Ανάκτηση Εικόνας Βάσει Περιεχομένου

Ευστάθιος Χατζηκυριακίδης

Η γενική εικόνα

Σε αυτήν την εργασία υλοποιήσαμε ένα σύστημα ανάκτησης εικόνων βάσει περιεχομένου. Πιο συγκεκριμένα, η βασική δομή του συστήματος ανάκτησης εικόνων που υλοποιήθηκε αποτελείται από την εξαγωγή των χαρακτηριστικών των εικόνων από την αναπαράσταση τους στα εσωτερικά και κρυφά επίπεδα ενός συνελικτικού νευρωνικού δικτύου που εκπαιδεύτηκε για κατηγοριοποίηση πολλών κλάσεων. Επίσης, έγινε η μέτρηση της απόστασης του διανύσματος χαρακτηριστικών της εικόνας που χρησιμοποιείται ως ερώτημα προς αναζήτηση από κάθε άλλη εικόνα της βάσης και χρήση της μεθόδου των K κοντινότερων γειτόνων για την εμφάνιση των εικόνων που μοιάζουν περισσότερο χρησιμοποιώντας ως μετρική την απόσταση συνημίτονου. Τέλος, έγινε εξαγωγή μετρικών ακρίβειας και ανάκλησης για διάφορα πειράματα χρησιμοποιώντας διαφορετικά κρυφά επίπεδα του συνελικτικού νευρωνικού δικτύου, διαφορετικών μετρικών απόστασης και διαφόρων K κοντινότερων γειτόνων.

Τεχνολογίες που χρησιμοποιήθηκαν

- Anaconda - 4.5.4
- Spyder - 3.2.8
- Tensorflow - 1.8.0
- Keras - 2.1.6
- Scikit-learn - 0.19.1
- Matplotlib - 2.2.2
- Python - 3.6.5
- NumPy - 1.14.3

Η βάση δεδομένων

Αρχικά, σκεφθήκαμε να χρησιμοποιήσουμε το σύνολο δεδομένων της κλασικής MNIST που περιέχει χειρόγραφα ψηφία αλλά τελικά χρησιμοποιήσαμε την Fashion MNIST που περιέχει ρουχισμό. Οι λόγοι που μας οδήγησαν στο να χρησιμοποιήσουμε τελικά την Fashion MNIST είναι οι εξής:

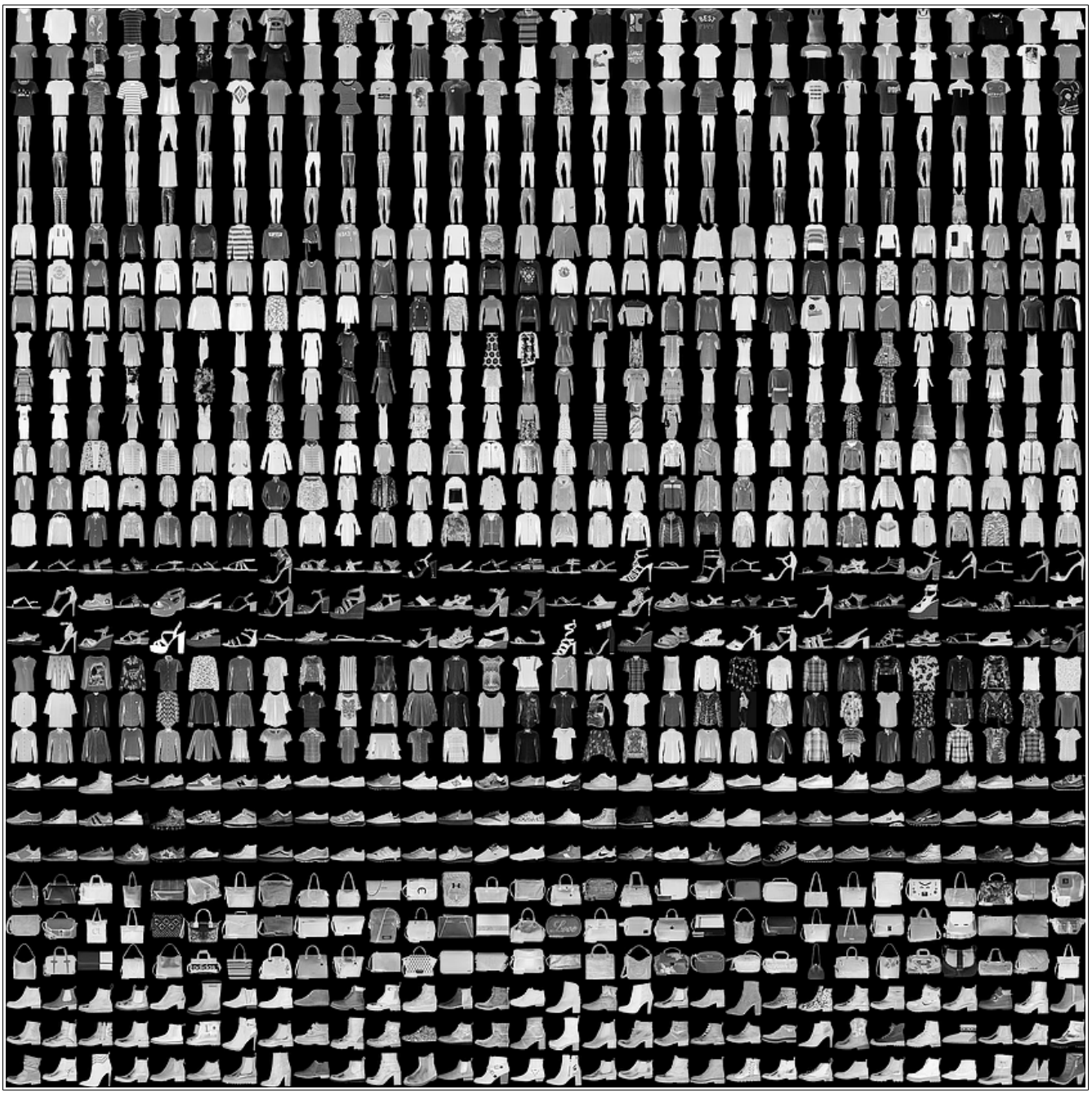
- Η κλασική MNIST είναι πολύ απλή και πολλά μοντέλα πλέον την μαθαίνουν αρκετά καλά
- Η κλασική MNIST χρησιμοποιείται συχνά ενώ υπάρχουν πλέον και άλλα σύνολα δεδομένων
- Η Fashion MNIST ενδείκνυται περισσότερο για προβλήματα τεχνητής όρασης

