\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ایده و راه حل کلی :

با توجه کمبودن داده های پروژه از آنجایی که منجر اور فیتتینگ میشد ابتدا سعی بر افزایش تعداد داده ها بود که چند تکنیک برای افزایش داده ها استفاده شده بود

یکی از این تکنیک ها که ایده آن را از مقاله ای برداشتم (لینک در انتهای گزارش قرار داده شده است) این بود که به طور رندوم 15 درصد فریم های صدا را حذف کنیم

ایده های دیگر همه از افکت های از پیش نوشته شده کتابخانه لیبروسا بودند که بعدا به با جزییات توضیح داده خواهند شد

بعد از افزایش تعداد صدا ها که البته قابل ذکر است که داده های مربوط به کلاس های

S, H, W

به نسبت بقیه داده ها از حجم کمتری برخوردار بودند که نیاز بود اندازه کلاس ها را تا حدی متوازن تر کنیم به همین منظور افزایش داده ها روی این 3 کلاس بیشتر انجام شده است

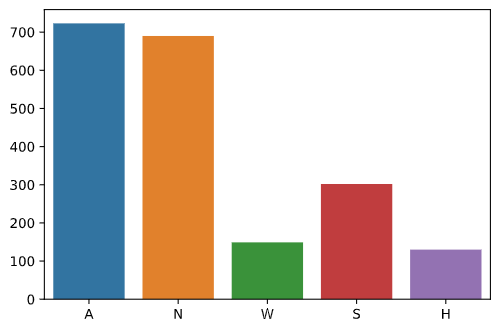
بعد از افزایش تعداد داده ها نوبت به استخراج ویژگی هایی از صدا ها داریم تا بتوانیم سری های زمانی داشته باشیم که به شبکه های با حافظه خود بدهیم

برای شبکه با حافظه از یک شبکه LSTM استفاده شده است

و یک شبکه BiDirectionalLSTM (به توصیه یکی از دوستان) استفاده شده است

اضافه کردن داده ها :

ابتدا نگاهی به توضیع داده های خام میاندازیم



همانطور که مشخص است سه کلاس آخر نیاز بیشتری به اضافه سازی داده دارند

برای اینکار روی این سه کلاس تکنیک های زیر را پیاده میکنیم

Librosa.effects.time\_strech :

این افکت همانطور که از اسمش پیداست به نوعی فاصله بین فریم ها را تغییر میدهد به این شکل که موقع اجرا آرام تر یا سریعتر اجرا شوند

این افکت را روی نیمی از داده های S بطور رندوم و تمام داده های W و H پیاده سازی میکنیم

Librosa.effects.pitch\_shift :

این افکت نیز همانطور که از اسمش پیداست به زبان ساده همه فرکانس های در هر لحظه را با نرخی که در آرگومانش به آن میدهیم به فرکنس های بالاتر یا پایینتر انتقال میدهد

البته در استفاده از این افکت انتظار میرود تشخیص جداگانه زن و مرد کمی به مشکل بخورد چرا که بطور کلی صدای زنان از فرکانس های بالاتری برخوردار است

اما بطور کلی میتوان انتظار داشت که در پیشبینی کلیت داده ها بهبود ایجاد کند

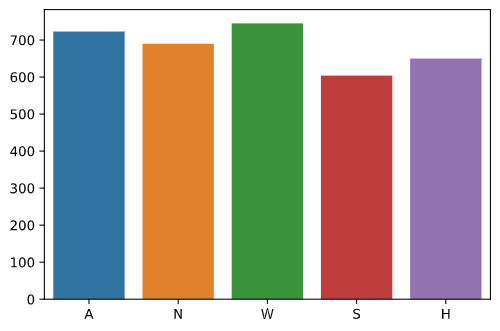
این روش را روی نیمی از کلاس S بطور رندوم و تمام داده های W و H اعمال میکنیم

Time Masking :

این روش که ایده آن را از مقاله ای برداشتم همانطور که در بالا توضیح داده شد به طور رندوم 15 درصد فریم های صدا را حذف میکند

این روش را روی تمامی کلاس ها اعمال میکنیم (این کار را در دو حالت وجود و نبودش امتحان کردم و وجودش تمام هزینه محاسباتی داشت و نه بهبودی در دقت. در حالت نبودش روش های بالا را بر سه کلاس S و H و W پیاده سازی کردم)

