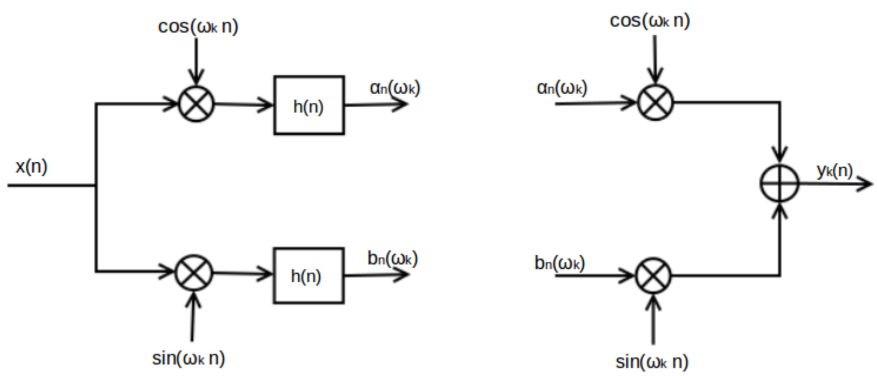
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Επεξεργασία Φωνής και Φυσικής Γλώσσας

Χειμερινό εξάμηνο 2020-2021

$2^η$ Σειρά Αναλυτικών Ασκήσεων

Άσκηση 1 (15 μονάδες)
Θεωρήστε την ανάλυση και σύνθεση του σήματος $x(n) = \cos(\omega_0 n)$. Το δίκτυο για το $k-$ οστό κανάλι φαίνεται στο σχήμα που ακολουθεί.



1. Ορίστε τα $\alpha_n(\omega_k)$, $b_n(\omega_k)$ για το σήμα εισόδου.

2. Υποθέτοντας οτι το h_n είναι ένα narrowband lowpass φίλτρο, απλοποιήστε τις εκφράσεις σας για τα $\alpha_n(\omega_k)$, $b_n(\omega_k)$, θεωρώντας οτι $(\omega_0 - \omega_k)$ πέφτει στην μπάντα διέλευσης του φίλτρου, και οτι $H(e^{j\omega}) \approx 1$ για τέτοιες συχνότητες.

3. Συνδιάζοντας τα $\alpha_n(\omega_k)$, $b_n(\omega_k)$, ορίστε τα $M_n(\omega_k)$ (magnitude) και $\phi_n^{'}(\omega_k)$ παράγωγο φάσης (phase derivative) για το συγκεκριμένο παράδειγμα.

4. Δείξτε οτι χρησιμοποιώντας το δίκτυο της εικόνας, το σήμα εξόδου είναι κατ' ουσίαν ίδιο με το σήμα εισόδου.

5. Η παράγωγος φάσης $\phi_n^{'}(\omega_k)$ υπολογίζεται από τον τύπο

$$\phi_n^{'}(\omega_k) = \frac{b_n(\omega_k)a_n^{'}(\omega_k) - a_n(\omega_k)b_n^{'}(\omega_k)}{[a_n(\omega_k)]^2 + [b_n(\omega_k)]^2} \tag{1}$$

   Λύστε ως προς $\phi_n^{'}(\omega_k)$ και συγκρίνεται το αποτέλεσμα με αυτό από το ερώτημα 3.

6. Τώρα υποθέστε οτι οι παράγωγοι του ερωτήματος 5 υπολογίζονται ως εξής

$$a^{'}(\omega_k) \approx \frac{1}{T}(a_n(\omega_k) - a_{n-1}(\omega_k)) \tag{2}$$

   όπου T είναι η περίοδος δειγματοληψίας στο πεδίο του χρόνου. Λύστε ως προς $\phi_n^{'}(\omega_k)$ και συγκρίνεται τα αποτελέσματα σας με αυτά από το ερώτημα 3. Κάτω από ποιές συνθήκες είναι περίπου ίδια?

