به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس شبکه‌های عصبی

استاد صفابخش

تمرین دوم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

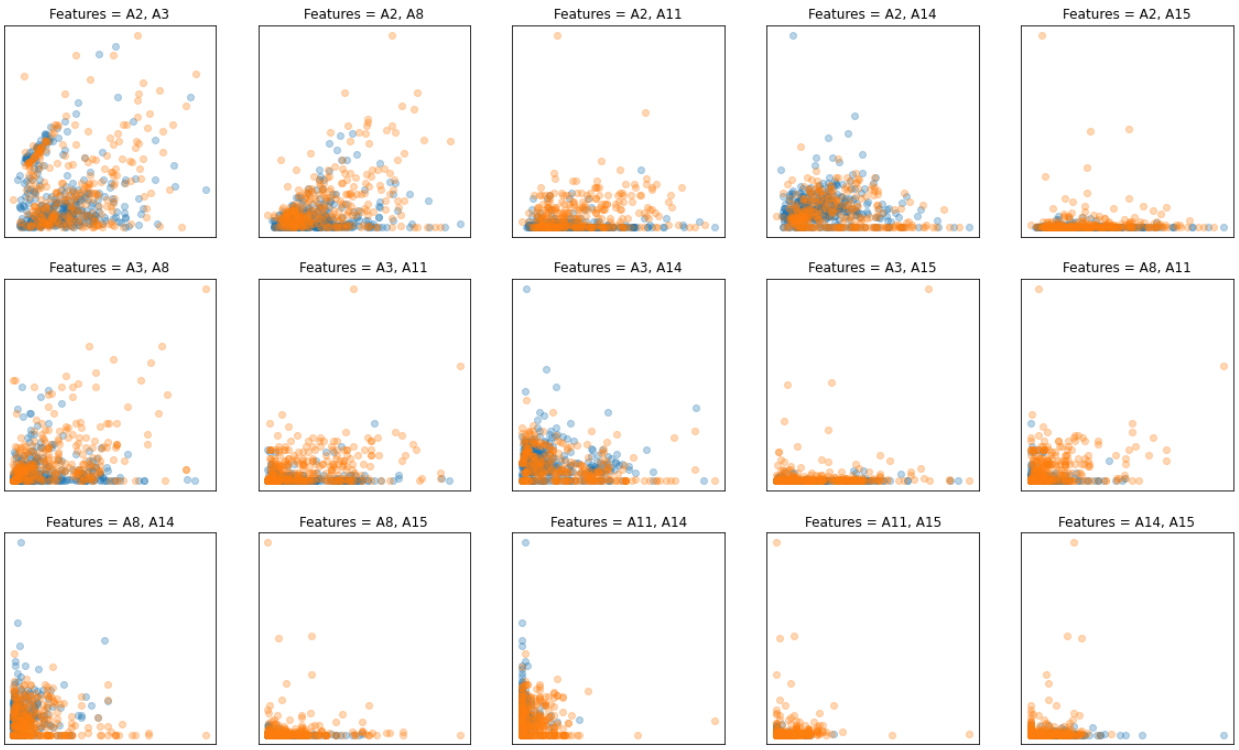
سوال ۱

مطابق با توضیحات موجود در مجموعه‌داده در حدود ۵ درصد داده‌ها یک یا چند ویژگی دارای مقدار تهی است. برای ستون‌هایی که دارای داده‌های پیوسته هستند از میانگین سایر داده‌ها و برای ستون‌هایی که دارای مقادیر گسسته هستند از ماکسیمم مقدار برای پر کردن مقدار از دست رفته بهره می‌گیریم. مقادیر + و - ستون کلاس را به مقدار ۰ و ۱ تبدیل کردم. برای تمام ستون‌های ویژگی یک نرمال‌سازی مطابق با فرمول زیر هم به کار می‌گیریم تا رنج مقادیر برای تمام این ستون‌های مشابه با یکدیگر باشد:

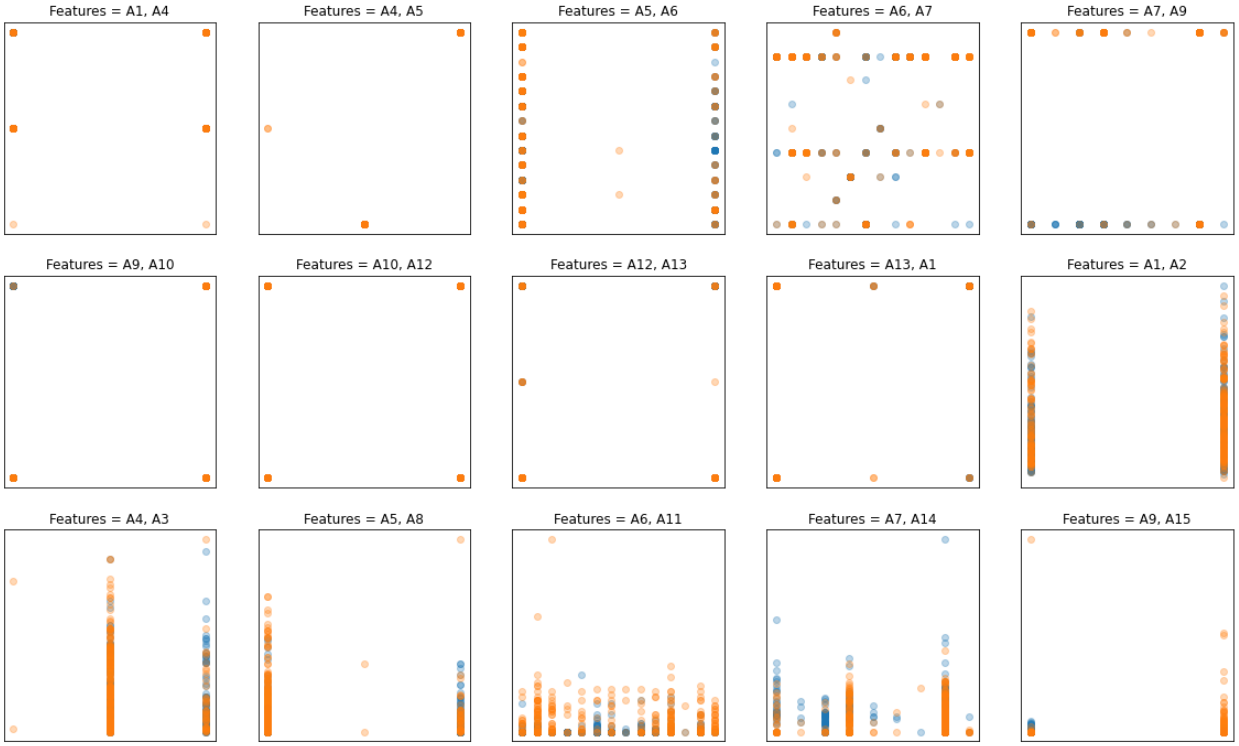
$$x_{normal} = \frac{x - mean(x)}{std(x)}$$