Το σύνολο δεδομένων Fashion MNIST έχει μοντελοποιηθεί χρησιμοποιώντας εικόνες από τα διάφορα άρθρα της ιστοσελίδας Zalando. Το σύνολο δεδομένων περιέχει 70000 εικόνες απόχρωσης του γκρι και διαστάσεων 28x28 εικονοστοιχείων. Από όλες αυτές τις εικόνες οι 60000 χρησιμοποιούνται για εκπαίδευση και οι 10000 για έλεγχο. Κάθε μία από αυτές τις εικόνες ανήκει αποκλειστικά σε μία και μόνο κατηγορία από ένα σύνολο 10 κατηγοριών. Στο σύνολο εικόνων εκπαίδευσης έχουμε για κάθε κατηγορία 6000 εικόνες, ενώ στο σύνολο εικόνων ελέγχου έχουμε για κάθε κατηγορία 1000 εικόνες.

Οι 10 κατηγορίες είναι οι εξής:

T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle boot

Παρακάτω ακολουθούν ενδεικτικές εικόνες για κάθε μία από τις 10 κατηγορίες:



Ενδεικτικές εικόνες ρουχισμού για τις 10 κατηγορίες (αλλαγή κατηγορίας ανά 3 γραμμές)

Εξαγωγή χαρακτηριστικών

Το διάνυσμα χαρακτηριστικών μιας εικόνας δεν προκύπτει με μεθόδους εξαγωγής χειρωνακτικά κατασκευασμένων χαρακτηριστικών αλλά προσπαθούμε να το εξάγουμε από την αναπαράσταση της στα κρυφά και εσωτερικά επίπεδα ενός συνελικτικού νευρωνικού δικτύου που εκπαιδεύεται για κατηγοριοποίηση πολλών κλάσεων.

Κατά την εκπαίδευση του συνελικτικού νευρωνικού δικτύου προσπαθούμε με αλγορίθμους βαθιάς μάθησης να μάθουμε τις εικόνες εκπαίδευσης με τέτοιο τρόπο ώστε το μοντέλο να μάθει να γενικεύει, να αποφεύγει την υπερ-εκπαίδευση και να μπορεί να προβλέπει με επιτυχία άγνωστα δείγματα. Με λίγα λόγια, προσπαθούμε να επιτύχουμε μία ικανοποιητική ακρίβεια κατηγοριοποίησης στις εικόνες ελέγχου.

Κατά την διάρκεια της εκπαίδευσης ρυθμίζονται κατάλληλα οι παράμετροι των κρυφών συνελικτικών επιπέδων (και όχι μόνο) του νευρωνικού δικτύου (πλαστικότητα νευρωνικών δικτύων) ώστε να κάνουν καλή εξαγωγή χαρακτηριστικών για την βελτιστοποίηση της ακρίβειας κατηγοριοποίησης στις εικόνες ελέγχου.

