NN for Images – Ex2

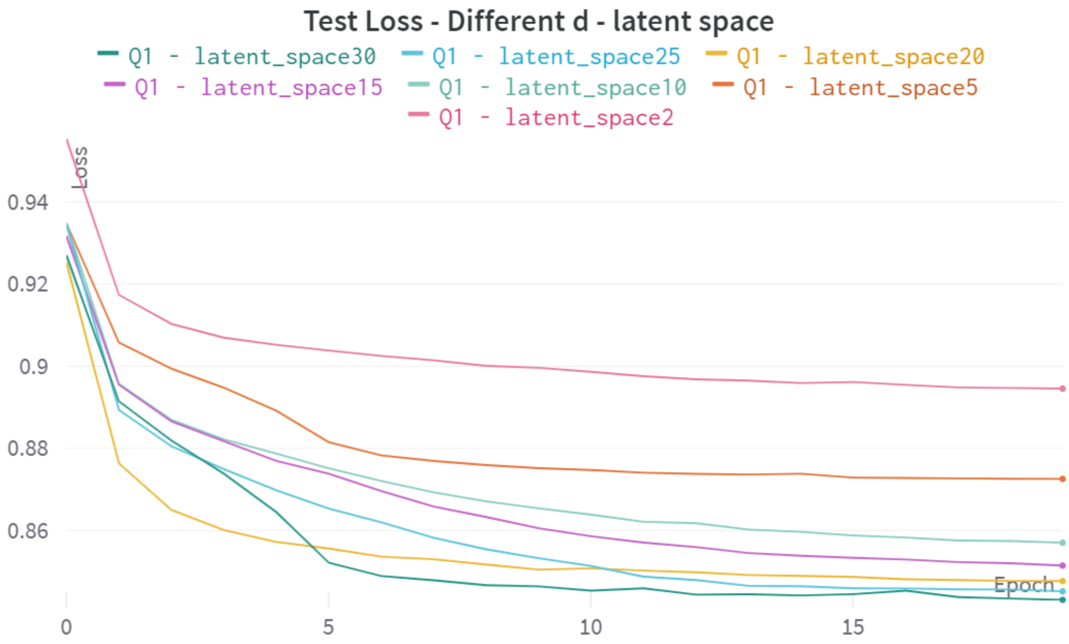
David Guedalia

1. Practical Task

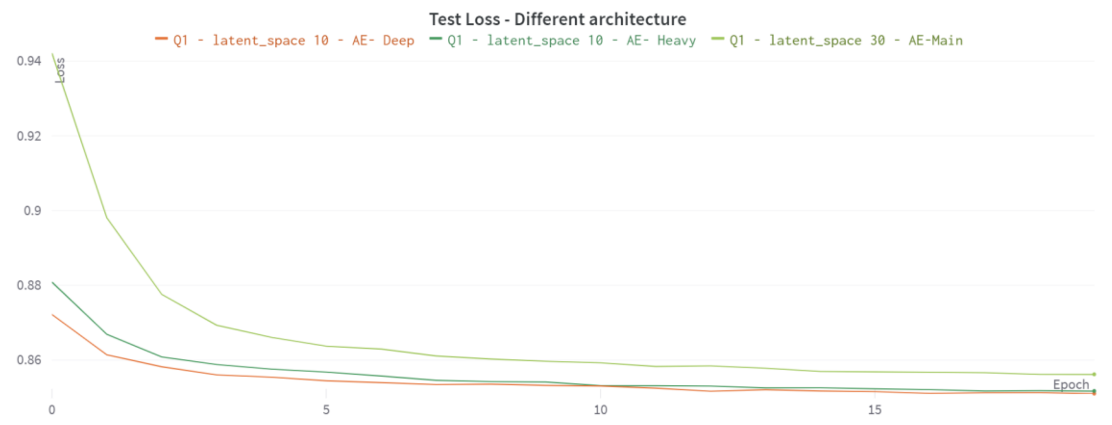
1.1 Auto-Encoding

- On this Section I tried different architecture, with different of convolution layers and change the inner dimensions and finally I decide to go forward with this architecture:
  The first net which is the Encoder and one for the Decoder, I decided that the encoder will contain three convolution layers, and two fully connect as I saw it gives performance that are good enough, Also I found that batch nor, layer improving the performance that make regularizes, as we saw in class the Decoder has the same architecture but in inverse order. Also, I used ReLU as the activation function, except of the last activation on the decoder that I used Sigmoid.

