Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Εαρινό Εξάμηνο 2023-2024

ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΚΑΙ ΒΑΘΙΑ ΜΑΘΗΣΗ

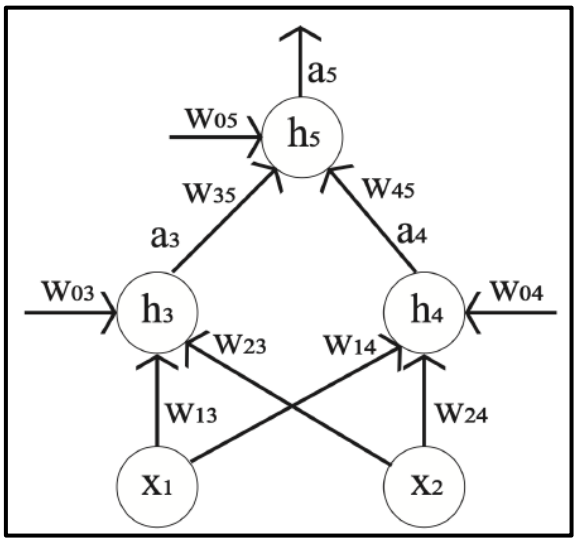
1η και 2η Σειρά Αναλυτικών Ασκήσεων 2023

Ιωάννης (Χουάν) Τσαντήλας
03120883

Contents

Άσκηση 1.1 (Multi Layer Perceptron- Regularization)

Δίνεται το πολυστρωματικό perceptron της εικόνας, το οποίο έχει 2 εισόδους x_1, x_2 στο στρώμα εισόδου, δύο νευρώνες h_3, h_4 στο κρυμμένο στρώμα και έναν νευρώνα h_5 στο στρώμα εξόδου.



Ερώτημα (α)

Έστω είσοδος $x_1 = 1, x_2 = -1$ και έστω Δw_{ij} η ανανέωση του βάρους w_{ij} κατά την οπισθοδιάδοση (backpropagation). Εάν $\Delta w_{03} = d_3$ και $\Delta w_{04} = d_4$, να εκφράσετε τα $\Delta w_{13}, \Delta w_{23}, \Delta w_{14}, \Delta w_{24}$ συνάρτηση των d_3, d_4.

Λύση

$$\Delta w_{03} = \eta \cdot \delta_3 \cdot x_0 \rightarrow \eta \cdot \delta_3 = \frac{\Delta w_{03}}{x_0} = \frac{d_3}{x_0}$$

$$\Delta w_{04} = \eta \cdot \delta_4 \cdot x_0 \rightarrow \eta \cdot \delta_4 = \frac{\Delta w_{04}}{x_0} = \frac{d_4}{x_0}$$

$$\Delta w_{13} = \eta \cdot \delta_3 \cdot x_1 = \frac{d_3 \cdot x_1}{x_0} = d_3$$

$$\Delta w_{23} = \eta \cdot \delta_3 \cdot x_2 = \frac{d_3 \cdot x_2}{x_0} = -d_3$$

$$\Delta w_{14} = \eta \cdot \delta_4 \cdot x_1 = \frac{d_4 \cdot x_1}{x_0} = d_4$$

$$\Delta w_{24} = \eta \cdot \delta_4 \cdot x_2 = \frac{d_4 \cdot x_2}{x_0} = -d_4$$

Ερώτημα (β)

Έστω ότι για τις παραπάνω εισόδους η επιθυμητή έξοδος-στόχος είναι $y = 1$. Δίνονται οι παρακάτω αρχικές τιμές βαρών και πολώσεων:

$$w_{03} = -1, w_{13} = 1, w_{23} = -1$$

$$w_{04} = 2, w_{14} = -1, w_{24} = 1$$

$$w_{05} = -2, w_{15} = 1, w_{25} = 1$$

Να υπολογίσετε την έξοδο του δικτύου και το σφάλμα. Στη συνέχεια, να υπολογίσετε τις ανανεωμένες τιμές των βαρών και πολώσεων (με ακρίβεια τριών δεκαδικών ψηφίων) μετά από ένα πέρασμα backpropagation. Τέλος, να υπολογίσετε τις νέες τιμές της εξόδου και του σφάλματος μετά την ανανέωση των βαρών. Θεωρήστε ότι ο ρυθμός μάθησης ισούται με 1.0 και ότι η συνάρτηση ενεργοποίησης είναι η σιγμοειδής $g(x) = 1/(1 + e^{-x})$.

Λύση

Επειδή δεν δίνονται, θεωρώ $w_{35} = w_{45} = 1$.

Υπολογισμοί

$$a_3 = w_{03} \cdot x_0 + w_{13} \cdot x_1 + w_{23} \cdot x_2 = (-1) \cdot 1 + 1 \cdot 1 + (-1) \cdot (-1) = 1$$

$$y_3 = \frac{1}{1 + e^{-1}} = 0.731$$

$$a_4 = w_{04} \cdot x_0 + w_{14} \cdot x_1 + w_{24} \cdot x_2 = 2 \cdot 1 + (-1) \cdot 1 + 1 \cdot (-1) = 0$$

$$y_4 = \frac{1}{1 + e^{0}} = 0.5$$

$$a_5 = w_{05} \cdot x_0 + w_{15} \cdot x_1 + w_{25} \cdot x_2 = (-2) \cdot 1 + 0.731 \cdot 1 + 0.5 \cdot 1 = -0.769$$

$$y_5 = \frac{1}{1 + e^{0.317}} = 0.317$$

Σφάλμα

$$y_{target} - y_5 = 1 - 0.317 = 0.983$$

Backward Pass

$$\delta_5 = (y_{target} - y_5) \cdot y_5 \cdot (1 - y_5) = 0.683 \cdot 0.317 \cdot 0.683 = 0.148$$

$$\delta_3 = y_3 \cdot (1 - y_3) \cdot w_{35} \cdot \delta_5 = 0.731 \cdot 0.269 \cdot 1 \cdot 0.148 = 0.037$$