Άσκηση 2 (15 μονάδες)

Θεωρήστε 2 πεπερασμένα σήματα φωνής $x(n)$ και $y(n)$, $0 \le n \le N-1$ (με μηδενικές τιμές εκτός του παραθύρου ανάλυσης). Για LPC ανάλυση με την autocorrelation function μέθοδο χρειάζονται οι αυτοσυσχετίσεις

$$R_x(k) = \sum_{n=0}^{N-1-k} x(n)x(n+k), \quad R_y(k) = \sum_{n=0}^{N-1-k} y(n)y(n+k) \tag{3}$$

οι οποίες με τη μέθοδο Levinson μας δίνουν τους αντίστοιχους βέλτιστους LPC συντελεστές

$$a_x = (a_{x0}, a_{x1}, \ldots, a_{xp}), \quad a_y = (a_{y0}, a_{y1}, \ldots, a_{yp}) \tag{4}$$

με $a_{x0} = a_{y0} = -1$.

1. Να αποδείξετε οτι η ενέργεια λάθους πρόβλεψης (για το $x(n)$) ισούται με

$$E_x = \sum_{n=0}^{N-1+p} \left(\sum_{k=0}^{p} a_{xk} x(n-k) \right)^2 = a_x R_x a_x^T \tag{5}$$

   όπου R_x είναι ένας $(p+1) \times (p+1)$ πίνακας.

2. Αν κάνετε γραμμική πρόβλεψη του σήματος $x(n)$ με τους βέλτιστους συντελεστές του σήματος $y(n)$, να αποδείξετε οτι η ενέργεια του νέου υβριδικού λάθους πρόβλεψης ισούται με

$$E_{xy} = \sum_{n=0}^{N-1+p} \left(\sum_{k=0}^{p} a_{yk} x(n-k) \right)^2 = a_y R_x a_y^T \tag{6}$$

3. Να βρείτε το πεδίο τιμών του λόγου E_{xy}/E_x

Άσκηση 3 (15 μονάδες)

Θεωρήστε σε μια ακολουθία φωνημάτων την μοντελοποίηση της εναλλαγής άφωνων (U=unvoiced) και έμφωνων (V=voiced) ήχων με ένα HMM μοντέλο (παραμέτρων λ) 4 καταστάσεων με τις εξής πιθανότητες

	State 1	State 2	State 3	State 4
P(V)	0.5	0.8	0.25	0.2
P(U)	0.5	0.2	0.75	0.8

Υποθέτουμε τις ακόλουθες πιθανότητες μετάβασης καταστάσεων

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} 0.25 & 0.2 & 0.3 & 0.25 \\ 0.2 & 0.25 & 0.3 & 0.25 \\ 0.4 & 0.2 & 0.2 & 0.2 \\ 0.25 & 0.3 & 0.2 & 0.25 \end{bmatrix} \tag{7}$$

και ίσες πιθανότητες αρχικής κατάστασης

$$\pi_i = 0.25, \quad i = 1,2,3,4. \tag{8}$$

Παρατηρούμε την ακολουθία $O_1 O_2 \cdots O_{10}$:

$$\mathbf{O} = (UVUVVVUUVU) \tag{9}$$

1. Να υπολογιστούν οι πιθανότητες

$$\delta_t(i) = \max_{q_1,q_2,\ldots,q_{t-1}} P[q_1 q_2 \ldots q_{t-1}, q_t = i, O_1 O_2 \ldots O_t | \lambda], \quad i = 1,2,3,4, \quad t = 1,\ldots,10 \tag{10}$$

2. Να βρεθεί η πιο πιθανή ακολουθία καταστάσεων $\mathbf{Q}^* = (q_1, q_2, \ldots, q_{10})$.

3. Να υπολογισθεί η πιθανότητα $P^* = (\mathbf{O}, \mathbf{Q}^*|\lambda)$

Για τα ερωτήματα (1) και (2) χρησιμοποιήστε τον αλγόριθμο Viterbi.

