Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Νευρωνικά Δίκτυα - Βαθιά Μάθηση

Εργασία 1

Κούτση Χριστίνα

ΑΕΜ: 9872

cvkoutsi@ece.auth.gr

Νοέμβριος, 2022

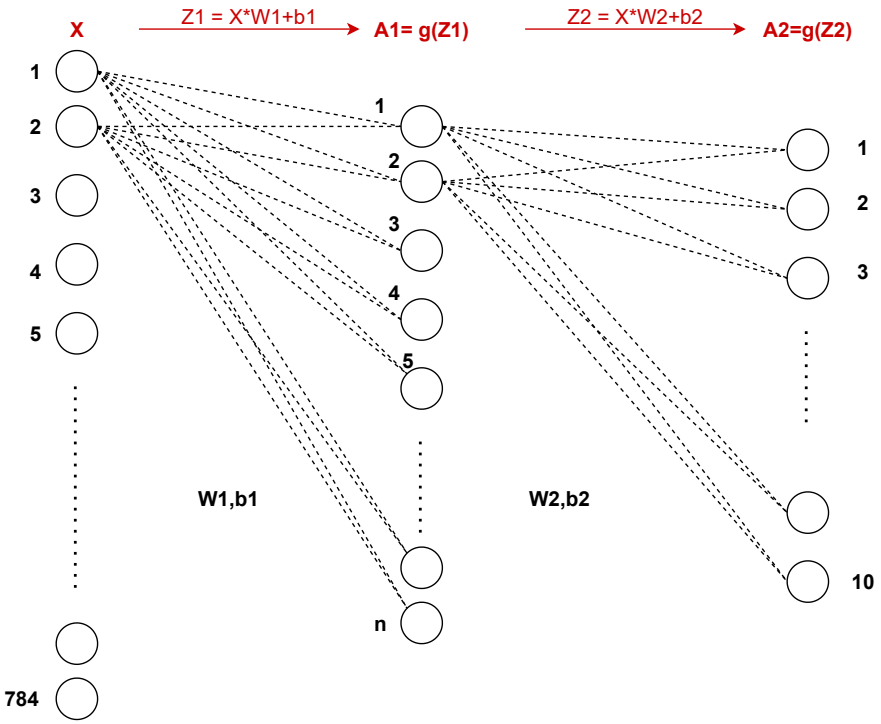
Περιεχόμενα

Κεφάλαιο 1

Δημιουργία Νευρωνικού δικτύου τριών στρωμάτων

Η παρούσα εργασία έχει αντικείμενο την δημιουργία ενός νευρωνικού δικτύου τριών στρωμάτων, με σκοπό την ανίχνευση δεκαδικών ψηφίων από εικόνες χειρόγραφων αριθμών. Το dataset που θα χρησιμοποιηθεί θα είναι το mnist dataset σε csv μορφή

Το νευρωνικό δίκτυο θα είναι πλήρως συνδεδεμένο, θα έχει 784 νευρώνες στην είσοδο, 10 νευρώνες στην έξοδο και θα εκπαιδευτεί με τον αλγόριθμο back propagation.



1.1 Φόρτωση και κανονικοποίηση δεδομένων

Το dataset περιέχει 60.000 παραδείγματα εκπαίδευσης και 10.000 παραδείγματα ελέγχου. Τα πρώτα 5 παραδείγματα εκπαίδευσης φαίνονται παρακάτω:

	label	1x1	1x2	1x3	1x4	1x5	1x6	1x7	1x8	1x9	...	28x19	28x20	\
0	5	0	0	0	0	0	0	0	0	0	...	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	
2	4	0	0	0	0	0	0	0	0	0	...	0	0	
3	1	0	0	0	0	0	0	0	0	0	...	0	0	
4	9	0	0	0	0	0	0	0	0	0	...	0	0	

	28x21	28x22	28x23	28x24	28x25	28x26	28x27	28x28
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0

[5 rows x 785 columns]

Βλέπουμε ότι η πρώτη στήλη περιέχει τα labels των εικόνων και οι στήλες 1 έως 784 περιέχουν τα pixels. Παρατηρώ ότι οι εικόνες έχουν ανασχηματιστεί από έναν πίνακα 28 x 28 pixels σε μία στήλη με 784 pixels.