سوال ۲

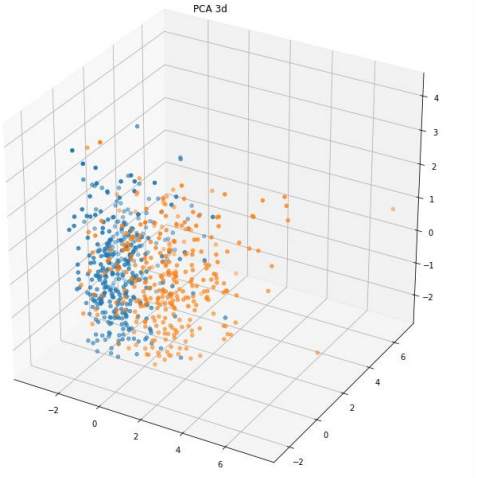
برای بررسی این مورد، چندین بررسی را انجام دادیم. اول آنکه ویژگی‌ها پیوسته را به صورت دو به دو در یک نمودار ترسیم کردیم ولی در هیچ کدام از نمودارها داده‌های دو کلاس از یکدیگر به صورت خطی جدا نشد:

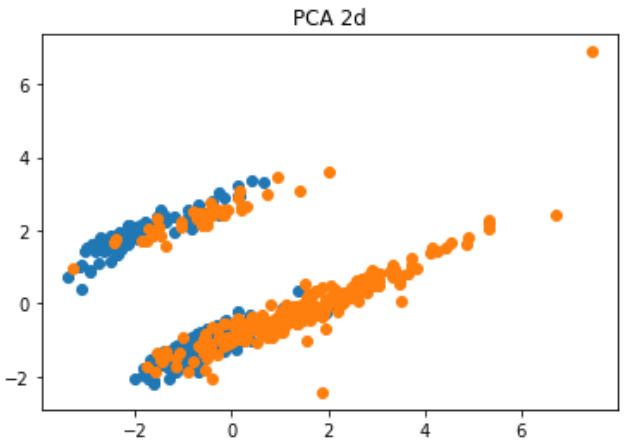


ویژگی‌های پیوسته با توجه به رنج وسیع‌تری که نسبت به ویژگی‌های گسسته دارند احتمال بیشتری برای جداکردن کلاس‌ها دارند ولی با این حال ممکن است ویژگی‌های گسسته داده‌ها را جداکنند. لذا تعدادی از ترکیب‌های ویژگی‌ها گسسته با هم و با ویژگی‌های پیوسته را بررسی کردیم ولی در این حالت هم موفقیتی حاصل نشد:



بار دیگر از روش PCA استفاده کردیم ولی این بار هم جداسازی داده‌ها انجام نشد:

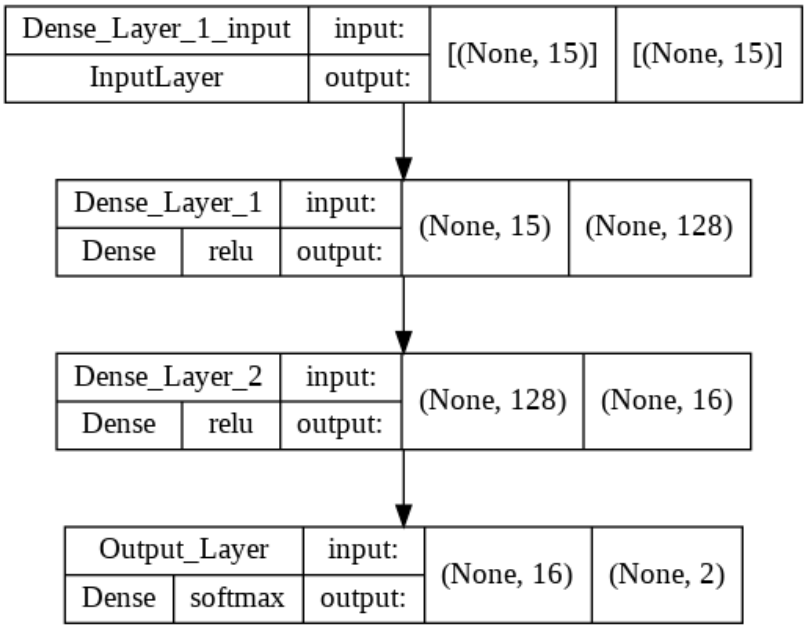




حال که این روش‌های اولیه جواب ندادند، نوبت به بررسی یک روش پیچیده‌تر ولی با جواب قطعی می‌رسد. این بار از روش SVM با حاشیه سخت استفاده می‌کنم. اگر بتوان یک مدل SVM با حاشیه سخت پیدا کرد که به دقت ۱۰۰٪ برسد یعنی داده‌ها خطی جداپذیرند و اگر چنین مدلی وجود نداشته باشد یعنی خطی جداپذیر نیستند. در بررسی‌ای که من انجام دادم چنین مدلی پیدا نشد و لذا داده‌ها خطی جدا ناپذیرند.

