Υπολογιστική Νοημοσύνη

3η εργασία

Επίλυση Προβλήματος Ταξινόμησης με χρήση MLP δικτύου

Γεώργιος Τσουμπλέκας, gktsoump@ece.auth.gr, ΑΕΜ: 9359, ΝΠΣ

---

0 Εισαγωγή

Στην συγκεκριμένη εργασία καλούμαστε να μελετήσουμε την λειτουργία των Multi-Layered-Perceptrons (MLPs) στην επίλυση προβλημάτων ταξινόμησης και πιο συγκεκριμένα στην ταξινόμηση χειρόγραφων ψηφίων από το 0-9 (MNIST dataset). Στο πρώτο σκέλος της εργασίας διερευνούνται και συγκρίνονται ως προς την απόδοση τους διάφορα μοντέλα MLP. Τα μοντέλα αυτά διαφέρουν σε διάφορες σχεδιαστικές λεπτομέρειες όπως τον ρυθμό μάθησης, την ύπαρξη κανονικοποίησης, την ύπαρξη αρχικοποίησης καθώς και την μέθοδο εκπαίδευσης μεταξύ άλλων. Στο δεύτερο σκέλος της εργασίας μελετάμε πώς μπορούμε να βελτιστοποιήσουμε ένα μοντέλο MLP με κατάλληλο tuning των υπερ-παραμέτρων του για ακόμα καλύτερη απόδοση. Συγκεκριμένα, για τον σκοπό αυτό θα χρησιμοποιήσουμε τον Hyperband tuner του Keras.

1 Διερεύνηση απόδοσης διαφορετικών MLP μοντέλων

Αρχικά, όλα τα μοντέλα που θα παρουσιαστούν στην συνέχεια αποτελούνται από 2 κρυφά στρώματα, το 1ο με 128 και το 2ο με 256 νευρώνες και συνάρτηση ενεργοποίησης την ReLU. Το στρώμα εξόδου αποτελείται από 10 νευρώνες με συνάρτηση ενεργοποίησης την softmax. Ως αντικειμενική συνάρτηση προς βελτιστοποίηση επιλέγεται η categorical cross-entropy ενώ ως μετρική αξιολόγησης χρησιμοποιείται η ακρίβεια (accuracy). Η εκπαίδευση διαρκεί 100 εποχές με ένα 20% του training set να χρησιμοποιείται για validation. Από κει και πέρα οι υπόλοιπες λεπτομέρειες των δικτύων μπορεί να διαφέρουν μεταξύ τους.

1.1 Μοντέλο 1

Για το μοντέλο αυτό χρησιμοποιήθηκαν τα default settings του Keras (RMSprop optimizer με $\eta = 0.001$ και $\rho = 0.9$) ενώ επιπλέον χρησιμοποιήθηκε η online διαδικασία μάθησης. Στα Σχ.1α,β φαίνονται οι καμπύλες ακριβείας και κόστους για τα training και validation sets.

(a) (b)

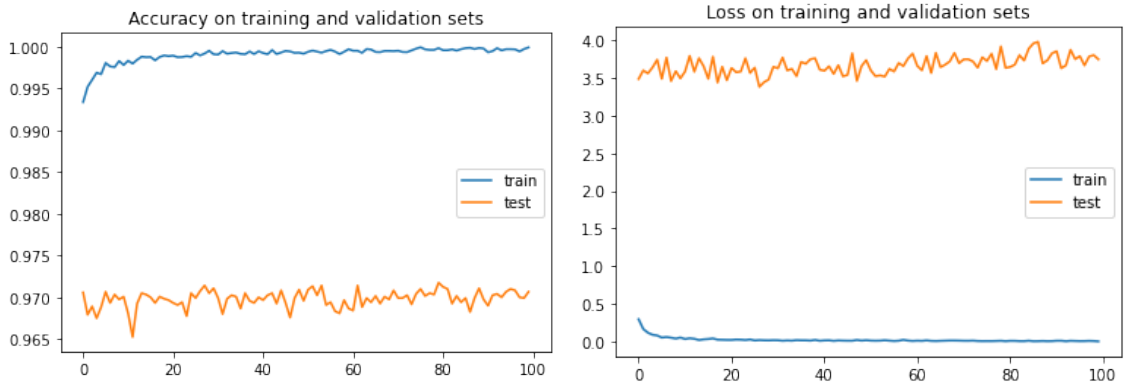
Figure 1: Καμπύλες ακριβείας και κόστους για το μοντέλο 1

Βλέπουμε πως το μοντέλο αυτό πετυχαίνει μεγάλη ακρίβεια για το training set ενώ στο validation set είναι λίγο μικρότερη (όπως και ήταν αναμενόμενο) αλλά πάλι πολύ υψηλή καθιστώντας το ένα υψηλά αποδοτικό μοντέλο. Βέβαια, όπως μπορούμε να δούμε από την καμπύλη κόστους, το κόστος στο validation set αρχίζει να αυξάνεται από νωρίς υποδεικνύοντας πως έχουμε overfitting του μοντέλου στα δεδομένα εκπαίδευσης. Για την αντιμετώπιση του προβλήματος θα μπορούσε να χρησιμοποιηθεί κάποια μέθοδος κανονικοποίησης, dropout, early stopping ή fine-tuning του αριθμού εποχών με κάποιον tuner. Αξίζει, τέλος, να σημειώσουμε, πως παρότι η απόδοση του μοντέλου είναι πολύ υψηλή, ο χρόνος εκπαίδευσης του είναι αρκετά μεγάλος και σε ακόμα μεγαλύτερα προβλήματα ίσως να ήταν μη διαχειρίσιμος.