Επομένως, ορίζω ως ετικέτες εκπαίδευσης και ελέγχου τις πρώτες στήλες των δεδομένων εκπαίδευσης και ελέγχου αντίστοιχα και ως δεδομένα εκπαίδευσης και ελέγχου τις υπόλοιπες στήλες. Στη συνέχεια κανονικοποιώ τα δεδομένα διαιρώντας με 255, καθώς αυτή είναι η υψηλότερη τιμή που μπορεί να πάρει η φωτεινότητα ενός pixel.

Εφόσον τα δεδομένα εισόδου έχουν διάσταση 784xM, το στρώμα εισόδου του νευρωνικού δικτύου θα έχει 784 νευρώνες, ενώ εφόσον οι ετικέτες έχουν διάσταση 10xM το στρώμα εξόδου θα έχει 10 νευρώνες.

1.2 Αρχικοποίηση βαρών

Αρχικοποιούμε τις παραμέτρους W1, W2 και b1,b2 ανάλογα με τον αριθμό των νευρώνων στο στρώμα εισόδου, στο κρυφό στρώμα και στο στρώμα εξόδου. Αν θεωρίσουμε ότι το κρυφό στρώμα έχει n νευρώνες:

- W1: Πίνακας διάστασης nx784
- W2: Πίνακας διάστασης 10xn
- b1: Πίνακας διάστασης nx1
- b2: Πίνακας διάστασης 10x1

Τα βάρη αρχικοποιούνται στο διάστημα [-0.5,0.5]

1.3 Ορισμός συναρτήσεων ενεργοποίησης και βοηθητικών συναρτήσεων

- ReLU: Χρησιμοποιείται ως συνάρτηση ενεργοποίησης στο κρυφό στρώμα και ορίζεται ως

$$\text{ReLU(Z)} = \begin{cases} Z, Z > 0 \\ 0, Z \leq 0 \end{cases}$$

Επιπλέον, χρησιμοποιείται και η παράγωγος της ReLU, η οποία ορίζεται ως

$$\frac{\mathrm{dReLU}}{\mathrm{dZ}} = \begin{cases} 1, Z > 0 \\ 0, Z \leq 0 \end{cases}$$

- Softmax: Χρησιμοποιείται ως συνάρτηση ενεργοποίησης στο στρώμα εξόδου και δίνει αποτέλεσμα στο [0,1]. Ορίζεται ως:

$$Softmax(Z) = \frac{e^{Z_i}}{\sum_{j=1}^{K} e^{Z_j}}$$

- One Hot Encoding: Χρησιμοποιείται για την κωδικοποίηση των labels και για έναν ακέραιο αριθμό εισόδου επιστρέφει ένα διάνυσμα μήκους M με μηδενικά και μονάδα στην θέση που ισούται με την τιμή του αριθμού.
  π.χ. για είσοδο 4 επιστρέφει το διάνυσμα [0,0,0,0,1,0,...,0]
- Mean Squared Error: Χρησιμοποιείται για την εκτίμηση των αποτελεσμάτων του νευρωνικού δικτύου και ορίζεται ως

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- Accuracy: Χρησιμοποιείται για την εκτίμηση των αποτελεσμάτων του νευρωνικού δικτύου και ορίζεται ως ο αριθμός των labels που ταξινομήθηκαν σωστά προς το συνολικό αριθμό των labels

1.4 Forward Propagation

Το Forward propagation υλοποιεί το πέρασμα των δεδομένων εισόδου στην έξοδο, αφού υλοποιηθούν οι πολλαπλασιασμοί με τα βάρη W1 και W2. Οι πράξεις που υλοποιούν την παραπάνω διαδικασία είναι:

$$\begin{cases} \mathbf{Z1} = \mathbf{W1} \times \mathbf{X} + \mathbf{b1} \\ \mathbf{A1} = g1(\mathbf{Z1}) \\ \mathbf{Z2} = \mathbf{W2} \times \mathbf{A1} + \mathbf{b2} \\ \mathbf{A2} = g2(\mathbf{Z2}) \end{cases}$$

1.5 Back Propagation

Ο αλγόριθμος του back propagation υλοποιεί την ανανέωση των βαρών W1,W1,b1,b2 προς την κατεύθυνση μείωσης του σφάλματος εξόδου. Ο ψευδοκώδικας που υλοποιεί τον αλγόριθμο back propagation είναι ο εξής:

Backward pass:

Compute error signals

$$\Delta^{(L)} = \begin{bmatrix} \delta^{(L,0)} & \dots & \delta^{(L,M)} \end{bmatrix}^{\top}$$

for $k \leftarrow L-1$ *to* 1 **do**

$$\Delta^{(k)} \leftarrow \mathrm{act}'(\boldsymbol{Z}^{(k)}) \odot (\Delta^{(k+1)} \boldsymbol{W}^{(k+1)\top}) \ ;$$

end

Return $\frac{\partial c^{(n)}}{\partial \boldsymbol{W}^{(k)}} = \sum_{n=1}^{M} \boldsymbol{a}^{(k-1,n)} \otimes \boldsymbol{\delta}^{(k,n)}$ for all k

Η εύρεση της μεταβολής των βαρών με στόχο την μείωση του σφάλματος υλοποιείται από τις παρακάτω πράξεις:

$$\begin{cases} \mathbf{dZ2} = \mathbf{A2} - \mathbf{Y} \\ \mathbf{dW2} = \frac{1}{\mathrm{m}} \mathbf{dZ2} \mathrm{x} \mathbf{A1}^{T} \\ \mathbf{db2} = \frac{1}{\mathrm{m}} \sum \mathbf{dZ2} \\ \mathbf{dA1} = \mathbf{W2}^{T} \mathrm{x} \mathbf{dZ2} \\ \mathbf{dZ1} = \mathbf{dA1} \mathrm{xg'(Z1)} \\ \mathbf{dW1} = \frac{1}{\mathrm{m}} \mathbf{dZ1} \mathrm{x} \mathbf{\Xi}^{T} \\ \mathbf{db1} = \frac{1}{\mathrm{m}} \sum \mathbf{dZ1} \end{cases}$$

1.6 Update Parameters

Αφού έχουμε υπολογίσει τα dW1,dW2,db1,db2 ανανεώνουμε τις παραμέτρους W1,W2,b2,b2 ως εξής:

$$\begin{cases} \mathrm{W1} = \mathrm{W1} - a * tldW1 \\ \mathrm{b1} = \mathrm{b1} - a * tldb1 \\ \mathrm{W2} = \mathrm{W2} - a * tldW2 \\ \mathrm{b2} = \mathrm{b2} - a * tldb2 \end{cases}$$

όπου α ο ρυθμός μάθησης.

1.7 Εκπαίδευση

Αφού υλοποιήσαμε τα παραπάνω, εκπαιδεύουμε το νευρωνικό δίκτυο με τον εξής τρόπο: Αρχικοποιούμε τα βάρη και για κάθε εποχή:

1. Κάνουμε shuffle τα δεδομένα εκπαίδευσης
2. Εισάγουμε στο νευρωνικό δίκτυο τα δεδομένα εισόδου με μέγεθος 784 x k, όπου k το batch size

3. Πραγματοποιούμε forward propagation
4. Πραγματοποιούμε back propagation
5. Ενημερώνουμε τις παραμέτρους

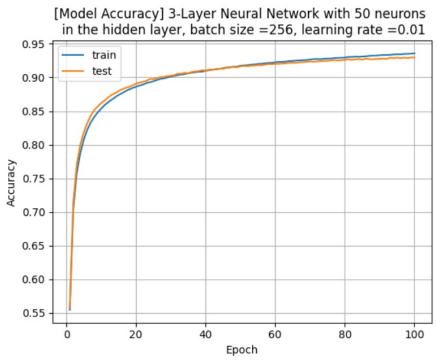
Κεφάλαιο 2

Απόδοση του νευρωνικού δικτύου σε δεδομένα εισόδου χωρίς μείωση διάστασης

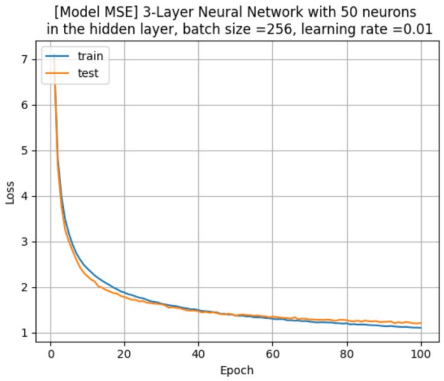
2.1 Δοκιμή Διάφορων αριθμών νευρώνων στο κρυφό στρώμα