- On this section I tried the architecture on different dimension d as the latent.
  Where the sizes were $d \in \{2,\ 5,\ 10,\ 15,\ 20,\ 25,\ 30,\}$
  And I concluded that the performances are well when $d = 10$ and that why I chose it, and if $d > 10$ the improvement is much smaller and less significant.



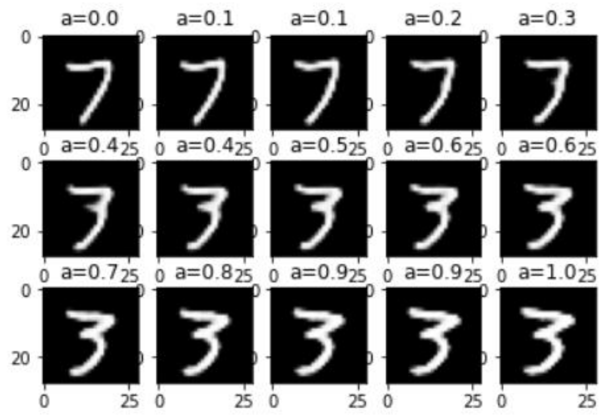
- I chose to work with MSE as the function loss and with this function I trained the model.
- Also, I decided to add on the output of the Decoder the activation function, Sigmoid, as to let the model make the pixels to be more significant, because the modt of pixels in dataset are strongly black or white.
- I compared different architecture where **Deep** is with four convolution and the **heavy** is with bigger weights.



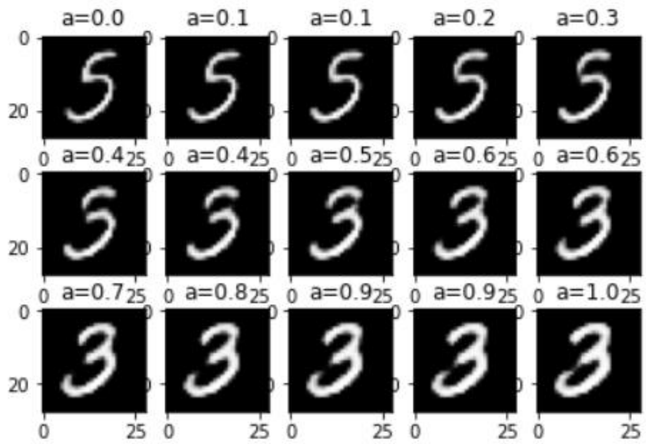
1.2 Interpolation

On this section after I trained the autoencoder, I interpolate between two different digits where I chose $\alpha\epsilon(0,1)$

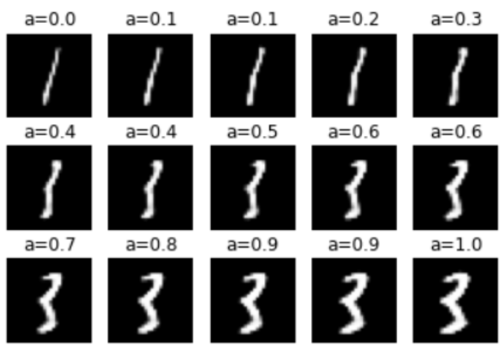
i. I add interpolate between two digits:



ii. Different pairs of digits:



iii. When I tried to use higher dimension where $d\sim 20$.



iv. Why the quality does not increase:

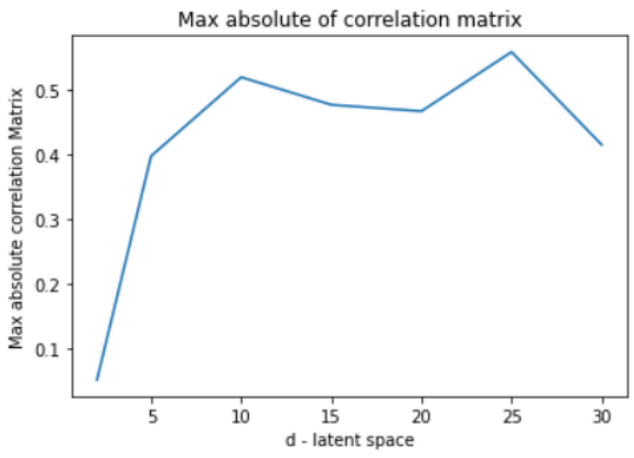
After several times I tried with $d \sim 20$ and I saw the results didn’t get better and that's because as much d is getting bigger there is more noise on the image and that’s why I get better performances for $d \sim 10$.

