Βαθιά Μάθηση στην Επεξεργασία Φυσικής Γλώσσας

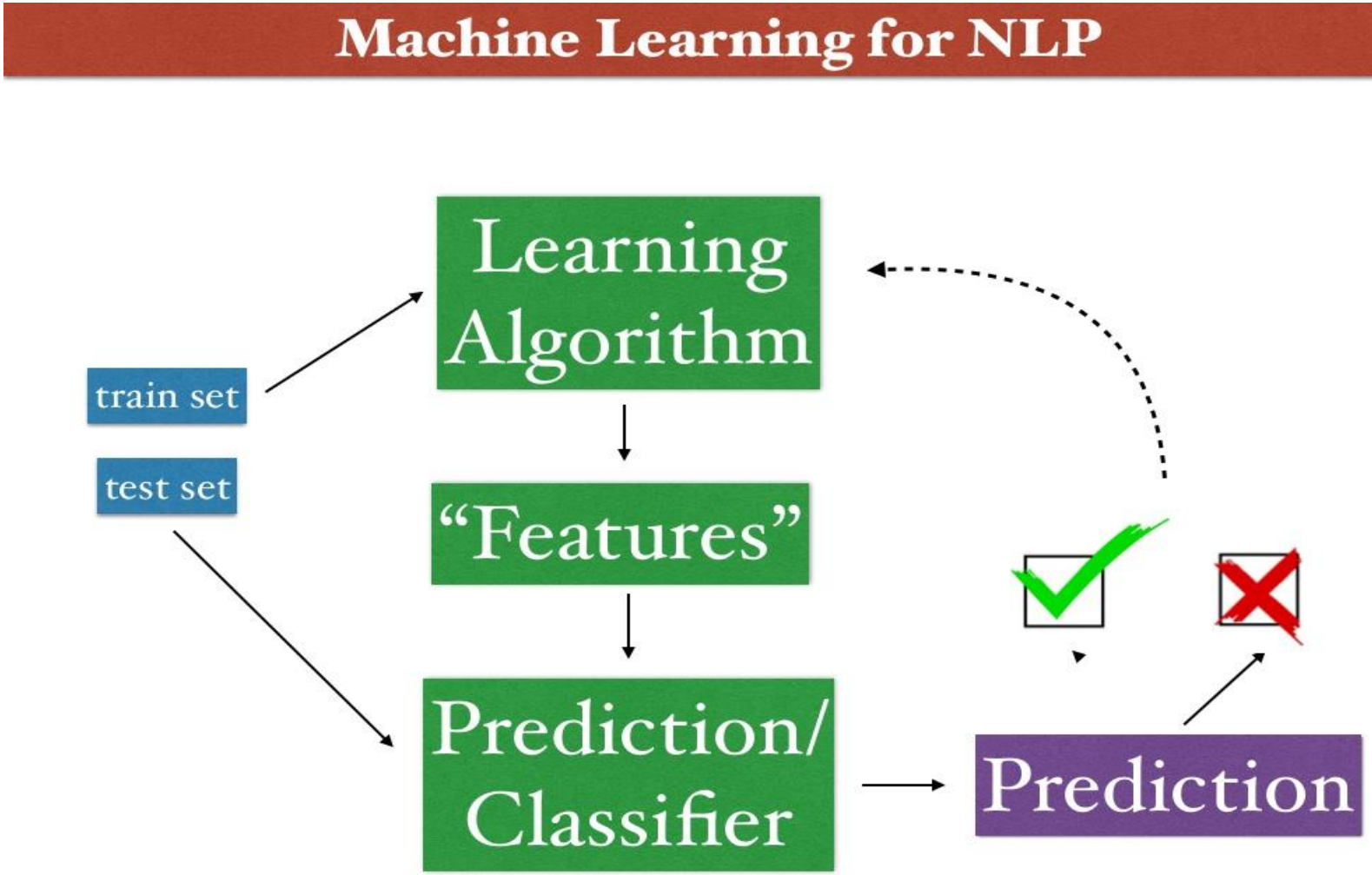
Deep Learning for NLP

Κάτια Κερμανίδου
kerman@ionio.gr

Μηχανική Μάθηση: Παραδοσιακή Προσέγγιση

- Για κάθε καινούρια εργασία:
  - Σύλλεξε όσο περισσότερα επισημειωμένα δεδομένα
  - Αφιέρωσε χρόνο στην μηχανική χαρακτηριστικών (feature engineering)
    - Feature extraction
  - Τρέξε πάνω στα δεδομένα αλγορίθμους μάθησης
  - Συνέχισε την μηχανική χαρακτηριστικών
    - Feature selection
    - Dimensionality reduction
  - Επανέλαβε

Μηχανική Μάθηση: Παραδοσιακή Προσέγγιση



Μηχανική Μάθηση: Παραδοσιακή Προσέγγιση

- Όταν δουλεύει καλά, αυτό οφείλεται στην χειρωνακτική σχεδίαση χαρακτηριστικών για την αναπαράσταση των γλωσσολογικών φαινομένων
  - Πχ χαρακτηριστικά για την αναγνώριση ονομάτων-οντοτήτων (τοπωνύμια, ονόματα οργανισμών κλπ)
- Προβλήματα
  - Τα χειρωνακτικά σχεδιασμένα σετ χαρακτηριστικών είναι συνήθως υπερ-εξειδικευμένα, μη ολοκληρωμένα, χρονοβόρα στην σχεδίαση και την επικύρωσή τους
  - Δεν είναι αυτός ο τρόπος που μαθαίνει ο άνθρωπος.

Feature	NER
Current Word	✓
Previous Word	✓
Next Word	✓
Current Word Character n-gram	all
Current POS Tag	✓
Surrounding POS Tag Sequence	✓
Current Word Shape	✓
Surrounding Word Shape Sequence	✓
Presence of Word in Left Window	size 4
Presence of Word in Right Window	size 4

Finkel, 2010

Πώς μαθαίνει ο άνθρωπος;

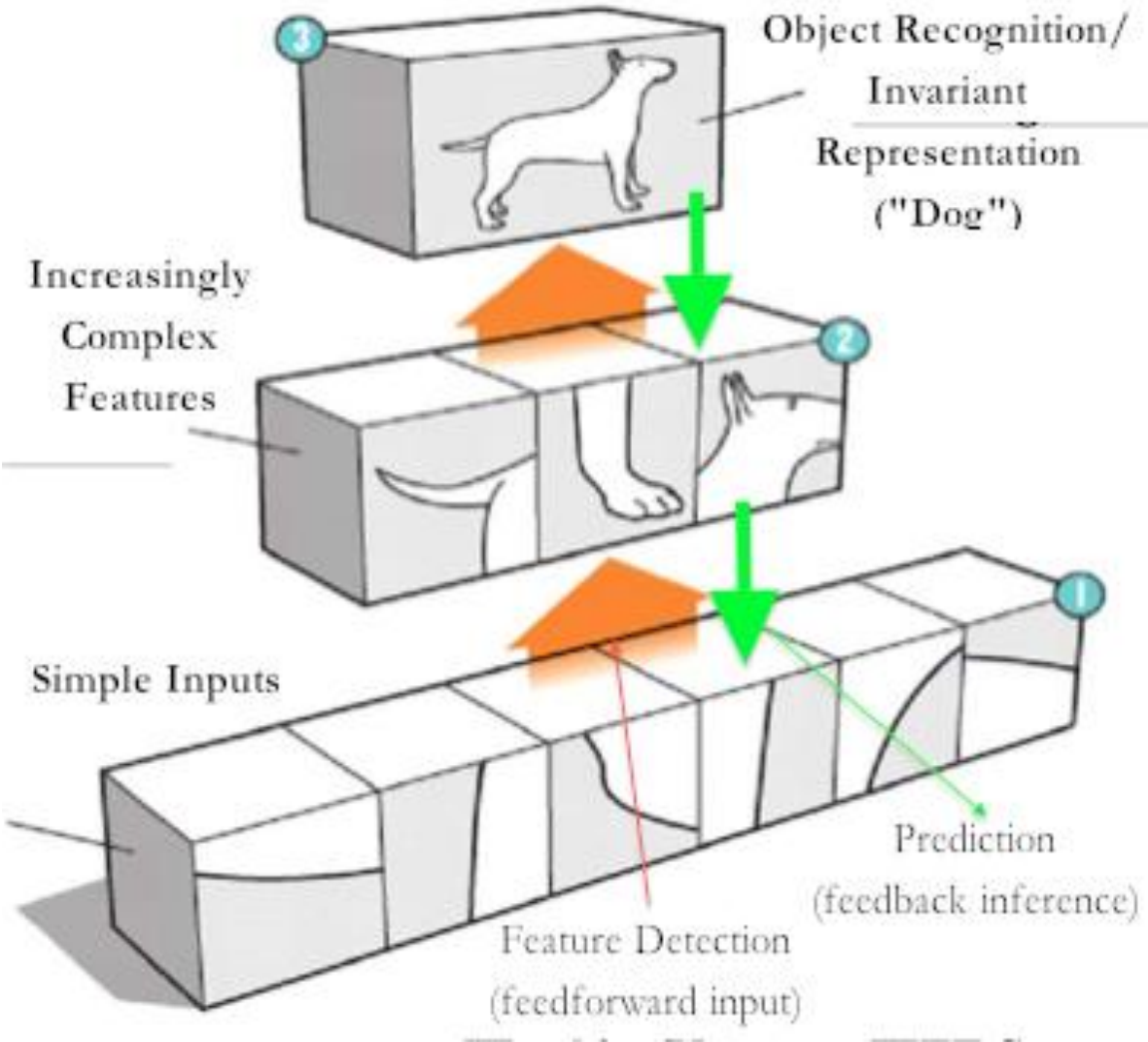
- Ο άνθρωπος μαθαίνει έννοιες και αποκτά ικανότητες και τις εφαρμόζει σε διαφορετικά προβλήματα
  - Μεταφορά (Transfer Learning)
- Ο άνθρωπος πρώτα μαθαίνει απλές έννοιες και μετά τις συνδυάζει για να μάθει πιο πολύπλοκες
- Υπάρχει η ένδειξη ότι ο πυρήνας του εγκεφάλου (cortex) έχει έναν μοναδικό αλγόριθμο μάθησης
  - Η είσοδος από τους οπτικούς νευρώνες κουναβιών δρομολογήθηκε στο τμήμα του πυρήνα που λαμβάνει την ακουστική είσοδο.
  - Είχαν την ικανότητα να μαθαίνουν να βλέπουν με αυτό το τμήμα του εγκεφάλου

Επομένως, αν θέλουμε έναν αλγόριθμο γενικό/καθολικό, αυτός θα πρέπει να:

- Μπορεί να δουλέψει με οποιοδήποτε τύπο δεδομένων
- Μπορεί να μαθαίνει από μη επισημειωμένα δεδομένα
- Μπορεί να εξάγει αυτόματα τα χαρακτηριστικά που χρειάζεται
  - Τα χαρακτηριστικά που μαθαίνονται αυτόματα, μαθαίνονται γρήγορα και προσαρμόζονται εύκολα
- Μπορεί να μεταφέρει αυτό που έμαθε σε καινούριες εφαρμογές/περιοχές
- Μπορεί να εφαρμόζει πολυπτροπική (multimodal) μάθηση
  - Να μαθαίνει ταυτόχρονα από διαφορετικές εισόδους (όραση, γλώσσα κλπ)

Ιεραρχική Αναπαράσταση

- Η αναπαράσταση του μικρόκοσμου της εφαρμογής μου γίνεται σε πολλαπλά επίπεδα
- Κάθε επίπεδο δημιουργεί καινούρια χαρακτηριστικά από συνδυασμό χαρακτηριστικών του προηγούμενου επιπέδου
- Κάθε επίπεδο είναι πιο «αφαιρετικό» (abstract) από το προηγούμενο επίπεδο

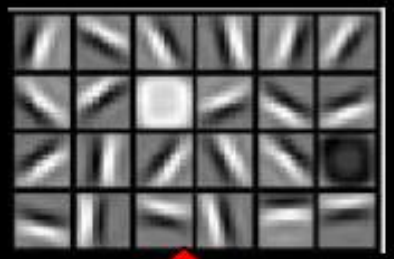


Hierarchical Sparse coding (Sparse DBN): Trained on face images

Training set: Aligned images of faces.

object models

object parts (combination of edges)



