاس، بالاتر معرفی می‌شود.

نحوه عملکرد شبکه عصبی کانولوشن چیست؟

مدل کانولوشنال طی یک فرایند آموزش (training) می‌تواند مقادیر مناسب را به فیلترها تخصیص دهد.این فرایند backpropagation نام دارد. در ابتدای کار، اعداد موجود در ماتریس فیلتر، رندم و تصادفی هستند. به مرور زمان و با آموزش صداهای مختلف به مدل، اعداد موجود در فیلتر تصحیح می‌شوند تا به یک عملکرد قابل قبول برسند.

تست شبکه عصبی CNN

پس از آن که مدل نهایی و آمده شد، وقت تست کردن فرا می‌رسد. برای تست مدل از تعدادی صدا که محتویات احساسات آن مشخص است، استفاده می‌شود. صدا را به ورودی مدل می‌دهیم تا خروجی را به ما نشان دهد؛ سپس خروجی را بررسی می‌کنیم تا ببینیم درست عمل شده است یا نه.

لایه های کانولوشنال

در معماری شبکه عصبی پیچشی سنتی ، لایه های دیگری نیز وجود دارند که بین این لایه های متقاطع پراکنده شده اند. در یک مفهوم کلی ، لایه ها، توابع غیر خطی و کنترل کننده ابعاد هستند که به بهبود یکپارچگی مدل و کنترل برازش بیش از حد overfittingکمک می کند. برازش بیش از حد یا overfitting زمانی اتفاق می افتد که مدل بتواند بر اساس داده های موجود در مجموعه آموزشی طبقه بندی یا پیش بینی کند ، اما در طبقه بندی داده هایی که بر روی آنها آموزش ندیده است ، خوب عمل نمی کند. بنابراین اساساً ، مدل از داده های موجود در آموزش بیش از حد برخوردار است .معماری کلاسیک CNN شبیه این خواهد بود.

```
Input -> Conv -> ReLU -> Conv -> ReLU -> Pool -> ReLU -> Conv -> ReLU -> Pool ->Fully Connected
```

فیلترهایی که در لایه کانولوشن اول برای تشخیص حاشیه و خطوط مرزی طراحی شده اند، مورد بررسی قرار گرفت. آنها ویژگی های سطح پایین مانند قدرت صدا و طول موج ها را تشخیص می دهند. همانطور که تصور می شود ، برای پیش بینی اینکه یک صدا چه احساسی دارد، ما به شبکه نیاز داریم تا بتوانیم ویژگی های سطح بالاتری مانند اجزا و مولفه های اصلی صدا مانند نوع احساسات را تشخیص دهیم. خروجی شبکه بعد از اولین لایه conv با حجم m*n خواهد بود (با فرض اینکه از فیلتر m * n استفاده شود). هنگامی که از یک لایه دیگر متقاطع عبور شود، خروجی اولین لایه تبدیل به ورودی لایه دوم کانولوشن تبدیل می شود. در مورد لایه اول، ورودی فقط صدای اصلی بود. با این حال ، هنگامی که در مورد لایه دوم متقاطع صحبت می شود ، ورودی نقشه (های) فعالسازی است که از لایه اول حاصل می شود. بنابراین هر لایه ورودی اساساً مکان هایی را در صدای اصلی توصیف می کند که در آن مشخصه های سطح پایین ظاهر می شوند. اکنون

هنگامی که مجموعه ای از فیلترها روی آن اعمال می شود (و از لایه دوم جابجایی عبور داده می شود) ، خروجی فعال سازی هایی است که ویژگی های سطح بالاتری را نشان می دهند.

انواع این ویژگی ها می تواند قدرت صدا (بم . یا زیر بودن) یا طول موج صدا مانند (کوتاه و بلند) باشد. همانطور که از شبکه عبور می کنید و لایه های کانولوشن بیشتری را طی می کنید ، نقشه های فعال سازی دریافت می شوند که ویژگی های پیچیده تر و پیچیده تری را نشان می دهد. با عمیق تر شدن در شبکه ، فیلترها دارای یک میدان پذیرش بزرگتر و بزرگتر می شوند ، به این معنی که آنها می توانند اطلاعات را از یک منطقه بزرگتر از حجم ورودی اصلی را در نظر بگیرند. این است که آنها به منطقه بزرگتری از فضای داریه های صدا پاسخ می دهند.

