به نام خدا

تمرین سری ششم درس یادگیری عمیق

دکتر محمدی

محمد یارمقدم

۹۶۴۶۲۱۰۴

سوال اول)

① الف)

Swish function:

$$f(x) = \text{Sigmoid}(x) \times x = \frac{x}{1+e^{-x}}$$

$$\text{Sigmoid}(x) = \frac{1}{1+e^{-x}}$$

Mish function:

$$f(x) = x \times \tanh(\text{softplus}(x)) = x \times \tanh(\ln(1+e^{x}))$$

$$\text{Softplus}(x) = \ln(1+e^{x})$$

ب)

derivitate of swish function:

$$f'(x) = \frac{\partial}{\partial x}(x \cdot \sigma(x)) = \sigma(x) + x\sigma'(x) = x\left[\sigma(x)\left(1-\sigma(x)\right)\right]$$

$$= x\sigma(x) + \sigma(x)\left(1 - x\sigma(x)\right) \Longrightarrow f'(x) = f(x) + \sigma(x)\left(1-\sigma(x)\right)$$

derivitate of Mish function:

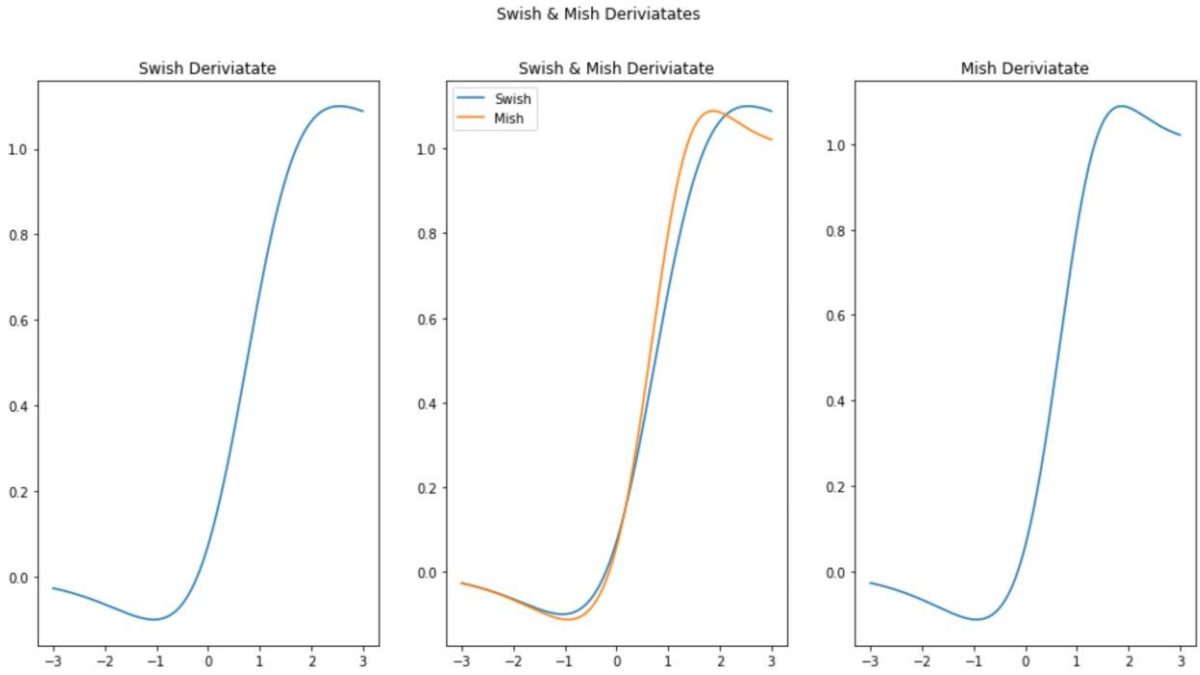
$$f'(x) = \frac{\partial}{\partial x}\left(x \cdot \tanh(\ln(1+e^{x}))\right) = x \times \frac{e^{x}}{1+e^{x}} \text{Sec}^{2}h\left(\ln(1+e^{x})\right)$$

