

# 1st Project

The background of the slide features a series of overlapping diagonal bands. A wide blue band runs from the bottom-left towards the top-right. Overlapping this from the top-left is an orange band. Another orange band runs parallel to the blue one, positioned below it. At the bottom, a light gray band runs parallel to the others.

Σπύρος Κούγιας

## Dataset:

- Επιλέχτηκε το cifar-10 που περιέχει 60000 εικόνες 32x32x3 που κατανέμονται σε 10 κλάσεις.
- Τα δεδομένα είναι ήδη χωρισμένα σε training batches(0-5) των 10000 εικόνων και σε ένα testing batch 10000 εικόνων.
- Περισσότερες πληροφορίες στο site του cifar-10:  
<https://www.cs.toronto.edu/~kriz/cifar.html>
- Τέλος η εισαγωγή των δεδομένων έγινε με την εγκατάστασή τους σε ανάλογο φάκελο, το path του οποίου μπαίνει ως είσοδος όταν κανείς τρέξει το script.

## Intermediate project:

### Ενδιάμεση Εργασία

Να γραφεί πρόγραμμα σε οποιαδήποτε γλώσσα επιθυμείτε το οποίο να συγκρίνει την απόδοση του κατηγοριοποιητή πλησιέστερου γείτονα με 1 και 3 πλησιέστερους γείτονες με τον κατηγοριοποιητή πλησιέστερου κέντρου στην βάση δεδομένων που θα επιλέξετε για την εργασία σας. Το πρόγραμμα δηλαδή αυτό θα πρέπει να διαβάζει τα δεδομένα εκπαίδευσης (training) και τα δεδομένα ελέγχου (test) και να μετράει την απόδοση των παραπάνω κατηγοριοποιητών.

Γράφτηκε script σε python που υλοποίησε τους τρεις κατηγοριοποιητές:

- 1-NN
- 3-NN
- N-Centroid

Σαν είσοδο δέχεται ολόκληρη την εικόνα.

Γίνεται μια προ επεξεργασία των δεδομένων και συγκεκριμένα η αφαίρεση του μέσου όρου και η μείωση της κλίμακας (0-255 => 0-1).

Χρησιμοποιήθηκε η pytorch και sklearn για γρηγορότερες πράξεις πινάκων στον υπολογισμό των αποστάσεων και για μέτρηση των αποδόσεων.

Πρέπει να σημειωθεί πως αποστάσεις υπολογίστηκαν με τον τύπο:

$$\|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2 \cdot x \cdot y$$

Επίσης προστέθηκε, προαιρετικά επιλογή για χρήση PCA και διαχωρισμός των δεδομένα σε chunks έτσι ώστε να γίνονται γρηγορότερα οι υπολογισμοί (όμως επιβαρύνεται η μνήμη).

## Πειράματα-Intermediate:

(Επανάληψη: 100 φορές το καθένα)

**PCA {50, 200, 300, 500}**

**CHUNKS {1, 200, 400}**

pca_dim	chunk_size	use_pca	1NN_acc	3NN_acc	centroid_acc
50	1	TRUE	0.29283	0.31696	0.261985
50	200	TRUE	0.29342	0.31688	0.262015
50	400	TRUE	0.29313	0.31761	0.261995
200	1	TRUE	0.27891	0.28945	0.262355
200	200	TRUE	0.27924	0.29004	0.26237
200	400	TRUE	0.27899	0.28922	0.262355
300	1	TRUE	0.26884	0.28493	0.263
300	200	TRUE	0.26875	0.28475	0.263
300	400	TRUE	0.26886	0.28467	0.263
500	1	TRUE	0.26271	0.27593	0.26279
500	200	TRUE	0.26255	0.27544	0.262835
500	400	TRUE	0.26294	0.27562	0.262845

Και τα αντίστοιχα stds:

1NN_std	3NN_std	centroid_std
0.002314011	0.002428451	0.000544624
0.002689974	0.002375294	0.000646226
0.002427348	0.002604949	0.000707062
0.00214191	0.002001893	0.000480895
0.002216194	0.002155191	0.000485125
0.002271897	0.002057654	0.000480207
0.001178254	0.001478841	0
0.001217507	0.001431076	0
0.001154875	0.001511271	0
0.001208514	0.001416034	0.00040636
0.001358661	0.001451714	0.000372989
0.001277782	0.001369067	0.000363662

## NO PCA

### CHUNKS {1, 200, 400}

use_pca	chunk_size	use_pca	1NN_acc	3NN_acc	centroid_acc
FALSE	1	FALSE	0.282	0.285	0.256
FALSE	200	FALSE	0.282	0.285	0.256
FALSE	400	FALSE	0.282	0.285	0.256

Αναμενόμενα, η επιλογή διαφορετικού chunk size δεν επηρεάζει σημαντικά την ευστοχία της κατηγοριοποίησης, καθώς είχε αποκλειστικό σκοπό να βελτιώσει την ταχύτητα με τίμημα την χρήση μνήμης.

Η εφαρμογή PCA βελτιώνει την ταχύτητα του αλγορίθμου αλλά φαίνεται να αυξάνει μερικώς και την ευστοχία του, πιθανότατα επειδή αποθρυβοποιεί τα δεδομένα.

Το συμπέρασμα που μπορούμε να τραβήξουμε από αυτά τα δεδομένα είναι ότι καλύτερα αποτελέσματα έχουμε με 3-NN (με Pca) που παρουσιάζει ευστοχία 31.7%.

Και 3-NN (χωρίς Pca) με ευστοχία 28.5%.

Επίσης παρατηρούμε ότι μικρότερες διαστάσεις pca, για τον 1-NN και 3NN, δίνουν καλύτερα αποτελέσματα (κοντά στις 50). Ενώ αντιθέτως ο centroid κατηγοριοποιητής δείχνει πολύ μικρή βελτίωση κοντά στις 300 διαστάσεις pca και όχι στις 50.