ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

A coin with a head on it

Description automatically generated

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«ΑΞΙΟΠΟΙΗΣΗ ΜΕΓΑΛΩΝ ΓΛΩΣΣΙΚΩΝ ΜΟΝΤΕΛΩΝ ΓΙΑ ΤΗΝ ΑΝΑΠΤΥΞΗ ΑΥΤΟΜΑΤΩΝ ΕΛΕΓΧΩΝ ΑΠΟΔΟΧΗΣ ΛΟΓΙΣΜΙΚΟΥ»

ΠΛΑΤΙΑΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

ΑΡΙΘΜΟΣ ΜΗΤΡΩΟΥ: P3200157

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:  
ΔΙΑΜΑΝΤΙΔΗΣ ΝΙΚΟΛΑΟΣ

ΑΘΗΝΑ, ΑΥΓΟΥΣΤΟΣ 2024

© Copyright

Κωνσταντίνος Πλατιάς

Σημείωμα Συγγραφέα

Το δοκίμιο αυτό αποτελεί πτυχιακή εργασία που συντάχθηκε για το Προπτυχιακό Πρόγραμμα Σπουδών του τμήματος Πληροφορικής του ΟΙΚΟΝΟΜΙΚΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΑΘΗΝΩΝ και υποβλήθηκε τον Αύγουστο του 2024.

Ο συγγραφέας βεβαιώνει ότι το περιεχόμενο του παρόντος έργου είναι αποτέλεσμα προσωπικής εργασίας και ότι έχει γίνει η κατάλληλη αναφορά στην εργασία τρίτων -όπου κάτι τέτοιο ήταν απαραίτητο-, σύμφωνα με τους κανόνες της ακαδημαϊκής δεοντολογίας.

Πινακασ περιεχομενων

[1. ΕΙΣΑΓΩΓΗ 3](#_Toc173000352)

[2. ΜΕΓΑΛΑ ΓΛΩΣΣΙΚΑ ΜΟΝΤΕΛΑ 4](#_Toc173000353)

[2.1 Τι είναι ένα νευρωνικό δίκτυο και πως λειτουργεί 4](#_Toc173000354)

[2.1.1 Ορισμός Νευρωνικών Δικτύων 4](#_Toc173000355)

[2.1.2 Ορισμός της μείωσης βαρών με καθοδική κλίση 5](#_Toc173000356)

[2.1.3 Ορισμός του αλγόριθμου Ανάστροφης Μετάδοσης 6](#_Toc173000357)

[2.2 Τι είναι ένα μεγάλο γλωσσικό μοντέλο 7](#_Toc173000358)

[2.2.1 Ορισμός Μεγάλου Γλωσσικού Μοντέλου 7](#_Toc173000359)

[2.2.2 Τι κάνει ένα Transformer 7](#_Toc173000360)

[2.2.3 Τι είναι το Attention layer 8](#_Toc173000361)

[2.2.4 Τι είναι το Feed Forward Step 9](#_Toc173000362)

[2.3 Ιστορία και εξέλιξη των μεγάλων γλωσσικών μοντέλων 9](#_Toc173000363)

[2.3.1 Το chatbot ELIZA 9](#_Toc173000364)

[2.3.2 Άνοδος των Νευρωνικών Δικτύων 9](#_Toc173000365)

[2.3.3 Δημιουργία των LSTM 10](#_Toc173000366)

[2.3.4 Δημιουργία Gated Recurrent Network 10](#_Toc173000367)

[2.3.5 Άνοδος του συστατικού Attention 10](#_Toc173000368)

[2.3.6 Η εφεύρεση των Transformers 11](#_Toc173000369)

[2.3.7 Εμφάνιση Μεγάλων Γλωσσικών Μοντέλων 11](#_Toc173000370)

[2.4 Κύριες εφαρμογές και χρήσεις 12](#_Toc173000371)

[2.4.1 Ανάλυση ήχου 12](#_Toc173000372)

[2.4.2 Δημιουργία περιεχομένου 12](#_Toc173000373)

[2.4.3 Υποστήριξη πελατών 13](#_Toc173000374)

[2.4.4 Μετάφραση γλωσσών 13](#_Toc173000375)

[2.4.5 Εκπαίδευση 13](#_Toc173000376)

[2.4.6 Κυβερνοασφάλεια 13](#_Toc173000377)

[2.5 Παραδείγματα Μεγάλων Γλωσσικών Μοντέλων 14](#_Toc173000378)

[2.5.1 BERT 14](#_Toc173000379)

[2.5.2 GEMINI 14](#_Toc173000380)

[2.5.3 GPT – 3 14](#_Toc173000381)

[2.5.4 GPT - 3.5 και GPT – 3.5 Turbo 15](#_Toc173000382)

[2.5.5 GPT – 4 15](#_Toc173000383)

[2.5.6 GPT – 4o 16](#_Toc173000384)

[3. ΜΕΓΑΛΑ ΓΛΩΣΣΙΚΑ ΜΟΝΤΕΛΑ ΣΤΗΝ ΑΝΑΠΤΥΞΗ ΛΟΓΙΣΜΙΚΟΥ 16](#_Toc173000385)

[4. ΒΙΒΛΙΟΓΡΑΦΙΑ 17](#_Toc173000386)

