Αναγνώριση Συναισθημάτων από την Ομιλία

Μετρήσεις και Έλεγχοι στη Βιοϊατρική Τεχνολογία
Ακαδημαϊκό Έτος 2021 - 2022

Γεώργιος Στεφανάκης
Αναστάσιος Παπαζαφειρόπουλος
Χαράλαμπος Μπότσας
Ραφαήλ Καρανίκας

Περιεχόμενα

1 Εισαγωγή

2 Μοντέλα Συναισθημάτων

3 Σύνολα Δεδομένων

4 Χαρακτηριστικά

5 Κλασικές Μέθοδοι Ταξινόμησης

6 Νευρωνικά Δίκτυα και Βαθιά Μάθηση

7 Συμπεράσματα

1

Εισαγωγή

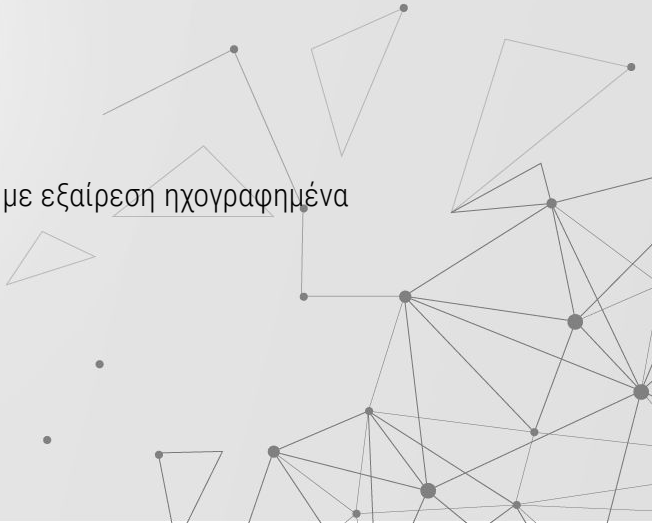
Νόημα του Λόγου και Περιεχόμενο

Η ομιλία είναι σύνθετο σήμα

Περιέχει πληροφορίες όπως:

- Το μήνυμα (σημασιολογικό περιεχόμενο)
- Για τον ομιλητή
- Το συναίσθημα του ομιλητή
- Τη γλώσσα και άλλα

Δυσκολία εντοπισμού και κατανόησης συναισθηματικού τόνου σε καθημερινή ομιλία με εξαίρεση ηχογραφημένα δείγματα από στούντιο.



Χαρακτηριστικά του Λόγου και Στόχοι του SER

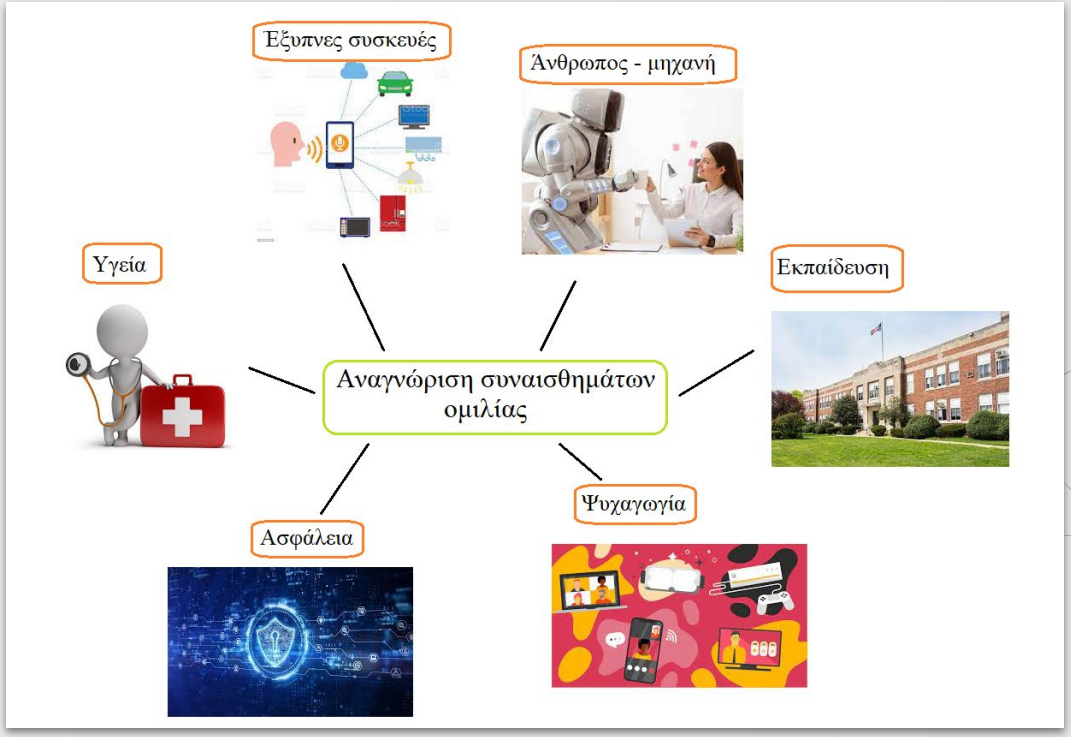
Ο τρόπος έκφρασης συναισθήματος ποικίλει ανάλογα με:

