Υπολογιστική Νοημοσύνη

Εργασία 3

Επίλυση προβλήματος ταξινόμησης με χρήση Multi-layer Perceptron δικτύου

Ονοματεπώνυμο: Σιδηρόπουλος Λεωνίδας
ΑΕΜ: 9818
email: leonsidi@ece.auth.gr
Περίοδος: Φεβρουάριος 2023

Διερεύνηση απόδοσης μοντέλου με διαφοροποίηση στο σχεδιασμό και τη διαδικασία εκπαίδευσης

Για μελέτη της λειτουργίας των MLP θα εκπαιδεύσουμε 9 διαφορετικά μοντέλα. Τα κοινά που έχουν αυτά τα μοντέλα μεταξύ τους είναι:

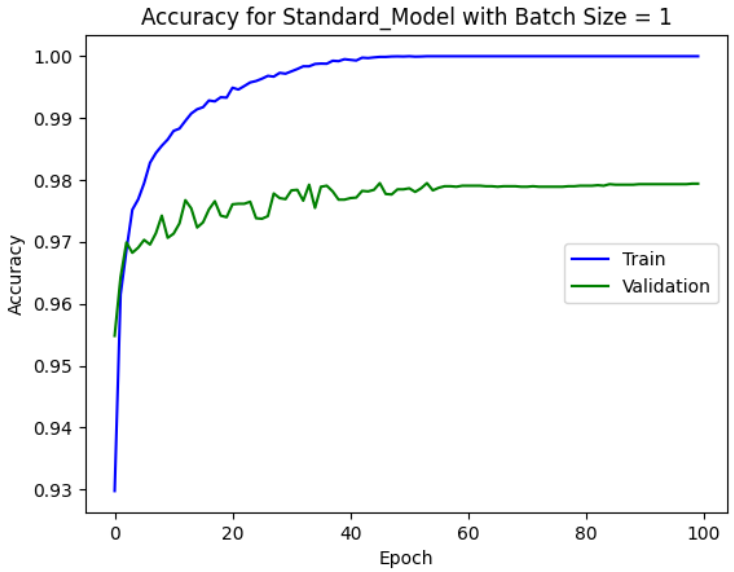
- Εποχές = 100
- Μετρική αξιολόγησης είναι η ακρίβεια (Accuracy)
- Τα 2 κρυφά στρώματα αποτελούνται απο 128 και 256 νευρώνες αντίστοιχα
- 20% του συνόλου των δεδομένων εκπαίδευσης χρησιμοποιείται για επικύρωση (validation)

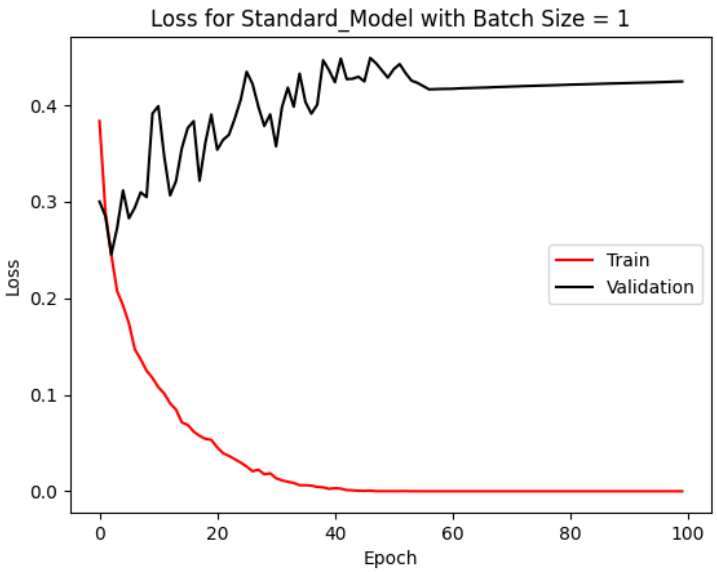
- Η αντικειμενική συνάρτηση προς βελτιστοποίηση είναι η categorical cross-entropy
- Η συνάρτηση ενεργοποίησης του στρώματος εξόδου είναι η softmax

Οι διαφορές των μοντέλων αφορούν κάποια σχεδιαστικά χαρακτηριστικά, όπως ο ρυθμός εκμάθησης, το batch size κτλ. Παρακάτω παρουσιάζονται τα μοντέλα με τις κυματομορφές τους, όπου στην αριστερή πλευρά είναι οι κυματομορφές ακρίβειας (Accuracy) και στη δεξιά πλευρά οι κυματομορφές απωλειών (Loss).

<u>Μοντέλο 1</u>

Χαρακτηριστικά: default δίκτυο, batch size = 1 (online μάθηση)