# ΕΙΣΑΓΩΓΗ

# ΜΕΓΑΛΑ ΓΛΩΣΣΙΚΑ ΜΟΝΤΕΛΑ

## Τι είναι ένα νευρωνικό δίκτυο και πως λειτουργεί

### Ορισμός Νευρωνικών Δικτύων

Ξεκινώντας την συζήτηση για τα μεγάλα γλωσσικά μοντέλα και την χρήση τους, δεν θα μπορούσε να παραληφθεί η αναφορά στην βασική συνιστώσα που τα αποτελεί, το ονομαζόμενο **«νευρωνικό δίκτυο»** αλλά και τους τρόπους με τους οποίους αυτό λειτουργεί. Ένα νευρωνικό δίκτυο, όπως γίνεται αντιληπτό και από το ίδιο το όνομα, είναι ένα δίκτυο πολλών συνδεδεμένων νευρώνων, χωρισμένων σε διαφορετικά «στρώματα», όπου κάθε νευρώνας μπορεί να συσχετιστεί ως ένα αντικείμενο το οποίο παίρνει κάποιες εισόδους, εκτελεί κάποιους πολύ απλούς υπολογισμούς με αυτές και στην συνέχεια παράγει κάποια έξοδο, ένα νούμερο, το οποίο με την σειρά του περνάει σαν είσοδο στους επόμενουA diagram of a neural network

Description automatically generatedς νευρώνες. Συγκεκριμένα, οι απλοί αυτοί υπολογισμοί είναι το άθροισμα των εισόδων που δέχθηκε, πολλαπλασιασμένοι με ένα νούμερο, το ονομαζόμενο «βάρος», το οποίο μαθαίνει το νευρωνικό δίκτυο από χιλιάδες δεδομένα κατά την εκπαίδευσή του. Στην συνέχεια, αφού εφαρμοστεί ένας αριθμητικός μετασχηματισμός, το αποτέλεσμα αυτό, όπως αναφέρθηκε και προηγουμένως, μεταφέρεται στους νευρώνες του επόμενου «στρώματος» σαν είσοδο και η διαδικασία αυτή σταματά στο τελευταίο «στρώμα» νευρώνων του νευρωνικού δικτύου. Για παράδειγμα, μπορεί το αποτέλεσμα που προωθεί ο νευρώνας στο επόμενο στρώμα να είναι μηδέν εάν το άθροισμα που υπολόγισε μέσω των εισόδων είναι αρνητικός αριθμός, διαφορετικά εάν είναι θετικός αριθμός ή μηδέν, να προωθεί απλά το ίδιο το άθροισμα χωρίς αλλαγές (ονομαζόμενη «συνάρτηση ενεργοποίησης»). Συνεχίζοντας, τα μοντέλα που χρησιμοποιούν βαθιά μάθηση, χρησιμοποιούνε πολλά εκατομμύρια νευρώνες, χωρισμένα σε πολυάριθμα στρώματα με δισεκατομμύρια βάρη και πολύπλοκες διατάξεις νευρώνων, όπου όμως η βασική ιδέα παραμένει η παραπάνω. (Ανδρουτσόπουλος, 2024)

Εικόνα 1. Παράδειγμα απλού νευρωνικού δικτύου

### Ορισμός της μείωσης βαρών με καθοδική κλίση

Αυτό που κάνει τα νευρωνικά δίκτυα τόσο ενδιαφέροντα, και ως αποτέλεσμα, και την μηχανική μάθηση, είναι ότι στην ουσία, ποτέ δεν γράφεται κάποιος αλγόριθμος ο οποίος ορίζει συγκεκριμένα τι θα πρέπει να κάνει ένα νευρωνικό δίκτυο και πως θα πρέπει να παράγει αποτελέσματα, αλλά από την άλλη, γράφεται ένας αλγόριθμος ο οποίος θα μπορεί μέσω της εισαγωγής εκατομμυρίων παραδειγμάτων και δεδομένων με τις ορθές «ετικέτες» τους (το επιθυμητό δηλαδή αποτέλεσμα), να μεταβάλλει με τέτοιο τρόπο τα εκατομμύρια βάρη από τα οποία αποτελείται, ώστε να αποδίδει καλύτερα στα παραδείγματα αυτά. Τα δεδομένα αυτά ονομάζονται «δεδομένα εκπαίδευσης», τα οποία παράλληλα με τα «δεδομένα δοκιμής», τα οποία είναι παραδείγματα όπου το νευρωνικό δίκτυο δεν έχει «ξαναδεί» και τα οποία χρησιμοποιούμε για να αξιολογήσουμε την αποτελεσματικότητα του νευρωνικού μας δικτύου, αποτελούν την συνολική αξιολόγηση του συστήματός μας. Το πρόβλημα αυτό της μεταβολής των βαρών που καλείται να λύσει ένας αλγόριθμος, καταλήγει τελικά να είναι η εύρεση του ελαχίστου μιας «συνάρτησης κόστους». Η συνάρτηση αυτή στην γενική εικόνας της υπολογίζεται με βάση τα αποτελέσματα που παράγει το νευρωνικό δίκτυο και τα επιθυμητά αποτελέσματα που έχουμε για κάθε παράδειγμα, και συνεπώς έχει μεγάλη τιμή εάν τα αποτελέσματα διαφέρουν σε μεγάλο βαθμό από τα επιθυμητά και μικρό στην αντίθετη περίπτωση. Σκοπός του αλγορίθμου είναι με βάση τον μέσο όρο όλων αυτών των τιμών κόστους για κάθε παράδειγμα, να προσπαθήσει να μεταβάλλει τις τιμές των βαρών της συνάρτησης κόστους, ώστε αυτή να φτάσει σε ένα τοπικό ελάχιστο. Η τεχνική αυτή ονομάζεται «**καθοδική κλίση»** ή όπως είναι γνωστό, gradient descent, αφού προσπαθεί να βρει ένα τοπικό ελάχιστο μίας συνάρτησης με πολλές χιλιάδες μεταβλητές (τις εισόδους και τα βάρη), η οποία εάν A computer graphics of a red and yellow wavy surface