$$\delta_4 = y_4 \cdot (1 - y_4) \cdot w_{45} \cdot \delta_5 = 0.5 \cdot 0.5 \cdot 1 \cdot 0.148 = 0.029$$

$$\Delta w_{03} = \eta \cdot \delta_3 \cdot x_0 = 1 \cdot 0.029 \cdot 1 = 0.029$$

$$\Delta w_{04} = \eta \cdot \delta_4 \cdot x_0 = 1 \cdot 0.037 \cdot 1 = 0.037$$

$$\Delta w_{05} = \eta \cdot \delta_5 \cdot x_0 = 1 \cdot 0.148 \cdot 1 = 0.148$$

$$\Delta w_{13} = \eta \cdot \delta_3 \cdot x_1 = 1 \cdot 0.029 \cdot 1 = 0.029$$

$$\Delta w_{14} = \eta \cdot \delta_4 \cdot x_1 = 1 \cdot 0.037 \cdot 1 = 0.037$$

$$\Delta w_{23} = \eta \cdot \delta_3 \cdot x_2 = 1 \cdot 0.029 \cdot (-1) = -0.029$$

$$\Delta w_{24} = \eta \cdot \delta_4 \cdot x_2 = 1 \cdot 0.037 \cdot (-1) = -0.037$$

$$\Delta w_{35} = \eta \cdot \delta_5 \cdot y_3 = 1 \cdot 0.148 \cdot 0.5 = 0.074$$

$$\Delta w_{45} = \eta \cdot \delta_5 \cdot y_4 = 1 \cdot 0.148 \cdot 0.731 = 0.108$$

Νέες Τιμές

$$w_{03} = -0.971, w_{13} = 1.029, w_{23} = -1.029$$

$$w_{04} = 2.037, w_{14} = -0.963, w_{24} = 0.963$$

$$w_{05} = -1.852, w_{35} = 1.108, w_{45} = 1.074$$

$$a_3 = w_{03} \cdot x_0 + w_{13} \cdot x_1 + w_{23} \cdot x_2 = (-0.971) \cdot 1 + 1.029 \cdot 1 + (-1.029) \cdot (-1) = 1.087$$

$$y_3 = \frac{1}{1 + e^{-1.087}} = 0.748$$

$$a_4 = w_{04} \cdot x_0 + w_{14} \cdot x_1 + w_{24} \cdot x_2 = 2.037 \cdot 1 + (-0.963) \cdot 1 + 0.983 \cdot (-1) = 0.111$$

$$y_4 = \frac{1}{1 + e^{-0.111}} = 0.528$$

$$a_5 = w_{05} \cdot x_0 + w_{15} \cdot x_1 + w_{25} \cdot x_2 = -1.852 \cdot 1 + 0.748 \cdot 1.108 + 0.528 \cdot 1.074 = -0.456$$

$$y_5 = \frac{1}{1 + e^{-0.456}} = 0.388$$

Σφάλμα

$$y_{target} - y_5 = 1 - 0.388 = 0.612$$

Μειώθηκε ελαφρώς.

Ερώτημα (γ)

Να επαναλάβετε τους υπολογισμούς του προηγούμενου ερωτήματος θεωρώντας τώρα ως συνάρτηση ενεργοποίησης την υπερβολική εφαπτομένη $g(x) = tanh(x)$.

Λύση

Υπολογισμοί

$$a_3 = w_{03} \cdot x_0 + w_{13} \cdot x_1 + w_{23} \cdot x_2 = (-1) \cdot 1 + 1 \cdot 1 + (-1) \cdot (-1) = 1$$

$$y_3 = \tanh(1) = 0.762$$

$$a_4 = w_{04} \cdot x_0 + w_{14} \cdot x_1 + w_{24} \cdot x_2 = 2 \cdot 1 + (-1) \cdot 1 + 1 \cdot (-1) = 0$$

$$y_4 = \tanh(0) = 0$$

$$a_5 = w_{05} \cdot x_0 + w_{35} \cdot y_3 + w_{45} \cdot y_4 = (-2) \cdot 1 + 1 \cdot 0.762 + 1 \cdot 0 = -1.238$$

$$y_5 = r = tanh(-1.238) = -0.845$$

Σφάλμα

$$y_{target} - y_5 = 1 + 0.845 = 1.845$$

Backward Pass

$$g(x) = tanh(x), \qquad g'(x) = 1 - tanh^2(x)$$

$$\Delta_{wkj} = \varepsilon \cdot (t_k - a_k) \cdot \left(1 - a_k^2\right) \cdot a_\text{j} \rightarrow$$

$$\rightarrow \delta_k = \left(y_{target} - y_k\right) \cdot \left(1 - y_k^2\right) \text{ \# output}$$

$$\rightarrow \delta_k = \left(1 - y_k^2\right) \cdot \sum\nolimits_\text{j} W_{kj} \cdot \delta_j \text{ \# hidden}$$

$$\delta_5 = \left(y_{target} - y_5\right) \cdot \left(1 - y_5^2\right) = 1.845 \cdot (1 - 0.845^2) = 0.528$$

$$\delta_3 = (1 - y_3\text{^}2) \cdot w_{35} \cdot \delta_5 = 0.419 \cdot 1 \cdot 0.528 = 0.221$$

$$\delta_4 = (1 - y_4\text{^}2) \cdot w_{45} \cdot \delta_5 = 1 \cdot 1 \cdot 0.528 = 0.528$$

$$\Delta w_{03} = \eta \cdot \delta_3 \cdot x_0 = 1 \cdot 0.221 \cdot 1 = 0.221$$

$$\Delta w_{04} = \eta \cdot \delta_4 \cdot x_0 = 1 \cdot 0.528 \cdot 1 = 0.528$$

$$\Delta w_{05} = \eta \cdot \delta_5 \cdot x_0 = 1 \cdot 0.528 \cdot 1 = 0.528$$

$$\Delta w_{13} = \eta \cdot \delta_3 \cdot x_1 = 1 \cdot 0.221 \cdot 1 = 0.221$$

$$\Delta w_{14} = \eta \cdot \delta_4 \cdot x_1 = 1 \cdot 0.528 \cdot 1 = 0.528$$

$$\Delta w_{23} = \eta \cdot \delta_3 \cdot x_2 = 1 \cdot 0.221 \cdot (-1) = -0.221$$

$$\Delta w_{24} = \eta \cdot \delta_4 \cdot x_2 = 1 \cdot 0.528 \cdot (-1) = -0.528$$

$$\Delta w_{35} = \eta \cdot \delta_5 \cdot y_3 = 1 \cdot 0.528 \cdot 0 = 0$$

$$\Delta w_{45} = \eta \cdot \delta_5 \cdot y_4 = 1 \cdot 0.528 \cdot 0.762 = 0.402$$

