Θέματα 2023-2024 Ιούνιος

Τα πολλαπλής είναι με κλειστές σημειώσεις

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΚΑΙ ΒΑΘΙΑ ΜΑΘΗΣΗ

ΚΑΝΟΝΙΚΗ ΕΞΕΤΑΣΗ — **ΘΕΜΑΤΑ ΠΟΛΛΑΠΛΗΣ ΕΠΙΛΟΓΗΣ**

21-06-2024 — **ΔΙΑΡΚΕΙΑ: 35 ΛΕΠΤΑ**

Να απαντήσετε στα παρακάτω ερωτήματα επιλέγοντας μία από τις τέσσερις επιλογές. Τα ερωτήματα είναι ισοδύναμα και έχουν μία μόνο σωστή επιλογή. Απαντάτε σωστά στο ερώτημα, αν κυκλώσετε μόνο την σωστή επιλογή. Κάθε άλλη απάντηση, θεωρείται λανθασμένη. Για κάθε σωστή απάντηση που δίνετε θα κερδίζετε 3 μονάδες. Για κάθε λανθασμένη απάντηση που δίνετε θα χάνετε 1 μονάδα. Μπορείτε να απαντήσετε σε όσα ερωτήματα επιθυμείτε (αν δεν απαντήσετε σε κάποιο ερώτημα, ούτε κερδίζετε, ούτε χάνετε μονάδες). Αν το τελικό αποτέλεσμα είναι αρνητικό, θα επηρεάσει αντίστοιχα το βαθμό όλου του διαγωνίσματος.

1. Ποιο πρόβλημα του απλού Q-learning αλγορίθμου αντιμετωπίζει ο DQN;
   - Α. Την αργή σύγκλιση στις βέλτιστες τιμές του πίνακα Q.
   - Β. Την εμφάνιση φαινομένων ταλαντώσεων ή αποκλίσεων όταν εφαρμόζεται στα νευρωνικά δίκτυα.
   - Γ. Την αστάθεια του απλού Q-learning αλγορίθμου όταν εφαρμόζεται σε δίκτυα που βασίζονται στην εκτίμηση πολιτικών.
   - Δ. Κανένα από τα παραπάνω.

2. Ας υποθέσουμε ότι ένας transformer με δύο κεφαλές προσοχής, ο ένας με projection matrices $W_Q^{(1)}, W_K^{(1)}, W_V^{(1)}$ και ο άλλος με projection matrices $W_Q^{(2)}, W_K^{(2)}, W_V^{(2)}$ όπου $W_Q^{(2)} = 2W_Q^{(1)}$, $W_K^{(2)} = W_K^{(1)}$, και $W_V^{(2)} = (1/2)\, W_V^{(1)}$. Έστω $P^{(1)}$ και $P^{(2)}$ οι πίνακες πιθανότητας προσοχής και $Z^{(1)}$ και $Z^{(2)}$ είναι τα context representations που λαμβάνονται από τις δύο κεφαλές προσοχής. Ποια είναι η σχέση μεταξύ $P^{(1)}$ και $P^{(2)}$ και μεταξύ $Z^{(1)}$ και $Z^{(2)}$ που ισχύει γενικά;
   - Α. $P^{(1)} = P^{(2)}$ και $Z^{(1)} = Z^{(2)}$.
   - Β. $P^{(1)} \neq P^{(2)}$ και $Z^{(1)} = Z^{(2)}$.
   - Γ. $P^{(1)} \neq P^{(2)}$ και $Z^{(1)} \neq Z^{(2)}$.
   - Δ. $P^{(1)} = P^{(2)}$ και $Z^{(1)} \neq Z^{(2)}$.

