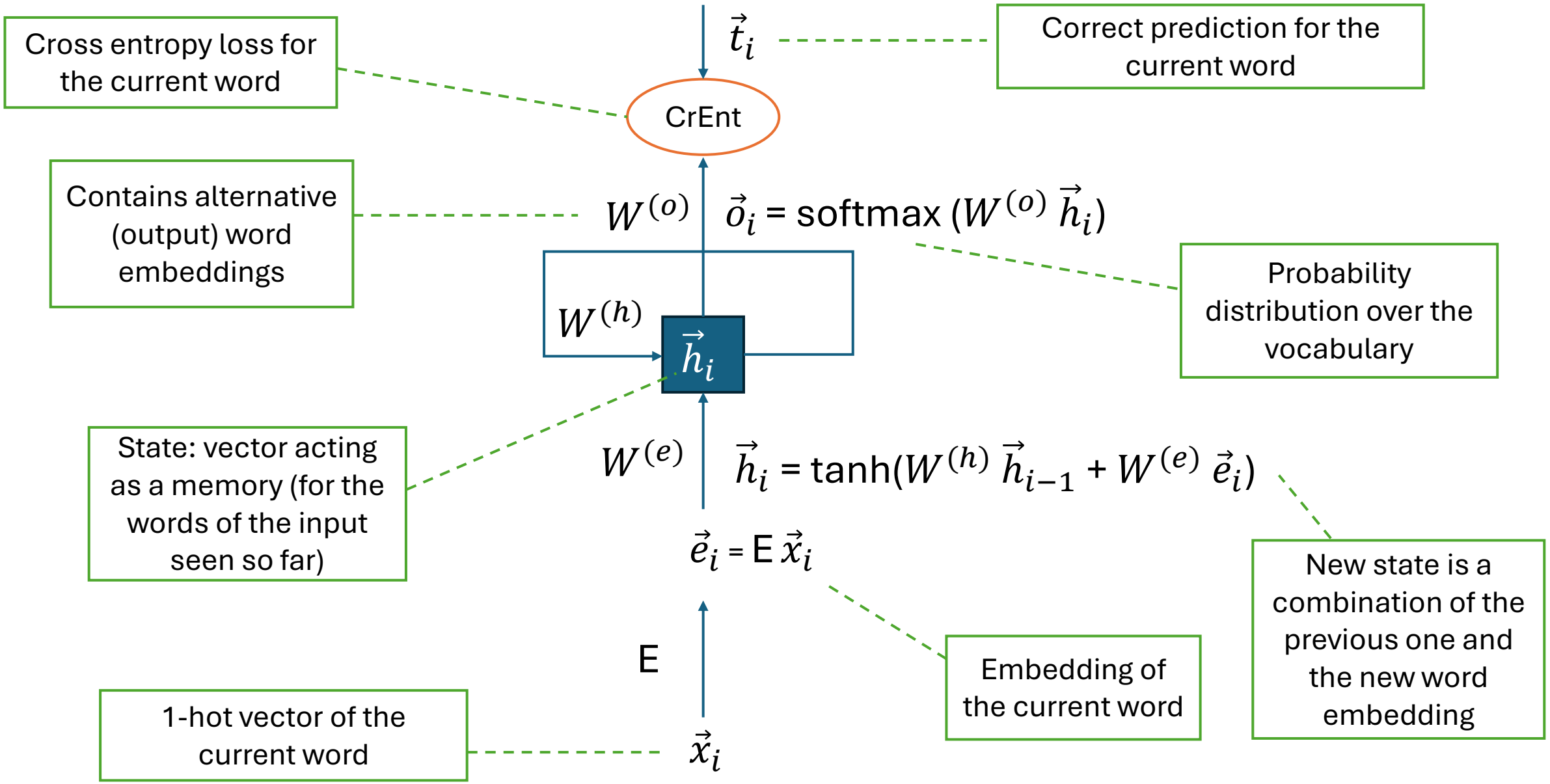
Τεχνητή νοημοσύνη

Φροντιστήριο 13

Ασκήσεις μελέτης της 22ης διάλεξης

Recurrent Neural Networks (RNNs)



Άσκηση 22.1

Θέλουμε να χρησιμοποιήσουμε το (RNN) των διαφανειών για να αναγνωρίζουμε ονόματα προσώπων, οργανισμών και τοποθεσιών. Χρησιμοποιούμε **7 κατηγορίες**.

Το μέγεθος του λεξιλογίου είναι **|V| = 100.000.**

Κάθε **word embedding είναι ένα διάνυσμα 300 διαστάσεων.**

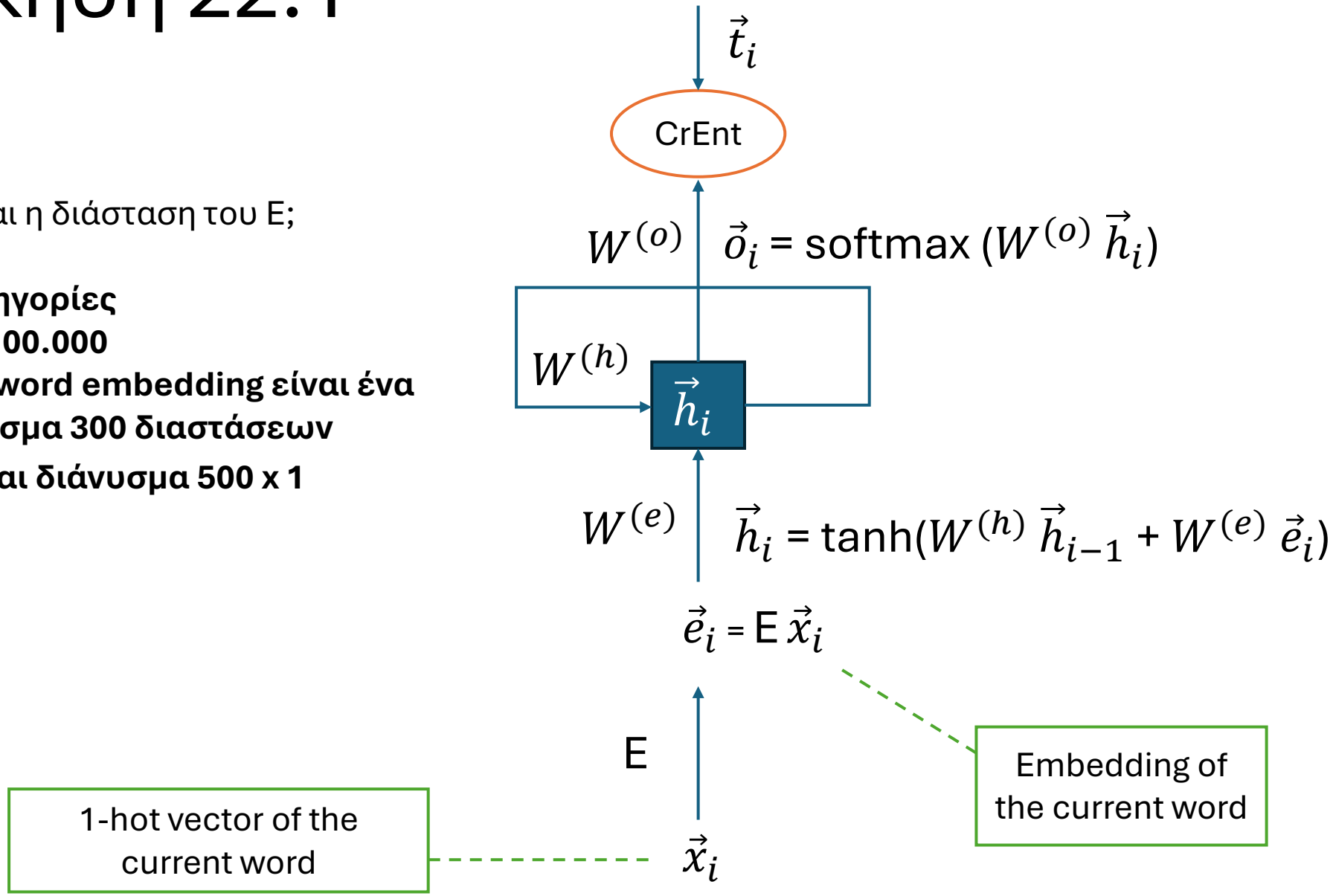
Το κρυφό επίπεδο (η κατάσταση του RNN) αποτελείται από 500 νευρώνες, δηλαδή το $\overrightarrow{\boldsymbol{h}_i}$ **είναι διάνυσμα 500 x 1.**