Νέες Τιμές

$$w_{03} = -0.779, w_{13} = 1.221, w_{23} = -1.221$$

$$w_{04} = 2.528, w_{14} = -0.472, w_{24} = 0.472$$

$$w_{05} = -1.472, w_{35} = 1.402, w_{45} = 1$$

$$a_3 = w_{03} \cdot x_0 + w_{13} \cdot x_1 + w_{23} \cdot x_2 = (-0.779) \cdot 1 + 1.221 \cdot 1 + (-1.221) \cdot (-1) = 1.663$$

$$y_3 = \tanh(1.663) = 0.931$$

$$a_4 = w_{04} \cdot x_0 + w_{14} \cdot x_1 + w_{24} \cdot x_2 = 2.528 \cdot 1 + (-0.472) \cdot 1 + 0.472 \cdot (-1) = 1.584$$

$$y_4 = \tanh(1.584) = 0.919$$

$$a_5 = w_{05} \cdot x_0 + y_3 \cdot w_{15} + y_4 \cdot w_{25} = -1.472 \cdot 1 + 0.931 \cdot 1.402 + 0.919 \cdot 1 = 0.752$$

$$y_5 = tanh(0.752) = 0.636$$

Σφάλμα

$$y_{target} - y_5 = 1 - 0.636 = 0.364$$

Μειώθηκε.

Ερώτημα (δ)

Έστω δίκτυο στην εκπαίδευση του οποίου γίνεται χρήση παραμέτρου ομαλοποίησης (regularization). Εξηγήστε ποια θα είναι η επίδραση της αύξησης της παραμέτρου ομαλοποίησης:

- στην ορθότητα επί των δεδομένων εκπαίδευσης (training accuracy)
- στην ορθότητα επί των δεδομένων ελέγχου (testing accuracy)

Λύση

Μείωση και Αύξηση.

Άσκηση 1.2 (Convolutional Neural Networks)

Ας υποθέσουμε ότι έχουμε ένα δίκτυο σαν το AlexNet (Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", NeurIPS 2012). Θεωρήστε πως έχουμε εικόνες διαστάσεων 227x227x3 (έγχρωμες με RGB channels) και φίλτρο 11x11x3 στο πρώτο convolutional layer. Το δίκτυο έχει συνολικά 96 φίλτρα, stride ίσο με 4 και μηδενικό padding.

a) Υπολογίστε τις διαστάσεις στην έξοδο του πρώτου convolutional layer.
b) Υπολογίστε τον αριθμό των layer στο πρώτο convolutional layer.
c) Υπολογίστε τον αριθμό των εκπαιδεύσιμων παραμέτρων του πρώτου convolutional layer με διαμοιρασμό βαρών.
d) Αν αντικαθιστούσαμε το CNN με ένα FeedForward layer με 256 units πόσες εκπαιδεύσιμες παραμέτρους θα είχαμε;

Λύση

- Διαστάσεις εικόνας εισόδου: $227 \times 227 \times 3$
- Μέγεθος φίλτρου: $F = 11 \times 11 \times 3$
- Αριθμός φίλτρων: $K = 96$
- Stride: $S = 4$
- Padding: $P = 0$
- Πλάτος, ύψος, βάθος εισόδου: $w_1 = 227, h_1 = 227, d_1 = 3$
- Πλάτος, ύψος, βάθος εξόδου: w_2, h_2, d_2

Ερώτημα (α)

Ο τύπος για τον υπολογισμό του πλάτους και του ύψους εξόδου για ένα συνελικτικό στρώμα είναι:

$$w_2 = \left(\frac{w_1 - F + 2 \cdot P}{S}\right) + 1 = \left(\frac{227 - 11 + 2 \cdot 0}{S}\right) + 1 = 55$$

$$h_2 = \left(\frac{h_1 - F + 2 \cdot P}{S}\right) + 1 = 55$$

$$d_2 = K = 96$$

Επομένως, οι διαστάσεις εξόδου του πρώτου επιπέδου συνελίξεων είναι $55 \times 55 \times 96$.

Ερώτημα (β)

Για να βρούμε τον αριθμό των μονάδων (ή ενεργοποιήσεων) στον όγκο εξόδου, πολλαπλασιάζουμε τις διαστάσεις μαζί:

$$Units = w_2 \cdot h_2 \cdot d_2 = 55 \cdot 55 \cdot 96 = 290.400$$

Ερώτημα (γ)

Ο αριθμός των εκπαιδεύσιμων παραμέτρων σε ένα συνελικτικό στρώμα υπολογίζεται ως εξής (ο όρος *+1* αντιπροσωπεύει τον όρο μεροληψίας σε κάθε φίλτρο):

$$Parameters = (F \cdot F \cdot d_1 + 1) \cdot K$$

Αντικαθιστώντας τις τιμές:

$$Parameters = (F \cdot F \cdot d_1 + 1) \cdot K = (11 \cdot 11 \cdot 3 + 1) \cdot 96 = 34{,}944$$

Ερώτημα (δ)