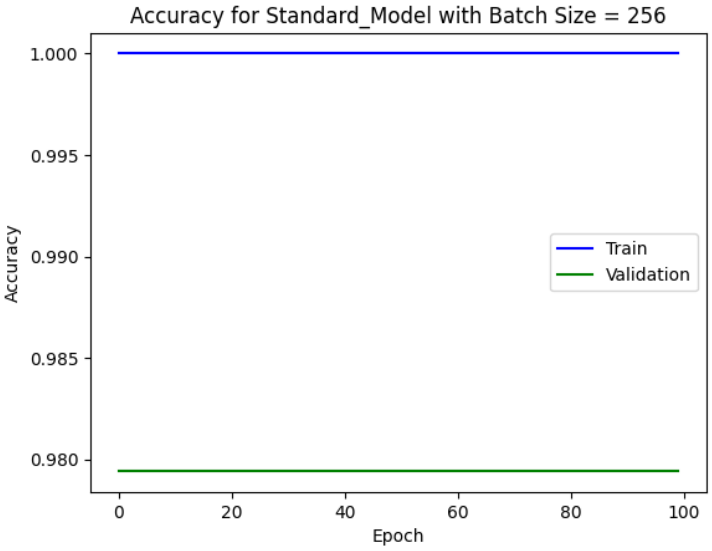


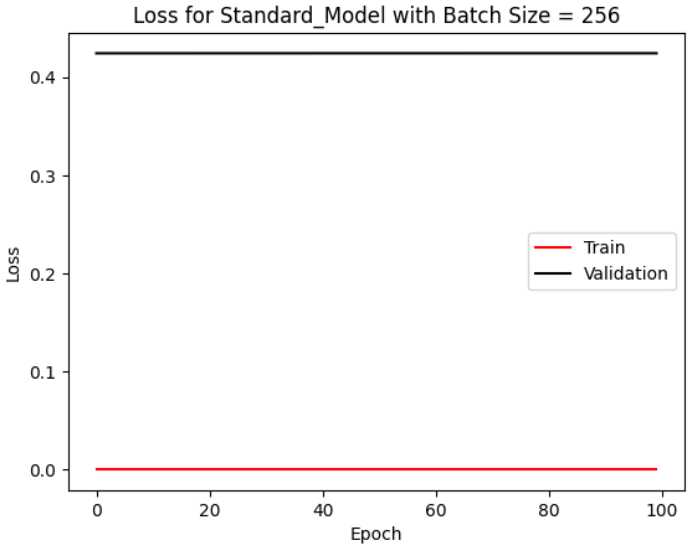
<u>Παρατηρήσεις/Αποτελέσματα:</u> Παρατηρούμε ότι το συγκεκριμένο μοντέλο πήρε πολύ χρόνο για να εκπαιδευτεί (training time = 15338,87 seconds), κάτι που το περιμέναμε εφόσον έχουμε ορίσει το batch size = 1. Γενικά, όσο πιο υψηλό είναι το batch size, τόσο ταχύτερη είναι και εκπαίδευση του μοντέλου, καθώς όσο μεγαλύτερο είναι το batch size τόσα περισσότερα δείγματα δέχεται το μοντέλο πριν κάνει

ενημέρωση των παραμέτρων του. Παρόλα αυτά, η εκπαίδευση του μοντέλου ήταν αρκετά αποτελεσματική, έχοντας πετύχει αρκετά υψηλές επιδόσεις στην ακρίβεια, τόσο στην εκπαίδευση όσο και στην επικύρωση. Ωστόσο, στην κυματομορφή των απωλειών, αν και η καμπύλη της εκπαίδευσης μειώνεται, η καμπύλη της επικύρωσης αυξάνεται με το πέρασμα τον εποχών. Αυτό μπορεί να προκαλέσει overfitting του μοντέλου, κάτι που δεν θέλουμε και για να αντιμετωπιστεί θα χρειαστεί να εφαρμόσουμε κάποια μέθοδο κανονικοποίησης(π.χ. early stopping κ.α.).

Μοντέλο 2

Χαρακτηριστικά: default δίκτυο, batch size = 256 (minibatch)