Έχοντας εκπαιδεύσει το μοντέλο μπορούμε να χρησιμοποιήσουμε την έξοδο αυτών των ενδιάμεσων κρυφών επιπέδων για την περιγραφή του διανύσματος χαρακτηριστικών. Ακολουθεί μια περιγραφή του συνελικτικού νευρωνικού δικτύου:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 28, 28, 64)        320
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 14, 14, 64)        0
_________________________________________________________________
dropout_1 (Dropout)          (None, 14, 14, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 14, 14, 32)        8224
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 7, 7, 32)          0
_________________________________________________________________
dropout_2 (Dropout)          (None, 7, 7, 32)          0
_________________________________________________________________
flatten_1 (Flatten)          (None, 1568)              0
_________________________________________________________________
dense_1 (Dense)              (None, 256)               401664
_________________________________________________________________
dropout_3 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 10)                2570
=================================================================
Total params: 412,778
Trainable params: 412,778
Non-trainable params: 0
```

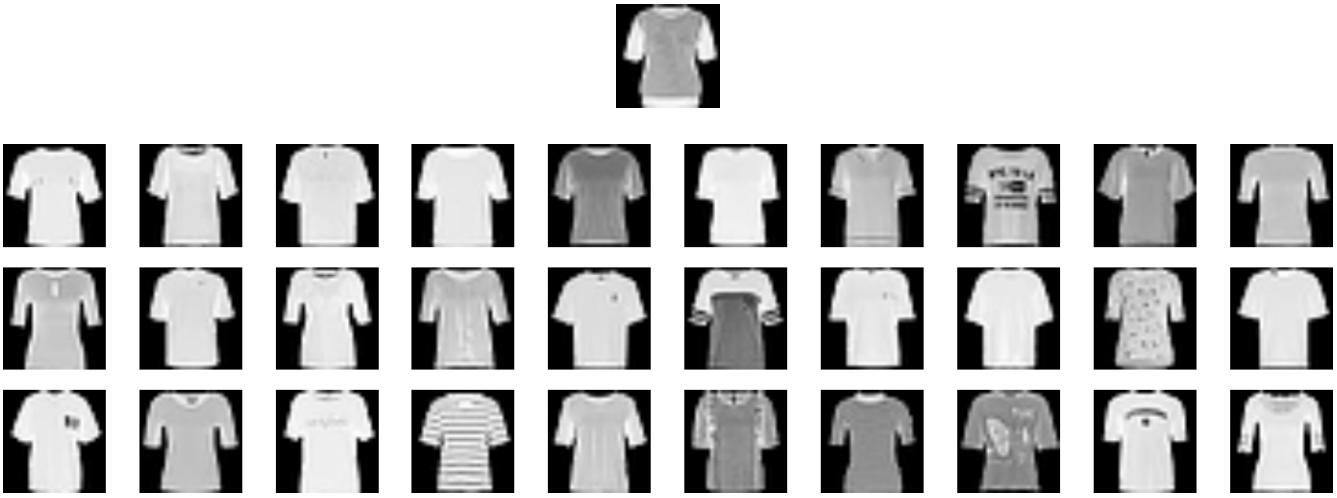
Περιγραφή του συνελικτικού νευρωνικού δικτύου

Από τα διάφορα κρυφά επίπεδα χρησιμοποιήθηκαν δύο επίπεδα (flatten_1 και dense_1) για την εξαγωγή του διανύσματος χαρακτηριστικών και μετρήθηκαν οι επιδόσεις του συστήματος με την χρήση καμπυλών ακρίβειας και ανάκλησης. Από τα αποτελέσματα προέκυψε πως το επίπεδο dense_1 ήταν καλύτερο από το επίπεδο flatten_1. Το διάνυσμα χαρακτηριστικών στο επίπεδο flatten_1 έχει διαστάσεις 1x1568 ενώ στο επίπεδο dense_1 έχει διαστάσεις 1x256.

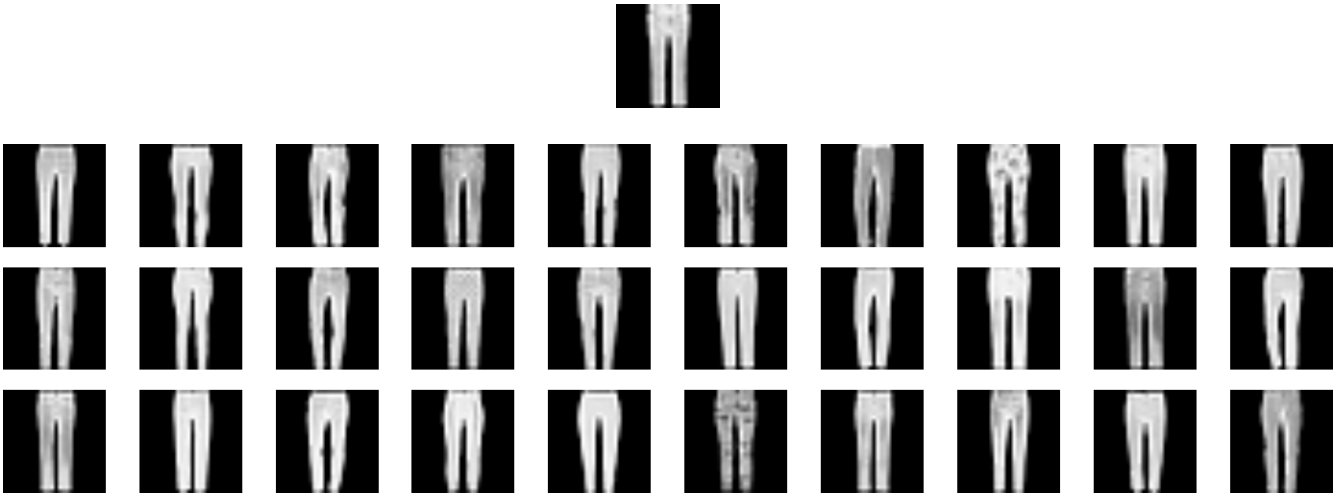
Ποιοτικά αποτελέσματα

Πριν την παρουσίαση των καμπυλών ακρίβειας και ανάκλησης κρίνεται σκόπιμο να παρουσιαστούν ποιοτικά κάποια ενδεικτικά αποτελέσματα από την αναζήτηση ορισμένων εικόνων κάθε κατηγορίας. Η πρώτη εικόνα που εμφανίζεται σε κάθε κατηγορία αποτελεί την εικόνα ερωτήματος, ενώ οι υπόλοιπες τα αποτελέσματα (30 κοντινότεροι γείτονες):