1.2 Μοντέλο 2

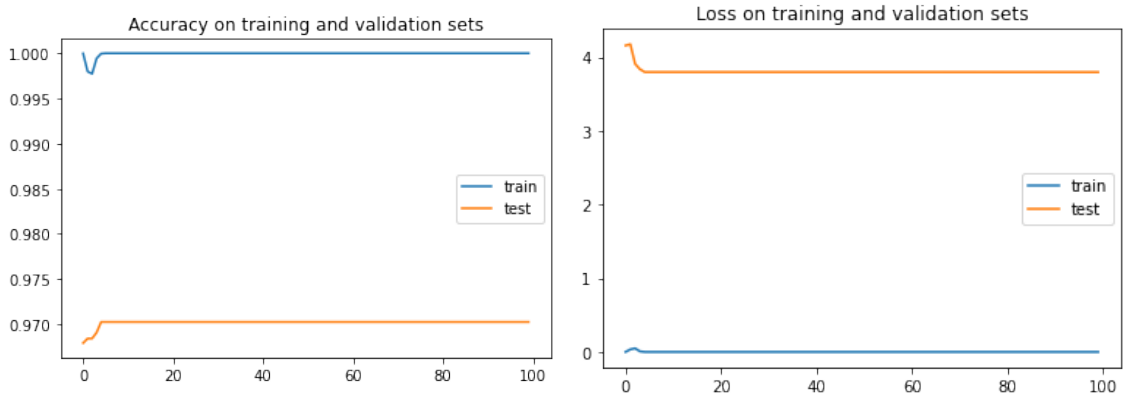
Για το μοντέλο αυτό χρησιμοποιήθηκαν τα default settings του Keras (RMSprop optimizer με $\eta = 0.001$ και $\rho = 0.9$) ενώ επιπλέον χρησιμοποιήθηκε η minibatch διαδικασία μάθησης με batch size $= 256$. Στα Σχ.2α,β φαίνονται οι καμπύλες ακριβείας και κόστους για τα training και validation sets.

Από το διάγραμμα ακριβείας, μπορούμε να δούμε πως η απόδοση και αυτού του μοντέλου είναι πολύ υψηλή τόσο για το training όσο και για το validation set. Επιπλέον, πολύ σημαντικό είναι και το γεγονός ότι οι τιμές της συνάρτησης κόστους στο validation set, ενώ παρουσιάζουν κάποιες διακυμάνσεις από εποχή σε εποχή, δεν φαίνεται να αυξάνονται με μεγάλο ρυθμό με το πέρασμα των εποχών. Συμπεραίνουμε, επομένως, πως στην περίπτωση αυτή δεν έχουμε overfitting και το μοντέλο εμφανίζει καλή ικανότητα γενίκευσης. Τέλος, παρατηρούμε πως σε σχέση με την περίπτωση της online μάθησης ο χρόνος εκπαίδευσης είναι αρκετά μικρότερος με αντίκτυπο την λίγο μικρότερη ακρίβεια του μοντέλου.



(a) (b)

Figure 2: Καμπύλες ακριβείας και κόστους για το μοντέλο 2



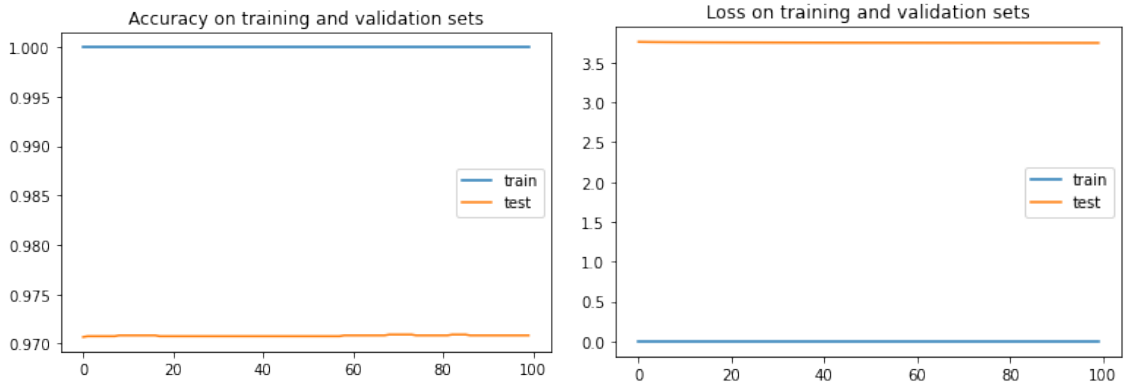
(a) (b)

Figure 3: Καμπύλες ακριβείας και κόστους για το μοντέλο 3

1.3 Μοντέλο 3

Για το μοντέλο αυτό χρησιμοποιήθηκαν τα default settings του Keras (RMSprop optimizer με $\eta = 0.001$ και $\rho = 0.9$) ενώ επιπλέον χρησιμοποιήθηκε η batch διαδικασία μάθησης. Στα Σχ.3α,β φαίνονται οι καμπύλες ακριβείας και κόστους για τα training και validation sets.

