



Παραγωγική Τεχνητή Νοημοσύνη: Generative AI

Κωνσταντίνος Καραμανής

The University of Texas at Austin & Archimedes/Athena RC

constantine@utexas.edu

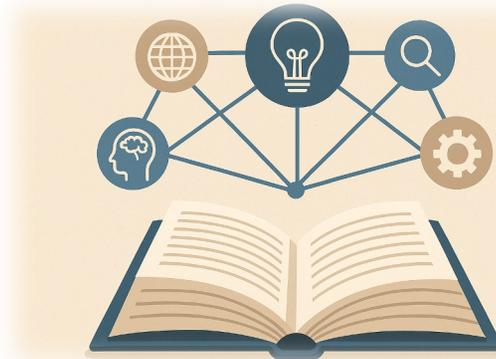
<https://caramanis.github.io/>





Ας θυμηθούμε τα
προηγούμενα...

Search (αναζήτηση) σε βάση γνώσεων



Επιστημονικά περιοδικά που περιέχουν χιλιάδες δημοσιευμένα άρθρα για παθήσεις του ανοσοποιητικού συστήματος

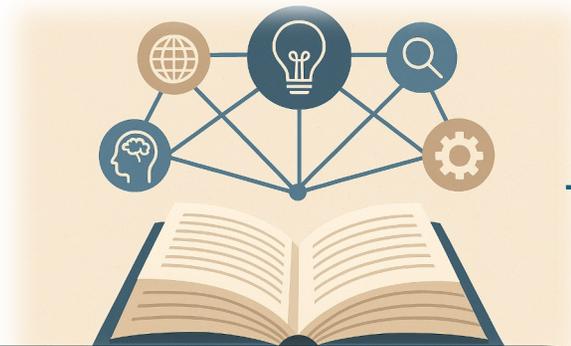
Ερώτημα χρήστη (query):

«Ποιες είναι οι χρήσεις και πιθανές παρενέργειες του IL-23 inhibitor biologic Skyrizi;»

Η «σωστή» απάντηση;

Recall@K NDCG@K

Search (αναζήτηση) σε βάση γνώσεων



Επιστημονικά περιοδικά που περιέχουν χιλιάδες δημοσιευμένα άρθρα για παθήσεις του ανοσοποιητικού συστήματος

Πώς εκτιμούμε ποια είναι τα σχετικά άρθρα στη βάση γνώσεων;

Ποιες είναι οι χρήσεις και πιθανές παρενέργειες του IL-23 inhibitor biologic Skyrizi;»

Η «σωστή» απάντηση;