Description automatically generated with medium confidenceαναπαρασταθεί σε ένα διανυσματικό χώρο, έχει την εικόνα ενός «γεωγραφικού τοπίου» στο οποίο πρέπει να βρεθεί ένα τοπικό «χαμηλότερο σημείο», όπως φαίνεται και στην *Εικόνα 2*. Χρησιμοποιώντας μαθηματικές έννοιες, όλα τα παραπάνω καταλήγουν τελικά στην εύρεση του αρνητικού του «ανάδελτα ή ∇» της συνάρτησης κόστους, το οποίο δείχνει προς την πιο απότομη μείωση μίας συνάρτησης (Sanderson, 2017).A black and white text with numbers and symbols

Description automatically generated with medium confidence

Εικόνα 2. Γεωγραφική αναπαράσταση ενός παραδείγματος Gradient descent

Εικόνα 3. Παράδειγμα μαθηματικής αναπαράστασης του ανάδελτα

Κατά συνέπεια, σε αυτό το σημείο θα αναφερθεί περιληπτικά ο ορισμός του αλγορίθμου που προσπαθεί να επιτύχει όλα τα παραπάνω, δηλαδή να υπολογίσει αυτό το τοπικό ελάχιστο της συνάρτησης κόστους, ο οποίος αναφέρεται ως «**αλγόριθμος ανάστροφης μετάδοσης**»

### Ορισμός του αλγόριθμου Ανάστροφης Μετάδοσης

Όπως παρουσιάστηκε παραπάνω, ο αλγόριθμος ανάστροφης μετάδοσης είναι ένας αλγόριθμος εύρεσης ενός τοπικού ελαχίστου μέσω του υπολογισμού του αρνητικού ανάδελτα μίας συνάρτησης κόστους. Αρχικά, ο αλγόριθμος αρχικοποιεί όλα τα βάρη του νευρωνικού δικτύου με τυχαίες μικρές τιμές και για μία δεδομένη είσοδο/παράδειγμα εκπαίδευσης, υπολογίζει το συνολικό σφάλμα στην τελική έξοδο, συγκρίνοντας την πραγματική έξοδο με την επιθυμητή έξοδο. Στην συνέχεια, το σφάλμα μεταδίδεται από την έξοδο προς την είσοδο υπολογίζοντας παράλληλα τους παραγώγους ως προς κάθε ξεχωριστό βάρος, με τον κανόνα της αλυσίδας, και κάθε βάρος ενημερώνεται ώστε να δείχνει (με την τεχνική που αναφέραμε νωρίτερα της καθοδικής κλίσης) προς την κατεύθυνση που μειώνεται το σφάλμα. Η παραπάνω διαδικασία γίνεται για κάθε παράδειγμα εκπαίδευσης που δίνεται στο νευρωνικό δίκτυο, οι οποίες ονομάζονται και «εποχές» και τελειώνει είτε μόλις το σύστημα ξεπεράσει έναν μέγιστο αριθμό εποχών, είτε εάν το συνολικό σφάλμα μειωθεί σε επιθυμητό αριθμό (Ανδρουτσόπουλος, 2023).

## Τι είναι ένα μεγάλο γλωσσικό μοντέλο

### Ορισμός Μεγάλου Γλωσσικού Μοντέλου

Όπως αναφέρει και ο Ανδρουτσόπουλος με απλά λόγια, ένα μεγάλο γλωσσικό μοντέλο είναι και αυτό ένα νευρωνικό δίκτυο, το οποίο όμως παίρνει σαν εισόδους λέξεις, ή καλύτερα tokens, σε μορφή αριθμών, τα οποία αποτελούν ίσως ένα ημιτελές κείμενο, και παράγει ως εξόδους όλες τις πιθανές λέξεις που θα μπορούσαν να είναι η επόμενη λέξη, επιλέγοντας αυτήν με την μεγαλύτερη πιθανότητα ορθότητας. Έτσι, βασιζόμενοι στο μεγάλο γλωσσικό μοντέλο, μπορούμε να πάρουμε την νέα αυτή πρόταση που παρήγαγε, να την ξαναδώσουμε σαν είσοδο και να παράγει ξανά άλλη μία επόμενη πιθανή λέξη και έτσι μετά από έναν αριθμό επαναλήψεων της διαδικασίας αυτής να παραχθεί μια ολοκληρωμένη απάντηση η οποία να βγάζει ένα νόημα (2024, σελίδα 9).