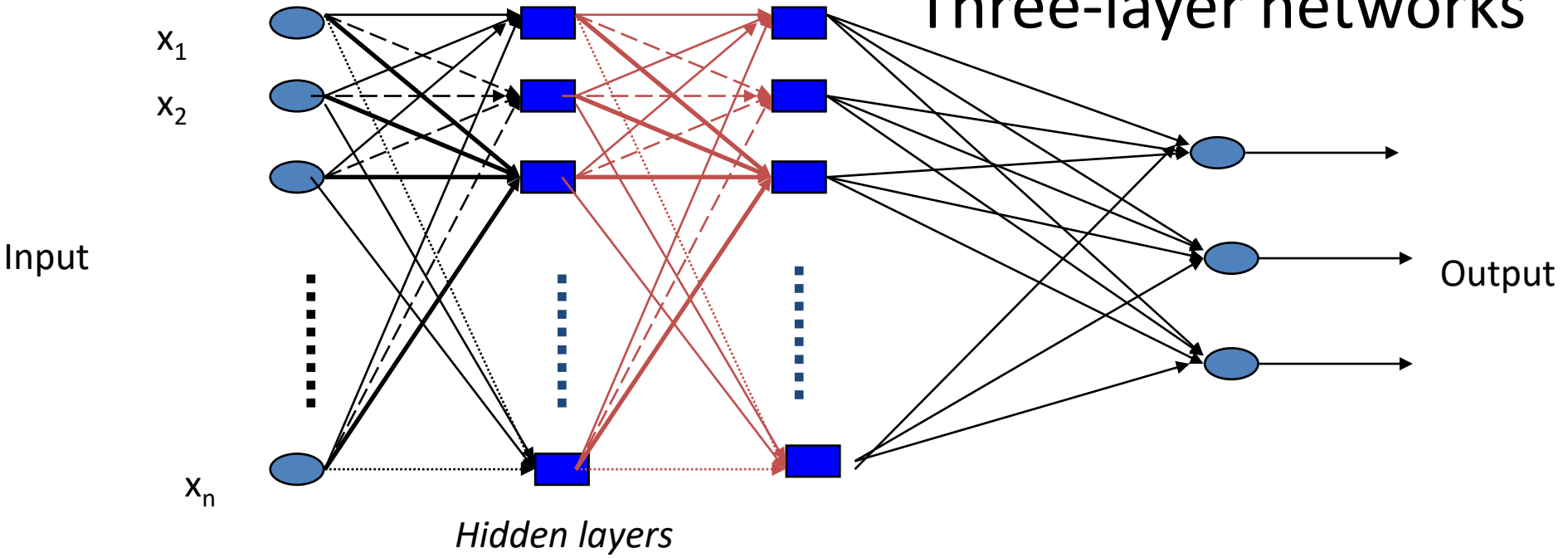
edges

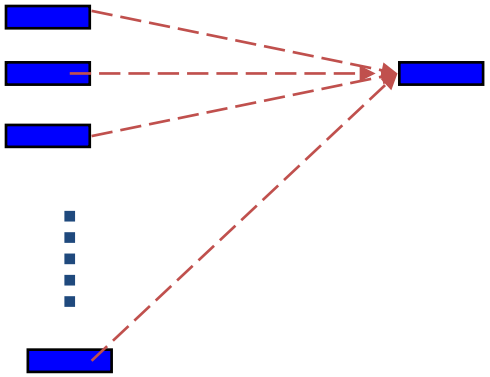
pixels

[Honglak Lee

Ρηχή Μάθηση:Feed-forward Neural Network

Three-layer networks





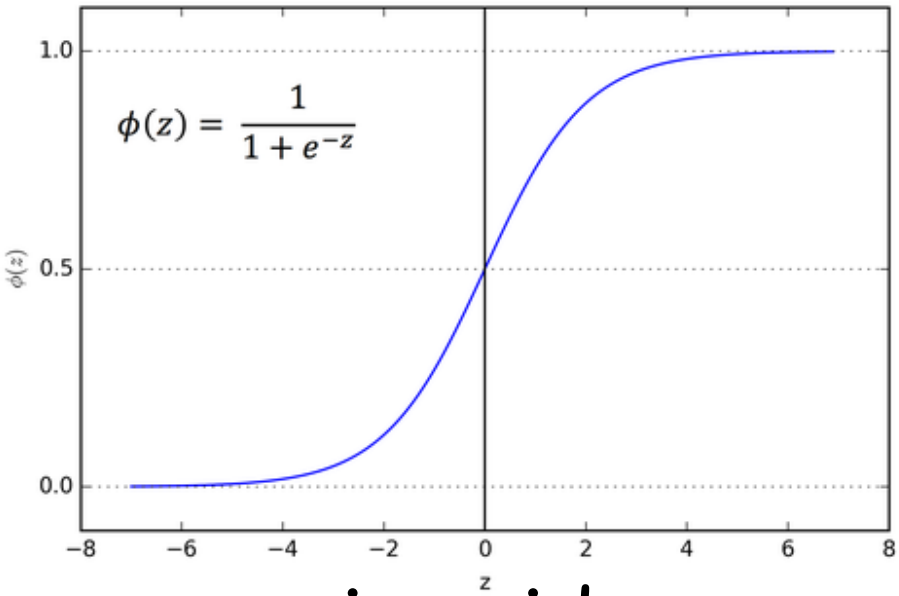
Activation function

$$y_i = f\left(\sum_{j=1}^{m} w_{ij} x_j + b_i\right)$$

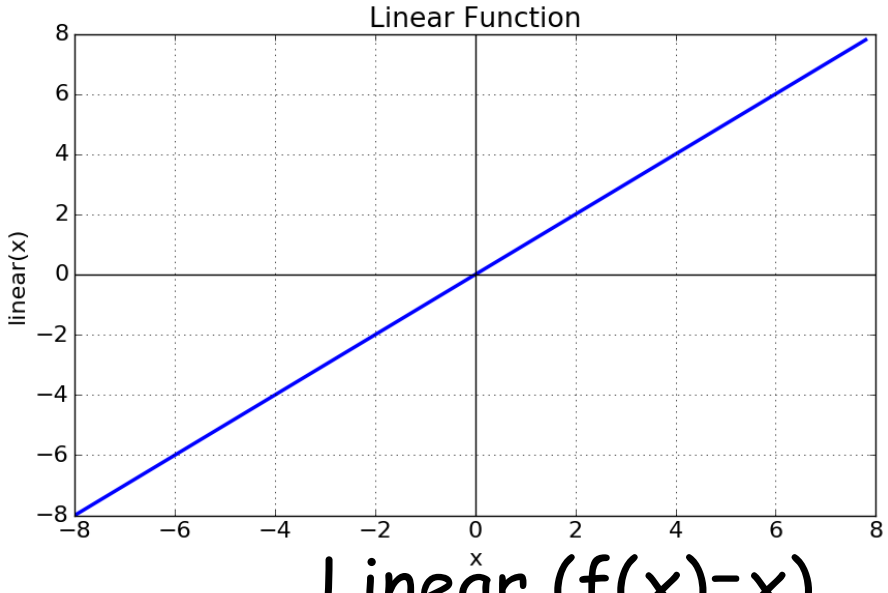
Ρηχή Μάθηση:Feed-forward Neural Network

- Σε ένα FFNN, σε κάθε κρυφό κόμβο
  - Υπολογίζεται το άθροισμα των σταθμισμένων με βάρη εισόδων του κόμβου
    - Τα βάρη αυτά είναι οι παράμετροι που πρέπει το δίκτυο να μάθει κατά την εκπαίδευσή του
  - Εφαρμόζεται στο άθροισμα αυτό μια συνάρτηση ενεργοποίησης (activation function), η οποία κανονικοποιεί το άθροισμα.
  - Η συνάρτηση αυτής μπορεί να είναι γραμμική ή μη γραμμική.

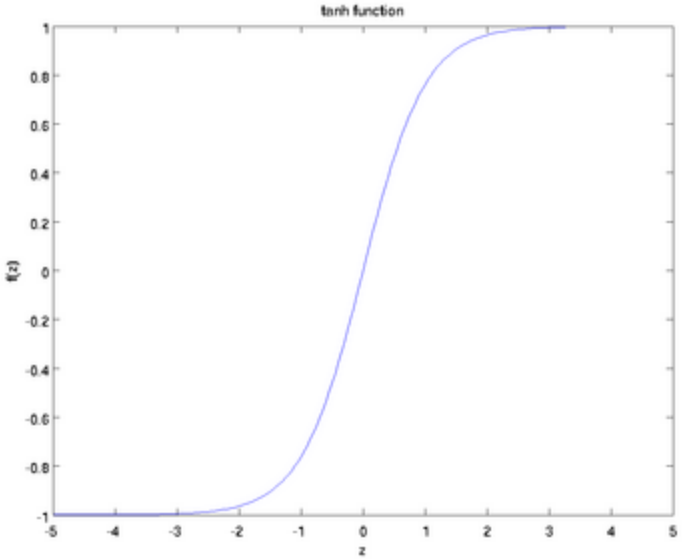
Activation functions



sigmoid



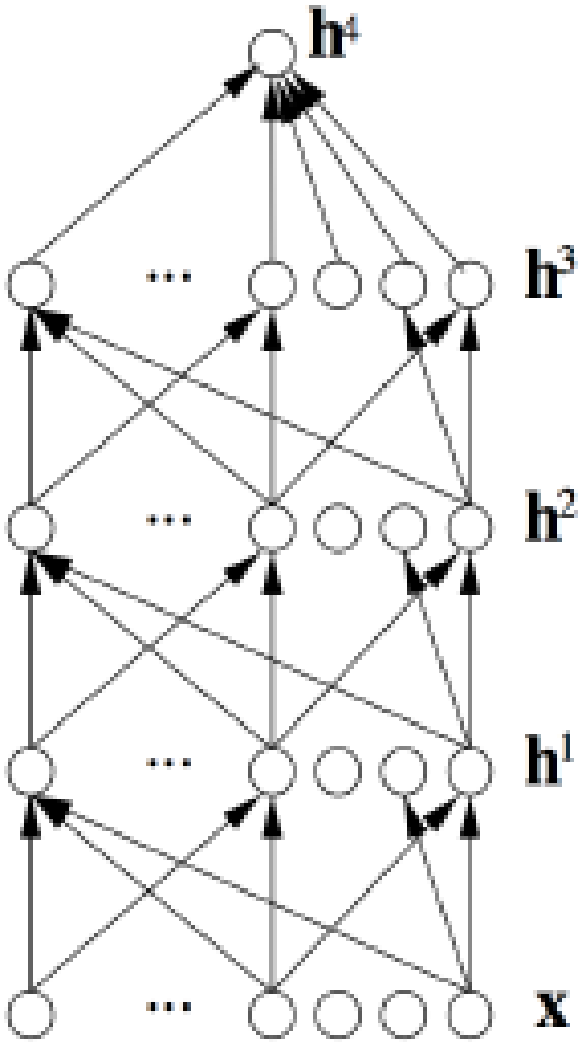
Linear (f(x)=x)



$$f(x) = tanh(x) = \frac{2}{1+e^{-2x}} - 1$$

Τι είναι Βαθιά Μάθηση (Deep Learning);

- Οι αλγόριθμοι βαθιάς μάθησης στοχεύουν να μάθουν πολλαπλά επίπεδα αναπαράστασης (χαρακτηριστικών) και μια έξοδο
- Από ένα σύνολο ωμών (raw) δεδομένων εισόδου **x** (πχ λέξεις)
- Η βασική οικογένεια αλγορίθμων είναι τα **νευρωνικά δίκτυα**



Γιατί Βαθιά Μάθηση;

- Οι αλγόριθμοι βαθιάς μάθησης
  - είναι και επιβλεπόμενοι και μη-επιβλεπόμενοι
  - Παρέχουν ένα καθολικό πλαίσιο για την πολυεπίπεδη αναπαράσταση γνώσης
  - Το 2006 ξεκίνησαν να αποδίδουν καλύτερα από την παραδοσιακή μάθηση
  - Σήμερα
    - Έχουν την δυνατότητα πρόσβασης σε περισσότερα δεδομένα
    - Έχουν πρόσβαση σε μεγαλύτερη υπολογιστική ισχύ
    - Στηρίζονται σε νέους αλγορίθμους και αρχιτεκτονικές

Deep Learning: Why for NLP ?

One Model rules them all ?