$$+ \tanh(\ln(1+e^{x}))$$

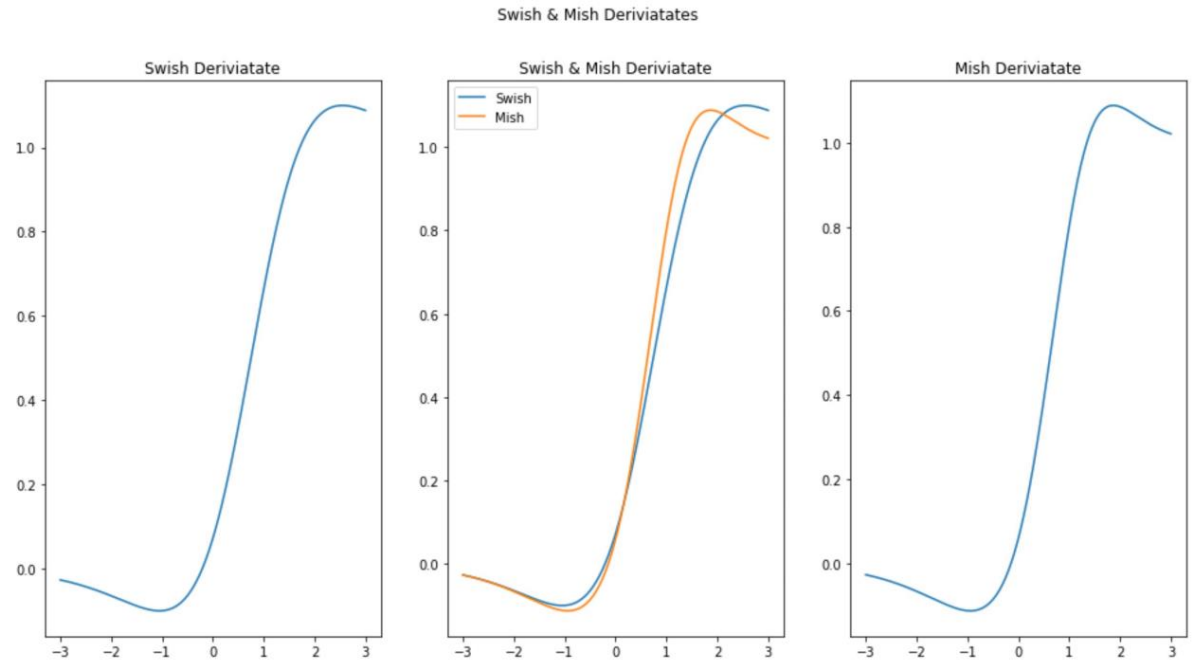
$$\Longrightarrow f'(x) = \frac{f(x)}{x} + \sigma(x) \times x \times \text{Sec}^{2}h\left(\ln(1+e^{x})\right)$$

$$= \frac{f(x)}{x} + \sigma(x) \times x \times \text{Sec}^{2}h\left(\text{softplus}(x)\right)$$

نمودار های قسمت الف



نمودار های قسمت ب



ت)

تابع relu در مقایسه دو تابع sigmoid و tanh دارای مزایای زیر است:

- از نظر پیچیدگی محاسباتی بهینه تر و ساده تر است.
- پراکنده است. سیگموید به احتمال زیادی مقادیر غیر صفر تولید میکند و منجر به نمایش متراکم دیتا میشود. نمایش پراکنده داده ها به وسیله relu مناسب تر است.
- یکی از مزایای مهم، کاهش احتمال vanish شدن گرادیان است.
- تابع relu بر خلاف دو تابع دیگر محدودیت مقدار از بالای محور y را ندارد. این امر از اشباع شدن گرادیان جلوگیری می کند.
- از آنجا که اگر گرادیان ورودی مثبت باشد، در این شبکه می تواند زیاد شود پس بهینگی آن بیشتر و سرعت و راحتی آن در نتیجه بیشتر است.

شباهت های relu با این توابع:

- این دو تابع همانند relu از بالا محدودیت ندارند که باعث جلوگیری از اشباع شدن گرادیان میشود.
- در مقادیر بزرگ مثبت این دو تابع روند و عملکرد مشابهی دارند.

