Άσκηση 3

Ταξινόμηση χρονοσειρών πολυφασματικών δορυφορικών δεδομένων

Αντικείμενο - Στόχοι

- ✔ Εξοικείωση με τηλεπισκοπικά δεδομένα χρονοσειρών
- ✔ Σχεδιασμός και υλοποίηση μιας μεθοδολογίας ταξινόμησης χρονοσειρών ΤΝΔ
- ✔ Εξοικείωση με αρχιτεκτονικές Transformer
- ✔ Μελέτη open source κώδικα
- ✔ Υλοποίηση μέρους επιστημονικής δημοσίευσης

Δεδομένα & Προεπεξεργασία

Καλείστε να εργαστείτε σε ένα υποσύνολο του σετ δεδομένων Timematch (περισσότερες πληροφορίες **εδώ**). Το Timematch αποτελείται από ετήσιες χρονοσειρές πολυφασματικών δεδομένων Sentinel-2 σε 4 περιοχές της Ευρώπης. Το πρόβλημα που πραγματεύεται το συγκεκριμένο σετ δεδομένων είναι η ταξινόμηση του εκάστοτε αγροτεμαχίου σε μία εκ των 15 διαθέσιμων κατηγοριών καλλιέργειας (crop type classification). Επιπρόσθετα παρέχεται η δυνατότητα μελέτης του προβλήματος της μη επιβλεπόμενης αλλαγής domain (unsupervised domain adaptation), πρόβλημα που δεν καλύπτεται από την παρούσα άσκηση.

Για την άσκηση παρέχεται ένα υποσύνολο του σετ δεδομένων που αφορά μία μεμονωμένη περιοχή (Δανία). Για τεχνικούς λόγους έχει αφαιρεθεί σημαντικός όγκος δεδομένων ώστε να καταστεί το πρόβλημα πιο διαχειρίσιμο για την άσκηση. Το υποσύνολο δεδομένων είναι διαθέσιμο **εδώ**.

Η δομή των δεδομένων έχει ως εξής:

- Κάθε αγροτεμάχιο (parcel), ως η βασική δομή του σετ, αποτελείται από ένα (μη σταθερό) πλήθος εικονοστοιχείων
- Για κάθε εικονοστοιχείο είναι διαθέσιμη μία χρονοσειρά πολυφασματικών δεδομένων (10 κανάλια) με σταθερό πλήθος διαθέσιμων ημερομηνιών (52)
- Η ακριβής ημερομηνία απόκτησης κάθε δείγματος της χρονοσειράς είναι διαθέσιμη στα μεταδεδομένα του κάθε αγροτεμαχίου και είναι κοινή για όλα τα parcels.

Στα πλαίσια της άσκησης καλείστε να πραγματοποιήσετε τα ακόλουθα:

- Ελέγξτε πόσες και ποιες κατηγορίες είναι διαθέσιμες στο υποσύνολο δεδομένων της άσκησης. Ορίστε νέο indexing για τις κατηγορίες.
- Φιλτράρετε και αφαιρέστε από τη συνέχεια τις κατηγορίες εκείνες που διαθέτουν λιγότερα από 200 παραδείγματα.
- Δημιουργείστε κατάλληλους αλγορίθμους τροφοδότησης δεδομένων
  - Κανονικοποιήστε τα δεδομένα
  - Χρησιμοποιήστε κατάλληλη τεχνική τυχαίας δειγματοληψίας ώστε να συλλέγετε σταθερό αριθμό εικονοστοιχείων ανά parcel (ενδεικτικά 32 εικονοστοιχεία). Προσοχή, η τυχαία δειγματοληψία αφορά μόνο το training set και εφαρμόζεται με τυχαίο τρόπο σε κάθε εποχή εκπαίδευσης.
  - Μπορείτε να μελετήσετε την επίσημη υλοποίηση από τους δημιουργούς του dataset **εδώ**.

Μεθοδολογία

Σχεδιάστε και αναπτύξτε ένα μοντέλο ταξινόμησης χρονοσειρών με τις εξής προδιαγραφές:

- Ως δεδομένα εισόδου θα λαμβάνετε το εκάστοτε parcel με συγκεκριμένο πλήθος εικονοστοιχείων κατά το training και μεταβλητό κατά το validation
- Σχεδιάστε μία αρχιτεκτονική Pixel Set Encoder παρόμοια με αυτή που περιγράφεται **εδώ (Figure 2 - pixel set encoder)**.
  - Στόχος αυτής της αρχιτεκτονικής είναι η ενσωμάτωση της χωρικής διάστασης των δεδομένων. Η έξοδος της έχει μόνο χρονική διάσταση και διάσταση χαρακτηριστικών (features).
- Χρησιμοποιήστε μία Transformer Encoder αρχιτεκτονική για την ενσωμάτωση της χρονικής διάστασης. Η έξοδος της αποτελείται μόνο από ένα διάνυσμα χαρακτηριστικών.
  - Χρησιμοποιήστε κατάλληλη ημιτονοειδή χρονική κωδικοποίηση (sinusoidal temporal positional encoding) αξιοποιώντας τα διαθέσιμα μεταδεδομένα με τις ημερομηνίες απόκτησης κάθε δείγματος της χρονοσειράς.
  - Χρησιμοποιήστε κατάλληλο **token** ταξινόμησης για να ενσωματώσετε τη χρονική διάσταση.
- Χρησιμοποιήστε ένα απλό MLP δικού σας σχεδιασμού για την ταξινόμηση των διανυσμάτων.

    - Ως συνάρτηση κόστους χρησιμοποιήστε την Categorical Cross Entropy
- Εκπαιδεύστε και αξιολογήστε το μοντέλο σας με τις συνήθεις μετρικές ταξινόμησης. (Προτείνεται η χρήση της βιβλιοθήκης torchmetrics)
    - Πίνακας σύγχυσης
    - Micro και weighted μετρικές F1, Accuracy, Precision, Recall
    - Ερμηνεύστε τα αποτελέσματα.

Στρατηγική εκπαίδευσης και ελέγχου

Υλοποιήστε μία στρατηγική k-fold cross validation (k=5) για να μελετήσετε την επίδοση του μοντέλου σας

- Χωρίστε τα δεδομένα σας σε 5 ίσα τμήματα
- Σε κάθε φάση θεωρήστε τα 4 ως σετ εκπαίδευσης και το 1 ως validation set.
- Για κάθε μετρική υπολογίστε μέσο όρο και τυπική απόκλιση μεταξύ των 5 πειραμάτων.

Ζητούμενα

εκπονήστε τεχνική έκθεση περιγράφοντας τις διαδικασίες που ακολουθήσατε, απαντώντας και στα παρακάτω ερωτήματα

1. (100%) Εκτελέστε τα παραπάνω βήματα.

Παράρτημα

Ιεραρχική δομή δεδομένων. Τα .zarr αρχεία περιέχουν τις χρονοσειρές για κάθε parcel. z Απαιτείται η βιβλιοθήκη zarr να είναι εγκατεστημένη (!pip install zarr). Το αρχείο **dates.json** αναφέρει τις ημερομηνίες απόκτησης κάθε δείγματος σε μορφή EEEEMMHH, χρησιμοποιήστε το για αναγωγή σε day-of-year 2017, δηλαδή πλήθος ημερών από 1/1/2017. Το αρχείο **labels.json** παρέχει την κατηγορία στην οποία ανήκει το κάθε parcel. Τα .json αρχεία μπορείτε να τα διαβάσετε με τη βιβλιοθήκη json της Python (std lib).