سوال ۳

به عنوان اولین تلاش شبکه عصبی زیر را ایجاد کردم:

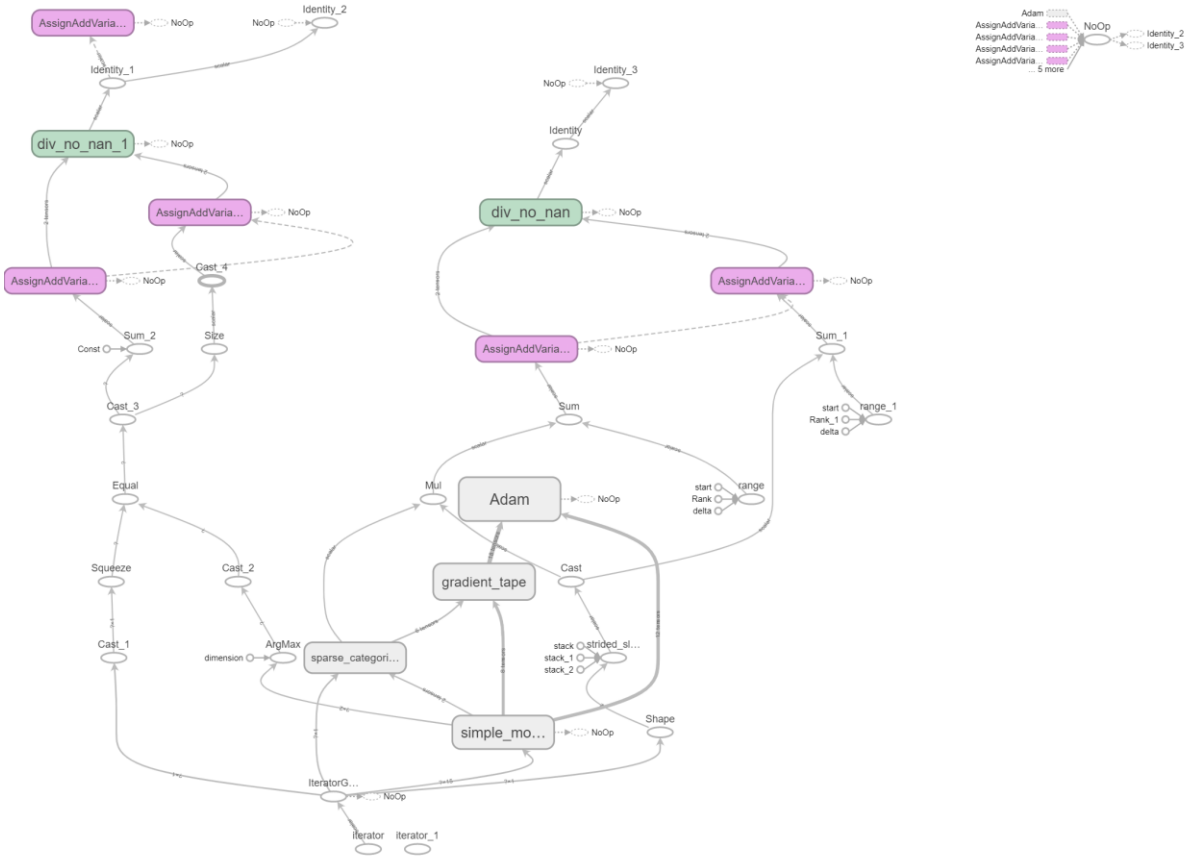


در این شبکه لایه ورودی شامل ۱۵ ویژگی است. سپس از این ۱۵ ویژگی توسط یک لایه Dense ۱۲۸ ویژگی استخراج می‌شود. در لایه Dense بعد ۱۶ ویژگی ترکیبی و پیچیده حاصل می‌شود و نهایتا در لایه خروجی دو ویژگی ایجاد می‌شود که هر کدام متناسب با احتمال تعلق داده به یکی از دو کلاس موجود است. از Optimizer آدام و از تابع خطای Sparse Categorical Cross Entropy کمک گرفته‌ام.

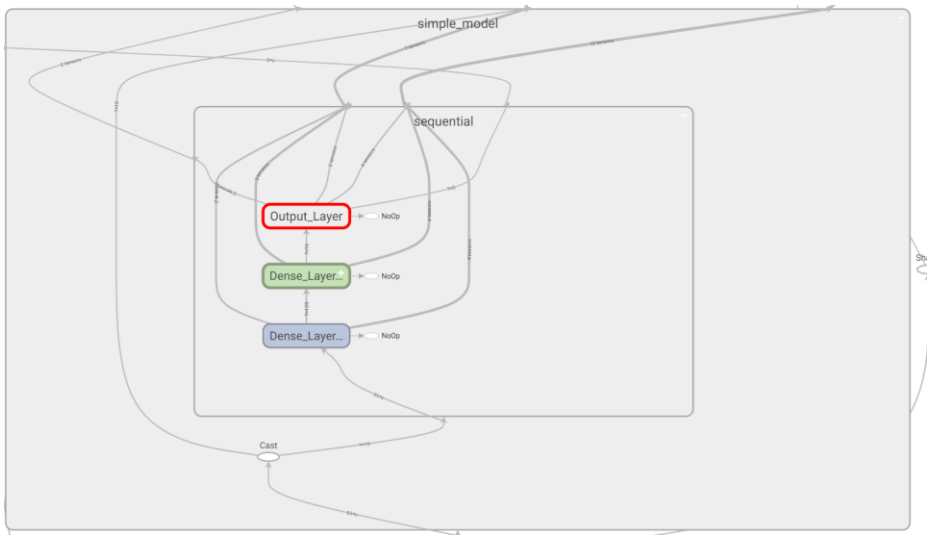
در لایه‌های Dense میانی از ReLU به عنوان تابع فعال‌سازی استفاده شده است تا عملکرد غیرخطی به مدل داده شود و در آخرین لایه که لایه خروجی باشد از یک لایه

Softmax استفاده شده است تا خروجی از جنس احتمال باشد و با تابع خطا استفاده می‌شود سازگار باشد. تعداد گام هم برابر با ۲۰ درنظر گرفته شده است.