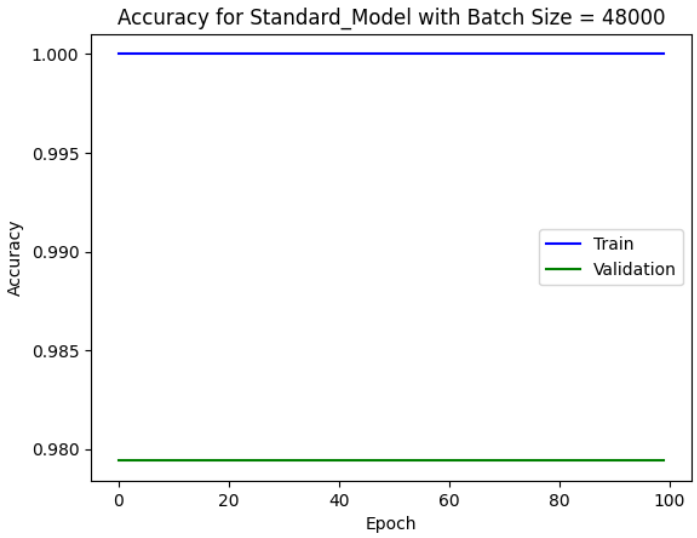


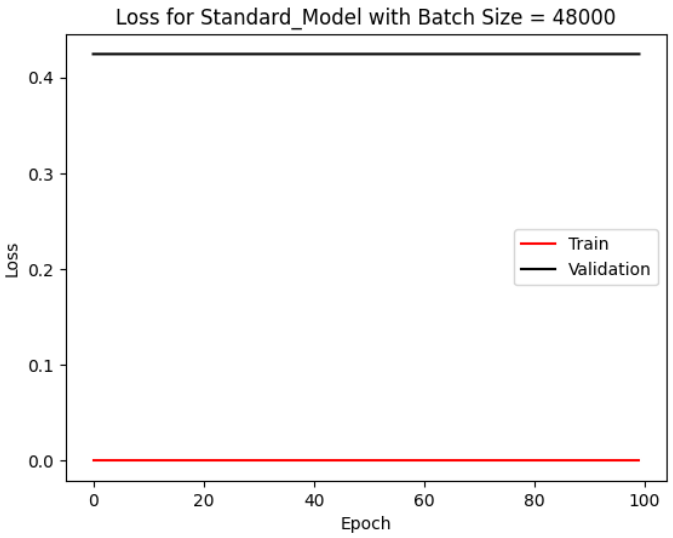


<u>Παρατηρήσεις/Αποτελέσματα:</u> Παρατηρούμε ότι το συγκεκριμένο μοντέλο πήρε πολύ λιγότερο χρόνο για να εκπαιδευτεί από το προηγούμενο (training time = 274,51 seconds), λόγω του αρκετά μεγαλύτερου batch size, όπως εξηγήθηκε παραπάνω. Επίσης, παρατηρούμε εξαιρετικά καλές επιδόσεις τόσο στην ακρίβεια όσο και στις απώλειες, χωρίς διακυμάνσεις στις τιμές κατα το πέρασμα των εποχών, κάτι που δείχνει ότι δεν έχουμε overfitting.

Μοντέλο 3

Χαρακτηριστικά: default δίκτυο, batch size = 48000 (N_{train}), όπου N_{train} είναι το συνολικό πλήθος δεδομένων εκπαίδευσης.

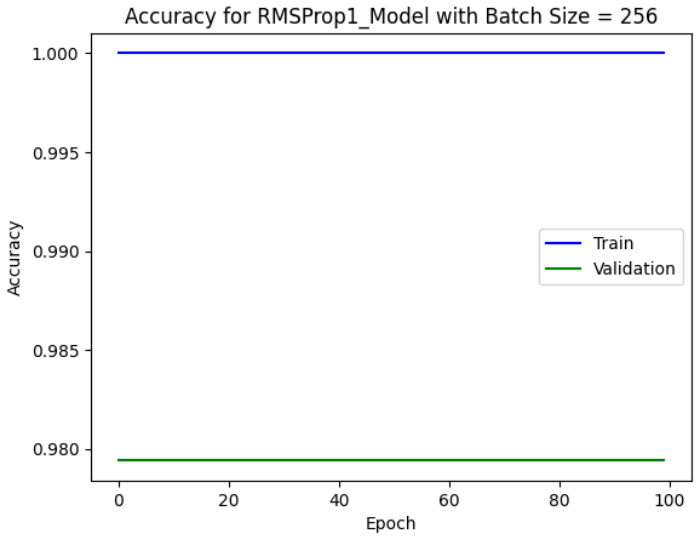


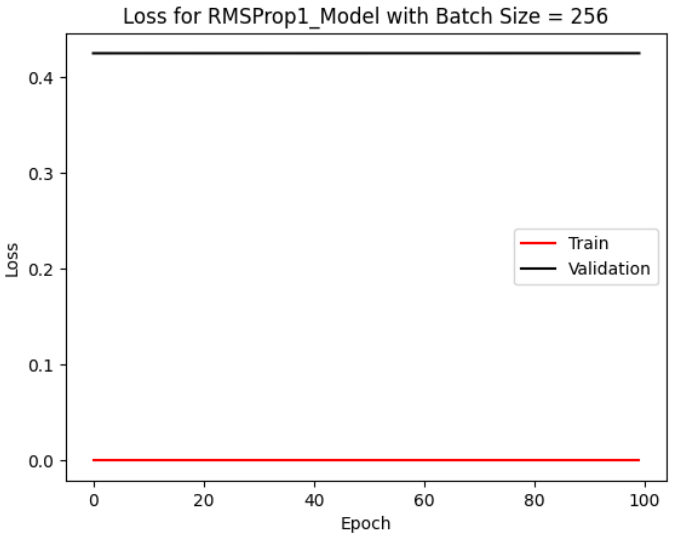


<u>Παρατηρήσεις/Αποτελέσματα:</u> Ακόμη ταχύτερη η εκπαίδευση του μοντέλου σε σύγκριση με τα προηγούμενα με training time = 29.03 seconds. Οι κυματομορφές παραμένουν ίδιες με του μοντέλου 2.

Μοντέλο 4

Χαρακτηριστικά: RMSProp optimizer, batch size = 256, ρ = 0.01, lr = 0.001.