Εκπαιδεύουμε το νευρωνικό δίκτυο για αριθμό νευρώνων στο κρυφό στρώμα n = [50,150,200]. Θέτουμε learning rate = 0.01, batch size = 256 και εκπαιδεύουμε το δίκτυο για 100 εποχές. Παρακάτω φαίνονται τα διαγράμματα του MSE και accuracy για κάθε περίπτωση:

- 50 νευρώνες στο κρυφό στρώμα
  Μέγιστη ακρίβεια για τα δεδομένα εκπαίδευσης = 0.9358
  Ελάχιστο σφάλμα για τα δεδομένα εκπαίδευσης= 1.1032
  Χρόνος εκπαίδευσης = 50.2007 s

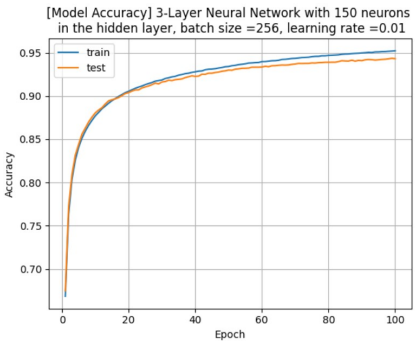


(α') Accuracy

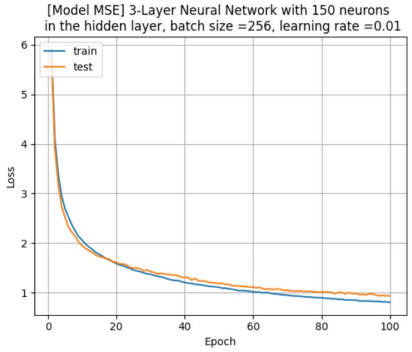


(β') MSE

- 150 νευρώνες στο κρυφό στρώμα
  Μέγιστη ακρίβεια για τα δεδομένα εκπαίδευσης = 0.9523
  Ελάχιστο σφάλμα για τα δεδομένα εκπαίδευσης = 0.8046
  Χρόνος εκπαίδευσης = 87.9747 s

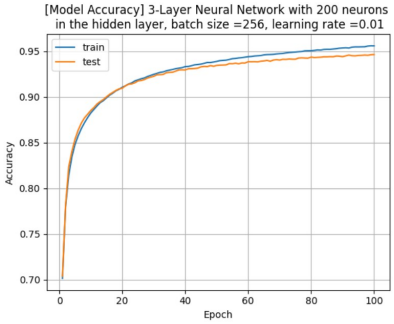


(γ') Accuracy

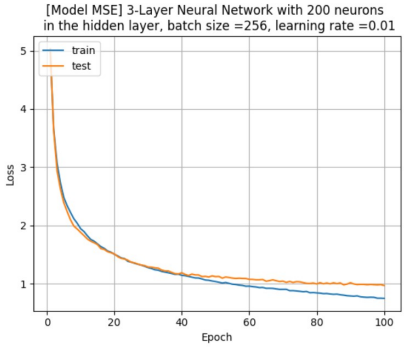


(δ') MSE

- 200 νευρώνες στο κρυφό στρώμα
  Μέγιστη ακρίβεια για τα δεδομένα εκπαίδευσης = 0.9557
  Ελάχιστο σφάλμα για τα δεδομένα εκπαίδευσης = 0.7475
  Χρόνος εκπαίδευσης = 132.5587 s



(ε') Accuracy



(ς') MSE

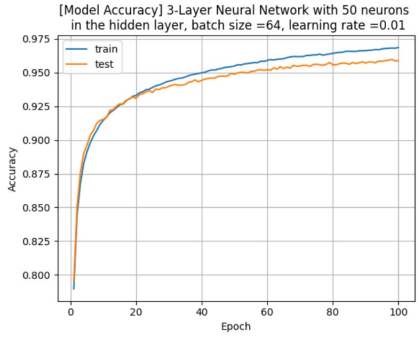
Παρατηρήσεις

- Η αύξηση του αριθμού των νευρώνων στο κρυφό στρώμα μπορεί να αυξήσει την απόδοση του νευρωνικού δικτύου
- Η αύξηση του αριθμού των νευρώνων στο κρυφό στρώμα αυξάνει τον χρόνο εκτέλεσης
- Η περεταίρω αύξηση του αριθμού των νευρώνων κρυφού στρώματος δεν αυξάνει απαραίτητα την απόδοση