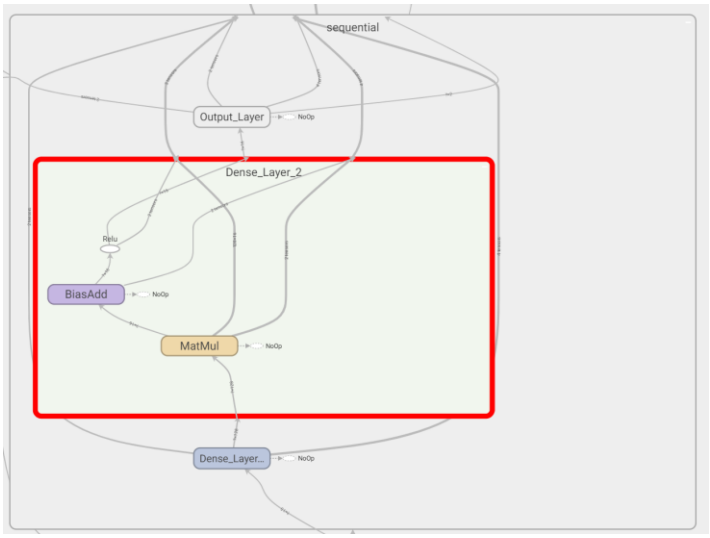
بعد از آموزش به صحت ۹۲٪ روی داده‌های آموزش و ۸۸٪ روی داده‌های اعتبارسنجی رسیدم. همچنین اگر تمایل داشته باشیم می‌توانیم اطلاعاتی را از روی tensor board هم ببینیم. مثلا گراف شبکه عبارت است از:



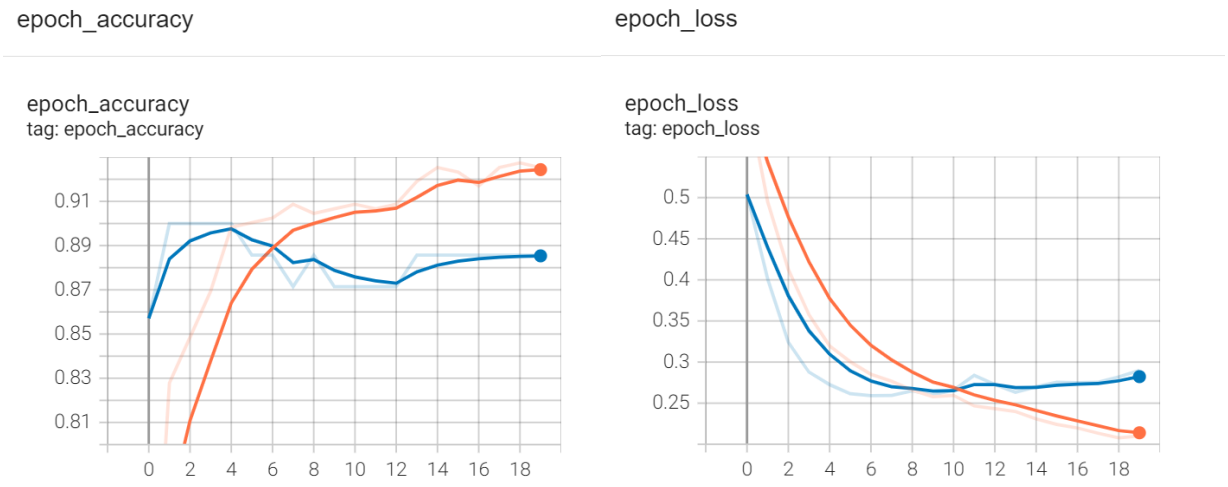
در قسمت پایین گراف کل مدل قرار گرفته است «simple_mo...» که با کلیک بر روی آن اطلاعات بیشتری راجع به خود مدل دریافت می‌کنیم:



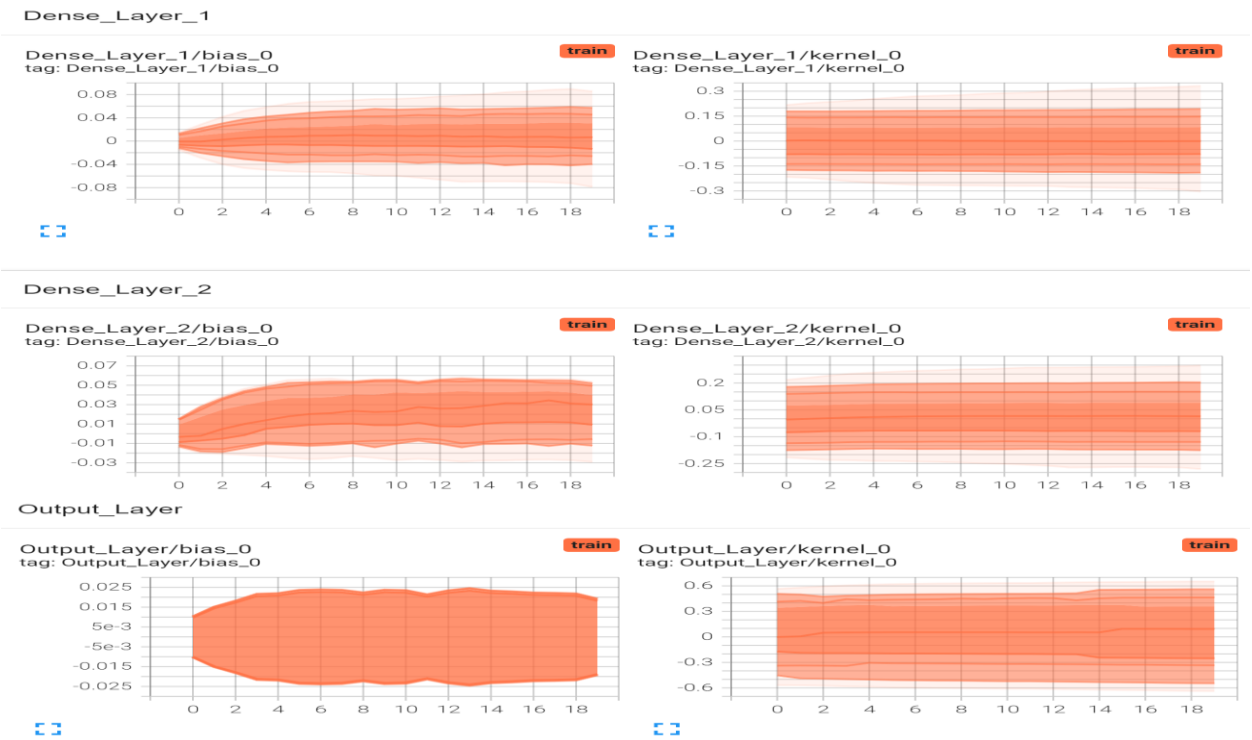
چنانچه روی یکی از لایه‌ها نظیر لایه وسط سبزرنگ کلیک کنیم اطلاع بیشتری راجع به خود لایه بدست می‌آوریم:



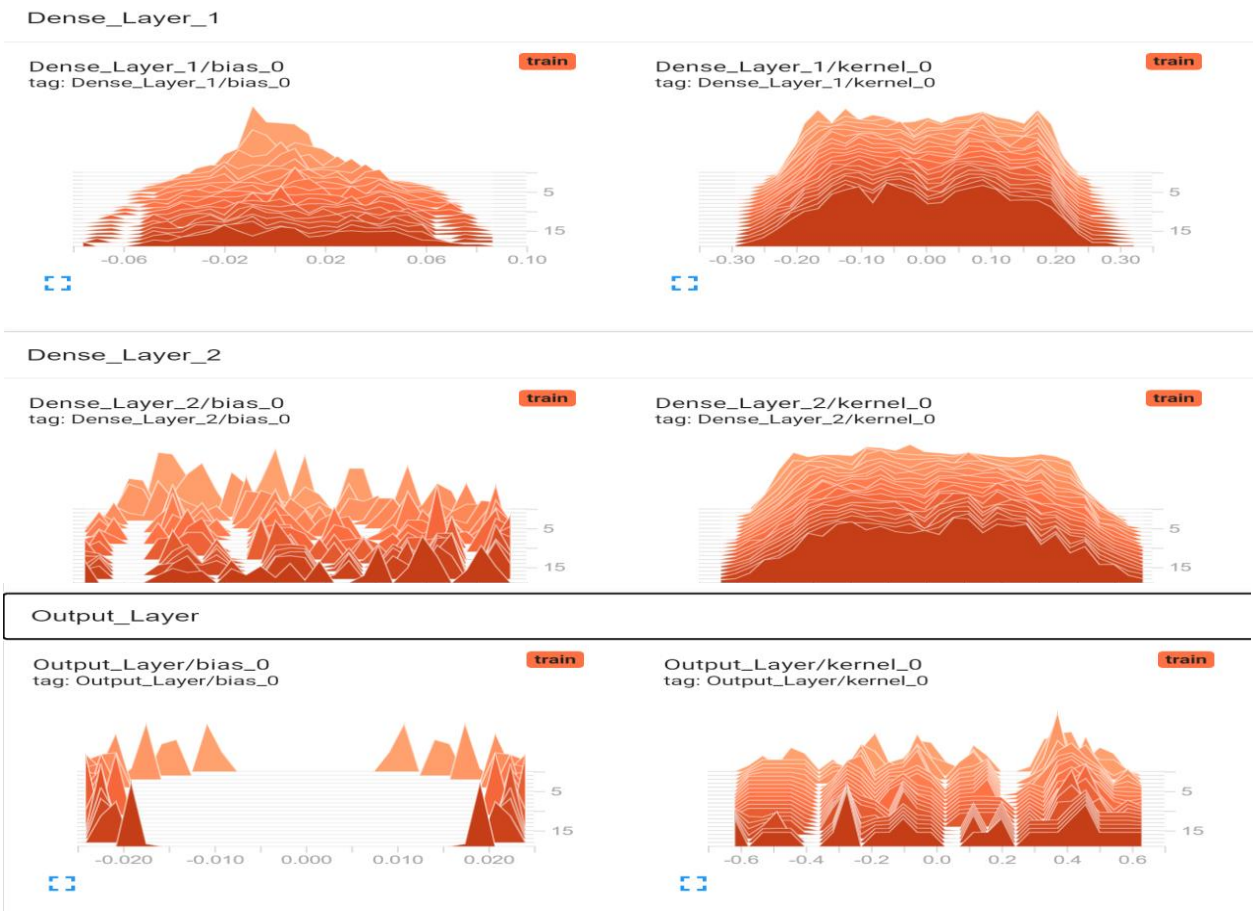
می‌توانیم بخش Scalars را هم بررسی کنیم:



متناسب با نتایج این قسمت به نظر می‌رسد که از گام ۱۲ به بعد مدل دچار بیش‌برازش شده است و بهتر بود که در همان قسمت یادگیری مدل متوقف شود. در قسمت بعدی tensor board یعنی Distribution نتایج زیر را می‌توان دید. نسبتا فراوانی مقدار پارامترها قابل قبول است و از این بابت مشکل جدی‌ای دیده نمی‌شود.



در قسمت Histogram نمودارهای زیر دیده می‌شود. به نظر bias هر سه لایه آموزش مناسبی دیده است و وزن‌های متنوعی را اختیار کرده‌اند. وزن‌های مربوط به کرنل خروجی هم قابل قبول است اما وزن‌های مربوط به کرنل دو لایه مخفی تغییر جدی‌ای پیدا نکرده‌اند که نکته مثبتی نیست.



