Υπολογιστική Νοημοσύνη

Εργασία 3

MLP

Κουτρουμπής Γεώργιος, ΑΕΜ: 9668

2022

Περιεχόμενα

1 Διερεύνηση Μοντέλων

1.1 Διαφορετικά Batch Sizes

Batch Size	Training Time (s)
1	13762.74
256	82.47s
60000	13.45

Πίνακας 1. *Ταχύτητα εκπαίδευσης για διαφορετικά batch sizes*

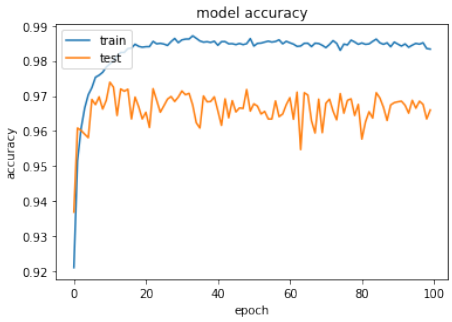
Σε αυτήν την περίπτωση εκπαιδεύτηκε ενα MLP δίκτυο με 3 διαφορετικές επιλογές batch size: online, mini-batch και full batch. Σε κάθε εποχή (epoch) της εκπαίδευσης του νευρωνικού δικτύου, εκτελούνται κάποια βήματα forward pass - backpropagation pass, όπου ανανεώνονται τα βάρη.

Στην περίπτωση του full-batch γίνεται μόνο ένα βήμα, καθώς γίνεται χρήση όλου του training dataset.
Στην περίπτωση του mini-batch γίνονται $dataset\_size/mini\_batch\_size$ βήματα, καθώς γίνεται χρήση ενός μέρους του dataset σε κάθε iteration.
Στην περίπτωση του online γίνονται dataset_size καθώς σε κάθε iteration χρησιμοποιείται μόνο παράδειγμα του dataset.

Στον πίνακα 1 φαίνονται οι χρόνοι εκπαίδευσης μοντέλων που έγινε χρήση full-batch, mini-batch και online μεθόδων. Όπως είναι αναμενόμενο, εφόσον μεταβάλλεται ο αριθμός των iterations που απαιτούνται για να ολοκληρωθεί μια εποχή ανάλογα την επιλεγμένη μέθοδο, η online μέθοδος πήρε τον περισσότερο χρόνο για να εκπαιδευτεί το μοντέλο, ακολουθώντας η mini-batch και τέλος full-batch.

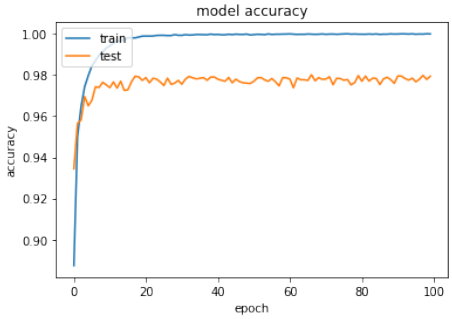
Σημείωση: για full-batch επιλογή μεγέθους batch size στην πραγματικότητα ήταν batch_size = 48000, καθώς ένα 20% των δεδομένων χρησιμοποιήθηκε ως validation set.

1.2 Καμπύλες ακρίβειας και κόστους

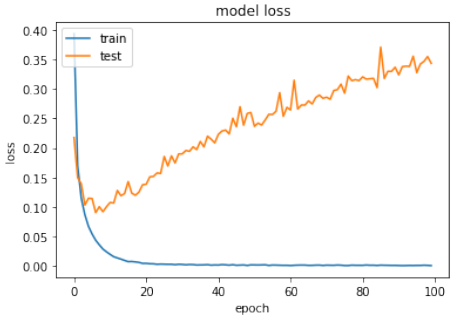


Εικόνα 1.
Καμπύλη ακρίβειας για batch_size=1

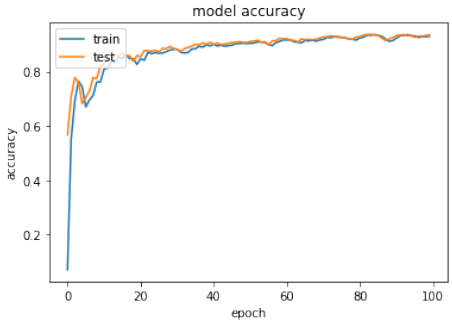
Εικόνα 2.
Καμπύλη κόστους για batch_size=1



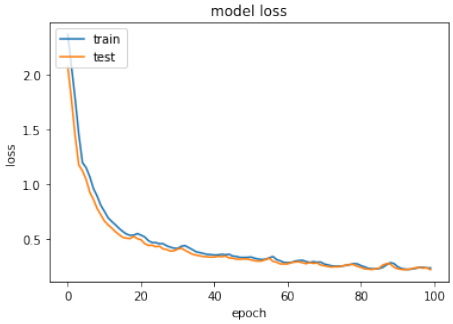
Εικόνα 3.
Καμπύλη ακρίβειας για batch_size=256