DL approaches have been successfully applied to:

- Automatic summarization
- Coreference resolution
- Discourse analysis
- Machine translation
- Morphological segmentation
- Named entity recognition (NER)
- Natural language generation
- Word sense disambiguation
- Relationship extraction
- Speech processing
- Part-of-speech tagging
- sentence boundary disambiguation
- Sentiment analysis
- Optical character recognition (OCR)
- Question answering
- Parsing
- Word segmentation
- Natural language understanding
- Information retrieval (IR)
- Speech recognition
- Topic segmentation and recognition
- Speech segmentation
- Information extraction (IE)

Βαθιά Μάθηση στην ΕΦΓ

- Μεγάλες βελτιώσεις τα τελευταία χρόνια
  - Στα διάφορα επίπεδα γλωσσολογικής πληροφορίας
    - φωνητικό
    - Μορφολογικό
    - Συντακτικό
    - Σημασιολογικό
  - Στις διάφορες εφαρμογές
    - Αυτόματη μετάφραση
    - Ανάλυση συναισθήματος
    - Συστήματα ερωταποκρίσεων

Διανύσματα λέξεων (Word vectors)

One-hot vectors

	abandon	ability	able	…	ants	…	zone
ants	0	0	0	0	1	0	0

Διανύσματα λέξεων (Word vectors)

"You shall know a word by the company it keeps" (J. R. Firth, 1957)

government debt problems turning into banking crises as has happened in

saying that Europe needs unified banking regulation to replace the hodgepodge

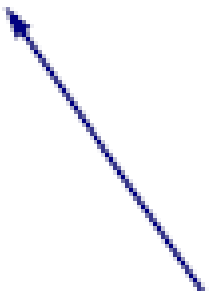
↖ These words will represent *banking* ↗

Μια από τις σημαντικότερες ιδέες της σύγχρονης στατιστικής ΕΦΓ

Διανύσματα λέξεων (Word vectors)

- Term-document matrices

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0



Boolean, συχνότητα, tf.idf

Διανύσματα λέξεων (Word vectors)

- Window-based co-occurrence matrices

  Example corpus:

  - I like deep learning.
  - I like NLP.
  - I enjoy flying.

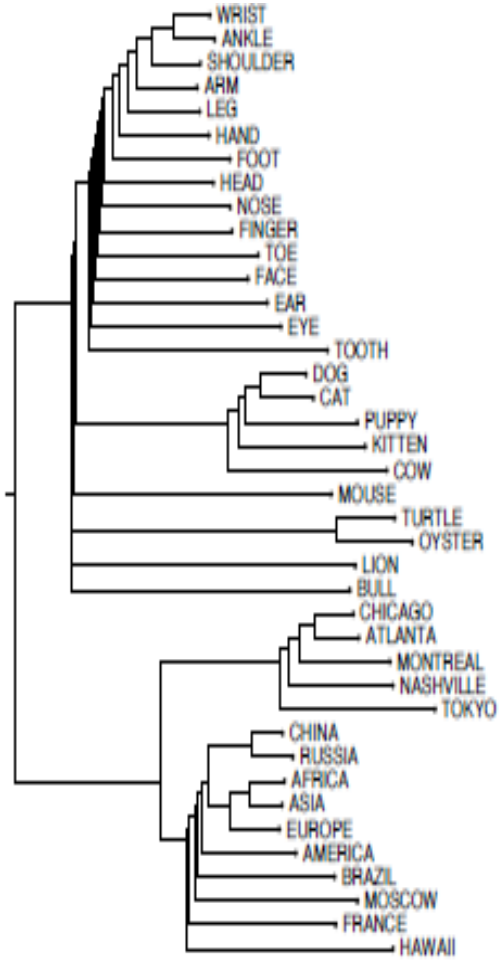
counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

Διανύσματα λέξεων (Word vectors)

- Οι πίνακες αυτοί
  - Αυξάνουν πολύ σε διαστατικότητα καθώς αυξάνεται το λεξικό
  - Απαιτούν μεγάλους πόρους για αποθήκευση
  - Έχουν προβλήματα πολύ αραιών εγγραφών (sparsity)
- Λύση
  - Κρατάμε μόνο την σημαντική πληροφορία μέσω ενός μικρού αριθμού διαστάσεων (25-1000 διαστάσεις)
    - Singular Value Decomposition
  - Οπότε κάθε λέξη αναπαρίσταται σαν ένα πυκνό διάνυσμα

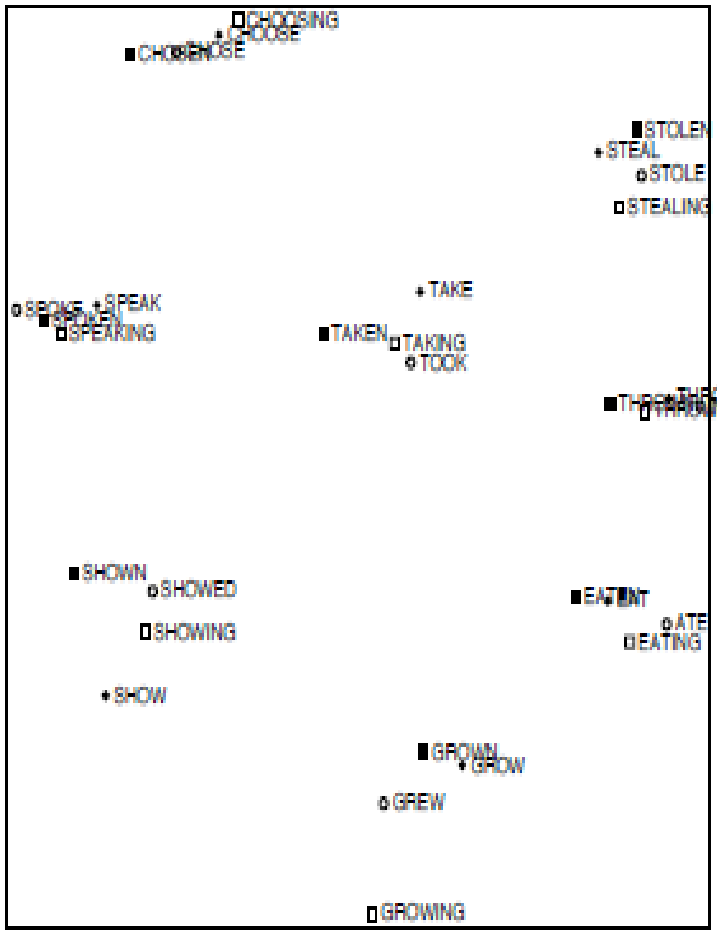
$$linguistics = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Ενδιαφέροντα αποτελέσματα



An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence
Rohde et al. 2005

- Σημασιολογικά



of Semantic Similarity Based on Lexical Co-Occurrence

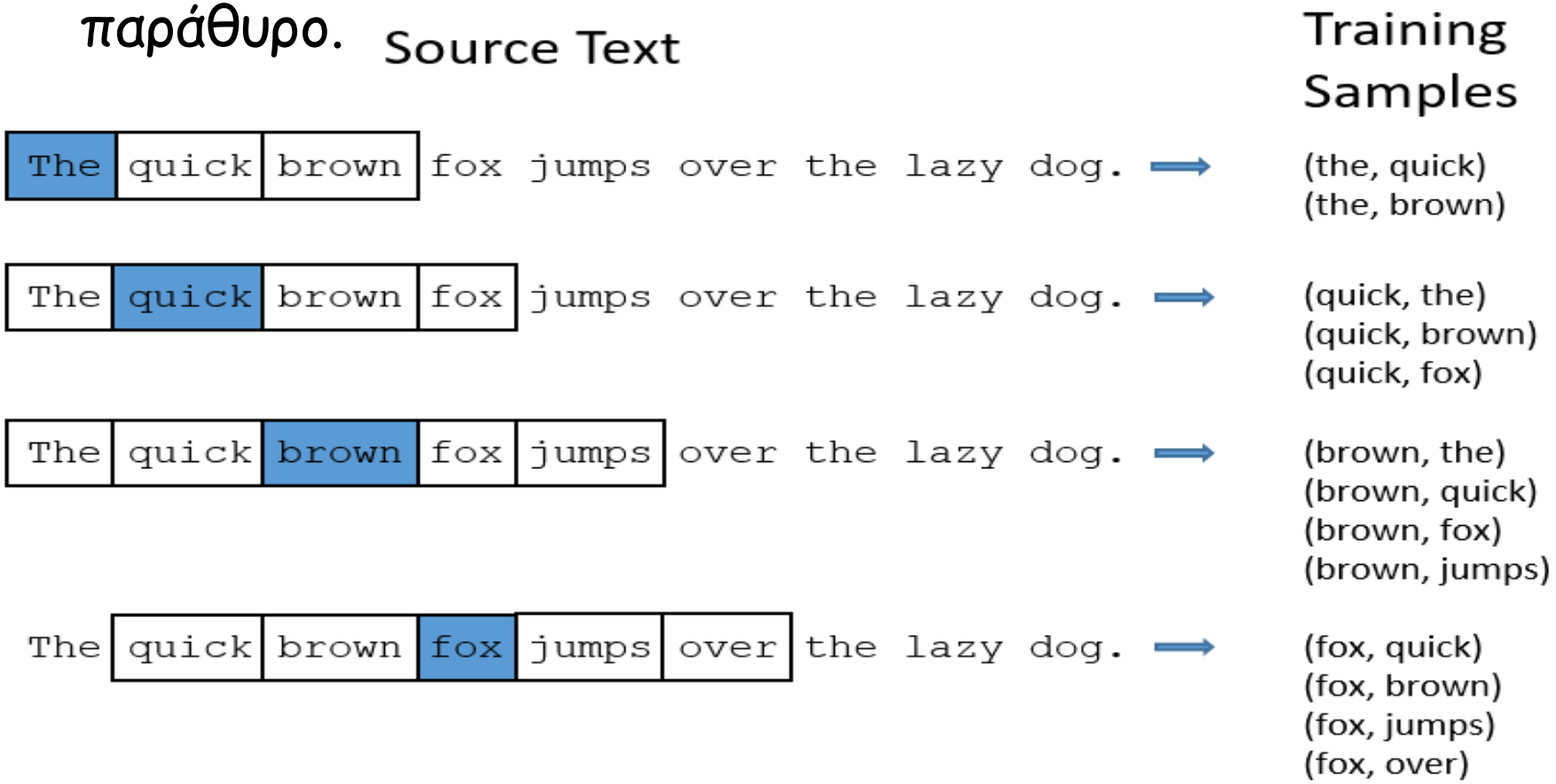
- Συντακτικά

Διανύσματα λέξεων (Word vectors)

- Προβλήματα με την μείωση διαστατικότητας
  - Πολύ μεγάλη υπολογιστική πολυπλοκότητα
  - Δύσκολα ενσωματώνονται καινούριες λέξεις ή έγγραφα
- Λύση
  - Αυτόματη μάθηση διανυσμάτων λέξεων χαμηλής διαστατικότητας
  - **Word2vec** (Mikolov et al. 2013)
    - Ρηχά μοντέλα που χρησιμοποιούνται για την παραγωγή word embeddings