Από τα παραπάνω διαγράμματα φαίνεται χαρακτηριστικά πως μετά την 6η εποχή, οι τιμές της ακρίβειας και του κόστους παραμένουν σταθερές σε κάθε εποχή. Όσον αφορά το training set, βλέπουμε ότι η ακρίβεια γίνεται 1 (άρα το κόστος 0) αποδεικνύοντας πως έχει προσαρμοστεί τέλεια στα δεδομένα εκπαίδευσης, ενώ για το validation set η ακρίβεια είναι γύρω στο 0.97. Επομένως, το μοντέλο φαίνεται να είναι αντίστοιχα αποδοτικό με τα 2 προηγούμενα. Επιπλέον, το γεγονός ότι δεν έχουμε αύξηση της τιμής της συνάρτησης κόστους στο validation set με το πέρασμα των εποχών δείχνει ότι το μοντέλο εμφανίζει καλή ικανότητα γενίκευσης (δεν έχουμε overfitting). Παρόλα αυτά, το γεγονός ότι η τιμή της ακρίβειας και του κόστους παραμένουν σταθερές από ένα σημείο και μετά μας κάνει να σκεφτούμε πως μάλλον δεν είναι απαραίτητη η



(a) (b)

Figure 4: Καμπύλες ακριβείας και κόστους για το μοντέλο 4

εκπαίδευση του μοντέλου για τόσες πολλές εποχές αφού δεν επιτυγχάνεται κάποια βελτίωση. Τέλος, όσον αφορά τον χρόνο εκπαίδευσης βλέπουμε ότι το συγκεκριμένο μοντέλο είναι με διαφορά το γρηγορότερο από τα 3 που μελετήθηκαν ως τώρα.

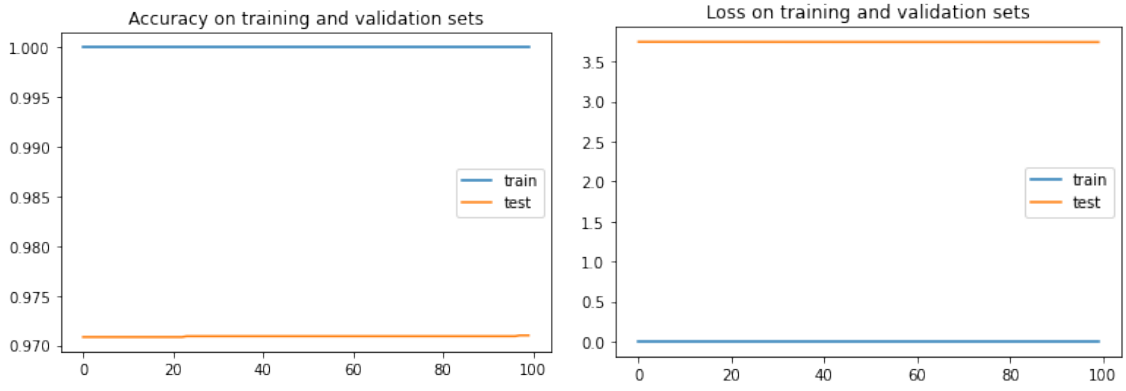
1.4 Μοντέλο 4

Για το μοντέλο αυτό χρησιμοποιήθηκε minibatch εκπαίδευση με batch size = 256 και ως μέθοδος εκπαίδευσης χρησιμοποιήθηκε η RMSprop με α=0.001 και ρ=0.01. Στα Σχ.4α,β φαίνονται οι καμπύλες ακριβείας και κόστους για τα training και validation sets.

Όπως φαίνεται από τα παραπάνω διαγράμματα, το μοντέλο φαίνεται να προσαρμόζεται από νωρίς τέλεια στα δεδομένα του συνόλου εκπαίδευσης ενώ και η ακρίβεια του για το σύνολο επικύρωσης είναι πολύ υψηλή (0.97). Γνωρίζουμε ότι μικρή τιμή του ρ ευνοεί την κίνηση του αλγορίθμου προς τις παλιότερες διευθύνσεις του αθροιστικού τετραγωνικού gradient με αποτέλεσμα, όπως φαίνεται και εδώ, να μην υπάρχει σημαντική βελτίωση με το πέρασμα των εποχών. Συνολικά, μπορούμε να πούμε ότι δεν έχουμε overfitting καθώς το κόστος στο validation set δεν αυξάνεται με το πέρασμα των εποχών και το μοντέλο έχει επίσης καλή απόδοση. Όμως, μπορούμε επίσης να καταλάβουμε πως η επιλογή μιας τόσο μικρής τιμής για το α δεν φαίνεται να είναι ιδιαίτερα καλή καθώς δεν επιτρέπει στο μοντέλο να βελτιωθεί με γρήγορο ρυθμό και έτσι αυτό μένει στάσιμο για μεγάλο αριθμό εποχών.

1.5 Μοντέλο 5

Για το μοντέλο αυτό χρησιμοποιήθηκε minibatch εκπαίδευση με batch size = 256 και ως μέθοδος εκπαίδευσης χρησιμοποιήθηκε η RMSprop με α=0.001 και ρ=0.99. Στα Σχ.5α,β φαίνονται οι καμπύλες ακριβείας και κόστους για τα training και validation sets.



