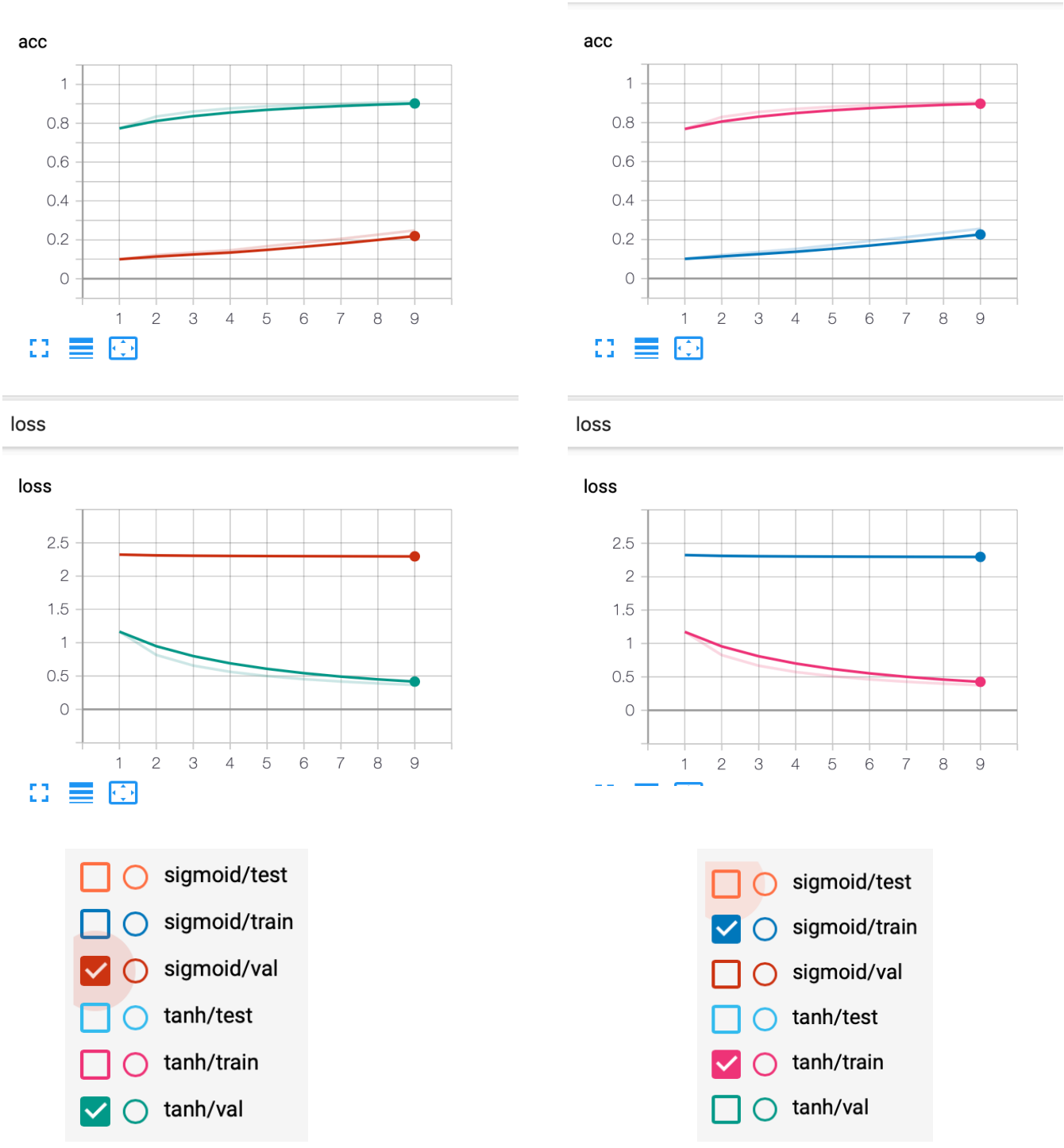
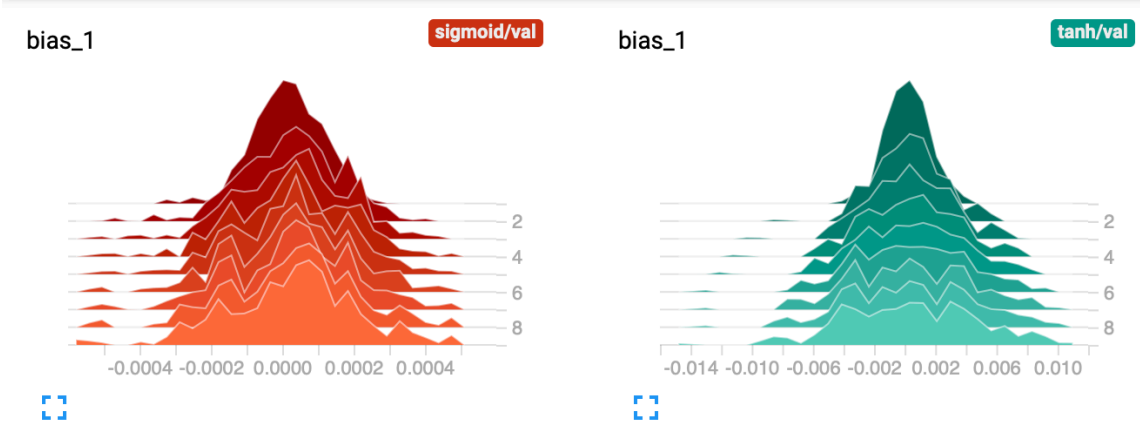
سوال اول)