Σε ένα πλήρως συνδεδεμένο επίπεδο, κάθε μονάδα εισόδου συνδέεται με κάθε μονάδα εξόδου. Ο αριθμός των εκπαιδεύσιμων παραμέτρων υπολογίζεται ως εξής (ο όρος *+1* αντιπροσωπεύει τον όρο μεροληψίας):

$$\text{Parameters} = (\text{Input Units} + 1) \cdot \text{Output Units}$$

Οι μονάδες εισόδου σε αυτή την περίπτωση είναι όλα τα εικονοστοιχεία της εικόνας εισόδου:

$$\text{Input units} = 227 \cdot 227 \cdot 3 = 154{,}588$$

Οι μονάδες εξόδου δίνονται 256. Επομένως, ο υπολογισμός είναι ο εξής:

$$\text{Parameters} = 154{,}588 \cdot 256 = 39{,}574{,}528$$

Άσκηση 1.3 (Recurrent Neural Networks)

Υποθέστε ότι θέλουμε να εκπαιδεύσουμε ένα μικρό δίκτυο RNN για την εργασία της ταξινόμησης σύντομου κειμένου (short text classification), το οποίο εκπαιδεύεται να ταξινομεί κείμενα με βάση το συναίσθημα με τις πιθανές κλάσεις να ανήκουν στο $C = \{0; 1; 2; \}$. Το δίκτυο είναι υλοποιημένο ως εξής:

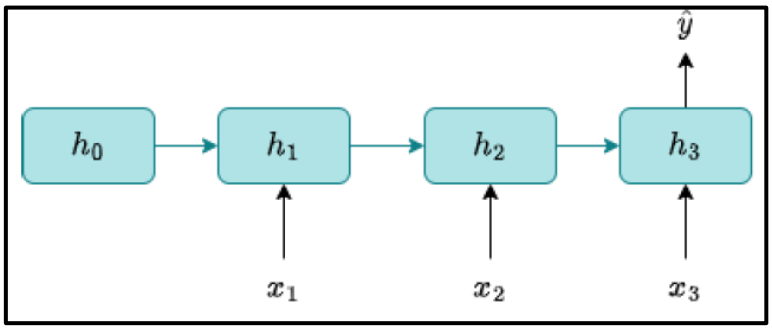
- Οι λέξεις κωδικοποιούνται με τη μέθοδο one-hot.
- Η αρχική κρυμμένη κατάσταση (hidden-state) του RNN h_0 είναι αρχικοποιημένη στο 0.
- Τα input-to-hidden matrix και hidden-to-output matrix είναι αντίστοιχα:

$$W_{wh} = \begin{bmatrix} 0 & -1 & 2 \\ 1 & -2 & 0 \end{bmatrix}, \qquad W_{hy} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & -1 \end{bmatrix}$$

- Ο πίνακας W_{hh} είναι ο μοναδιαίος πίνακας.
- Όλες οι πολώσεις (bias) b ισούνται με 0.
- Το RNN δίκτυο χρησιμοποιεί τη συνάρτηση αρχικοποίησης ReLU.
- Χρησιμοποιείται η τελευταία κρυμμένη κατάσταση h_3 για τον υπολογισμό της εξόδου.

Το μοντέλο χρησιμοποιεί μία αναπαράσταση 3×1 για κάθε λέξη (token) και κάνει χρήση ενός προϋπολογισμένου λεξικού $\mathcal{V}$, μέρος του οποίου παρατίθεται ακολούθως:

$$\mathcal{V} = \begin{bmatrix} 0 & -1 & 2 \\ 1 & -2 & 0 \\ 3 & 1 & -2 \\ -2 & -1 & 1 \\ 4 & 1 & -2 \\ -2 & -2 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{matrix} \#\ this \\ \#\ is \\ \#\ good \\ \#\ bad \\ \#\ great \\ \#\ horrible \\ \#\ <UNK> \end{matrix}$$



Ερώτημα 1

Πώς θα ταξινομήσει το μοντέλο την πρόταση "This was horrible"; Κάνετε αναλυτικά τους υπολογισμούς.

Λύση

$$h_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad h_t = g(V \cdot x_t + U \cdot h_{t-1} + C), \qquad y = softmax(W \cdot h_t + b)$$

$$Bias\ C = b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad U = W_{hh} = I, \qquad V = \begin{bmatrix} 0 & -1 & 2 \\ 1 & -2 & 0 \end{bmatrix}, \qquad W = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & -1 \end{bmatrix}$$

$$g = ReLU, ReLU: f(x) = max(0, x)$$

Κωδικοποίηση Λέξεων

$$This: x_1 \begin{bmatrix} 0 \\ -1 \\ 2 \end{bmatrix}, \qquad Was: x_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad Horrible: x_3 = \begin{bmatrix} -2 \\ -2 \\ 1 \end{bmatrix}$$

Υπολογισμός Κρυμμένων Καταστάσεων

$$h_1 = ReLU\left(\begin{bmatrix} 0 \cdot 0 + (-1) \cdot (-1) + 2 \cdot 2 \\ 1 \cdot 0 + (-2) \cdot (-1) + 0 \cdot 2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

$$h_2 = ReLU\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 5 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

$$h_3 = ReLU\left(\begin{bmatrix} 0 \cdot (-2) + (-1) \cdot (-2) + 2 \cdot 1 \\ 1 \cdot (-2) + (-2) \cdot (-2) + 0 \cdot 1 \end{bmatrix} + \begin{bmatrix} 5 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 4 \\ 2 \end{bmatrix} + \begin{bmatrix} 5 \\ 2 \end{bmatrix} = \begin{bmatrix} 9 \\ 4 \end{bmatrix}$$

$$y = softmax\left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & -1 \end{bmatrix} \cdot \begin{bmatrix} 9 \\ 4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = softmax\left(\begin{bmatrix} 4 \\ 9 \\ 14 \end{bmatrix}\right)$$

Η μεγαλύτερη τιμή είναι 14, δηλαδή η κλάση είναι 2. Άρα Αρνητικό Συναίσθημα.

Ερώτημα 2

Είναι το μοντέλο καλά εκπαιδευμένο; Αιτιολογήστε την απάντησή σας.

Λύση

Παρατηρούμε ότι για το συγκεκριμένο παράδειγμα το μοντέλο δίνει σωστό αποτέλεσμα. Όμως τρέξαμε μόνο 1 παράδειγμα άρα δεν ξέρουμε σίγουρα ότι δουλεύει γενικά για όλες τις περιπτώσεις. επίσης το λεξικό V περιέχει λίγες λέξεις που δεν καλύπτουν γενικά πολλές περιπτώσεις. Στο παράδειγμα η λέξη “was” δεν υπήρχε στο λεξικό V. Δεν μπορούμε να πούμε 100% ότι το μοντέλο είναι καλά εκπαιδευμένο.