### Τι κάνει ένα Transformer

Πιο συγκεκριμένα, τα μεγάλα γλωσσικά μοντέλα σπάνε το κείμενο που δέχονται σαν είσοδο σε διαφορετικά tokens, τα οποία δεν είναι απαραίτητα λέξεις, αλλά μπορεί να αποτελούν και κομμάτια λέξεων, και στην συνέχεια τα αναπαριστά σε διανύσματα. Τα διανύσματα αυτά αποτελούνται από πολλές χιλιάδες μεταβλητές και αριθμούς και τοποθετούνται σε έναν πολυδιάστατο χώρο, όπου λέξεις με παρόμοια σημασία όπως για παράδειγμα «γάτα» και «σκύλος» βρίσκονται πολύ κοντά. Η χρήση διανυσμάτων επιτρέπει στα μεγάλα γλωσσικά μοντέλα να πραγματοποιούν μαθηματικές πράξεις που αποκαλύπτουν σχέσεις μεταξύ λέξεων. Για παράδειγμα έχει διαπιστωθεί ότι αν από το διάνυσμα της λέξης "μεγαλύτερος" αφαιρεθεί το διάνυσμα της λέξης "μεγάλος" και προστεθεί το διάνυσμα της λέξης "μικρός", το αποτέλεσμα θα είναι το διάνυσμα της λέξης "μικρότερος". Τα μεγάλα γλωσσικά μοντέλα μπορούν να αντιπροσωπεύουν τις λέξεις με διαφορετικά διανύσματα ανάλογα με τα συμφραζόμενα. Αυτό επιτυγχάνεται με τη χρήση ενός αρχιτεκτονικού μοντέλου νευρωνικού δικτύου, γνωστού ως «**transformer»,** που ενημερώνει τα διανύσματα των λέξεων μέσω πολλαπλών επιπέδων. A close-up of words

Description automatically generated Κάθε μεγάλο γλωσσικό μοντέλο αποτελείται από πολλά στρώματα transformers συνδεδεμένα μεταξύ τους, όπου σκοπός του κάθε ενός είναι να εμπλουτίσει με πληροφορία κάθε ένα token με βάση τα γύρω και έτσι να αλλάξει ως προς την σωστή «κατεύθυνση» το διάνυσμα του για να έχει την σωστή σημασία που αποκαλύπτεται από τα συμφραζόμενα. Η διαδικασία αυτή επαναλαμβάνεται για κάθε transformer το οποίο προσθέτει και βελτιώνει τα διανύσματα έως το τελικό στρώμα (Lee, 2023).

Εικόνα 4. Παράδειγμα επικοινωνίας των transformers σε ένα μεγάλο γλωσσικό μοντέλο

### Τι είναι το Attention layer

Ένα κύριο συστατικό το οποίο αποτελεί το transformer και του επιτρέπει να εμπλουτίζει τα διανύσματα, είναι το «attention layer». Συγκεκριμένα, το συστατικό αυτό επιτρέπει στα tokens να ανταλλάσσουν πληροφορίες μεταξύ τους και να εμπλουτίζουν το ένα το άλλο, ώστε να προκύπτει η σωστή σημασία κάθε ενός. Για παράδειγμα, η λέξη «μοντέλο» έχει διαφορετική σημασία ανάλογα τα συμφραζόμενα, όπως για παράδειγμα στην πρόταση «μοντέλο μαθηματικών» και «μοντέλο του Χόλυγουντ». Η λειτουργία του attention layer είναι να γίνει ο σωστός αυτός διαχωρισμός για κάθε διαφορετικό token με βάση τα γύρω του. Η διαδικασία αυτή επιτρέπει στα μεγάλα γλωσσικά μοντέλα να μαντεύουν σωστά την επόμενη λέξη σε κάθε κείμενο (Lee, 2023)

### Τι είναι το Feed Forward Step

Μετά τη μεταφορά πληροφοριών ανάμεσα σε διανύσματα λέξεων από τα attention heads, to transformer αποτελείται και από ένα ακόμη συστατικό, ένα στρώμα που ονομάζεται feed forward, το οποίο «σκέφτεται» κάθε διάνυσμα λέξης και προσπαθεί να προβλέψει την επόμενη λέξη. Σε αυτό το στάδιο δεν γίνεται ανταλλαγή πληροφοριών μεταξύ των λέξεων, το στρώμα feed forward αναλύει κάθε λέξη μεμονωμένα. Ωστόσο, το στρώμα feed forward έχει πρόσβαση σε οποιαδήποτε πληροφορία αντιγράφηκε προηγουμένως από όλα τα προηγούμενα attention heads ώστε να μπορέσει αποτελεσματικότερα να προβλέψει την επόμενη λέξη (Lee, 2023).