لایه کاملاً متصل اساساً یک حجم ورودی می گیرد (خروجی کنولوشن یا ReLU یا لایه pool قبل از آن) و یک بردار ابعادی N را خروجی می دهد که N تعداد کلاس هایی است که برنامه باید از بین آنها انتخاب کند. برای مثال ، اگر برنامه طبقه بندی قطعه صدا از حیث احساسات باشد، N=5 خواهد بود زیرا 5 کلاس احساسات اعم از شادی، غم، ترس، عصبانیت و آرامش وجود دارد. هر عدد در این بردار ابعادی N نشان دهنده احتمال یک کلاس خاص است. نحوه عملکرد این لایه کاملاً متصل این است که به خروجی لایه قبلی (که باید نقشه های فعال سازی ویژگی های سطح بالا را نشان دهد) نگاه می کند و مشخص می کند که کدام ویژگی ها بیشتر با یک کلاس خاص مرتبط هستند، مثلا قطعه صدای حاوی قدرت صدای بالا است که ویژگی عصبانیت را متمایز می نمایاند. به عنوان مثال ، اگر برنامه پیش بینی کند که برخی از صداها عصبانی هستند ، در نقشه های فعال سازی که نشان دهنده ویژگی های سطح بالا مانند قدرت صدا و غیره که نمادهای عصبانیت بودن است ، مقادیر بالایی خواهد داشت.

به طور مشابه ، اگر برنامه پیش بینی کند که برخی از صداها با آرامش هستند ، در نقشه های فعال سازی که نشان دهنده ویژگی های سطح بالا مانند آرام بودن قدرت صوت و آوا و غیره که نمادهای آرامش است ، مقادیر بالایی خواهد داشت. اساساً ، یک لایه FC به ویژگیهای سطح بالا که بیشترین ارتباط را با یک کلاس خاص دارد و وزنهای خاصی دارد نگاه می کند ، به طوری که وقت محاسبه خروجی ها بین وزنها و لایه قبلی ، احتمالات صحیح برای کلاسهای مختلف را بدست می آورید.

در حال حاضر ، این یکی از جنبه های شبکه های عصبی این است که چگونه فیلترهای اولین لایه کانولوشن می دانند که به دنبال قدرت صدا و یا طول موج ها هستند؟ چگونه لایه کاملاً متصلFC می داند که باید به

کدام نقشه های فعال سازی نگاه کرد؟ چگونه فیلترهای هر لایه می دانند چه مقادیری باید داشته باشند؟ روشی که کامپیوتر می تواند مقادیر فیلتر (یا وزن) خود را تنظیم کند از طریق یک فرآیند آموزشی به نام انتشار مجدد عقبگردBack Propagation است.

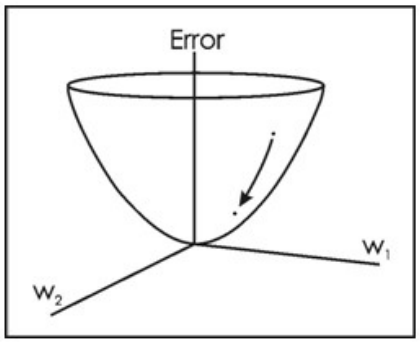
وزنها یا مقادیر فیلتر تصادفی می شوند. فیلترها نمی دانند که به دنبال قدرت و طول موج هستند. فیلترهای لایه های بالاتر نمی دانند که باید به دنبال چه احساساتی باشند. با افزایش دوره های آموزشی epoch ها و گذر زمان صداهای متفاوتی به مدل داده شده است و برچسب مربوطه را به آنها الصاق می گردد. این ایده برای دادن صداهای متفاوت و برچسب ، فرآیند آموزشی است که CNN ها طی می کنند. یک مجموعه آموزشی وجود دارد که دارای هزاران صدا از احساسات متفاوت اعم از شادی، غم، ترس، عصبانیت و آرامش است و هر یک از صداها دارای برچسب است که آن صدا چه احساساتی دارد.