Ποιες είναι οι διαστάσεις των E, $\vec{e}_i$, $W^{(h)}$, $W^{(e)}$, $W^{(o)}$, $\vec{o}_i$;

Άσκηση 22.1

Ποια είναι η διάσταση του E;

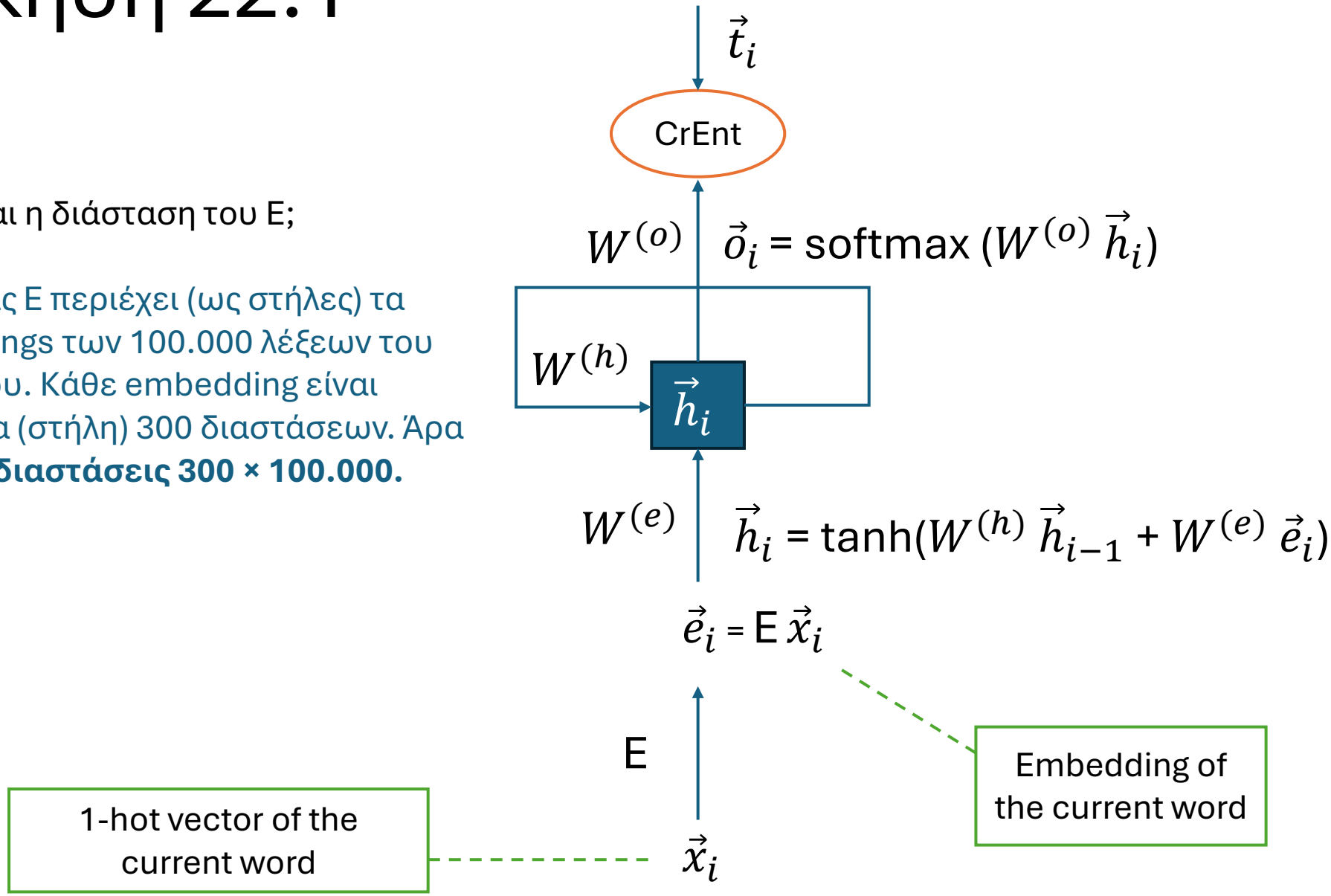
- **7 κατηγορίες**
- **|V| = 100.000**
- Κάθε **word embedding είναι ένα διάνυσμα 300 διαστάσεων**
- $\vec{h}_i$ **είναι διάνυσμα 500 x 1**



Άσκηση 22.1

Ποια είναι η διάσταση του E;

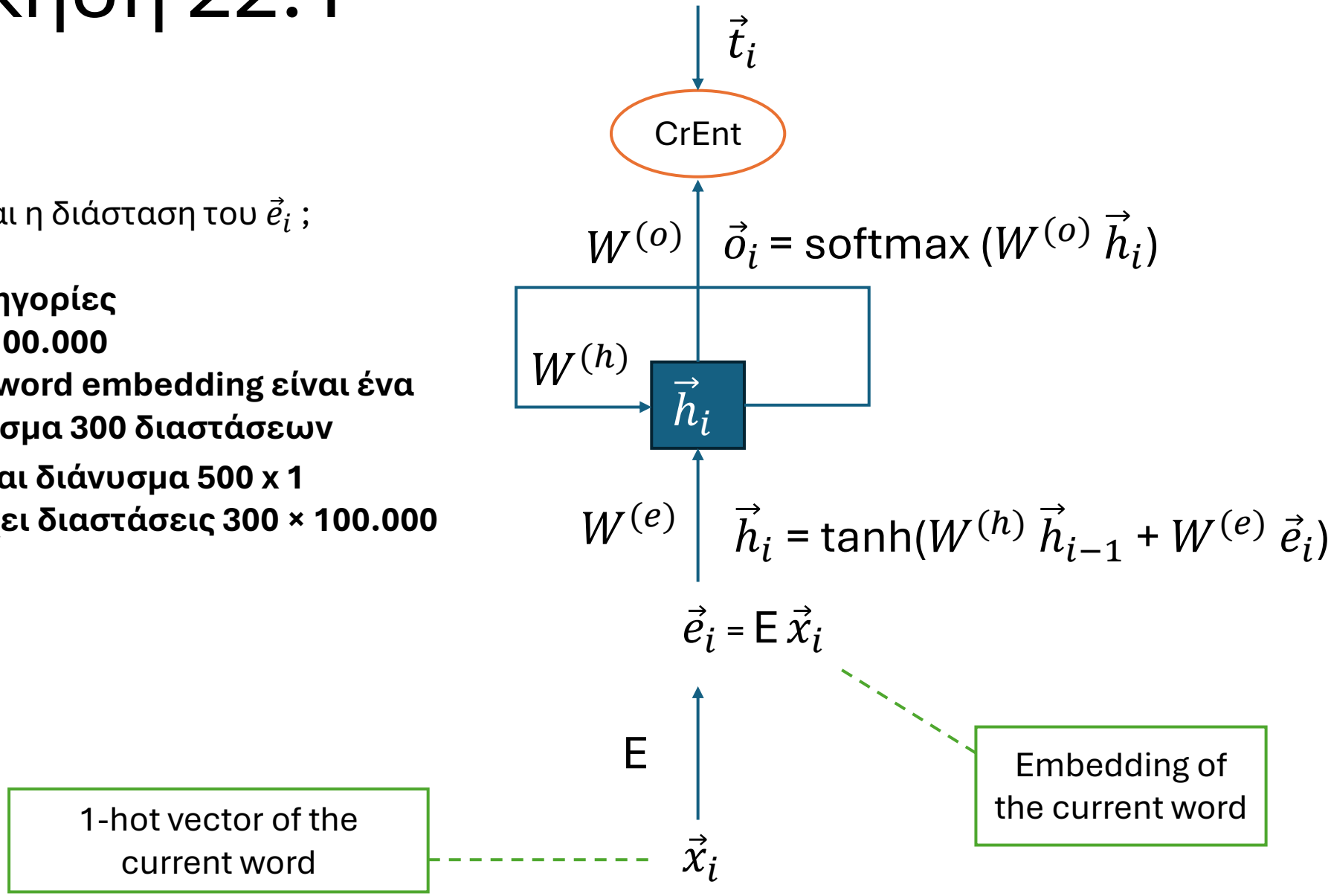
Ο πίνακας E περιέχει (ως στήλες) τα embeddings των 100.000 λέξεων του λεξιλογίου. Κάθε embedding είναι διάνυσμα (στήλη) 300 διαστάσεων. Άρα **ο E έχει διαστάσεις 300 × 100.000.**