Ερώτημα 3

Υποθέστε ότι αντί για το τελευταίο hidden state το μοντέλο χρησιμοποιεί average pooling από τα h_1, h_2 και h_3. Πώς θα αλλάξει η πρόβλεψη για το ερώτημα 1;

Λύση

Η πρόβλεψη θα είναι:

$$h = \frac{h_1 + h_2 + h_3}{3} = \begin{bmatrix} \frac{5 + 5 + 9}{3} \\ \frac{2 + 2 + 4}{3} \end{bmatrix} = \begin{bmatrix} \frac{19}{3} \\ \frac{8}{3} \end{bmatrix}$$

$$y = softmax\left(\begin{bmatrix}0 & 1\\ 1 & 0\\ 2 & -1\end{bmatrix}\cdot\begin{bmatrix}19/3\\ 8/3\end{bmatrix}+\begin{bmatrix}0\\ 0\end{bmatrix}\right) = softmax\left(\begin{bmatrix}8/3\\ 19/3\\ 10\end{bmatrix}\right) = softmax\left(\begin{bmatrix}2.67\\ 6.33\\ 10\end{bmatrix}\right)$$

Η μεγαλύτερη τιμή είναι 10, δηλαδή η κλάση είναι 2. Άρα Αρνητικό Συναίσθημα όπως πριν.

Άσκηση 1.4 (Representation Learning – Autoencoders)

Έστω ότι έχουμε πρόσβαση σε αυξημένες υπολογιστικές υποδομές και εκπαιδεύουμε ένα skipgram μοντέλο για ένα μεγαλύτερο λεξιλόγιο $\mathcal{V}'$. Το $\mathcal{V}'$ περιέχει αναπαραστάσεις 1500 λέξεων (μαζί με τα special tokens) και η διάσταση των διανυσμάτων u_o και u_c είναι (256×1). Στη συνέχεια χρησιμοποιούμε τα skipgram vectors και εκπαιδεύουμε έναν αυτοκωδικοποιητή (auto-encoder) με 5 κρυμμένο στρώμα διαστάσεων $[500,250,50,250,500]$ αντιστοίχως. Απαντήστε στα παρακάτω ερωτήματα:

a) Ποια είναι η διάσταση των χαρακτηριστικών εισόδου x_i στον auto-encoder;
b) Ποια είναι η διάσταση των χαρακτηριστικών εξόδου y_i του auto-encoder;
c) Ποια είναι η διάσταση της λανθάνουσας αναπαράστασης (latent representation) του auto-encoder;
d) Αν θέλουμε να χρησιμοποιήσουμε τον εκπαιδευμένο auto-encoder κάνοντας fine-tuning για το πρόβλημα της άσκησης 1.3, περιγράψτε τη διαδικασία και τις σχετικές συναρτήσεις και διαστάσεις κάθε στρώματος.

Λύση

Ερώτημα (α)

- **Μέγεθος λεξιλογίου ($\mathcal{V}'$)**: 1500 λέξεις
- **Διάσταση διανύσματος**: Κάθε διάνυσμα λέξεων έχει μέγεθος 256×1.
- **Διανύσματα πλαισίου και στόχου**: Χρησιμοποιούνται τόσο τα διανύσματα u_o (έξοδος) όσο και τα διανύσματα u_c (πλαίσιο), δηλαδή 2 διανύσματα ανά λέξη.

Έτσι, για κάθε λέξη, έχουμε:

$$\text{Total Dimension per Word} = 256\ (\text{από } u_o) + 256\ (\text{από } u_c) = 512$$

Δεδομένου ότι το μέγεθος του λεξιλογίου είναι 1500, η συνολική διάσταση των χαρακτηριστικών εισόδου x_i είναι:

$$\text{Dimension of } x_i = 1500 \cdot 512 = 768{,}000$$

Ερώτημα (β)

Σε έναν αυτόματο κωδικοποιητή, τα χαρακτηριστικά εξόδου y_i είναι ανακατασκευασμένες είσοδοι. Επομένως, η διάσταση των χαρακτηριστικών εξόδου y_i θα είναι η ίδια με τη διάσταση των χαρακτηριστικών εισόδου x_i.

Συνεπώς, η διάσταση των χαρακτηριστικών εξόδου y_i είναι επίσης 768,000.

Ερώτημα (γ)

Η λανθάνουσα αναπαράσταση του αυτοκωδικοποιητή καθορίζεται από το μέγεθος του μικρότερου κρυμμένου στρώματος του δικτύου, το οποίο είναι συνήθως το στρώμα συμφόρησης. Δεδομένων των διαστάσεων του κρυμμένου στρώματος: $[500,250,50,250,500]$. Η μικρότερη διάσταση είναι 50. Συνεπώς, η διάσταση της λανθάνουσας αναπαράστασης είναι 50.

Ερώτημα (δ)

Άσκηση 2.1 (Attention – Transformers)

Ερώτημα 1

Έστω η εξής ακολουθία μήκους 4:

$$x^{(1)} = [-2 \quad 1 \quad 0.5]^T, \qquad x^{(2)} = [1 \quad 1.5 \quad -0.5]^T,$$
$$x^{(3)} = [-1.5 \quad 1 \quad -0.5]^T, \qquad x^{(4)} = [-2 \quad -2.5 \quad 1.5]^T$$

της οποίας κάθε στοιχείο είναι ένα διάνυσμα (vector) στο R^3. Έστω επίσης ότι με βάση τα παραπάνω διανύσματα έχουμε έναν πίνακα εισόδου (input matrix) $X = \left[x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}\right]^T \in R^{4\times3}$. Έστω ότι θέλουμε να χρησιμοποιήσουμε ένα επίπεδο προσοχής (attention layer) στη συγκεκριμένη είσοδο με query vector $q = [-2.0, -1.0, -1.0]^T$. Υπολογίστε τις πιθανότητες προσοχής (attention probabilities) και το διάνυσμα εξόδου (output vector) με βάση το query X και κάνοντας χρήση scaled dot product attention.