- Τη γλώσσα (τονική ή μη)
- Λέξεις πολύσημες
- Τον ομιλητή
- Την κουλτούρα κ.α.

Βασικοί στόχοι της επεξεργασίας σήματος:

- Κατανόηση των συναισθημάτων που εμφανίζονται στην ομιλία
- Σύνθεση των επιθυμητών συναισθημάτων στην ομιλία σύμφωνα με το επιδιωκόμενο μήνυμα

SER και Εφαρμογές σε Ποικίλους Τομείς



Το Ζήτημα του SER και η Ανάπτυξή του

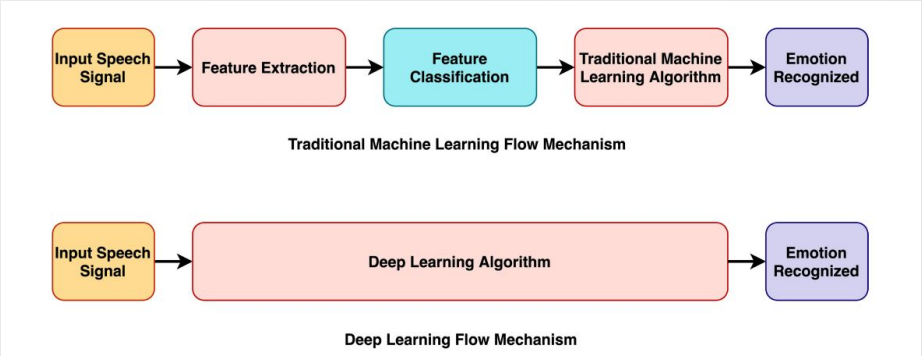
Περίπλοκο πρόβλημα με ιδιότροπη διαδικασία κατα περίπτωση.

Ζήτημα: ο ορισμός βασικών κατηγοριών συναισθημάτων και κριτήρια επιλογής αυτών με κατάλληλα χαρακτηριστικά.

Για την ανάπτυξη του SER κάποια τυπικά βήματα είναι:

1. Καλή προεπεξεργασία των ηχητικών δειγμάτων
2. Feature engineering
3. Χρήση διάφορων παραδοσιακών τεχνικών μηχανικής μάθησης (Hidden Markov Models (HMM), Gaussian Mixture Models (GMM) και Support Vector Machines (SVM)
4. Συνδυασμός με πληθώρα βάσεων δεδομένων για την εκπαίδευση και δοκιμή των διατάξεων που έχουν κατασκευαστεί

Εξέλιξη του SER



Ανάγκη περαιτέρω έρευνας σε αναδυόμενες τεχνολογίες βαθιάς μάθησης και νευρωνικών δικτύων λόγω προχωρημένων εφαρμογών SER.

