

Νευρωνικά Δίκτυα - Βαθιά Μάθηση

2η Υποχρεωτική Εργασία
Παπαδόπουλος Παναγιώτης 10697 ΗΜΜΥ

Εισαγωγή

Το παρόν έργο αποσκοπεί στην ανάπτυξη ενός νευρωνικού δικτύου ικανού να παράγει εικόνες που αντιπροσωπεύουν τα αποτελέσματα απλών μαθηματικών πράξεων. Συγκεκριμένα, το δίκτυο λαμβάνει ως είσοδο τρεις εικόνες: δύο εικόνες ψηφίων από το γνωστό σύνολο δεδομένων *MNIST* και μία εικόνα που περιέχει το μαθηματικό σύμβολο της πράξης από το σύνολο δεδομένων *HASYv2*. Το δίκτυο εξάγει δύο εικόνες, οι οποίες αντιστοιχούν στα ψηφία του αποτελέσματος της πράξης. Το έργο βασίζεται σε αρχιτεκτονική *encoder – decoder*, όπου τα χαρακτηριστικά των εικόνων εισόδου κωδικοποιούνται σε μία λανθάνουσα αναπαράσταση (*latent representation*) και στη συνέχεια αποκωδικοποιούνται για να παραχθούν οι εικόνες εξόδου.

Σετ Δεδομένων

Το σύνολο δεδομένων *MNIST* χρησιμοποιείται για την αναπαράσταση των αριθμητικών ψηφίων, καθώς περιέχει εικόνες ψηφίων 28×28 σε ασπρόμαυρη κλίμακα. Για τα μαθηματικά σύμβολα, χρησιμοποιείται το *HASYv2*, το οποίο περιλαμβάνει εικόνες μαθηματικών συμβόλων σε ανάλυση 32×32 και μετασχηματίζεται κατάλληλα ώστε να εναρμονιστεί με τη μορφή του *MNIST*. Ο μετασχηματισμός αυτός εμπεριέχει:

- Μετατροπή της διάστασης των εικόνων σε 28×28 για να συμβαδίζει με του *MNIST*
- Αντιστροφή χρώματος, από μαύρους χαρακτήρες και λευκό φόντο σε λευκούς χαρακτήρες και μαύρο φόντο.

Το δίκτυο εκπαιδεύεται να εκτελεί τέσσερις βασικές πράξεις: πρόσθεση, αφαίρεση, πολλαπλασιασμό και διαιρεση. Κατά την εκπαίδευση, χρησιμοποιούνται δεδομένα εισόδου-στόχου, όπου τα αποτελέσματα των πράξεων χωρίζονται στα δύο ψηφία που τα απαρτίζουν. Αξίζει να σημειωθεί ότι:

- Η πράξη της αφαίρεσης έχει ως αποτέλεσμα την απόλυτη τιμή της διαφοράς των δύο αριθμών
- Η πράξη της διαιρεσης είναι στην πραγματικότητα ακέραια διαιρεση, τα δεκαδικά ψηφία δεν υπολογίζονται

Για την δημιουργία του σετ δεδομένων υπολογίζεται τυχαία μια μαθηματική εξίσωση που υπακούει τους παραπάνω κανόνες και τυχαία επιλέγονται εικόνες από τα αντίστοιχα προϋπάρχοντα σετ δεδομένων. Κάθε εικόνα κανονικοποιείται, με τις τιμές των πίειλ στο εύρος $[0,1]$.

Εργαλεία

Για την υλοποίηση του δικτύου επιλέχθηκε η γλώσσα προγραμματισμού *python* και η βιβλιοθήκη *keras* για την ανάπτυξή της. Άλλες βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι εξής:

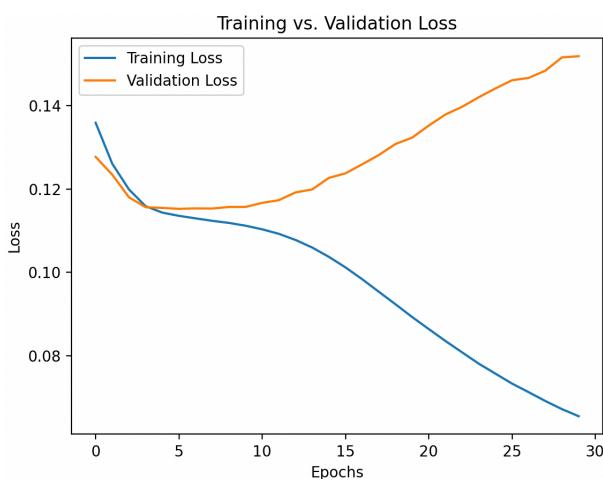
- *numpy*, για την επεξεργασία των δεδομένων και την οργάνωση των εικόνων σε πολυδιάστατους πίνακες
- *deeplake*, για την πρόσβαση στο σύνολο δεδομένων *HASYv2*, το οποίο περιέχει τα μαθηματικά σύμβολα “+”, “-”, “x” και “/”.
- *opencv*, για την επεξεργασία των εικόνων από το *HASYv2*, ώστε να προσαρμοστούν στις διαστάσεις και την κανονικοποίηση του *MNIST*

Ανάπτυξη του Δικτύου

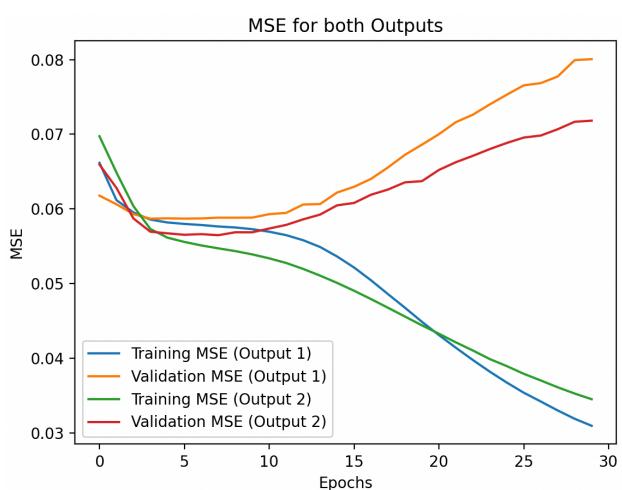
Για τον καθορισμό του βέλτιστου συνδυασμού υπερπαραμέτρων πραγματοποιηθήκαν αρχικές εκπαίδευσεις δικτύων με τα παρακάτω κοινά χαρακτηριστικά:

- *3x3 kernel size*, ισορροπώντας ακρίβεια με απόδοση
- *2x2 pool size* για την μείωση της διάστασης του χώρου εξόδου
- *ReLU* συνάρτηση ενεργοποίησης, ιδανική για συνελικτικά νευρωτικά δίκτυα
- Λανθάνουσα μεγέθους 256
- *batch size* 32
- 30 εποχές
- Οι αρχικές εκπαίδευσεις έγιναν με 50.000 αριθμό δειγμάτων

Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιείται η απώλεια (*loss*) και *mean square error* κάθε εξόδου.



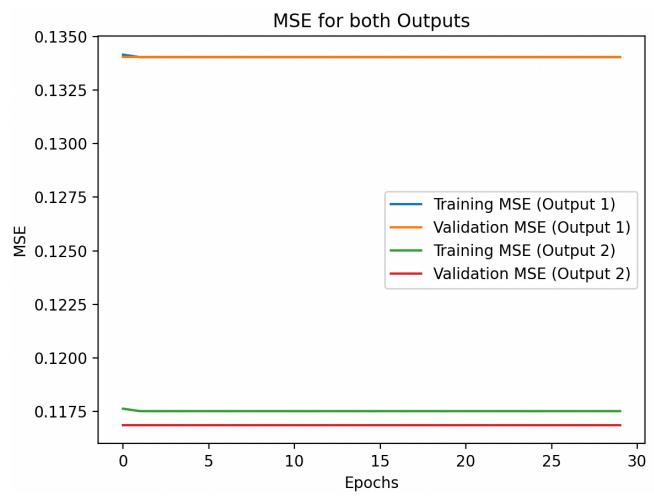
Απώλεια για 3 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, ρυθμό εκμάθησης 0.001



MSE για 3 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, ρυθμό εκμάθησης 0.001



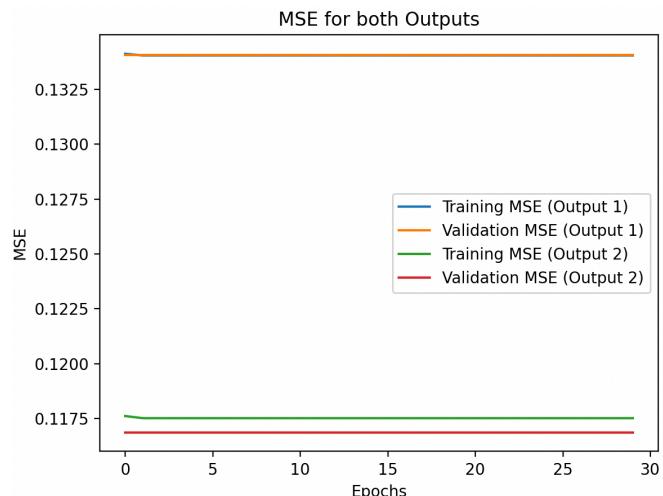
Απώλεια για 3 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, ρυθμό εκμάθησης 0.01



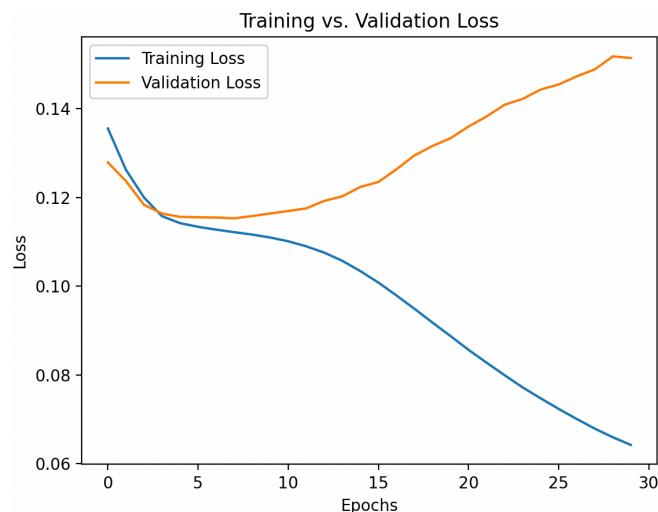
MSE για 3 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, ρυθμό εκμάθησης 0.01



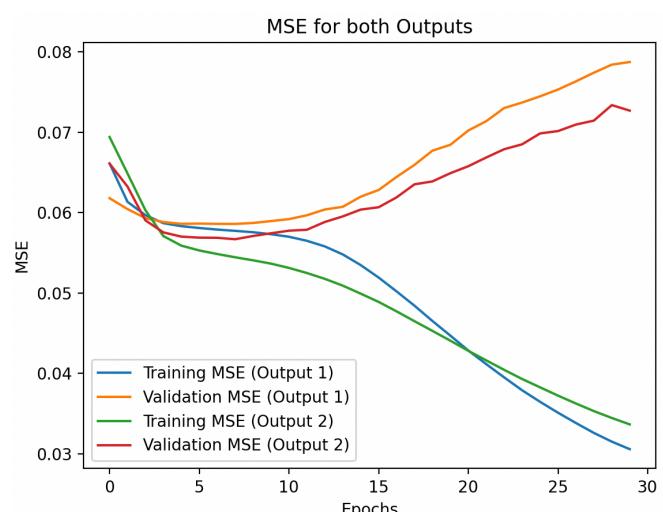
Απώλεια για 3 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, ρυθμό εκμάθησης 0.1



