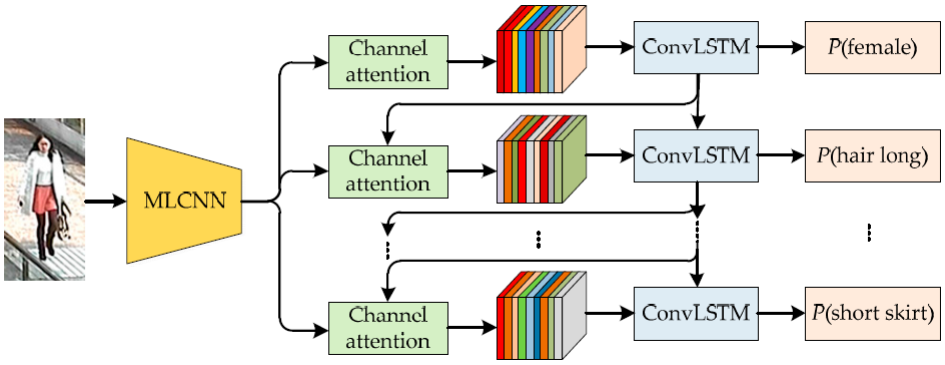
تمرین چهار

شبکه عصبی

۱ تشخیص صفات عابر پیاده

در این مجموعه داده هدف یافتن صفات عابر پیاده است که این صفات به صورت یک دنباله مکانی_زمانی در نظر گرفته شده است. به طور مثال احتمال ظهور تصویر یک مرد و دامن بسیار پایین است، یعنی فرض می‌شود این صفات همبستگی بسیار بالایی با یکدیگر دارند. برای تشخیص این صفات مدلی با ساختار زیر پیاده‌سازی شود. در این مدل



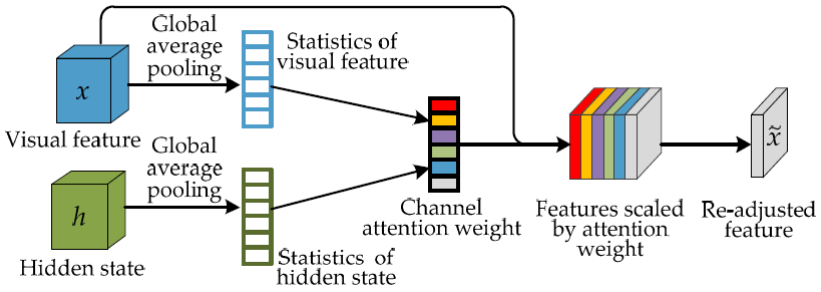
شکل ۱: ساختار شبکه خواسته شده.

بخش MLCNN برای استخراج ویژگی‌های بصری تصویر استفاده می‌شود. قسمت بعدی یک مکانیسم توجه خود سازگار [1] است که تاثیر ویژگی‌های بصری را طوری تغییر می‌دهد که بهترین ویژگی‌ها برای تشخیص صفات مشخص شود. ساختار این بخش در تصویر ۲ مشخص شده است. بخش بعدی مدل لایه‌های ConvLSTM هستند که با استفاده از ویژگی‌های بصری و حافظه نهان مرحله به مرحله صفات تصویر را استخراج می‌کند.

۱.۱ ConvLSTM theory

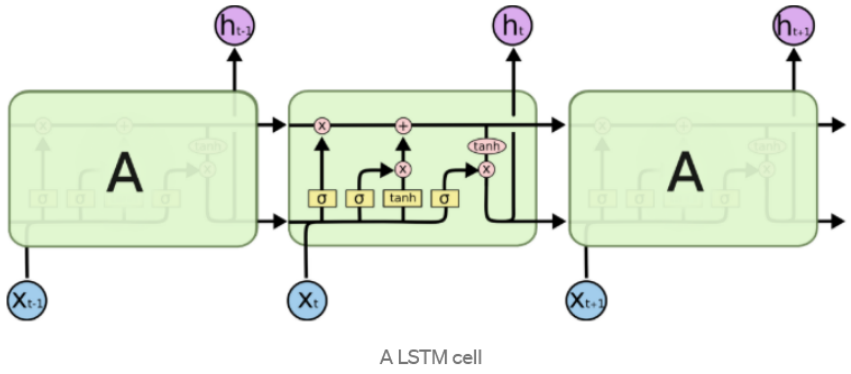
در دنیای امروز ما با داده های سری بسیاری سرکار داریم همانند داده های مربوط به ویدیو ، عکس های ماهواره ای ، دوربین های مدار بسته و داده های بورس و ... .
به صورت کلی داده های جمع آوری شده در طول زمان را یک سری زمانی [2] می نامند . در چنین حالتی چون داده های زمان حال تاثیری از داده ها گذشته میگیرند مثلا داده های بورس را در نظر بگیرید که قیمت امروز ارتباطاتی با