Η παρούσα κατάσταση της τεχνολογίας (State-of-the-art) SER ως πρόβλημα ταξινόμησης με εκμετάλλευση:

- Τυπικών αλγορίθμων μηχανικής μάθησης
- Βαθιών νευρωνικών δικτύων (κατασκευή ικανοποιητικών μοντέλων SER)

Τυπικη Μεθοδολογία

1) Κατάτμηση των δειγμάτων σε μικρά χρονικά πλαίσια και εξαγωγή ιδιοτήτων χαμηλού επιπέδου, όπως η θεμελιώδης συχνότητα ***Fo***, Mel-Frequency Cepstral συντελεστές (MFCCs) κ.ά.
2) Επιλογή κάποιων ιδιοτήτων (features) και συναλήθευση αυτών με χρήση στατιστικών μεθόδων => αναγωγή σε πρόβλημα λιγότερων διαστάσεων απ' το αρχικό (dimensionality reduction)

Προεπεξεργασία στα ακατέργαστα δεδομένα:

- Οριοθέτηση της ακρίβειας των εξαγόμενων μοντέλων,
- Δύσκολη η γενίκευσή τους σε πραγματικές εφαρμογές που απαιτούν αποτελέσματα σε πραγματικό χρόνο.

Αλλαγές αντιμετώπισης SER

Προσπάθειες για αντικατάσταση της τυπικής μεθοδολογία με αλγορίθμους end-to-end μηχανικής μάθησης.

Αλγορίθμους end-to-end μηχανικής μάθησης: προσαρμογή σε ιδιότητες χαμηλού επιπέδου και έπειτα ικανότητα εξαγωγή νέων ιδιοτήτων υψηλού επιπέδου.

Αποφυγή υπερεξάρτησης των μοντέλων από την επιλογή των χαρακτηριστικών και από άλλα βήματα προεπεξεργασίας, πράγμα που αποτελεί και το στόχο της βαθιάς μάθησης.

2

Μοντέλα Συναισθημάτων

Μοντέλα συναισθημάτων

- **Πολυδιάστατα μοντέλα συναισθημάτων**
  - ➔ Χρήση μικρού αριθμού διαστάσεων για τον χαρακτηρισμό των συναισθημάτων:
    - δυναμικότητα (valence)
    - δραστηριοποίηση (activation)
    - κυριαρχία (dominance)
  - ➔ Μειονεκτήματα
    - Δύσκολη τιτλοφόρηση λόγω ανεπαρκούς διαισθητικότητας
    - Απαγορευτική πολυπλοκότητα για υλοποίηση με μηχανική μάθηση
- **Διακριτά μοντέλα συναισθημάτων**
  - ➔ Κατάταξη συναισθημάτων σε πεπερασμένο πλήθος κατηγοριών
  - ➔ Μοντέλο Paul Ekman: λύπη, χαρά, φόβος, θυμός, αποστροφή και έκπληξη

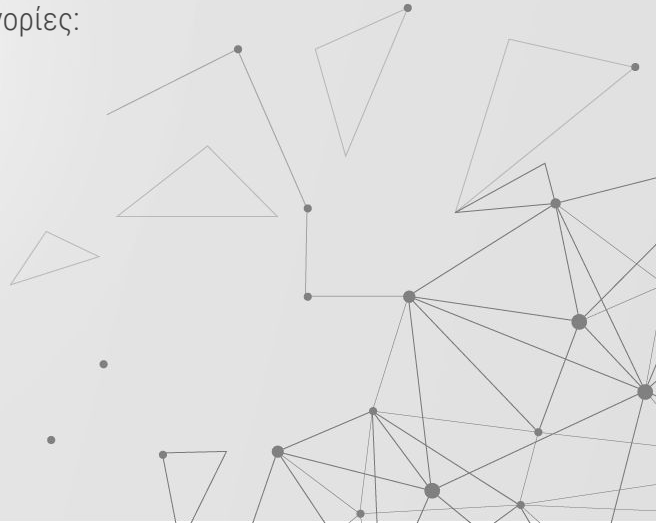
3

Σύνολα Δεδομένων

Σύνολα Δεδομένων - Datasets

Τα **σύνολα δεδομένων**:

- Περιέχουν (ηχητικά) δείγματα για την **εκπαίδευση**, την **επαλήθευση** και τη **δοκιμή** των μοντέλων
- Είναι πολύ σημαντικά καθώς από εκείνα εξαρτάται σημαντικά η απόδοση των ταξινομητών
- Περιέχουν δείγματα τα οποία έχουν ήδη ταξινομηθεί στην κλάση που ανήκουν, είτε από τους κατασκευαστές τους, είτε με τη μέθοδο του πληθοπορισμού (crowdsourcing)
- Στην αναγνώριση συναισθημάτων μέσω ομιλίας, διακρίνονται σε τρεις κατηγορίες:
    - **Προσομοιωμένα** (simulated)
    - **Ημι-φυσικά** (semi-natural ή induced)
    - **Φυσικά** (natural)