سوال ۴

در شبکه ساده قسمت قبل دو لایه میانی با تعداد نورون ۱۲۸ و ۱۶ داشتیم. از شبکه قبل تنظیمات لایه خروجی، Optimizer، تابع خطا، نوع توابع فعال‌ساز لایه‌های میانی و تعداد گام آموزش را بدون تغییر نگه می‌دارم. تعداد لایه‌های میانی و تعداد نورون هر لایه را به عنوان پارامترهای ورودی در نظر می‌گیرم. برای این دو پارامتر طبیعتا بی‌نهایت عدد را می‌توان تست کرد و نمی‌توان به پارامترهای بهینه رسید؛ اما می‌توان

به یکی از جواب‌های نزدیک به حالت بهینه دست پیدا کرد. در جدول زیر ۵۷ اجرای مختلف آورده شده است. تعداد لایه‌های مخفی از ۰ تا ۷ متغیر است. توجه کنید که برای هر حالت تنها یک بار اجرا انجام شده است و کاملا محتمل است که در اجراهای بعد نتایج متفاوت باشد.

ردیف	تعداد لایه مخفی	تعداد نورون لایه‌های مخفی	صحت اعتبارسنجی
۱	۰	-	۸۸٫۵۷٪
۲	۱	۱۶	۸۸٫۵۷٪
۳	۱	۳۲	۸۴٫۲۹٪
۴	۱	۶۴	۸۴٫۲۹٪
۵	۱	۱۲۸	۸۸٫۵۷٪
۶	۱	۲۵۶	۸۵٫۷۱٪
۷	۲	۱۶-۱۶	۸۸٫۵۷٪
۸	۲	۱۶-۶۴	۸۴٫۲۹٪
۹	۲	۱۶-۱۲۸	۸۴٫۲۹٪
۱۰	۲	۶۴-۱۶	۸۵٫۷۱٪
۱۱	۲	۶۴-۶۴	۸۷٫۱۴٪
۱۲	۲	۶۴-۱۲۸	۸۴٫۲۹٪
۱۳	۲	۱۲۸-۱۶	۸۵٫۷۱٪
۱۴	۲	۱۲۸-۶۴	۸۲٫۸۶٪
۱۵	۲	۱۲۸-۱۲۸	۸۸٫۵۷٪
۱۶	۳	۲۵۶-۱۲۸-۶۴	۸۵٫۷۱٪
۱۷	۳	۱۲۸-۶۴-۳۲	۸۵٫۷۱٪
۱۸	۳	۶۴-۳۲-۱۶	۸۸٫۵۷٪
۱۹	۳	۶۴-۱۲۸-۲۵۶	۸۵٫۷۱٪
۲۰	۳	۳۲-۶۴-۱۲۸	۸۷٫۱۴٪
۲۱	۳	۱۶-۳۲-۶۴	۸۵٫۷۱٪
۲۲	۳	۱۲۸-۱۲۸-۱۲۸	۸۸٫۵۷٪
۲۳	۳	۱۶-۱۶-۱۶	۸۴٫۲۹٪

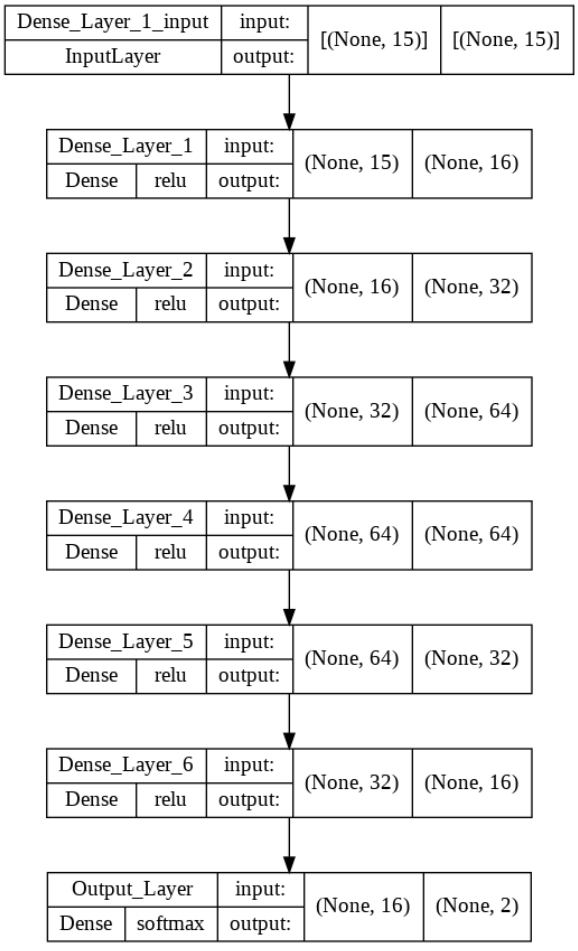
۲۴	۳	۱۲۸-۱۶-۱۲۸	۸۵/۷۱٪
۲۵	۳	۱۶-۱۲۸-۱۶	۸۷/۱۴٪
۲۶	۳	۱۲۸-۶۴-۳۲-۱۶	۸۴/۲۹٪
۲۷	۴	۲۵۶-۱۲۸-۶۴-۳۲	۸۲/۸۶٪
۲۸	۴	۲۵۶-۶۴-۳۲-۸	۸۷/۱۴٪
۲۹	۴	۱۶-۳۲-۶۴-۱۲۸	۸۵/۷۱٪
۳۰	۴	۳۲-۶۴-۱۲۸-۲۵۶	۸۲/۸۶٪
۳۱	۴	۸-۳۲-۶۴-۲۵۶	۸۵/۷۱٪
۳۲	۴	۲۵۶-۲۵۶-۲۵۶-۲۵۶	۸۵/۷۱٪
۳۳	۴	۱۶-۱۶-۱۶-۱۶	۸۸/۵۷٪
۳۴	۵	۱۲۸-۶۴-۳۲-۱۶-۸	۸۱/۴۳٪
۳۵	۵	۲۵۶-۱۲۸-۶۴-۳۲-۱۶	۸۴/۲۹٪
۳۶	۵	۲۵۶-۱۲۸-۶۴-۶۴-۱۶-۸	۸۵/۷۱٪
۳۷	۵	۸-۱۶-۳۲-۶۴-۱۲۸	۸۲/۸۶٪
۳۸	۵	۱۶-۳۲-۶۴-۱۲۸-۲۵۶	۸۵/۷۱٪
۳۹	۵	۸-۱۶-۶۴-۱۲۸-۲۵۶	۸۵/۷۱٪
۴۰	۵	۱۲۸-۱۲۸-۱۲۸-۱۲۸-۱۲۸	۸۷/۱۴٪
۴۱	۵	۱۶-۱۶-۱۶-۱۶-۱۶	۸۴/۲۹٪
۴۲	۵	۱۶-۳۲-۶۴-۳۲-۱۶	۸۷/۱۴٪
۴۳	۵	۶۴-۳۲-۱۶-۳۲-۶۴	۹۰/۰٪
۴۴	۶	۱۲۸-۶۴-۳۲-۱۶-۸-۴	۸۷/۱۴٪
۴۵	۶	۲۵۶-۱۲۸-۶۴-۳۲-۱۶-۸	۸۲/۸۶٪
۴۶	۶	۴-۸-۱۶-۳۲-۶۴-۱۲۸	۸۵/۷۱٪
۴۷	۶	۸-۱۶-۳۲-۶۴-۱۲۸-۲۵۶	۸۷/۱۴٪
۴۸	۶	۱۶-۳۲-۶۴-۶۴-۳۲-۱۶	۹۱/۴۳٪
۴۹	۶	۶۴-۳۲-۱۶-۱۶-۳۲-۶۴	۸۷/۱۴٪
۵۰	۶	۱۲۸-۱۲۸-۱۲۸-۱۲۸-۱۲۸-۱۲۸	۸۷/۱۴٪
۵۱	۶	۱۶-۱۶-۱۶-۱۶-۱۶-۱۶	۸۵/۷۱٪
۵۲	۷	۲۵۶-۱۲۸-۶۴-۳۲-۱۶-۸-۴	۸۸/۵۷٪
۵۳	۷	۴-۸-۱۶-۳۲-۶۴-۱۲۸-۲۵۶	۸۰/۰٪