T-shirt/Top

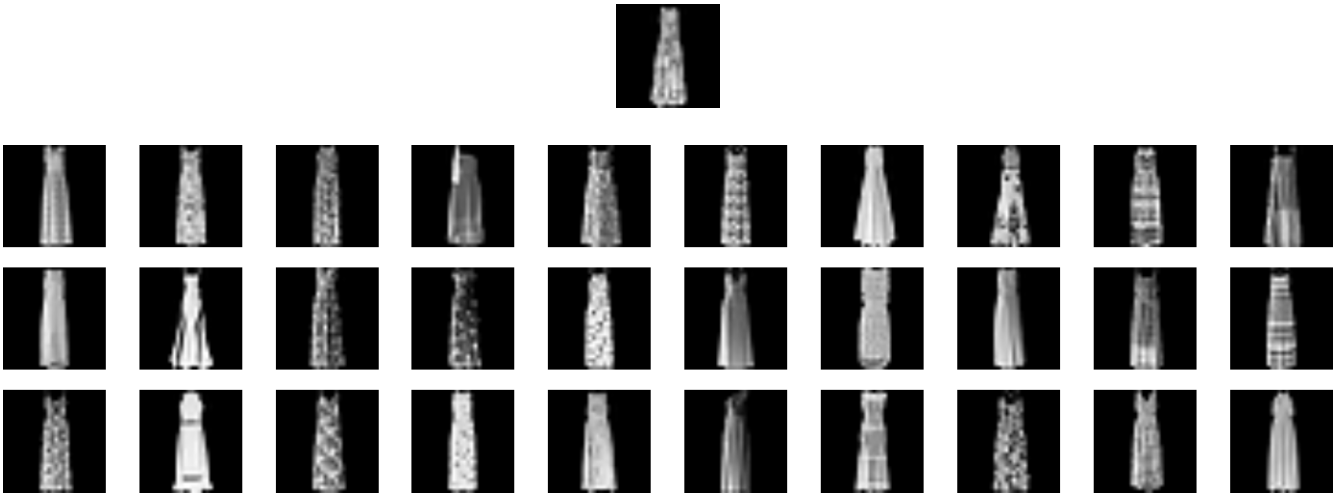


Trouser

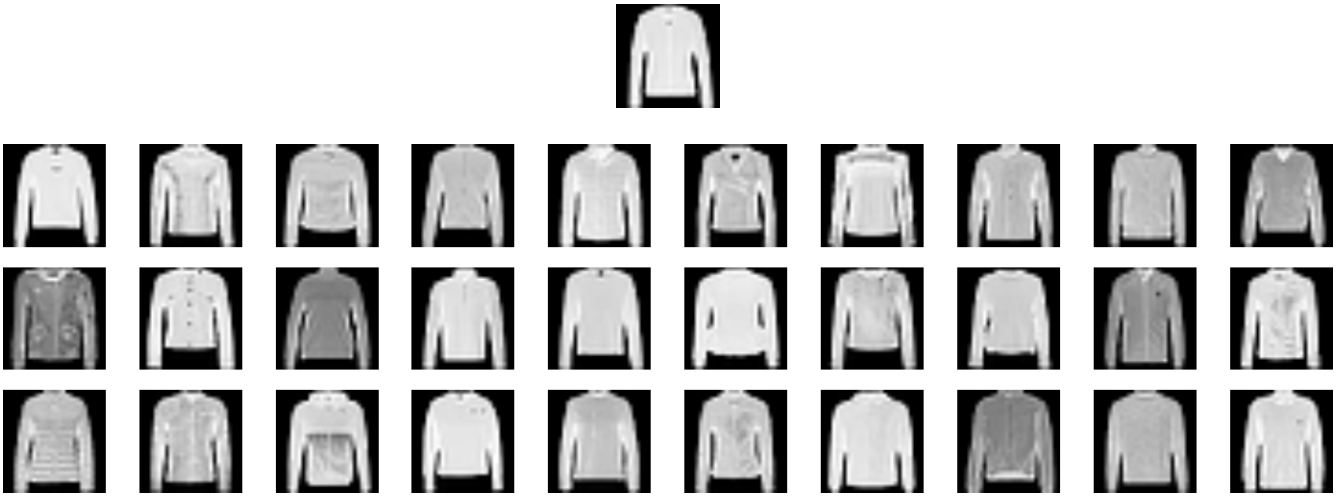


Pullover

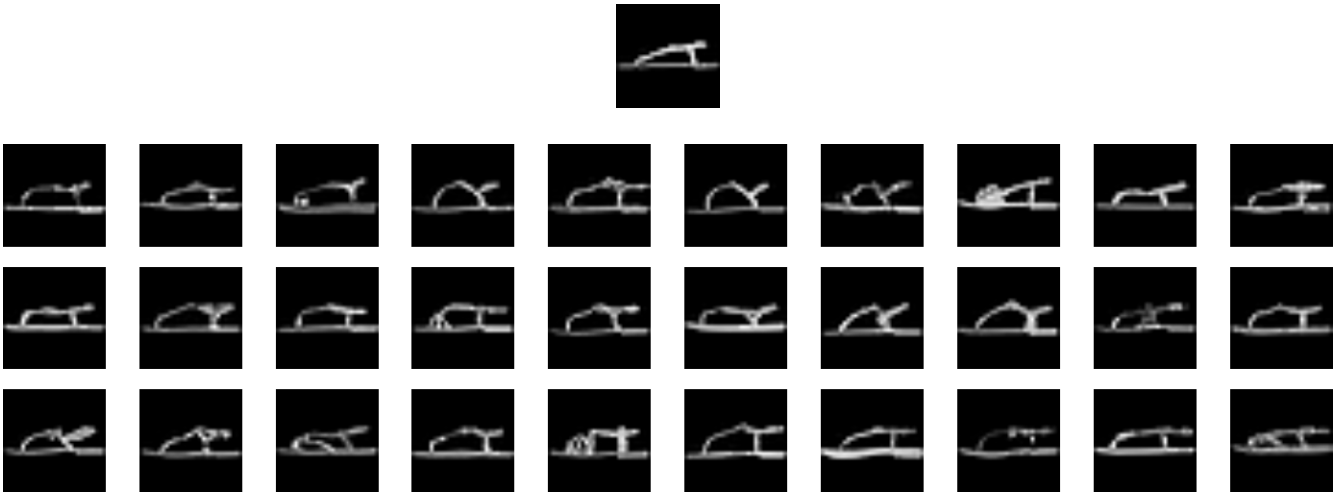
Dress



Coat

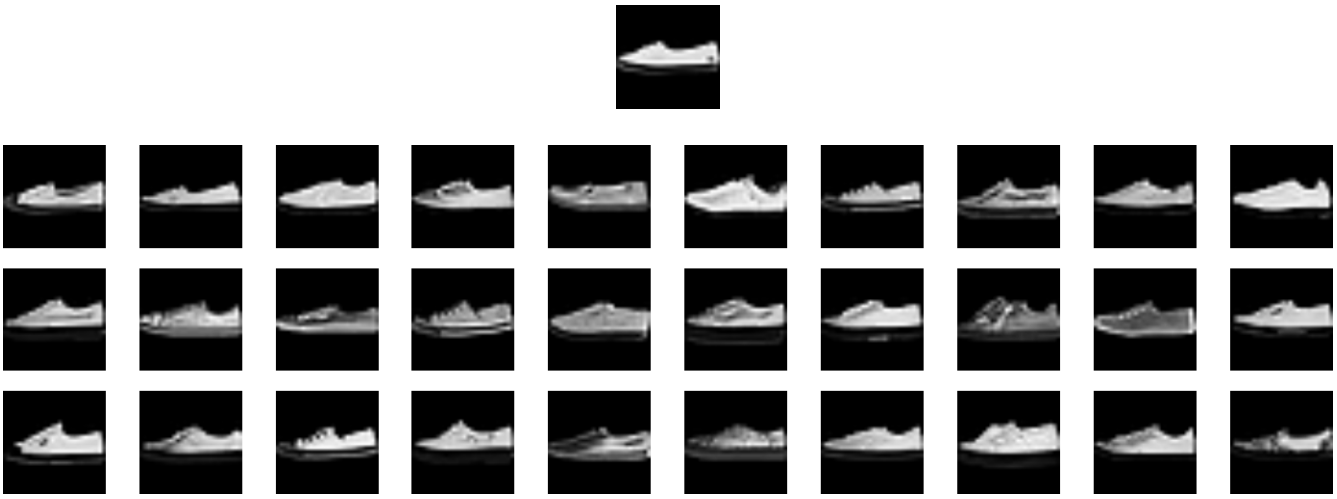


Sandal



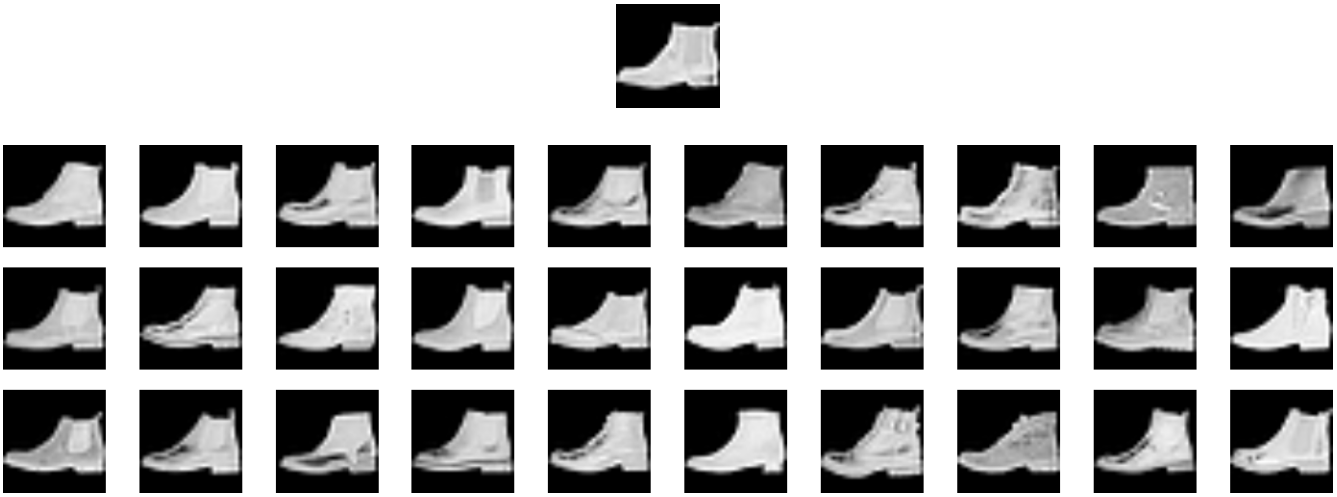
Shirt

Sneaker



Bag

Ankle Boot



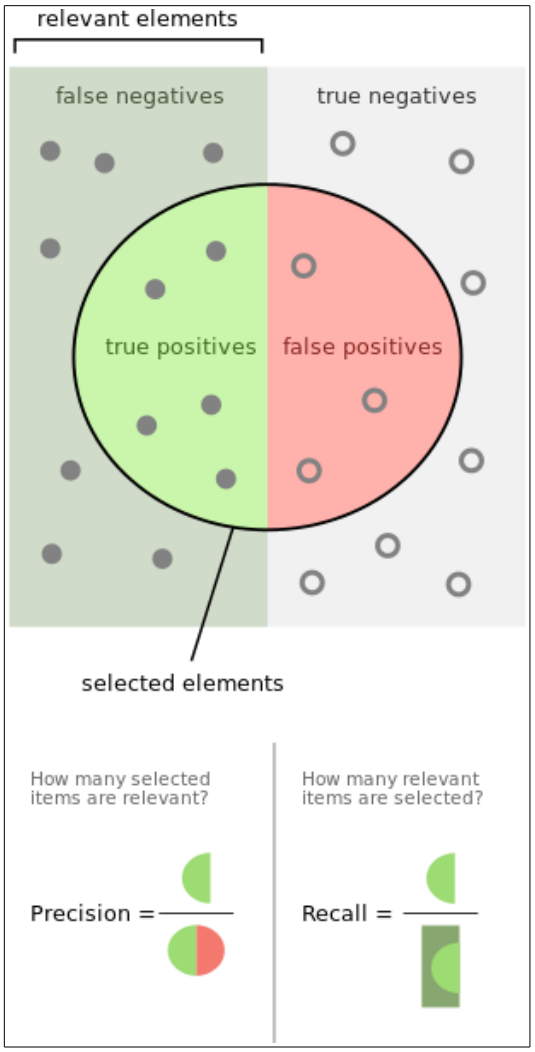
Μετρικές ακρίβειας και ανάκλησης

Η ακρίβεια και η ανάκληση μας δίνουν μια αίσθηση του πόσο σχετικά είναι τα αποτελέσματα που επιστρέφονται. Οι τιμές ακρίβειας και ανάκλησης παίρνουν πραγματικές τιμές από 0 έως 1 και ορίζονται ως εξής:

$$ακρίβεια = \frac{αριθμός\ σωστών\ αποτελεσμάτων\ του\ ερωτήματος}{αριθμός\ αποτελεσμάτων\ του\ ερωτήματος}$$

$$ανάκληση = \frac{αριθμός\ σωστών\ αποτελεσμάτων\ του\ ερωτήματος}{αριθμός\ σχετικών\ με\ το\ ερώτημα\ εικόνων\ στην\ βάση}$$

Ακολουθεί μια ακόμη βοηθητική εικόνα για την κατανόηση της ακρίβειας και της ανάκλησης:



Συμπεράσματα