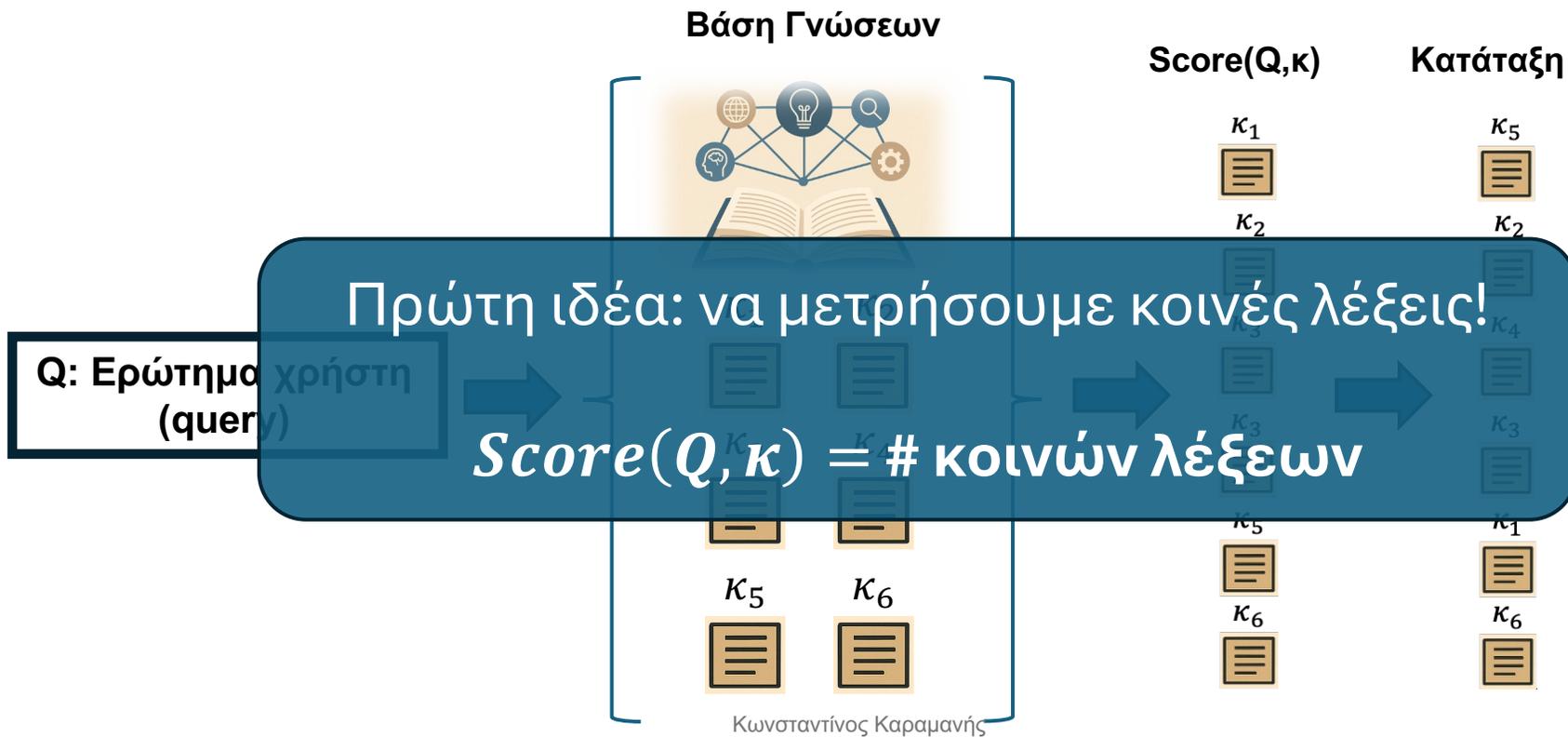
Recall@K NDCG@K

1. Bag of Words
2. TF-IDF/BM25

3. Semantic Search



Αναζήτηση σε Βάση Γνώσεων



Αναζήτηση σε Βάση Γνώσεων

Βάση Γνώσεων



Score(Q,κ)

Κατάταξη

κ₁

κ₅

κ₂

κ₂

Πρώτη ιδέα: να μετρήσουμε κοινές λέξεις!

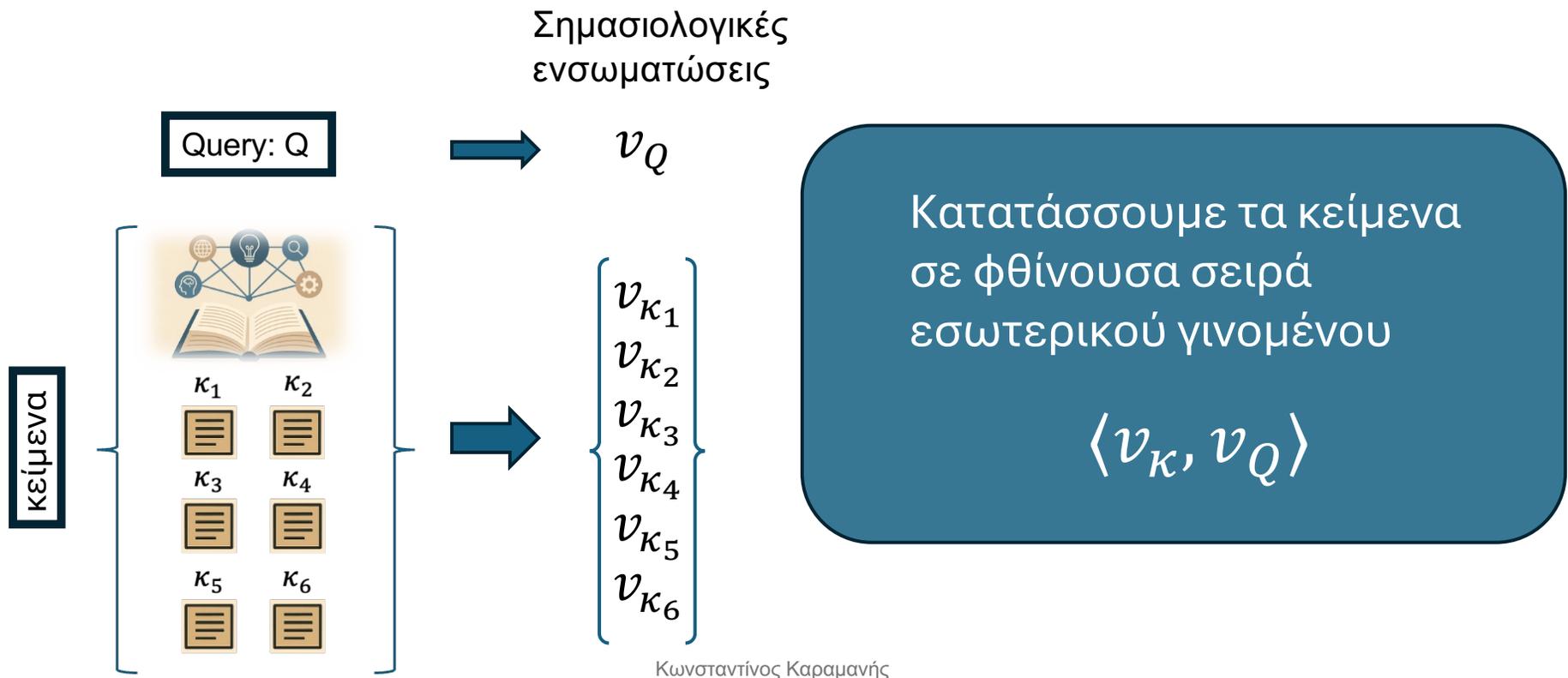
Q: Ερώτημα χρήστη
(query)