۵۴	۷	۱۶-۳۲-۶۴-۱۲۸-۶۴-۳۲-۱۶	۸۷/۱۴٪
۵۵	۷	۱۲۸-۶۴-۳۲-۱۶-۳۲-۶۴-۱۲۸	۹۰/۰٪
۵۶	۷	۱۲۸-۱۲۸-۱۲۸-۱۲۸-۱۲۸-۱۲۸-۱۲۸	۸۷/۱۴٪
۵۷	۷	۱۶-۱۶-۱۶-۱۶-۱۶-۱۶-۱۶	۹۰/۰٪

در این جدول ۴ شبکه برتر عبارت است از:

ردیف	تعداد لایه مخفی	تعداد نورون لایه‌های مخفی	صحت اعتبارسنجی
۴۸	۶	۱۶-۳۲-۶۴-۶۴-۳۲-۱۶	۹۱/۴۳٪
۴۳	۵	۶۴-۳۲-۱۶-۳۲-۶۴	۹۰/۰٪
۵۵	۷	۱۲۸-۶۴-۳۲-۱۶-۳۲-۶۴-۱۲۸	۹۰/۰٪
۵۷	۷	۱۶-۱۶-۱۶-۱۶-۱۶-۱۶-۱۶	۹۰/۰٪

معماری برترین شبکه یعنی مدل ۴۸-ام به صورت مقابل است. با این مدل می‌توان به صحت این مدل روی داده‌های آموزش برابر با ۸۵/۵۱٪ درصد است. که ۵ درصد کمتر از صحت اعتبارسنجی است ولی باز فاصله زیادی با آن ندارد.



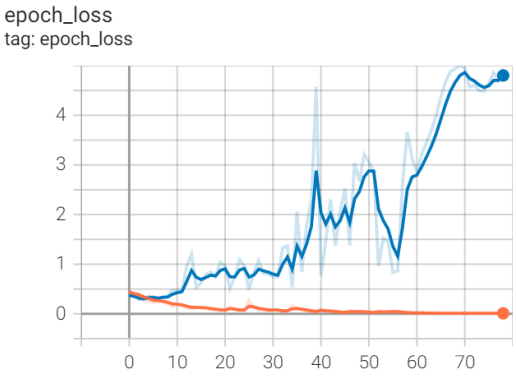
با بررسی جدول نتایج موجود در گزارش می‌توان دید که مدل‌هایی که یک لایه مخفی دارند یا اصلا لایه مخفی‌ای ندارند به طور کلی صحت پایینی دارند و مدل‌هایی که ۷ لایه دارند (بر خلاف انتظار من!) از نظر میانگین بهترین نتایج را داشته‌اند. اولین لایه برخی از مدل‌ها دارای تعداد نورون پایینی مانند ۴ نورون است (مدل ۴۶ یا ۵۳). در