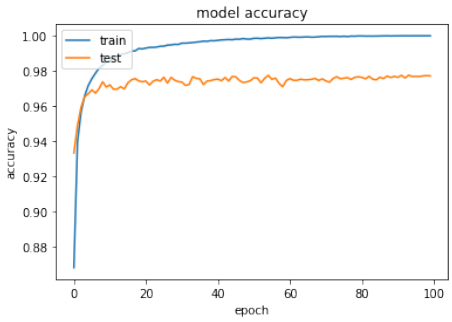
Εικόνα 4.
Καμπύλη κόστους για batch_size=256



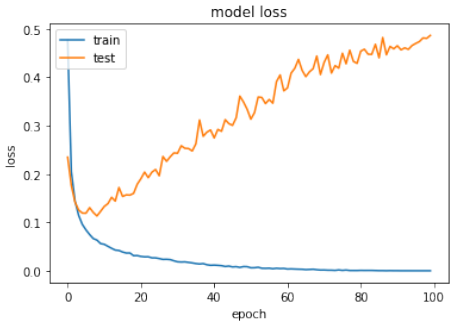
Εικόνα 5.
Καμπύλη ακρίβειας για batch_size=n_train



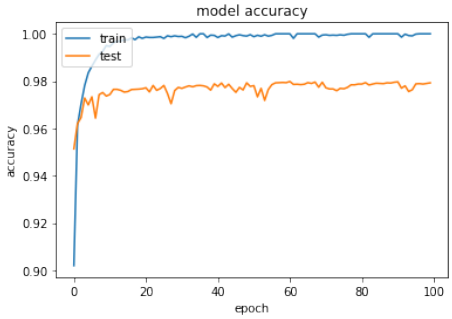
Εικόνα 6.
Καμπύλη κόστους για batch_size=n_train



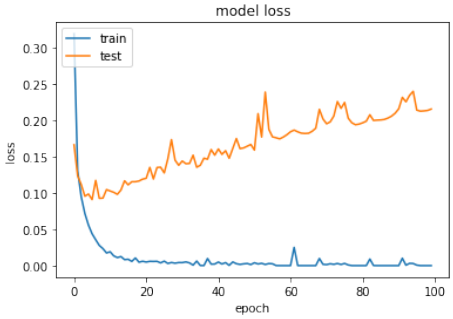
Εικόνα 7.
Καμπύλη ακρίβειας για RMSProp, ρ=0.01



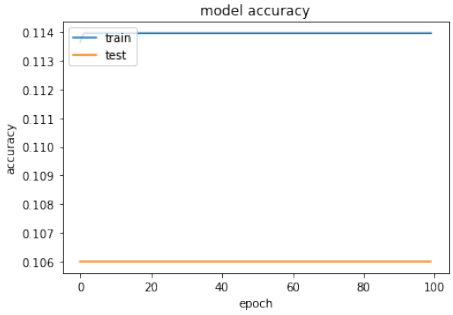
Εικόνα 8.
Καμπύλη κόστους για RMSProp, ρ=0.01



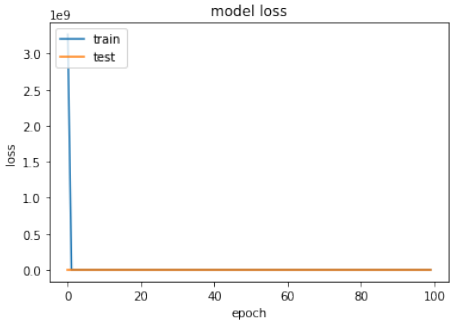
Εικόνα 9.
Καμπύλη ακρίβειας για RMSProp, ρ=0.99



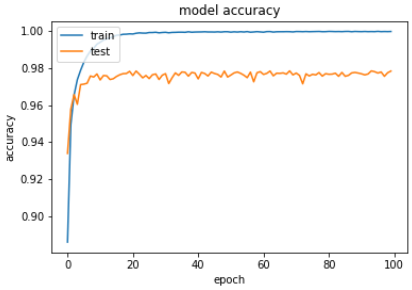
Εικόνα 10.
Καμπύλη κόστους για RMSProp, ρ=0.99



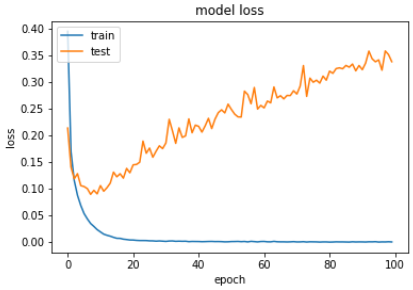
Εικόνα 11.
Καμπύλη ακρίβειας για SGD και αρχικοποίηση