MSE για 3 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, ρυθμό εκμάθησης 0.1



Απώλεια για 3 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, ρυθμό εκμάθησης 0.001 και 100.000 δείγματα



MSE για 3 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, ρυθμό εκμάθησης 0.001 και 100.000 δείγματα

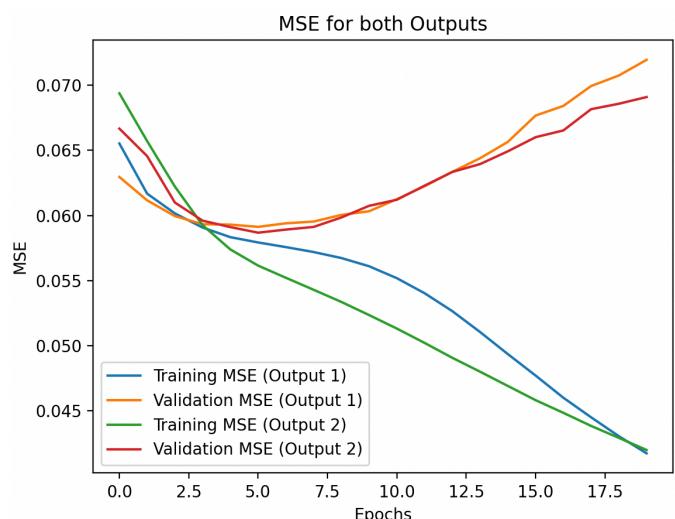
Μετά από χρόνο εκπαίδευσης 8 ωρών προκύπτουν οι παραπάνω κυματομορφές. Είναι εμφανές ότι το δίκτυο έχει την ικανότητα να μαθαίνει μόνο για ρυθμό εκμάθησης 0.001 από αυτούς που δοκιμάστηκαν. Επίσης, δεν υπάρχει φανερή διαφορά για την αύξηση του σετ δεδομένων. Για τα διαγράμματα *MSE* και των δύο εξόδων:

Η εκπαίδευση δείχνει συνεχή μείωση στο *MSE* για και τις δύο εξόδους, κάτι που υποδηλώνει ότι το μοντέλο μαθαίνει να παράγει πιο ακριβείς προβλέψεις με κάθε εποχή. Ωστόσο, το *validation MSE* (ειδικά για την πρώτη έξοδο) αρχίζει να αυξάνεται μετά από περίπου 10-15 εποχές, ενώ για τη δεύτερη έξοδο, η αύξηση είναι πιο έντονη.

Αυτή η αύξηση στο *validation MSE* είναι ένδειξη υπερπροσαρμογής (*overfitting*), όπου το μοντέλο αρχίζει να αποδίδει καλά στα δεδομένα εκπαίδευσης αλλά όχι στα δεδομένα επικύρωσης.

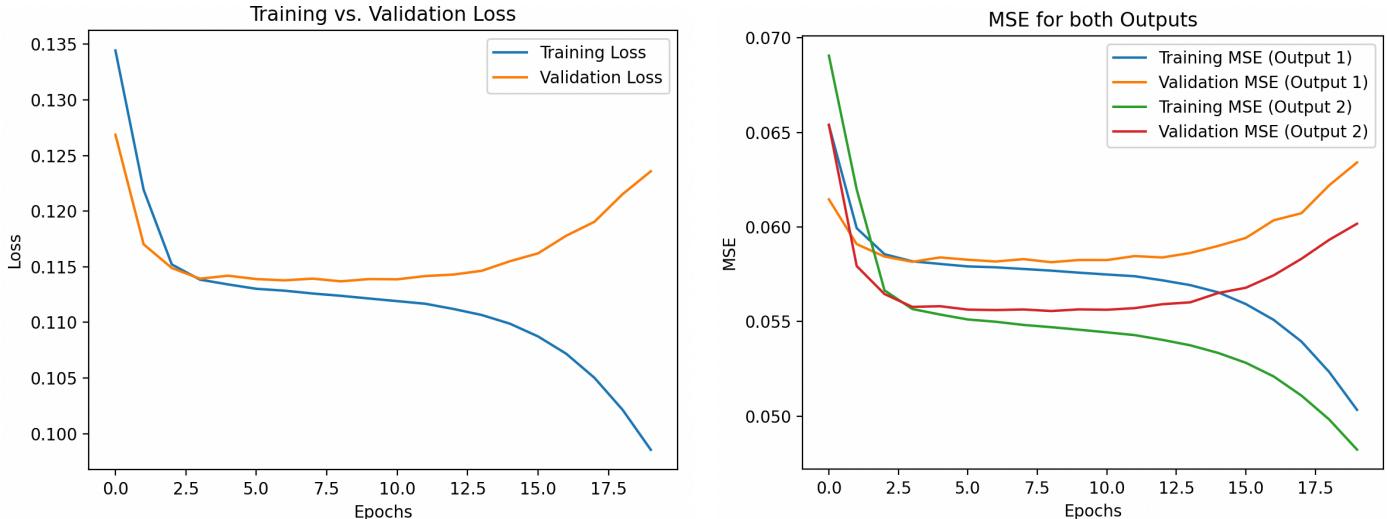
Από την άλλη, η εκπαίδευση έχει συνεχώς φθίνουσα απώλεια, γεγονός που σημαίνει ότι το μοντέλο βελτιστοποιεί τη συνάρτηση κόστους, ενώ η απώλεια επικύρωσης (*Validation Loss*) ακολουθεί το ίδιο μοτίβο με το *validation MSE*. Αυξάνεται μετά από περίπου 10 – 15 εποχές, υποδηλώνοντας το ίδιο πρόβλημα υπερπροσαρμογής.