Λύση

Υπολογισμός Πιθανότητας Προσοχής (attention probabilities)

$$\text{Score} = X \cdot q = \begin{bmatrix} -2 & 1 & 0.5 \\ 1 & 1.5 & -0.5 \\ -1.5 & 1 & -0.5 \\ -2 & -2.5 & 1.5 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 4 + 1 - 0.5 \\ -2 + 1.5 + 0.5 \\ 3 + 1 + 0.5 \\ 4 - 2.5 - 1.5 \end{bmatrix} = \begin{bmatrix} 4.5 \\ 0 \\ 4.5 \\ 0 \end{bmatrix}$$

Scaled Dot Product

$$d_k = columns = 3$$

$$Scale = \frac{\text{score}}{\sqrt{d_k}} = \begin{bmatrix} \frac{4.5}{\sqrt{3}} \\ 0 \\ \frac{4.5}{\sqrt{3}} \\ 0 \end{bmatrix} = \begin{bmatrix} 2.598 \\ 0 \\ 2.598 \\ 0 \end{bmatrix}$$

Softmax

$$\sum_{i=1}^{4} e^{xi} = e^{\frac{4.5}{\sqrt{3}}} + e^0 + e^{\frac{4.5}{\sqrt{3}}} + e^0 = 28.876$$

$$softmax([2.598 \quad 0 \quad 2.598 \quad 0]) = \left[\frac{2.598}{28.876} \quad \frac{1}{28.876} \quad \frac{2.598}{28.876} \quad \frac{1}{28.876}\right] =$$
$$= [0.465 \quad 0.035 \quad 0.465 \quad 0.035]$$

Διανύσματα Εξόδου

$$y_1 = 0.465 \cdot x^{(1)} = [0.93 \quad 0.465 \quad 0.232]$$
$$y_2 = 0.035 \cdot x^{(2)} = [0.035 \quad 0.525 \quad -0.018]$$
$$y_3 = 0.465 \cdot x^{(3)} = [-0.698 \quad 0.465 \quad -0.232]$$

$$y_4 = 0.035 \cdot x^{(4)} = [\text{-}0.07 \quad \text{-}0.088 \quad 0.018]$$

Ερώτημα 2

Ας υποθέσουμε στη συνέχεια ότι για τη χρήση ενός query vector θέλουμε να υπολογίσουμε (single-head) self-attention για την ίδια είσοδο. Έστω ότι τα projection matrices για τα queries, keys, και values είναι αντίστοιχα:

$$W_Q = \begin{bmatrix} 1 & -1.5 \\ 0 & 2 \\ -0.5 & 1 \end{bmatrix}, \qquad W_K = \begin{bmatrix} -1.5 & -1 \\ 2.5 & 0 \\ 0.5 & -1 \end{bmatrix}, \qquad W_V = \begin{bmatrix} 1 & 2.5 \\ -0.5 & -2 \\ 0 & -1 \end{bmatrix}$$

Υπολογίστε τους πίνακες query, key, και value (Q, K, V), καθώς και το διάνυσμα εξόδου Z.

Λύση

$$Q = X \cdot W_Q = \begin{bmatrix} \text{-}2 & 1 & 0.5 \\ 1 & 1.5 & \text{-}0.5 \\ \text{-}1.5 & 1 & \text{-}0.5 \\ \text{-}2 & \text{-}2.5 & 1.5 \end{bmatrix} \cdot \begin{bmatrix} 1 & -1.5 \\ 0 & 2 \\ -0.5 & 1 \end{bmatrix} = \begin{bmatrix} \text{-}2.25 & 4.5 \\ 1.25 & 2 \\ \text{-}1.25 & 4.75 \\ \text{-}2.75 & \text{-}3.5 \end{bmatrix}$$

$$\mathrm{K} = X \cdot W_{\mathrm{K}} = \begin{bmatrix} \text{-}2 & 1 & 0.5 \\ 1 & 1.5 & \text{-}0.5 \\ \text{-}1.5 & 1 & \text{-}0.5 \\ \text{-}2 & \text{-}2.5 & 1.5 \end{bmatrix} \cdot \begin{bmatrix} -1.5 & -1 \\ 2.5 & 0 \\ 0.5 & -1 \end{bmatrix} = \begin{bmatrix} 5.75 & 1.5 \\ 2 & \text{-}0.5 \\ 4.5 & 2 \\ \text{-}2.5 & 0.5 \end{bmatrix}$$

$$\mathrm{V} = X \cdot W_{\mathrm{V}} = \begin{bmatrix} \text{-}2 & 1 & 0.5 \\ 1 & 1.5 & \text{-}0.5 \\ \text{-}1.5 & 1 & \text{-}0.5 \\ \text{-}2 & \text{-}2.5 & 1.5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2.5 \\ -0.5 & -2 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} \text{-}2.5 & \text{-}6.5 \\ 0.25 & \text{-}1 \\ \text{-}2 & \text{-}6.25 \\ 0.75 & 1.5 \end{bmatrix}$$

$$z = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \qquad d_k = 2$$

$$QK^T = \begin{bmatrix} \text{-}2.25 & 4.5 \\ 1.25 & 2 \\ \text{-}1.25 & 4.75 \\ \text{-}2.75 & \text{-}3.5 \end{bmatrix} \begin{bmatrix} 5.75 & 2 & 4.5 & \text{-}2.5 \\ 1.5 & \text{-}0.5 & 2 & 0.5 \end{bmatrix} = \begin{bmatrix} \text{-}6.188 & \text{-}6.75 & \text{-}1.125 & 7.878 \\ 10.188 & 1.5 & 9.625 & \text{-}2.125 \\ \text{-}0.03 & \text{-}4.875 & \text{-}3.875 & 5.5 \\ \text{-}21.063 & \text{-}3.75 & \text{-}19.375 & 5.125 \end{bmatrix}$$