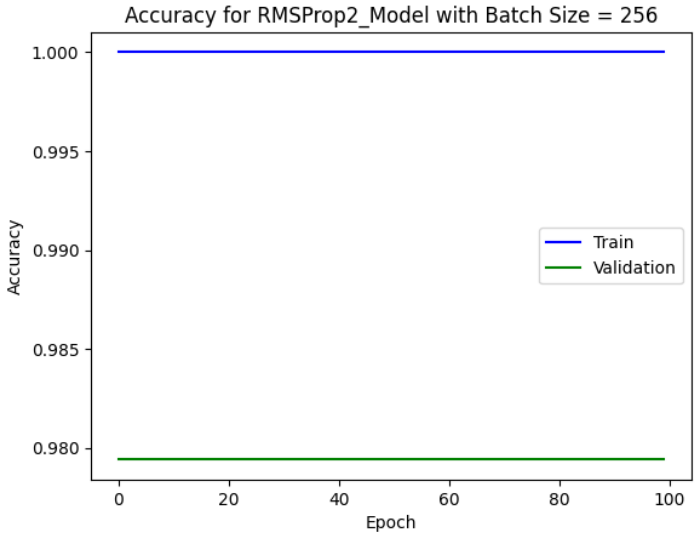


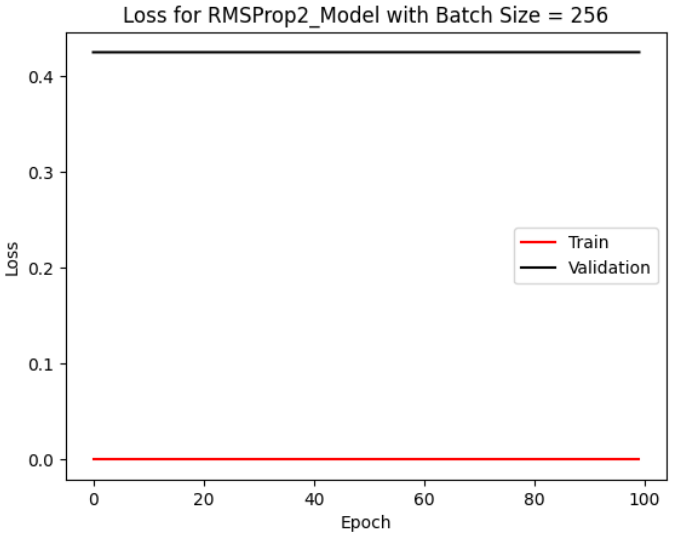


Παρατηρήσεις/Αποτελέσματα: Ο χρόνος εκπαίδευσης του μοντέλου είναι 250.44 seconds.

Μοντέλο 5

Χαρακτηριστικά: RMSProp optimizer, batch size = 256, ρ = 0.99, lr = 0.001.

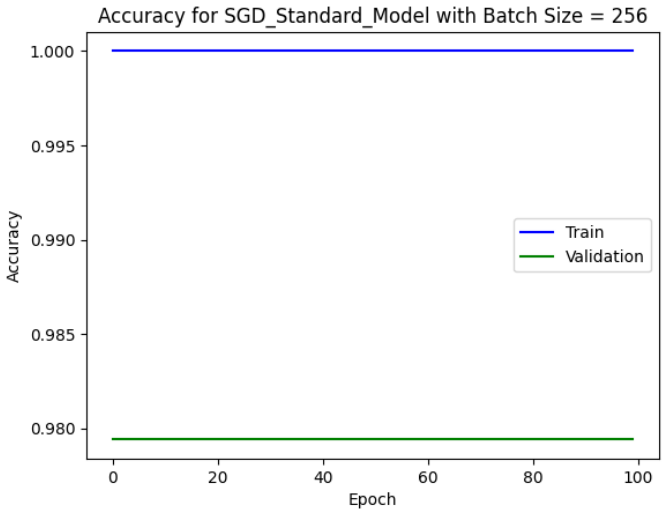




Παρατηρήσεις/Αποτελέσματα: Ο χρόνος εκπαίδευσης του μοντέλου είναι 246.89 seconds (ελάχιστα πιο γρήγορη εκπαίδευση από το προηγούμενο μοντέλο). Παρατηρούμε από τις κυματομορφές των μοντέλων 4 και 5 ότι δεν έχουμε κάποια μεταβολή, έχοντας αλλάξει το ρ από 0.01 σε 0.99. Αυτό συμβαίνει επειδή το μοντέλο έχει ταχεία προσαρμογή στα δεδομένα, πετυχαίνοντας πολυ υψηλές επιδόσεις ήδη από την αρχή των εποχών, κάτι που δεν αφήνει περιθώρια βελτίωσης.

Μοντέλο 6

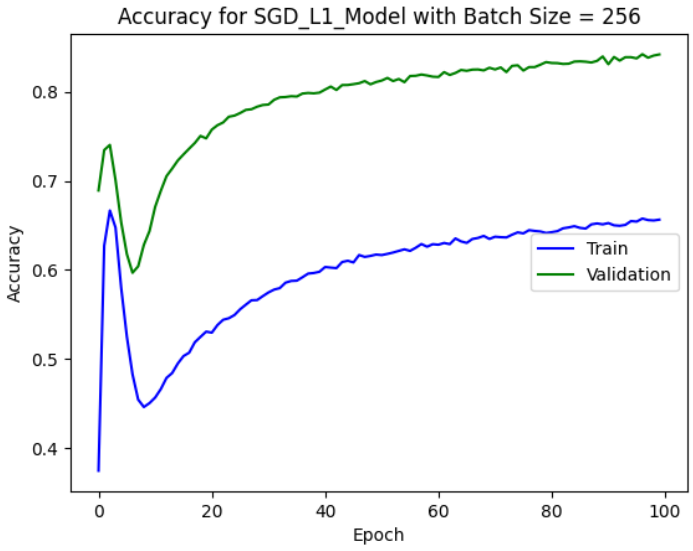
Χαρακτηριστικά: Standard SGD optimizer, batch size = 256, lr = 0.01, με αρχικοποίηση των βαρών κάθε στρώματος με βάση μια κανονική κατανομή με μέση τιμή 10.