## Ιστορία και εξέλιξη των μεγάλων γλωσσικών μοντέλων

### Το chatbot ELIZA

Το 1966 επιβεβαιώνεται ότι ήταν το έτος της γέννησης του «ELIZA», που θεωρείται το πρώτο chatbot που κατασκευάστηκε από ανθρώπους και δημιουργήθηκε από τον Joseph Weizenbaum στο πανεπιστήμιο του MIT αποτελώντας ένα πρωτοποριακό πείραμα της εποχής του, επιτρέποντας την αλληλεπίδραση ανθρώπου-υπολογιστή. Παρόλο που το chatbot δεν κατανοούσε τις συνομιλίες με τους ανθρώπους σε τέτοιο βάθος όπως οι ίδιοι άνθρωποι ή όπως το ChatGPT σήμερα, μπορούσε όμως να δημιουργήσει την ψευδαίσθηση συνομιλίας με την τεχνική της αναδιατύπωσης δηλώσεων των χρηστών ως ερωτήσεις. Εκείνη την εποχή, δημιουργήθηκαν πολλές παραλλαγές του συγκεκριμένου chatbot οι οποίες λειτουργούσαν με παρόμοιο τρόπο και μια από τις πιο γνωστές ονομάζεται «DOCTOR», η οποία ανταποκρινόταν σαν ψυχοθεραπευτής. Αυτή η αρχή έβαλε τις βάσεις για περαιτέρω έρευνα στον τομέα των chatbots και της επεξεργασίας φυσικής γλώσσας. (Pi, 2024)

### Άνοδος των Νευρωνικών Δικτύων

Όπως αναφέρει ο Pi στο άρθρο του, προς τα τέλη του 20ου αιώνα, εμφανίστηκαν τα νευρωνικά δίκτυα, εμπνευσμένα βαθιά από τον ανθρώπινο εγκέφαλο, όπως φαίνεται και από το όνομα, και τους διασυνδεδεμένους νευρώνες του. Το 1986, τα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNN) κέρδισαν άμεση δημοτικότητα. Σε αντίθεση με τα παραδοσιακά προωθητικά νευρωνικά δίκτυα, όπου η ροή των πληροφοριών είχε μονάχα μία κατεύθυνση, τα RNN μπορούσαν να θυμούνται προηγούμενες εισόδους και να απαντούν με βάση το πλαίσιο και έτσι εκπαιδεύονταν να επεξεργάζονται και να μετατρέπουν μια ακολουθία δεδομένων εισόδου σε συγκεκριμένη ακολουθία δεδομένων εξόδου. Ωστόσο, τα RNN είχαν περιορισμούς, ειδικά με μεγάλες προτάσεις, και ουσιαστικά είχαν το μεγάλο μειονέκτημα στο να «θυμούνται» πράγματα, κάτι σαν το σημερινό context size των μεγάλων γλωσσικών μοντέλων, το οποίο τα κάνει να φαίνεται σαν να «ξεχνάνε» πληροφορίες από προηγούμενα μηνύματα(2024).

### Δημιουργία των LSTM

Το 1997 εμφανίστηκε η Μνήμη Μακράς Βραχείας Διάρκειας (LSTM), μια εξειδικευμένη μορφή RNN, που μπορούσε να θυμάται πληροφορίες για μεγάλες ακολουθίες προτάσεων. Συγκεκριμένα, το εργαλείο αυτό είχε μια μοναδική αρχιτεκτονική που αποτελούνταν από πύλες εισόδου, λήθης και πύλες εξόδου, οι οποίες καθόριζαν πόσες πληροφορίες να θυμούνται, να απορρίπτουν ή να εξάγουν σε κάθε βήμα. (Pi, 2024)

### Δημιουργία Gated Recurrent Network

Το 2014, εμφανίστηκαν οι Μονάδες Επαναλαμβανόμενων Δικτύων με Πύλες (GRU), σχεδιασμένες για να επιλύουν τα ίδια προβλήματα με τα LSTM, αλλά με απλούστερη δομή. Οι GRU χρησιμοποιούσαν δύο πύλες: μια πύλη ενημέρωσης και μια πύλη επαναφοράς. (Pi, 2024)

### Άνοδος του συστατικού Attention

Τελικά, οι τεχνολογίες RNN, LSTM και GRU φάνηκε να μην είναι τόσο καλές στην διατήρηση των συμφραζόμενων, όταν αυτό μεγαλώνει σε μεγάλο βαθμό προσθέτοντας πληροφορίες. Έτσι, δημιουργήθηκε ο μηχανισμός ο οποίος ονομάστηκε **«Attention»**, όπου αναφέρθηκε και παραπάνω και ο οποίος προσέφερε μία νέα οπτική στα μεγάλα γλωσσικά μοντέλα. Συγκεκριμένα, το attention επέτρεψε στο μοντέλο να «κοιτάει» πίσω σε ολόκληρο το υλικό που έχει διαθέσιμο δυναμικά, και να επιλέγει τα πιο σημαντικά κομμάτια που προσθέτουν σημασία στις υπόλοιπες προτάσεις (Pi, 2024).

Εικόνα 5. Μέτρηση απόδοσης Attention και RNN

### Η εφεύρεση των Transformers

Το 2017 ήταν το έτος στο οποίο πρωτοεμφανίστηκε η έννοια των **«Transformers»,** στο paper με όνομα «Attention is all you need» από τον Vaswani και τους συνεργάτες του στην Google. Η νέα αυτή αρχιτεκτονική, όπως εξηγήθηκε και στην προηγούμενη υποενότητα, χρησιμοποιούσε ως βασικό της εργαλείο τον μηχανισμό attention για να επεξεργαστεί τα δεδομένα εισόδου και ήταν ικανή να επεξεργάζεται ακολουθίες παράλληλα, θέτοντας τα θεμέλια για μεταγενέστερα μοντέλα όπως το BERT και το GPT (Pi, 2024).