Σύνολα Δεδομένων - Datasets

- **Προσομοιωμένα Σύνολα:**

  Εκπαιδευμένοι ηθοποιοί φωνής καταγράφονται ενώ διαβάζουν μεμονωμένες προτάσεις με συγκεκριμένο συναίσθημα (έκπληξη, θυμός κ.λπ.)

  EMO-DB, DES, RAVDESS, eNTERFACE, SAVEE, CREMA-D

- **Ημι-φυσικά Σύνολα:**

  Εκπαιδευμένα και μη άτομα διαβάζουν ένα σενάριο ομιλίας που περιέχει διάφορα συναισθήματα

  IEMOCAP

- **Φυσικά Σύνολα:**

  Τα δείγματα εξάγονται από πραγματικές συνομιλίες (τηλεοπτικών εκπομπών, τηλεφωνικών κέντρων κ. λπ.) και επιλέγονται τα πιο ξεκάθαρα από αυτά

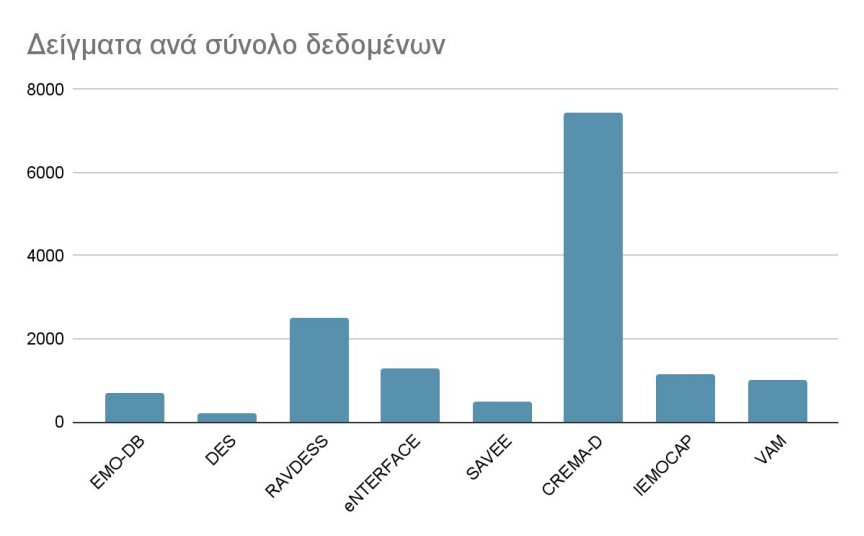
  VAM

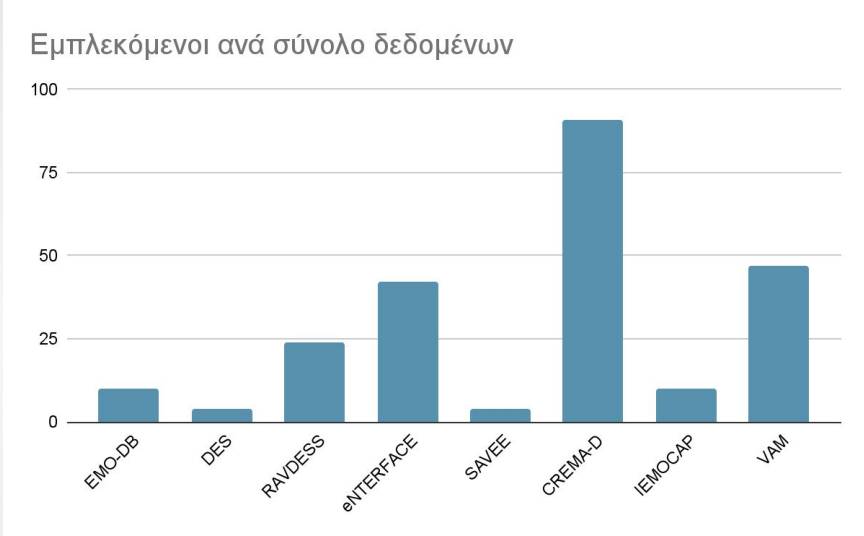
Σύνολα Δεδομένων - Datasets

Οι βασικές κλάσεις συναισθημάτων που εμφανίζονται στα προηγούμενα σύνολα είναι οι εξής:

- **Θυμός, χαρά, λύπη (7)**
- **Ουδετερότητα, φόβος, αποστροφή (6)**
- **Έκπληξη (5)**
- **Ανία, ηρεμία, δυναμικότητα, δραστηριοποίηση, κυριαρχία (2)**
- **Απογοήτευση, ενθουσιασμός (1)**

Σύνολα Δεδομένων - Datasets





4

Χαρακτηριστικά

Επιλογή χαρακτηριστικών

- **Προσωδικά χαρακτηριστικά (Prosodic features)**
  - θεμελιώδης συχνότητα (F0)
  - ενέργεια
  - διάρκεια
- **Φασματικά χαρακτηριστικά (Spectral features)**
  - Mel-Frequency Cepstral Coefficients (MFCC)
  - Linear Prediction Cepstral Coefficients (LPCC)
  - Log-Frequency Power Coefficients (LFPC)
  - Gammatone Frequency Cepstral Coefficients (GFCC)
  - Formants
- **Χαρακτηριστικά ποιότητας φωνής (Voice Quality features)**
  - Jitter
  - Shimmer
  - λόγος αρμονικών προς θόρυβο (harmonics to noise ratio – HNR)

5

Κλασικές Μέθοδοι Ταξινόμησης

Κλασικές Μέθοδοι Ταξινόμησης

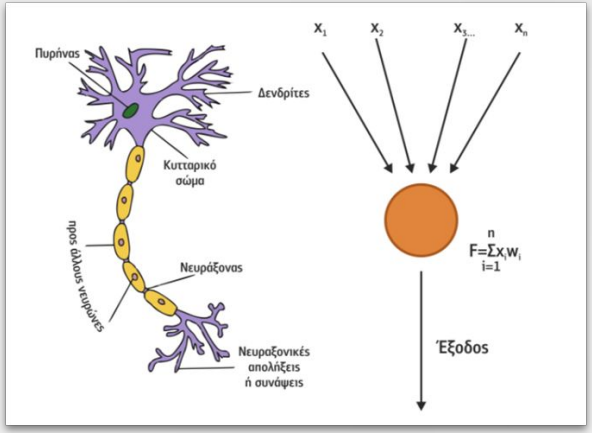
- **Support Vector Machines (SVMs)**
  - Στόχος: εύρεση βέλτιστου υπερεπιπέδου που διαχωρίζει γραμμικά τα δεδομένα
  - Από τους πιο διαδεδομένους και ακριβείς ταξινομητές
- **Hidden Markov Models (HMMs)**
  - Ευρεία χρήση στην αναγνώριση φωνής
  - Από τις πρώτες μεθόδους που χρησιμοποιήθηκαν στο SER
- **Gaussian Mixture Models (GMMs)**
  - το σύνολο των δεδομένων έχει προκύψει από τον συνδυασμό ενός πεπερασμένου αριθμού Γκαουσιανών κατανομών με άγνωστες παραμέτρους
  - αναζήτηση των αγνώστων παραμέτρων

6

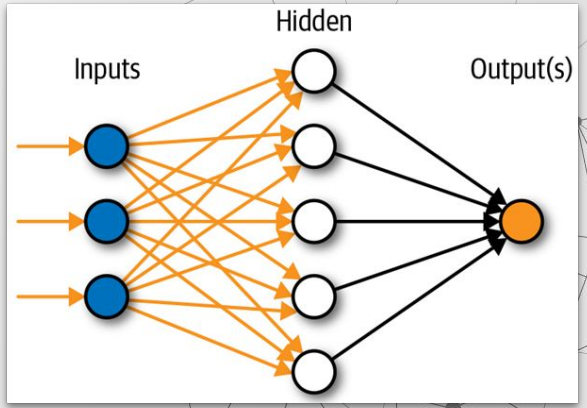
Νευρωνικά Δίκτυα και Βαθιά Μάθηση

Artificial Neural Networks (ANNs)