$$z = softmax\left(\frac{\begin{bmatrix} \text{-}6.188 & \text{-}6.75 & \text{-}1.125 & 7.878 \\ 10.188 & 1.5 & 9.625 & \text{-}2.125 \\ \text{-}0.03 & \text{-}4.875 & \text{-}3.875 & 5.5 \\ \text{-}21.063 & \text{-}3.75 & \text{-}19.375 & 5.125 \end{bmatrix}}{\sqrt{2}}\right) \cdot V =$$

$$= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0.6 & 0 & 0.4 & 0 \\ 0.02 & 0 & 0 & 0.98 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \text{-}2.5 & \text{-}6.5 \\ 0.25 & \text{-}1 \\ \text{-}2 & \text{-}6.25 \\ 0.75 & 1.5 \end{bmatrix} = \begin{bmatrix} \text{-}0.75 & 1.5 \\ \text{-}2.3 & \text{-}6.4 \\ \text{-}0.785 & 1.34 \\ \text{-}0.75 & 1.5 \end{bmatrix}$$

Άσκηση 2.2 (Generative Models)

Έστω ότι θέλετε να δημιουργήσετε ρεαλιστικές εικόνες χταποδιών χρησιμοποιώντας Generative Adversarial Networks (GANs). Ωστόσο δυστυχώς δεν έχετε καταφέρει να συγκεντρώσετε μεγάλο αριθμό φωτογραφιών από αληθινά χταπόδια, και γι' αυτό αποφασίζετε να χρησιμοποιήσετε τεχνικές επαύξησης δεδομένων (data augmentation) με την ελπίδα ότι με αυτόν τον τρόπο θα βελτιώσετε την ποιότητα της εκπαίδευσης του δικτύου σας. Δοκιμάζετε τρεις δημοφιλείς τεχνικές επαύξησης:

1. Θόλωμα εικόνων.
2. Αλλαγή χρώματος των pixels.
3. Περιστροφή (flip) ως προς τον κεντρικό κατακόρυφο άξονα της εικόνας (άξονας "αριστερά-δεξιά").

Ποιες από τις παραπάνω τεχνικές θεωρείτε ότι θα ήταν κατάλληλες για να βελτιώσουν την ποιότητα των εικόνων που παράγει ο Generator; Αιτιολογήστε αναλυτικά την απάντησή σας. Εάν μπορείτε, προτείνετε κάποια επιπλέον τεχνική που θεωρείτε ότι θα ήταν αποτελεσματική για τον ίδιο σκοπό.

Λύση

Θόλωση εικόνων

- **Καταλληλότητα**: Η θόλωση μπορεί μερικές φορές να βοηθήσει τους GAN εισάγοντας μεταβλητότητα στην ευκρίνεια της εικόνας, η οποία μπορεί να κάνει τη γεννήτρια ανθεκτική σε μικρές μεταβολές στην εστίαση. Ωστόσο, η υπερβολική θόλωση μπορεί να υποβαθμίσει την ποιότητα και την ευκρίνεια των παραγόμενων εικόνων, πράγμα ανεπιθύμητο για τη δημιουργία ρεαλιστικών εικόνων.
- **Αποτελεσματικότητα**: Γενικά, αυτή η τεχνική προτιμάται λιγότερο για τη βελτίωση της απόδοσης του GAN, επειδή δεν ενισχύει σημαντικά την ποικιλομορφία των δεδομένων εκπαίδευσης.

Αλλαγή του χρώματος των εικονοστοιχείων

- **Καταλληλότητα**: Οι αλλαγές χρώματος μπορούν να βοηθήσουν τη γεννήτρια να μάθει το αναλλοίωτο χρώματος και να βελτιώσει την ικανότητά της να γενικεύει σε διαφορετικές συνθήκες φωτισμού και χρωματικές παραλλαγές. Ωστόσο, αν δεν γίνει προσεκτικά, μπορεί να εισαγάγει μη ρεαλιστικά χρωματικά μοτίβα που δεν εμφανίζονται στη φύση.
- **Αποτελεσματικότητα**: Αυτή η τεχνική μπορεί να είναι μέτρια αποτελεσματική, ειδικά όταν ο στόχος είναι να διασφαλιστεί ότι η GAN δεν προσαρμόζεται υπερβολικά στα συγκεκριμένα χρώματα που υπάρχουν στο περιορισμένο σύνολο δεδομένων.

Ανατροπή της εικόνας κατά μήκος του κεντρικού κατακόρυφου άξονα (άξονας αριστερά-δεξιά)

- **Καταλληλότητα**: Η αναστροφή είναι μια απλή αλλά ισχυρή τεχνική επαύξησης που διπλασιάζει τον όγκο των δεδομένων εκπαίδευσης χωρίς να μεταβάλλει τα βασικά χαρακτηριστικά της εικόνας. Για συμμετρικά αντικείμενα όπως τα χταπόδια, η αναστροφή μπορεί να αυξήσει αποτελεσματικά την ποικιλομορφία των παραδειγμάτων εκπαίδευσης.
- **Αποτελεσματικότητα**: Πρόκειται για μια εξαιρετικά αποτελεσματική τεχνική για την εκπαίδευση του GAN, επειδή διατηρεί τις χωρικές σχέσεις εντός της εικόνας, ενώ αυξάνει το μέγεθος και την ποικιλομορφία του συνόλου δεδομένων.

Τυχαία περικοπή και κλιμάκωση

- **Καταλληλότητα**: Η τυχαία περικοπή και κλιμάκωση μπορεί να εισαγάγει μεταβλητότητα στη θέση και την κλίμακα του αντικειμένου μέσα στην εικόνα. Αυτό βοηθά τη γεννήτρια να μάθει να παράγει εικόνες με το αντικείμενο σε διαφορετικές θέσεις και μεγέθη, ενισχύοντας την ανθεκτικότητά της.
- **Αποτελεσματικότητα**: Αυτή η τεχνική έχει αποδειχθεί ιδιαίτερα αποτελεσματική στη βελτίωση της απόδοσης των GAN, παρέχοντας ποικίλα δείγματα εκπαίδευσης που αντιπροσωπεύουν καλύτερα τις παραλλαγές που απαντώνται στις εικόνες του πραγματικού κόσμου.