### Εμφάνιση Μεγάλων Γλωσσικών Μοντέλων

Με την μεγάλη επιτυχία των transformers, το επόμενο λογικό βήμα ήταν η μεγέθυνση της κλίμακας. Αυτό ξεκίνησε με το μοντέλο BERT της Google το 2018 και συνεχίστηκε με την κυκλοφορία των GPT-2 και T5 το 2019,του GPT-3 το 2020 αλλά και των νέων εκδόσεων αυτού όπως το GPT-3.5, GPT-4 και GPT-4o. Αυτά τα μοντέλα είναι ικανά να εκτελούν αμέτρητες εργασίες σημειώνοντας μια μεγάλη πρόοδο στις δυνατότητες της τεχνητής νοημοσύνης (Pi, 2024).

## Κύριες εφαρμογές και χρήσεις

Με την τεράστια εξέλιξη των μεγάλων γλωσσικών μοντέλων, αυτά έχουν γίνει καθημερινό εργαλείο για πολλούς ανθρώπους, γλυτώνοντας χρόνο, αφού το μέγεθός τους από δεδομένα και η πολυπλοκότητά τους επιτρέπει σε στιγμιαίες απαντήσεις χωρίς καθυστέρηση, κόστος, αφού τα περισσότερα είναι δωρεάν στην απλούστερη αλλά και πάρα πολύ αποτελεσματική μορφή τους αλλά και κόπο, αφού απαιτούν ελάχιστες γνώσεις για να τα χρησιμοποιήσεις ορθά. Όλα τα παραπάνω έχουν ωθήσει τα μεγάλα γλωσσικά μοντέλα στο να έχουν πολυάριθμες χρήσεις σε πολλούς τομείς. Όπως αναφέρει και ο Sumrak, κάποια από τα δημοφιλέστερα παραδείγματα χρήσης των μεγάλων γλωσσικών μοντέλων είναι στην ανάλυση ήχου, στην δημιουργία περιεχομένου, στην υποστήριξη πελατών, στην μετάφραση γλωσσών, στην εκπαίδευση αλλά και στην κυβερνοασφάλεια (2024).

### Ανάλυση ήχου

Συγκεκριμένα, τα μεγάλα γλωσσικά μοντέλα αναθεωρούν τον τρόπο με τον οποίο οι άνθρωποι χειρίζονταν τα δεδομένα ήχου, αφού έχουν την δυνατότητα να ακούνε πολύωρες συζητήσεις και να παράγουν αποτελεσματικά περιλήψεις, σημαντικά σημεία των συζητήσεων αλλά και να απαντούν ερωτήσεις σχετικά με την συνάντηση. Ακόμη, μπορούν να παίρνουν σημαντικές αποφάσεις από πάρα πολύ μεγάλο αριθμό κλήσεων όπως για παράδειγμα με πελάτες και να εξάγουν αποτελέσματα με βάση αυτά (Sumrak, 2024).

### Δημιουργία περιεχομένου

Επιπρόσθετα, τα μεγάλα γλωσσικά μοντέλα χρησιμοποιούνται αποτελεσματικά από συγγραφείς και εμπόρους για την δημιουργία αρχικών σχεδίων(drafts), για την πρόταση διάφορων αλλαγών και για την γρήγορη εύρεση άρθρων και reports στο διαδίκτυο. Όλα τα παραπάνω επιταχύνουν με πολύ μεγάλους ρυθμούς την παραγωγικότητα των ατόμων, επιτρέποντάς τους να επικεντρωθούν παραπάνω στο πιο απαιτητικό και δημιουργικό κομμάτι της δουλείας τους και να αφήσουν το μεγάλο γλωσσικό μοντέλο να ασχοληθεί με τα μηχανικά στοιχεία της εργασίας τους. (Sumrak, 2024).

### Υποστήριξη πελατών

Ένας ακόμη τομέας στον οποίο τα μεγάλα γλωσσικά μοντέλα έχουν κάνει ραγδαία άνοδο και χρησιμοποιούνται κατά κόρον, είναι στην εξυπηρέτηση πελατών. Συγκεκριμένα, εταιρείες τηλεφωνίας, κυβερνητικές σελίδες του δημοσίου τομέα αλλά και πολύ μεγάλες επιχειρήσεις έχουν υιοθετήσει ένα μοντέλο εξυπηρέτησης πελατών μέσω των μεγάλων γλωσσικών μοντέλων, το οποίο είναι διαθέσιμο 24 ώρες και κάθε ημέρα της εβδομάδας χωρίς βοήθεια από κάποιο ανθρώπινο παράγοντα, το οποίο επιτρέπει την συνεχή παροχή πολύτιμης βοήθειας σε χρήστες από όλο τον κόσμο. Αυτό επιτρέπει σε μία αψεγάδιαστη υποστήριξη πελατών χωρίς σπατάλη πολλών πόρων (Sumrak, 2024).

### Μετάφραση γλωσσών

Παράλληλα, τα μεγάλα γλωσσικά μοντέλα βοηθούν τις επιχειρήσεις στην άρση των γλωσσικών φραγμών και δίνουν την δυνατότητα στις επιχειρήσεις να προσεγγίζουν πελάτες αλλά και να προσλαμβάνουν άτομα από όλες τις χώρες του κόσμου. Αυτά τα μοντέλα προσφέρουν ακριβείς υπηρεσίες μετάφρασης σε πραγματικό χρόνο, κάνοντας ιστότοπους, εφαρμογές και ψηφιακό περιεχόμενο παγκοσμίως προσβάσιμα (Sumrak, 2024).