(a) (b)

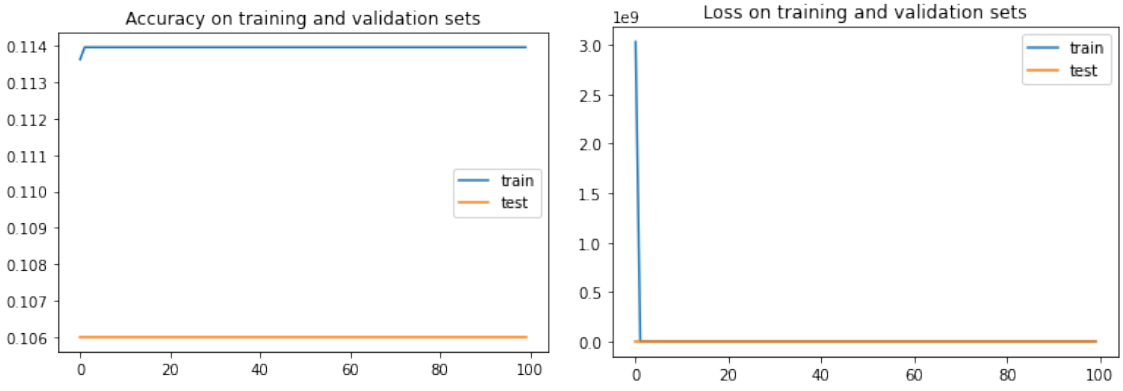
Figure 5: Καμπύλες ακριβείας και κόστους για το μοντέλο 5

Βλέπουμε πως είναι εμφανής η ομοιότητα των διαγραμμάτων αυτών με των αντίστοιχων του μοντέλου 4 επομένως μπορούμε να καταλήξουμε σε αντίστοιχα συμπεράσματα και εδώ. Αξίζει βέβαια να τονίσουμε πως στην περίπτωση αυτού του μοντέλου έχουμε επιλέξει μια μεγάλη τιμή για το ρ κάτι το οποίο σημαίνει ότι ευνοείται η κίνηση προς την κατεύθυνση των πιο πρόσφατων τιμών του αθροιστικού τετραγωνικού gradient στον αλγόριθμο RMSprop. Όμως, παρόλα αυτά, έχουμε αντίστοιχη συμπεριφορά με πριν, κάτι που οφείλεται στο ότι το μοντέλο προσαρμόζεται ήδη από τις πρώτες εποχές πολύ καλά στα δεδομένα χωρίς να αφήνει σημαντικά περιθώρια βελτίωσης.

1.6 Μοντέλο 6

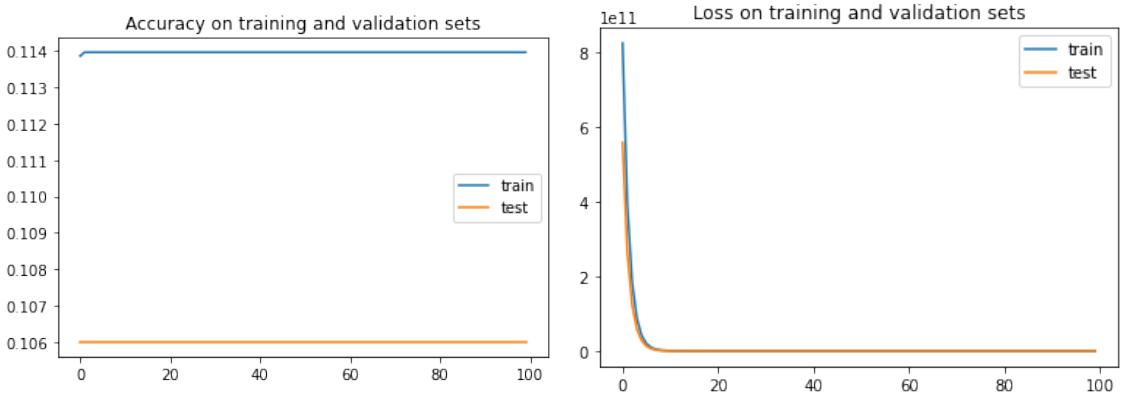
Για το μοντέλο αυτό εφαρμόστηκε αρχικοποίηση των βαρών των στρωμάτων με τυχαίες τιμές από κανονική κατανομή με μέση τιμή 10. Ως μέθοδος βελτιστοποίησης χρησιμοποιήθηκε το Stochastic Gradient Descend (SGD) με η=0.01. Στα Σχ.6α,β φαίνονται οι καμπύλες ακριβείας και κόστους για τα training και validation sets.

Όπως γίνεται εύκολα αντιληπτό, το μοντέλο αυτό δεν φαίνεται να προσαρμόζεται καθόλου καλά στα δεδομένα (underfitting) αφού η ακρίβεια του για το training set είναι γύρω στο 0.114 ενώ για το validation set 0.106. Αυτό κατά πάσα πιθανότητα οφείλεται στην αρχικοποίηση των βαρών που πραγματοποιήθηκε. Λόγω της μεγάλης μέσης τιμής της κανονικής κατανομής, οι αρχικές τιμές των βαρών είναι σχετικά μεγάλες με αποτέλεσμα να έχουμε το φαινόμενο των exploding gradients που οδηγεί στο να έχουμε έντονες ταλαντώσεις και αδυναμία σύγκλισης στο ελάχιστο. Επομένως, η χρήση των παραπάνω παραμέτρων για το συγκεκριμένο μοντέλο κρίνεται μη αποτελεσματική.



(a) (b)

Figure 6: Καμπύλες ακριβείας και κόστους για το μοντέλο 6



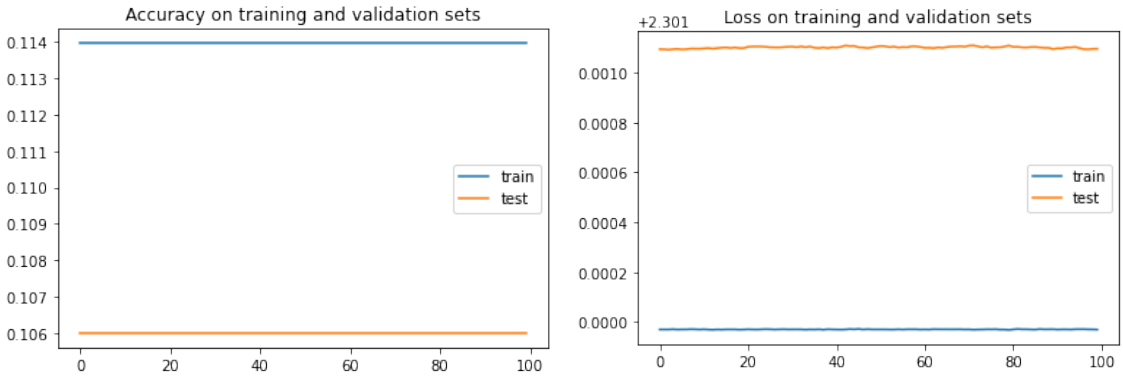
(a) (b)