3. Πώς σχετίζονται τα Langevin dynamics με τα Score-Based Generative Models;
   - Α. Ορίζουν τη συνάρτηση θορύβου που χρησιμοποιείται για την αλλοίωση των δεδομένων κατά τη διάρκεια της εκπαίδευσης.
   - Β. Χρησιμοποιούνται για την άμεση αξιολόγηση της πιθανότητας κατανομής δεδομένων ενός δείγματος.
   - Γ. Χρησιμοποιούνται για την προ-εκπαίδευση του score network.
   - Δ. Καθοδηγούν τη διαδικασία δειγματοληψίας με επαναληπτική αποθορυβοποίηση του τυχαίου θορύβου.

4. Ποια από τις παρακάτω προτάσεις για τα επίπεδα pooling ενός CNN (Convolutional Neural Network) είναι λανθασμένη;
   - Α. Η έξοδος των συνελικτικών σταδίων τροφοδοτεί τα στάδια pooling.
   - Β. Η χρήση ενός σταδίου pooling διατηρεί σταθερό το βάθος του χάρτη (τένσορα) χαρακτηριστικών.
   - Γ. Μειώνουν τον χρόνο εκπαίδευσης.
   - Δ. Καμία από τις παραπάνω.

5. Ποιο από τα παρακάτω προεκπαιδευμένα μοντέλα είναι πιο κατάλληλο για μια ένα downstream task που απαιτεί παραγωγή κειμένου με με αυτόματη παλινδρόμηση;
   - Α. Ένα encoder-only μοντέλο όπως το BERT.
   - Β. Είτε ένα μοντέλο encoder-only μοντέλο όπως το BERT είτε ένα μοντέλο encoder-decoder όπως το T5.
   - Γ. Είτε ένα μοντέλο decoder-only όπως το GPT-3 είτε ένα μοντέλο encoder-decoder όπως το T5.
   - Δ. Ένα μοντέλο decoder-only όπως το GPT-3.

6. Ποια από τις παρακάτω προτάσεις σχετικά με τα autoregressive RNN μοντελά είναι σωστή;
   - Α. Τα RNN έχουν απεριόριστη μνήμη, ωστόσο για μεγάλες ακολουθίες έχουν μια τάση να «θυμούνται» με μικρότερη ακρίβεια τις αρχικές λέξεις που παράγουν.
   - Β. Ακριβώς όπως τα n-gram μοντέλα, τα RNN μπορούν μόνο να θυμηθούν τις τελευταίες n λέξεις που παράγουν, για ένα δεδομένο $n \in N$.
   - Γ. Τα RNN έχουν απεριόριστη μνήμη, ωστόσο έχουν την τάση να «θυμούνται» με λιγότερη ακρίβεια τις πιο πρόσφατες λέξεις που παράγουν καθώς το βάρος των αρχικών λέξεων μεγαλώνει όσο ανατροφοδοτείται στα RNN units.
   - Δ. Τα RNN υποφέρουν από vanishing gradients, αλλά αυτό μπορεί να μετριαστεί με gradient clipping.

7. Ποια είναι η βασική διαφορά μεταξύ των loss functions που χρησιμοποιούνται για τον generator και τον discriminator σε ένα GAN που εκπαιδεύεται με minimax loss;

Α. Το discriminator loss χρησιμοποιεί cross-entropy, ενώ το generator loss χρησιμοποιεί το μέσο τετραγωνικό σφάλμα.
Β. Και οι δύο συναρτήσεις προέρχονται από την ίδια αρχική διατύπωση minimax, αλλά το generator loss είναι απλοποιημένο λόγω της αδυναμίας του generator να επηρεάσει την πιθανότητα των πραγματικών δεδομένων.
Γ. Το discriminator loss ενθαρρύνει την ταξινόμηση των πραγματικών δεδομένων ως πραγματικών, ενώ το generator loss στοχεύει στην ταξινόμηση των παραγόμενων δεδομένων ως ψεύτικων.
Δ. Το discriminator loss ελαχιστοποιεί τη διαφορά μεταξύ πραγματικών και παραγόμενων δεδομένων, ενώ το generator loss τη μεγιστοποιεί.

