UNIVERSITY OF THESSALY
SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL
& COMPUTER ENGINEERING

Νευρο-Ασαφής Υπολογιστική

Χειμερινό Εξάμηνο 2023-2024

Δημήτριος Κατσαρός

Coding project

Ημέρα ανακοίνωσης: Thursday, November 30, 2023

Προθεσμία παράδοσης: Κυριακή, Φεβρουάριος 18, 2024

Περιγραφή προβλήματος

Η κατηγοριοποίηση κειμένου είναι μια από τις θεμελιώδεις εργασίες στον τομέα της μάθησης. Στο παρόν έργο ανάπτυξης κώδικα καλείστε ν' αναπτύξετε έναν κατηγοριοποιητή (classifier) για την κατηγοριοποίηση σχετικά μικρού κειμένου σε κατηγορίες και υποκατηγορίες. Καλείστε ν' αναπτύξετε ένα νευρωνικό δίκτυο για να φέρετε σε πέρας την εργασία. Το είδος του νευρωνικού δικτύου του κατηγοριοποιητή, η αρχιτεκτονική του κ.τ.λ. είναι ελεύθερο να επιλεγούν από την κάθε ομάδα.

Το σύνολο δεδομένων εκπαίδευσης και testing βρίσκεται εδώ https://courses.e-ce.uth.gr/CE418/nfc_fall23/news-classification.csv: Είστε ελεύθεροι να χρησιμοποιήσετε όπως επιθυμείτε τα δεδομένα αυτά για την εκπαίδευση, έλεγχο. Το σύνολο αυτό περιέχει περίπου 10917 άρθρα-ειδήσεις με ιεραρχικές κατηγορίες ειδήσεων τα οποία έχουν συλλεγεί το 2019 και έχουν ταξινομηθεί με την ταξινόμηση NewsCodes Media Topic. Τα δεδομένα έχουν κατηγοριοποιηθεί χειρωνακτικά σε 17 κατηγορίες πρώτου επιπέδου και 109 κατηγορίες δευτέρου επιπέδου.

Πριν την προθεσμία παράδοση θα σάς δοθεί ένα άλλο σύνολο δεδομένων όπου θα κάνετε και την τελική αποτίμηση του νευρωνικού σας δικτύου.

Οι δυο αρχιτεκονικές με την ακριβέστερη πρόβλεψη θα έχουν το προνόμιο ν' αριστεύσουν στο μάθημα, ανεξάρτητα των επιδόσεων των μελών της ομάδας στις άλλες συνιστώσες του μαθήματος.

Ως παραδοτέο ζητείται, να γράψετε μια αναφορά που, εκτός των όποιων γενικών πληροφοριών, θα περιέχει τα ακόλουθα:

- λεπτομέρειες της αρχιτεκτονικής του νευρωνικού δικτύου,
- τον αλγόριθμο εκπαίδευσής του,
- τις τελικές (εκπαιδευμένες) τιμές των παραμέτρων του,
- αποτίμηση της επίδοσής του πάνω στα δεδομένα που έχετε διαθέσιμα,
- καταγραφή των χρόνων εκπαίδευσης (training time) του δικτύου ως συνάρτηση του μεγέθους των δεδομένων εισόδου (παρουσιάζοντας σταδιακά τα διαθέσιμα δεδομένα),
- καταγραφή του χρόνου απόκρισης (inference time) του δικτύου.

Παρακαλείστε να δημιουργήσετε τον κώδικα κατά τέτοιον τρόπο, ώστε το module το οποίο διαβάζει τα input data να είναι ανεξάρτητο από το υπόλοιπο που εκτελεί το training και το inference του νευρωνικού δικτυου. Με τον τρόπο αυτό επιτυγχάνουμε ευελιξία για την χρήση διαφορετικών μορφών δεδομένων εισόδου, αφού αρκεί ένας απλός wrapper για την μετατροπή των σε δεδομένα "καταναλώσιμα" από το νευρωνικό μας δίκτυο.

Για όσους επιλέξουν να παραδόσουν το coding project τον Σεπτέμβριο, θα είναι το ίδιο. Είναι προφανές ότι η παράδοση κώδικα ο οποίος αντλήθηκε αυτούσιος ή σχεδ'ν αυτούσιος από το Διαδίκτυο δεν ειναι αποδεκτή. Η χρήση όμως ξένων κομματιών κώδικα – τα οποία πρέπει ρητά να αναγνωριστούν ως τέτοια – είναι επιτρεπτή.

Χρηστικές πληροφορίες:

Η προθεσμία παράδοσης είναι αυστηρή. Είναι όμως δυνατή η παροχή παράτασης (μέχρι 4 ημέρες), αλλά μόνο αφού δώσει ο διδάσκων την έγκρισή του, και αυτή η παράταση στοιχίζει 10% ποινή στον τελικό βαθμό της. Η παράδοση γίνεται με email (dkatsar@e-ce.uth.gr) του πηγαίου κώδικα, καθώς της αναφοράς που περιέχει την (σύντομη) περιγραφή του κώδικα, και των αποτελεσμάτων της πειραματικής αξιολόγησης. Το subject του μηνύματος πρέπει να είναι: CE418-Project: AEMx-AEMy

Ερμηνεία συμβόλου:

Απαιτεί την ανάπτυξη κώδικα σε PyTorch, Matlab, Tensorflow (Keras, …), κ.τ.λ. Εάν χρησιμοποιήσετε έτοιμο κώδικα από κάποια πηγή απαιτείται να δηλώσετε την πηγή, καθώς και σε ποιο σημείο του project τον χρησιμοποιήσατε.