ε.δε.μ²

ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ & ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΕΡΓΑΣΙΑ:
Επιτάχυνση εκπαίδευσης νευρωνικού δικτύου σε αρχιτεκτονικές κοινής μνήμης με OpenMP και CUDA

Εργαστήριο Υπολογιστικών Συστημάτων (CSLab)

ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση

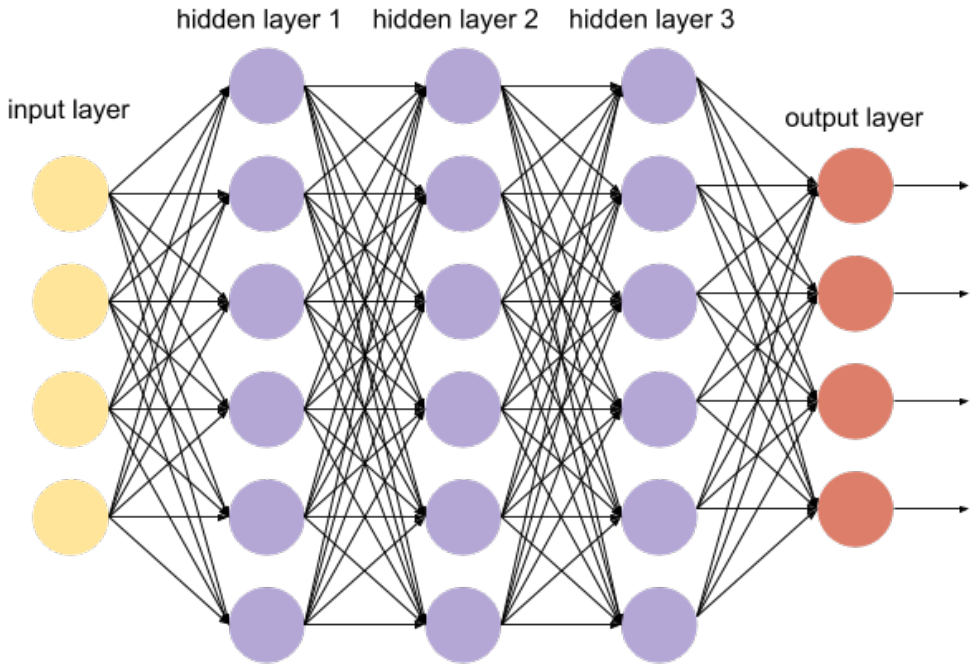
Σχολή ΗΜΜΥ, Ε.Μ.Π.

2020

Εργασία

Εισαγωγικά

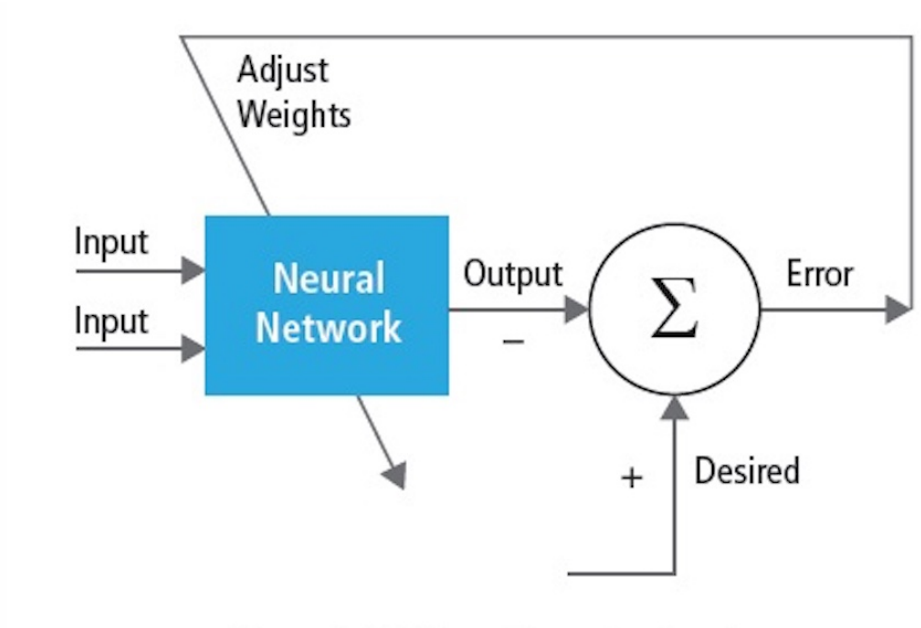
- Σας δίνεται νευρωνικό δίκτυο με 3 κρυφά επίπεδα
- Ζητείται η εκπαίδευση του στη βάση δεδομένων MNIST για την αναγνώριση ψηφίων