پس از اتمام این کار توضیع داده ها به شکل زیر است



سپس نوبت به استخراج ویژگی ها میرسد

برای این کار از 2 تابع زیر استفاده شده است

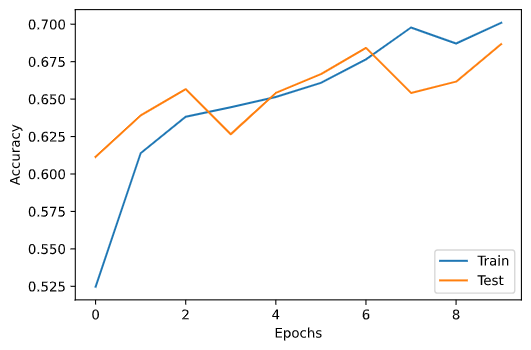
Librosa.features.mfcc

Librosa.features.chroma\_cens

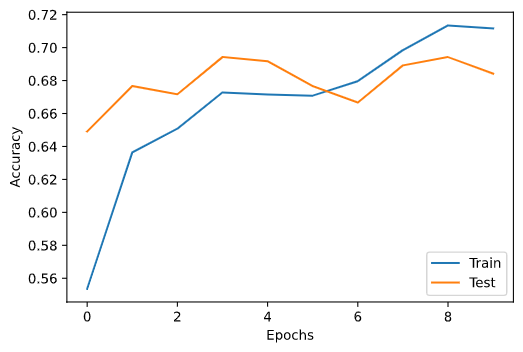
(توابع دیگری نیز در سایت لیبروسا مطالعه کردم اما تقریبا هیچکدام مثل این دو سری زمانی خروجی نمیداد و به کار ما نمیامد)

در این دو ویژگی تعداد گام های زمانی مختلف را امتحان کردم که بهترین آنها 72 بود.

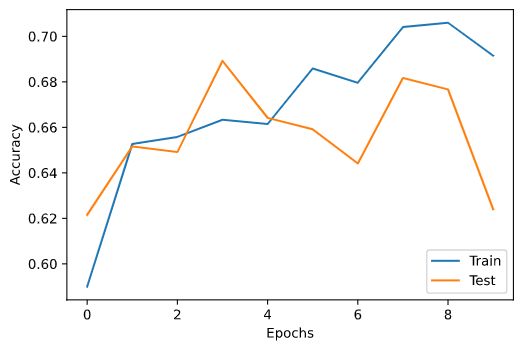
در اصل سه گام 36 و 72 و 128 امتحان شد که نتایج آن ها را در زیر میبینیم. البته درست است که دقت 72 کمی بهتر از 36 بود اما استفاده از 72 نه تنها هزینه بیشتری داشت بلکه کمی متمایل به اورفیت شدن بود و تصمیم بهینه همان 36 بود. 128 نیز نه تنها از 72 نتیجه بهتری نداشت بلکه افت دقت را نیز در کنار افزایش هزینه در بر داشت



36 : در این حالت میبینیم نسبتا بهترین جواب را گرفته ایم. البته قابل به ذکر است در این سه عکس تنها از فیچر اول که استخراج شده بود استفاده شده و بعد نتایج استففاده از هر را اعمال خواهیم کرد

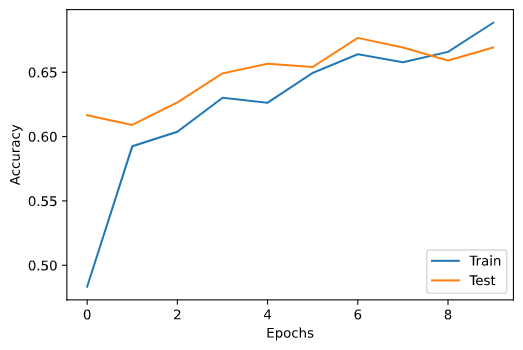


72 : این حالت کمی و فقط کمی از 36 بهتر بود و در اصل از ابتدا حالتی که انتظار میرفت این بود که با افزایش این پارامتر جزییات بیشتری را حفظ کنیم که نتیجه آن این بود که در مواقعی مانند این مورد استفاده از جزییات بیشتر لزوما خوب نیست و باید تعادلی در جزییات و مفید بودن آن ها برقرار باشد



128 : همانطور که در بالا نیز گفته شد نبود این تعادل باعث شده است شبکه ما در کمتر از 5 ایپاک به سمت اورفیتینگ شدیدی برود و ازین به بعد استفاده از 36 مورد صلاح است

در عکس زیر نیز تاثیر استفاده از هر دو فیچر را بجای فقط mfcc میبینیم. درصد دقت در 5 ایپاک لزوما بهتر نشد اما روند آموزش و بهبود شبکه روند با ثبات تری را طی کرد. میتواند به این معنی باشد که با این کار هویت داده ها بهتر حفظ شده است. البته افزایش ایپاک نیز امتحان شد. ثبات هنوز محسوس بود اما دقت بیشتر از 67 نمیرفت و بعد از آن به روند کمی نامحسوس از اورفیتینگ برمیخورد



نتایج تمام داده های آموزش :

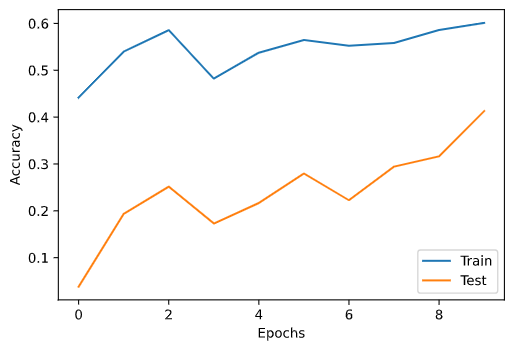
حال داده های افزایشی را امتحان میکنیم

با اینکار انتظار میرفت دقت دو کلاس H و W و S بهبود پیدا کند که در انتها نتایج آن را در یک confusion plot میبینیم.