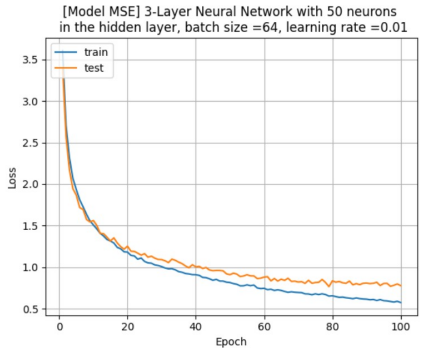
2.2 Δοκιμή Διάφορων batch sizes

Δοκιμάζουμε batch size = [64,128,256,60000] με learning rate = 0.01, n_h = 50 και εκπαιδεύουμε το δίκτυο για 100 εποχές:

- batch size = 64
  Μέγιστη ακρίβεια για τα δεδομένα εκπαίδευσης = 0.9685
  Ελάχιστο σφάλμα για τα δεδομένα εκπαίδευσης = 0.5743
  Χρόνος εκπαίδευσης = 95.7317 s

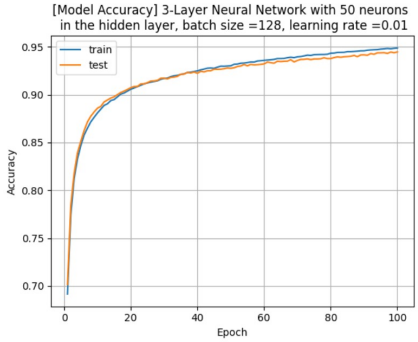


(ζ) Accuracy

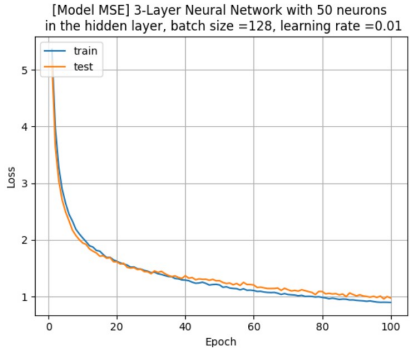


(η) MSE

- batch size = 128
  Μέγιστη ακρίβεια για τα δεδομένα εκπαίδευσης = 0.9488
  Ελάχιστο σφάλμα για τα δεδομένα εκπαίδευσης = 0.8957
  Χρόνος εκπαίδευσης = 60.5556 s

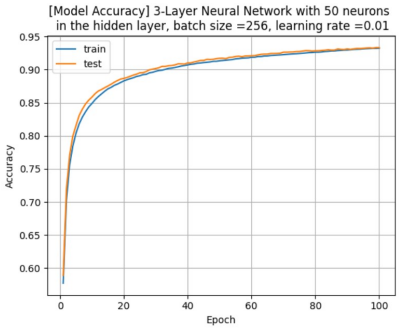


(θ) Accuracy

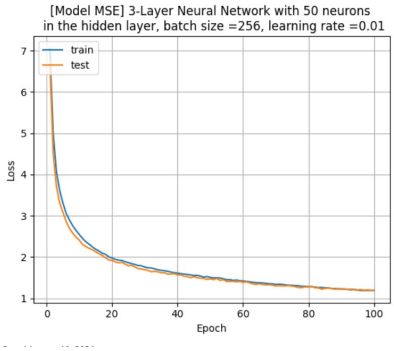


(ι) MSE

- batch size = 256
  Μέγιστη ακρίβεια για τα δεδομένα εκπαίδευσης = 0.9324
  Ελάχιστο σφάλμα για τα δεδομένα εκπαίδευσης = 1.1927
  Χρόνος εκπαίδευσης = 50.6881 s