[1] self-adaption
[2] Time Series



شکل ۲: ساختار مکانیسم توجه

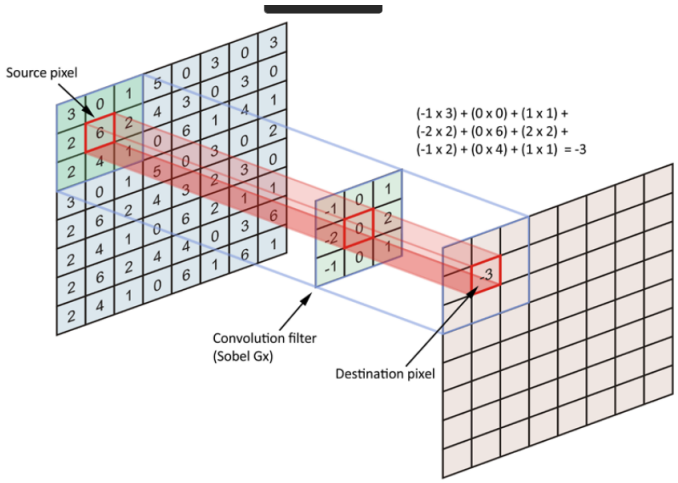
داده های روز های گذشته خواهد داشت و یا در فیلم ها که فریم فعلی ارتباطی با فریم های قبل و بعد از خود خواهد داشت به همین دلیل استفاده از شبکه های با حافظه همانند Long Short Term Memory [۳] می تواند ما را در تحلیل داده ها کمک کند.
ساختار شبکه LSTM در کلاس درسی توضیح داده شده است اما به صورت خلاصه اگر بخواهیم این شبکه ها را بررسی کنیم میتوان گفت که در این شبکه ها ، حالت پنهان قبلی را به مرحله بعدی دنباله منتقل می کند. بنابراین نگهداری اطلاعات در این شبکه ها مشاهده میشود . در طرف دیگر ما شبکه های کانولوشنی [۴] را داریم که برای



شکل ۳: ساختار یک شبکه LSTM

پردازش های عکس ها استفاده میکنیم . در این شبکه ها ما با فیلتر های کانولوشنی سرکار داریم که در هر مرحله بر روی عکس ورودی اعمال میشوند و هدف استخراج فیچرهای مهمی از تصویر میباشد . در خصوص جزییات شبکه های عصبی کانولوشنی در کلاس درس و جلسه دوم حل تمرین توضیحات تکمیلی داده شده است .

[3] LSTM
[4] CNN



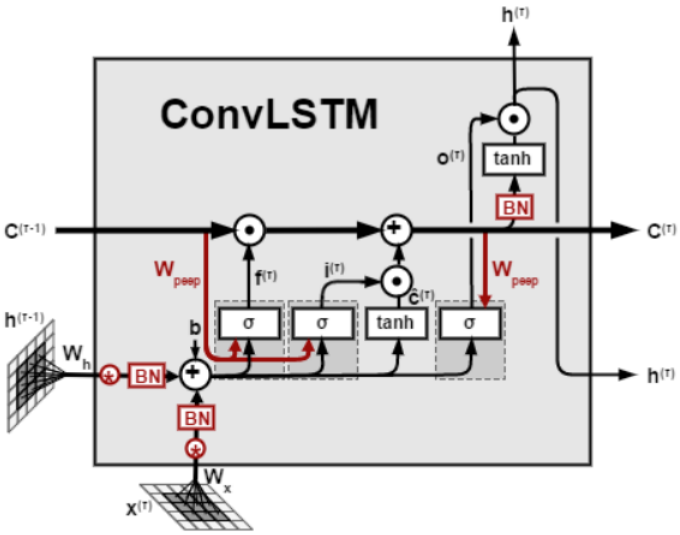
شکل ۴: اعمال یک فیلتر کانولوشنی بر روی عکس در شبکه های کانولوشنی