Άσκηση 22.1

Ποια είναι η διάσταση του $\vec{e}_i$;

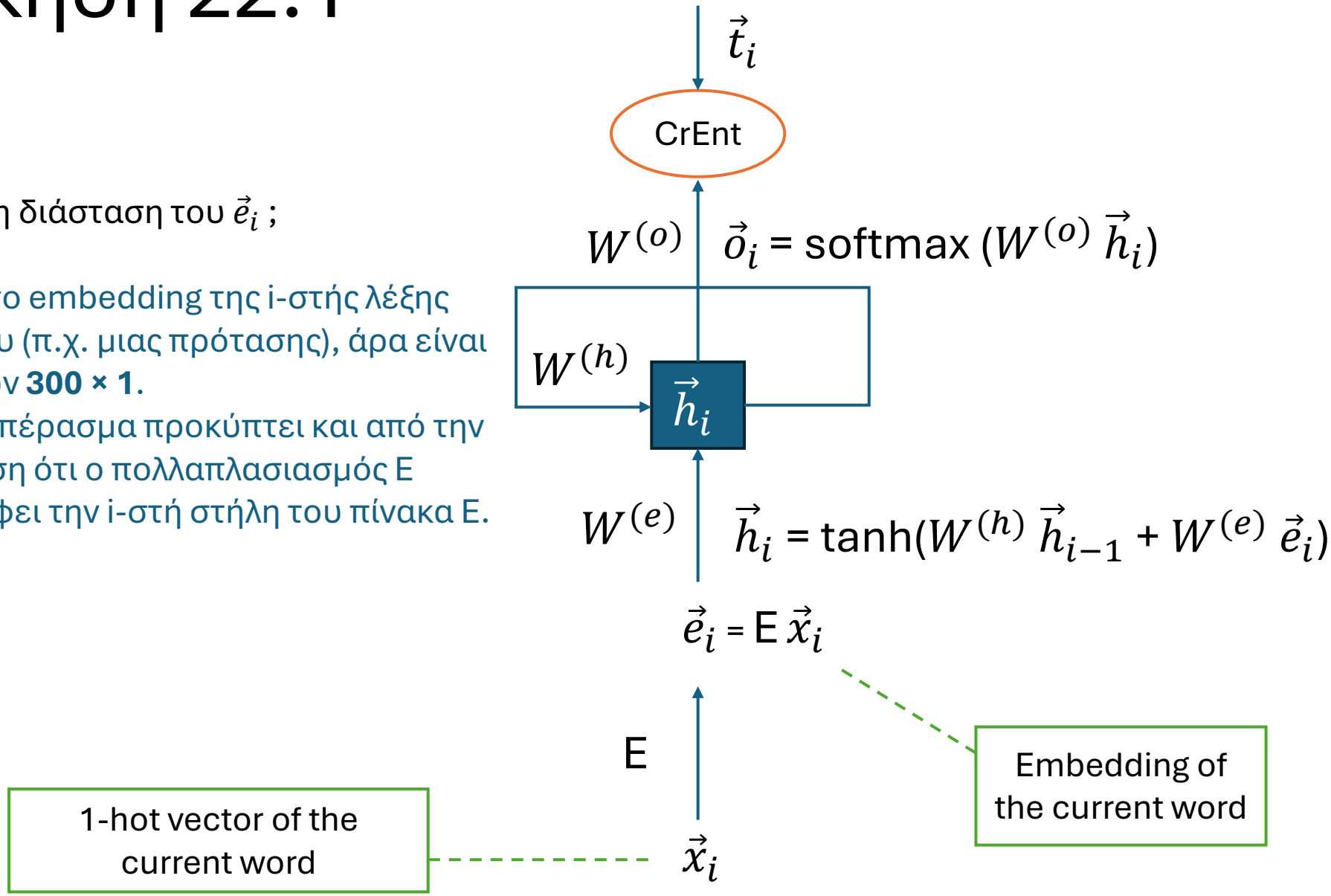
- **7 κατηγορίες**
- **|V| = 100.000**
- Κάθε **word embedding είναι ένα διάνυσμα 300 διαστάσεων**
- $\vec{h}_i$ **είναι διάνυσμα 500 x 1**
- **ο Ε έχει διαστάσεις 300 × 100.000**



Άσκηση 22.1

Ποια είναι η διάσταση του $\vec{e}_i$;

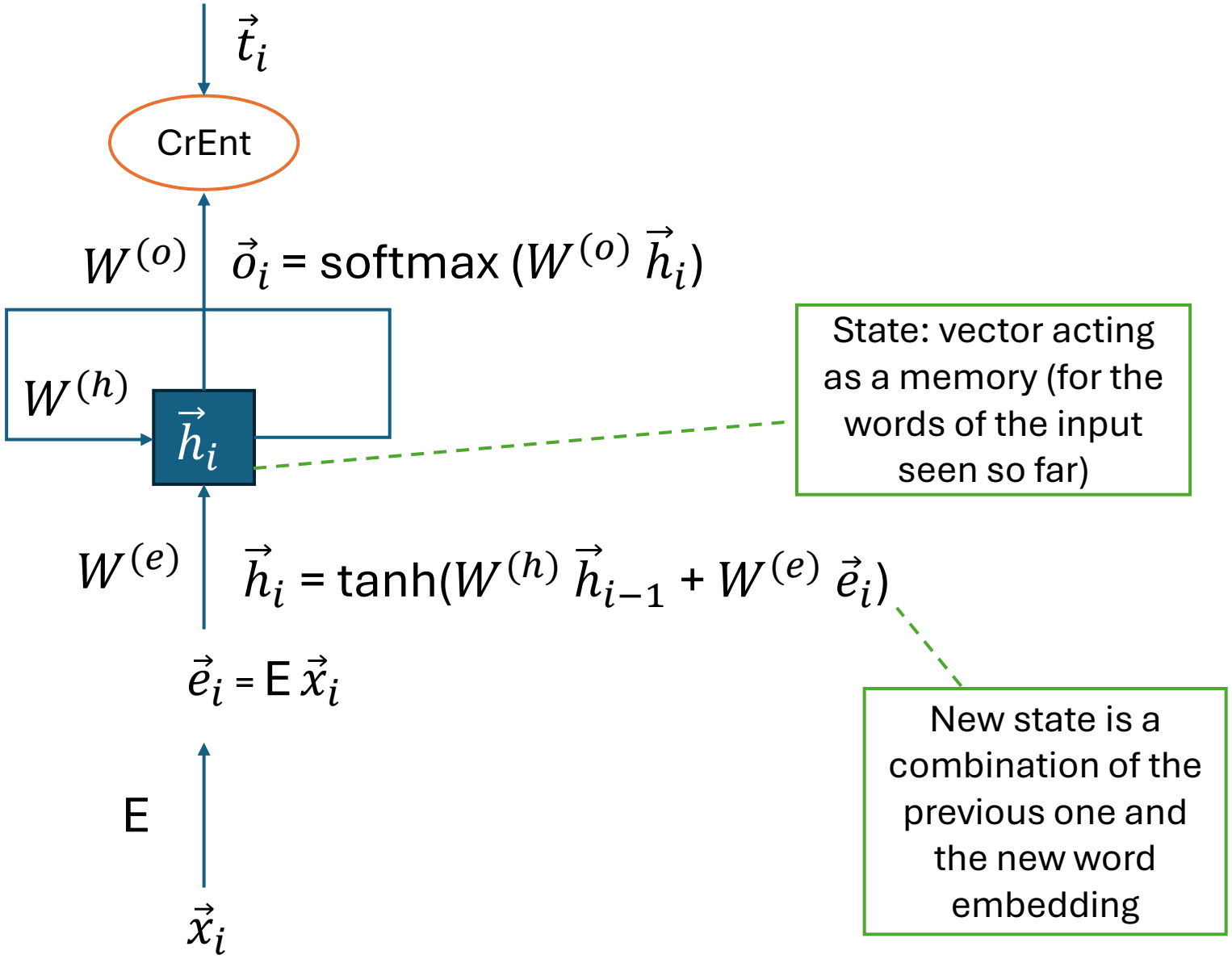
Το $\vec{e}_i$ είναι το embedding της i-στής λέξης της εισόδου (π.χ. μιας πρότασης), άρα είναι διαστάσεων **300 × 1**.
Το ίδιο συμπέρασμα προκύπτει και από την παρατήρηση ότι ο πολλαπλασιασμός E $\vec{x}_i$ επιστρέφει την i-στή στήλη του πίνακα E.