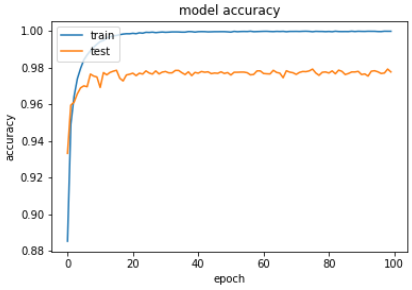
Εικόνα 12.
Καμπύλη κόστους για SGD και αρχικοποίηση



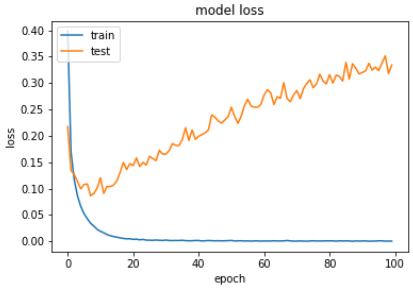
Εικόνα 13.
Καμπύλη ακρίβειας για L2 (a=1e-3), batch_size=256



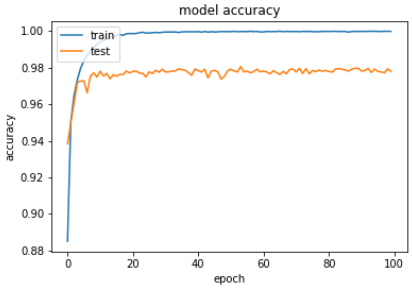
Εικόνα 14.
Καμπύλη κόστους για L2 (a=1e-3), batch_size=256



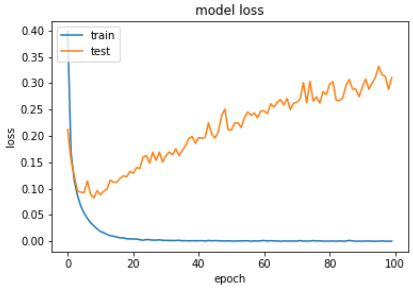
Εικόνα 15.
Καμπύλη ακρίβειας για L2 (a=1e-2), batch_size=256



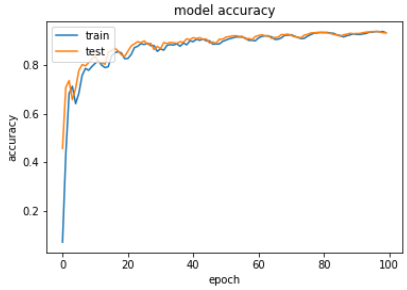
Εικόνα 16.
Καμπύλη κόστους για L2 (a=1e-2), batch_size=256



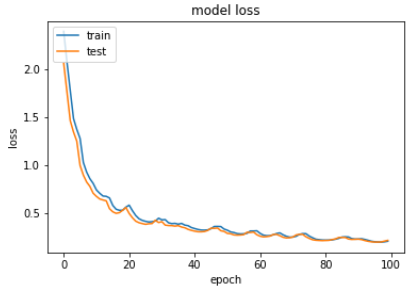
Εικόνα 17.
Καμπύλη ακρίβειας για L2 (a=1e-1), batch_size=256



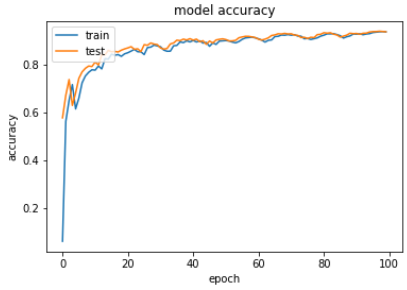
Εικόνα 18.
Καμπύλη κόστους για L2 (a=1e-1), batch_size=256



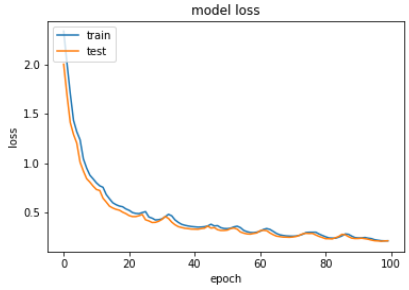
Εικόνα 19.
Καμπύλη ακρίβειας για L2 (a=1e-3), batch_size=n_train



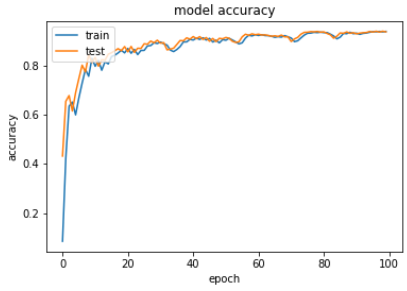
Εικόνα 20.
Καμπύλη κόστους για L2 (a=1e-3), batch_size=n_train