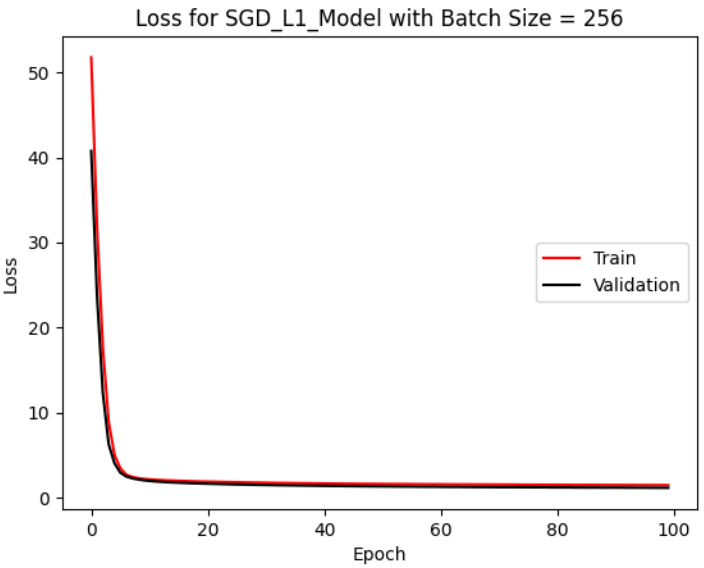


<u>Παρατηρήσεις/Αποτελέσματα:</u> Ο χρόνος εκπαίδευσης του μοντέλου είναι 200.26 seconds, δηλαδή αρκετά πιο γρήγορο από τον RMSProp optimizer για το ίδιο batch size.

<u>Μοντέλο 7</u>

Χαρακτηριστικά: SGD optimizer με L1 κανονικοποίηση, batch size = 256, lr = 0.01, α = 0.01 και dropout probability = 0.3.

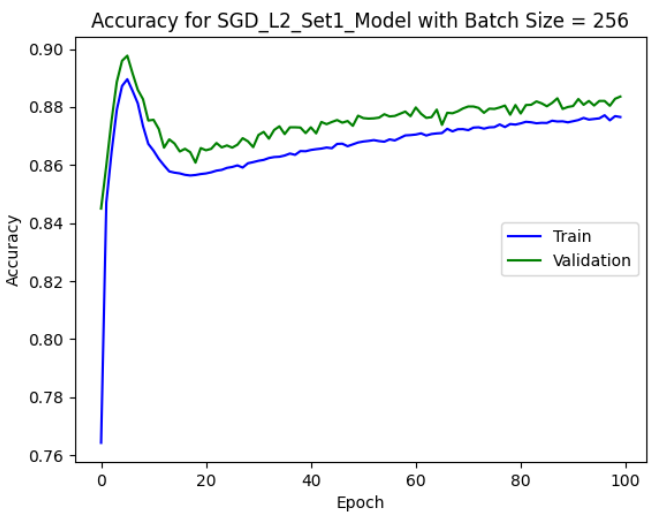


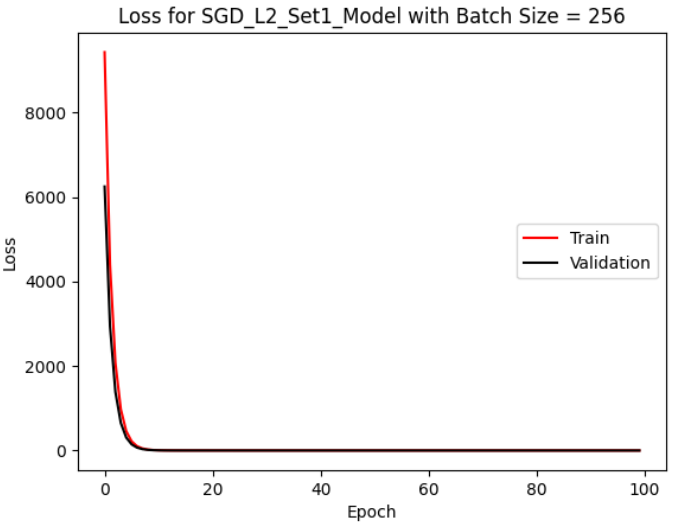


Παρατηρήσεις/Αποτελέσματα: Ο χρόνος εκπαίδευσης του μοντέλου είναι 258.39 seconds (περίπου ίδιος με τον RMSProp optimizer). Παρατηρούμε μια κυμάτωση στις κυματομορφές του accuracy για τα Train και Validation στην αρχή των εποχών αλλά σιγά σιγά το μοντέλο προσαρμόζεται και επιτυγχάνει όλο και μεγαλύτερη ακρίβεια. Ίσως, αν το εκπαιδεύαμε για περισσότερες εποχές να είχαμε ακόμα υψηλότερες επιδόσεις. Σχετικά με τις κυματομορφές των απωλειών στην αρχή έχουμε υψηλές απώλειες, ωστόσο με το πέρασμα των εποχών οι απώλειες μειώνονται πολύ γρήγορα μέχρι να σταθεροποιηθούν.

Μοντέλο 8

Χαρακτηριστικά: SGD optimizer με L2 κανονικοποίηση, batch size = 256, lr = 0.01, α = 0.1.