بنا به کاربرد ممکن است داده های تصویری ما وابسته به زمان باشند لذا دراین گونه مواقع ما نیاز به رویکردی شبیه LSTM خواهیم داشت . در این متن دو رویکرد در خصوص داده های تصویری متناسب با زمان مطرح شده است . لازم است خارج از این دو رویکرد ، رویکردهای دیگری نیز وجود دارد که استفاده از آن رویکرد ها میتواند برای شما **نمره امتیازی** به همراه داشته باشد .

۲.۱ رویکرد شماره ۱

رویکرد اول استفاده از لایه های ConvLSTM میباشد . در اصل این لایه ها ، لایه های بازگشتی [۵] همانند LSTM می باشند با این تفاوت که به جای ضرب داخلی ماتریس ها ، ما از عملیات کانولوشنی بهره میبریم . با جایگزینی عملیات کانولوشنی ، جریان دیتاهای ما به جای این که یک بعدی باشند ، ابعاد ورودی را حفظ میکنند که برای یک عکس سه رنگی که ورودی ۳ بعدی است ، این سه بعد در طول جریان داده ها حفظ میگردد .

[5] Recurrent layers



شکل ۵:

۳.۱ رویکرد شماره ۲

رویکرد متفاوتی دیگری نیز وجود دارد که در آن، شما یک مدل Convolutional-LSTM دارید یعنی ترکیبی از یک شبکه کانولوشنی و یک شبکه با حافظه . در این حالت عکس ها از مجموعه از فیلتر های کانولوشنی عبور میکنند سپس نتیجه ی حاصل flatten شده و تبدیل به برداری یک بعدی میگردد . همانند فرایندی که در تمرین شماره ۲ برای عکس های راهنمایی رانندگی طی کردیم . هنگامی که ما فرایند گفته شده را برای تمامی عکس های یک time set را اجرا کنیم ، خروجی کار ما مجموعه ای ویژگی ها در طول زمان میگردد که حال این فیچر ها در طول زمان را به عنوان ورودی به یک شبکه LSTM می دهیم.

توجه ۱: در صورتی که شما هر دو رویکرد بیان شده را پیاده سازی کنید ، **نمره امتیازی** به شما تعلق میگیرد .

توجه ۲ : لایه ConvLSTM مربوط به رویکرد اول در tensorflow وجود دارد . توصیه میشود اول چگونه ورودی دادن به یک لایه LSTM در tensorflow بررسی کنید سپس به سراغ لایه ConvLSTM بروید .

توجه ۳ : در صورت مشخص کردن تاثیر مکانیسم توجه در یک تصویر مانند تصویر ۶ ، **نمره امتیازی** به شما تعلق می‌گیرد.

(a) Original (b) Head-shoulder (c) Upper-body (d) Lower-body (e) Footwear (f) Accessory

شکل ۶: تاثیر مکانیسم توجه

۲ تشخیص احساسات

۱.۲ مقدمه و تعریف مسئله

امروز بحث تشخیص احساسات کاربردهای فراوانی در حوزه های مختلفی از قبیل بازاریابی ، پزشکی ، روانپزشکی و ... پیدا کرده است . در تسک جاری شما با تشخیص احساسات از روی صدا [۶] کار میکنید به این صورت که صدایی را به عنوان ورودی دریافت کرد و حال باید کلاس بندی بکنید که این صدا مربوط به کدام کلاس از احساسات می‌باشد . جهت راحت تر شدن مسئله ما تنها ۵ کلاس احساسات داریم .
دیتاستی که به شما داده شده است به صورت زیر است .

```
------ train.zip
------------ train
```

در فولدر train شما ۱۹۹۴ فایل صوتی با فرمت wav. دارید . اسم گذاری هر فایل صوتی به این صورت است که یک عدد که همان Id صدا میباشد و یکتا است و یک بخش حرفی که شامل دو حرف انگلیسی میباشد . حرف دوم بیانگر کلاس احساسات همان لیبل کلاس و حرف اول بیان گر دیتای اضافه جنسیت کسی که صدا را تولید کرده است میباشد .
به طور مثال نام یکی از فایل های صوتی به صورت زیر است .

$$۲۷۰۵FN.wav \rightarrow \underbrace{۲۷۰۵}_{\text{Id}} \quad \underbrace{F}_{\text{Gender}} \quad \underbrace{N}_{\text{label}} \tag{۱}$$

شما در این دیتاست با ۵ کلاس احساسات و دو جنسیت زن و مرد که با M،F نمایش داده میشود ، سرکار دارید که در جدول زیرتوضیحات مربوط به ۵ کلاس احساسات قابل مشاهده است .
این دیتاست از دیتاست موسوم به ShEMO [۷] استخراج شده است که جهت سادگی کار فقط بخشی از دیتاست انتخاب شده است .