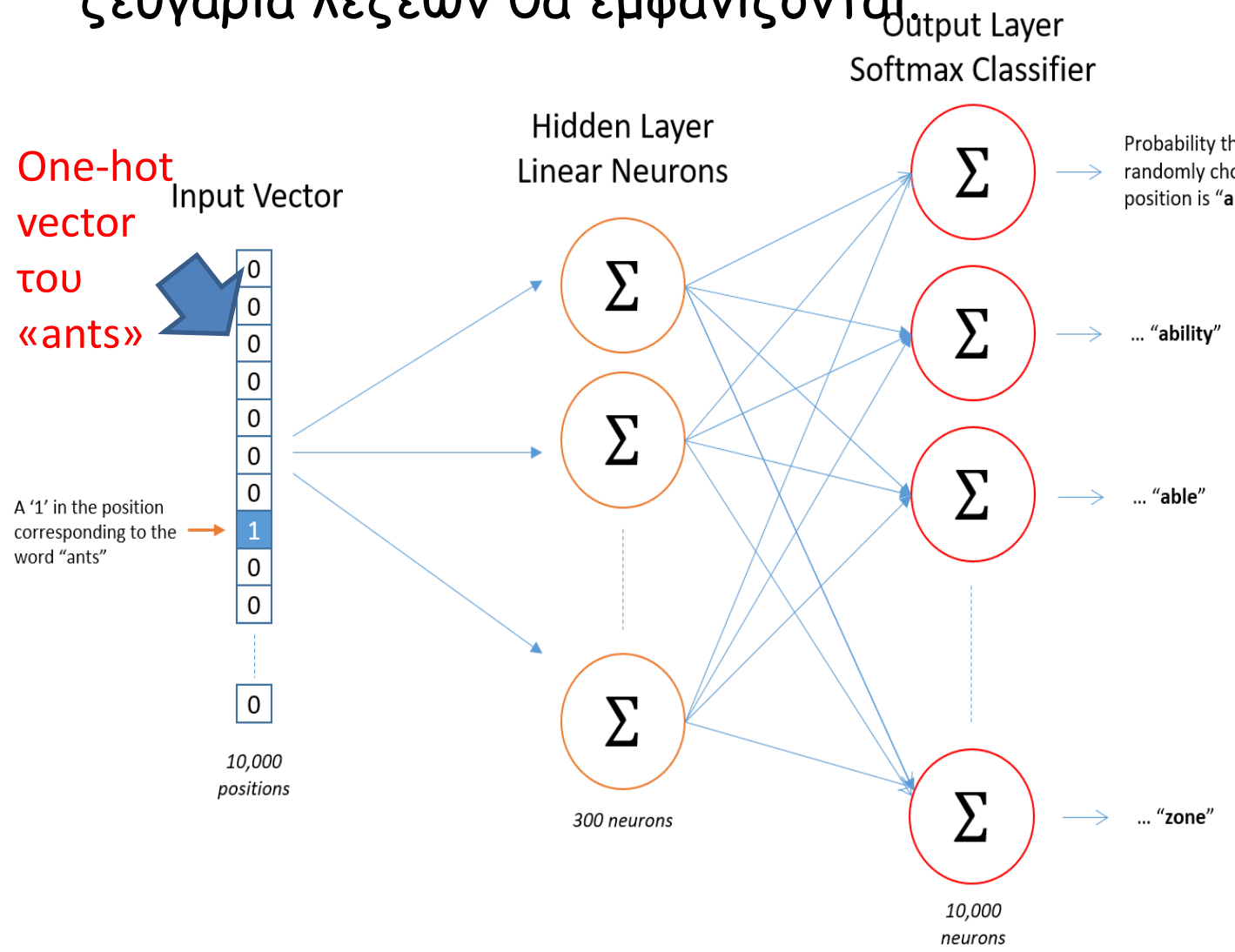
Word2vec: Training

- Δεδομένης μιας λέξης (input word) και ενός συγκεκριμένου παραθύρου συμφραζομένων (εδώ 2) φτιάξε παραδείγματα με ζεύγη λέξεων, όπου σε κάθε ζεύγος η μια λέξη θα είναι το input word (μπλε) και η άλλη θα είναι μια λέξη μέσα στο παράθυρο.



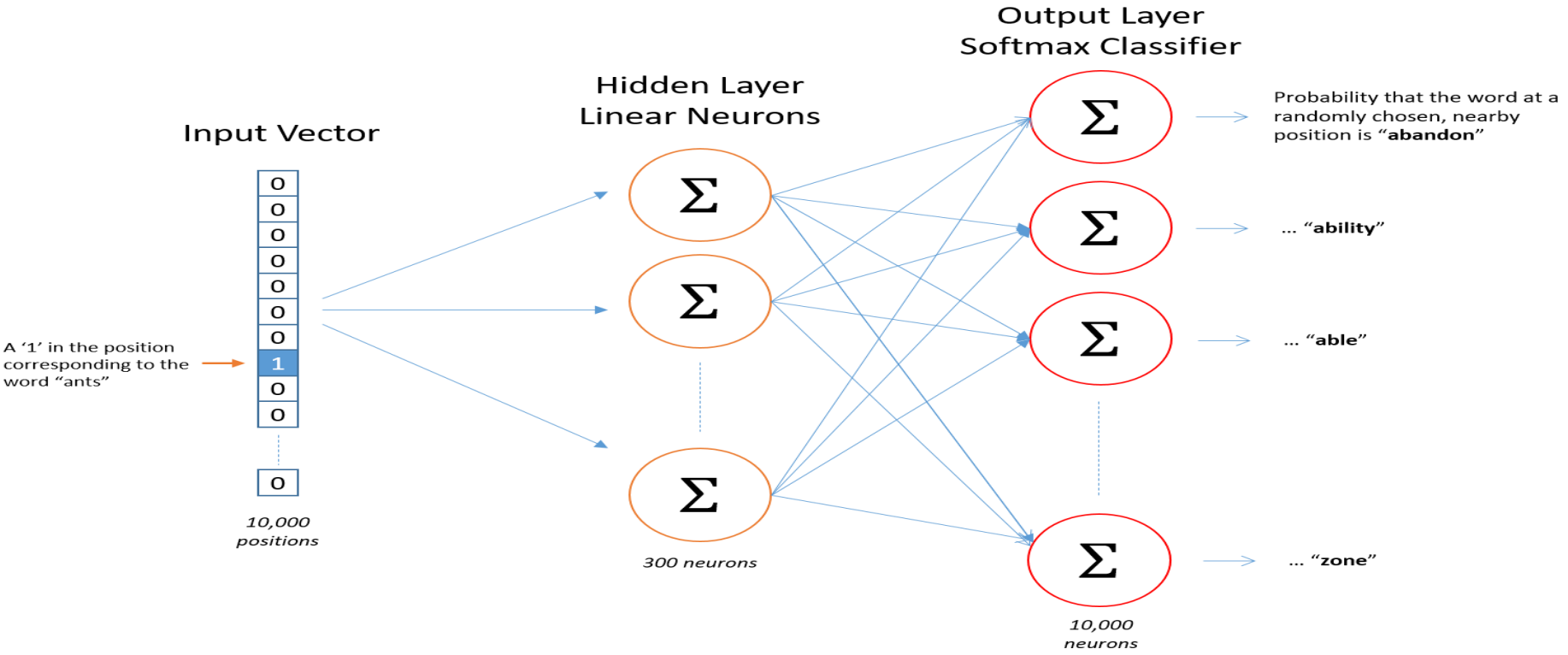
Word2vec: Training

- Το νευρωνικό δίκτυο θα εκπαιδευτεί από αυτά τα παραδείγματα βάσει του αριθμού των φορών που αυτά τα ζευγάρια λέξεων θα εμφανίζονται.



Η έξοδος κάθε κόμβου του τελευταίου επιπέδου θα είναι ένα one-hot vector της δεύτερης λέξης.

Word2vec: Training



Το κρυμμένο επίπεδο λειτουργεί σαν ένας πίνακας με 10,000 γραμμές (μια για κάθε λέξη του λεξικού), και πχ 300 στήλες (μια για κάθε νευρώνα του επιπέδου). Ο αριθμός των νευρώνων του επιπέδου είναι ερευνητικό ζητούμενο και εξαρτημένος εφαρμογής.

Στόχος της εκπαίδευσης είναι να «μαθευτούν» οι τιμές (βάρη) στα κελιά αυτού του πίνακα.

Word2vec Testing: Hidden Layer

$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

1 x 10,000 — 10,000 x 300 — 1 x 300

είσοδος — κρυφό επίπεδο — έξοδος κρυφού επιπέδου

Η έξοδος του κρυμμένου επιπέδου είναι το διάνυσμα της λέξης εισόδου.

Word2vec Testing: Output Layer

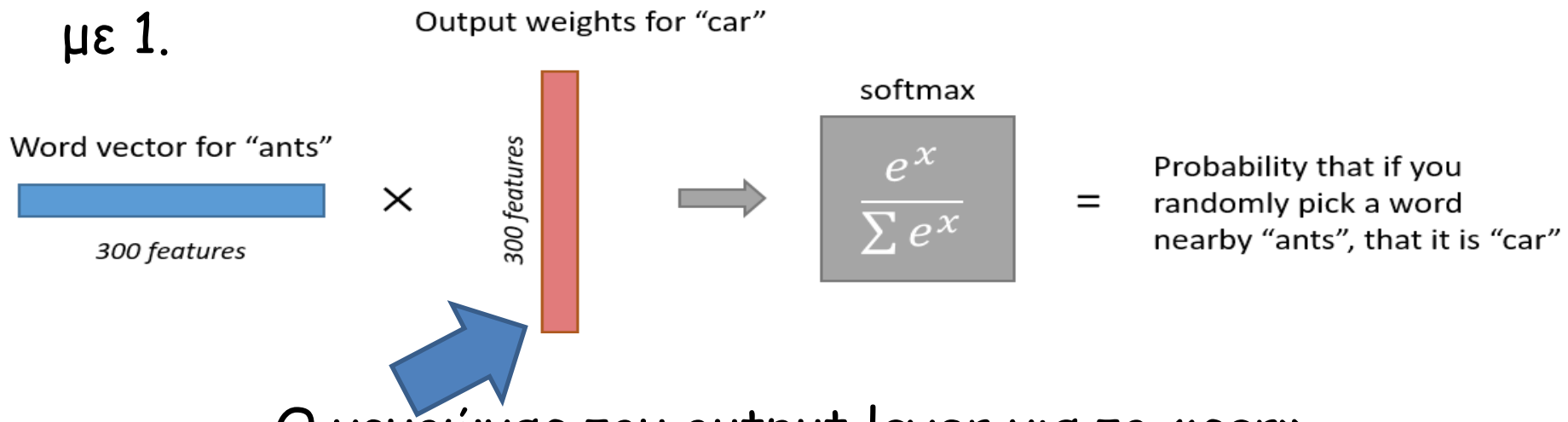
Η έξοδος του κρυμμένου επιπέδου, δηλ, το 1x300 διάνυσμα της λέξης "ants" εισάγεται στο output layer.

Στο output layer υπάρχει ένας νευρώνας για κάθε λέξη του λεξικού.

Ο κάθε νευρώνας θα παράξει μια έξοδο, η οποία μέσω της συνάρτησης softmax() θα πάρει τιμή στο διάστημα [0,1] .

Η τιμή αυτή είναι η πιθανότητα η συγκεκριμένη λέξη του λεξικού να βρίσκεται κοντά στο «ants»

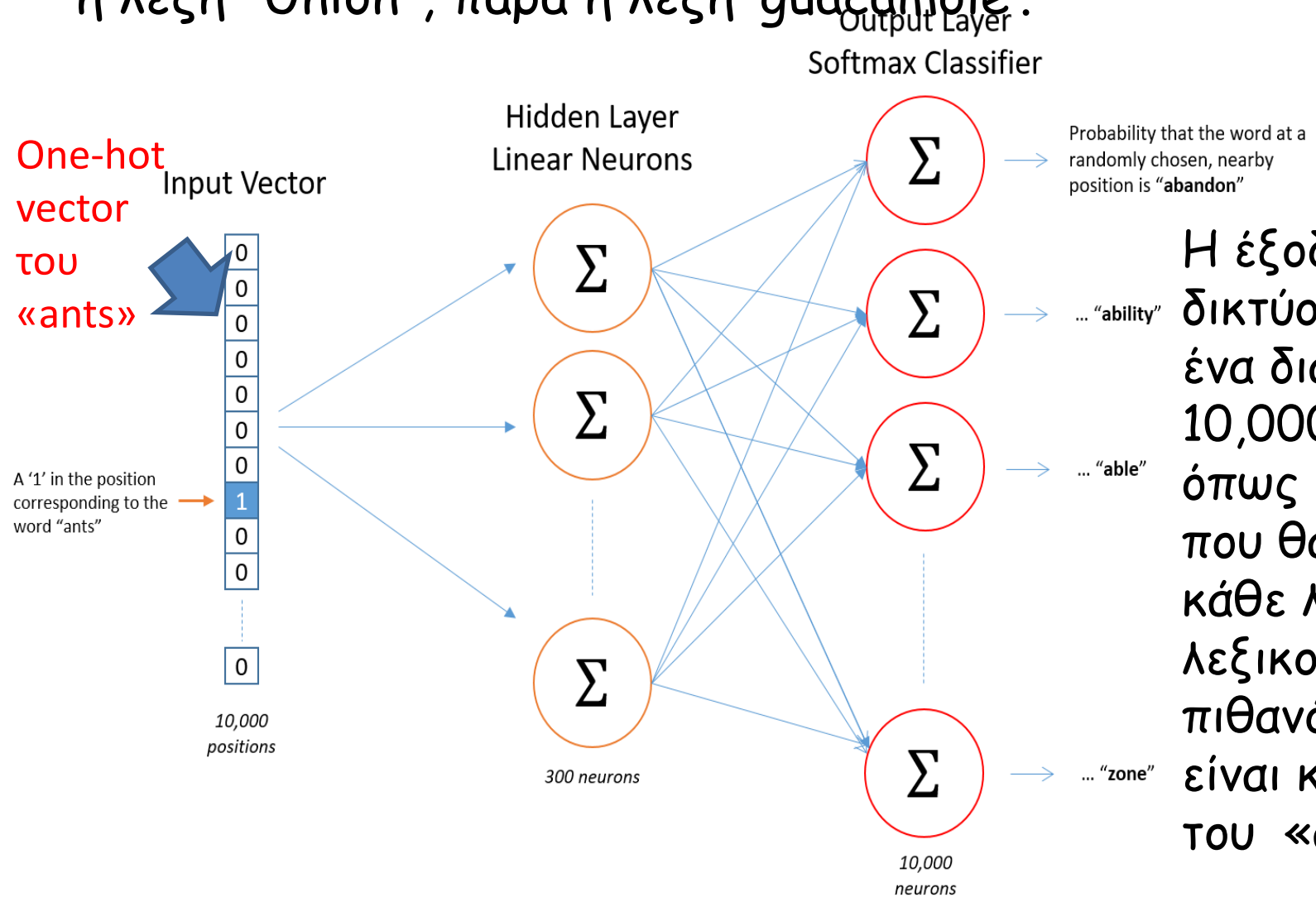
Το άθροισμα όλων αυτών των εξόδων (πιθανοτήτων) θα ισούται με 1.