Άσκηση 22.1

Ποια είναι η διάσταση του $W^{(h)}$, $W^{(e)}$;

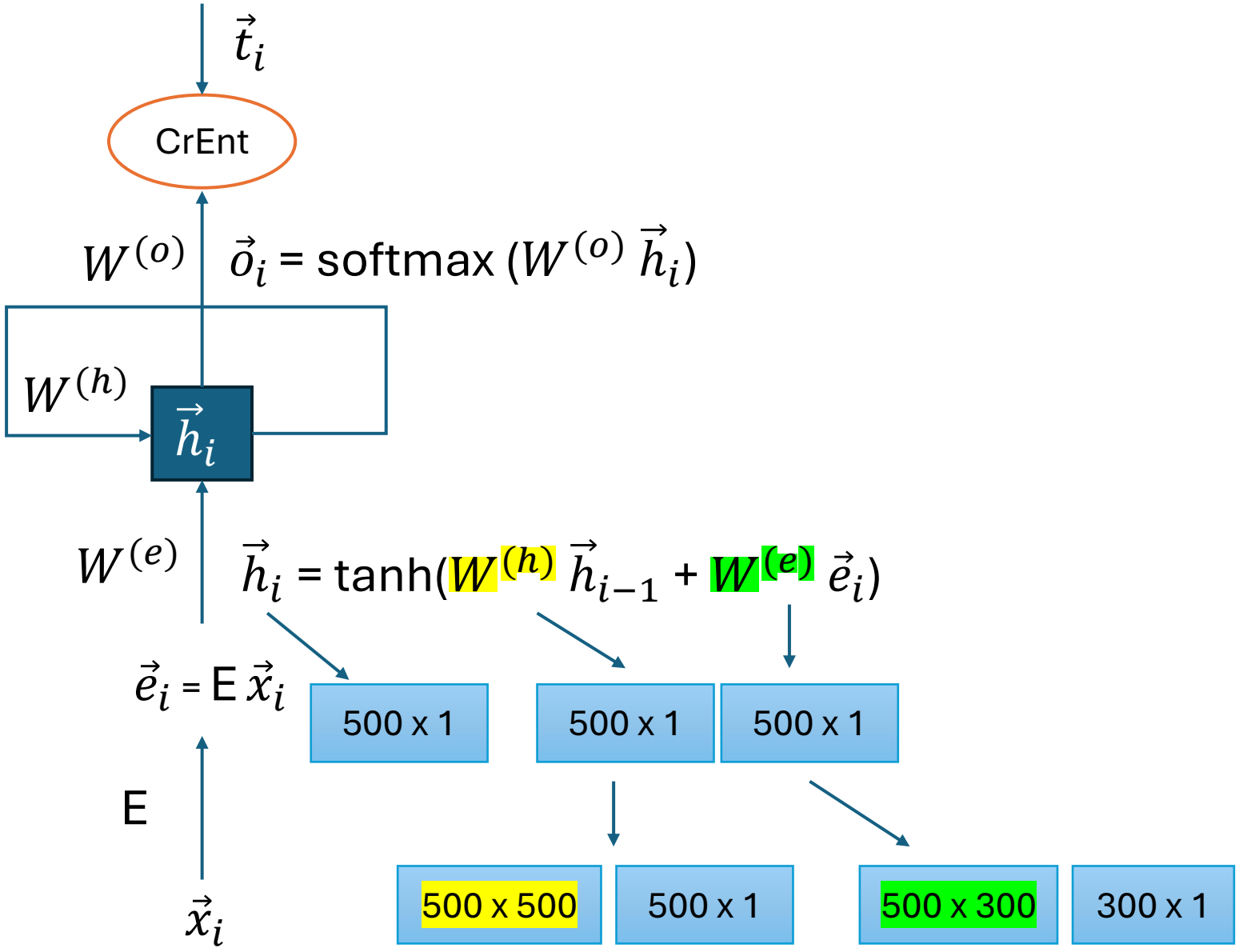
- **7 κατηγορίες**
- **|V| = 100.000**
- Κάθε **word embedding είναι ένα διάνυσμα 300 διαστάσεων**
- **$\vec{h}_i$ είναι διάνυσμα 500 x 1**
- **ο E έχει διαστάσεις 300 × 100.000**
- **Το $\vec{e}_i$ είναι διαστάσεων 300 × 1.**



Άσκηση 22.1

Ποια είναι η διάσταση του $W^{(h)}$, $W^{(e)}$;

- **7 κατηγορίες**
- **|V| = 100.000**
- Κάθε **word embedding είναι ένα διάνυσμα 300 διαστάσεων**
- **$\vec{h}_i$ είναι διάνυσμα 500 x 1**
- **ο E έχει διαστάσεις 300 × 100.000**
- **Το $\vec{e}_i$ είναι διαστάσεων 300 × 1.**