Figure 7: Καμπύλες ακριβείας και κόστους για το μοντέλο 7

1.7 Μοντέλο 7

Για το μοντέλο αυτό χρησιμοποιήθηκαν οι ίδιες παράμετροι με το μοντέλο 6 ενώ εφαρμόζουμε επιπρόσθετα και L2-κανονικοποίηση στα στρώματα του δικτύου με α=0.1. Στα Σχ.7α,β φαίνονται οι καμπύλες ακριβείας και κόστους για τα training και validation sets.

Όπως φαίνεται από τα παραπάνω διαγράμματα, το μοντέλο αυτό είναι μη αποτελεσματικό στα συγκεκριμένα δεδομένα καθώς η ακρίβεια του παραμένει μικρή καθόλη την διάρκεια εκπαίδευσης του χωρίς να δείχνει βελτίωση με το πέρασμα των εποχών. Ειδικά στις πρώτες εποχές, η τιμή της συνάρτησης κόστους τόσο στο training όσο και στο validation set είναι υπερβολικά μεγάλη (λογώ της έντονης επιβολής ποινής στις μεγάλες αρχικές τιμές των βαρών) και παρότι μειώνεται γρήγορα και σε σημαντικό βαθμό, δεν συνεπάγεται και αποδοτικότητα του μοντέλου. Από τα παραπάνω, επομένως, συμπεραίνουμε ότι έχουμε underfitting του μοντέλου στα δεδομένα.



(a) (b)

Figure 8: Καμπύλες ακριβείας και κόστους για το μοντέλο 8

1.8 Μοντέλο 8

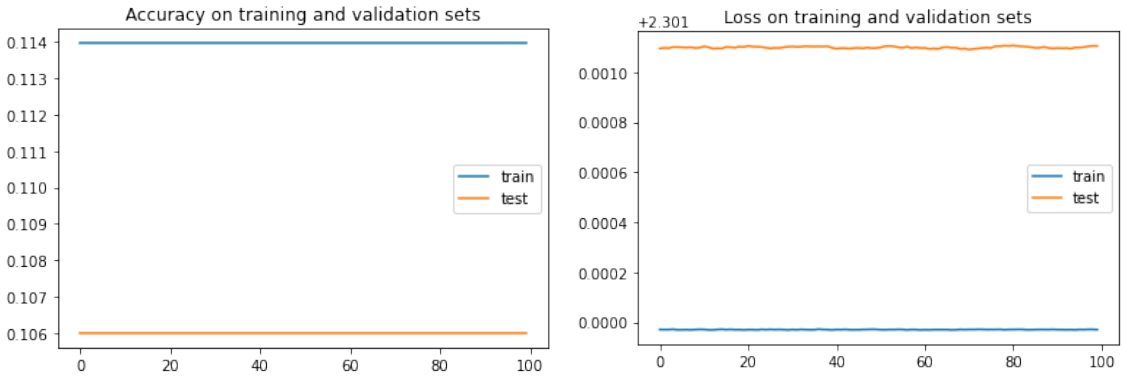
Για το μοντέλο αυτό χρησιμοποιήθηκαν οι ίδιες παράμετροι με το μοντέλο 6 ενώ εφαρμόζουμε επιπρόσθετα και L2-κανονικοποίηση στα στρώματα του δικτύου με α=0.01. Στα Σχ.8α,β φαίνονται οι καμπύλες ακριβείας και κόστους για τα training και validation sets.

Από τα παραπάνω διαγράμματα φαίνεται ότι ούτε αυτό το μοντέλο είναι ιδιαίτερα αποδοτικό καθώς η ακρίβεια του παραμένει μικρή καθόλη την διάρκεια εκπαίδευσης χωρίς να βελτιώνεται με το πέρασμα των εποχών. Σε αντίθεση με το μοντέλο 7, εδώ οι τιμές της συνάρτησης κόστους τόσο στο training όσο και στο validation set παραμένουν από την αρχή σχετικά σταθερές. Αυτό οφείλεται στο γεγονός ότι η παράμετρος α της κανονικοποίησης έχει μικρότερη τιμή με αποτέλεσμα στον υπολογισμό της συνάρτησης κόστους να δίνεται μεγαλύτερη έμφαση στο σφάλμα των εκτιμήσεων παρά στις (αρχικά μεγάλες) τιμές των βαρών του μοντέλου. Σε κάθε περίπτωση όμως, έχουμε και πάλι underfitting του μοντέλου στα δεδομένα.