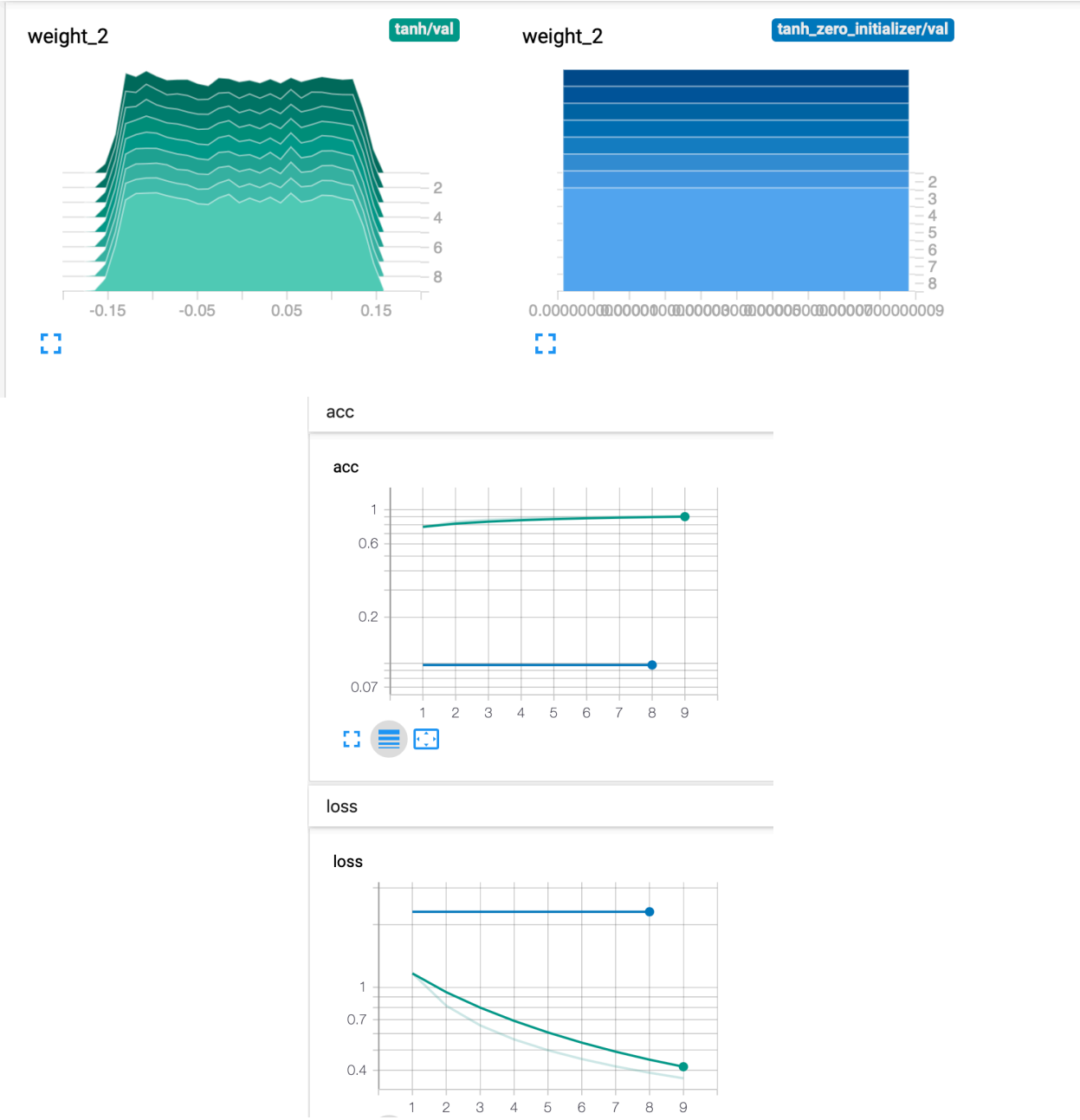

acc
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9
acc
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9
loss
loss
2.5
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9
loss
2.5
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9
sigmoid/test
sigmoid/train
sigmoid/val
tanh/test
tanh/train
tanh/val
sigmoid/test
sigmoid/train
sigmoid/val
tanh/test
tanh/train
tanh/val



همانطور که مشاهده می کنید و نتایج عددی تست tanh خیلی بهتر عمل کرد. در تمرین ۲ به تفصیل توضیح داده شد که چرا این تابع فعالیت از sigmoid می تواند بهتر عمل کند. وزن ها هم در کل در tanh کمی واریانس کمتری داشتند و نزدیک به همتر بودند.