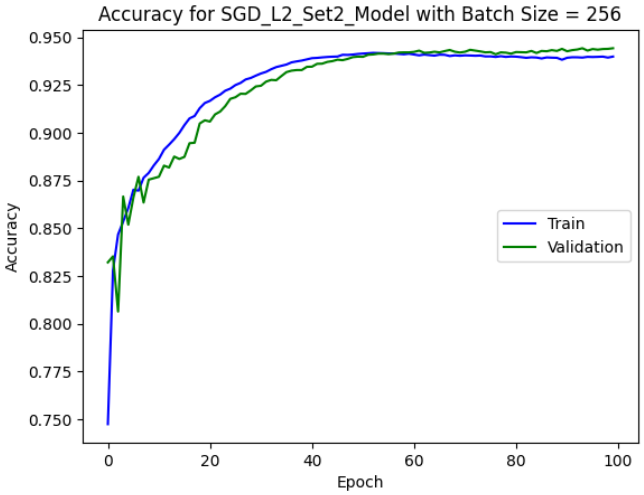


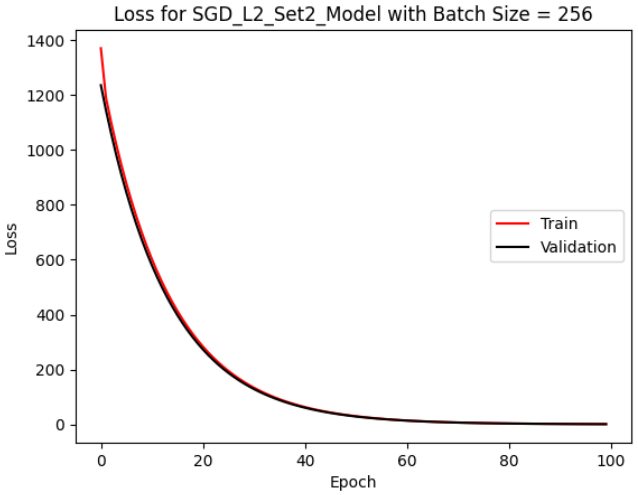


Παρατηρήσεις/Αποτελέσματα: Ο χρόνος εκπαίδευσης του μοντέλου είναι 214.57 seconds (πιο γρήγορο από την L1 κανονικοποίηση). Παρατηρούμε υψηλότερα ποσοστά ακρίβειας αλλά και αρκετά υψηλότερα ποσοστα απωλειών σε σύγκριση με την L1 κανονικοποίηση, στην αρχή των εποχών. Παρόλα αυτά το μοντέλο προσαρμόζεται και πάλι σχετικά γρήγορα και βελτιώνεται συνεχώς με την πάροδο των εποχών.

Μοντέλο 9

Χαρακτηριστικά: SGD optimizer με L2 κανονικοποίηση, batch size = 256, lr = 0.01, α = 0.01.

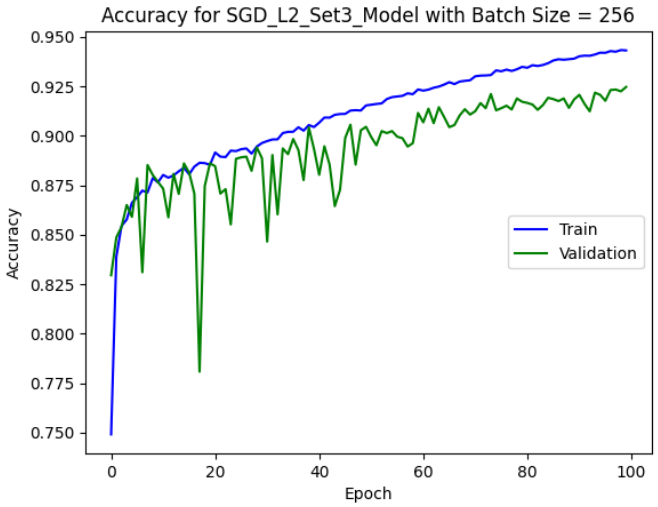


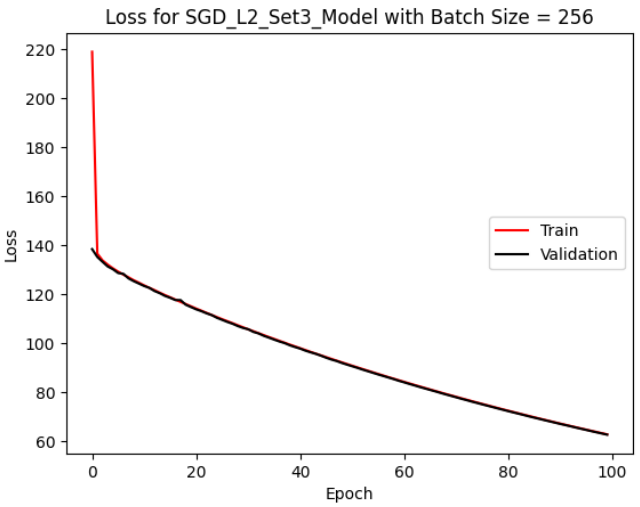


<u>Παρατηρήσεις/Αποτελέσματα:</u> Ο χρόνος εκπαίδευσης του μοντέλου είναι 205.49 seconds (ελάχιστα πιο γρήγορο από το μοντέλο 8). Παρατηρούμε ότι το μοντέλο προσαρμόζεται σταδιακά και ομαλά σε σύγκριση με τα 2 προηγούμενα, καθώς και έχει μεγαλύτερη ακρίβεια και λιγότερες απώλειες, σε σύγκριση με το μοντέλο 8.

Μοντέλο 10

Χαρακτηριστικά: SGD optimizer με L2 κανονικοποίηση, batch size = 256, lr = 0.01, α = 0.001.