Άσκηση 4 (15 μονάδες)

1. Δίνεται το ακόλουθο φωνητικό λεξικό:

any	eh n iy
e.	iy
many	m eh n iy
men	m eh n
per	p er
persons	p er s uh n z
sons	s uh n z
suns	s uh n z
to	t uw
tomb	t uw m
too	t uw
two	t uw

2. Υπολογίστε την φωνολογική απόσταση μεταξύ των λέξεων του λεξικού. Θεωρήστε ότι η απόσταση ανάμεσα στα φωνήματα {uh, uw} και στα φωνήματα {er, eh} είναι η μισή της απόστασης μεταξύ δύο τυχαίων φωνημάτων. Το ίδιο για τα κρουστικά σύμφωνα {p,t}, τα ένρινα {m,n} και τα συριστικά {s,z}.
   Η απόσταση για διαγραφή (deletion) ή πρόσθεση (insertion) φωνήματος είναι 1.3, ενώ το κόστος της αντικατάστασης (substitution) δύο τυχαίων φωνημάτων είναι 1.0.
   Σημείωση: Σχεδιάστε την μηχανή πεπερασμένης κατάστασης που υπολογίζει την (μικρότερη) φωνολογική απόσταση ανάμεσα σε δύο λέξεις.

Να λυθεί χρησιμοποιώντας τη βιβλιοθήκη προγραμμάτων μηχανών πεπερασμένης κατάστασης - OpenFst .

Άσκηση 5 (10 μονάδες)

1. Το λεξικό μιάς φανταστικής παιδικής γλώσσας αποτελείται από τις ακόλουθες συλλαβές: {Μπα, Ντα, Γκα, Τσα}. Ένας γλωσσολόγος συνέλεξε τα παρακάτω δεδομένα από παιδιά ηλικίας ενός έτους:

   Μπα Μπα Τσα Ντα Ντα Τσα Γκα Τσα Μπα Μπα Γκα Γκα Τσα Ντα Ντα Γκα Ντα Ντα Ντα Ντα Μπα Μπα Τσα Γκα Τσα Ντα Ντα Ντα Μπα Μπα Τσα Μπα Μπα Ντα Ντα Τσα Τσα Μπα Μπα Γκα Γκα Τσα Ντα Ντα Γκα Ντα Ντα Ντα Ντα Μπα Μπα Τσα Γκα Τσα Ντα Ντα Ντα Μπα Μπα Τσα

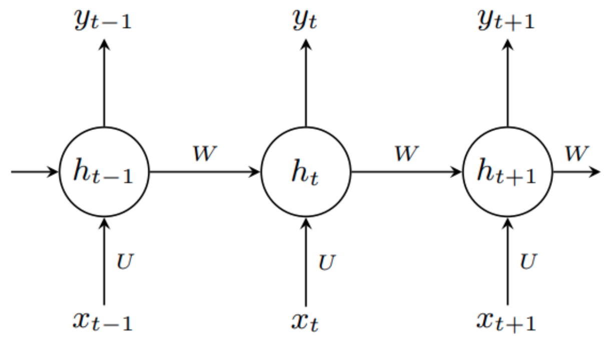
   Ο γλωσσολόγος ξέρει ότι όλες οι λέξεις αυτής της παιδικής γλώσσας περιέχουν μία ή δύο συλλαβές, π.χ, Μπα, Ντα, Γκα, Τσα, ΜπαΜπα, ΜπαΝτα, ΜπαΓκα, ΜπαΤσα, ΝταΜπα, ΝταΝτα. Οι λέξεις με μία συλλαβή εμφανίζονται εξίσου συχνά με τις λέξεις με δύο συλλαβές, δηλαδή, P(Μπα) + P(Ντα)+P(Γκα)+P(Τσα) = 0.5

   (a) Υπολογίστε την πιο πιθανή λέξη δύο συλλαβών.
   (b) Υπολογίστε την πιο πιθανή σειρά από λέξεις στην παραπάνω σειρά από συλλαβές. που συνέλεξε ο γλωσσολόγος

Να λυθεί χρησιμοποιώντας τη βιβλιοθήκη προγραμμάτων μηχανών πεπερασμένης κατάστασης - OpenFst .