بنابراین انتشار عقب Back Propagation را می توان به 4 بخش مجزا ، دریافت صدا و آموزش مدل رو به جلو ، تخمین ضرر ، محاسبات مجدد وزن ها در پیمایش نورونها در انتشار عقبگرد و ارتقاء وزن تقسیم کرد. در طول انتشار رو به جلو ، یک صدای آموزشی گرفته می شود که یک آرایه m * n از اعداد است و آن را از کل شبکه عبور می دهد. از آنجا که همه وزنها یا مقادیر فیلتر به طور تصادفی راه اندازی شدند ، خروجی احتمالاً چیزی شبیه [.1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1] خواهد بود. خروجی که به هیچ عدد خاصی ترجیح داده نمی شود و این شبکه ، با وزن فعلی خود ، نمی تواند به دنبال ویژگی های سطح پایین باشد بنابراین نمی تولند نتیجه گیری منطقی در مورد طبقه بندی داشته باشد. این به بخش تابع از دست دادن back propagation می رود. آنچه در حال حاضر از آن استفاده می شود ، داده های آموزشی است. این داده ها دارای یک صدا و یک برچسب هستند. یک تابع محاسبه اشتباه را می توان به روش های مختلف تعریف کرد ، اما یک تابع رایج MSE (خطای میانگین مربع) است که برابر مربع (واقعی - مقدار پیش بینی شده)½ است.

$$Cost\ Function = \sum \frac{1}{2}\ (target - output)^2$$

مقدار اولیه تابع اشتباه برای اولین صدای آموزشی بسیار زیاد خواهد بود. مقصود رسیدن به جایی است که برچسب پیش بینی شده (خروجی ConvNet) با برچسب آموزشی یکسان باشد (این بدان معناست که شبکه پیش بینی خود را به درستی انجام داده است). برای رسیدن به آنجا ، می خواهیم میزان ضرر و اشتباه به

حداقل برسد. با تصور این مسئله فقط به عنوان یک مشکل بهینه سازی در محاسبات ، یافت می شود که کدام ورودی ها (وزن ها) به طور مستقیم به (خطا) شبکه کمک کرده اند.

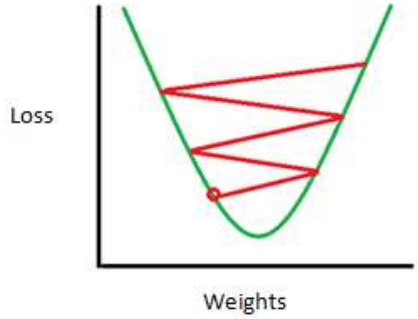


این معادل ریاضی dL/dW است که در آن W وزن یک لایه خاص است. در حال حاضر ، هدف این است که از شبکه به عقب حرکت کنیم ، که تعیین می کند کدام وزنه ها بیشترین ضرر یا خطا را داشته اند و راه هایی برای تنظیم آنها به گونه ای پیدا می شود که از دست دادن کاهش یابد.

$$w = w_i - \eta \frac{dL}{dW}$$

w = Weight
w_i = Initial Weight
η = Learning Rate

هنگامی که این مشتق محاسبه می شود، مدل آموزشی به آخرین مرحله که بروزرسانی وزن است می رود. اینجاست که همه وزن فیلترها را گرفته و آنها را به روز می کند تا در جهت مخالف گرادیان تغییر کنند. میزان یادگیری Learning Rate پارامتری است که توسط برنامه نویس انتخاب می شود. نرخ یادگیری بالا به این معنی است که گام های بزرگتری در به روزرسانی وزن برداشته می شود و بنابراین ، ممکن است زمان کمتری طول بکشد تا مدل در یک مجموعه بهینه از وزنه ها همگرا شود. با این حال ، میزان یادگیری بیش از حد بالا می تواند منجر به جهش هایی شود که بسیار بزرگ هستند و به اندازه کافی دقیق نیستند تا به نقطه مطلوب برسند.