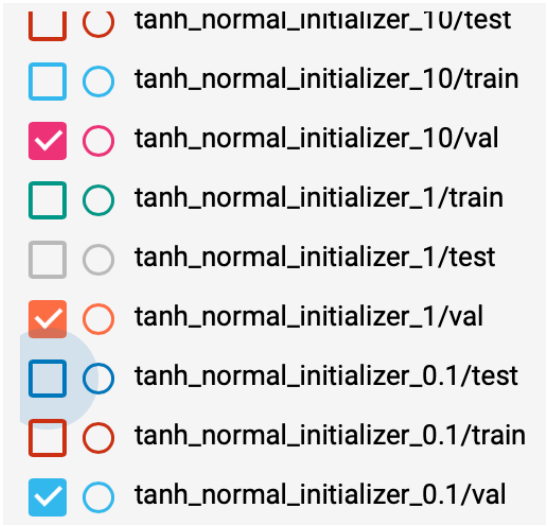
ه

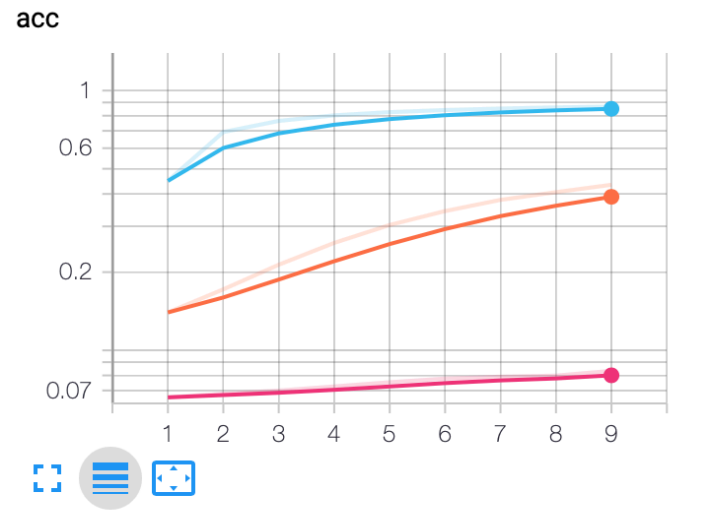
سوال دوم)



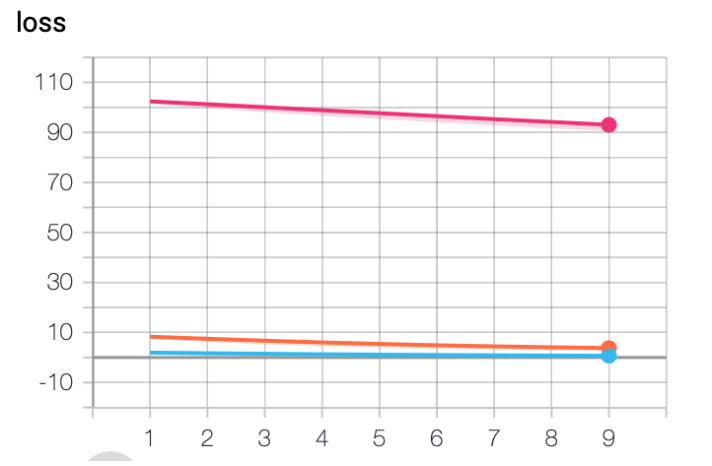
همانطور که می بینید چون وضعیت اولیه همه ی نورو ها یکی بوده، همه شبیه هم عمل کرده و متقارن مثل هم مانده اند و وزن ها به صورت یکنواخت پراکندگی دارند و نورن ها نتوانسته اند متفاوت شوند. از روی نمودار دقت و لاس می عملکرد ضعیف را می توان فهمید.

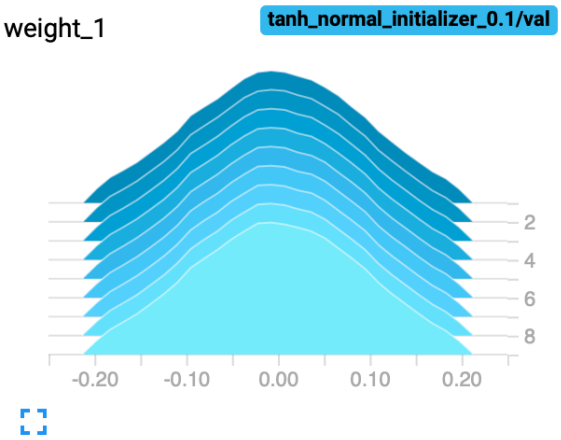
سوال سوم)