1.9 Μοντέλο 9

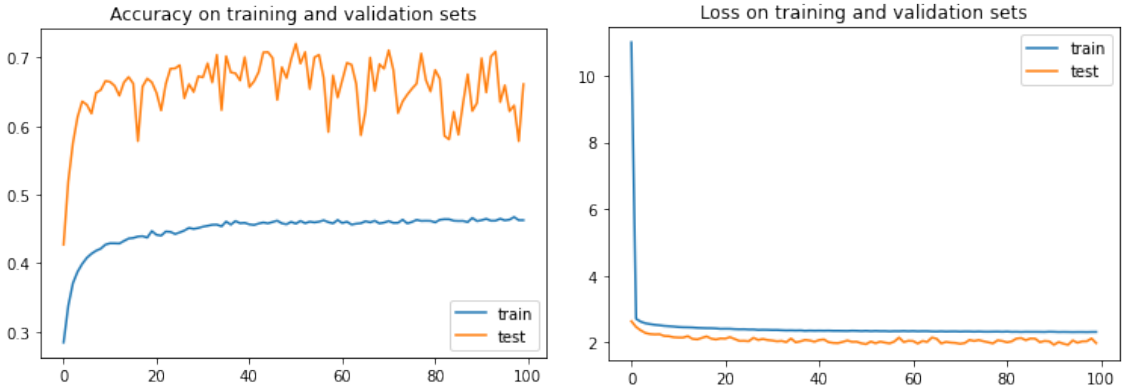
Για το μοντέλο αυτό χρησιμοποιήθηκαν οι ίδιες παράμετροι με το μοντέλο 6 ενώ εφαρμόζουμε επιπρόσθετα και L2-κανονικοποίηση στα στρώματα του δικτύου με α=0.001. Στα Σχ.9α,β φαίνονται οι καμπύλες ακριβείας και κόστους για τα training και validation sets.

Όπως μπορούμε εύκολα να διαπιστώσουμε, τα διαγράμματα που έχουμε για αυτό το μοντέλο μοιάζουν σε πολύ μεγάλο βαθμό με τα διαγράμματα του μοντέλου 8. Ακριβώς επειδή έχουμε και εδώ πολύ μικρή τιμή για το α, οι τιμές των βαρών δεν λαμβάνονται σχεδόν καθόλου υπόψιν στον υπολογισμό της συνάρτησης κόστους. Αυτό σε συνδυασμό με το φαινόμενο των exploding gradients λόγω των μεγάλων αρχικών τιμών που δεν επιτρέπουν στο μοντέλο να συγκλίνει στις πραγματικές τιμές έχει ως αποτέλεσμα η ακρίβεια του μοντέλου να παραμένει μικρή και στάσιμη με το πέρασμα των εποχών. Συνεπώς, και στο μοντέλο αυτό έχουμε ξεκάθαρο underfitting.



(a) (b)

Figure 9: Καμπύλες ακριβείας και κόστους για το μοντέλο 9



(a) (b)

Figure 10: Καμπύλες ακριβείας και κόστους για το μοντέλο 9

1.10 Μοντέλο 10

Στο μοντέλο αυτό εφαρμόστηκε L1-κανονικοποίηση στα κρυφά στρώματα με παράμετρο α=0.001 σε συνδυασμό με dropout με πιθανότητα 0.3. Ως μέθοδος βελτιστοποίησης χρησιμοποιήθηκε η RMSprop με η=0.001 και ρ=0.9. Στα Σχ.10α,β φαίνονται οι καμπύλες ακριβείας και κόστους για τα training και validation sets.

Όπως μπορούμε να δούμε, το μοντέλο εμφανίζει σχετικά μέτρια απόδοση μιας και δεν επιτυγχάνει ιδιαίτερα υψηλές τιμές ακρίβειας όυτε για το training ούτε για το validation set. Εντύπωση προκαλέι επίσης το γεγονός ότι η ακρίβεια του μοντέλου είναι μεγαλύτερη για το validation set από ότι για το training set ενώ αντίστοιχα έχουμε μικρότερη τιμή της συνάρτησης κόστους στο validation set από ότι στο training set. Χαρακτηριστικές είναι επίσης οι διακυμάνσεις της ακρίβειας που βλέπουμε μεταξύ διαδοχικών εποχών και οι οποίες οφείλονται στον τυχαίο τρόπο με τον οποίο αφαιρούνται ή όχι από κάθε κρυφό στρώμα νευρώνες. Συνολικά, πάντως, μπορούμε να πούμε πως έχουμε underfitting του μοντέλου μιας και αυτό δεν καταφέρνει να πετύχει υψηλές τιμές ακρίβειας για τα δοσμένα δεδομένα εισόδου.

2 Fine tuning δικτύου

2.1 Το μοντέλο

Στην προηγούμενη ενότητα μελετήθηκαν διάφοροι συνδυασμοί μεθόδων και παραμέτρων αυτών που εφαρμόζονται στα MLP δίκτυα με άλλους να αποδεικνύονται περισσότερο και άλλους λιγότερο αποτελεσματικοί. Στην παρούσα ενότητα αυτό που μας ενδιαφέρει είναι πώς για μια συγκεκριμένη μορφή ενός MLP δικτύου μπορούμε να επιλέξουμε κατάλληλα τις διάφορες υπερπαραμέτρους του για να πετύχουμε όσο το δυνατόν καλύτερη απόδοση.

Πιο συγκεκριμένα, θέλουμε να δημιουργήσουμε ένα MLP με 2 κρυφά στρώματα τα οποία θα εκπαιδεύσουμε με τον αλγόριθμο RMSprop. Επιπλέον, εφαρμόζουμε L2-κανονικοποίηση στα βάρη των στρωμάτων ενώ, τέλος, η αρχικοποίηση των βαρών γίνεται με την χρήση του He initialization για αποφυγή φαινομένων vanishing ή exploding gradients. Οι υπερπαράμετροι που προκύπτουν προς βελτιστοποίηση είναι το πλήθος νευρώνων του 1ου και 2ου στρώματος, ο ρυθμός εκμάθησης και η τιμή α της κανονικοποίησης. Ακόμα, αξίζει να σημειώσουμε πως η μετρική αξιολόγησης του μοντέλου που επιλέχθηκε είναι το F-measure που γενικά θεωρείται πιο robust και έμπιστο από το accuracy, ειδικά σε unbalanced datasets.