Δεύτερη ιδέα: χρησιμοποιούμε
σημασιολογικές ενσωματώσεις!

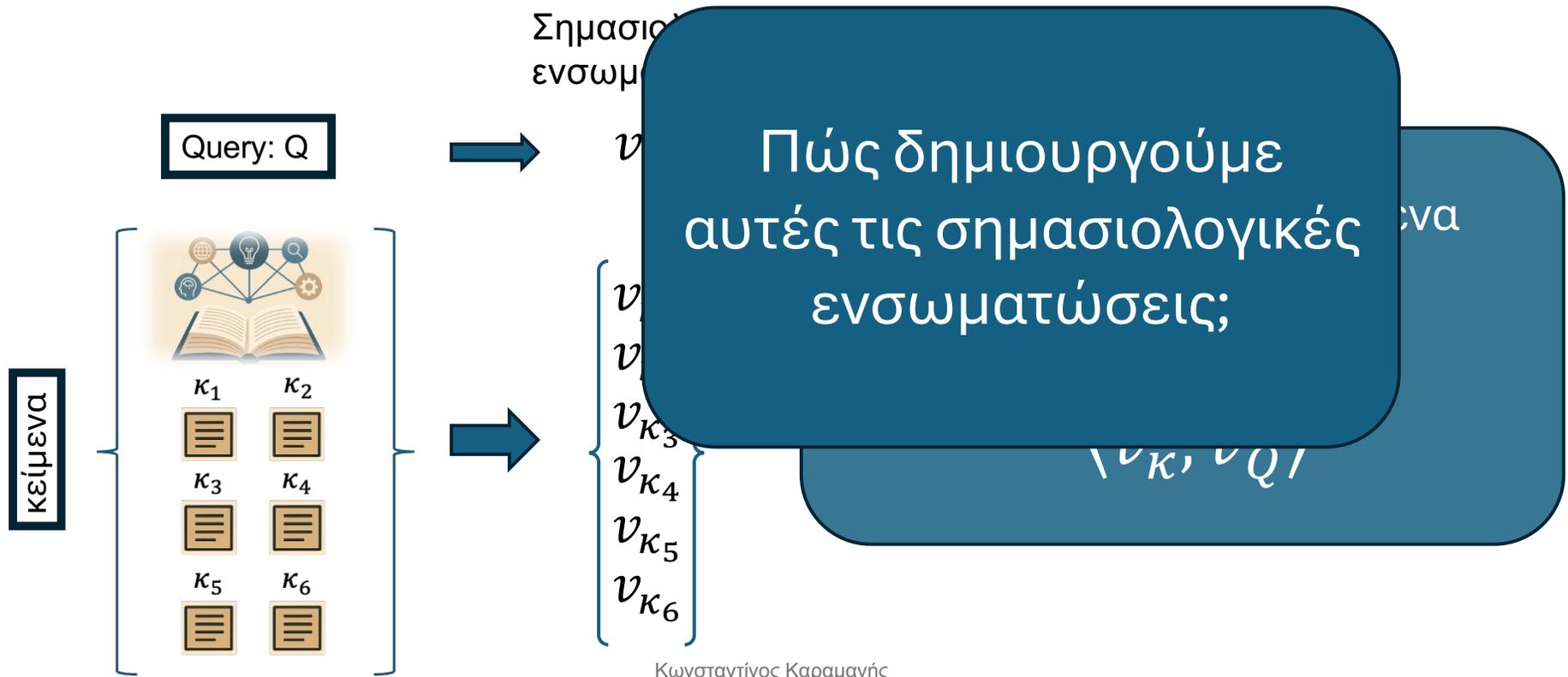
$$Score(Q, \kappa) = \langle v_Q, v_\kappa \rangle$$