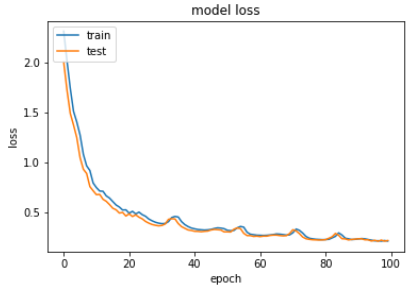
Εικόνα 21.
Καμπύλη ακρίβειας για L2 (a=1e-2), batch_size=n_train



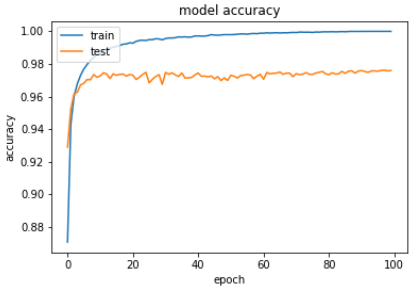
Εικόνα 22.
Καμπύλη κόστους για L2 (a=1e-2), batch_size=n_train



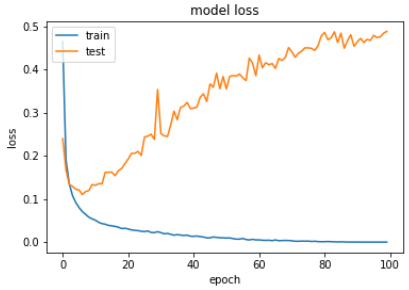
Εικόνα 23.
Καμπύλη ακρίβειας για L2 (a=1e-1), batch_size=n_train



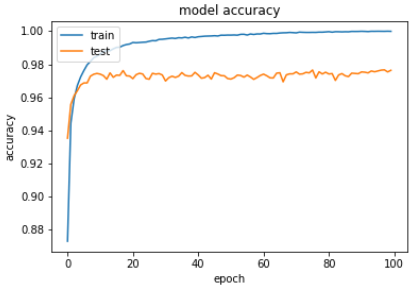
Εικόνα 24.
Καμπύλη κόστους για L2 (a=1e-1), batch_size=n_train



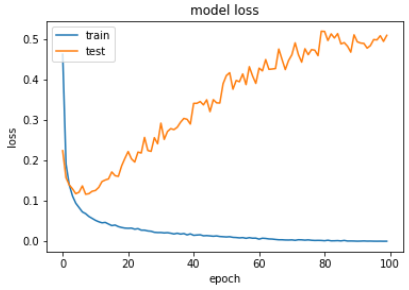
Εικόνα 25.
Καμπύλη ακρίβειας για L2 (a=1e-3), RM-SProp ρ=1e-2



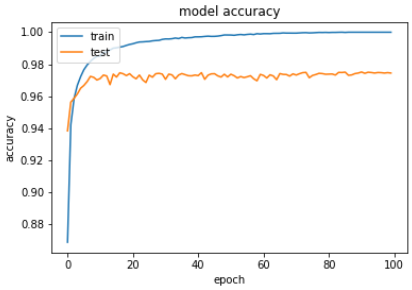
Εικόνα 26.
Καμπύλη κόστους για L2 (a=1e-3), RM-SProp ρ=1e-2



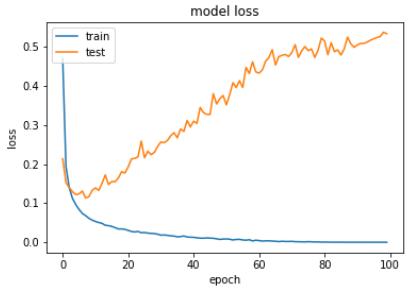
Εικόνα 27.
Καμπύλη ακρίβειας για L2 (a=1e-2), RM-SProp ρ=1e-2



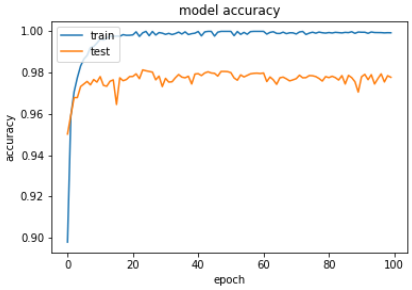
Εικόνα 28.
Καμπύλη κόστους για L2 (a=1e-2), RM-SProp ρ=1e-2



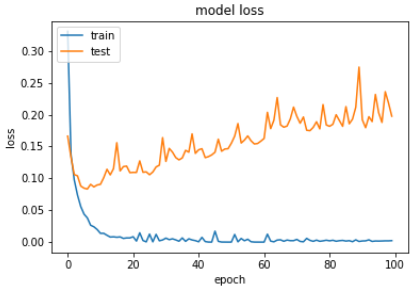
Εικόνα 29.
Καμπύλη ακρίβειας για L2 (a=1e-1), RM-SProp ρ=1e-2



