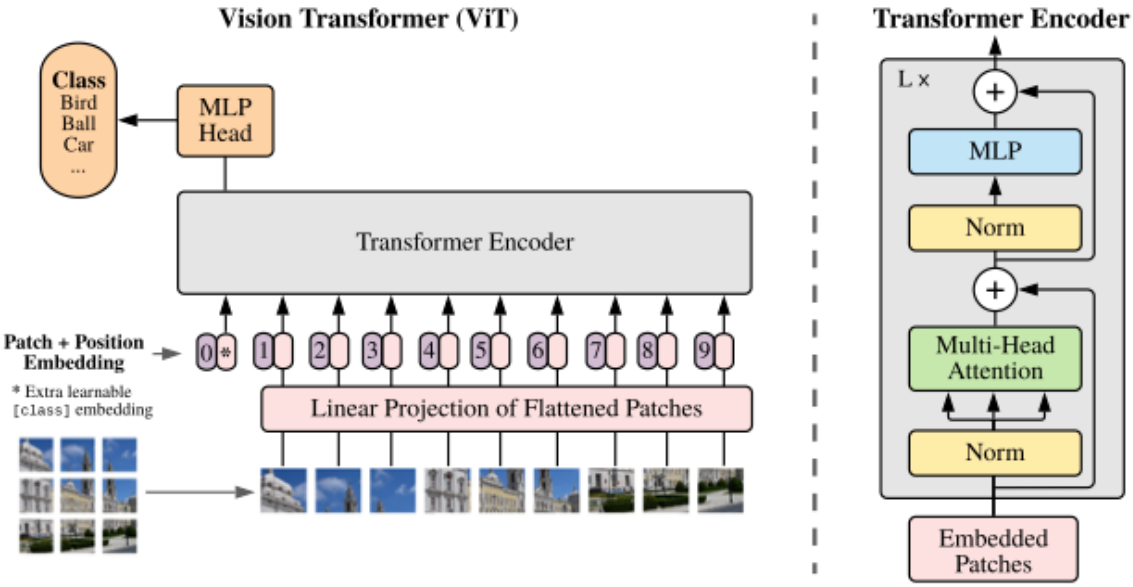
ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
HAROKOPIO UNIVERSITY

Μάθημα: Υπολογιστική όραση

Εργασία #2

Ημερομηνία παράδοσης: 02/02/2024

Άσκηση #1, Αρχιτεκτονική Vision Transformer (ViT): Θεωρήστε την αρχιτεκτονική νευρωνικών δικτύων 'Vision Transformer', όπως αυτή αναλύεται στην ακόλουθη δημοσίευση[1] και παρατίθεται στο παρακάτω σχήμα: Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020). Στόχος της είναι η αυτόματη ταξινόμηση μιας εικόνας εισόδου σε μια κλάση από ένα προκαθορισμένο σύνολο σημασιολογικών εννοιών.



Η αρχιτεκτονική στηρίζεται στα ακόλουθα βασικά δομικά στοιχεία (στάδια επεξεργασίας):

1. Μετατροπή της εικόνας εισόδου σε τεμάχια (patches)
2. Χρήση γραμμικής αντιστοίχισης για τη δημιουργία δειγμάτων (tokens) από τα τεμάχια
3. Δημιουργία εκπαιδεύσιμων δειγμάτων ταξινόμησης
4. Προσθήκη κωδικοποίησης/πληροφορίας θέσης σε κάθε δείγμα
5. Εφαρμογή στρώματος κωδικοποίησης (encoder)
6. Εκτέλεση σταδίου ταξινόμησης (classification)

Σε περιβάλλον Google Colab[2] φορτώστε το notebook 'ViT_tutorial.ipynb', το οποίο περιέχει την αναλυτική υλοποίηση και επεξήγηση για την εκπαίδευση (δημιουργία) και εκτέλεση (παραγωγή προβλέψεων ταξινόμησης) του οπτικού ταξινομητή 'ViT'.

Ζητούμενα:

A. Εκπαιδεύστε έναν 'ViT' ταξινομητή για το σύνολο δεδομένων MNIST[3], όπου στόχος είναι η αναγνώριση χειρόγραφων ψηφίων. Χρησιμοποιήστε τις προκαθορισμένες παραμέτρους εκπαίδευσης του μοντέλου.

[1] https://arxiv.org/pdf/2010.11929.pdf
[2] https://colab.research.google.com
[3] http://yann.lecun.com/exdb/mnist/

Β. Μελετήστε όλα τα επιμέρους στάδια επεξεργασίας και τα ενδιάμεσα αποτελέσματα που υπολογίζονται.
Γ. Υπολογίστε και τυπώστε προβλέψεις ταξινόμησης για τυχαία επιλεγμένες εικόνες του συνόλου δεδομένων MNIST.

Άσκηση #2, Εκπαίδευση αρχιτεκτονικής ViT στο σύνολο δεδομένων 'Fashion MNIST': Φορτώστε το notebook 'ViT_cifar_fashion_MNIST.ipynb', το οποίο περιέχει την υλοποίηση ενός οπτικού ταξινομητή 'ViT' στο σύνολο δεδομένων 'CIFAR-10'[4]. Το σύνολο 'CIFAR-10' αποτελείται από 60,000 έγχρωμες εικόνες διαστάσεων 32x32 που ανήκουν σε 10 σημασιολογικές κατηγορίες (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck).

Ζητούμενα:

Α. Εκπαιδεύστε έναν 'ViT' ταξινομητή για το σύνολο δεδομένων 'Fashion MNIST'[5] τροποποιώντας κατάλληλα τον κώδικα (οι ρουτίνες που είναι απαραίτητες για τη φόρτωση των δεδομένων του 'Fashion MNIST' παρέχονται σε μορφή σχολίων). Χρησιμοποιήστε τις προκαθορισμένες παραμέτρους εκπαίδευσης του μοντέλου, δηλαδή τις ίδιες με αυτές για το σύνολο δεδομένων 'CIFAR-10'. Το σύνολο 'Fashion MNIST' αποτελείται από 70,000 γκρι (grayscale) εικόνες διαστάσεων 28x28 που ανήκουν σε 10 σημασιολογικές κατηγορίες σχετικές με το πεδίο της μόδας (T-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, ankle boot).
Β. Εκπαιδεύστε τον ταξινομητή 'ViT' με διαφορετικές τιμές για τις βασικές παραμέτρους της διαδικασίας εκπαίδευσης ('N_EPOCHS': πλήθος εποχών εκπαίδευσης) και της αρχιτεκτονικής του νευρωνικού δικτύου ('num_layers': πλήθος στρωμάτων, 'num_heads': πλήθος κεφαλών). Υπολογίστε και τυπώστε για κάθε διαφορετική παραμετροποίηση τα αποτελέσματα της διαδικασίας εκπαίδευσης (καμπύλες των τιμών 'accuracy' και 'loss').
Γ. Για την καλύτερη παραμετροποίηση εκπαίδευσης που έχει προκύψει στο ζητούμενο Β, υπολογίστε και τυπώστε προβλέψεις ταξινόμησης για τυχαία επιλεγμένες εικόνες του συνόλου δεδομένων 'Fashion MNIST'.

[4] https://www.cs.toronto.edu/~kriz/cifar.html
[5] https://www.kaggle.com/datasets/zalando-research/fashionmnist