Από τα αποτελέσματα που εξάγαμε σχετικά με τις μετρικές ακρίβειας και ανάκλησης διαπιστώνουμε πως όσο αυξάνουμε τον αριθμό των K κοντινότερων γειτόνων τόσο χάνουμε σε ακρίβεια από τα αποτελέσματα μας εφόσον αρχίζουν να επιστρέφονται όλο και περισσότερο εικόνες που δεν ανήκουν στην κατηγορία της εικόνας του ερωτήματος.

Επιπλέον, θα παρατηρήσουμε πως αυξάνει και η τιμή της ανάκλησης, κάτι που είναι απολύτως λογικό καθώς δίνουμε περισσότερες ευκαιρίες σε παρατηρήσεις που κατατάχθηκαν πιο μακριά να «μπουν» μέσα στα αποτελέσματα.

Επίσης, θα διαπιστώσουμε πως το κρυφό επίπεδο dense_1 του συνελικτικού νευρωνικού δικτύου παράγει διανύσματα χαρακτηριστικών που πετυχαίνουν καλύτερη ακρίβεια και ανάκληση σε σχέση με αυτά του κρυφού επιπέδου flatten_1.

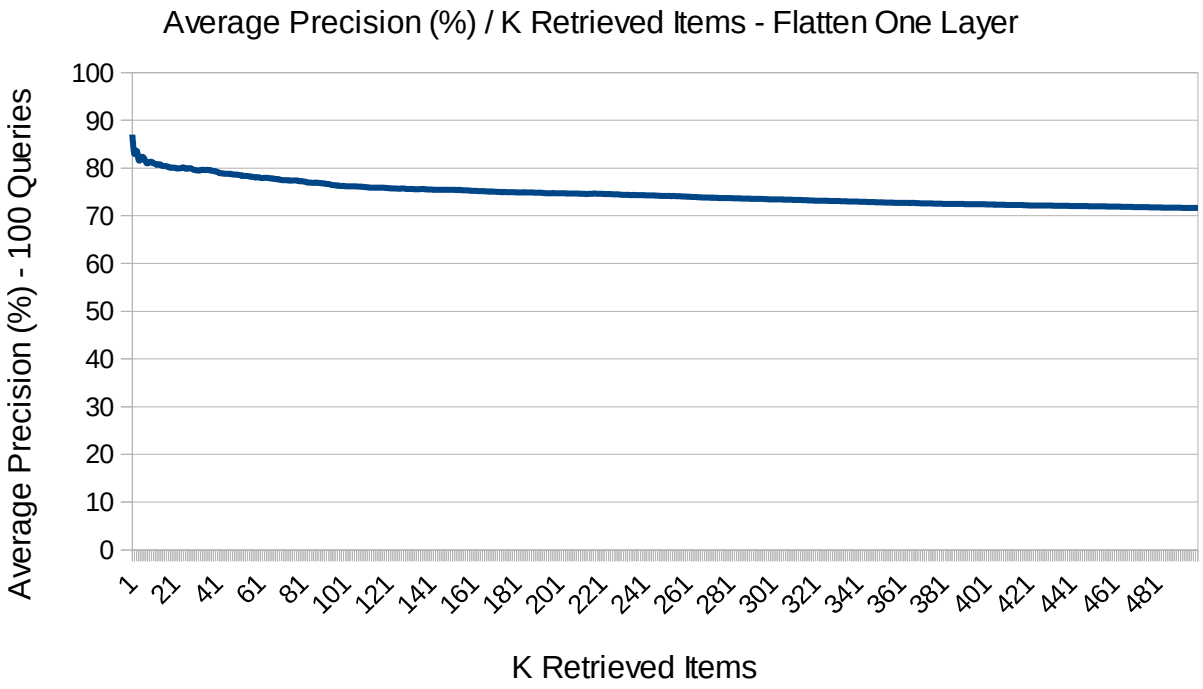
Να αναφέρουμε σε αυτό το σημείο ότι οι τιμές ακρίβειας και ανάκλησης που υπολογίστηκαν είναι ο μέσος όρος που προήλθε από 100 αναζητήσεις με τυχαίες εικόνες από το σύνολο εικόνων ελέγχου και με K κοντινότερους γείτονες ξεκινώντας από το 1 έως το 500.

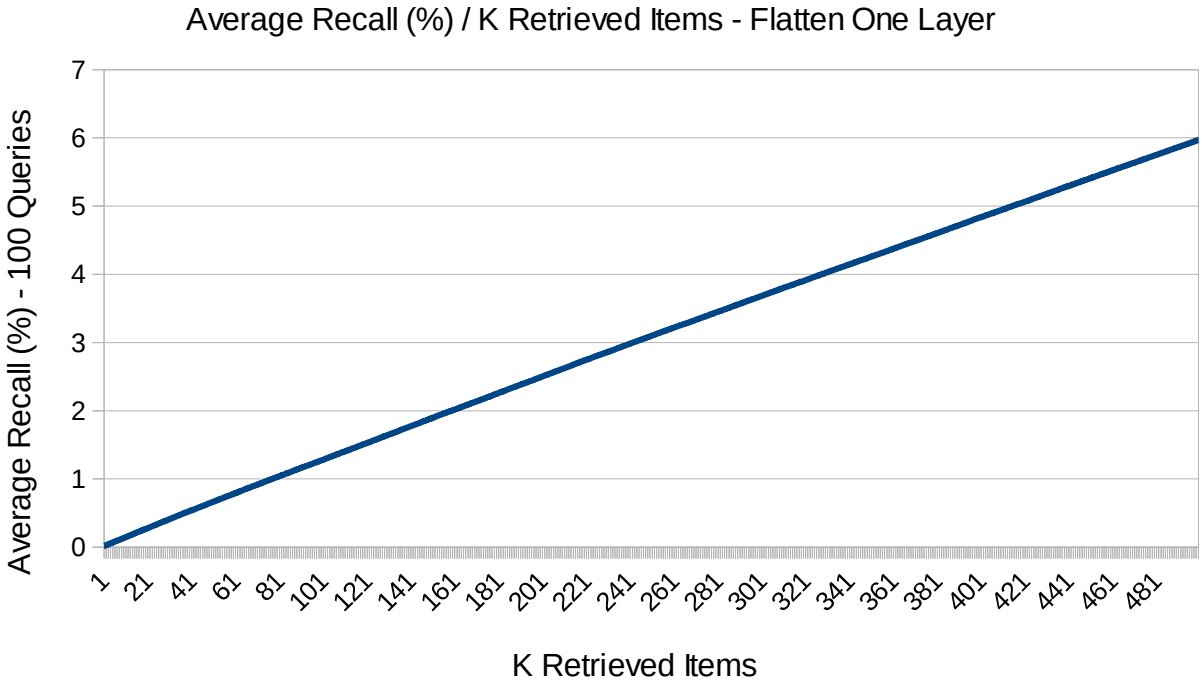