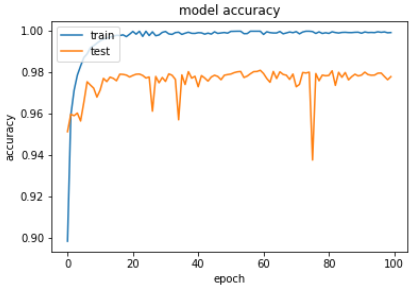
Εικόνα 30.
Καμπύλη κόστους για L2 (a=1e-1), RM-SProp ρ=1e-2



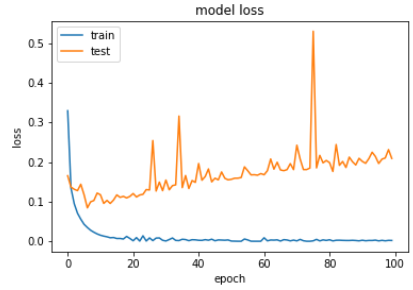
Εικόνα 31.
Καμπύλη ακρίβειας για L2 (a=1e-3), RM-SProp ρ=0.99



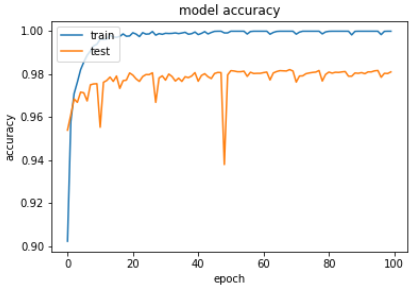
Εικόνα 32.
Καμπύλη κόστους για L2 (a=1e-3), RM-SProp ρ=0.99



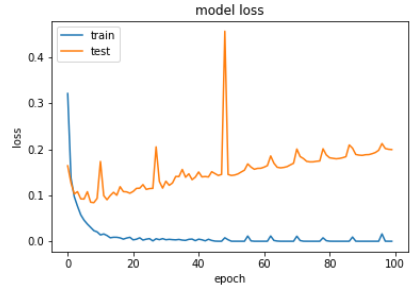
Εικόνα 33.
Καμπύλη ακρίβειας για L2 (a=1e-2), RM-SProp ρ=0.99



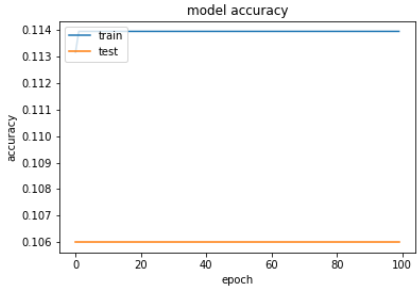
Εικόνα 34.
Καμπύλη κόστους για L2 (a=1e-2), RM-SProp ρ=0.99



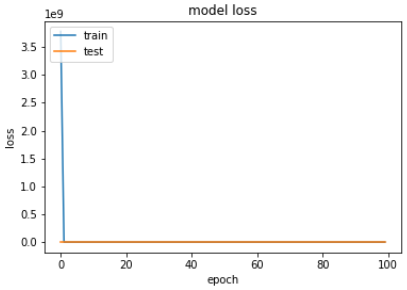
Εικόνα 35.
Καμπύλη ακρίβειας για L2 (a=1e-1), RM-SProp ρ=0.99



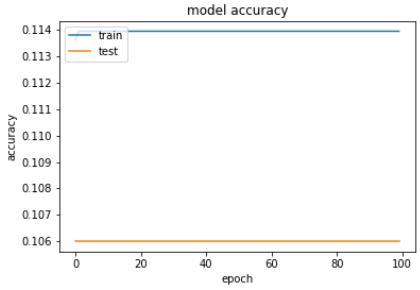
Εικόνα 36.
Καμπύλη κόστους για L2 (a=1e-1), RM-SProp ρ=0.99



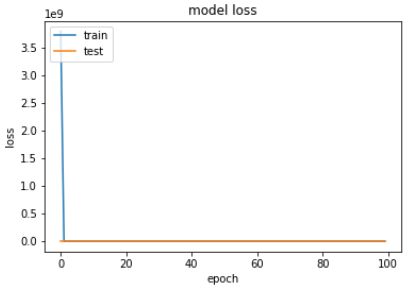
Εικόνα 37.
Καμπύλη ακρίβειας για L2 (a=1e-3), SGD με αρχικοποίηση