8. Ποια από τις παρακάτω στρατηγικές προσδιορισμού βέλτιστης πολιτικής για έναν πράκτορα αυτοενισχυόμενης μάθησης (reinforcement learning) συγκλίνει εγγυημένα σε βέλτιστες τιμές;

Α. Η επανάληψη πολιτικών (policy iteration)
Β. Η επανάληψη τιμών (value iteration)
Γ. Και οι δύο
Δ. Καμία από τις δύο

ΚΑΛΗ ΕΠΙΤΥΧΙΑ

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΚΑΙ ΒΑΘΙΑ ΜΑΘΗΣΗ

ΚΑΝΟΝΙΚΗ ΕΞΕΤΑΣΗ Καρκαν Δημητ 03400216 (Εδ αμμ)
21-06-2024

ΑΝΟΙΚΤΑ ΘΕΜΑΤΑ
ΔΙΑΡΚΕΙΑ 1:30

Θέμα 1 [18 μονάδες]

Υποθέστε οτι έχουμε ένα μοντέλο ταξινόμησης κειμένου βασισμένο σε autoregressive RNN generation, που ταξινομεί σχόλια ως προς το toxicity ενός κειμένου. Για την ταξινόμηση ενός σχολίου χρησιμοποιούμε prompting, με το template:

$$[X] \text{ Το σχόλιο είναι } [Y]$$

όπου το $[X]$ γεμίζει με το αρχικό σχόλιο προς ταξινόμηση και το $[Y]$ με την απόφαση ως προς το περιεχόμενο του σχολίου. Έστω t το βήμα που αντιστοιχεί στην παραγωγή του $[Y]$ και $w_1, ..., w_t$ οι λέξεις που έχουν παραχθεί συνολικά με βάση το παραπάνω prompt template. Η απόφαση για το αν είναι αποδεκτό το σχόλιο δίνεται από:

$$\hat{y} = \begin{cases} +1 & \text{if } P(w_t = \text{“αποδεκτό”} \mid w_1, ..., w_{t-1}) \geq P(w_t = \text{“τοξικό”} \mid w_1, ..., w_{t-1}), \\ -1 & \text{otherwise.} \end{cases}$$

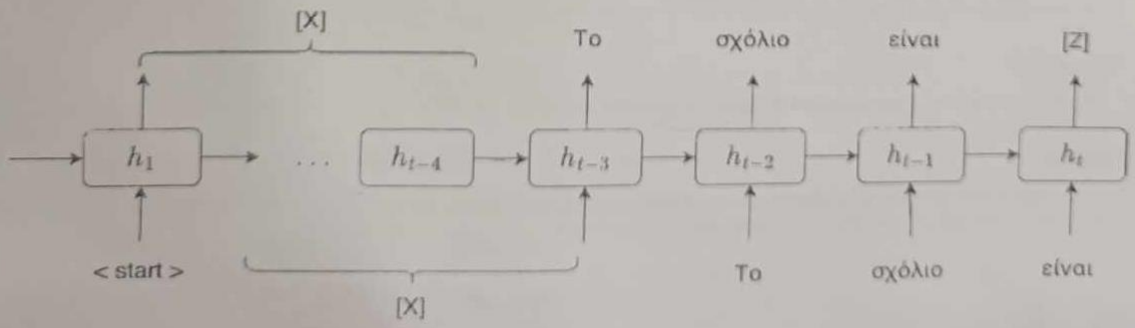
Ερώτημα 1.1:

Ας υποθέσουμε ότι το σχόλιο προς ταξινόμηση είναι “Every man must die.” Τα embeddings είναι

$$\begin{aligned} &x_{\langle\text{start}\rangle} = [0,0,0]^\top, \quad x_{\text{Every}} = [-1,0,0]^\top, \quad x_{\text{man}} = [1,0,0]^\top, \quad x_{\text{must}} = [0,0,0]^\top, \\ &x_{\text{die.}} = [0,5,0]^\top, \quad x_{\text{Το}} = [0,0,0]^\top, \quad x_{\text{σχόλιο}} = [-2,0,0]^\top, \quad x_{\text{είναι}} = [0,0,0]^\top, \\ &x_{\text{αποδεκτό}} = [1,2,3]^\top, \quad x_{\text{τοξικό}} = [-1,-2,-3]^\top, \end{aligned}$$