Παρατηρήσεις/Αποτελέσματα: Ο χρόνος εκπαίδευσης του μοντέλου είναι 210.30 seconds. Παρατηρούμε, ότι το μοντέλο αυτό μετά από λίγες εποχές στην αρχή, προσαρμόζεται και βελτιώνεται (σχεδόν) γραμμικά στην κυματομορφή Accuracy για το Train, πετυχαίνοντας υψηλές επιδόσεις. Επίσης, παρατηρούμε μια διακύμανση των τιμών της κυματομορφής του Validation, κάτι που οφείλεται στον τυχαίο τρόπο με τον οποίο αφαιρούνται ή όχι από κάθε κρυφό στρώμα νευρώνες. Τέλος, στο διάγραμμα των απωλειών, η κυματομορφή του Loss για το Train μετά από ένα σημείο μειώνεται επίσης γραμμικά και έχουμε λιγότερες απώλειες σε σύγκριση με το μοντέλο 9.

Σύνοψη

Οι χρόνοι εκπαίδευσης των μοντέλων συνοψίζονται στον παρακάτω πίνακα.

Αριθμός Μοντέλου	Όνομα Μοντέλου	Χρόνος εκπαίδευσης (seconds)
1	Standard_Model, batch size = 1	15338.87
2	Standard_Model, batch size = 256	274.51
3	Standard_Model, batch size = 48000	29.03
4	RMSProp1_Model, batch size = 256	250.44
5	RMSProp2_Model, batch size = 256	246.89
6	SGD_Standard_Model, batch size = 256	200.26
7	SGD_L1_Model, batch size = 256	258.39
8	SGD_L2_Set1_Model, batch size = 256	214.57
9	SGD_L2_Set2_Model batch size = 256	205.49
10	SGD_L2_Set3_Model batch size = 256	210.30

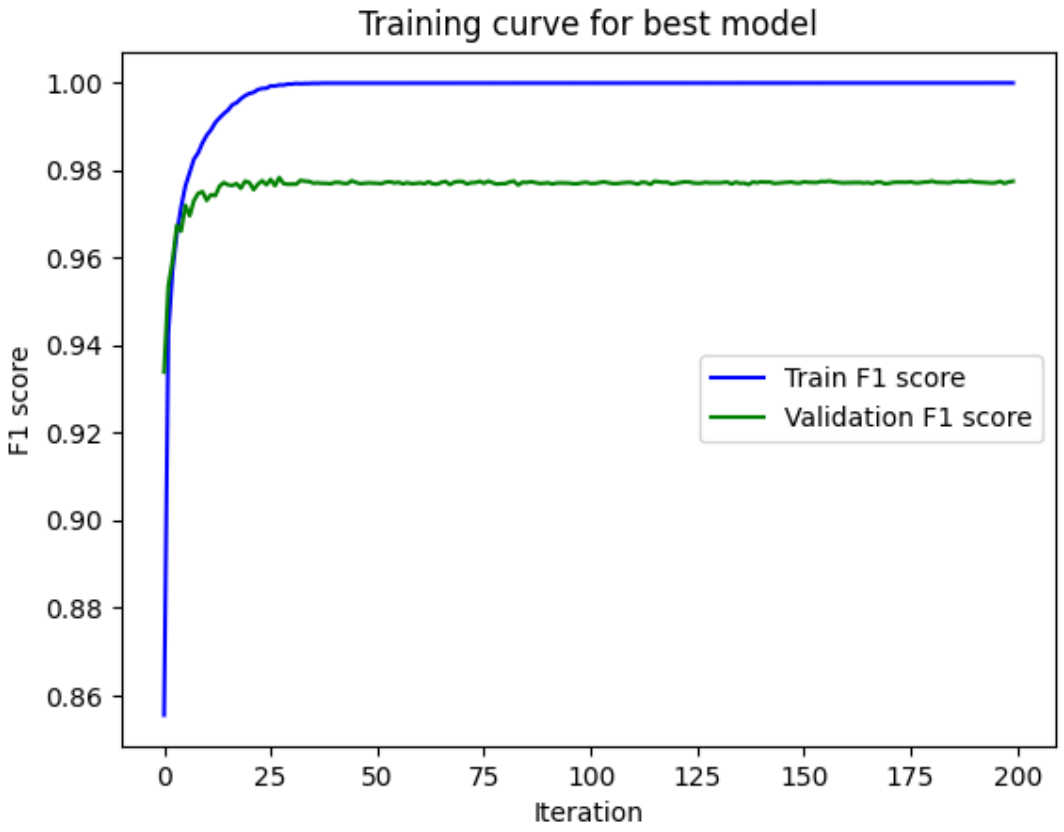
Fine Tuning

Συνοψίζονται οι υπερπαράμετροι του καλύτερου μοντέλου στον παρακάτω πίνακα:

Layer 1 size	Layer 2 size	Learning Rate	Loss	Accuracy	Recall Score	Precision Score	f1 score
128	512	0.1	0.0906	0.97979	0.97991	0.97933	0.98051