- Σκοπός εμφάνισης: Μάθηση
- Εμπνευσμένα από Κεντρικό Νευρικό Σύστημα του ανθρώπου



- Οργάνωση σε επίπεδα
- Κάθε επίπεδο έχει πολλούς κόμβους που συνδέονται μεταξύ τους και με κόμβους διαφορετικών επιπέδων
- Οι κόμβοι αλληλεπιδρούν μεταξύ τους
- Επίπεδο εισόδου, Κρυμμένα επίπεδα, επίπεδο εξόδου
- Γρήγορη εκπαίδευση
- Δε μπορεί να λύσει πολύπλοκα και μη γραμμικά προβλήματα

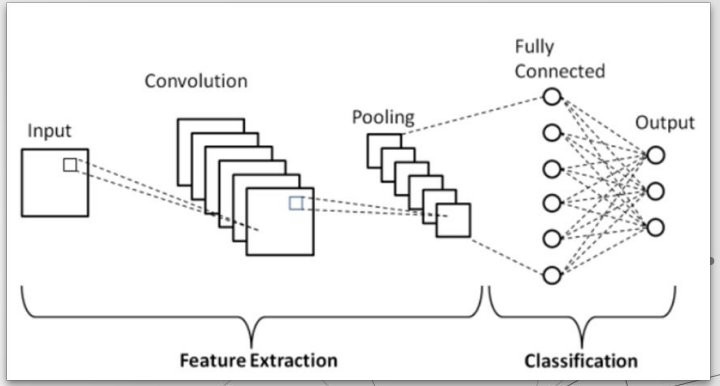


Σημαντικότερη Έρευνα:
Emotion recognition from Marathi speech database using adaptive artificial neural network, Darekar, and Dhande, 2018

Χαρακτηριστικά: MFCC, NMF, Τόνος , Υλοποίηση μαζί με αλγόριθμο PSO Feedforward
Αποτελέσματα: RAVDESS: 88.7%

Convolutional Neural Networks (CNNs)

Διαφοροποίηση από ANN:
Στο κρυμμένο τους επίπεδο έχουν διαφορετικά φίλτρα ή περιοχές που αποκρίνονται σε ένα συγκεκριμένο χαρακτηριστικό του σήματος εισόδου.



Πλεονεκτήματα:

- Μαθαίνουν χαρακτηριστικά από πολυδιάστατα δεδομένα εισόδου
- Μαθαίνουν την εμφάνιση μικρών παραλλαγών και διαστρεβλώσεων

Μειονεκτήματα:

- Προϋπόθεση ύπαρξης κατάλληλου αποθηκευτικού χώρου την ώρα εκτέλεσης

Σημαντικότερες Έρευνες:

A First Look Into A Convolutional Neural Network For Speech Emotion Detection, Bertero, and Fung, 2017
Συναισθήματα: θυμός, χαρά, λύπη Υλοποίηση με Theano toolkit Αποτελέσματα: TEDLIUM2: 66.1%

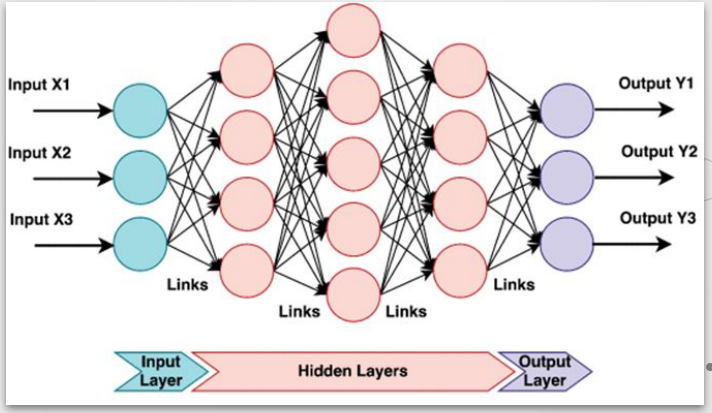
S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom, “Negative Emotion Recognition using Deep Learning for Thai Language,” 2020
Ταξινόμηση αρνητικών συναισθημάτων από σύνολο δεδομένων στην Ταϋλανδέζικη γλώσσα,
Αποτελέσματα: 96.60%