και οι παράμετροι του RNN είναι

$$\mathbf{W}_{hx} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & -1 \end{bmatrix}, \quad \mathbf{W}_{hh} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{W}_{yh} = \begin{bmatrix} 0 & 0 \\ 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \\ 9 & 10 \\ 11 & 12 \\ 13 & 14 \\ 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{matrix} \text{\# ⟨start⟩} \\ \text{\# Every} \\ \text{\# man} \\ \text{\# must} \\ \text{\# die.} \\ \text{\# Το} \\ \text{\# σχόλιο} \\ \text{\# είναι} \\ \text{\# αποδεκτό} \\ \text{\# τοξικό} \end{matrix}$$



Σχήμα 1: RNN αρχιτεκτονική του μοντέλου ταξινόμησης.

Όλα τα biases είναι 0. Η αρχική κατάσταση του RNN είναι $\mathbf{h}_0 = [0, 0]^\top$ και η συνάρτηση ενεργοποίησης για το RNN είναι η ReLU. Ποια θα είναι η τελική πρόβλεψη; Δείξτε όλους τους υπολογισμούς.

Ερώτημα 1.2:

Τώρα ας υποθέσουμε ότι αντικαθιστούμε το RNN και χρησιμοποιούμε ως γλωσσικό μοντέλο έναν (πολύ απλό) transformer (στην πραγματικότητα, μόνο ένα self-attention layer με μία μόνο κεφαλή προσοχής, χωρίς feed-forward layers και χωρίς residual σύνδεσεις. Τα word embeddings είναι οι ίδια με πριν, τα position embeddings ειναι:

$$p_1 = [0, 0, 0]^\top, \quad p_2 = [0, 0, 1]^\top, \quad p_3 = [0, 0, 2]^\top, \quad \ldots, \quad p_t = [0, 0, t-1]^\top.$$

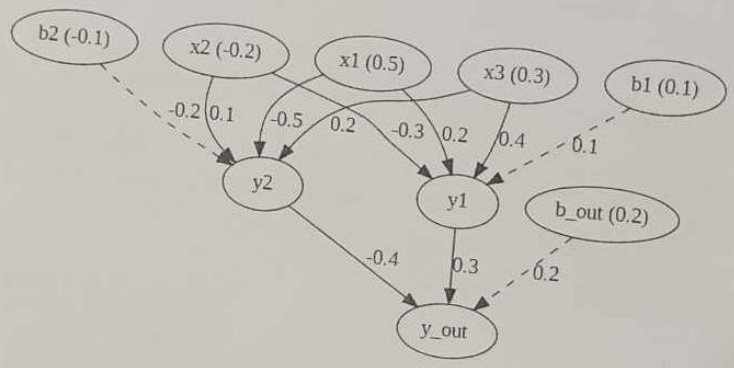
και τα projection matrices:

$$W_Q = W_K = W_V = \begin{bmatrix} 0.1 & 0 \\ 0 & -0.1 \\ 0 & 0.1 \end{bmatrix}.$$

Χρησιμοποιούμε scaled dot product attention για τον υπολογισμό των πιθανοτήτων προσοχής. Το linear layer εξόδου χρησιμοποιεί τον ίδιο πίνακα W_{yh} όπως παραπάνω. Ποια θα είναι η πρόβλεψη; Υποδείξτε όλους τους υπολογισμούς.
(Υπόδειξη: παρατηρήστε ότι δεν χρειάζεται να υπολογίσετε τον πλήρη πίνακα πιθανοτήτων προσοχής, μόνο τις πιθανότητες προσοχής που σχετίζονται με την τελευταία λέξη.)

Θέμα 2 [16 μονάδες]

Δίνεται το Multi-Layer Perceptron του Σχήματος.