Άσκηση 6 (30 μονάδες)

Back propagation through time: Σας δίνεται το ακόλουθο RNN



Κάθε κατάσταση h_t δίνεται από το ακόλουθο ζεύγος εξισώσεων

$$h_t = \sigma(Wh_{t-1} + Ux_t), \quad \sigma(z) = \frac{1}{1+e^{-z}}$$

Έστω L η συνάρτηση σφάλματος, η οποία ορίζεται ως το άθροισμα πάνω σε όλα τα επιμέρους χρονικά σφάλματα L_t σε κάθε χρονική στιγμή t μέχρι το χρονικό ορίζοντα T. Δηλαδή, $L = \sum_{t=0}^{T} L_t$, όπου το κάθε επιμέρους χρονικό σφάλμα εξαρτάται από την κατάσταση h_t.

Με βάση τα παραπάνω να εξάγετε την παράγωγο της συνάρτησης σφάλματος ως προς τον πίνακα βαρών W.

α) Δοθέντος ότι $y = \sigma(Wx)$ όπου $y \in \mathbb{R}^n, x \in \mathbb{R}^d$ και $W \in \mathbb{R}^{n\times d}$. Δείξτε ότι για την Ιακωβιανή ισχύει $\frac{\partial y}{\partial x} = \mathrm{diag}(\sigma')W \in \mathbb{R}^{n\times d}$

β) Δείξτε ότι ισχύει $\frac{\partial L}{\partial W} = \sum_{t=0}^{T}\sum_{k=1}^{t} \frac{\partial L_t}{\partial h_t}\frac{\partial h_t}{\partial h_k}\frac{\partial h_k}{\partial W}$

Vanishing / Exploding Gradients Σε αυτό το σκέλος θα εξετάσουμε τα προβλήματα που εμφανιζονται στις vanilla RNN αρχιτεκτονικές και συγκεκριμένα το πρόβλημα των vanishing και exploding gradients. Θα βασιστούμε στις ιδιοτιμές του πίνακα βαρών προκειμένου να μελετήσουμε τα προβλήματα αυτά.

γ) Για τιμή χρονικού ορίζοντα $T = 3$ γράψτε τη συνολική μορφή της εξίσωσης του ερωτήματος β. Δείξτε εποπτικά πως εαν θέλαμε να εκτελέσουμε backpropagation σε n το πλήθος χρονικές στιγμές θα έπρεπε να πολλαπλασιάσουμε το μητρώο $diag(\sigma' W)$ με τον εαυτό του n φορές.

δ) Κάθε διαγωνοποιήσιμος πίνακας M μπορεί να αναπρασταθεί μέσω των ιδιοτιμών και των ιδιοδιανυσμάτων του και συγκεκριμένα στη μορφή $M = Q\Lambda Q^{-1}$ όπου Q είναι ο πίνακας του οποίου η i-στη στήλη είναι το $i-$στο ιδιοδιάνυσμα του M και Λ είναι ένας διαγώνιος πίνακας με τις αντίστοιχες ιδιοτιμές πάνω στη διαγώνιο. Δείξτε ότι το γινόμενο $\prod_{i=1}^{n} M$ μπορεί να γραφτεί ως $M^n = Q\Lambda^n Q^{-1}$

ε) Θεωρείστε τον πίνακα βαρών $W = \begin{pmatrix} 0.58 & 0.24 \\ 0.24 & 0.72 \end{pmatrix}$. Η ιδιοδιάσπαση του πίνακα είναι:

$$W = Q\Lambda Q^{-1} = \begin{pmatrix} 0.8 & -0.6 \\ 0.6 & 0.8 \end{pmatrix}\begin{pmatrix} 0.4 & 0 \\ 0 & 0.9 \end{pmatrix}\begin{pmatrix} 0.8 & 0.6 \\ -0.6 & 0.8 \end{pmatrix}$$

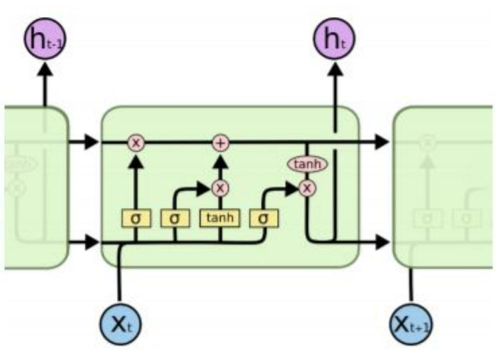
Υπολογίστε το W^{30}. Τι παρατηρείτε; Στη γενική περίπτωση, τι συμβαίνει όταν η απόλυτη τιμή των ιδιοτιμών του W είναι μικρότερη, μεγαλύτερη ή ίση με 1; Αναλύστε τις τρεις περιπτώσεις.