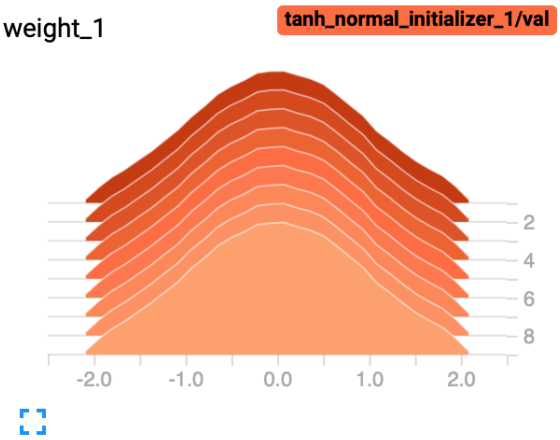


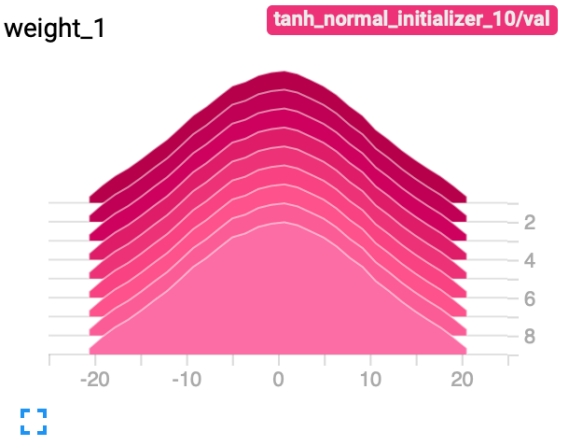


loss





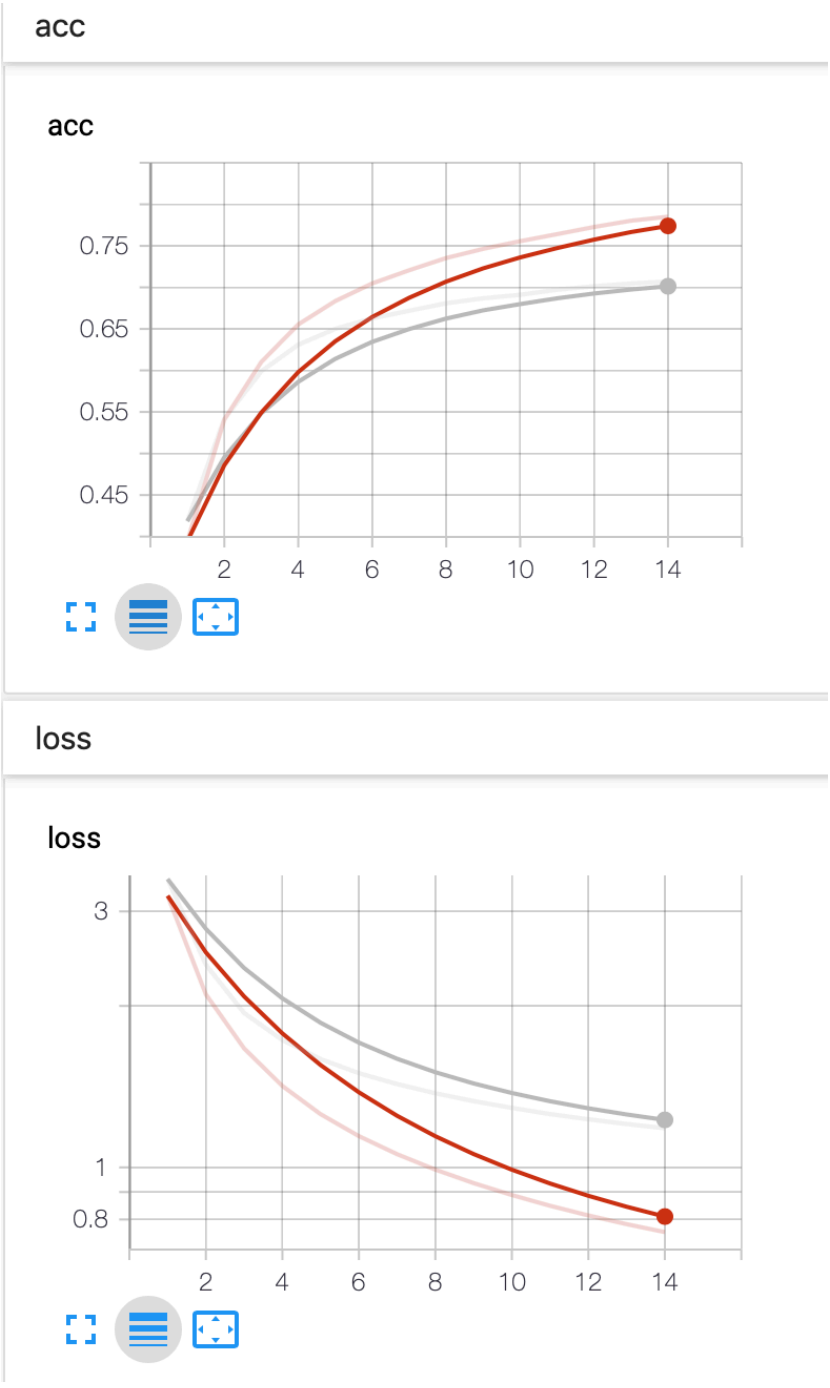


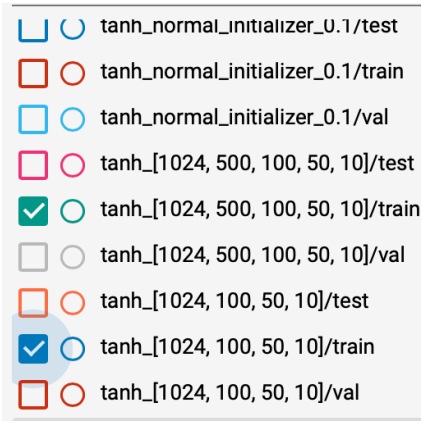


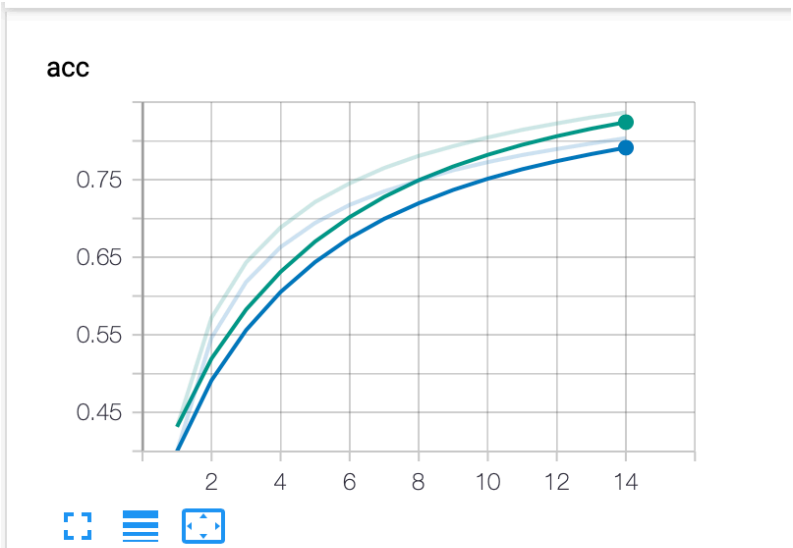
همانطور که مشاهد می کنید هرچه استاندارد دویژن کمتر بود، عملکرد شبکه بهتر شد و همچنین پراکندگی وزن ها واقعا مانند نمودار نرمال شده و با توجه به استاندارد دویژن هر مورد، پراکنده تر یا جمع تر هستند.(دو برابر استاندارد دویوژن هر سناریو از مرکز تا چپ و راست، تقریبا کل جمعیت است ۹۶٪ ...)

سوال چهارم)

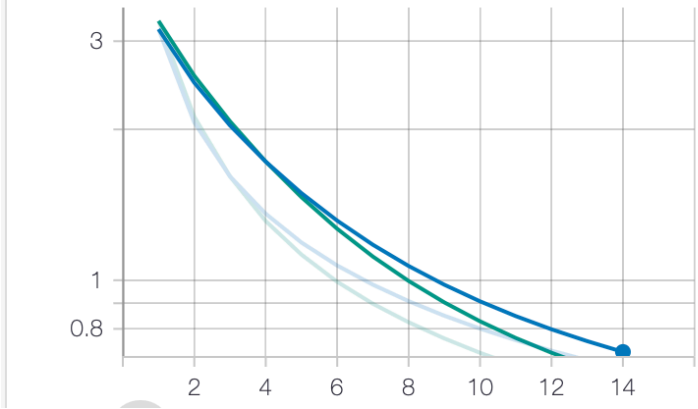

tanh_normal_initializer_0.1/test
tanh_normal_initializer_0.1/train