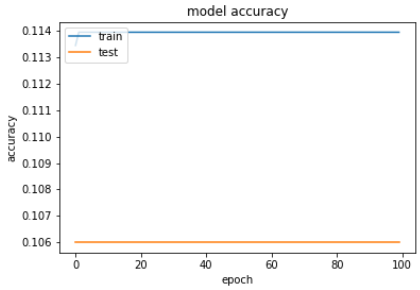
Εικόνα 38.
Καμπύλη κόστους για L2 (a=1e-3), SGD με αρχικοποίηση



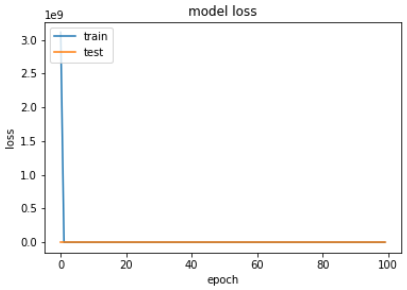
Εικόνα 39.
Καμπύλη ακρίβειας για L2 (a=1e-2), SGD με αρχικοποίηση



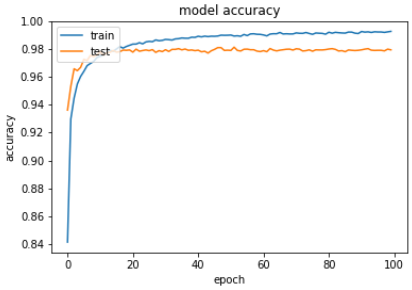
Εικόνα 40.
Καμπύλη κόστους για L2 (a=1e-2), SGD με αρχικοποίηση



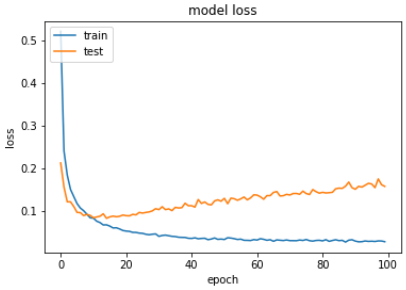
Εικόνα 41.
Καμπύλη ακρίβειας για L2 (a=1e-1), SGD με αρχικοποίηση



Εικόνα 42.
Καμπύλη κόστους για L2 (a=1e-1), SGD με αρχικοποίηση



Εικόνα 43.
Καμπύλη ακρίβειας για L1 με dropout



Εικόνα 44.
Καμπύλη κόστους για L1 με dropout

1.3 Σχολιασμός Αποτελεσμάτων

Για τα 3 μοντέλα που έχουν διαφορετικά batch sizes και τις default παραμέτρους της βιβλιοθήκης keras, παρατηρούμε ότι για online εκμάθηση (batch_size = 1) και για mini-batch υπάρχει overfitting, με το ποσοστό accuracy όμως στο validation set να είναι υψηλό. Αντίστοιχα, για full batch μέθοδο, οι καμπύλες των train και validation sets συμβαδίζουν και συγκλίνουν σε υψηλό ποσοστό accuracy. Αυτό παρατηρείται καθώς έγινε χρήση συγκεκριμένου αριθμού εποχών (epochs), που σημαίνει ότι οι μέθοδοι με μικρότερα batch sizes, καθώς ανανεώνουν πολλαπλές φορές τα βάρη τους ανά εποχή, φτάνουν και πιο γρήγορα σε overfitting, σε αντίθεση με το full batch που θα χρειαστεί παραπάνω εποχές για να φτάσει σε overfitting.
Για τις online και mini-batch μεθόδους, αυτό θα μπορούσε να διορθωθεί με χρήση early stopping αντί συγκεκριμένου αριθμού εποχών, ώστε να σταματήσει η εκπαίδευση όταν σταματήσει να παρατηρείται σημαντική βελτίωση μεταξύ των εποχών.
Για τα μοντέλα που χρησιμοποιούν RMSProp ως optimizer, παρατηρούμε ότι για rho=0.99 έχουμε καλύτερα αποτελέσματα από ότι για rho=0.01, γεγονός που βρίσκεται σύμφωνο με τη βιβλιογραφία που προτείνει τιμή rho=0.9.
Παρατηρούμε ότι στο μοντέλο που γίνεται χρήση SGD αλλά και αρχικοποίηση των συναπτικών βαρών με normal distribution που έχει mean = 10, πρακτικά δεν εκπαιδεύεται το μοντέλο. Αυτό συνέβη καθώς η αρχικοποίηση των βαρών ήταν πολύ μακριά από τιμές που θα ελαχιστοποιούσαν την παράγωγο (στην καμπύλη κόστους βλέπουμε ότι στην αρχή το κόστος είναι πολύ μεγάλο), οπότε δεν έγινε κάποιο ουσιαστικό βήμα, εκτός της πρώτης εποχής, όπου έπεσε πολύ σημαντικά το κόστος. Το ίδιο συμβαίνει και με χρήση L2 regularization.
Παρατηρούμε ότι στα περισσότερα μοντέλα υπάρχει overfitting, που με τη χρήση L2-regularization δεν διορθώνεται σε μεγάλο βαθμό. Όπως αναφέρθηκε και πιο πάνω, επειδή έχουμε συγκεκριμένο αριθμό εποχών εκπαίδευσης για τα μοντέλα, και αυτά φτάνουν σε υψηλά ποσοστά ακρίβειας σχετικά νωρίς στην εκπαίδευσή τους, παρατηρείται το overfitting, σε διαφορετικό βαθμό για κάθε μοντέλο.
Αντιθέτως, όμως, στο μοντέλο με L1-regularization και χρήση dropout, επιτυγχάνονται υψηλά ποσοστά, και οι καμπύλες μεταξύ των δεδομένων εκπαίδευσης και επικύρ-