Κωνσταντίνος Καραμανής

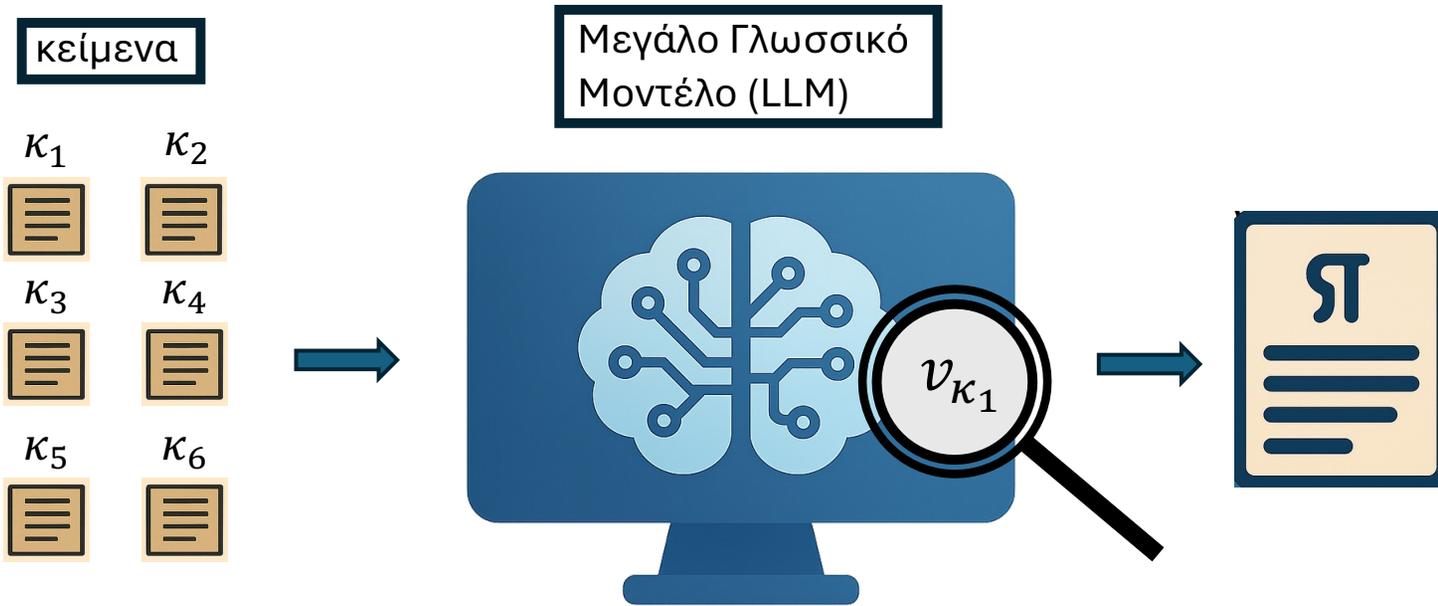
Αναζήτηση σε Βάση Γνώσεων



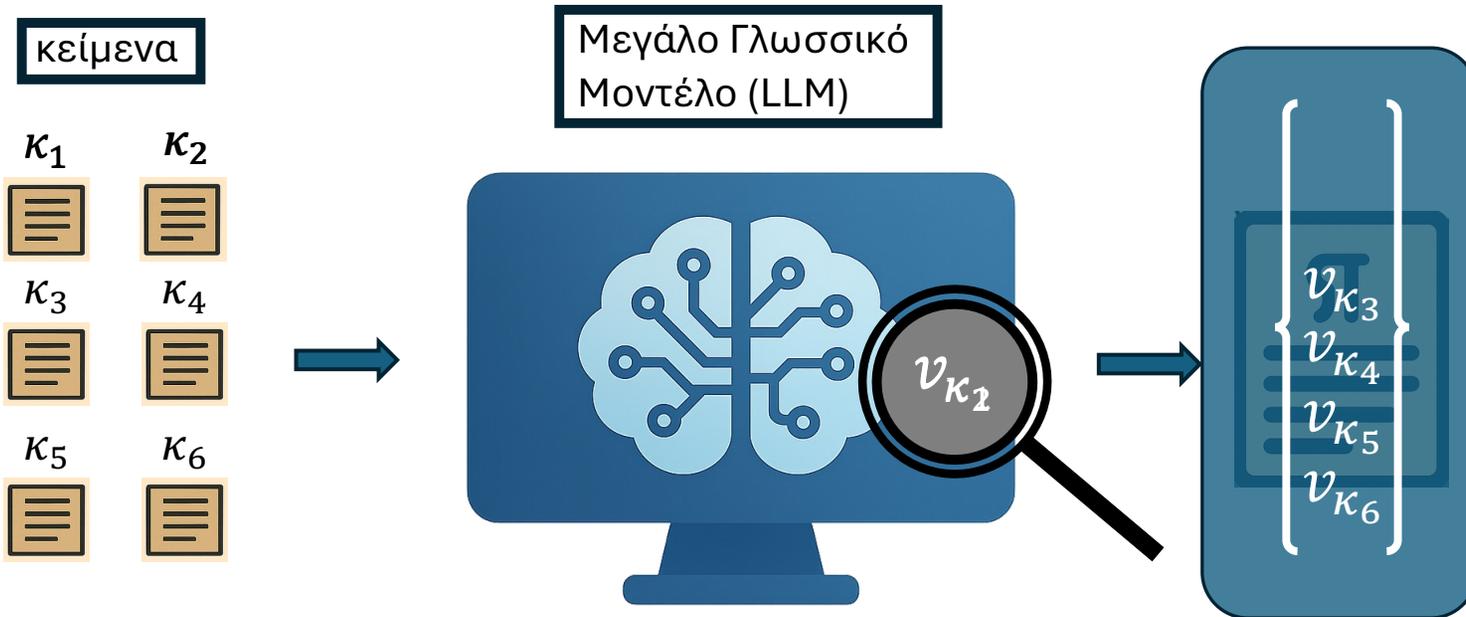
Αναζήτηση σε Βάση Γνώσεων



Σημασιολογικές Ενσωματώσεις Κειμένου



Σημασιολογικές Ενσωματώσεις Κειμένου



Σημασιολογικές Ενσωματώσεις Κειμένου

1. Μετατρέπουμε την Βάση Γνώσης μας σε συλλογή μικρότερων κειμένων, 300-500 λέξεων.

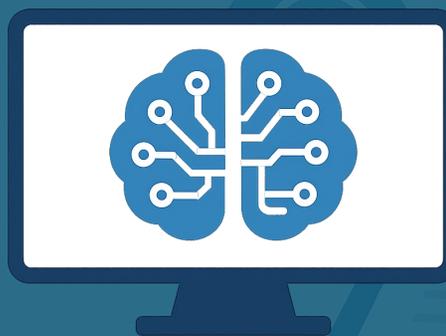
Π : {πλήρες κείμενο διαλέξεων} \Rightarrow {κομμάτια κειμένου}

2. Περνάμε τα κομμάτια $\{\kappa_i\}$ από ένα LLM: υπολογίζουμε ενσωματώσεις $\{v_i\}$

3. Με “query” Q , υπολογίζουμε την ενσωμάτωση v_Q και κατατάσσουμε τα κομμάτια $\{\kappa_i\}$ σε φθίνουσα σειρά: $\langle v_Q, v_i \rangle$

RAG: Retrieval Augmented Generation

Query: «Πώς λειτουργούν
τα συνελκτικά δίκτυα;»



Τα συνελκτικά νευρωνικά δίκτυα (CNNs) είναι ένας τύπος νευρωνικών δικτύων ειδικά σχεδιασμένα για την επεξεργασία εικόνων και χωρικών δεδομένων, εμπνευσμένα από το οπτικό σύστημα του εγκεφάλου. [umite](#)

Βασική Αρχιτεκτονική

Ένα CNN αποτελείται από στρώματα συνέλιξης (convolutional layers), pooling και πλήρως συνδεδεμένα (fully connected) στρώματα. Τα στρώματα συνέλιξης εφαρμόζουν φίλτρα (kernels) που γλιστρούν πάνω στην είσοδο (π.χ. εικόνα) και υπολογίζουν το εσωτερικό γινόμενο για να εξαγάγουν τοπικά χαρακτηριστικά όπως άκρες ή υφές. Τα pooling στρώματα μειώνουν τις διαστάσεις, διατηρώντας τα πιο σημαντικά χαρακτηριστικά. [strouthopoulos_lhu_v1](#)