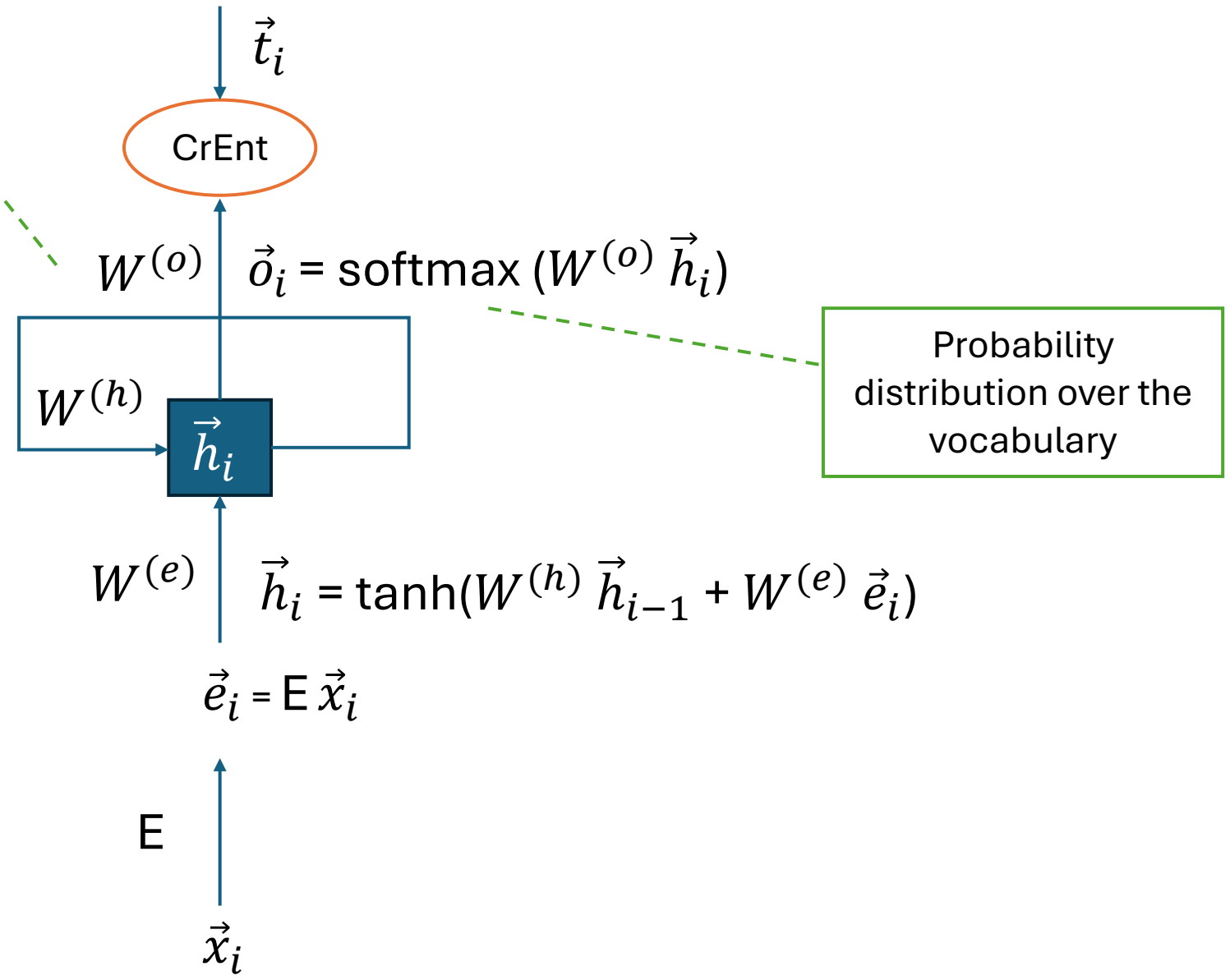


Άσκηση 22.1

Contains alternative (output) word embeddings

Ποια είναι η διάσταση του $W^{(o)}$, $\vec{o}_i$;

- **7 κατηγορίες**
- **|V| = 100.000**
- Κάθε **word embedding είναι ένα διάνυσμα 300 διαστάσεων**
- **$\vec{h}_i$ είναι διάνυσμα 500 x 1**
- **ο E έχει διαστάσεις 300 × 100.000**
- **Το $\vec{e}_i$ είναι διαστάσεων 300 × 1.**
- **Το $W^{(h)}$ είναι διαστάσεων 500 x 500.**
- **Το $W^{(e)}$ είναι διαστάσεων 500 x 300.**



Probability distribution over the vocabulary

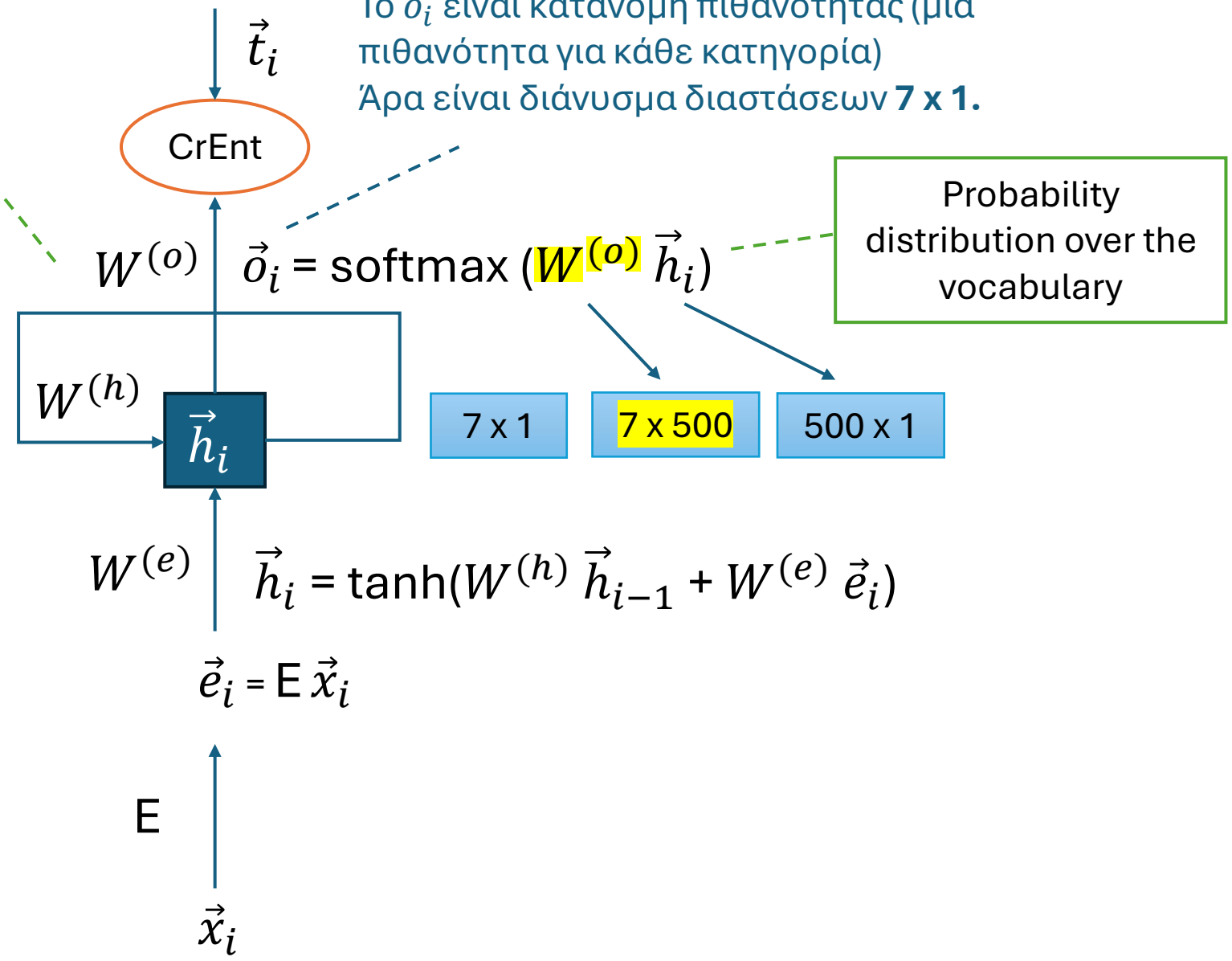
Άσκηση 22.1

Ποια είναι η διάσταση του $W^{(o)}$, $\vec{o}_i$;

- **7 κατηγορίες**
- **|V| = 100.000**
- Κάθε **word embedding είναι ένα διάνυσμα 300 διαστάσεων**
- **$\vec{h}_i$ είναι διάνυσμα 500 x 1**
- **ο E έχει διαστάσεις 300 × 100.000**
- **Το $\vec{e}_i$ είναι διαστάσεων 300 × 1.**
- **Το $W^{(h)}$ είναι διαστάσεων 500 x 500.**
- **Το $W^{(e)}$ είναι διαστάσεων 500 x 300.**

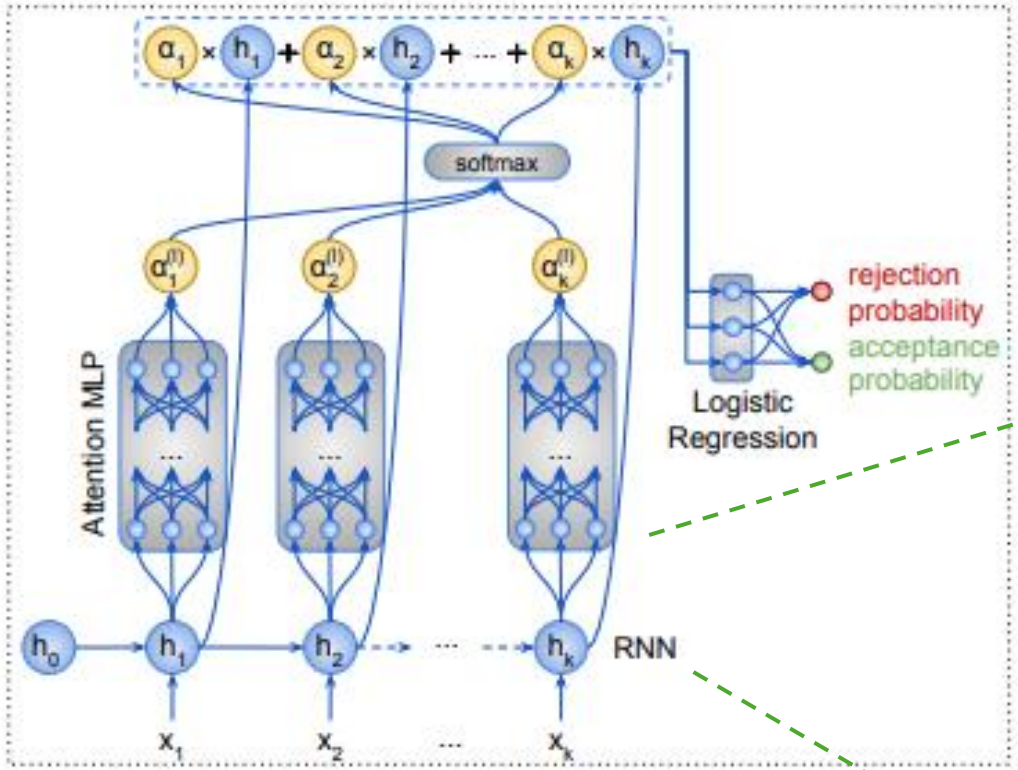
Contains alternative (output) word embeddings