LSTMs: Μια αρχιτεκτονική αναδρομικών δικτύων που λύνει το πρόβλημα της εξαφάνισης ή έκρηξης παραγώγων (vanishing / exploding gradients) είναι τα δίκτυα Βραχέας-Μακράς Μνήμης (Long Short Term Memory networks - LSTM). Η αρχιτεκτονική και οι πράξεις που πραγματοποιεί το δίκτυο φαίνονται στην εικόνα (το σύμβολο $\odot$ υποδηλώνει τον πολλαπλασιασμό στοιχείο προς στοιχείο - hadamard product):

στ) Διαβάστε αυτό το άρθρο και εξηγείστε εν συντομία το ρόλο των πυλών f_t, i_t και o_t

ζ) Εξηγείστε ποιες από τις ποσότητες είναι πάντα θετικές (ή μηδέν)

Για να κατανοήσουμε το πώς προσεγγίζει το LSTM πρόβλημα εξαφάνισης παραγώγων χρειάζεται να υπολογίσουμε τις μερικές παραγώγους $\frac{\partial L}{\partial \theta}$, όπου θ οι παράμετροι του δικτύου (W_f, W_o, W_i, W_c). Στην περίπτωση του LSTM αντί για την κρυφή κατάσταση



$$f_t = \sigma(W_f h_{t-1} + U_f x_t)$$
$$i_t = \sigma(W_i h_{t-1} + U_i x_t)$$
$$o_t = \sigma(W_o h_{t-1} + U_o x_t)$$
$$\tilde{C}_t = \tanh(W_c h_{t-1} + U_c x_t)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$
$$h_t = o_t \odot \tanh(C_t)$$

h_t ενδιαφερόμαστε για την κατάσταση κελιού C_t. Όπως και το h_t στα απλά RNN έτσι και το C_t εξαρτάται από προηγούμενες τιμές $C_{t-1}, \ldots C_0$ και οδηγούμαστε σε μια απλοποιημένη εξίσωση της μορφής:

$$\frac{\partial L}{\partial W} = \sum_{t=0}^{T} \sum_{k=1}^{t} \frac{\partial L}{\partial C_t} \frac{\partial C_t}{\partial C_k} \frac{\partial C_k}{\partial W}$$

η) Η εξίσωση είναι απλοποιημένη, καθώς αγνοούμε τις εξαρτήσεις του C_t από τους όρους $f_t, i_t, \tilde{C}_t$. Μας ενδιαφέρει η εξάρτηση από αυτούς τους όρους για να μελετήσουμε το φαινόμενο εξαφάνισης παραγώγων; Γιατί;

ϑ) Δεδομένου ότι:

$$\frac{\partial C_t}{\partial C_k} = \prod_{i=k+1}^{t} \frac{\partial C_i}{\partial C_{i-1}}$$

και αν θεωρήσετε ότι $f_t = 1$ και $i_t = 0$ υπολογίστε την ποσότητα $\frac{\partial C_t}{\partial C_k}$.

ι) (Bonus) Δείξτε ότι στη γενική περίπτωση η αναδρομική σχέση είναι της μορφής

$$\frac{\partial C_t}{\partial C_{t-1}} = \sigma'() \cdot W_f \cdot \delta \odot C_{t-1} + f_t + \sigma'() \cdot W_i \cdot \delta \cdot \tilde{C}_t + i_t \odot tanh'()\delta,$$

όπου $\delta = o_{t-1} \odot tanh'(C_{t-1})$.

Γιατί εν τέλει είναι καλύτερο να χρησιμοποιούμε το cell state από το hidden state (σχετικά με τα vanishing gradients);

Hint: Θυμηθείτε τον κανόνα παραγώγισης γινομένων. Ισχύει και για το hadamard product: $(x \odot f(x))' = x' \odot f(x) + x \odot f'(x)$