1.3 Decorrelation

In this section, I decided to look on the highest absolute value of the correlation matrix, I calculate this with the average of the all matrix in the dataset, where I trained over ~ 3300.

I expected to see that as much the latent space d is bigger the highest absolute value of the correlation matrix is getting higher also, and I think it's because when d is low we won't keep dependent variables.

I add a plot of the trend in correlation versus d:



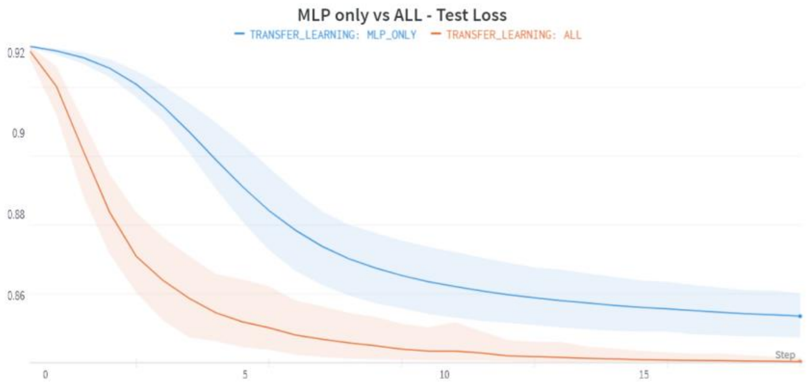
1.4 Transfer Learning

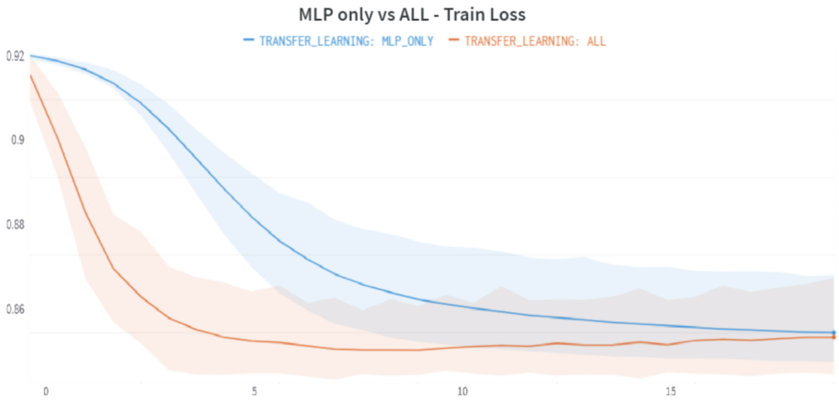
On this section I used the trained encoders from the first question and I connect them a MLP net that make a classification.

One time the weights of the Encoder were trained and second time only MLP net were trained, I chose the loss function to be entropy loss.

And after we tried on small data ~ 80.

And this is the results I got:





From these graphs we can obtain that when the weights of the encoder are trained, we getting faster to overfit, also after ~ 15 epochs the test getting higher, also after try several d I got that I get the best performance for $d \sim 10$.

I saw it also on the accuracy where it was almost 1 on the accuracy and ~ 0.84.

And the others model where better on the test although they got worse on the train.

תרגיל 2 - תיאורטי

שאלה 1

עפ"י השאלה נניח כי f, g הינן פונקציות לינאריות

כך ש $f: V \to W$ ו $g: W \to U$.

כעת נשים לב כי מתקיים:

$$g \circ f(v_1 + \alpha v_2) \overset{\text{לינאריות}}{=} g(f(v_1) + \alpha f(v_2))$$

$$= g \circ f(v_1) + \alpha [g \circ f(v_2)]$$

מכאן קיבלנו כי זה אכן עונה להגדרה של פונק' לינארית.

אם f, g הינן פונק' אפיניות אזי קיימות פונקציות כך ש:

$g(w) = g^*(w) + u$, $f(v) = f^*(v) + w$ (וכן ש f^* ו g^* לינאריות).

אז נקבל כי: $g \circ f(v) = g^*(f^*(v)) + g^*(w) + u$

וגם פה נקבל כי $g \circ f$ הינה פונקציה אפינית מאחר ו- $g^*(w) + u$ קבוע ב V.

והפונקציות g^* ו f^* לינאריות.

שאלה 2

סעיף א׳

— נשים לב כי אם f קמורה אזי θ^n היא מינימום מקומי.