Το $\vec{o}_i$ είναι κατανομή πιθανότητας (μία πιθανότητα για κάθε κατηγορία)
Άρα είναι διάνυσμα διαστάσεων **7 x 1.**

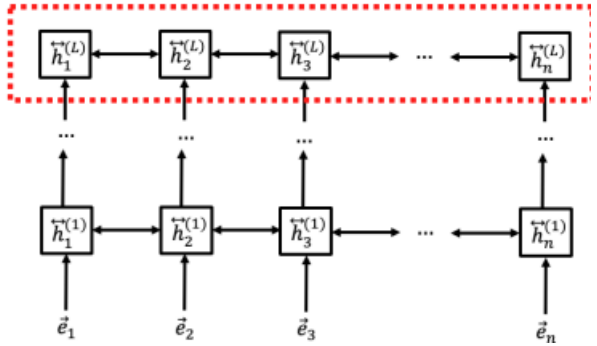


RNN with deep self-attention

The entire input text is now represented by the weighted (by a_i scores) sum of the revised embeddings of its words.

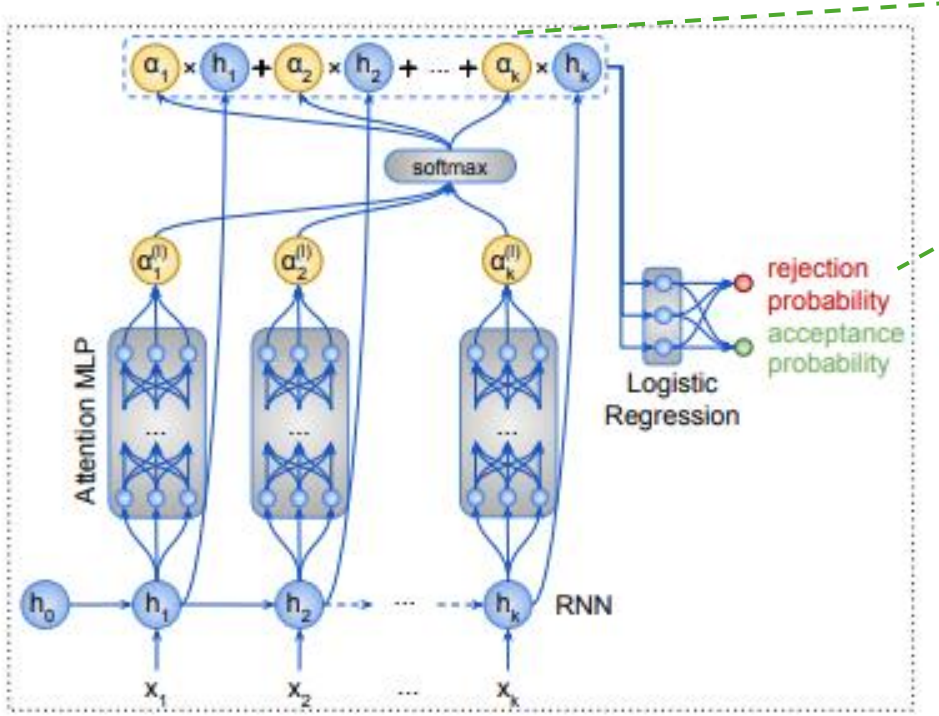


We use an MLP to obtain an **attention score** (importance) a_i for each word from its revised embedding h_i



Could be the top-level revised embeddings of a stacked biRNN

RNN with deep self attention



$$h_{sum} = \sum_{t=1}^{k} a_t h_t$$

$$P_{a-\text{RNN}}(reject|c) = \sigma(W_p h_{sum} + b_p)$$

$$\begin{aligned} a_t^{(1)} &= \text{RELU}(W^{(1)} h_t + b^{(1)}) \\ &\cdots \\ a_t^{(l-1)} &= \text{RELU}(W^{(l-1)} a_t^{(l-2)} + b^{(l-1)}) \\ a_t^{(l)} &= W^{(l)} a_t^{(l-1)} + b^{(l)} \end{aligned}$$

MLP layers

$$a_t = \texttt{softmax}(a_t^{(l)}; a_1^{(l)}, \ldots, a_k^{(l)})$$

The softmax ensures all scores are between 0 and 1, and that they sum to 1

$$\begin{aligned} \tilde{h}_t &= \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \\ z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \end{aligned}$$

Update gate: how much of the previous hidden state should be carried forward to the current time step

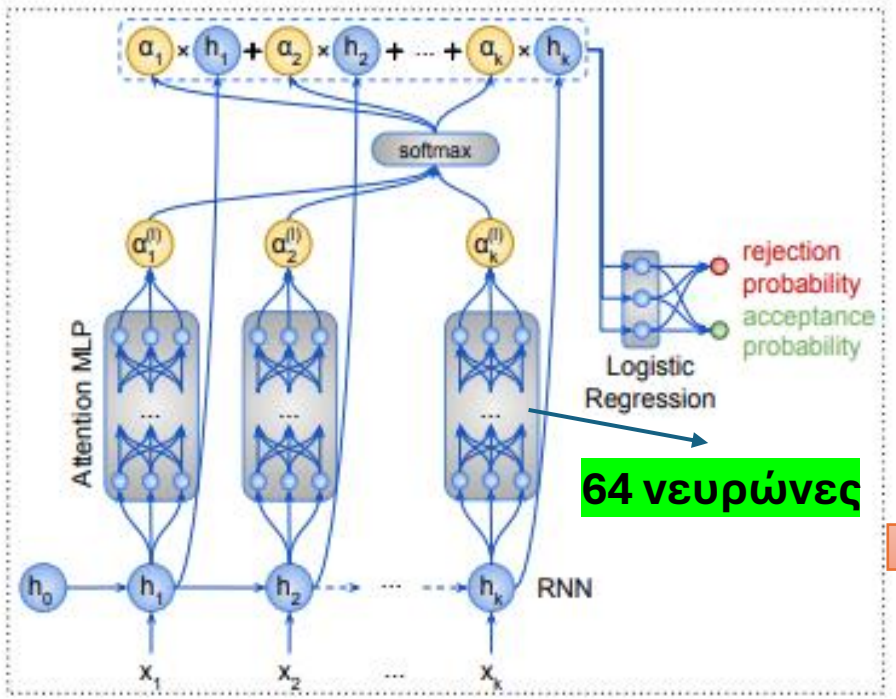
Reset gate: how much of the previous hidden state is used when calculating $\tilde{h}_t$

Άσκηση 22.2 (α)

Στο νευρωνικό δίκτυο RNN with deep self attention, οι καταστάσεις $\boldsymbol{h_1, h_2, \dots, h_k}$ **του RNN είναι διανύσματα 128 διαστάσεων** (η κάθε μία). Τα κρυφά επίπεδα (1), …, (l − 1) του Attention MLP **έχουν 64 νευρώνες** το καθένα και οι έξοδοι των κρυφών επιπέδων είναι $a_t^{(1)}, \dots, a_t^{(l-1)}$. Το επίπεδο εξόδου του Attention MLP έχει έναν μόνο νευρώνα με έξοδο $a_t^{(l)}$. Τι διαστάσεις θα έχουν οι πίνακες $W^{(1)}, \dots, W^{(l)}$ και τα διανύσματα $b^{(1)}, \dots, b^{(l)}$;

Άσκηση 22.2 (α)