tanh_normal_initializer_0.1/val
tanh_[1024, 500, 100, 50, 10]/test
tanh_[1024, 500, 100, 50, 10]/train
tanh_[1024, 500, 100, 50, 10]/val
tanh_[1024, 100, 50, 10]/test
tanh_[1024, 100, 50, 10]/train
tanh_[1024, 100, 50, 10]/val


acc
acc
0.75
0.65
0.55
0.45
2
4
6
8
10
12
14
loss
loss
3
1
0.8
2
4
6
8
10
12
14

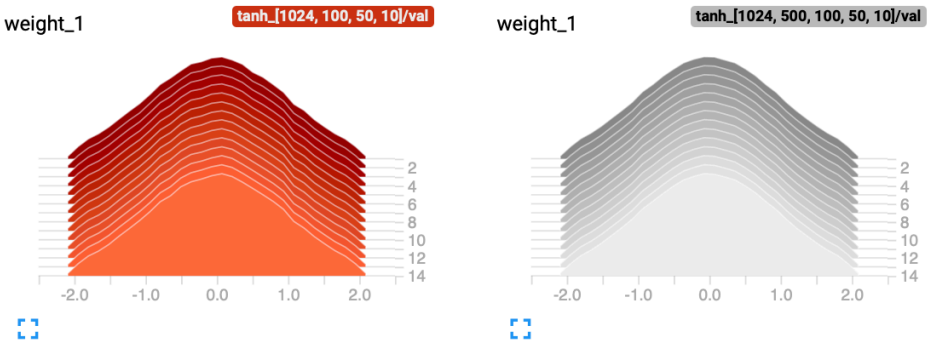

tanh_normal_initializer_0.1/test
tanh_normal_initializer_0.1/train
tanh_normal_initializer_0.1/val
tanh_[1024, 500, 100, 50, 10]/test
tanh_[1024, 500, 100, 50, 10]/train
tanh_[1024, 500, 100, 50, 10]/val
tanh_[1024, 100, 50, 10]/test
tanh_[1024, 100, 50, 10]/train
tanh_[1024, 100, 50, 10]/val