— נשים לב כי תנאי העצירה הוא $\nabla f_{\theta^n}(x)=0$, מאחר ונשים לב כי כאשר התנאי הזה מתקיים נקבל כי:

$$\theta^{n+1}=\theta^n-0$$

$$\Rightarrow \quad \nabla f_{\theta^{n+1}}(x)=0$$

נמשיך ונקבל:

$$\theta^n=\theta^{n+1}=\theta^{n+2}=\theta^{n+3}=\dots$$

כלומר אין מה להמשיך עם האיטרציות וזהו תנאי העצירה.

סעיף ב׳

נשים לב כי אם $\nabla f(x)=0$ ו $H\succ 0$ (או לחילופין $H\prec 0$)
אזי x היא נקודת מינימום או מקסימום מקומי.

- נשים כי התנאים מתקיימים

שאלה 3

פסודו - קוד :

$$\frac{\alpha_1 \cdot \pi}{180} = \theta_1$$

$$\frac{\alpha_2 \cdot \pi}{180} = \theta_2$$

$$\begin{bmatrix} \cos(\theta_1) \\ \sin(\theta_1) \end{bmatrix} = A$$

$$\begin{bmatrix} \cos(\theta_2) \\ \sin(\theta_2) \end{bmatrix} = B$$

ונחזיר $L = \|A-B\|$ (כמובן שבפייתון נעשה חישוב שונה.)

נשים לב כי L היא רציפה ומחזורית ואף זוגית בהפרש בין α ל 360 שקולה בין 0 ל 2π.
בנוסף היא מונו' עולה בין 0 ל 180 ויורדת בין 180 ל 360

שאלה 4

סעיף א'

$g(x) = (x+y, 2x, z)$

$$\frac{\partial}{\partial x} f(x+y, 2x, z) = D\big[f(g(x))\big] \overset{\text{כלל השרשרת}}{=} D\big[f(g(x))\big] D\big[g(x)\big]$$

$$= D\big[f(x+y, 2x, z)\big] \cdot \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

סעיף ב'

ניזכר בכלל השרשרת:

$$D\big[f_1(\dots f_n(x))\big]$$

$$= D\big[f_1(f_2(\dots f_n(x)))\big] \cdot D\big[f_2(\dots f_n(x))\big]$$

$$= D\big[f_1(f_2(\dots f_n(x)))\big] \cdot D\big[f_2(\dots f_n(x))\big] \cdot D\big[f_3(\dots f_n(x))\big]$$

- נמשיך
- בצורה
- אינדוקטיבית ונקבל:

$$= D\big[f_1(f_2(\dots f_n(x)))\big] \cdot D\big[f_2(\dots f_n(x))\big] \cdot \dots \; D\big[f_{n-1}(f_n(x))\big]$$

סעיף ג

עבור $1<t$ נגדיר

$$g_t(x) = (x, f_t(x \ldots f_{n-1}(x, f_n(x)))$$

וכעת נגזור את הפונק' הבאה:

$$D[f_1(g_2(x))] =$$

$$= D[f_1(g_2(x))] \cdot D[g_2(x)]$$

$$= D[f_1(g_2(x))] \cdot \begin{bmatrix} 1 & 0 \\ 0 & D[f_2(g_3(x))] \end{bmatrix}$$

$$\vdots$$

$$= D[f_1(g_2(x))] \begin{bmatrix} 1 & 0 \\ 0 & D[f_2(g_3(x))] \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & D[f_3(g_3(x))] \end{bmatrix} \cdot \begin{bmatrix} I_{n-1} & 0 \\ 0 & D[f_n(x)] \end{bmatrix}$$

סעיף ג

באופן דומה לסעיפים הקודמים, נגדיר:

$$Z(x) = x + g(x + h(x))$$

בנוסף:

$$T(x) = x + h(x)$$

לכן:

$$D\left[f(x+g(x+h(x))\right] =$$

$$= D\left[f(Z(x))\right]$$

$$= D\left[f(Z(x))\right] \cdot D\left[Z(x)\right]$$

$$= D\left[f(x+g(x+h(x))\right] \cdot D\left[x + g(T(x))\right]$$

$$= D\left[f(x+g(x+h(x))\right] \cdot \left(I_n + D\left[g(x+h(x))\right] \cdot D\left[x + h(x)\right]\right.$$

$$= D\left[f(x+g(x+h(x))\right] \cdot \left(I_n + D\left[g(x+h(x))\right] \cdot \left(I_n + D\left[h(x)\right]\right)\right)$$