Επομένως, διατηρείται ο ρυθμός εκμάθησης 0.001 και 50.000 στοιχεία για το σετ δεδομένων και δοκιμάζονται δύο νέα δίκτυα. Το πρώτο έχει αυξημένο αριθμό (4) συνελικτικών επιπέδων που επιτρέπουν στο δίκτυο να εξάγει πιο περίπλοκες ιεραρχικές ιδιότητες από τις εικόνες εισόδου. Επιπλέον, το μέγεθος του λανθάνοντος χώρου αυξάνεται από 256 σε 512, παρέχοντας έναν πιο εκφραστικό χώρο για την αναπαράσταση των πληροφοριών. Το *dense size* αναπροσαρμόστηκε σύμφωνα με τον αυξημένο αριθμό συνελικτικών επιπέδων και λειτουργιών *pooling*. Στη δεύτερη περίπτωση, ο αριθμός των συνελικτικών επιπέδων μειώθηκε (2), επιτρέποντας στο δίκτυο να μαθαίνει απλούστερες ιεραρχίες χαρακτηριστικών, μειώνοντας έτσι την πολυπλοκότητα της λανθάνουσας αναπαράστασης. Το μέγεθος του λανθάνοντος χώρου μειώθηκε σε 128 για να περιοριστεί η πολυπλοκότητα, ενώ το *dense size* αναπροσαρμόστηκε σύμφωνα με τον μειωμένο αριθμό συνελικτικών επιπέδων και λειτουργιών *pooling*. Τέλος, ο αριθμός των *transposed convolutional layers* μειώθηκε σε δύο, ώστε να ταιριάζει με τον απλούστερο κωδικοποιητή και τη λιγότερο λεπτομερή λανθάνουσα αναπαράσταση.



Απώλεια για 2 συνελικτικές, 2 αντίστροφες συνελικτικές στρώσεις, μέγεθος λανθάνουσας

MSE για 2 συνελικτικές, 2 αντίστροφες συνελικτικές στρώσεις, μέγεθος λανθάνουσας



Απώλεια για 4 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, μέγεθος λανθάνουσας
512

MSE για 4 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, μέγεθος λανθάνουσας
512

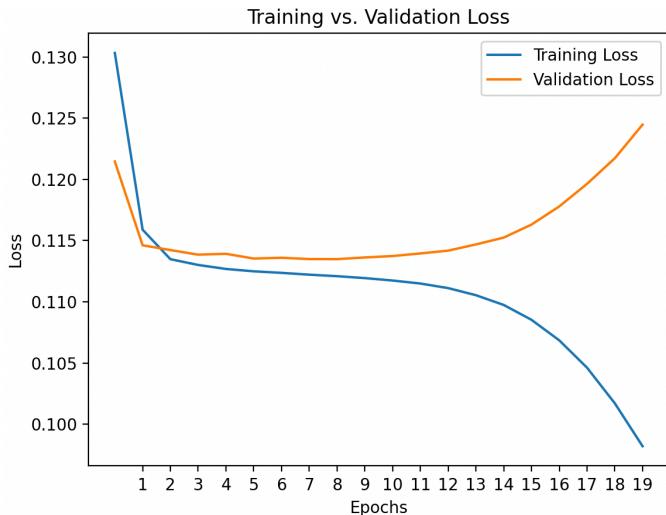
Συνολικός χρόνος εκπαίδευσης: 2 ώρες

Η υψηλή τιμή του validation loss σε σχέση με το training loss δείχνει ότι το απλουστευμένο μοντέλο δεν έχει επαρκή δυνατότητα να μάθει τα χαρακτηριστικά των δεδομένων. Ο μικρός αριθμός επιπέδων και φίλτρων στις συνελικτικές στρώσεις έχει ως αποτέλεσμα να μην ανιχνεύονται πολύπλοκα μοτίβα στα δεδομένα.
Το απλοποιημένο μοντέλο είναι ανεπαρκές για την επίλυση του προβλήματος. Ενώ μπορεί να μαθαίνει τα δεδομένα εκπαίδευσης (χαμηλό training loss), δεν μπορεί να γενικεύσει αποτελεσματικά στα δεδομένα επικύρωσης, με αποτέλεσμα υψηλότερο validation loss. Αυτό καθιστά αναγκαία τη χρήση πιο σύνθετων δομών.

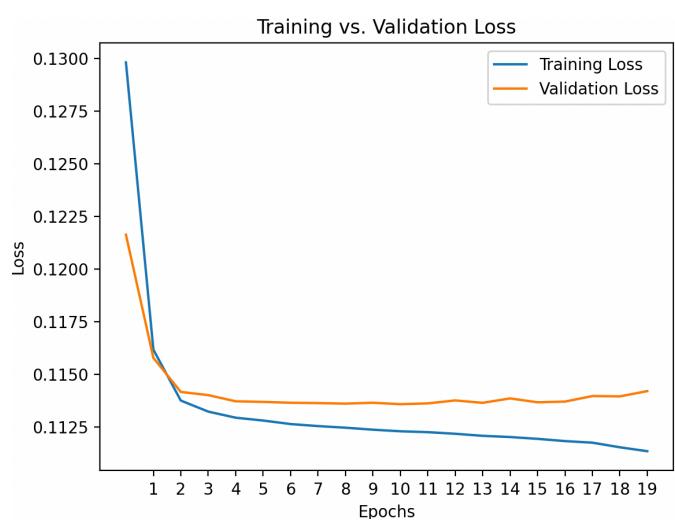
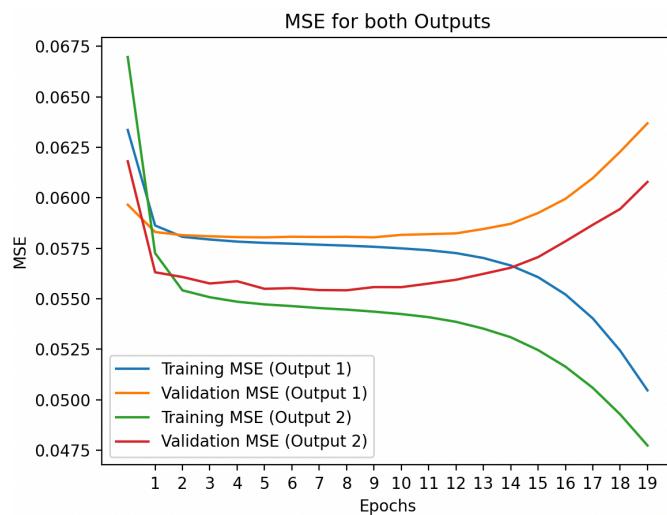
Το σύνθετο μοντέλο καταφέρνει να αποδώσει καλύτερα από τα άλλα μοντέλα στο validation set, γεγονός που υποδηλώνει ότι η αύξηση του αριθμού των επιπέδων και του μεγέθους του *latent space* βοήθησε στην αποδοτικότερη εκμάθηση χαρακτηριστικών.
Παρόλα αυτά, η σχετικά υψηλότερη τιμή του *training loss* δείχνει ότι το μοντέλο δεν έχει μάθει πλήρως τα δεδομένα εκπαίδευσης, πιθανώς λόγω του μεγάλου αριθμού παραμέτρων που πρέπει να εκπαιδευτούν. Επομένως, διατηρείται αυτό το μοντέλο και γίνονται οι εξής περεταίρω αλλαγές:

- Αύξηση των δεδομένων σε 100.000 δείγματα
- Μείωση ρυθμού εκμάθησης
- Διαφοροποιήσεις στην διάσταση λανθάνουσας

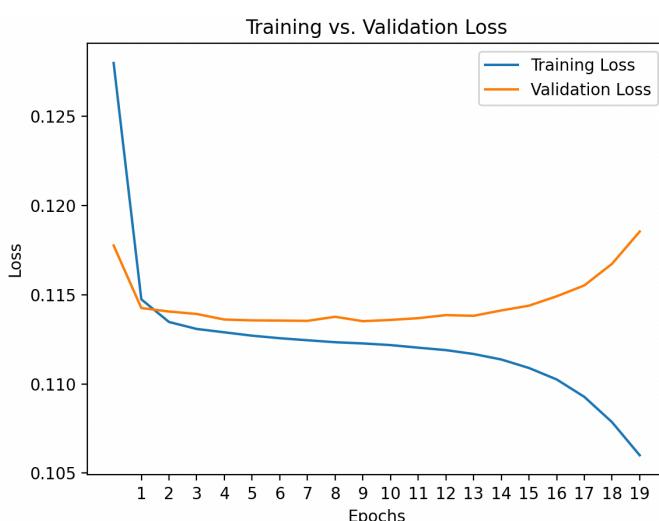
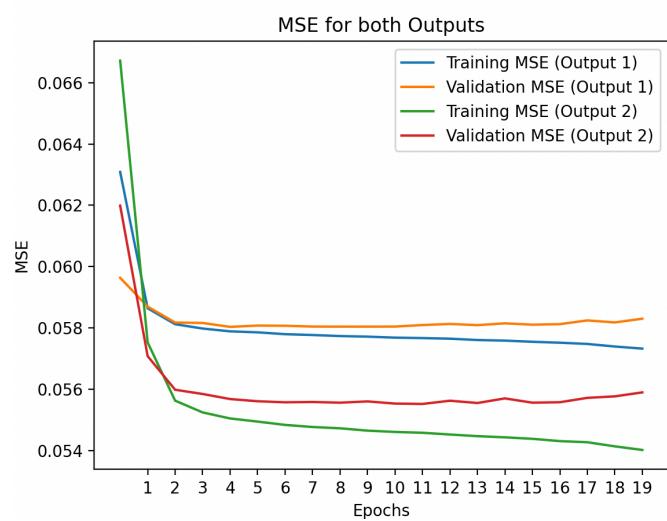
Μετά από 16 ώρες εκπαίδευσης:



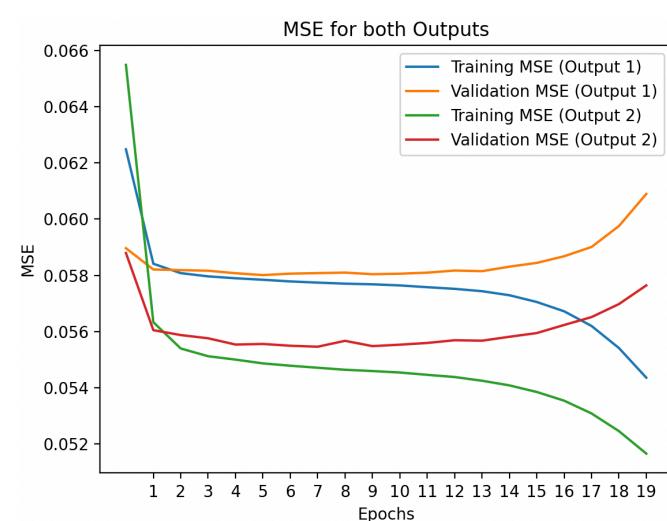
Απώλεια για 4 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, μέγεθος λανθάνουσας 512, ρυθμό εκμάθησης 0.0005



Απώλεια για 4 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, μέγεθος λανθάνουσας 128

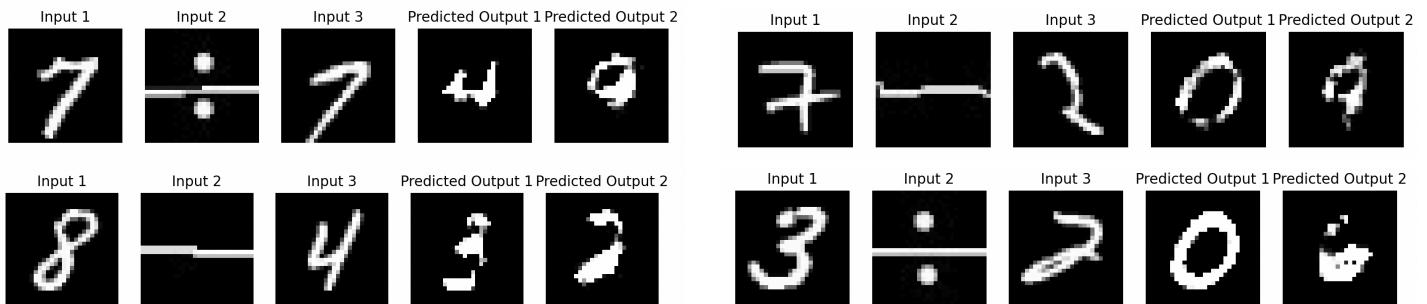


Απώλεια για 4 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, μέγεθος λανθάνουσας 1024