acc
0.75
0.65
0.55
0.45
2
4
6
8
10
12
14

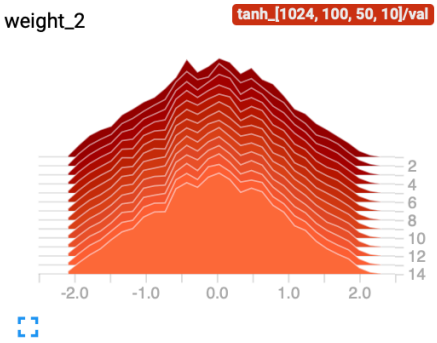
loss

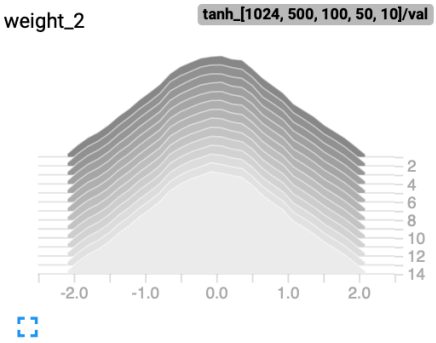

loss
3
1
0.8
2
4
6
8
10
12
14

weight_1

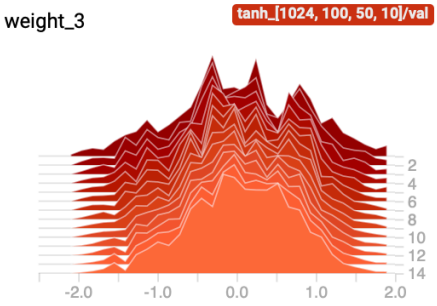


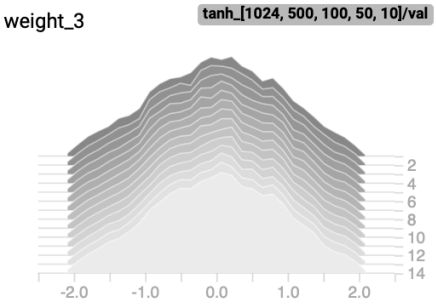
weight_2



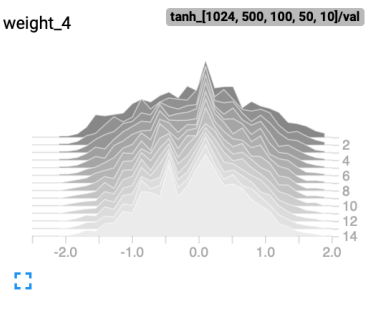


weight_3





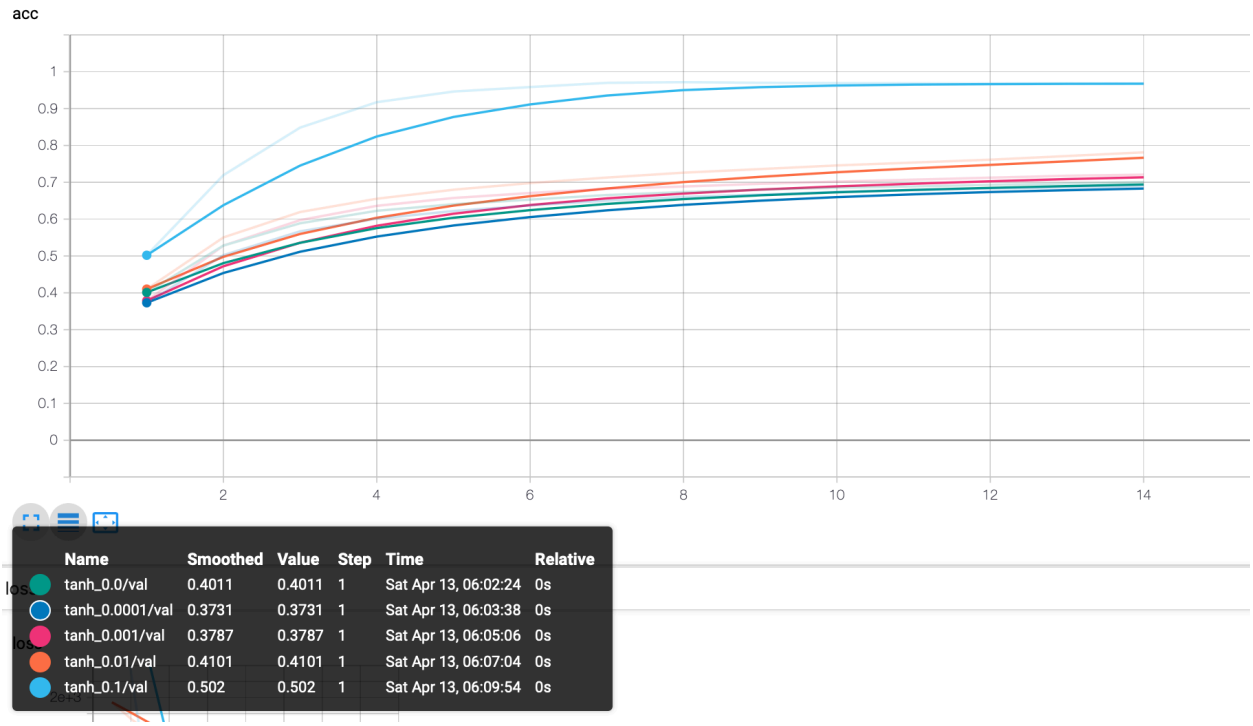
weight_4



مدل پیچیده تر روی داده ی آموزش بهتر عمل کرد اما مدل ساده تر روی داده ی تست و ولیدیشن بهتر بود، به عبارت دیگر مدل پیچیده روی داده ی آموزش اورفیت شد و فضای فرضیه اش خیلی بزرگتر بود و به فرضیه مناسبی نرسید(در مقایسه با دیگری)