مزیت های دو تابع نسبت به relu:

- دارای ویژگی smoothing وابستگی این دو تابع نسبت به relu به وزن های اولیه شبکه و نرخ یادگیری که مقدار دهی شده اند کمتر است. پس در این توابع سریع تر میتوان به نقطه بهینه رسید.
- به علت ویژگی non-monotonicity این تابع مقادیر منفی کوچکی تولید می کند که باعث گرادیان در مقادیر منفی هم فعالیت داشته باشد. در حالیکه در relu گرادیان در مقادیر منفی صفر می شود و ناپدید می شود.
- قابلیت تعمیم دارند.

ث)

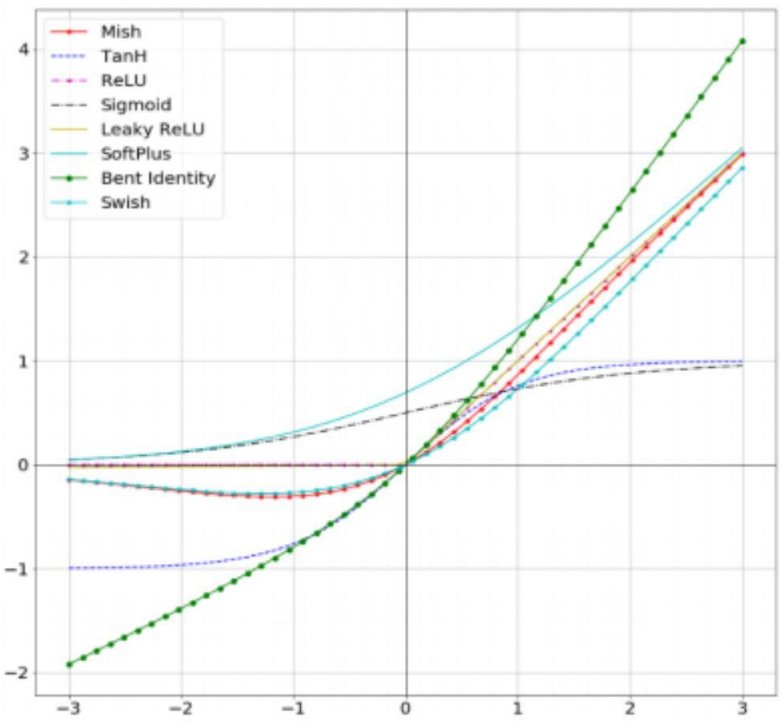
مزایای Swish نسبت به relu:

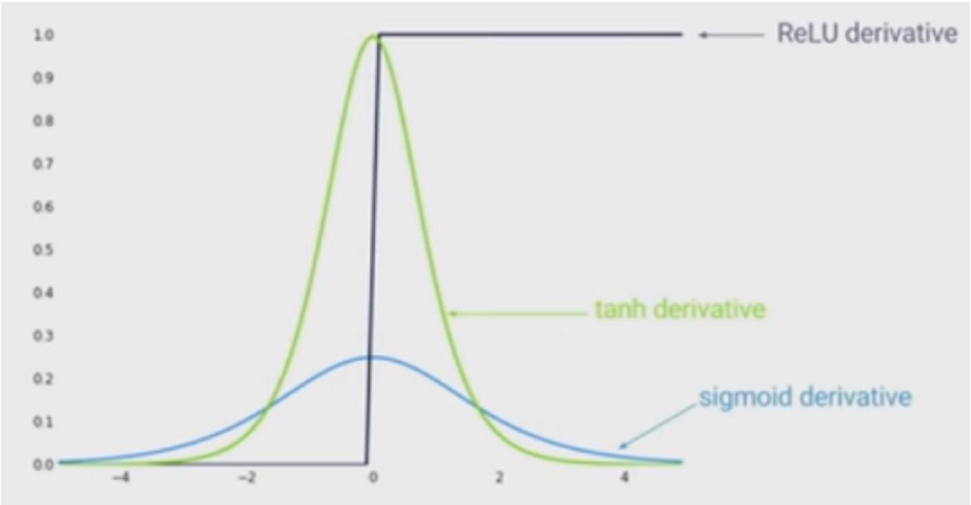
- در شبکه های خیلی عمیق، این تابع باعث می شود به دقت بیشتری داده های تست برسیم.
- در هر batch تابع Swish نبست به relu بهتر عمل می کند.
- محدود بودن در مقادیر زیر صفر شباهت است. این نیز مزیت محسوب می شود زیرا باعص اثرات منظم قوی می شود و از همه مهم تر باعث کاهش overfitting می شود.
- محدود بودن در مقادیر زیر صفر شباهت است. این نیز مزیت محسوب می شود زیرا باعص اثرات منظم قوی می شود و از همه مهم تر باعث کاهش overfitting می شود.
- وجود متغیر بتا در تعیین نوع خروجی می تواند اختیار را به ما بدهد. اگر به صفر آن را میل دهیم خروجی خطی و در عوض اگر آن را به بی نهایت میل دهیم خروجی به relu میل می کند.