Δοσμένων διάφορων πιθανών τιμών για καθεμία εξ αυτών των υπερπαραμέτρων, χρησιμοποιούμε το Hyperband tuner του Keras για να βρούμε ποιό από όλα τα πιθανά μοντέλα είναι το πιο αποδοτικό. Μιας και ο αριθμός εποχών που έχει οριστεί είναι αρκετά μεγάλος (1000) εφαρμόζουμε επιπρόσθετα και early stopping για να κάνουμε πιο γρήγορη και αποτελεσματική την όλη διαδικασία. Στόχος του tuner είναι να βρει ποιό από τα υποψήφια μοντέλα έχει την υψηλότερη τιμή για την μετρική F-measure στο validation set. Τρέχοντας, επομένως, τον tuner προκύπτει ότι το καλύτερο δυνατό μοντέλο είναι αυτό με 128 νευρώνες στο 1ο στρώμα, 512 στο 2ο στρώμα, ρυθμό μάθησης η=0.001 και παράμετρο κανονικοποίησης α=0.000001 για το οποίο επιτυγχάνει τιμή για το F-measure στο validation set 0.97. Παρατηρούμε ότι ο tuner επιλέγει το μοντέλο που έχει τους περισσότερους νευρώνες και στα 2 στρώματα, κάτι το οποίο είναι αναμενόμενο αφού γενικά μεγάλο πλήθος νευρώνων σημαίνει μεγαλύτερη ικανότητα προσαρμογής στα δεδομένα. Βέβαια, αυτό πολλές φορές συνεπάγεται την εμφάνιση του φαινομένου του overfitting το οποίο εν προκειμένω μπορούμε να αντιμετωπίσουμε ως ένα βαθμό με την κανονικοποίηση που εφαρμόζουμε. Βλέπουμε, πάντως, ότι η τιμή α της κανονικοποίησης έχει πολύ μικρή τιμή κάτι που υποδεικνύει ότι στην περίπτωση μας δεν έχουμε overfitting και έτσι δεν είναι σημαντικό να επιβάλλουμε μεγάλες ποινές στα βάρη των νευρώνων. Τέλος, όσον αφορά τον ρυθμό μάθησης, επιλέγεται η μικρότερη από τις υποψήφιες τιμές εξασφαλίζοντας έτσι την ευσταθή σύγκλιση του μοντέλου ακόμα και αν χρειαστούν περισσότερες εποχές εκπαίδευσης.

2.2 Απόδοση του μοντέλου

Αρχικά, θα μελετήσουμε τις καμπύλες εκμάθησης (learning curves) για τα training και validation sets.

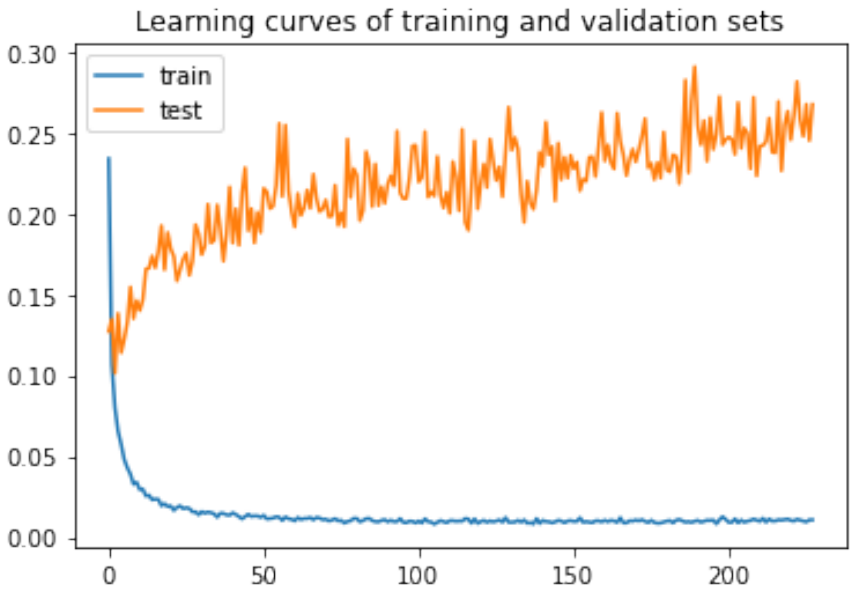


Figure 11: Καμπύλη μάθησης (learning curve) του τελικού μοντέλου

Όπως μπορούμε να δούμε στο Σχ.11, η τιμές της συνάρτησης κόστους τόσο για το training όσο και για το validation set είναι αρκετά μικρές με την τιμή να είναι μεγαλύτερη για το validation set όπως και είναι αναμενόμενο. Αξιοσημείωτο είναι το γεγονός ότι η τιμή του κόστους αρχίζει να αυξάνει για το validation set από σχετικά μικρή εποχή με αποτέλεσμα τον κίνδυνο εμφάνισης overfitting. Όμως, κάτι τέτοιο αποτρέπεται σε σημαντικό βαθμό από την χρήση του early stopping κατά την εκπαίδευση του μοντέλου. Έτσι, αντί το μοντέλο να εκπαιδευτεί για συνολικά 1000 όπως είχε οριστεί αρχικά, εν τέλει εκπαιδεύτηκε για μόλις 310. Ο λόγος που η τιμή του κόστους φαίνεται και πάλι να αυξάνει για αρκετές εποχές οφείλεται στο γεγονός ότι το patience για το early stopping ορίστηκε στις 200 εποχές. Πιθανώς, για μικρότερη τιμή του patience να πετυχαίναμε ακόμα καλύτερα αποτελέσματα. Μπορούμε, όμως, να πούμε ότι προλαβαίνουμε το overfitting πρωτού γίνει ιδιαίτερα έντονο και έτσι το μοντέλο μας είναι αρκετα αποδοτικό.