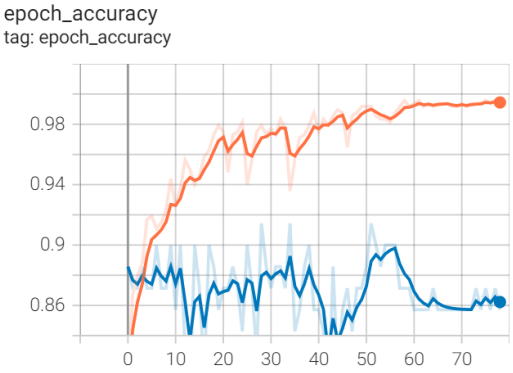
این موارد بخش مهمی از دانش شبکه در اولین لایه از بین می‌رود و خروجی هم مناسب نخواهد بود. همچنین برخی از مدل‌های نسبتا پیچیده مانند مدل ۳۲ تنها دارای صحت ۸۵٪ روی مجموعه اعتبارسنجی هستند ولی صحت ۹۹٪ای روی مجموعه آموزشی دارند! این نشان می‌دهد که این مدل‌های پیچیده بر روی مجموعه آموزشی بیش‌برازش شده است که در سوال بعدی دقیق‌تر بررسی می‌شوند. نهایتا باید توجه کرد که حتی مدل ۷ یا ۱۵ با داشتن دو لایه مخفی صحتی نزدیک به بهترین صحت داشته است (۸۸/۵۷٪). این نشان می‌دهد که برای این مسئله حتی با مدل‌های ساده هم می‌توان به صحت قابل قبولی رسید و اگر قرار باشد در مصالحه پیچیدگی مدل، زمان اجرا و صحت یک مدل را انتخاب کنیم، قطعا این مدل دو لایه بهتر از مدل‌های شش هفت لایه خواهد بود.

سوال ۵

برای دست‌یابی به یک مدل بیش‌برازش شده باید یک شبکه پیچیده با تعداد پارامتر بسیار زیاد در نظر بگیریم تا به یک صحت زیاد روی داده‌های آموزش ولی صحت کم روی داده‌های اعتبارسنجی برسیم. مثلا یک شبکه با هشت لایه مخفی که در هر لایه ۵۱۲ نورون داشته باشد به نظر مناسب می‌آید. سایر تنظیمات مانند دو سوال قبل است با این تفاوت که ۲۰۰ گام برای آموزش درنظر گرفته‌ام تا مدل زمان کافی برای بیش‌برازش شدن روی داده‌های آموزش را داشته باشد. نهایتا آنکه از یک call back از نوع Early stopping روی صحت آموزشی استفاده کرده‌ام.

پس از آموزش، صحت مدل روی داده‌های آموزشی برابر با ۹۹/۱۷٪ است که عدد بسیار مناسبی است اما صحت همین مدل روی داده‌های اعتبارسنجی و تست که در زمان آموزش دیده نشده است به ترتیب عبارت است از ۸۷/۱۴٪ و ۸۱/۱۶٪. اگر مقدار خطا را در نظر بگیریم اختلاف شدیدتر هم می‌شود؛ مقدار خطا بر روی داده‌های آموزشی، اعتبارسنجی و تست به ترتیب برابر است با ۰/۰۱۳۰، ۳/۶۶۲۳ و ۵/۲۶۴۳. این‌ها همه حاکی از آن است که مدل بیش‌برازش شده است. چرا که یک مدل بیش‌برازش‌شده روی مجموعه‌داده آموزشی دقت بسیار خوبی دارد ولی روی مجموعه‌داده تست خیر. این مسئله در قسمت tensor board هم به خوبی عیان است:



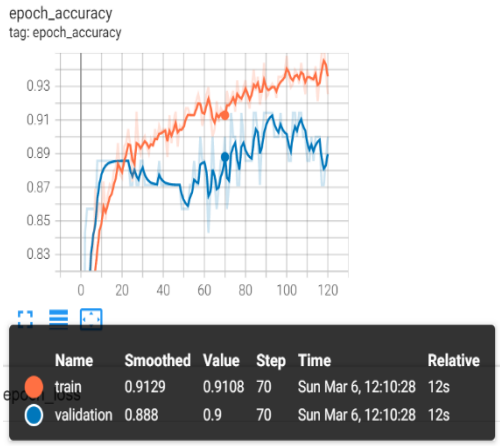


همچنین می‌دانیم مدل بیش‌برازش‌شده دارای تعداد پارامتر زیاد است. باتوجه به اینکه مدل حتی می‌تواند با دو لایه ساده با تعداد نورون نسبتا کم مخفی می‌تواند به صحت مناسبی برسد طبیعی است که این تعداد از لایه‌های با تعداد نورون بالا از نظر تئوری مدل را بیش از حد پیچیده می‌کند.

سوال ۶

برای آنکه تعمیم‌پذیری مناسب داشته باشیم شرطش آن است که پیچیدگی مدل مناسب باشد و صحت مدل روی مجموعه اعتبارسنجی بالا باشد. لذا یکی از مدل‌های مناسب سوال ۴ مانند مدل هفت لایه شماره ۵۵ را انتخاب کردم چراکه مطمئن هستیم دقت روی داده‌های اعتبارسنجی بالایی دارد. (علت عدم انتخاب مدل ۴۸ صحت پایین آن در اجرای مجدد بوده است!)؛ در عین حال در میان لایه‌های مخفی، لایه‌های Dropout برای تعمیم‌پذیری بیشتر اضافه کردم. از یک callback از نوع Early stopping و بر روی خطای اعتبارسنجی کمک گرفتم تا مطمئن شوم مدل زیاد از حد روی داده‌های آموزش پیش نمی‌رود.

پس از آموزش، مقدار accuracy بر روی سه مجموعه آموزشی، اعتبارسنجی و تست به ترتیب عبارت است از: ۹۴/۴۰٪، ۹۰٪ و ۸۴/۷۸٪. برای اطمینان به سراغ tensor board هم می‌رویم:



باتوجه به آنکه مقدار ۵۰ برای patient تنظیم شده است، خروجی فاز آموزش، خروجی مدل در گام ۷۰ خواهد بود. در این گام مقدار خطا در یک مینیمم مناسب قرار دارد و مقدار Accuracy هم قابل قبول است. اما اگر از callback استفاده نمی‌کردیم خروجی گام‌های آخر را می‌داشتیم که دچار بیش‌برازش شده است چراکه در گام ۱۲۰ خطای آموزش شدیدا کاهش یافته است ولی میزان خطای اعتبارسنجی در حال افزایش بوده است. در گام‌های آخر میزان accuracy داده‌های اعتبارسنجی هم دیگر از روال افزایشی خود خارج شده است.