Διαδικασία Συνέλιξης

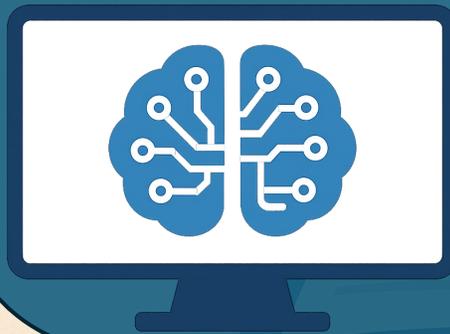
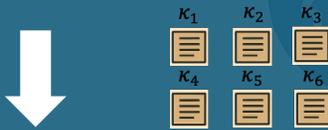
Η συνέλιξη μετατρέπει pixels σε χάρτες χαρακτηριστικών (feature maps): ένα φίλτρο 3x3, π.χ., πολλαπλασιάζει τις τιμές του με τμήμα της εικόνας και προσθέτει τα αποτελέσματα, παράγοντας μια νέα τιμή. Αυτό επαναλαμβάνεται με stride (βήμα μετακίνησης), ενώ πολλαπλά φίλτρα δημιουργούν κανάλια για διαφορετικά μοτίβα. Στα βαθιά στρώματα, απλά χαρακτηριστικά (γραμμές) ενώνονται σε σύνθετα (πρόσωπα). [umite_v1](#)

Εκπαίδευση και Λειτουργία

Κατά την εκπαίδευση, το δίκτυο προσαρμόζει τα βάρη των φίλτρων μέσω backpropagation και optimizer (π.χ. Adam), ελαχιστοποιώντας σφάλμα ταξινόμησης. Οι χάρτες από τα συνελκτικά στρώματα "flatten" και εισέρχονται σε fully connected layers για τελική απόφαση. Παράδειγμα: Σε αναγνώριση εικόνων, τα πρώτα στρώματα ανιχνεύουν άκρες, τα επόμενα σχήματα. [e1cca_v2](#)

RAG: Retrieval Augmented Generation

Query: «Πώς λειτουργούν
τα συνελκτικά δίκτυα;»



Τα **συνελκτικά νευρωνικά δίκτυα (convolutional neural networks - CNN)** είναι μια ειδική κατηγορία δικτύων σχεδιασμένη κυρίως για τη **μηχανική όραση**, η οποία προσομοιώνει τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος αναγνωρίζει αντικείμενα εστιάζοντας σε συγκεκριμένα χαρακτηριστικά ¹.

Η λειτουργία τους βασίζεται στις εξής βασικές αρχές:

- **Τοπική Συνδεσιμότητα (Local Connectivity):** Σε αντίθεση με τα επίπεδα πλήρους σύνδεσης (fully connected), όπου κάθε νευρώνας "βλέπει" όλα τα δεδομένα εισόδου, στα συνελκτικά επίπεδα ο κάθε νευρώνας **υπολογίζεται χρησιμοποιώντας μόνο ένα συγκεκριμένο κομμάτι της εικόνας** ¹. Για παράδειγμα, ένας νευρώνας μπορεί να "βλέπει" μόνο τα pixels x_1, x_2, x_3 και όχι ολόκληρη την είσοδο ².

- **Κοινή χρήση Βαρών (Weight Sharing):** Όλοι οι νευρώνες ενός συνελκτικού επιπέδου χρησιμοποιούν τις **ίδιες παραμέτρους (βάρη)** ². Αυτό σημαίνει ότι το δίκτυο εφαρμόζει την ίδια "σκέψη" (π.χ. αναζήτηση για ένα μάτι ή μια γωνία) σε ολόκληρη την έκταση της εικόνας ^{2 3}.

- **Κανάλια (Channels):** Η πληροφορία επεξεργάζεται σε πολλαπλά κανάλια ³. Μια έγχρωμη εικόνα εισόδου έχει τρία κανάλια (RGB - Κόκκινο, Πράσινο, Μπλε), αλλά τα ενδιάμεσα επίπεδα μπορούν να έχουν δεκάδες ή εκατοντάδες κανάλια, όπου το καθένα ανιχνεύει διαφορετικά χαρακτηριστικά, όπως μάτια, μύτες ή αυτιά ^{3 4}.

- **Φίλτρα και Πυρήνες (Kernels):** Η διαδικασία της συνέλιξης ορίζεται από ένα "**κουτάκι**" (kernel) συγκεκριμένου μεγέθους (π.χ. 5×5 ή 7×7 pixels) που σαρώνει την εικόνα ^{4 5}. Οι παράμετροι που καθορίζουν τη σάρωση είναι το **stride** (πόσα pixels μετακινείται το κουτί κάθε φορά) και το **zero padding** (προσθήκη μηδενικών στις άκρες της εικόνας για τη διατήρηση του μεγέθους) ^{5 6}.