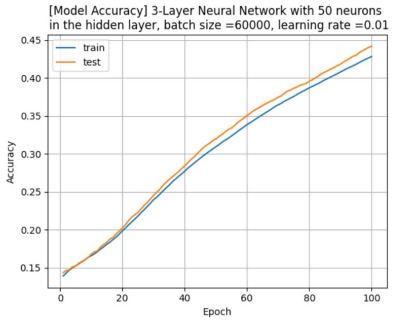


(ια΄) Accuracy

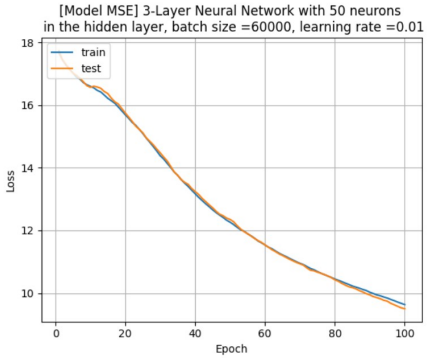


(ιβ΄) MSE

- batch size = 60000
  Μέγιστη ακρίβεια για τα δεδομένα εκπαίδευσης = 0.4283
  Ελάχιστο σφάλμα για τα δεδομένα εκπαίδευσης = 9.6323
  Χρόνος εκπαίδευσης = 46.3931 s



(ιγ΄) Accuracy



(ιδ΄) MSE

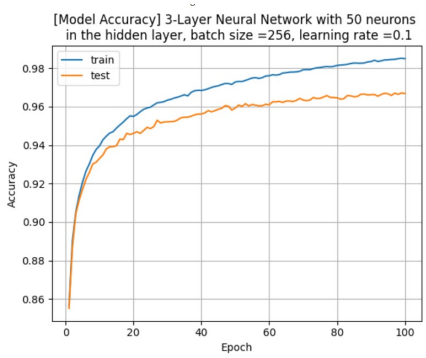
Παρατηρήσεις

- Η χρήση batch size μικρότερου του αριθμού των δεδομένων μπορεί να βελτιώσει δραματικά την απόδοση του αλγορίθμου σε συγκεκριμένο αριθμό εποχών.
- Η μείωση του batch size αυξάνει τον χρόνο εκπαίδευσης.
- Η χρήση μικρού batch size μπορεί να οδηγήσει σε υπερπροσαρμογή στα δεδομένα εκπαίδευσης.

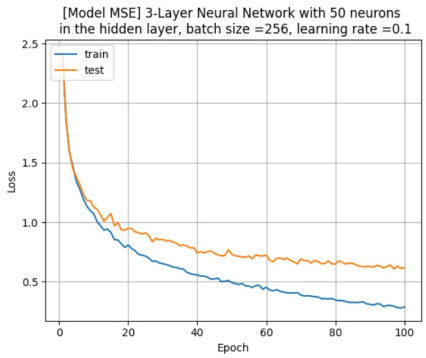
2.3 Δοκιμή Διάφορων learning rates

Δοκιμάζουμε learning rate = [0.1,0.01,0.001] με batch size = 256, n_h = 50 και εκπαιδεύουμε το δίκτυο για 100 εποχές:

- learning rate = 0.1
  Μέγιστη ακρίβεια για τα δεδομένα εκπαίδευσης = 0.9849
  Ελάχιστο σφάλμα για τα δεδομένα εκπαίδευσης = 0.2871
  Χρόνος εκπαίδευσης = 49.6446 s

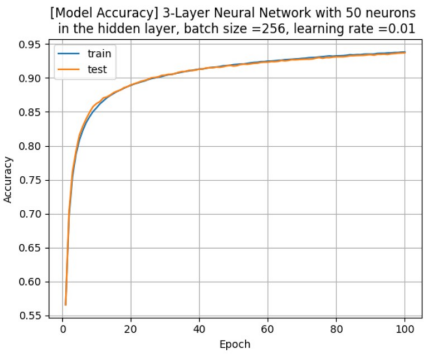


(ιε΄) Accuracy

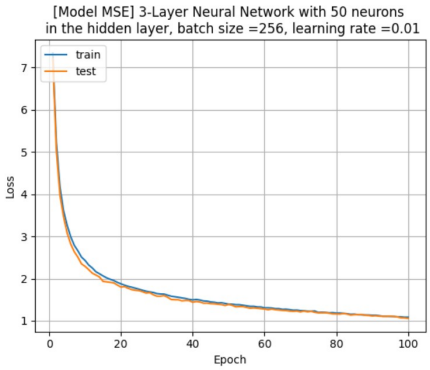


(ιϛ΄) MSE

- learning rate = 0.01
  Μέγιστη ακρίβεια για τα δεδομένα εκπαίδευσης = 0.9382
  Ελάχιστο σφάλμα για τα δεδομένα εκπαίδευσης = 1.0863
  Χρόνος εκπαίδευσης = 47.8854 s