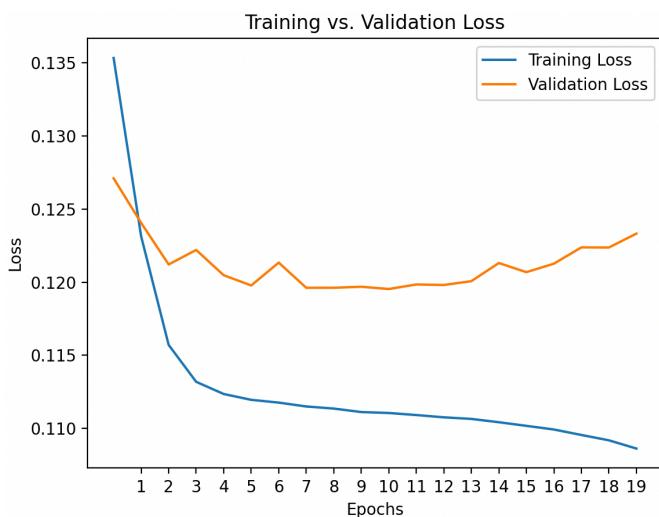


Απώλεια για 4 συνελικτικές, 3 αντίστροφες συνελικτικές στρώσεις, μέγεθος λανθάνουσας 1024

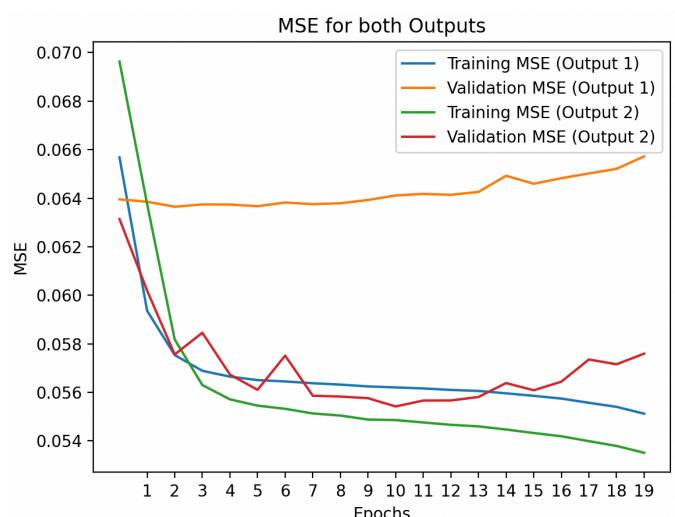
Είναι εμφανές ότι με μειωμένο ρυθμό εκμάθησης 0.0005 η απώλεια επαλήθευσης αρχίζει και πάλι να αυξάνεται, σε μικρότερο, ωστόσο, βαθμό. Το ίδιο συμβαίνει και για την αύξηση της λανθάνουσας σε 1024. Όμως, για μείωση της λανθάνουσας σε 128 φαίνεται να σταθεροποιείται η απώλεια και το *mean square error* κάθε εξόδου. Με δοκιμές χρησιμοποιώντας νέα δεδομένα από αυτά της εκπαίδευσης και επικύρωσης, παρατηρείται ότι το δίκτυο για πολλούς συνδυασμούς αριθμών θα εμφάνιζε σωστό αποτέλεσμα, αν ο τελεστής ήταν διαφορετικός.



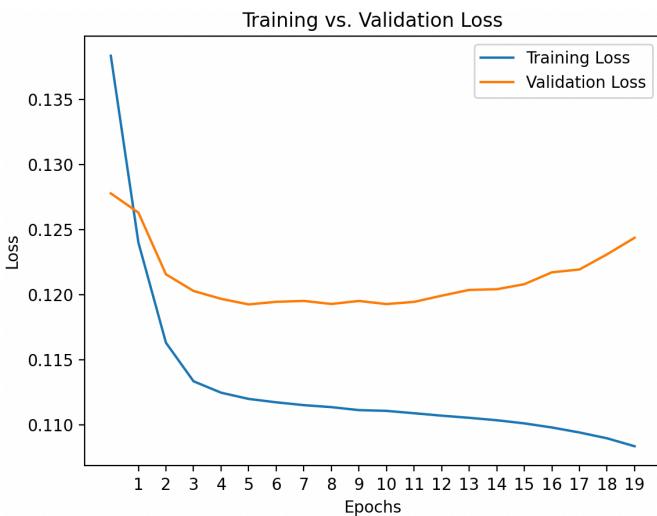
Επομένως, θα δοκιμαστεί η εισαγωγή βάρους στην είσοδο των δεδομένων της εικόνας του τελεστή. Μετά από 6 ώρες προκύπτουν οι μορφές:



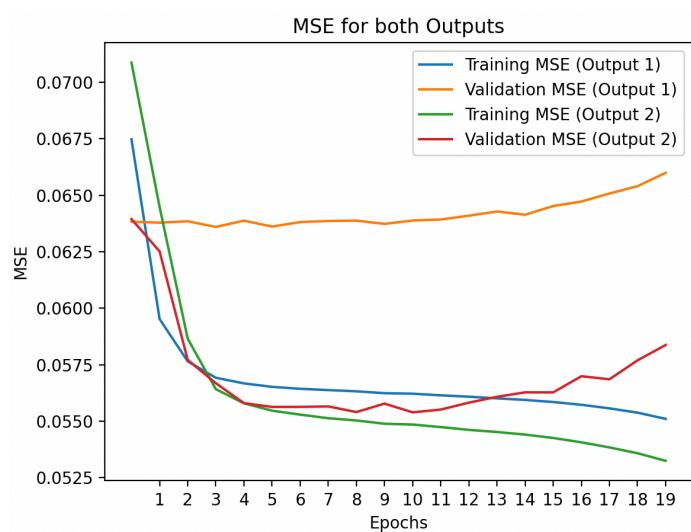
Απώλεια για συντελεστή βάρους 0.8



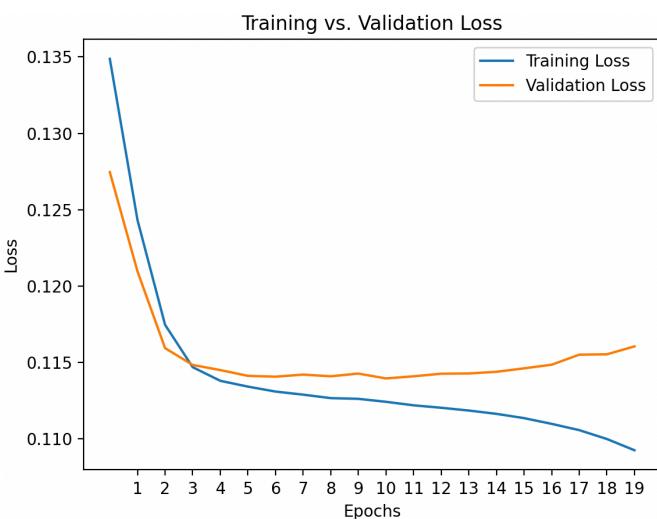
MSE για συντελεστή βάρους 0.8



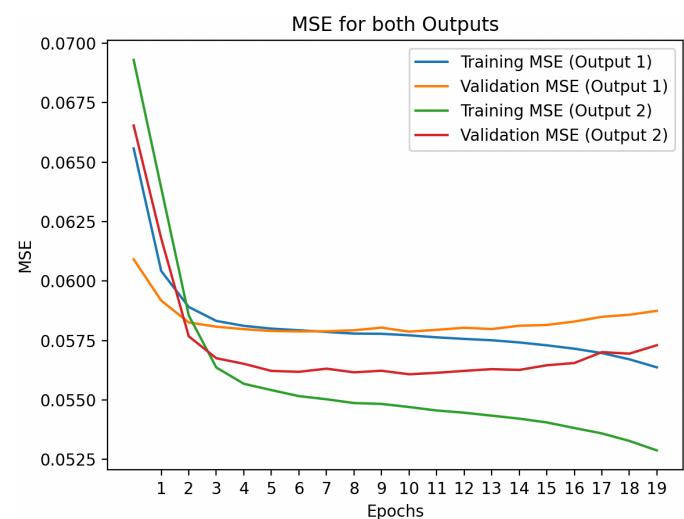
Απώλεια για συντελεστή βάρους 1.2



Απώλεια για συντελεστή βάρους 1.2

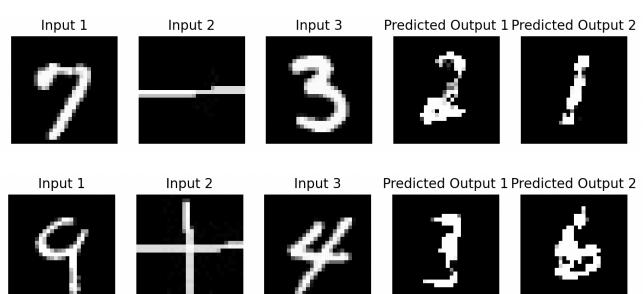
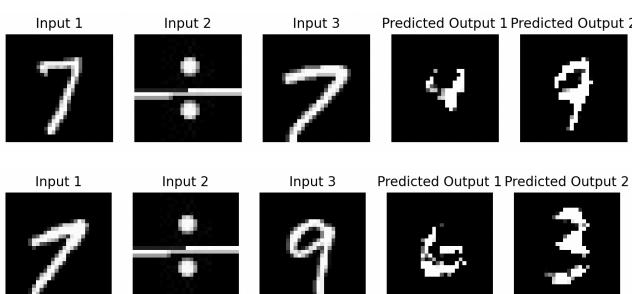