### Εκπαίδευση

Ένας από τους σημαντικότερους τομείς στον οποίο γίνεται χρήση των μεγάλων γλωσσικών μοντέλων είναι αυτός της εκπαίδευσης, όπου μπορούν να χρησιμοποιηθούν για την παροχή εξατομικευμένης εκπαίδευσης προσαρμόζοντας το περιεχόμενο στις ατομικές ανάγκες των μαθητών αλλά να προσφέρουν φροντιστηριακή υποστήριξη σε άτομα που δεν διαθέτουν την οικονομική δυνατότητα, να δημιουργούν ερωτήσεις πρακτικής και να παρέχουν λεπτομερείς εξηγήσεις προσαρμοσμένες σε αυτά που οι μαθητές μαθαίνουν ή δυσκολεύονται, όπως αυτό έχει γίνει κατανοητό από την συζήτηση με το κάθε άτομο (Sumrak, 2024).

### Κυβερνοασφάλεια

Τέλος, τα μεγάλα γλωσσικά μοντέλα μπορούν να χρησιμοποιηθούν στην ανάλυση και την ερμηνεία μεγάλων ποσών δεδομένων κυβερνοασφάλειας, ώστε να προβλέπουν, να αναγνωρίζουν και να ανταποκρίνονται σε πιθανές απειλές ασφαλείας. Επίσης, λόγω της στοχευμένης εκπαίδευσή τους, επιτρέπουν την ταχύτερη και πιο ακριβή ανίχνευση και ανταπόκριση στις απειλές, ενισχύοντας την ασφάλεια των επιχειρήσεων (Sumrak, 2024).

## Παραδείγματα Μεγάλων Γλωσσικών Μοντέλων

Ολοκληρώνοντας το κεφάλαιο 2 για τα μεγάλα γλωσσικά μοντέλα, δεν θα μπορούσε παραλειφθεί μία αναφορά σε κάποια από τα πιο πολυχρησιμοποιούμενα μεγάλα γλωσσικά μοντέλα και οι λειτουργίες και διαφοροποιήσεις του κάθε ενός.

### BERT

Το **«BERT»** είναι ένα μεγάλο γλωσσικό μοντέλο το οποίο αναπτύχθηκε από την Google το έτος 2018. Το BERT είναι ένα μοντέλο που χρησιμοποιεί την αρχιτεκτονική με τα πολυάριθμα στρώματα transformers συνδεδεμένα μεταξύ τους και 342 εκατομμύρια παραμέτρους για την επεξεργασία εισόδων, ενώ εκπαιδεύτηκε και σε πολλά εκατομμύρια δεδομένα για να παράγει απαντήσεις σε φυσική γλώσσα κατανοητή από τους ανθρώπους. Η κύρια χρήση του ήταν η βελτίωση της κατανόησης των ερωτημάτων που του γινόταν σε μία αναζήτηση με το Google το 2019 (Lutkevich, 2024).

### GEMINI

Συνεχίζοντας με την Google, το νέο ανανεωμένο μοντέλο της ονομάστηκε **«GEMINI»**, ίδιο με το όνομα του chatbot που προσφέρει η εταιρεία στους χρήστες της. Το Gemini αποτελεί ένα πολυτροπικό μοντέλο, δηλαδή μπορεί να χειριστεί ήχο, εικόνες, βίντεο αλλά και κείμενο, σε αντίθεση με πολλά άλλα γλωσσικά μοντέλα που λειτουργούν μόνο βάση κειμένου και έχει τρία «μεγέθη», το Ultra, το Pro και το Nano, από το μεγαλύτερο και πιο ικανό προς το μικρότερο και λιγότερο ικανό. Από πολλές πηγές αναφέρεται ότι το Gemini υπερτερεί σε δύναμη από το μοντέλο GPT -4 της OpenAI που θα αναφερθούμε αργότερα (Lutkevich, 2024).

### GPT – 3

Το GPT - 3 αποτελεί το πρώτο δυνατό μοντέλο της OpenAI το οποίο παρουσιάστηκε το 2020 με περισσότερες από 175 δισεκατομμύρια παραμέτρους (το BERT έχει 342 εκατομμύρια), το οποίο χρησιμοποιεί την αρχιτεκτονική των συνδεδεμένων στρωμάτων transformers. Το GPT – 3 είναι δέκα φορές μεγαλύτερο από τον προκάτοχο του, είναι εκπαιδευμένο από εκατομμύρια δεδομένα συμπεριλαμβανομένων αρχείων από το Wikipedia, Common Crawl, WebText2, Books1, Books2 και αποτελεί το τελευταίο μοντέλο από την σειρά μοντέλων που παρήγαγε η OpenAI, για το οποίο ήταν δημόσια γνωστός ο αριθμός των παραμέτρων που χρησιμοποιούσε (Lutkevich, 2024).