از روی وزن ها من این را برداشت می کنم لایه های آخری در هر دو مدل نمودار پراکندگیشان از حالت نمودار نرمال در آمده است و نموداری پیچیده و دندانه دار شده اند، اما مدل پیچیده تر، لایه اولش هنوز به فرم پراکندگی نرمال است و انگار نورون های لایه اولش به اندازه ی کافی داده ترین نشده اند تا هرکدام ویژگی مند و کاراکترستیک شوند، به عبارت دیگر انگار لایه اول در مدل پیچیده هنوز به خوبی train نشده است. اینکه چرا این اتفاق برای لایه اول و نه لایه آخر افتاده ، حدس می زنم دلیلش جهت برعکس روند backprop باشد. انگار جریان backprop چون اول از آنها رد می شود سریعتر یاد می گیردند و اثر backrop وقتی به لایه های اولی می رسد به نوعی ضعیفتر می شود و سخت تر یادگیری می شود.

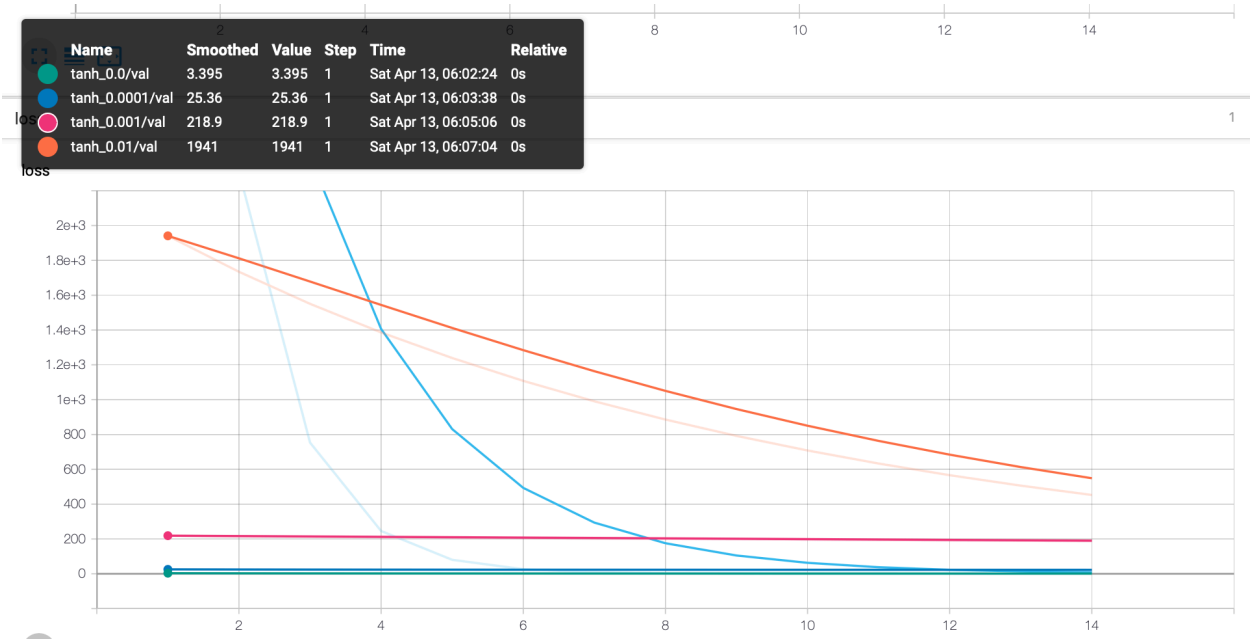
سوال پنجم)

Accuracy :