Το «trade-off» ανάμεσα στην ακρίβεια και στην ανάκληση βρίσκεται στο γεγονός πως όσο ζητάμε περισσότερες εικόνες από τη βάση, τόσο το σύστημα αναγκάζεται να μας φέρει πιο «μακρινές» εικόνες και έτσι αναγκαστικά μειώνεται η ακρίβεια.

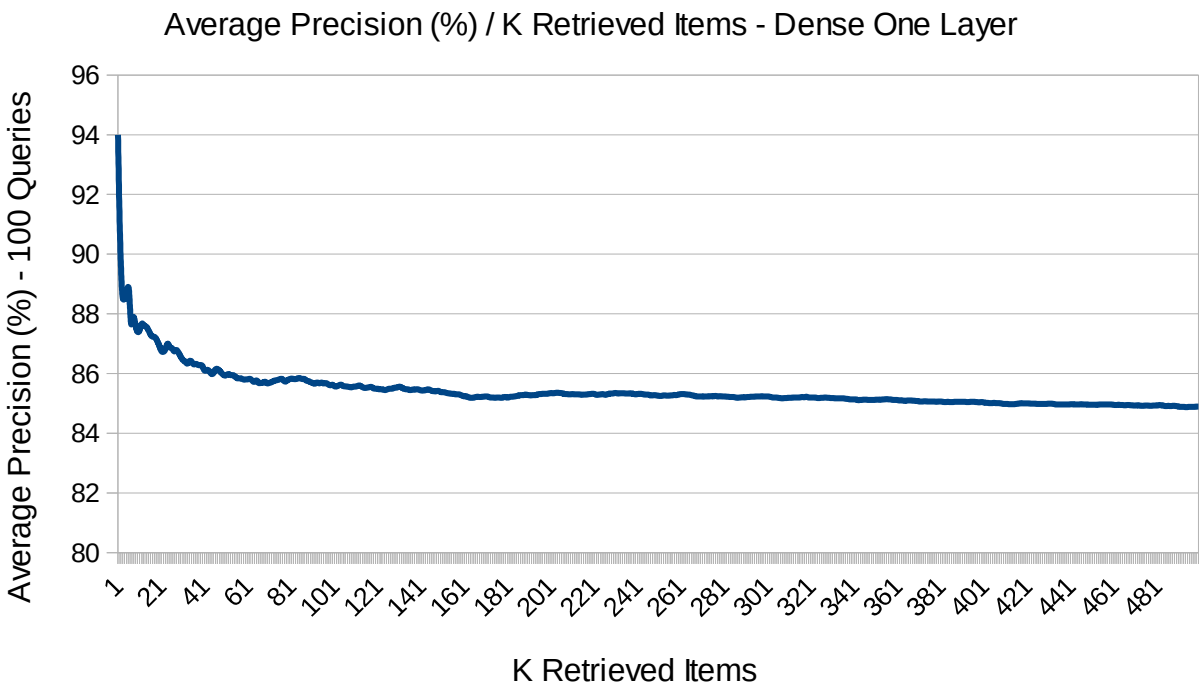
Η τιμή της μέσης ακρίβειας δεν μπορεί να ξεκινάει με 100% ακόμα και για K = 1 γιατί τα διανύσματα χαρακτηριστικών που χρησιμοποιήθηκαν προέρχονται από ένα νευρωνικό μοντέλο που κάνει και λάθη πρόβλεψης (ακρίβεια 91.16% στις εικόνες ελέγχου). Έτσι, η τιμή της ακρίβειας στην περίπτωση μας για K = 1 ξεκινάει με 94% και φθάνει στο 84.44% για K = 500. Όσο πιο καλή ακρίβεια νευρωνικού μοντέλου επιτυγχάνουμε τόσο καλύτερη ακρίβεια θα πετυχαίνουμε και στις αναζητήσεις μας γιατί θα έχουμε καλύτερα διανύσματα χαρακτηριστικών.

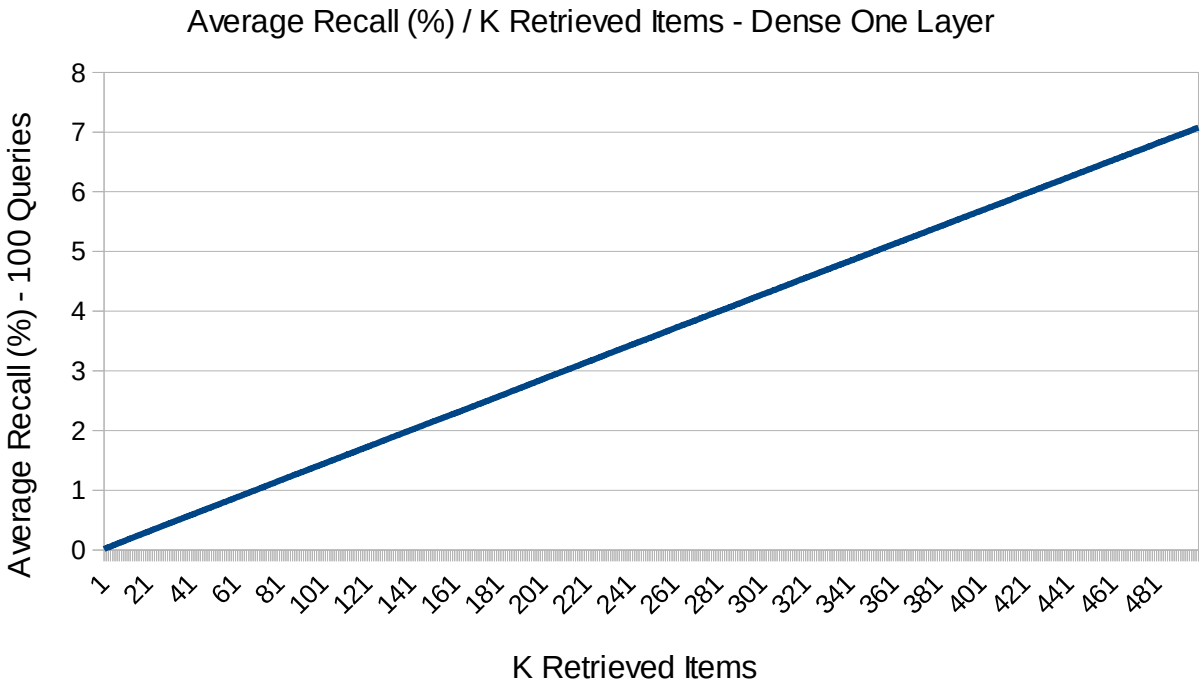
Τέλος, η μέση τιμή της ανάκλησης έστω και για K = 500 δεν ξεπερνάει το 7.07% και αυτό γιατί ο αριθμός των σχετικών εικόνων που επιστρέφονται είναι πολύ μικρότερος από τον αριθμό των συνολικών σχετικών εικόνων που υπάρχουν στην κατηγορία της εικόνας του ερωτήματος.

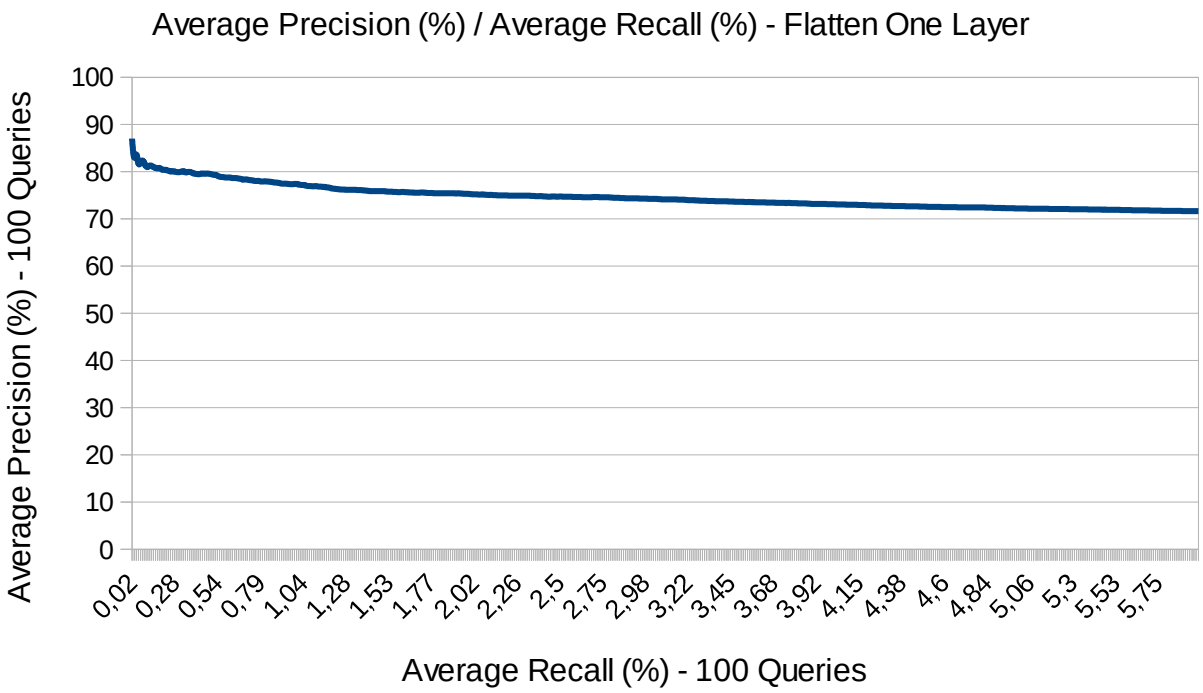
Γραφικές παραστάσεις ακρίβειας και ανάκλησης