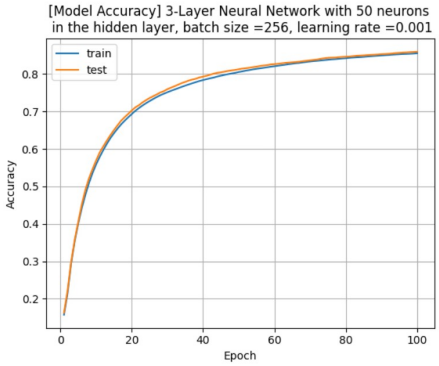


(ιζ΄) Accuracy

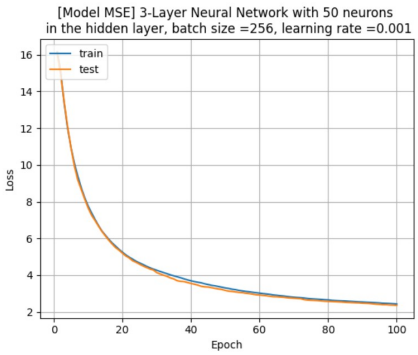


(ιη΄) MSE

- learning rate = 0.001
  Μέγιστη ακρίβεια για τα δεδομένα εκπαίδευσης = 0.8557
  Ελάχιστο σφάλμα για τα δεδομένα εκπαίδευσης = 2.4318
  Χρόνος εκπαίδευσης = 50.6385 s



(ιθ΄) Accuracy



(κ΄) MSE

Παρατηρήσεις

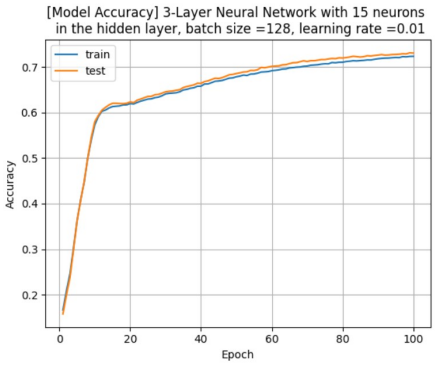
- Μικρό learning rate μπορεί να οδηγήσει σε μεγάλη ακρίβεια σε μικρό αριθμό εποχών, μπορεί όμως να οδηγήσει σε υπερπροσαρμογή
- Για μικρότερο learning rate το νευρωνικό δίκτυο χρειάζεται περισσότερες εποχές εκπαίδευσης ώστε να αποκτήσει μεγάλη ακρίβεια

Κεφάλαιο 3

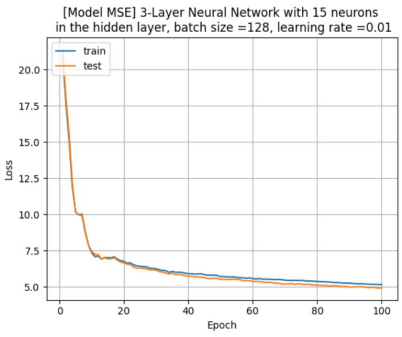
Απόδοση του νευρωνικού δικτύου σε δεδομένα εισόδου με μείωση διάστασης

3.1 Χρήση μέσης φωτεινότητας σειράς

Τροποποιούμε τα δεδομένα μας έτσι, ώστε τα δεδομένα εισόδου να αποτελούν την μέση φωτεινότητα σειράς κάθε εικόνας. Έτσι, τα δεδομένα εισόδου έχουν διάσταση 28xM. Εκπαιδεύουμε το νευρωνικό δίκτυο για αριθμό νευρώνων στο κρυφό στρώμα n = 15, θέτουμε learning rate = 0.01, batch size = 128 και εκπαιδεύουμε το δίκτυο για 100 εποχές. Παρακάτω φαίνονται τα διαγράμματα του MSE και accuracy:
Μέγιστη ακρίβεια για τα δεδομένα εκπαίδευσης = 0.7233
Ελάχιστο σφάλμα για τα δεδομένα εκπαίδευσης = 5.1392
Χρόνος εκπαίδευσης = 11.9861 s