ج)

مزایای Mish نسبت به relu:

- وجود دلتا باعث میشود که بهینه سازی شبکه سریع تر و بهتر و آسان تر انجام شود. زیرا این متغیر باعث میشود که میزان smoothing گرادیان خروجی بیشتر شود.





سوال دوم)

الف)

در ابتدا خروجی شبکه کاملا رندوم است. زیرا شبکه هنوز train نشده است. بنابراین می توان برای این امر احتمال score ها را 0.5 در نظر گرفت. پس:

در محاسبات با توجه به فرض بیان شده به این نتیجه رسیدم مقادیر اولیه برای مقدار خطا در MSE و BCE به ترتیب 0.5 و 0.69 می باشد.

ب)

از آنجا که مقدار خروجی تابع در حالتی که تابع خطا MSE است عددی بین ۰ تا ۱ است پس میزان خطا در هر مرحله حداکثر یک واحد میتواند افزایش یابد. درحالیکه این بازه برای binary cross entropy بزرگ تر استپس می تواند میزان خطا بسیار بازه بزرگ تری داشته باشد. هم چنین مقدار گرادیان های این حالت بزرگ تر است و همگرایی در آن سریع تر است. در نتیجه همگرای در داده های validation زودتر رخ می دهد. پس از آن شبکه به سمت overfitting حرکت می کند و در این حالت خطا در حالت آموزش کمتر می شود ولی در حالت اعتبارسنجی دقت تغییری نمی کند. اما همانطور که بالاتر بیان شد گرادیان های MSE کوچک تر است و همگرایی شبکه کند تر است و تا epoch های بالاتری همچنان آموزش و بهینه تر شدن ادامه دارد.

در MSE در epoch صدم هنوز همگرایی به اتمام نرسیده و ادامه دارد و فاصله نمودار ها کم است چون آموزش و بهینه شدن در هر دو حالت ادامه دارد. در BCE در epoch شصت الی شصت و پنج overfitting رخ می دهد.

بنابراین خطا در داده های validation در تابع دوم خطا می تواند مقدار بزرگ تری به خود بگیرد.

ت)

همانطور که در توضیح مورد قبلی نیز اشاره شد، برای تابع MSE فرآیند آموزش تا آخرین epoch ادامه دارد. زیرا مقدار خطا برای هر دو داده های آموزش و اعتبارسنجی در حال کاهش است. پس شبکه هم چنان در حال train است و تعمیم خوبی در انواع داده دارد. بنابراین بهترین حالت طبق نمودار عددی بیشتر از ۱۰۰ باید باشد تا به بهترین میزان خطا دست یابیم طبق روند عددی بین ۱۱۵ تا ۱۲۰ میتواند مناسب باشد.