Ο νευρώνας του output layer για το «car»

Word2vec: Testing

- Το νευρωνικό δίκτυο, αφού εκπαιδευτεί, θα έχει μάθει ότι πχ είναι πολύ πιο πιθανό κοντά στην λέξη «Soviet» να εμφανιστεί η λέξη "Union", παρά η λέξη 'guacamole'.



Η έξοδος του δικτύου θα είναι ένα διάνυσμα 10,000 θέσεων, όπως η είσοδος, που θα έχει για κάθε λέξη του λεξικού την πιθανότητα να είναι κοντινή λέξη του «ants»

Word2vec: παράδειγμα εξόδου

- Here's a list of words associated with "Sweden" using Word2vec, in order of proximity:

Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408

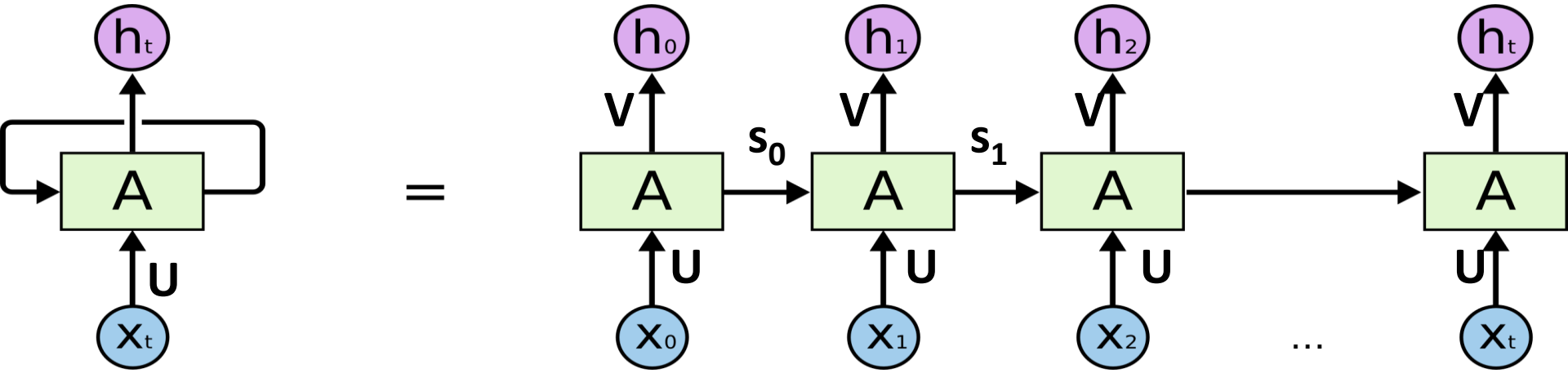
Βαθιές Αρχιτεκτονικές

- Είδη νευρωνικών δικτύων
  - Recursive neural networks (RNN)
    - Recurrent neural networks
    - Long short term memory neural networks (LSTM)
  - Convolutional neural networks (CNN)
  - Sequence-to-sequence models

Φαινόμενα ακολουθίας στην γλώσσα (Sequential data)

- Για την αποσαφήνιση του νοήματος σε προτάσεις, λέξεις, χαρακτήρες χρειάζονται τα συμφραζόμενά τους.
- Μηχανική Μετάφραση
  - Μια λέξη έχει διαφορετικό νόημα ανάλογα με τα συμφραζόμενά της
- Ανάλυση Συναισθήματος
  - Η εμφάνιση επιρρημάτων και λέξεων άρνησης (όπως "very", "not", και "a bit too") στα συμφραζόμενα της λέξης που κρύβει το συναίσθημα επηρεάζουν την ένταση, την πόλωση ή την αντιστροφή του συναισθήματος.
- Διαλογικά συστήματα
  - Το επόμενο βήμα σε έναν διάλογο καθορίζεται από τα προηγούμενα βήματα του διαλόγου και τον στόχο που έχει ο διάλογος.
- Tokenization
  - Οι προηγούμενοι και οι επόμενοι χαρακτήρες χρησιμοποιούνται για να αναγνωριστεί η έναρξη μιας καινούριας λέξης.

Recurrent NNS



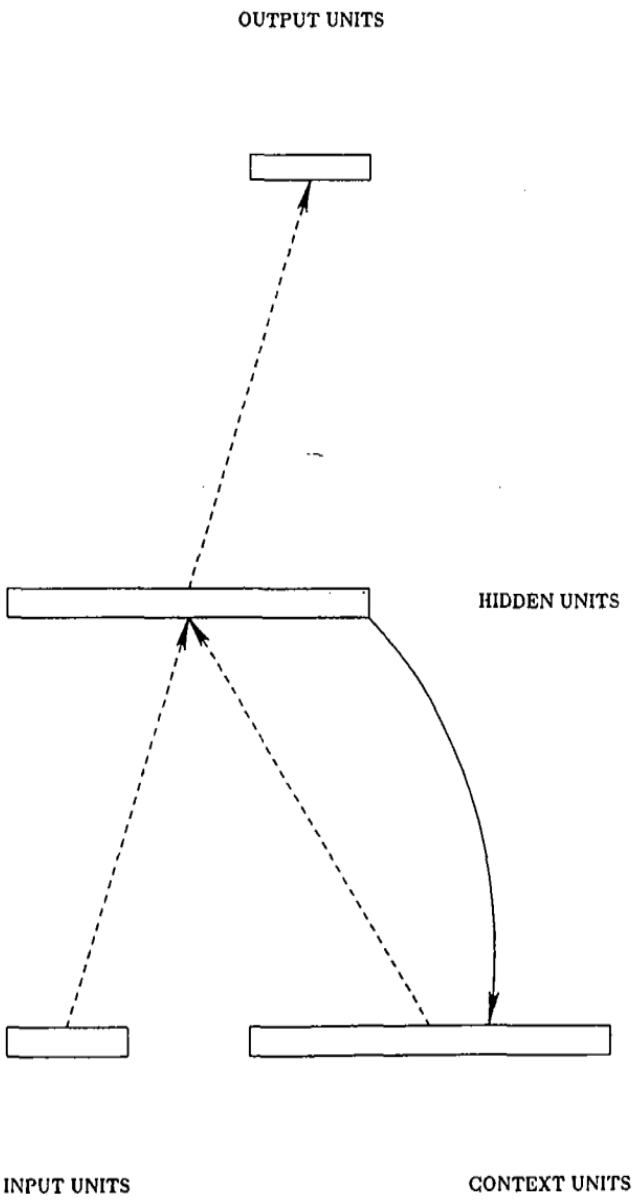
Ένα RNN είναι μια αλυσίδα αντιγράφων του ίδιου δικτύου. Οι ίδιες συνάψεις, τα ίδια βάρη, εφαρμόζονται σε καινούρια είσοδο (πχ καινούρια λέξη) σε κάθε χρονικό βήμα.
Τα RNNs συνδυάζουν την τρέχουσα είσοδο με την κατάσταση του προηγούμενου βήματος σε μια συνάρτηση η οποία παράγει την καινούρια τρέχουσα κατάσταση.

$$s_t = f(Ux_t + Ws_{t-1})$$

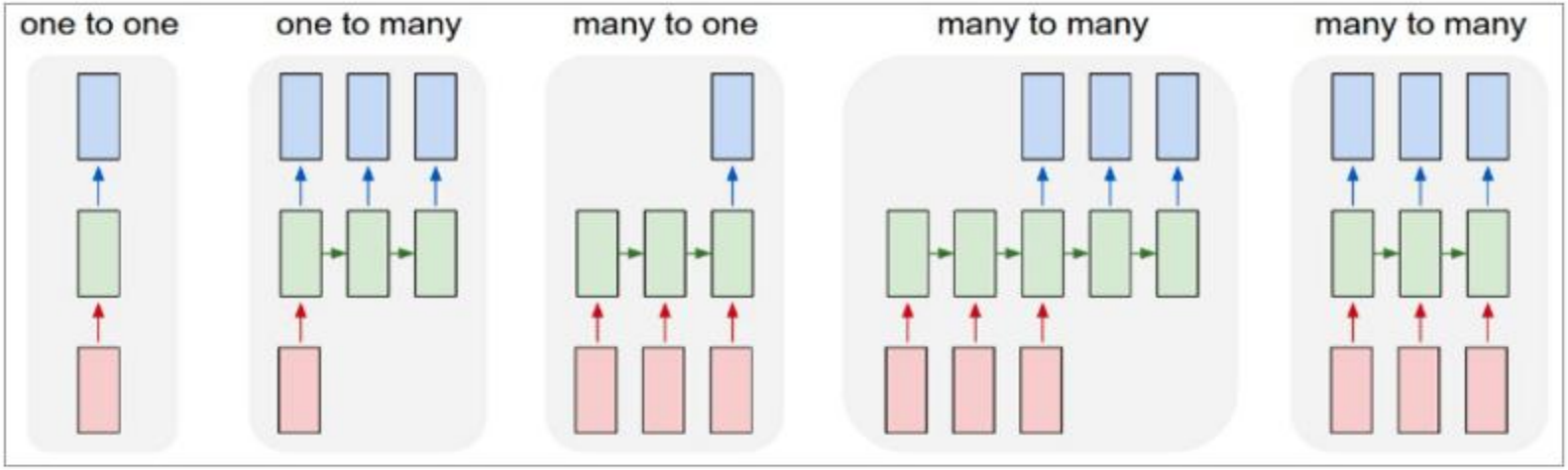
h_t= softmax(Vs_t)

Recurrent NNS

- Το πλεονέκτημα των RNN είναι η ικανότητά τους να αντιμετωπίζουν ακολουθιακά δεδομένα, χάρη στη «μνήμη» τους. Ενώ τα νευρωνικά δίκτυα δεν έχουν αίσθηση του χρόνου, και η πρόβλεψή τους εξαρτάται από την τωρινή τους είσοδο μόνο, τα RNNs λαμβάνουν υπόψη τους και την τωρινή είσοδο και την «είσοδο συμφραζομένων» ("context unit"), η οποία «χτίζεται» βάσει των όσων έχουν δει προηγούμενα.
- Έτσι, η πρόβλεψη που πραγματοποιείται την στιγμή Τ επηρεάζεται από αυτήν που πραγματοποιήθηκε την στιγμή Τ-1.



Αρχιτεκτονικές RNN



- Vanilla mode of processing without RNN, from fixed-sized input to fixed-sized output (e.g. image classification).
- Sequence output (e.g. image captioning takes an image and outputs a sentence of words).
- Sequence input (e.g. sentiment analysis where a given sentence is classified as expressing positive or negative sentiment)
- Sequence input and sequence output (e.g. Machine Translation: an RNN reads a sentence in English and then outputs a sentence in French).
- Synced sequence input and output (e.g. video classification where we wish to label each frame of the video).

Trigram RNN για POS tagging

Here, $h^{(t)}$ not only depends on the previous hidden state $h^{(t-1)}$, but also directly depends on $h^{(t-2)}$. We hope that this extra dependency can help to catch longer windows in the sentence.