Οι καταστάσεις $h_1, h_2, \dots, h_k$ του RNN είναι **διανύσματα 128 διαστάσεων**



$$h_{sum} = \sum_{t=1}^{k} a_t h_t$$

$$P_{a-\text{RNN}}(reject|c) = \sigma(W_p h_{sum} + b_p)$$

64 x 1 | 64 x 128 | 128 x 1

$$a_t^{(1)} = \text{RELU}(W^{(1)} h_t + b^{(1)})$$

64 x 1 | 64 x 64 | 64 x 1

$$a_t^{(l-1)} = \text{RELU}(W^{(l-1)} a_t^{(l-2)} + b^{(l-1)})$$

$$a_t^{(l)} = W^{(l)} a_t^{(l-1)} + b^{(l)}$$

1 x 1

$$a_t = \text{softmax}(a_t^{(l)}; a_1^{(l)}, \dots, a_k^{(l)})$$

Πραγματικοί αριθμοί (1x1)

1 x 1 | 1 x 64 | 64 x 1

$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

Άσκηση 22.2 (β)

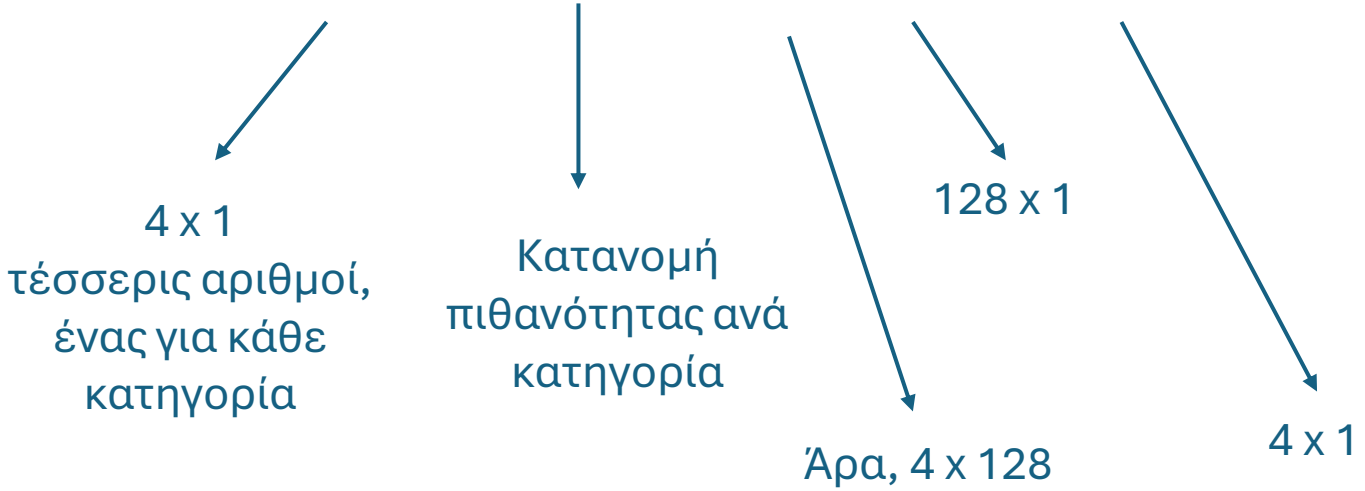
Θέλουμε να χρησιμοποιήσουμε το νευρωνικό δίκτυο των διαφανειών 19–21, τώρα για να κατατάξουμε tweets (που αναφέρονται σε ένα προϊόν) στις **κατηγορίες c_1 (θετική γνώμη), c_2 (αρνητική γνώμη), c_3 (ουδέτερη γνώμη), c_4 (θετική και αρνητική γνώμη μαζί)**.Κάθε tweet θα κατατάσσεται σε ακριβώς μία κατηγορία.

Αντικαθιστούμε τον τύπο: $P_{a-\text{RNN}}(reject|c) = \sigma(W_p h_{sum} + b_p)$ με τον παρακάτω τύπο που θα πρέπει να παράγει (στο αριστερό του μέρος) ένα διάνυσμα p $\in \mathbb{R}^4$, το οποίο θα περιέχει τις πιθανότητες (κατά το νευρωνικό δίκτυο) το εισερχόμενο tweet να ανήκει σε κάθε μία από τις τέσσερις κατηγορίες.

Συμπληρώστε το δεξί μέρος του τύπου, αναφέροντας τις διαστάσεις κάθε πίνακα και διανύσματος που θα εμφανίζεται στο δεξί μέρος του τύπου. Αιτιολογήστε σύντομα την απάντησή σας.

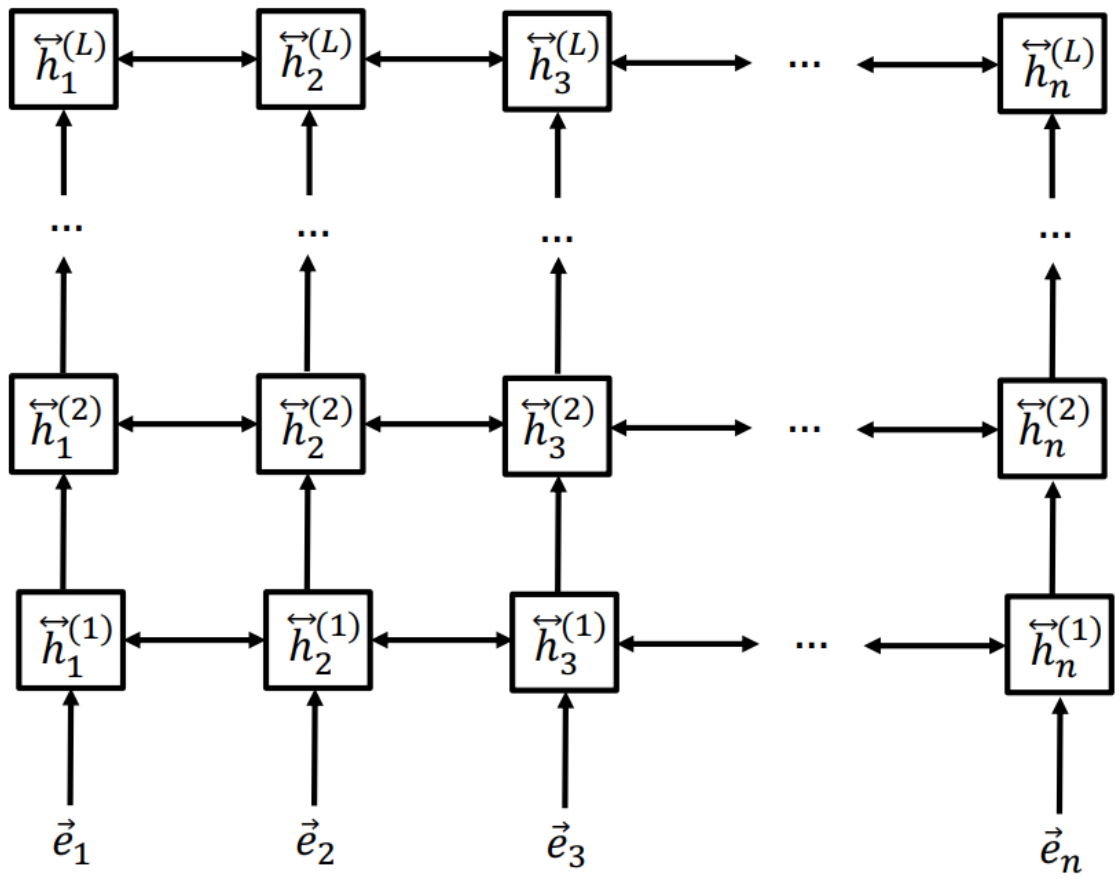
Άσκηση 22.2 (β)

Ο νέος τύπος θα είναι $p = \text{softmax}(W_p h_{sum} + b_p)$



Άσκηση 22.3

Γράψτε τις εξισώσεις για μια τροποποιημένη έκδοση του RNN with deep self attention, όπου το uni-directional RNN με GRU cells αντικαθίσταται από ένα stacked bi-directional RNN με GRU cells. Χρησιμοποιήστε το GRU($h_{t-1,}\ \tau_t$) για να δηλώσετε την καινούρια κατάσταση του GRU cell με προηγούμενη κατάσταση h_{t-1} και είσοδο τ_t.