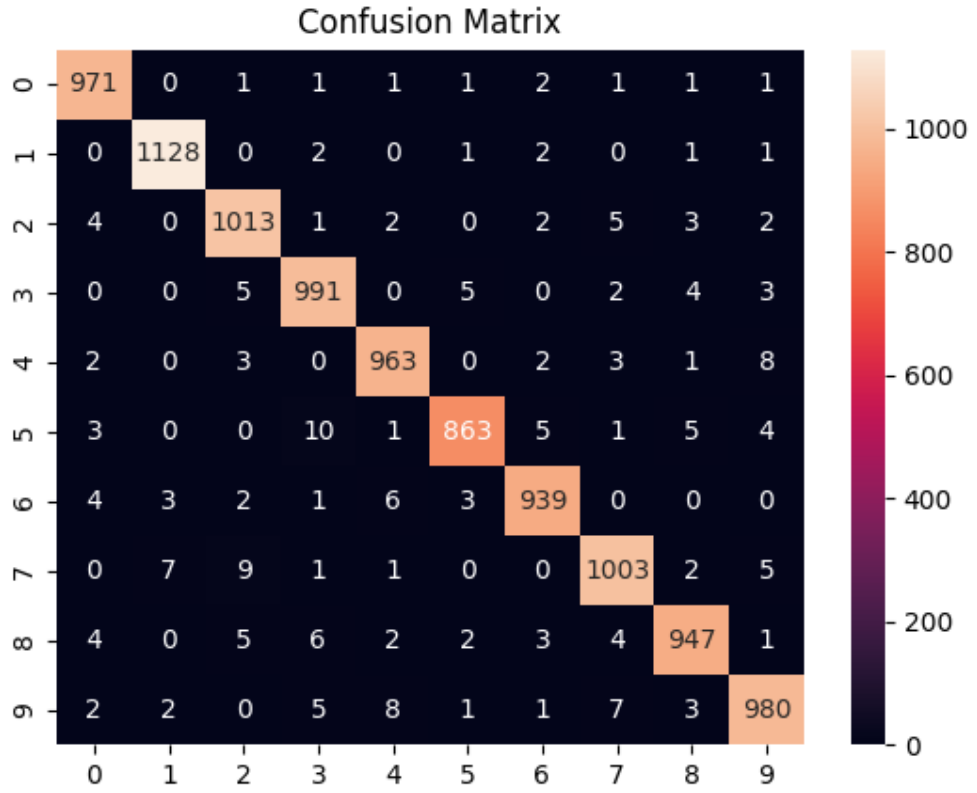
Παρατηρούμε ότι ο αριθμός των νευρώνων (128, 512), που επιλέγει ο tuner και στα 2 στρώματα, είναι ο μεγαλύτερος για τις τιμές του πλήθους νευρώνων που μας δίνονται για τα κρυφά στρώματα 1 και 2 αντίστοιχα. Φυσικά, κάτι τέτοιο ήταν αναμενόμενο, καθώς όσο μεγαλύτερο είναι το πλήθος νευρώνων στο δίκτυο, τόσο μεγαλύτερη είναι η ικανότητα προσαρμογής του στα δεδομένα. Όσον αφορά τις υπόλοιπες μετρικές, έχουμε αρκετά χαμηλές απώλειες και υψηλά ποσοστά για recall, Precision και f1 score, κάτι που είναι θεμιτό και καθιστά την εκπαίδευση του μοντέλου με fine tuning αποτελεσματική.

Παρακάτω παρουσιάζονται οι καμπύλες εκμάθησης και επικύρωσης για τις μετρικές του καλύτερου μοντέλου:



Από το παραπάνω διάγραμμα, παρατηρούμε ότι για μικρές εποχές, έχουμε ραγδαία αύξηση της κυματομορφής Validation κάτι που μπορεί να οδηγήσει σε overfitting. Ωστόσο κάτι τέτοιο αποτρέπεται, διότι κατά την εκπαίδευση του μοντέλου, έχουμε ορίσει να γίνεται early stopping. Έτσι, η εκπαίδευση του μοντέλου γίνεται αρκετά πιο γρήγορη και για περίπου 200 εποχές, αντί για 1000 που είχε αρχικά οριστεί. Ως αποτέλεσμα, γίνεται πρόληψη του overfitting και βελτιώνεται η απόδοση του μοντέλου.

Στη συνέχεια παρουσιάζεται ο πίνακας σύγχυσης, με οριζόντιο άξονα το y predicted label (πιθανό y) και κατακόρυφο άξονα το y actual label (αληθές y):



Σκοπός του πίνακα σύγχυσης είναι να μας δίνει τη δυνατότητα να δούμε πόσα στοιχεία του test set έχουν ταξινομηθεί σωστά και πόσα λανθασμένα. Το γεγονός ότι τα περισσότερα στοιχεία βρίσκονται πάνω στην κύρια διαγώνιο του πίνακα, υποδηλώνει ότι η διαδικασία ταξινόμησης των στοιχείων έγινε σωστά. Ωστόσο, πρέπει να σημειωθεί πως υπάρχει μία μικρή μειοψηφία στοιχείων τα οποία ταξινομούνται λανθασμένα, κάτι που υποδηλώνει ότι υπάρχει ακόμα χώρος για περαιτέρω βελτίωση στην απόδοση του μοντέλου. Οι αριθμοί στο μαύρο πλαίσιο του πίνακα σύγχυσης μας δείχνουν τον αριθμό των ψηφίων που ταξινομήθηκαν λανθασμένα και η θέση τους στο μαύρο πλαίσιο, μας δείχνει το ψηφίο χ του κατακόρυφου άξονα που ταξινομήθηκε λανθασμένα ως το ψηφίο ψ του οριζόντιου

άξονα. Για παράδειγμα, στην γραμμή με το ψηφίο 5 του κατακόρυφου άξονα και στην στήλη με το ψηφίο 3 του οριζόντιου άξονα, το ψηφίο 5 ταξινομείται λανθασμένα ως 3, 10 φορές. Παρομοίως και για τις υπόλοιπες περιπτώσεις.