Εργασία

Εκπαίδευση Νευρωνικών Δικτύων

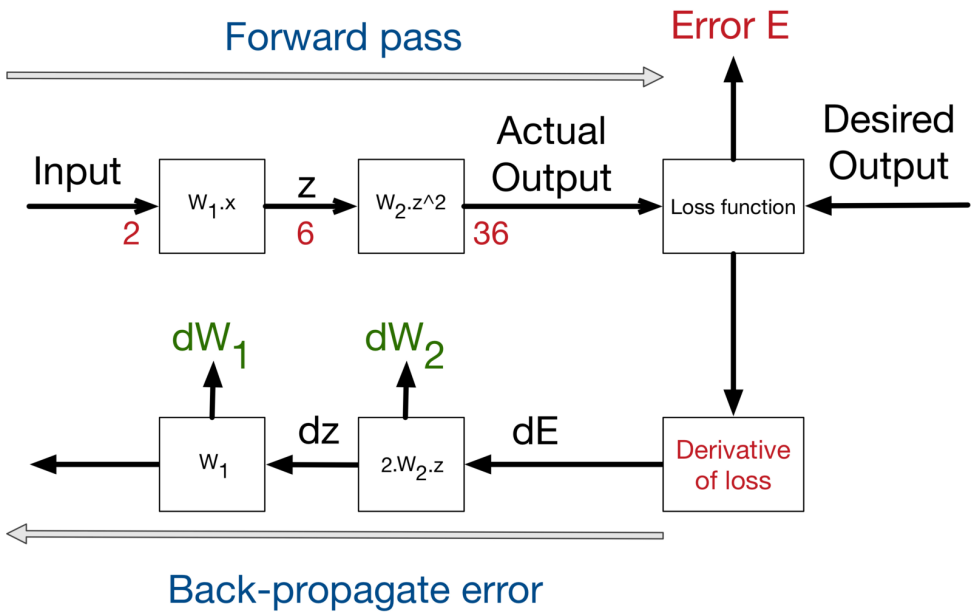
- Πώς γίνεται η εκπαίδευση;
  - Ορίζουμε μία συνάρτηση σφάλματος (loss function)
  - Αρχικοποιούμε τα βάρη του μοντέλου με τυχαίες τιμές
  - Προσαρμόζουμε σταδιακά τις τιμές των βαρών με στόχο την ελαχιστοποίηση της συνάρτησης σφάλματος σε μία βάση δεδομένων εκπαίδευσης (training set)



Εργασία

Εκπαίδευση Νευρωνικών Δικτύων

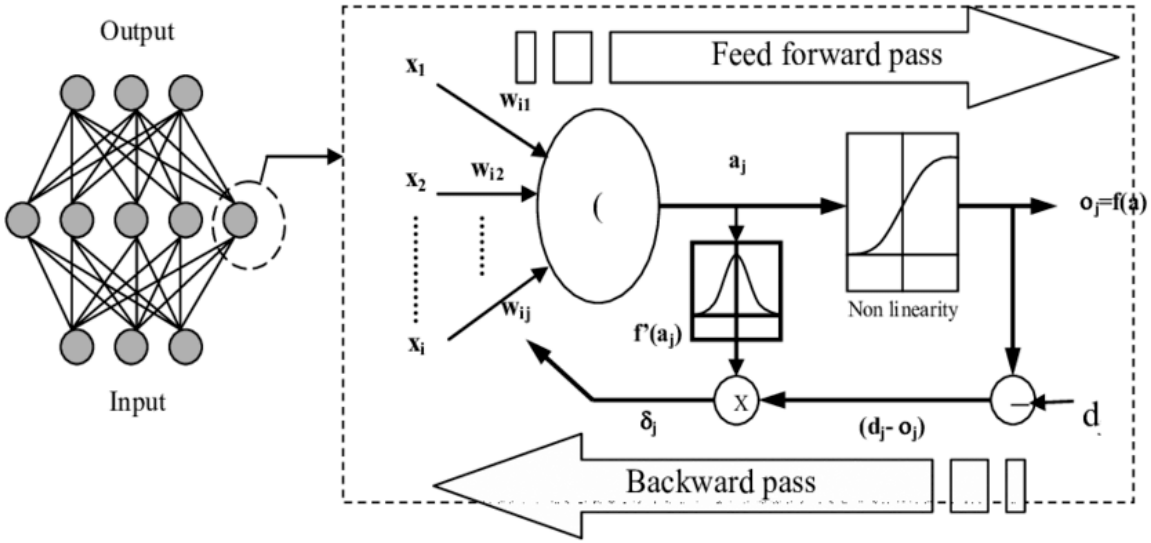
- Πώς προσαρμόζουμε τα βάρη ώστε να ελαχιστοποιήσουμε το σφάλμα;
  - Για κάθε δεδομένο της βάσης εκπαίδευσης υπολογίζουμε την έξοδο (forward pass)
  - Με βάση την πραγματική και επιθυμητή έξοδο υπολογίζουμε το σφάλμα
  - Με βάση το σφάλμα αλλάζουμε τις τιμές στα βάρη (back-propagation)