[6] Speech Emotion Detection
[7] https://arxiv.org/abs/1906.01155

کلاس احساسات	لیبل کلاس احساسات
خشم	A(anger
شادی	H(happiness)
غم	S(sadness)
شگفتی	W(surprise
بی تفاوت	N(neutral)

۲.۲ راهنمایی

در این تسک شما در ابتدا باید فایل های wav. را بخوانید که نیاز به کتابخانه های حوزه صدا دارید که یکی از کتابخانه های خوب در این زمینه کتابخانه Librosa می باشد .
در گام بعدی شما نیاز دارید که فیچرهایی را از صدا استخراج کنید که به این کار Feature extraction میگویند که روش های مختلفی برای این کار وجود دارد [۸]
یکی از روش های خوب و ساده میتواند روش MFCC [۹] باشد . [۱۰] (البته شما مجاز به انتخاب هر روش دلخواه برای استخراج ویژگی از صدا میباشید.)
پس از استخراج ویژگی از اصوات ما باید از شبکه های با حافظه بهره ببریم چرا که اصوات انسان از نظر زمانی رابطه تنگاتنگی را دارا می باشد . برای این بخش شما میتوانید از LSTM بهره ببرید . در کنار این شبکه با حافظه شما نیاز به یک Classifier نیز دارید همانند پروژه ۲ که شما اطلاعات عکس را به یک فضای جدید برده سپس در فضای جدید آنها را کلاس بندی میکردید .
توصیه میشود در صورت داشتن وقت ،تاثیر داده اضافه جنسیت را برروی نتایج کار خود بسنجید . اگر مدل انتخابی شما در کلاس خاصی از صدا بهتر عمل میکند سعی کنید برای آن استدلال و دلیل بیاورید . سعی کنید از روش های استخراج ویژگی متفاوتی و یا ترکیب آنها استفاده کنید و خروجی کار را با هم دیگر مقایسه کنید . کدام روش ها بهتر عمل میکنند ، کدام روش ها روی برخی از کلاس ها بهتر عمل می کنند .(این بخش ها دارای **نمره امتیازی** خواهند بود .)

---

[۸] یکی از مطالب خوب جهت معرفی این زمینه میتواند مقاله مروری "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011." با شماره DOI ، 10.1007/s10462-012-9368-5 باشد.

[9] Mel-frequency cepstral coefficients

[۱۰] اصوات تولید شده توسط انسان به وسیله شکل مجرای صوتی که شامل زبان ، دهان ، دندان ها و ... فیلتر میشود . این شکل مطرح میکند که صدا خروجی به چه صورتی باشد . حال اگر ما بتوانیم این شکل را تعیین کنیم میتوانیم اطلاعات ارزشمندی در خصوص واج در حال تولید به دست بیاوریم . به کمک MFCC ما میتوانیم اطلاعات ارزشمندی در خصوص این شکل به دست بیاوریم .

توجه :

- اگر از روش های دیگری برای استخراج ویژگی بهره میبرید لطفا مقاله مرتبط یا متنی که از روی آن ، متد را مطالعه کرده اید را نیز ارسال کنید و یا در گزارش خود مشخصات آن را درج کنید .
- توصیه میشود برای شروع کار سعی کنید نمونه کد های مشابه مثلا نمونه کدها و کار های در سایت Kaggle را ببینید و ایده بگیرید . توجه دارید که هدف کپی کردن کامل کدهای موجود در github یا Kaggle نمی باشد . درصورتی که استفاده زیادی از کدهای موجود در Kaggle و یا github داشته اید **درج منبع کد الزامی است** هم چنین لازم است آن بخش به صورت کامل در گزارش نهایی توضیح داده شود .