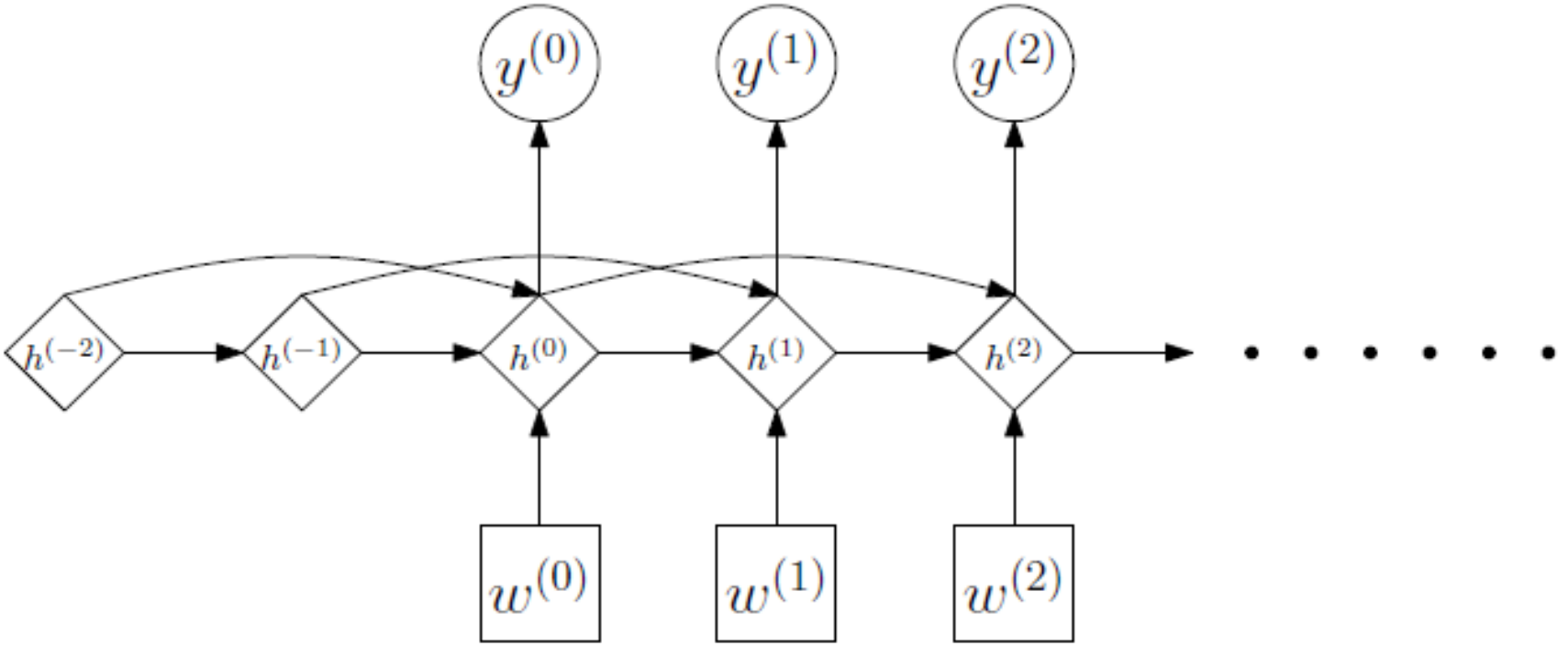
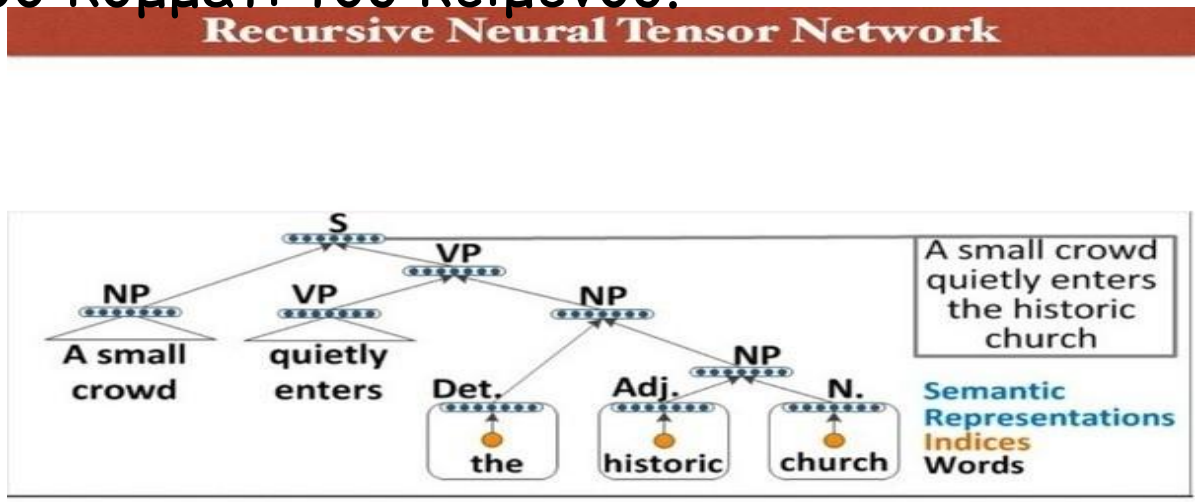


Figure 2: Trigram RNN

- https://cs224d.stanford.edu/reports/QinLonglu.pdf

Recursive NNS

- Στα Recursive NNs δεν έχω την έννοια των χρονικών βημάτων.
- Είναι ιεραρχικά δίκτυα στα οποία η είσοδος πρέπει να υποστεί ιεραρχική επεξεργασία σε μορφή δενδρικής δομής.
- Στην παρακάτω εικόνα φαίνεται πώς ένα recursive NN μαθαίνει το συντακτικό δέντρο μιας πρότασης παίρνοντας αναδρομικά την έξοδο της πράξης που πραγματοποιήθηκε σε ένα μικρότερο κομμάτι του κειμένου.



Long short term memory NNS

- Στα Recurrent NNs το να λάβω υπόψη πολλά χρονικά βήματα πριν μπορεί να
  - Προκαλέσει exploding gradients (αύξηση βαρών πολύ απότομη λόγω του επαναλαμβανόμενου πολλαπλασιασμού των βαρών που είναι μεγαλύτερα της μονάδας)
  - Προκαλέσει vanishing gradients (μείωση βαρών πολύ απότομη λόγω του επαναλαμβανόμενου πολλαπλασιασμού των βαρών που είναι μικρότερα της μονάδας)
- Για αυτό έχουν προταθεί τα Long short term memory NNS
- Είναι RNNs που περιλαμβάνουν ένα κύτταρο (LSTM unit)
- Επιτρέπουν στα RNNs να μαθαίνουν για πολλά χρονικά βήματα (πάνω από 1000).
- Το κύτταρο ρυθμίζει την διέλευση της πληροφορίας μέσα στο δίκτυο. Αποφασίζει τι θα αποθηκεύσει, τι θα διαβαστεί, τι θα διαγραφεί μέσω πυλών που ανοιγοκλείνουν.

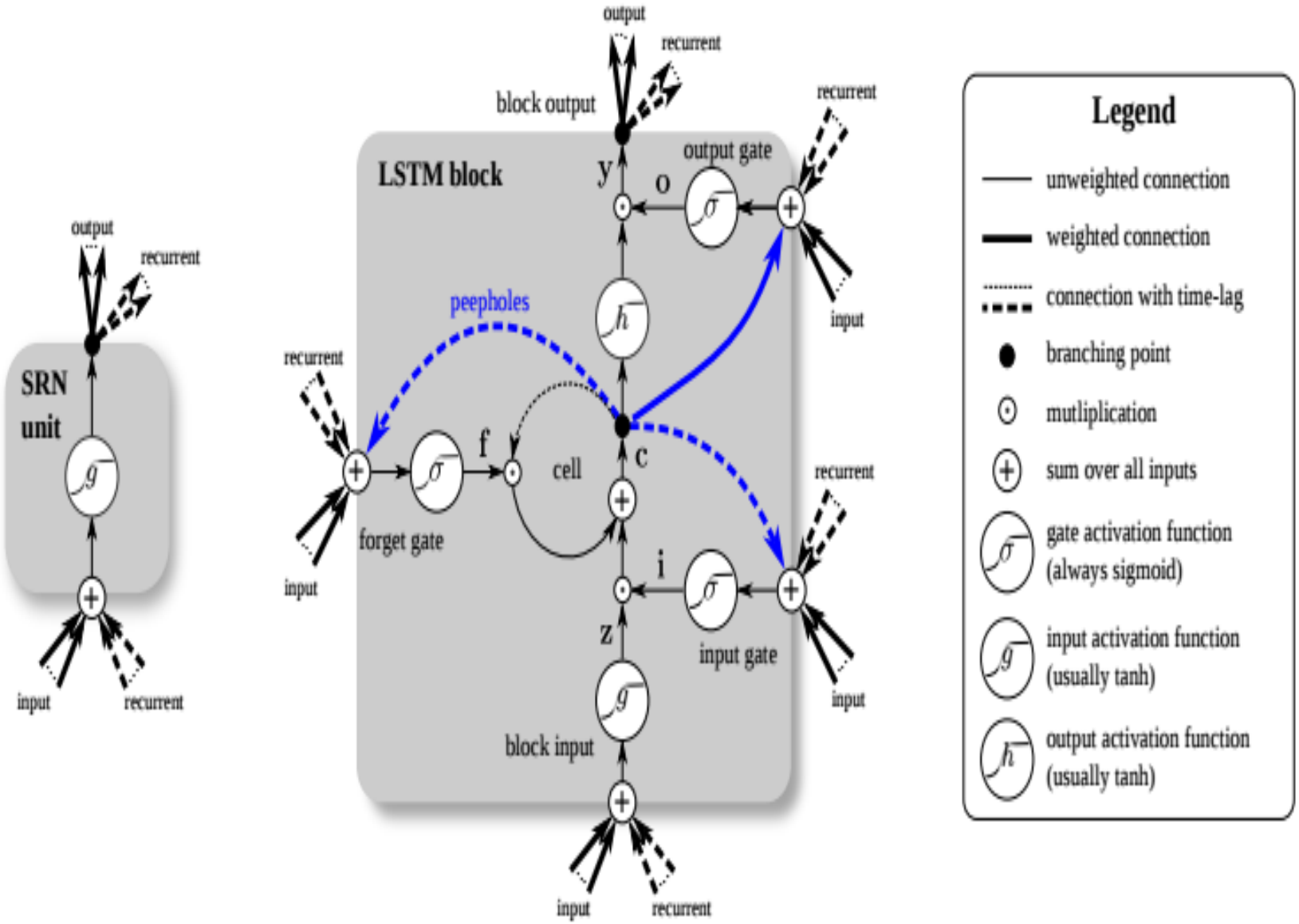


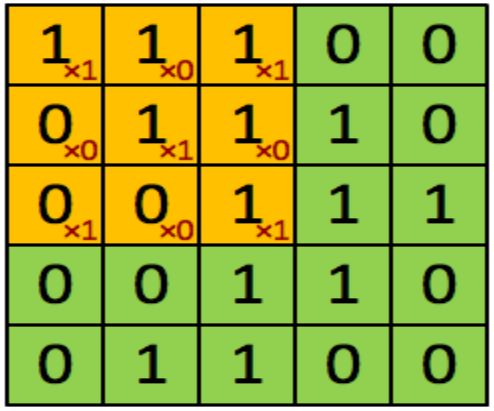
Figure 1. Detailed schematic of the Simple Recurrent Network (SRN) unit (left) and a Long Short-Term Memory block (right) as used in the hidden layers of a recurrent neural network.

Convolution – Συνέλιξη

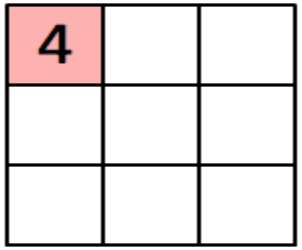
Το sliding window

1	**0**	**1**
0	1	0
1	0	1

ονομάζεται πυρήνας (kernel), φίλτρο (filter), ή ανιχνευτής χαρακτηριστικών (*feature detector*). Εδώ χρησιμοποιείται ένα φίλτρο 3×3, οι τιμές του πολλαπλασιάζονται κελί-κελί με τον αρχικό πίνακα, και αθροίζονται. Για την πλήρη συνέλιξη πραγματοποιείται αυτό για κάθε στοιχείο, «τσουλώντας» το φίλτρο πάνω από όλον τον αρχικό πίνακα.