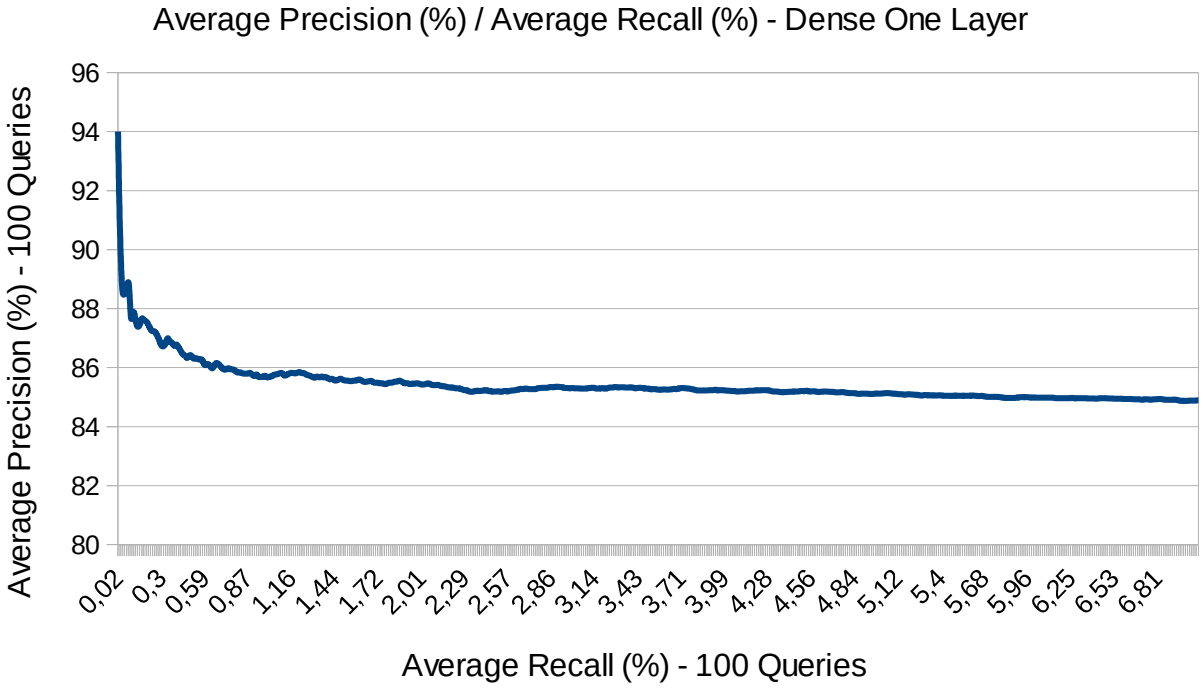





Average Precision (%) / K Retrieved Items - Dense One Layer
Average Precision (%) - 100 Queries
96
94
92
90
88
86
84
82
80
1
21
41
61
81
101
121
141
161
181
201
221
241
261
281
301
321
341
361
381
401
421
441
461
481
K Retrieved Items


Average Recall (%) / K Retrieved Items - Dense One Layer
Average Recall (%) - 100 Queries
8
7
6
5
4
3
2
1
0
1
21
41
61
81
101
121
141
161
181
201
221
241
261
281
301
321
341
361
381
401
421
441
461
481
K Retrieved Items


Average Precision (%) / Average Recall (%) - Flatten One Layer
Average Precision (%) - 100 Queries
100
90
80
70
60
50
40
30
20
10
0
0,02
0,28
0,54
0,79
1,04
1,28
1,53
1,77
2,02
2,26
2,5
2,75
2,98
3,22
3,45
3,68
3,92
4,15
4,38
4,6
4,84
5,06
5,3
5,53
5,75
Average Recall (%) - 100 Queries


Average Precision (%) / Average Recall (%) - Dense One Layer
Average Precision (%) - 100 Queries
96
94
92
90
88
86
84
82
80
0,02
0,3
0,59
0,87
1,16
1,44
1,72
2,01
2,29
2,57
2,86
3,14
3,43
3,71
3,99
4,28
4,56
4,84
5,12
5,4
5,68
5,96
6,25
6,53
6,81
Average Recall (%) - 100 Queries