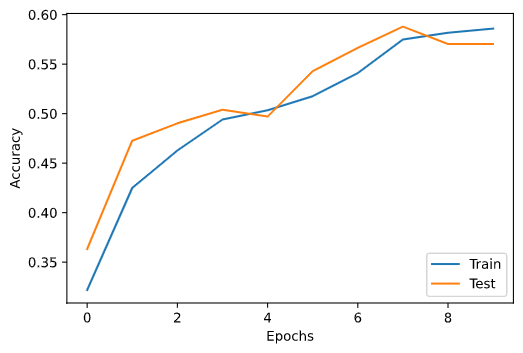
اما آنطور که انتظار میرفت پیش نرفت

اولین حالت زمانی است که نرخ pitch\_shift برابر2+ بود یعنی همه فرکانس ها به اندازه 2 pitch زیر تر شدند. و همچنین نرخ time\_strech برابر 1.1 و 0.9 بود و نرخ time\_masking برابر 10 درصد احتمال

نتایج آن به شکل زیر است

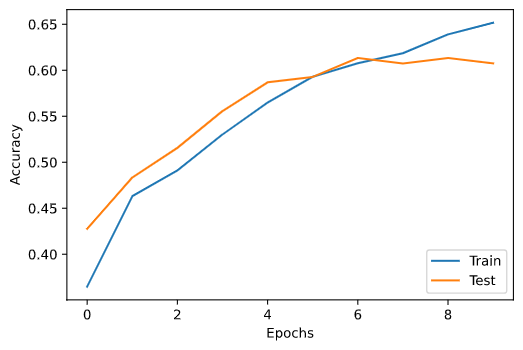


در این حالت نتیجه خوبی را در بر نداشتیم و البته نکته ای قابل ذکر است. این حالت را به اشتباه بدون بر زدن دیتا ها و در حال داشتن دیتا ها بطور تریتیب اولیه augmentation استفاده شده بود که البته خود نکته آموزشی مفیدی در بر داشت در عکس زیر با کتابخانه sklearn داده ها را بر زده ایم

در این عکس پارامتر ها همان قبلی و فقط داده ها بر زده شدند و اینجا میبینیم که چقدر وجود یه ترتیب تصادفی در روند آموزش مهم است

در حالت دوم :

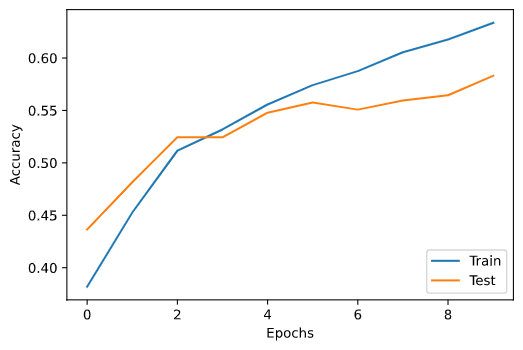
در این حالت نرخ pitch\_shift برابر4+ بود یعنی همه فرکانس ها به اندازه 4 pitch زیر تر شدند. و همچنین نرخ time\_strech برابر 1.3 و 0.7 بود و نرخ time\_masking برابر 20 درصد احتمال

دقت به نسب حالت قبل بهتر شد اما در نمودار heatmap میبنییم که با افزایش تعداد داده ها نه تنها تشخیص کلاس ها را راحت تر کردیم بلکه دشوار تر نیز شده است !

به نظر میرسید استفاده از تکنیک های augmentation در موارد این گونه مانند صدا کار راحتی نیست و استفاده درست از این تکنیک ها نیازمند تجبره بیشتر است

در حالت بعدی نرخ های تکنیک ها افزایش داده را باز هم تغییر دادیم (به ترتیب 8+ و 1.5 و 0.5 و 30 درصد)

اما شروع به کاهش دقت داشتیم و دیگر ادامه ندادم



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

برای نتیجه کار میتوان گفت استفاده از داده های افزایش یافته لزوما کار را بهتر نمیکند و بلکه برای بهبود نتایج باید دانست از چه تکنیک هایی در مواقع مناسبشان برای افزایش داده استفاده کرد

همچنین نکته جالبی که قابل توجه بود این بود که در اولین مرحله از استفاده از داده های افزایشی استفاده از test\_train\_split و بر زدن داده ها چقدر در نتیجه نهایی تاثیر مثبت داشت.

لینک مقاله ذکر شده : https://towardsdatascience.com/data-augmentation-for-speech-recognition-e7c607482e78