(α') Accuracy



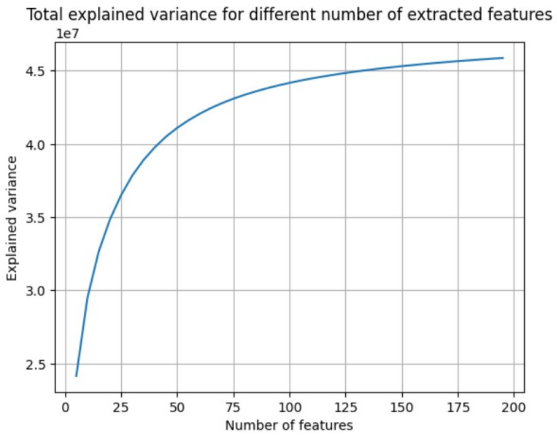
(β') MSE

Παρατηρήσεις

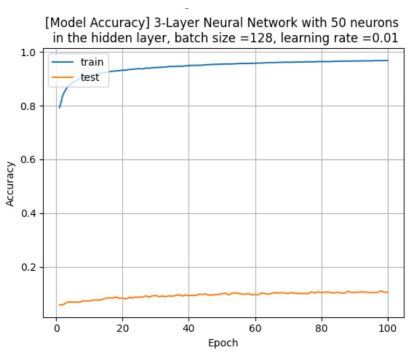
- Παρατηρώ ότι ο χρόνος εκπαίδευσης έχει μειωθεί σημαντικά, σε σύγκριση με τον χρόνο εκπαίδευσης του νευρωνικού δικτύου χωρίς μείωση διάστασης
- Το δίκτυο χρειάζεται περισσότερες εποχές εκπαίδευσης ώστε να καταφέρει ικανοποιητική ακρίβεια.

3.2 Χρήση αλγορίθμου PCA

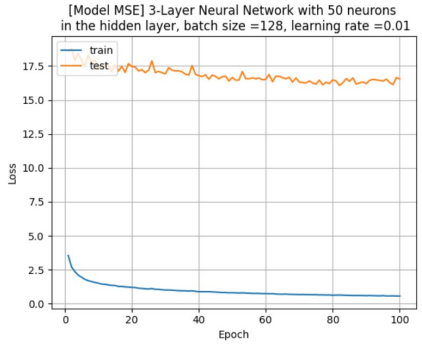
Υλοποιώ τον αλγόριθμο PCA και στη συνέχεια τρέχω τον αλγόριθμο για διάφορα κ, όπου κ είναι ο αριθμός των χαρακτηριστικών μετά την μείωση διάστασης. Παρακάτω φαίνεται η συνολική διακύμανση των δεδομένων συναρτήσει του κ:



Επιλέγω κ = 125.
Τα δεδομένα εισόδου έχουν πλέον διάσταση 125xM. Εκπαιδεύουμε το νευρωνικό δίκτυο για αριθμό νευρώνων στο κρυφό στρώμα n = 50, θέτουμε learning rate = 0.01, batch size = 128 και εκπαιδεύουμε το δίκτυο για 100 εποχές. Παρακάτω φαίνονται τα διαγράμματα του MSE και accuracy:
Μέγιστη ακρίβεια για τα δεδομένα εκπαίδευσης = 0.9684
Ελάχιστο σφάλμα για τα δεδομένα εκπαίδευσης = 0.5571
Χρόνος εκπαίδευσης = 96.7826 s



(γ') Accuracy



(δ') MSE

Παρατηρήσεις

- Το δίκτυο πετυχαίνει ικανοποιητική ακρίβεια για τα δεδομένα εκπαίδευσης,
- Το δίκτυο υπερπροσαρμόζεται στα δεδομένα εκπαίδευσης με αποτέλεσμα να μην έχουμε καλή ακρίβεια στα δεδομένα ελέγχου.