اما در تابع Binary cross entrpy از یک جا فواصل بین نمودار خطای داده های آموزش و اعتبارسنجی زیاد می شود. پس overfitting رخ داده است. در این حالت نتیجه بهینه نخواهد بود زیرا شبکه فقط الگو داده های آموزش را یادگیری می کند و در صورتی که داده های اعتبارسنجی را با آن بدهیم خطای زیادی خواهد داشت زیرا فقط در داده های آموزش fit شده است. از انجا که بهترین میزان loss برای داده های اعتبارسنجی در epoch ۶۰ اتفاق افتاده است پس بهترین حالت است و قبل از آن نیز شبکه underfit است.

سوال سوم)

در این تمرین از دیتاست mnist استفاده کردیم. این دیتاست ۶۰۰۰۰ داده آموزش و ۱۰۰۰۰ داده تست دارد که در مجموع ۷۰۰۰۰ داده میشود.

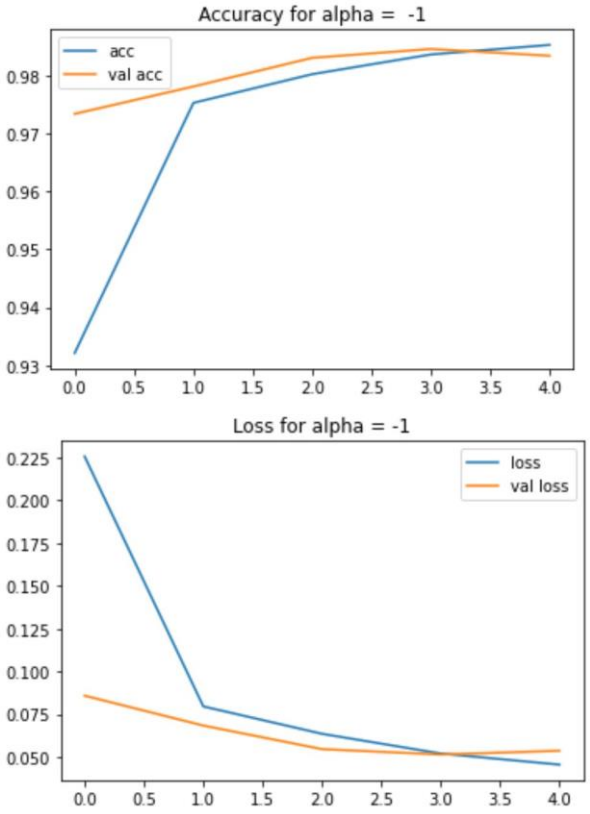
در مرحله اول مقدار هر پیکسل را طبق گفته سوال از ۲۵۵ کم کردیم تا حالت های متضاد عکس های موجود در دیتاست را تولید کنیم و تعداد داده ها را افزایش دهیم. در این حالت رنگ پس زمینه عکس ها از حالت مشکی به سفید و بالعکس تبدیل میشوند و حالت های حدید در دیتاست خواهیم داشت.

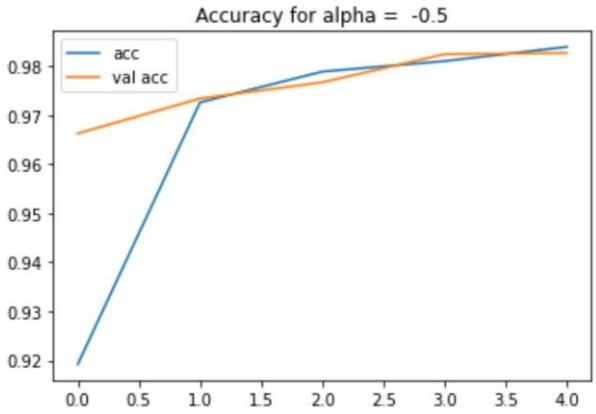
پس با عوض کردن پس زمینه هر عکس ما دو حالت برای هر عکس خواهیم داشت پس تعداد داده های دیتاست برابر ۱۲۰۰۰۰ داده آموزش و ۲۰۰۰۰ داده تست میشود. در مجموع برای این تمرین ۱۴۰۰۰۰ داده خواهیم داشت.