Εργασία

Εκπαίδευση Νευρωνικών Δικτύων

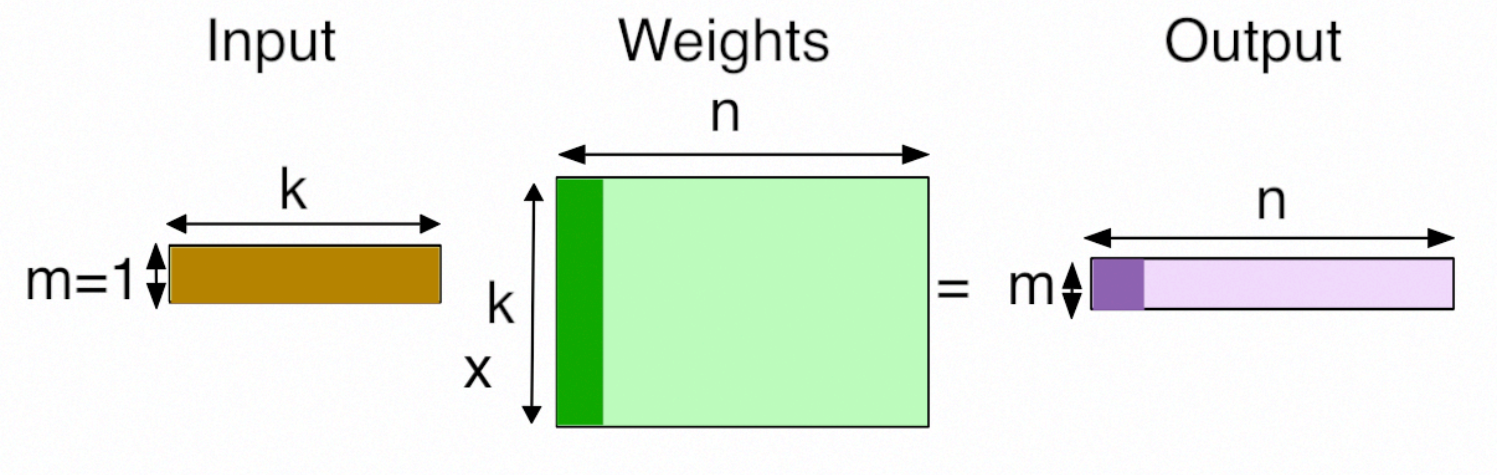
- Πώς προσαρμόζουμε τα βάρη με βάση την τιμή του σφάλματος;
  - Χρησιμοποιούμε τον αλγόριθμο Gradient Descent
    - Υπολογίζουμε την παράγωγο της συνάρτησης σφάλματος ως προς τα βάρη
    - Προσαρμόζουμε τις τιμές των βάρων προς την αντίθετη κατεύθυνση
  - Για τον υπολογισμό των παραγώγων χρησιμοποιούμε την τεχνική του back-propagation



Εργασία

Εκπαίδευση Νευρωνικών Δικτύων

- Το μεγαλύτερο μέρος του χρόνου εκπαίδευσης αναλώνεται σε πράξεις γραμμικής άλγεβρας (πολλαπλασιασμός πινάκων)
- Με βάση το νόμο του Amdahl αξίζει να προσπαθήσουμε να βελτιώσουμε την επίδοσή τους



Εργασία

Ζητούμενα

- Σας δίνεται ο σκελετός της άσκησης ο οποίος υλοποιεί όλα τα βήματα εκπαίδευσης:
  - σε CPU με χρήση του προγραμματιστικού μοντέλου OpenMP
  - σε GPU με χρήση του προγραμματιστικού μοντέλου CUDA
- Ζητείται η παραλληλοποίση των υπολογισμών γραμμικής άλγεβρας που χρησιμοποιούνται κατά την εκπαίδευση σε OpenMP και CUDA
- Ζητείται η σύγκριση των επιδόσεων ανάμεσα σε υλοποιήσεις της CPU
- Ζητείται η σύγκριση των επιδόσεων ανάμεσα σε υλοποιήσεις της GPU
- Ζητείται η σύγκριση των επιδόσεων ανάμεσα σε CPU και GPU