فرایند انتشار به جلو ، محاسبه خطا در وزنها، انتشار به عقب و به روزرسانی پارامترها یک تکرار آموزشی است. برنامه این فرایند را برای تعداد تکرار ثابت برای هر مجموعه از صدای آموزشی (که معمولاً دسته ای نامیده می شود) تکرار می کند. پس از اتمام به روزرسانی پارامتر در آخرین مثال آموزشی ، امید است شبکه به اندازه کافی آموزش ببیند تا وزن لایه ها به درستی تنظیم شود.

2-3- کتابخانه پای تورچ (PyTorch)

2-4- مروری بر منابع

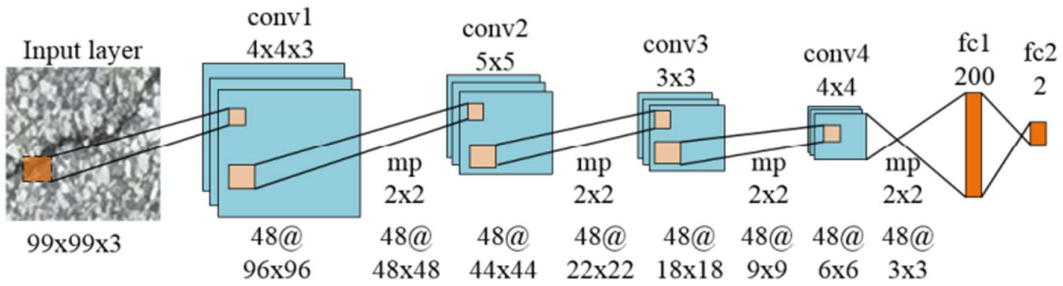
2-5- نتیجه‌گیری و نوآوری

فصل 3:

روش تحقیق

3-1- مقدمه

با توجه به تصویر 1-2، هدف از تشخیص احساسات در صدا ، تعیین این است که آیا پیکسل خاصی دارای قدرت و طول موج خاصی است یا خیر. برای حل این مشکل ، راه حل پیشنهادی مبتنی بر ConvNet است ، که از تکه های صدا اطلاعات لازمه را دریافت می کند، سپس برای طبقه بندی تکه ها با احساسات متفاوت آموزش داده می شود.



شکل (1-3) معماری شبکه کانولوشنال نتورک ConvNet پیشنهادی[1]

معماری ConvNet در شکل 1-2 نشان داده شده است ، جایی که conv، mp و fc به ترتیب نمایانگر لایه های کانولوشن ، حداکثر ترکیب و کاملاً متصل هستند. به طور کلی ، ConvNet به عنوان یک استخراج کننده ویژگی های سلسله مراتبی در نظر گرفته می شود ، که ویژگی های سطوح مختلف انتزاعی را استخراج می کند و شدت درایه های خام درایه های صدا را توسط چندین لایه کاملاً متصل به یک بردار ویژگی ترسیم می کند. همه پارامترها به طور مشترک از طریق به حداقل رساندن خطای طبقه بندی نادرست در مجموعه آموزش از طریق روش انتشار عقبگرد بهینه می شوند.

همه عناصر هسته فیلتر حرکتی با یادگیری از مجموعه نمونه های برچسب گذاری شده توسط Train Set از مجکوعه داده های آموزشی به شیوه ای تحت نظارت، آموزش داده می شوند. در هر لایه کانولوشن ConvNet، عملیات جمع آوری حداکثر انجام می شود تا پاسخ های ویژگی را در درایه های مجاور خلاصه کند.