Απώλεια για συντελεστή βάρους 2

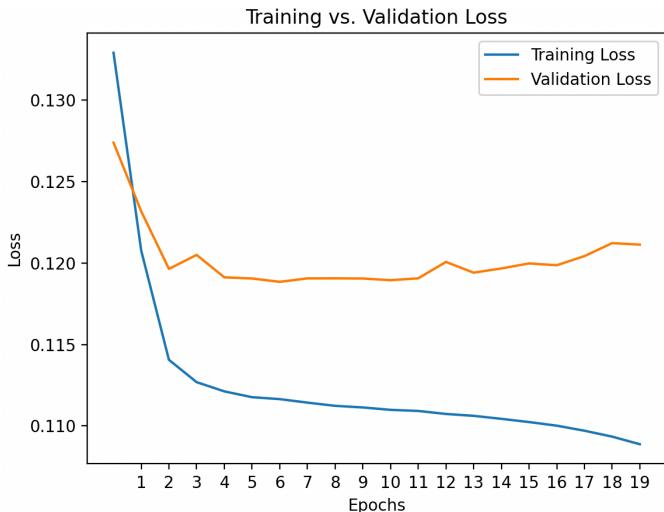


MSE για συντελεστή βάρους 2

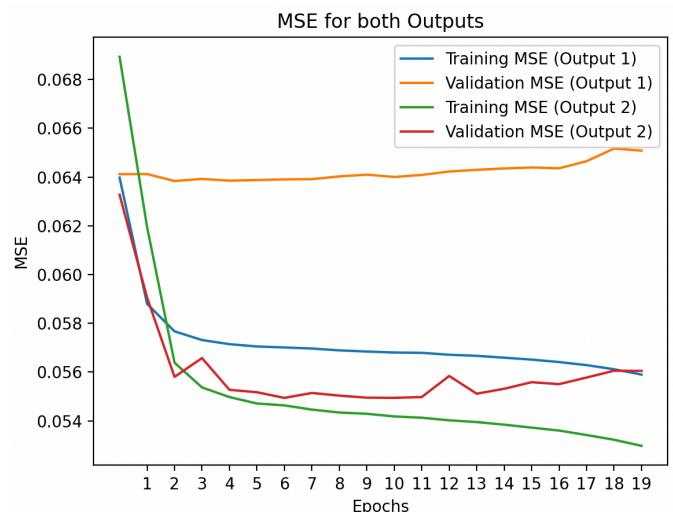
Αυτές οι τεχνικές δεν λειτούργησαν αφού παρατηρείται αύξηση της απώλειας συγκριτικά με το προηγούμενο μοντέλο. Με παραπάνω ελέγχους φαίνεται η πράξη του πολλαπλασιασμού να κυριαρχεί στα αποτελέσματα, ακόμη κι όταν δεν θα έπρεπε:



Με μείωση των δειγμάτων πολλαπλασιασμού στο σετ δεδομένων και μετά από 4 ώρες προκύπτει:

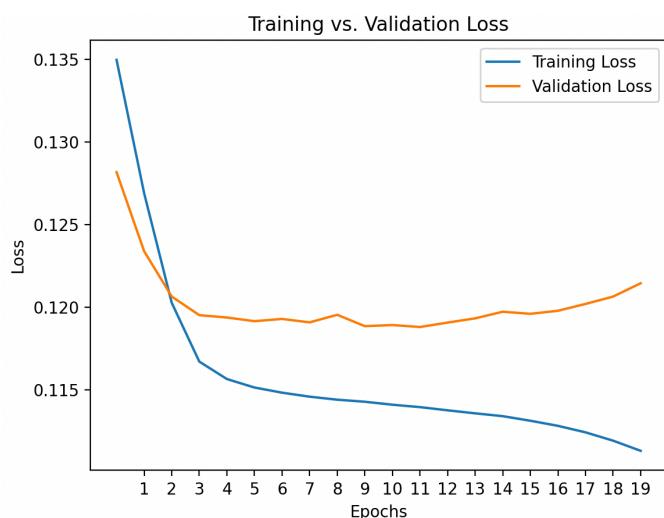


Απώλεια για αναλογία δειγμάτων πολ/μού 1 / 3

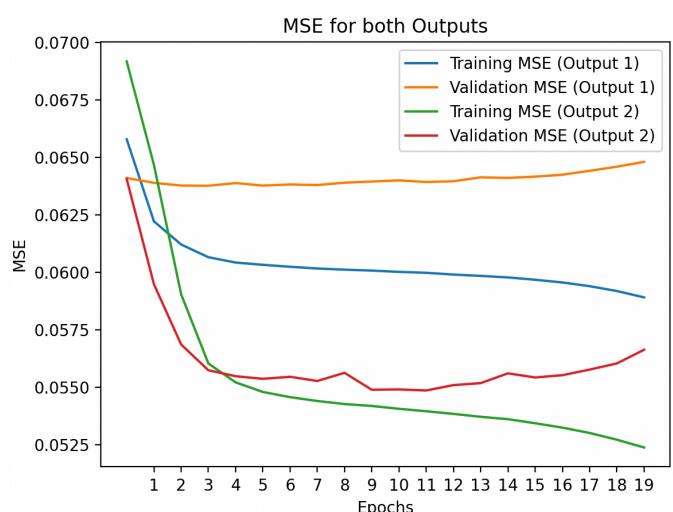


MSE για αναλογία δειγμάτων πολ/μού 1 / 3

Μετά από δοκιμές φαίνεται να κυριαρχεί η πράξη της πρόσθεσης, οπότε αφαιρούνται δείγματα και από αυτή την πράξη και μετά από 3 ώρες προκύπτει:



Απώλεια για αναλογία δειγμάτων πολ/μού και πρόσθεσης 1 / 3



MSE για αναλογία δειγμάτων πολ/μού και πρόσθεσης 1 / 3

Είναι εμφανές ότι οι τεχνικές αυτές δεν συμβάλουν στην μείωση της απώλειας. Επίσης, οι εικόνες εξόδου ήταν χαμηλότερες ποιότητας λόγω των περιορισμένων δειγμάτων. Εικάζεται ότι το φαινόμενο κατά το οποίο το νευρωνικό δίκτυο προτιμά να εκτελεί πράξεις πολλαπλασιασμού, ακόμη και όταν οι εισαγόμενες πράξεις είναι διαφορετικές, μπορεί να εξηγηθεί από τον τρόπο που υπολογίζεται η απώλεια και τη φύση των αριθμητικών πράξεων. Ο πολλαπλασιασμός παράγει γενικά μεγαλύτερα αποτελέσματα από τις υπόλοιπες πράξεις, όπως η πρόσθεση ή η αφαίρεση. Αυτό έχει ως αποτέλεσμα η απώλεια, που βασίζεται στη μέση τετραγωνική απόκλιση (Mean Squared Error - MSE), να ευνοεί μικρότερες αποκλίσεις όταν η πρόβλεψη είναι πιο κοντά σε αποτελέσματα πολλαπλασιασμού.

Άρα θα δοκιμαστεί η εισαγωγή βαρών στην απώλεια για να δοθεί μεγαλύτερη σημασία στις πράξεις που δεν αντιπροσωπεύονται καλά. Θα δημιουργηθεί ένα loss function που λαμβάνει υπόψη την πράξη.