Άσκηση 22.3

Στο πρώτο επίπεδο του GRU RNN έχουμε για t = 1, 2, … , k:

$$\vec{h}_t^{(1)} = \mathrm{GRU}\left(\vec{h}_{t-1}^{(1)}, x_t\right)$$

$$\overleftarrow{h}_t^{(1)} = \mathrm{GRU}\left(\overleftarrow{h}_{t+1}^{(1)}, x_t\right)$$

$$h_t^{(1)} = [\vec{h}_t^{(1)}; \overleftarrow{h}_t^{(1)}]$$

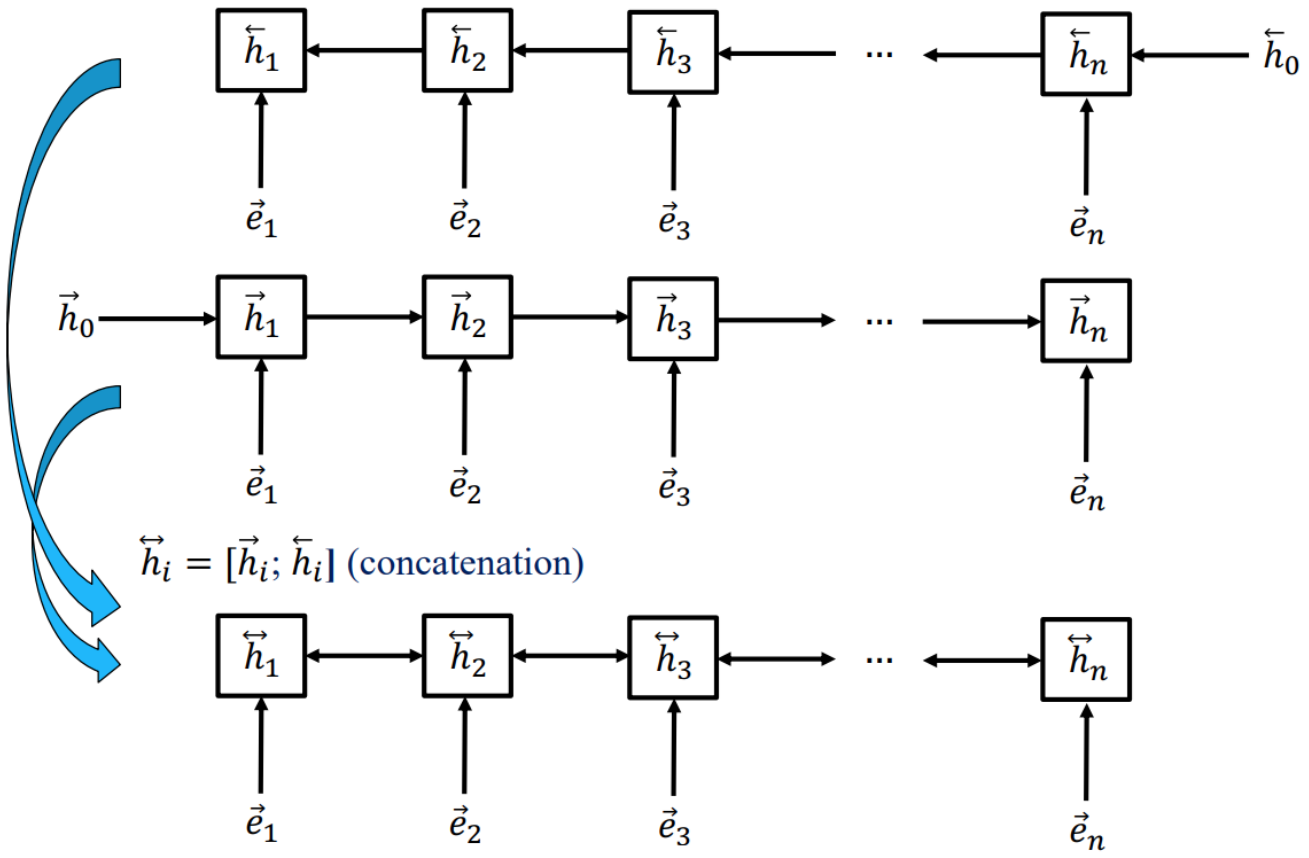
Αντίστοιχα, για το m-επίπεδο του GRU RNN:

$$\vec{h}_t^{(m)} = \mathrm{GRU}\left(\vec{h}_{t-1}^{(m)}, h_t^{(m-1)}\right)$$

$$\overleftarrow{h}_t^{(m)} = \mathrm{GRU}\left(\overleftarrow{h}_{t+1}^{(m)}, h_t^{(m-1)}\right)$$

$$h_t^{(m)} = [\vec{h}_t^{(m)}; \overleftarrow{h}_t^{(m)}]$$

Οι υπόλοιπες εξισώσεις μένουν όπως πριν.



Άσκηση 22.4

Τροποποιήστε τις εξισώσεις του νευρωνικού δικτύου από την προηγούμενη άσκηση ώστε να υποστηρίζουν ταξινόμηση με πολλαπλές ετικέτες (**multilabel classification**), δηλαδή περιπτώσεις όπου το ίδιο κείμενο (π.χ. ένα tweet) μπορεί να ανήκει σε πολλαπλές κατηγορίες (ετικέτες).

Ως μια διαφοροποίηση, χρησιμοποιήστε ένα ξεχωριστό self-attention-head για κάθε ετικέτα, το οποίο θα παράγει μια διαφορετική κατανομή από **attention scores** $\boldsymbol{a_{c,1}, \ldots, a_{c,k}}$ όπου k είναι όπως πριν το μέγεθος του κειμένου εισόδου μετρημένο σε λέξεις και έχουμε ένα **διαφορετικό** $\boldsymbol{h_{sum,c}}$ **για κάθε κλάση c.**

Δώστε το $h_{sum,c}$ κάθε κατηγορίας c σε ένα ξεχωριστό (διαφορετικό για κάθε κατηγορία) dense layer με μια sigmoid, για να παραχθεί η πιθανότητα ότι το κείμενο εισόδου πρέπει να ταξινομηθεί στην κατηγορία c.

Άσκηση 22.4

Έστω C το σύνολο όλων των πιθανών κλάσεων.

Τροποποιούμε το self-attention MLP των διαφανειών 19-21, έτσι ώστε το $a_t^{(l)}$ να είναι διάνυσμα και όχι πραγματικός αριθμός, το οποίο θα περιέχει |C| attention scores $a_{1,t}, \ldots, a_{|C|,t}$ για τη λέξη στη θέση t .

Για να το πετύχουμε αυτό τροποποιούμε τις διαστάσεις των $W^{(l)}$ και $b^{(l)}$ του επιπέδου l του self-attention MLP, ώστε να είναι |C| x d και |C| αντίστοιχα, όπου d είναι η διάσταση του προηγούμενου επιπέδου $a_t^{(l-1)}$.

Η softmax εφαρμόζεται ανά label, στα attention scores μιας συγκεκριμένης κλάσης και δημιουργούμε ένα διαφορετικό $h_{sum,c}$ για κάθε δυνατή κλάση c, όπου το M είναι ο αριθμός των προηγούμενων stacked GRU RNNs.

Οι υπόλοιπες εξισώσεις μένουν όπως πριν.

$$h_{sum,c} = \sum_{t=1}^{k} a_{c,t} h_t^{(M)}$$

$$h_{sum} = \sum_{t=1}^{k} a_t h_t$$

$$P_{a-\text{RNN}}(reject|c) = \sigma(W_p h_{sum} + b_p)$$

$$P(c|x_1, \ldots, x_k) = \sigma(W_{p,c} h_{sum,c} + b_{p,c})$$

|C| x d

d x 1

|C| x 1

|C| x 1

$$a_t^{(l)} = W^{(l)} a_t^{(l-1)} + b^{(l)}$$

$$a_t = \texttt{softmax}(a_t^{(l)}; a_1^{(l)}, \ldots, a_k^{(l)})$$

$$a_{c,t} = \text{softmax}\left(a_{c,t}^{(l)}; a_{c,1}^{(l)}, \ldots, a_{c,k}^{(l)}\right) = \frac{\exp(a_{c,t}^{(l)})}{\sum_{t'=1}^{k} \exp(a_{c,t'}^{(l)})}$$