Άσκηση 2.3 (Self-Supervised Learning)

Εξηγήστε τις βασικές αρχές της αυτοεπιβλεπόμενης αντιθετικής μάθησης (self-supervised contrastive learning). Μετά από κατάλληλη αναζήτηση και μελέτη στην πρόσφατη βιβλιογραφία, περιγράψτε και συγκρίνετε δύο τεχνικές αυτοεπιβλεπόμενης αντιθετικής μάθησης με εφαρμογή στο πεδίο της ανάλυσης εικόνας / όρασης υπολογιστών και δύο (διαφορετικές) τεχνικές με εφαρμογή στο πεδίο της γλωσσικής μοντελοποίησης / επεξεργασίας φυσικής γλώσσας.

Λύση

Η αυτοεπιβλεπόμενη αντιθετική μάθηση είναι μια τεχνική που χρησιμοποιείται για την εκπαίδευση μοντέλων χωρίς την ανάγκη για δεδομένα με ετικέτες. Η βασική ιδέα είναι η εκμάθηση ενός χώρου ενσωμάτωσης, όπου παρόμοια σημεία δεδομένων (θετικά) συγκεντρώνουν, ενώ ανόμοια (αρνητικά) απομακρύνονται. Αυτό επιτυγχάνεται με τη δημιουργία θετικών ζευγών μέσω επαυξήσεων του ίδιου σημείου δεδομένων και με τη χρήση άσχετων σημείων δεδομένων ως αρνητικών. Ο στόχος εκπαίδευσης περιλαμβάνει συνήθως μια συνάρτηση απώλειας αντίθεσης, όπως η απώλεια InfoNCE, η οποία μεγιστοποιεί την ομοιότητα μεταξύ θετικών ζευγών και την ελαχιστοποιεί μεταξύ αρνητικών ζευγών.

Αυτοεπιβλεπόμενη αντιθετική μάθηση στην Ανάλυση Εικόνας/Οπτική Υπολογιστών

SimCLR (Simple Framework for Contrastive Learning of Visual Representations)

- **Αρχή**: Το SimCLR χρησιμοποιεί επαυξήσεις δεδομένων για να δημιουργήσει θετικά ζεύγη από την ίδια εικόνα και θεωρεί διαφορετικές εικόνες στη δέσμη ως αρνητικά ζεύγη. Χρησιμοποιεί έναν κωδικοποιητή ResNet για την εξαγωγή χαρακτηριστικών, τα οποία στη συνέχεια αντιστοιχίζονται σε ένα χώρο ενσωμάτωσης χρησιμοποιώντας μια κεφαλή προβολής. Η απώλεια αντίθεσης εφαρμόζεται σε αυτές τις ενσωματώσεις για να επιβάλει την επιθυμητή ομοιότητα και ανομοιότητα.
- **Δυνατά σημεία**: Το SimCLR είναι απλό και αποτελεσματικό, επιτυγχάνοντας επιδόσεις κοντά στα μοντέλα μάθησης με επίβλεψη χωρίς επισημειωμένα δεδομένα. Αξιοποιεί ένα μεγάλο μέγεθος παρτίδας για να εξασφαλίσει ένα ποικίλο σύνολο αρνητικών δειγμάτων.

MoCo (Momentum Contrast)

- **Αρχή**: Η MoCo διατηρεί ένα δυναμικό λεξικό με μια ουρά και έναν κωδικοποιητή κινητού μέσου όρου. Χρησιμοποιεί την τρέχουσα παρτίδα και το λεξικό για να δημιουργήσει ένα μεγάλο σύνολο αρνητικών δειγμάτων. Ο κωδικοποιητής του μοντέλου ενημερώνεται με την ορμή των βαρών του προηγούμενου κωδικοποιητή για να σταθεροποιηθεί η εκπαίδευση.
- **Δυνατά σημεία**: Το MoCo χειρίζεται αποτελεσματικά μικρότερα μεγέθη παρτίδων και παρέχει συνεπή αρνητικά δείγματα, οδηγώντας σε καλύτερες αναπαραστάσεις σε περιβάλλοντα με περιορισμένους πόρους.

Αυτοεπιβλεπόμενη αντιθετική μάθηση στη μοντελοποίηση γλωσσών/επεξεργασία φυσικής γλώσσας

SimCSE (Simple Contrastive Sentence Embeddings)

- **Αρχή**: Το SimCSE παράγει θετικά ζεύγη περνώντας την ίδια πρόταση δύο φορές από στρώματα διακοπής σε ένα μοντέλο BERT, αντιμετωπίζοντας τις διαφορετικές μάσκες διακοπής ως επαυξήσεις. Χρησιμοποιεί αρνητικά σε παρτίδα και μια αντιθετική απώλεια για να εκπαιδεύσει ενσωμάτωση προτάσεων.

- **Δυνατά σημεία**: Αυτή η μέθοδος αξιοποιεί αποτελεσματικά τα προ-εκπαιδευμένα γλωσσικά μοντέλα και δεν απαιτεί πρόσθετη αύξηση δεδομένων, καθιστώντας την υπολογιστικά αποδοτική.

CPC (Contrastive Predictive Coding)

- **Αρχή**: Η CPC χρησιμοποιεί ένα αυτοπαλίνδρομο μοντέλο για να προβλέψει μελλοντικές αναπαραστάσεις σε μια ακολουθία από παρελθούσες αναπαραστάσεις πλαισίου. Μεγιστοποιεί την αμοιβαία πληροφορία μεταξύ του πλαισίου και των μελλοντικών παρατηρήσεων χρησιμοποιώντας μια αντιθετική απώλεια.
- **Δυνατά σημεία**: Η CPC είναι αποτελεσματική στην εξαγωγή χρήσιμων αναπαραστάσεων από διαδοχικά δεδομένα και έχει εφαρμοστεί με επιτυχία τόσο σε δεδομένα ομιλίας όσο και σε δεδομένα κειμένου.