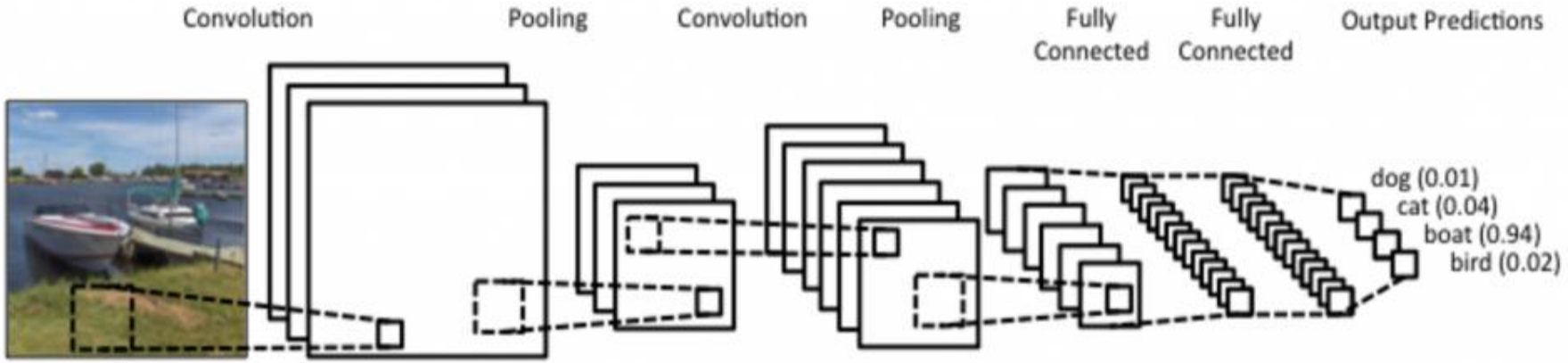


Image



Convolved Feature

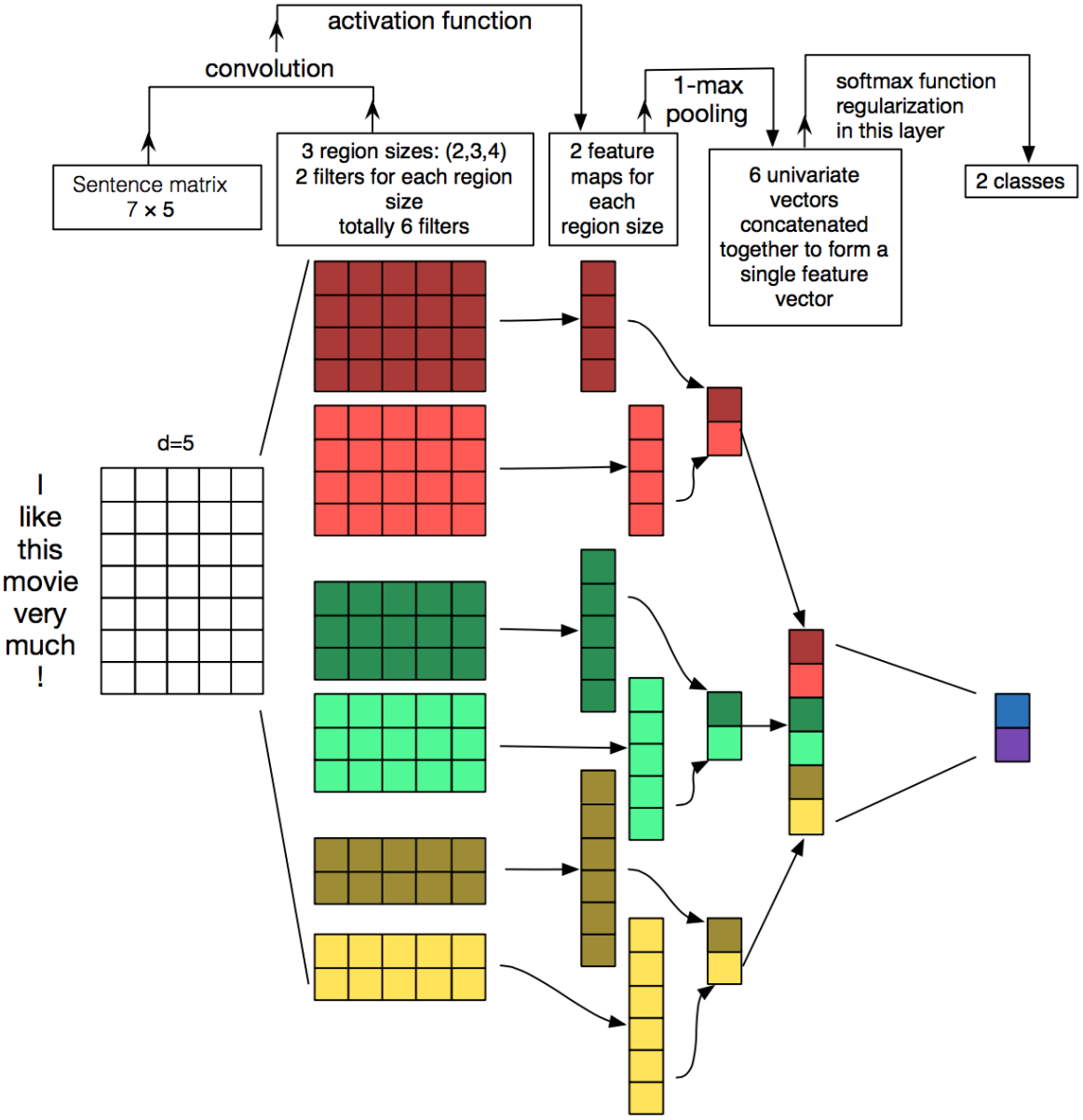
Convolutional NNs



Σε ένα convolutional layer (CL) δεν έχω πλήρεις συνδέσεις, αλλά πραγματοποιώ συνελίξεις πάνω στο επίπεδο εισόδου και η έξοδος του CL είναι το αποτέλεσμα της συνέλιξης.
Δημιουργούνται μόνο τοπικές συνδέσεις, όπου κάθε περιοχή της εισόδου συνδέεται με έναν νευρώνα της εξόδου.
Κατά την εκπαίδευση μαθαίνονται οι τιμές του φίλτρου.
Για παράδειγμα, στην ταξινόμηση εικόνων, σε ένα πρώτο CL μαθαίνεται η αναγνώριση ακμών, σε δεύτερο CL χρησιμοποιούνται οι ακμές για να μάθει το δίκτυο απλά σχήματα, και σε τρίτο CL χρησιμοποιούνται τα σχήματα για να μάθει το δίκτυο πιο υψηλού επιπέδου στοιχεία της εικόνας, όπως πχ σχήματα ανθρώπινου προσώπου.

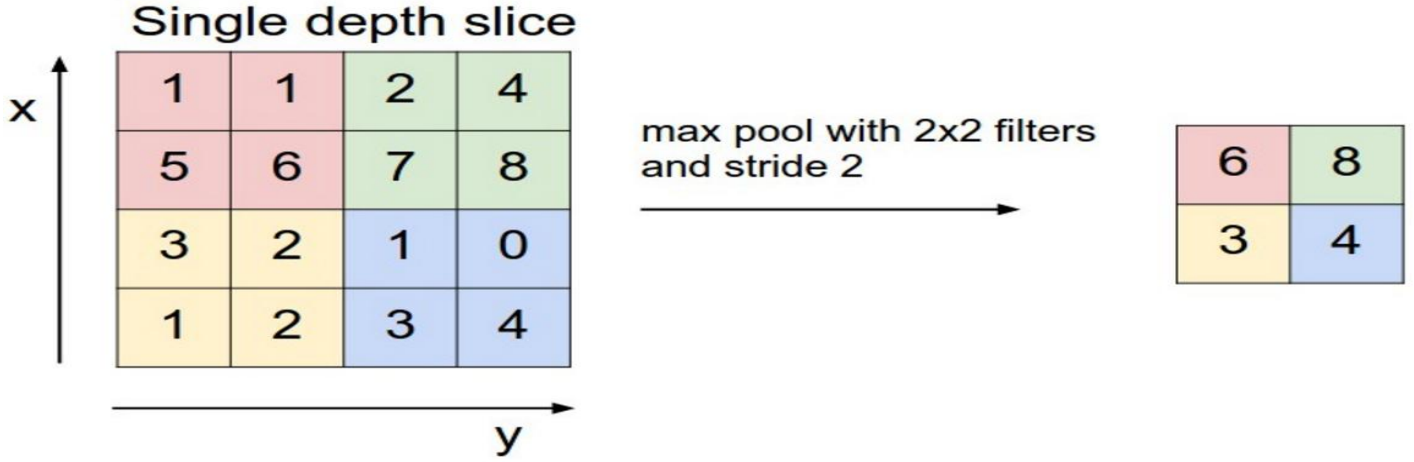
Convolutional NNs στην ΕΦΓ

Εδώ ο αρχικός μου πίνακας δεν είναι τιμές pixel, αλλά γραμμές από word embeddings Κάθε γραμμή είναι το διάνυσμα μιας λέξης σε μια πρόταση/κείμενο.



Convolutional NNs στην ΕΦΓ - Pooling

Max pooling for a 2×2 window (in NLP we typically apply pooling over the complete output, yielding just a single number for each filter.



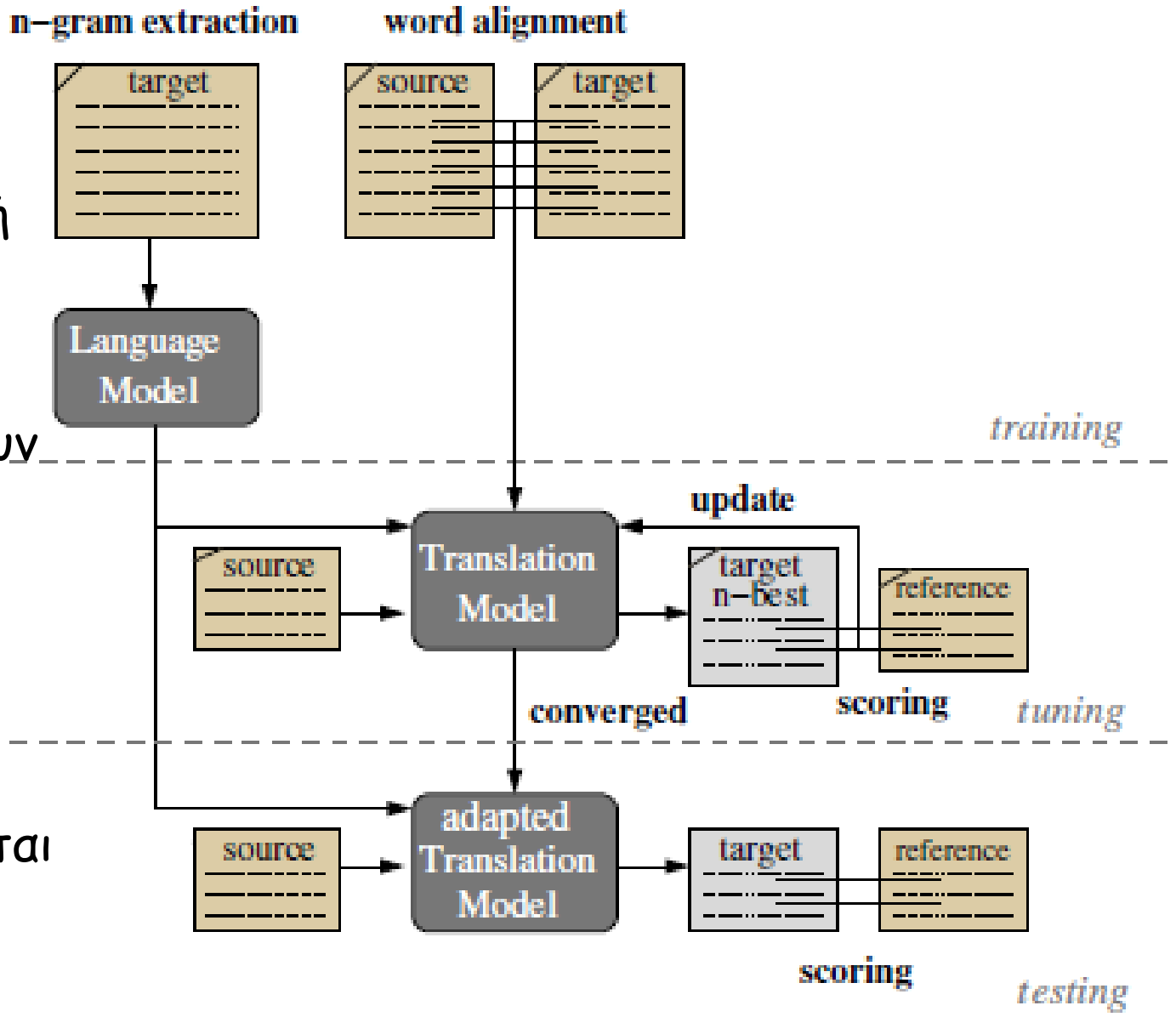
Με το pooling κάνω υποδειγματοληψία της εξόδου του CL. Με αυτό πετυχαίνω:

- Σταθερού μεγέθους έξοδο (αν πχ θέλω ταξινόμηση σε κλάση δύο τιμών μπορώ να ρυθμίσω της έξοδο ώστε να είναι διάνυσμα δύο θέσεων), ανεξάρτητα από το μέγεθος της εισόδου (το μήκος των προτάσεών μου)
- Να μειώνω την διαστατικότητά μου, διατηρώντας την σημαντική πληροφορία

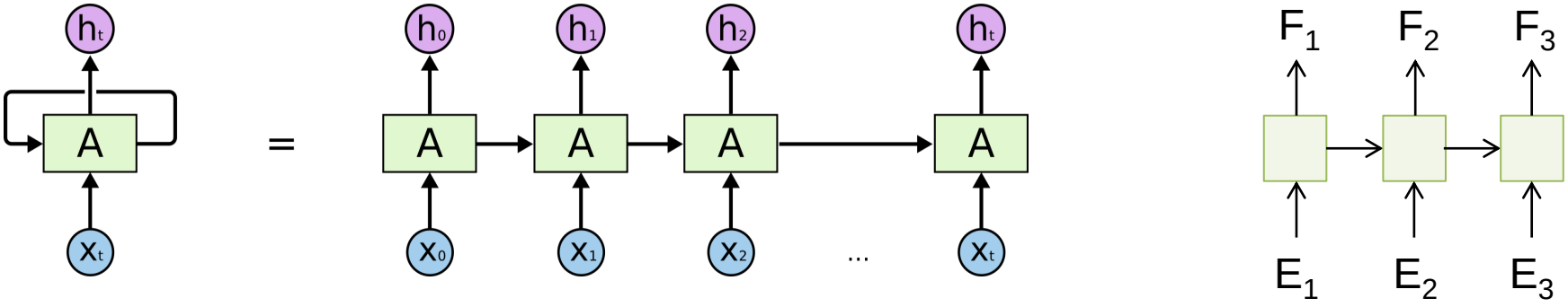
Στατιστική Μηχανική Μετάφραση

SMT Architecture

- Πολλά δομικά στοιχεία
- Αρχιτεκτονική pipeline
- Μετάφραση ανεξάρτητη συμφραζομένων
- Ανεξάρτητα μοντέλα (γλώσσας, μετάφρασης)
- Κάθε παράμετρος βελτιστοποιείται τοπικά



RNN – Μηχανική Μετάφραση

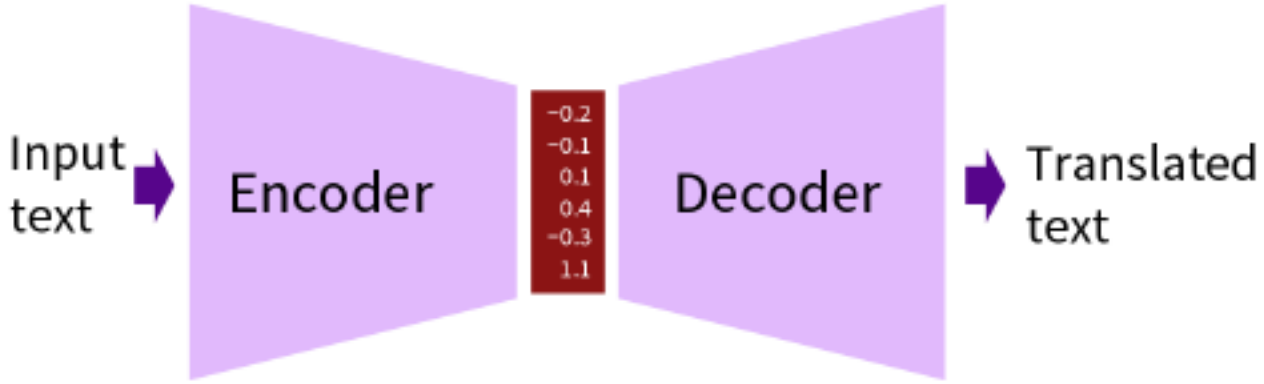


Μια RNN αρχιτεκτονική σαν τις παραπάνω δεν είναι κατάλληλη για μετάφραση μιας πρότασης από μια γλώσσα σε μια άλλη γιατί

- Οι δυο προτάσεις μπορεί να είναι διαφορετικού μήκους
- Οι λέξεις πολύ πιθανό να μην είναι ευθυγραμμισμένες

Οπότε πρέπει να διαβαστεί ολόκληρη η πρόταση πηγή προτού μεταφραστεί.

Encoder-Decoder Αρχιτεκτονικές



- Όλα σε ένα μεγάλο μοντέλο
- Καθολική βελτιστοποίηση των παραμέτρων
- Δεν υπάρχει ρητή κατάτμηση σε διακριτά μοντέλα

Μοντέλα Seq2Seq – Μηχανική Μετάφραση

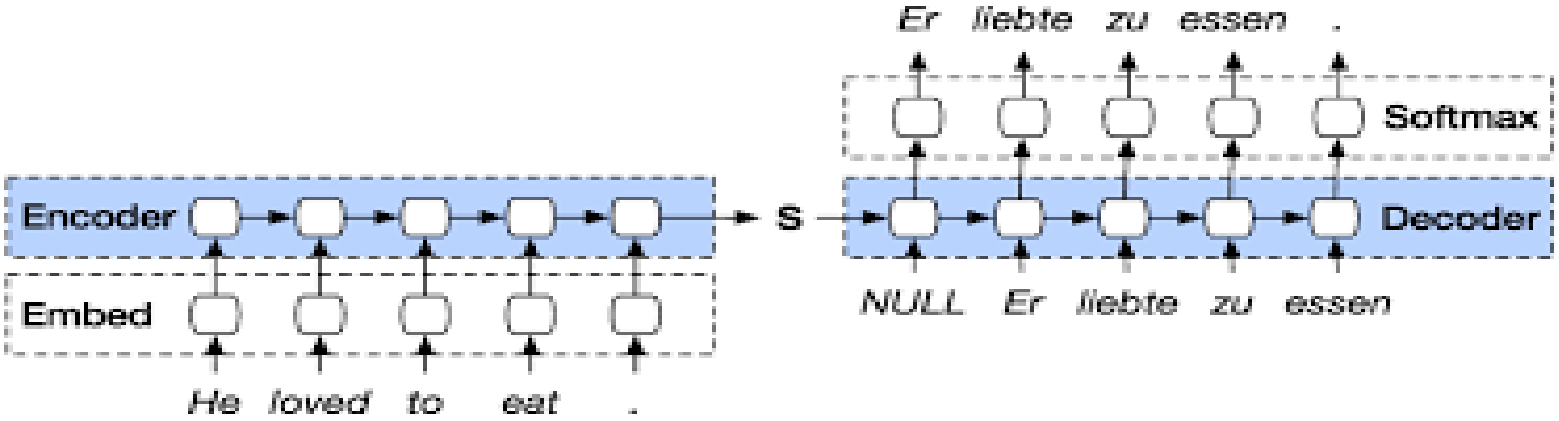
Για Sequence to sequence modeling έχουν προταθεί encoder-decoder αρχιτεκτονικές:

Οι λέξεις εισόδου μετατρέπονται σε word embeddings συγκεκριμένης διάστασης

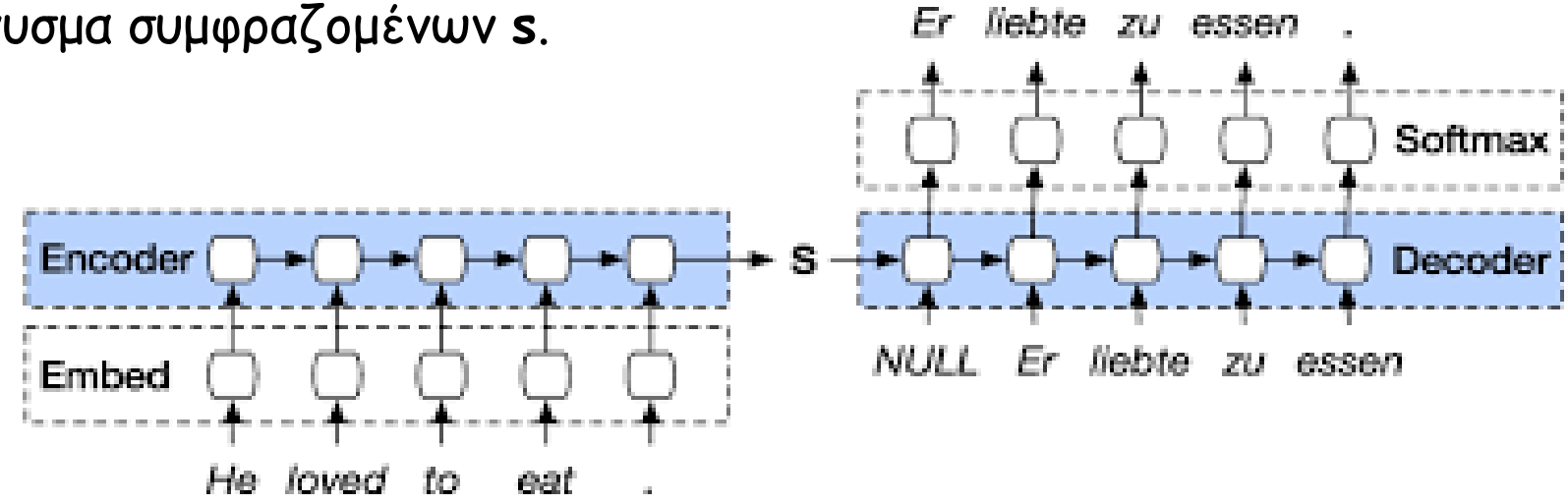
Το encoding NN (RNN ή CNN) παίρνει τα embeddings και τα αναπαριστά μέσω κρυφών καταστάσεων. Η έξοδος κάθε κρυφής κατάστασης εξαρτάται

- από την τωρινή είσοδο
- από την έξοδο της προηγούμενης κρυφής κατάστασης

Η τελευταία έξοδος είναι το διάνυσμα (νόημα) της πρότασης εισόδου (s) και είναι η είσοδος στο decoding NN (RNN ή CNN).



- Ο decoder θα ενεργοποιηθεί με το που θα δει το σύμβολο NULL (EOS).
- Στον πρώτο του κρυφό κόμβο θα πάρει σαν είσοδο το διάνυσμα **s**.
- Διατρέχονται όλα τα επίπεδα, εφαρμόζεται η softmax() στη έξοδο του τελευταίου επιπέδου και έτσι προβλέπεται η πρώτη λέξη εξόδου.
- Η λέξη αυτή επανατροφοδοτείται σαν είσοδος στο NN, μαζί με το διάνυσμα **s**, διατρέχονται πάλι όλα τα επίπεδα, εφαρμόζεται η softmax() στη έξοδο του τελευταίου επιπέδου και έτσι προβλέπεται η δεύτερη λέξη εξόδου.
- Με τον ίδιο τρόπο προβλέπονται και οι υπόλοιπες λέξεις της εξόδου.
- Εκπαίδευση: Κατά το backpropagation «μαθαίνονται» τα βάρη στον encoder, προκειμένου να μάθει καλύτερες διανυσματικές αναπαραστάσεις για τις προτάσεις και ταυτόχρονα «μαθαίνονται» τα βάρη του decoder για να μάθει να παράγει γραμματικά σωστές προτάσεις που είναι σχετικές με το διάνυσμα συμφραζομένων **s**.



Βιβλιογραφία/Δικτυογραφία

- http://cs224d.stanford.edu/
- http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
- https://recast.ai/blog/ml-spotlight-rnn/
- https://skymind.ai/wiki/lstm
- https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/
- http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/
- http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 2015
- https://web.stanford.edu/class/cs224n/archive/WWW_1617/lecture_notes/cs224n-2017-notes6.pdf