Εκτός από τα συνελκτικά επίπεδα, ένα δίκτυο CNN περιλαμβάνει συνήθως:

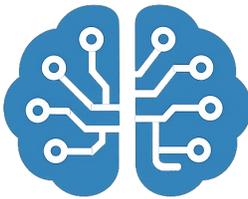
1. **Επίπεδα ReLU:** Μη γραμμικές συναρτήσεις που είναι απαραίτητες για να αποκτήσει το δίκτυο περιγραφική δύναμη και να μην ισοδυναμεί με ένα απλό γραμμικό μοντέλο ^{7 8}.

2. **Επίπεδα Pooling (Max/Average Pooling):** Χρησιμοποιούνται για τη **μείωση του αριθμού των παραμέτρων** και του μεγέθους των επιπέδων ⁶. Το **Max Pooling** κρατά τη μέγιστη τιμή από ένα "κουτάκι" pixels, ενώ το **Average Pooling** υπολογίζει τον μέσο όρο ^{9 10}.

3. **Επίπεδα Πλήρους Σύνδεσης (Fully Connected):** Συνήθως βρίσκονται στο τέλος του δικτύου, αφού η πληροφορία έχει μικρύνει σε μέγεθος, για να βγάλουν το τελικό αποτέλεσμα της ταξινόμησης ^{11 12}.

RAG: Retrieval Augmented Generation

Query: «Πώς λειτουργούν
τα συνελκτικά δίκτυα;»



Τα **συνελκτικά νευρωνικά δίκτυα (convolutional neural networks - CNN)** είναι μια ειδική κατηγορία δικτύων σχεδιασμένη κυρίως για τη **μηχανική όραση**, η οποία προσομοιώνει τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος αναγνωρίζει αντικείμενα εστιάζοντας σε συγκεκριμένα χαρακτηριστικά ¹.

Η λειτουργία τους βασίζεται στις εξής βασικές αρχές:

1. Επίπεδο Συνδεσιμότητας (Connectivity): Σε αντίθεση με τα επίπεδα πλήρους σύνδεσης (fully connected) όπου όλα τα δεδομένα εισόδου, στα συνελκτικά επίπεδα ο κάθε νευρώνας δίνει έμφαση σε **ένα συγκεκριμένο κομμάτι της εικόνας** ¹. Για παράδειγμα, ένας νευρώνας μπορεί να εστιάζει σε συγκεκριμένα pixels x_1, x_2, x_3 και όχι ολόκληρη την είσοδο ¹.

από δεξιά παρόλο που έχει παραπάνω παραμέτρους δεν υπερβαίνει την εκφραστική δύναμη δυνατότητα του απλού νευρωνικού δικτύου από αριστερά και άμα το σκεφτείτε λιγάκι θα δείτε ότι και παραπάνω νευρώνες να είχα αν είχα h1 h2 h3 h4 h5 και παραπάνω επίπεδα εάν είχα θα μπορούσαμε να κάνουμε ακριβώς τον ίδιο αλγεβρικό ϵ την τις αλγεβρικές πράξεις και θα βλέπαμε ότι και αυτό το δίκτυο παρόλο που είχε πιο πολλές παραμέτρους και πιο πολλά weights και biases και πιο πολλούς νευρώνες τελικά είναι ισοδύναμο με αυτό που έχουμε αριστερά το πολύ απλό πώς λοιπόν θα χτίσουμε επίπεδα πώς θα χτίσουμε πώς θα χρησιμοποιήσουμε πολλαπλά επίπεδα για να χτίσουμε κάτι πιο περίπλοκο μας λείπει ένα βασικό στοιχείο πρέπει να σώσουμε ένα μη γραμμικό επίπεδο γιατί αυτό που έχουμε δει εδώ είναι ότι αν βάζουμε πολλά **μη γραμμικά επίπεδα** αυτό ένα απλό γραμμικό μοντέλο και

από συνελκτικό επίπεδο χρησιμοποιούν εφαρμογή της "σκέψης" (π.χ. εικόνα ² ³) .
3 κανάλια ³ . Μια έγχρωμη εικόνα σε ενδιάμεσα επίπεδα μπορούν να έχουν διαφορετικά χαρακτηριστικά, όπως μάτια,

οριζείται από ένα "**κουτάκι**" (kernel) που κινείται πάνω στην εικόνα ⁴ ⁵ . Οι παράμετροι που ορίζονται στο κουτί κάθε φορά) και το **zero padding** (για να κρατήσουμε τον μέγεθος) ⁵ ⁶ .