Ερώτημα 2.1:

Θεωρώντας ότι οι συναρτήσεις ενεργοποίησης είναι η sigmoid για το κρυφό επίπεδο και η γραμμική (linear) για το επίπεδο εξόδου, να υπολογίσετε την έξοδο y_{out}.

Ερώτημα 2.2:

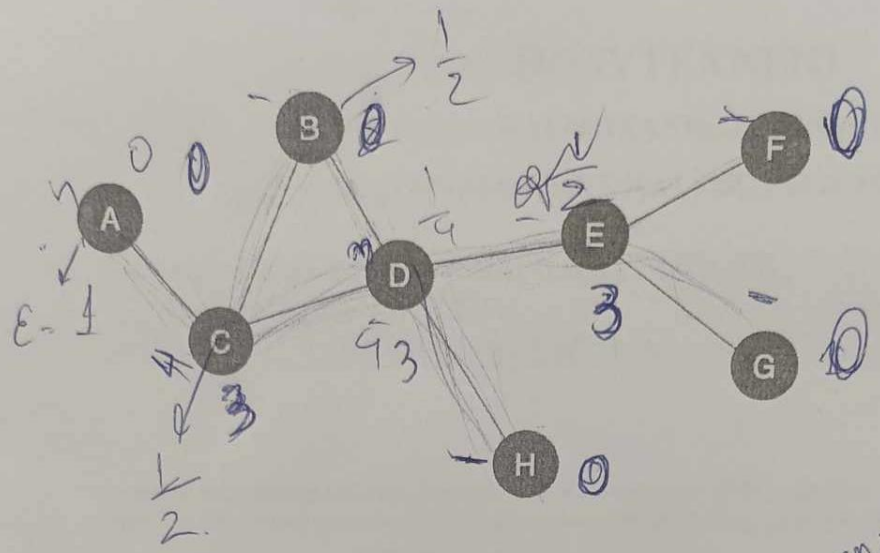
Αν η επιθυμητή έξοδος είναι y_{target}=0.21 να υπολογιστεί το loss function χωρίς regularization.

Ερώτημα 2.3:

Χρησιμοποιήστε το L2 regularization (Ridge regularization) με λ= 0.01 και υπολογίστε το συνολικό loss function με L2 regularization.

Ερώτημα 2.4:

Πώς επηρεάζει η επιλογή της συνάρτησης ενεργοποίησης τη συμπεριφορά και την απόδοση ενός MLP; Συγκρίνετε τη συνάρτηση σιγμοειδούς (sigmoid), την υπερβολική εφαπτομένη (tanh) και τη Rectified Linear Unit (ReLU).



Θέμα 3 [12 μονάδες]

Για τον γράφο του παραπάνω σχήματος να υπολογιστεί ο συντελεστής συσταδοποίησης και η ενδιαμεσότητα κάθε κόμβου. Με βάση τα μέτρα αυτά, τι συμπεράσματα θα μπορούσατε να βγάλετε σχετικά την σημαντικότητα των κόμβων;

ΚΑΛΗ ΕΠΙΤΥΧΙΑ

Θέμα 1

$h_0 = [0\ 0]^T$

Αρχικά έχω $h_1 = ReLU(W_{hx} \cdot x_{every} + W_{hh} \cdot h_0) =$

$= ReLU\left(\begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & -1 \end{bmatrix}\begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}\right) = ReLU\left(\begin{bmatrix} -1 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$h_2 = ReLU(W_{hx} \cdot x_{man} + W_{hh} \cdot h_1) = ReLU\left(\begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right) = ReLU\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$h_3 = ReLU(W_{hx} \cdot x_{must} + W_{hh} \cdot h_2) = ReLU\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$h_4 = ReLU(W_{hx} \cdot x_{die} + W_{hh} \cdot h_3) = ReLU\left(\begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & -1 \end{bmatrix}\begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) =$

$= ReLU\left(\begin{bmatrix} 5 \\ -5 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 6 \\ 0 \end{bmatrix}$ $h_5 = ReLU\left(\begin{bmatrix} 6 \\ 0 \end{bmatrix}\right.$