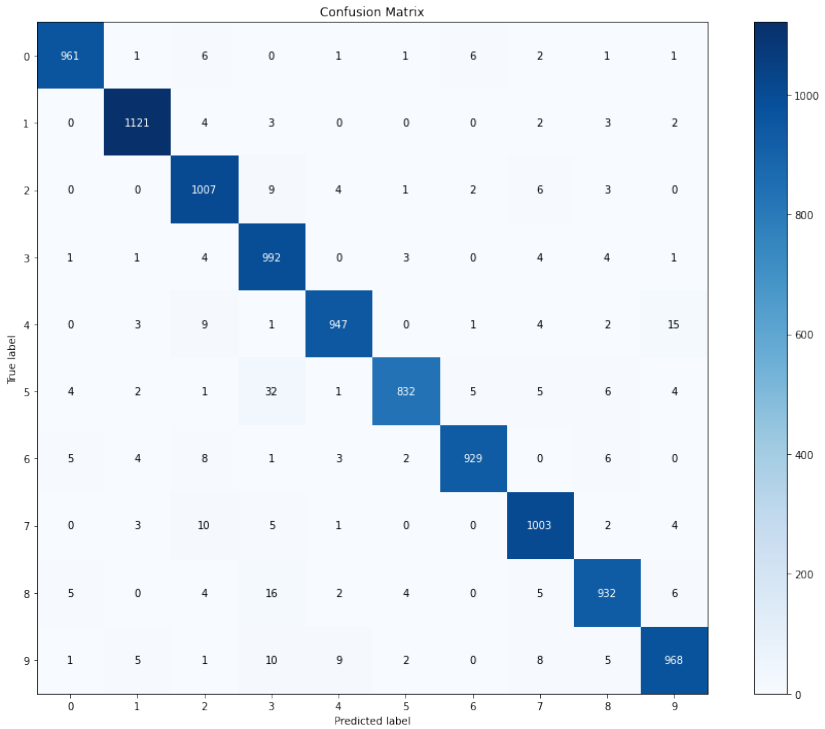
ωσης είναι πολύ κοντά μεταξύ τους. Από αυτό, θα μπορούσαμε να βγάλουμε 2 συμπεράσματα για τα μοντέλα μας: Αρχικά για το L1-regularization οτί μπορεί να συμβάλει στην εξάλειψη του overfitting καθώς εξορισμού προσθέτει ένα είδος feature-selection στο μοντέλο, αφού πλέον τα βάρη μπορούν να πάρουν μηδενικές τιμές, ενώ το dropout μπορεί να μας φανερώνει ότι χρησιμοποιήσαμε πολλούς νευρώνες στα κρυφά στρώματα του μοντέλου, και αυτά είναι που προκάκλεσαν το overfitting (μιας και το dropout απενεργοποιεί τυχαία νευρώνες του δικτύου).

2 Fine Tuning Δικτύου

Με τη χρήση του Keras Tuner και της μεθόδου Hyperband, έχοντας ως objective function το F-measure του συνόλου επικύρωσης, επιλέχθηκαν οι εξής υπερπαράμετροι, όπως φαίνεται στον πίνακα 2.

1st Layer Neurons	2nd Layer Neurons	RMSProp Learning Rate	L2 alpha
128	512	1e-3	1e-6

Πίνακας 2. *Υπερπαράμετροι τελικού μοντέλου*



Εικόνα 45.
Πίνακας σύγχυσης του τελικού μοντέλου

Από τον πίνακα σύγχυσης (2), μπορούμε να εξάγουμε το precision, recall και F1-Score για κάθε κλάση, και έπειτα το macro average αυτών των μετρικών.
Θα αναλυθεί η εύρεση αυτών των μετρικών για τις κλάσεις '0' και '4'.