να, ένα δίκτυο CNN με 3 κανάλια:

- 1. Επίπεδα ReLU:** Μη γραμμικές συναρτήσεις που είναι απαραίτητες για να αποκτήσει το δίκτυο περιγραφική δύναμη και να μην ισοδυναμεί με ένα απλό γραμμικό μοντέλο ⁷ ⁸.
- 2. Επίπεδα Pooling (Max/Average Pooling):** Χρησιμοποιούνται για τη **μείωση του αριθμού των παραμέτρων** και του μεγέθους των επιπέδων ⁶ . Το **Max Pooling** κρατά τη μέγιστη τιμή από ένα "κουτάκι" pixels, ενώ το **Average Pooling** υπολογίζει τον μέσο όρο ⁹ ¹⁰ .
- 3. Επίπεδα Πλήρους Σύνδεσης (Fully Connected):** Συνήθως βρίσκονται στο τέλος του δικτύου, αφού η πληροφορία έχει μικρύνει σε μέγεθος, για να βγάλουν το τελικό αποτέλεσμα της ταξινόμησης ¹¹ ¹² .

RAG: Retrieval Augmented Generation

Ο στόχος: Ένα σύστημα τεχνητής νοημοσύνης που απαντά σε ερωτήσεις του χρήστη με απαντήσεις βασισμένες σε μια δεδομένη βάση γνώσεων και, επιπλέον, τεκμηριωμένες με κείμενα από τη βάση γνώσεων.

Πλεονεκτήματα:

1. Παρέχει πρόσβαση σε ενημερωμένα ή εξειδικευμένα δεδομένα χωρίς επανεκπαίδευση του μοντέλου.
2. Βελτιώνει την ακρίβεια των απαντήσεων και περιορίζει τις «παραισθήσεις» (hallucinations) των μεγάλων γλωσσικών μοντέλων (LLMs)

1. Bag of Words

2. TF-IDF/BM25

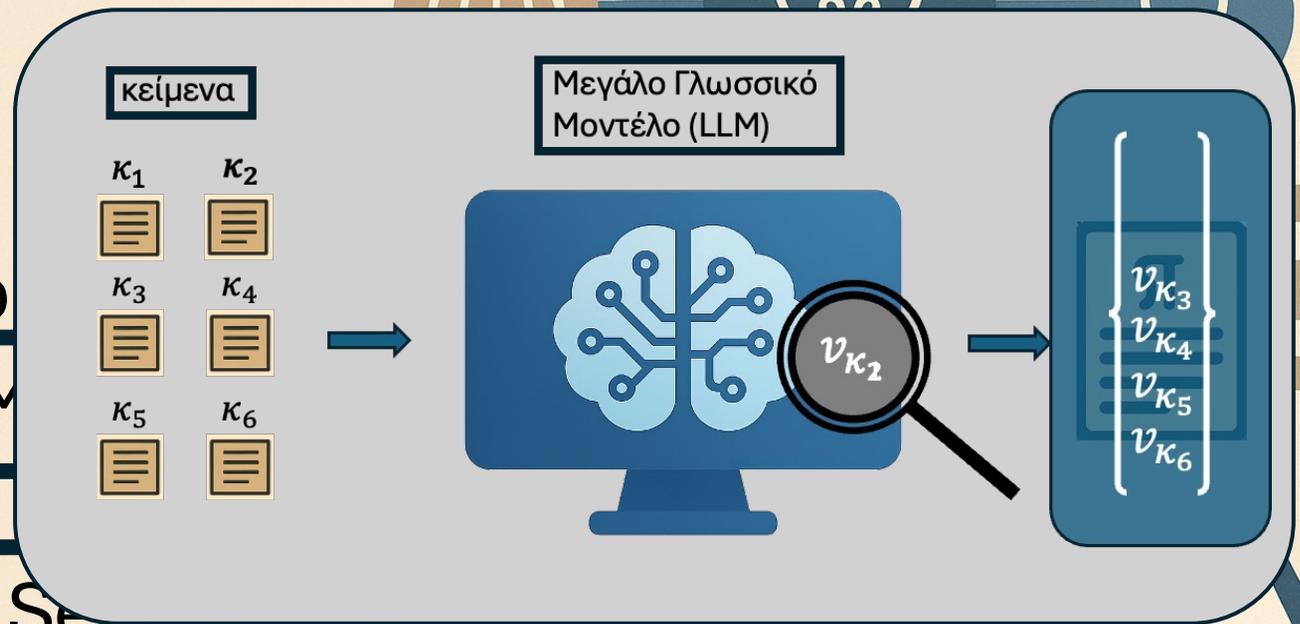
3. Semantic Search



1. Bag of Words

2. TF-IDF/BM

3. Semantic Search



1. Bag of Wo

2. TF-IDF/BM

3. Seman