### GPT - 3.5 και GPT – 3.5 Turbo

Το GPT – 3.5 αποτελεί την ενημερωμένη έκδοση του GPT – 3 με λιγότερες παραμέτρους αλλά πιο ποιοτικά δεδομένα εκπαίδευσης, το οποίο βρίσκεται πίσω από την λειτουργία της πλατφόρμας Chat GPT η οποία το 2023 άλλαξε ριζικά την δημοτικότητα και την χρήση των μεγάλων γλωσσικών μοντέλων προς το καλό. Το GPT – 3 είναι εκπαιδευμένο με γνώσεις μέχρι και τον Σεπτέμβρη του 2021 και δεν έχει την δυνατότητα εισχώρησης στο διαδίκτυο συγκριτικά με άλλα μεγάλα γλωσσικά μοντέλα. Παράλληλα, μια δυνατότερη έκδοση του ονομάζεται «Turbo» και χρησιμοποιείται από το πολύ γνωστό chatbot της Microsoft, το GitHub COPILOT, με σκοπό την υποβοήθηση των προγραμματιστών στην αποτελεσματικότερη παραγωγή κώδικα και την διόρθωση σφαλμάτων (Lutkevich, 2024).

### GPT – 4

Το GPT-4 αποτελεί το μεγαλύτερο μοντέλο στη σειρά GPT της OpenAI, το οποίο κυκλοφόρησε το 2023. Όπως και τα προηγούμενα, είναι ένα μοντέλο βασισμένο σε transformers. Ωστόσο, ο αριθμός των παραμέτρων του δεν έχει δημοσιοποιηθεί, αν και φήμες λένε ότι έχει πάνω από 170 **τρισεκατομμύρια**. Η OpenAI περιγράφει το GPT-4 ως ένα πολυτροπικό μοντέλο(όπως και το GEMINI), πράγμα που σημαίνει ότι μπορεί να επεξεργάζεται κείμενο, ήχο και εικόνες, σε αντίθεση με τα προηγούμενα μοντέλα της OpenAI που περιορίζονταν μόνο στο κείμενο. Το GPT- 4, όπως αναφέρεται σε άρθρα αλλά και μελέτες, επέδειξε ανθρώπινη απόδοση σε πολλαπλές ακαδημαϊκές εξετάσεις, κάτι το οποίο το κάνει ένα πολύ σημαντικό εργαλείο. Κατά την κυκλοφορία του μοντέλου, κάποιοι υποστήριξαν ότι το GPT-4 πλησίασε την τεχνητή γενική νοημοσύνη (AGI), που σημαίνει ότι είναι εξίσου έξυπνο ή εξυπνότερο από έναν άνθρωπο. Το GPT-4 τροφοδοτεί την αναζήτηση Microsoft Bing, είναι διαθέσιμο στο ChatGPT Plus και θα ενσωματωθεί μελλοντικά στα προϊόντα Microsoft Office. Τέλος, το μοντέλο έχει την δυνατότητα της εύρεσης πληροφοριών από το διαδίκτυο, σε αντίθεση με το GPT – 3.5, κάτι το οποίο το κάνει ακόμη πιο δυνατό (Lutkevich, 2024).

### GPT – 4o

Το GPT-4 Omni (GPT-4o) είναι ο διάδοχος του GPT-4 της OpenAI και προσφέρει αρκετές βελτιώσεις σε σχέση με το προηγούμενο μοντέλο. Το GPT-4o λέγεται ότι δημιουργεί μια πιο φυσική ανθρώπινη αλληλεπίδραση για το Chat GPT και είναι και αυτό ένα μεγάλο πολυτροπικό μοντέλο, δεχόμενο διάφορες εισόδους, συμπεριλαμβανομένων ήχου, εικόνας και κειμένου. Οι συνομιλίες επιτρέπουν στους χρήστες να αλληλεπιδρούν όπως θα έκαναν σε μια κανονική ανθρώπινη συνομιλία. Το GPT-4o μπορεί να δει φωτογραφίες ή οθόνες και να κάνει ερωτήσεις σχετικά με αυτές κατά τη διάρκεια της αλληλεπίδρασης (Lutkevich, 2024).

# ΜΕΓΑΛΑ ΓΛΩΣΣΙΚΑ ΜΟΝΤΕΛΑ ΣΤΗΝ ΑΝΑΠΤΥΞΗ ΛΟΓΙΣΜΙΚΟΥ

# ΒΙΒΛΙΟΓΡΑΦΙΑ

Lee, T. B. (2023, Ιούλιος 31). Ανάκτηση από arstechnica.com: https://arstechnica.com/science/2023/07/a-jargon-free-explanation-of-how-ai-large-language-models-work/6/

Pi, W. (2024, Μάϊος 7). *Research Graph.* Ανάκτηση από Medium: https://medium.com/@researchgraph/brief-introduction-to-the-history-of-large-language-models-llms-3c2efa517112

Sanderson, G. (2017, Οκτώβριος 16). *3blue1brown.* Ανάκτηση από 3blue1brown: https://www.3blue1brown.com/lessons/gradient-descent#another-way-to-think-about-the-gradient

Sumrak, J. (2024, Μάρτιος 11). *7 LLM use cases and applications in 2024.* Ανάκτηση από AssemblyAI: https://www.assemblyai.com/blog/llm-use-cases/

Ανδρουτσόπουλος, Ί. (2024, Φεβρουάριος). Τεχνητή Νοημοσύνη και Μεγάλα Γλωσσικά Μοντέλα. *ΟΠΑ News Εφημερίδα Οικονομικού Πανεπιστημίου Αθηνών Τεύχος 51*, σσ. 8-9.