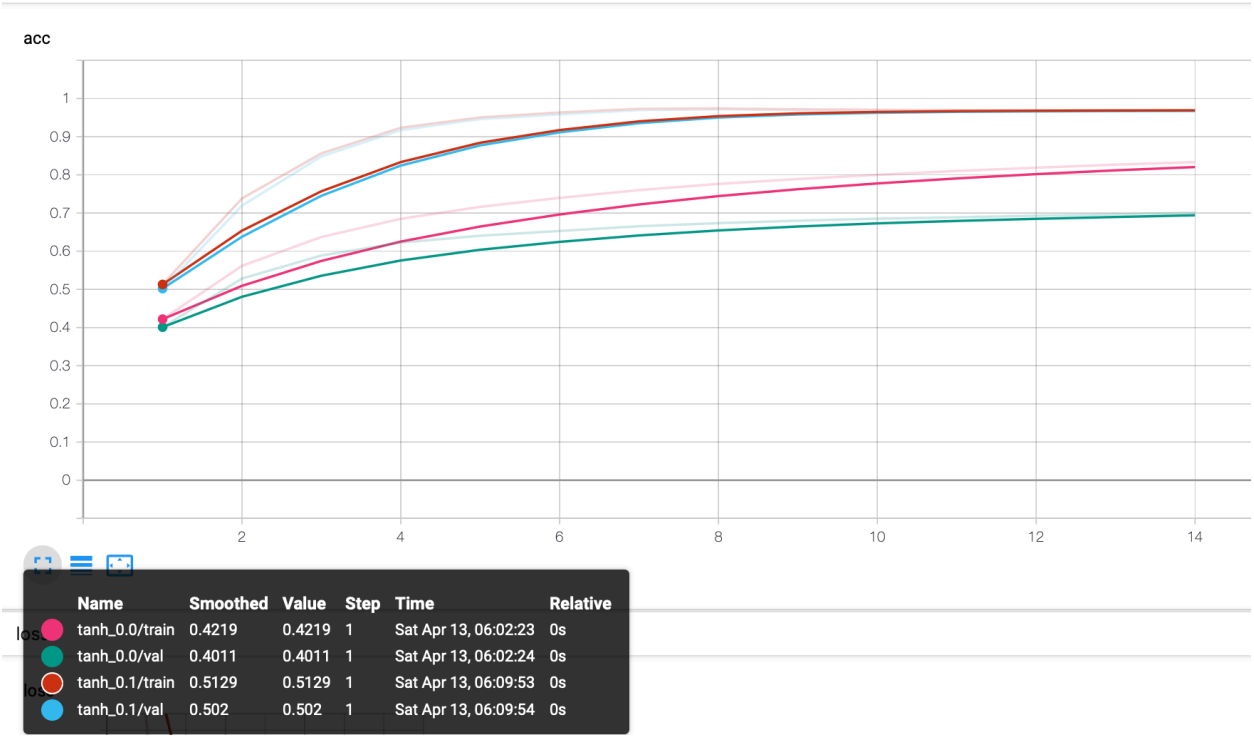
Name	Smoothed	Value	Step	Time	Relative
tanh_0.0/val	0.4011	0.4011	1	Sat Apr 13, 06:02:24	0s
tanh_0.0001/val	0.3731	0.3731	1	Sat Apr 13, 06:03:38	0s
tanh_0.001/val	0.3787	0.3787	1	Sat Apr 13, 06:05:06	0s
tanh_0.01/val	0.4101	0.4101	1	Sat Apr 13, 06:07:04	0s
tanh_0.1/val	0.502	0.502	1	Sat Apr 13, 06:09:54	0s

Loss:



Name	Smoothed	Value	Step	Time	Relative
tanh_0.0/val	3.395	3.395	1	Sat Apr 13, 06:02:24	0s
tanh_0.0001/val	25.36	25.36	1	Sat Apr 13, 06:03:38	0s
tanh_0.001/val	218.9	218.9	1	Sat Apr 13, 06:05:06	0s
tanh_0.01/val	1941	1941	1	Sat Apr 13, 06:07:04	0s

Differenc between Validation and Train:



از آنجا که مقدار رگولاریزشن در لاس موثر است، هرچه ضریب بیشتر شود، لاس بیشتر می شود و لذا نمی توانیم ضرایب مختلف را از روی میزان لاسشان باهم مقایسه کنیم تا ضریب بهتر را پیدا کنیم.
اما در کل هرچه ضریب رگولاریزشن بیشتر بود، لاس به سرعت بیشتر کاهش پیدا می کرد و دلیلش حساسیت بیشتر لاس به اندازه ضرایب است که با کمتر کردن اندازه ها لاس قابل توجه کم می شد.

از نظر دقت ضرایب رگولاریزشن بزرگتر، بهتر عمل کردند.
نکته ی دیگر، اینکه هرچه ضریب رگولاریزشن کمتر بود، دقت روی train و validation به مرور تفاوت بیشتری می گرفت و این بیانگر تاثیر رگولاریزشن روی کاهش overfit است.
و نکته ي آخر اینکه l2 تلاش می کند اندازه ی وزن ها را کمتر کند و در عکس این مساله قابل مشاهده است.