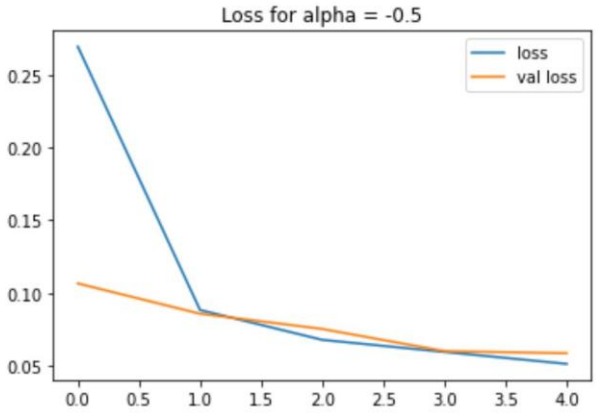
برای آنکه تصادفی بودن داده های در دیتاست رعایت شود با استفاده از متد shuffle که در کتابخانه sklearn موجود است داده ها را مخلوط می کنیم.

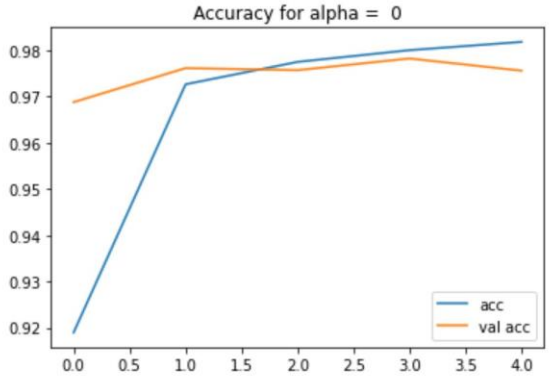
در مرحله بعد برای استفاده داده های سوال در لایه های Conv باید از خاصیت reshape بهره ببریم تا شکل داده را از 28 در 28 به 28 در 28 در 1 تبدیل کنیم.

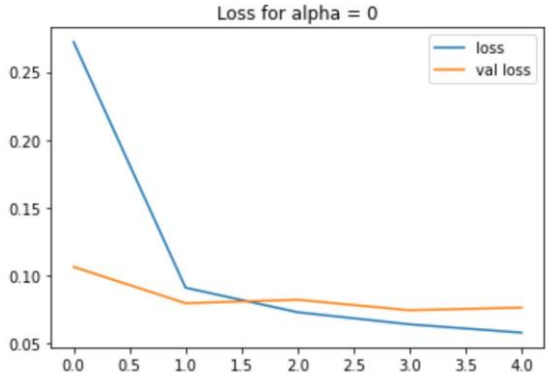
سپس داده ها را نرمالایز می کنیم تا در بازه ۰ تا ۱ قرار گیرند تا محاسبات ساده تر انجام پذیرد. هم چنین label عکس ها را به حالت کتگوریکال در می آوریم.

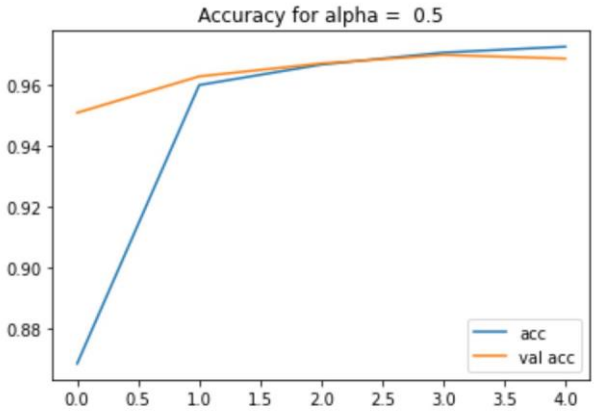


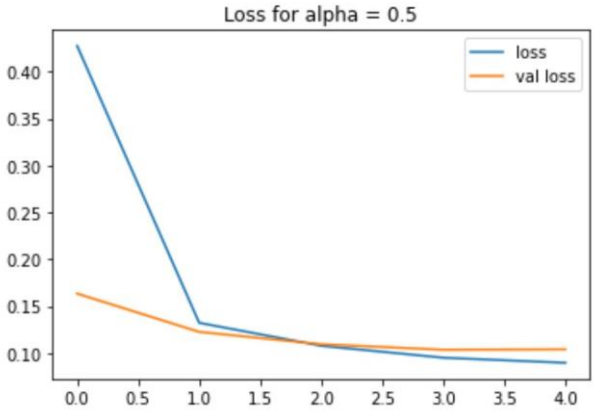

Accuracy for alpha = -0.5
acc
val acc
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