Ένα άλλο σημαντικό κριτήριο αξιολόγησης που χρησιμοποιείται για μοντέλα ταξινόμησης είναι η μελέτη του πίνακα σύγχυσης (confusion matrix), στον οποίον μπορούμε να δούμε το πλήθος των στοιχείων του test set που ταξινομήθηκαν σωστά και το πλήθος αυτών που ταξινομήθηκαν λανθασμένα σε κάθε περίπτωση. Για την περίπτωση μας ο πίνακας αυτός φαίνεται στο Σχ.12.

Όπως μπορούμε να δούμε, η πλειοψηφία των στοιχείων βρίσκεται πάνω στην κύρια διαγώνιο επιβεβαιώνοντας ότι γίνεται σωστή ταξινόμηση των ψηφίων στην πλειοψηφία των περιπτώσεων. Μερικές χαρακτηριστικές περιπτώσεις λαθών είναι το 4 που ταξινομείται ως 9 (12 περιπτώσεις), το 5 που ταξινομείται ως 3 (11 περιπτώσεις), το 6 που ταξινομείται ως 0 (10 περιπτώσεις), το 8 που ταξινομείται ως 3 (11 περιπτώσεις) και το 9 που ταξινομείται ως 4 (13 περιπτώσεις). Βλέπουμε, λοιπόν, ότι υπάρχουν διάφορες περιπτώσεις που παραμένουν ακόμα σχετικά challenging ως προς την σωστή τους ταξινόμηση από το μοντέλο μας υποδεικνύοντας ότι παρά την καλή του απόδοση υπάρχει χώρος για βελτίωση.

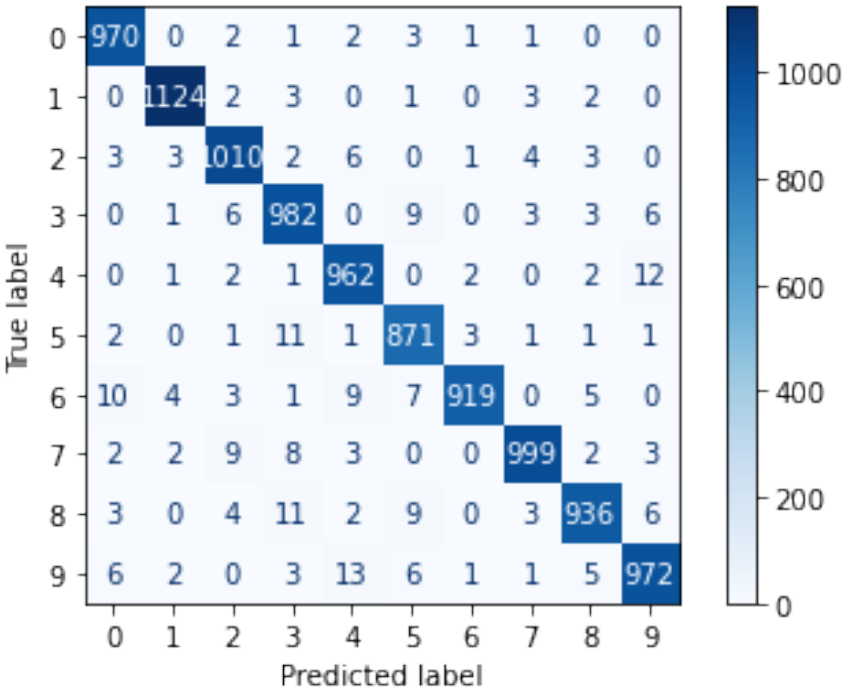


Figure 12: Πίνακας σύγχυσης για πρόβλεψη στο test set με το τελικό μοντέλο

Τέλος, στο Σχ.13 μπορούμε να δούμε μερικές κλασσικες μετρικές που χρησιμοποιούνται στην αξιολόγηση ταξινομητών και οι οποίες προκύπτουν από τον πίνακα σύγχυσης. Αυτές είναι τα accuracy, precision, recall και F-measure.

	precision	recall	f1-score	support
0	0.97	0.99	0.98	980
1	0.99	0.99	0.99	1135
2	0.94	0.98	0.96	1032
3	0.98	0.95	0.96	1010
4	0.98	0.96	0.97	982
5	0.99	0.95	0.97	892
6	0.97	0.98	0.97	958
7	0.96	0.97	0.97	1028
8	0.97	0.97	0.97	974
9	0.97	0.96	0.96	1009
accuracy			0.97	10000
macro avg	0.97	0.97	0.97	10000
weighted avg	0.97	0.97	0.97	10000

Figure 13: Μετρικές αξιολόγησης του τελικού μοντέλου

Μπορούμε με ευκολία να συμπεράνουμε ότι για κάθε κλάση, αλλά και συνολικά, το μοντέλο επιτυγχάνει πολύ υψηλές επιδόσεις με το μοντέλο να φτάνει σε accuracy το 0.97. Επομένως, μπορούμε να πούμε ότι το fine tuning των υπερπαραμέτρων του δικτύου ήταν αποτελεσματικό.