Για την κλάση '0', δηλαδή όταν το ψηφίο είναι το '0', βλέπουμε ότι έχουν γίνει 961 σωστές προβλέψεις ότι το ψηφίο είναι το '0' (true positive, κελί [0,0]), 16 λανθασμένες (false positive, άθροισμα στήλης 0, χωρίς το κελί [0,0]) ότι είναι το '0' (ενώ δεν ήταν) και 19 προβλέψεις ότι δεν είναι το '0' (false negative, άθροισμα σειράς 0, χωρίς το κελί [0,0]) (ενώ ήταν).

Το precision ορίζεται ως:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

δηλαδή η μετρική της αναλογίας σωστών προβλέψεων για την κλάση προς όλων των προβλέψεων ότι ένα δείγμα των δεδομένων ανήκει σε αυτή την κλάση. Το recall ορίζεται ως:

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

δηλαδή η μετρική της αναλογίας σωστών προβλέψεων για την κλάση προς το άθροισμα του αριθμού των δειγμάτων που ανήκουν σε αυτή την κλάση.
Το F1-Score ορίζεται ως:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

δηλαδή ουσιαστικά δίνει μια ισορροπημένη μετρική των precision recall.

Σύμφωνα με τα παραπάνω, για την κλάση '0' έχουμε:

$$TP_0 = 961, FP_0 = 16, FN_0 = 19$$

$$Precision_0 = \frac{961}{961 + 16} = 0.98$$

$$Recall_0 = \frac{961}{961 + 19} = 0.98$$

$$F1_0 = 2 * \frac{0.98 * 0.98}{0.98 + 0.98} = 0.98$$

Αντίστοιχα για την κλάση '4', δηλαδή όταν το ψηφίο είναι το '4', βλέπουμε ότι έχουν γίνει 947 σωστές προβλέψεις ότι το ψηφίο είναι το '4' (true positive, κελί [4,4]), 21 λανθασμένες (false positive, άθροισμα στήλης 4, χωρίς το κελί [4,4]) ότι είναι το '4' (ενώ δεν ήταν) και 35 προβλέψεις ότι δεν είναι το '4' (false negative, άθροισμα σειράς 4, χωρίς το κελί [4,4]) (ενώ ήταν).

Σύμφωνα με τα παραπάνω, για την κλάση '4' έχουμε:

$$TP_4 = 947, FP_4 = 21, FN_4 = 35$$

$$Precision_4 = \frac{947}{947 + 21} = 0.98$$

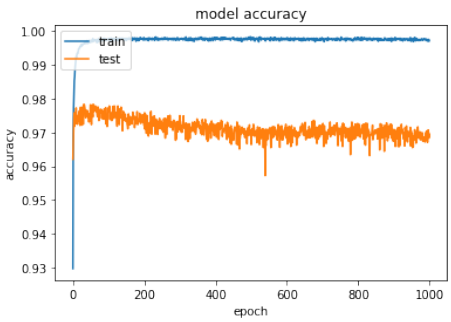
$$Recall_4 = \frac{947}{947 + 35} = 0.96$$

$$F1_4 = 2 * \frac{0.98 * 0.96}{0.98 + 0.96} = 0.97$$

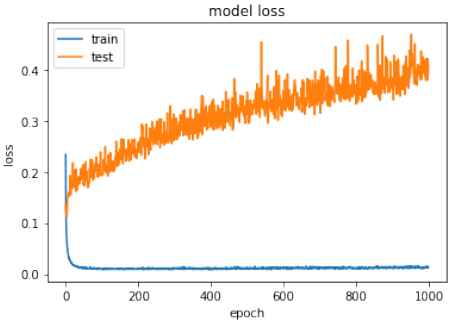
Συνοπτικά παρουσιάζονται στον παρακάτω πίνακα (3) οι μετρικές για κάθε κλάση, καθώς και το macro average αυτών:

Class	Precision	Recall	F1-Score
0	0.98	0.98	0.98
1	0.98	0.99	0.99
2	0.96	0.98	0.97
3	0.93	0.98	0.95
4	0.98	0.96	0.97
5	0.98	0.93	0.96
6	0.99	0.97	0.98
7	0.97	0.98	0.97
8	0.97	0.96	0.96
9	0.97	0.96	0.96
Avg	0.97	0.97	0.97

Πίνακας 3. *Precision, Recall F1-Score*



Εικόνα 46.
Καμπύλη ακρίβειας τελικού μοντέλου



Εικόνα 47.
Καμπύλη κόστους τελικού μοντέλου

Από τις καμπύλες εκμάθησης του μοντέλου παρατηρούμε ότι φτάνει νωρίς σε ικανοποιητική ακρίβεια και κόστος, αλλά καθώς το μοντέλο εκπαιδεύεται για 1000 εποχές, τείνει να κάνει overfit. Παρόλα αυτά, οι τελικές μετρικές (F1-Score και Accuracy) είναι αρκετά υψηλές, που σημαίνει ότι το μοντέλο αποδίδει καλά.