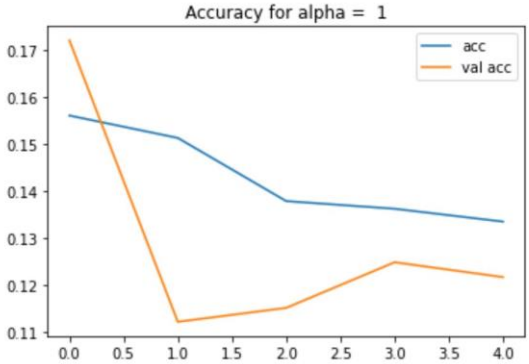

Loss for alpha = -0.5
loss
val loss
0.25
0.20
0.15
0.10
0.05
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

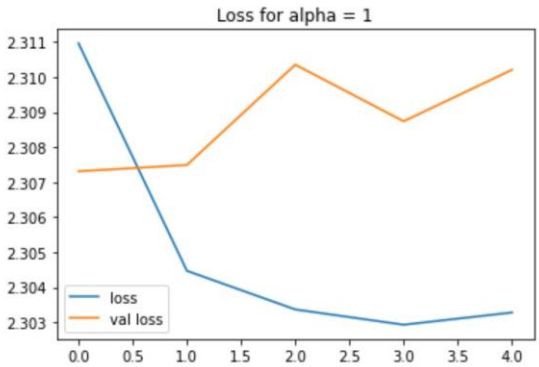

Accuracy for alpha = 0
0.98
0.97
0.96
0.95
0.94
0.93
0.92
acc
val acc
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Loss for alpha = 0
loss
val loss
0.25
0.20
0.15
0.10
0.05
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Accuracy for alpha = 0.5
acc
val acc
0.96
0.94
0.92
0.90
0.88
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Loss for alpha = 0.5
loss
val loss
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Accuracy for alpha = 1
acc
val acc
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Loss for alpha = 1
loss
val loss
2.311
2.310
2.309
2.308
2.307
2.306
2.305
2.304
2.303
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

نمودار پنج حالت دقت و خطا برای آلفا های گوناگون در بالا نمایش داده شده است.

برای مقایسه این حالات ابتدا لازم مقادیر زیر بررسی شود:

```
Alpha = -1 train acc = 0.9853214025497437
Alpha = -1 test acc = 0.982200026512146
Alpha = -1 val acc = 0.9834444522857666
-----------------------------------------
Alpha = -0.5 train acc = 0.9839047789573669
Alpha = -0.5 test acc = 0.9814500212669373
Alpha = -0.5 val acc = 0.9826666712760925
-----------------------------------------
Alpha = 0 train acc = 0.9818333387374878
Alpha = 0 test acc = 0.9789000153541565
Alpha = 0 val acc = 0.9756110906600952
-----------------------------------------
Alpha = 0.5 train acc = 0.9725714325904846
Alpha = 0.5 test acc = 0.9742000102996826
Alpha = 0.5 val acc = 0.968666672706604
-----------------------------------------
Alpha = 1 train acc = 0.13341666758060455
Alpha = 1 test acc = 0.13384999334812164
Alpha = 1 val acc = 0.12158333510160446
-----------------------------------------
```

طبق جدول بالا بهترین نتیجه مربوط به مقدار 0.5- می باشد. این مقدار دارای بهترین نتیجه کلی از نظر دقت های داده های آموزش، تست و اعتبارسنجی است. مقدار 1 دارای دقت بسیار پایینی می باشد. علت آن است که این تابع در این حالت درواقعا یک تابع activation خطی می باشد. به همین دلیل است که شبکه ساده است و عمق خاصی ندارد و دقت و خطا مقادیر ضعیفی در آن دارند.