چنین عملیاتی به ConvNet اجازه می دهد تا ویژگیهایی را که از نظر مکانی تغییر ناپذیر هستند یاد بگیرد ، یعنی از نظر موقعیت قدرت و طول موج در صداها تغییر نمی کند. در نهایت ، لایه های کاملاً متصل برای طبقه بندی استفاده می شود. به دلیل خاصیت متقابل منحصر به فرد مشکل تشخیص احساسات، یک لایه

softmax به عنوان آخرین لایه ConvNets برای محاسبه احتمال هر کلاس با توجه به یک وصله ورودی استفاده می شود.

3-2- دیـتـا ست

مجموعه آموزشی داده شده بصورت زیر است:

$$S = \{ x^{(i)}, y^{(i)} \} \qquad m\ audios\ x^{(i)}\ is\ i^{th}\ audio$$

$$y^{(i)} \in \{1, 2, 3, 4, 5\}$$

$$If\ y^{(i)} = 1 \Rightarrow x^{(i)}\ is\ Sad$$

$$If\ y^{(i)} = 2 \Rightarrow x^{(i)}\ is\ Happy$$

$$If\ y^{(i)} = 3 \Rightarrow x^{(i)}\ is\ Angry$$

$$If\ y^{(i)} = 4 \Rightarrow x^{(i)}\ is\ Calm$$

$$If\ y^{(i)} = 5 \Rightarrow x^{(i)}\ is\ Fear$$

در مجموعه آموزشی تعداد i =20000 قطعه صدا یا احساسات متفاوت داریم. این صدا ها که به 25 نفر تعلق دارند هر کدام در پوشه مخصوصی با نام (Actor_01 … Actor_25) دسته بندی می شوند، در نتیجه برچسب آنها براساس نوع احساسات نامگذاری می شود.

آموزش ConvNet توسط واحدهای پردازش گرافیکی (GPU) و (FPGA)تسریع می شود. افزایش بیشتر سرعت در هر دو مرحله آموزش و ارزیابی، با استفاده از واحدهای خطی اصلاح شده (ReLU) که به عنوان تابع فعال سازی حاصل می شود، موثرتر از توابع مماس هذلولی tanh (x) و تابع سیگموید$(1 + e^{-x})^{-1}$ که مورد استفاده در مدلهای عصبی کلاسیک است.ConvNets با استفاده از روش نزول شیب تصادفی (SGD) با اندازه 48 نمونه ، momentum 0.9 و decay weight 0.0005 آموزش دیده است. کمتر از 20 epochs برای رسیدن به حداقل در مجموعه اعتبارسنجی مورد نیاز است.

شکل (2-3) وارد نمودن کتابخانه ها

شکل (3-3)

عنوان پایان‌نامه/ تشخیص احساسات گفتاری با استفاده از شبکه عصبی عمیق کانولوشنال

فصل 4:

نتایج و تفسیر آن‌ها

4-1- مقدمه

4-2- صحه‌گذاری[1]

[1] Validation

4-2-1- ارائۀ نتایج

فصل 5:

نتیجه‌گیری و پیشنهادها

5-1- مقدمه

5-2 نتایج حاصل از شبیه سازی

1- سازه و

مراجع و منابع

[1] L.

پیوست‌ها

عنوان پایان‌نامه/ تشخیص احساسات گفتاری با استفاده از شبکه عصبی عمیق کانولوشنال

پیوست الف

Abstract:

The main goal of this research is based on Emotional Speech Detection, which is related to the different actors as the inputs. In this thesis, there is Convolutional Neural Network. The most important things here is to compute how much CNN can improve the acceleration and accuracy for model learning instead of using another algorithm. The purpose is to predict the growth of velocity parameter while holing on precision parameters. In this research, the mentioned parameters are extracted from Google Colab.

Keywords:

BSc/ MSc/ PhD Thesis Title

Emotional Speech Detection By Convolutional Neural Network

A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Bachelor of Science / Master of Science / Doctor of Philosophy in - engineering - Orientation

By:

Supervisor:

Dr.

Advisor:

Dr.

October 2021