Βαθιά Νευρωνικά Δίκτυα

- Νευρωνικά δίκτυα που περιέχουν **πολλαπλά κρυφά επίπεδα**
- Η πληροφορία τροφοδοτείται από την είσοδο του πρώτου επιπέδου **προς μία κατεύθυνση**
- Κάθε έξοδος του προηγούμενου επιπέδου τροφοδοτείται στην είσοδο κάθε νευρώνα του τρέχοντος επιπέδου
- Η έξοδος κάθε νευρώνα ανακατευθύνεται σε μία **μη γραμμική συνάρτηση** (συνήθως **σιγμοειδή** ή συνάρτηση πυροδότησης **Rectified Linear Unit - ReLU**)
- Για πρόβλημα ταξινόμησης σε πολλές κλάσεις υπολογίζεται στο τελευταίο επίπεδο (softmax) του δικτύου η **πιθανότητα το δείγμα να ανήκει σε κάθε κλάση**
- Ακολουθείται **αλγόριθμος οπισθοδρόμησης** (backpropagation) μέσω του οποίου βελτιστοποιούνται οι παράμετροι του δικτύου (βάρη ακμών ανάμεσα στα επίπεδα, αυθαίρετες σταθερές)
- Για να αποφευχθεί η υπερεκπαίδευση επιστρατεύονται **αλγόριθμοι εξομάλυνσης του δικτύου**

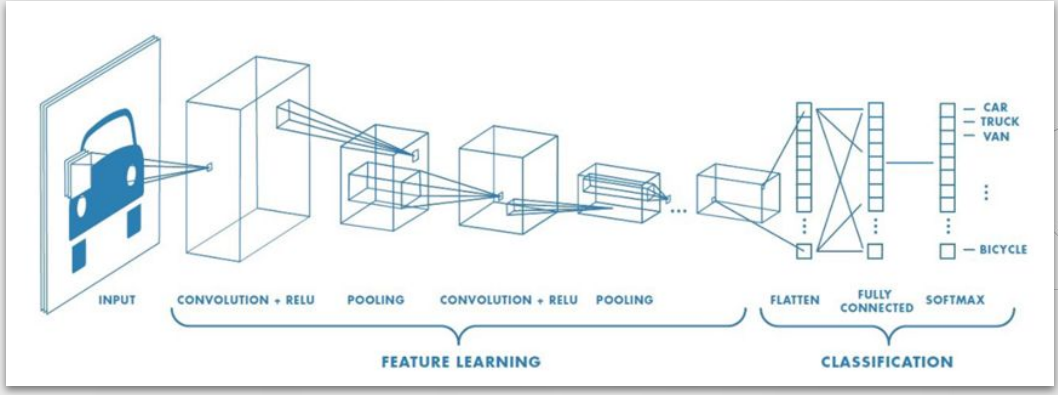
Βαθιά Νευρωνικά Δίκτυα

- H. M. Fayek, M. Lech, and L. Cavedon, “Towards real-time Speech Emotion Recognition using deep neural networks”, 2015
- Συνολική ακρίβεια μελέτης:
    - **eNTERFACE: 60.53%**
    - **SAVEE: 59.7%**
- Σχετικά χαμηλότερα ποσοστά επιτυχίας σε σχέση με άλλες μεθόδους
- Απλή διάταξη που θα μπορούσε να χρησιμοποιηθεί σε συστήματα πραγματικού χρόνου



Βαθιά Συνελικτικά Νευρωνικά Δίκτυα

- Υποκατηγορία των γενικών βαθιών νευρωνικών δικτύων
- **Επίπεδα συνελικτικών κόμβων**, στους οποίους η είσοδος συνελίσεται με διάφορα φίλτρα
- Τα φίλτρα αυτά αποσκοπούν στην **αναγνώριση συγκεκριμένων χαρακτηριστικών** του δείγματος
- Ακολουθούν ένα ή περισσότερα **πλήρως συνδεδεμένα επίπεδα** για να ολοκληρωθεί η ταξινόμηση



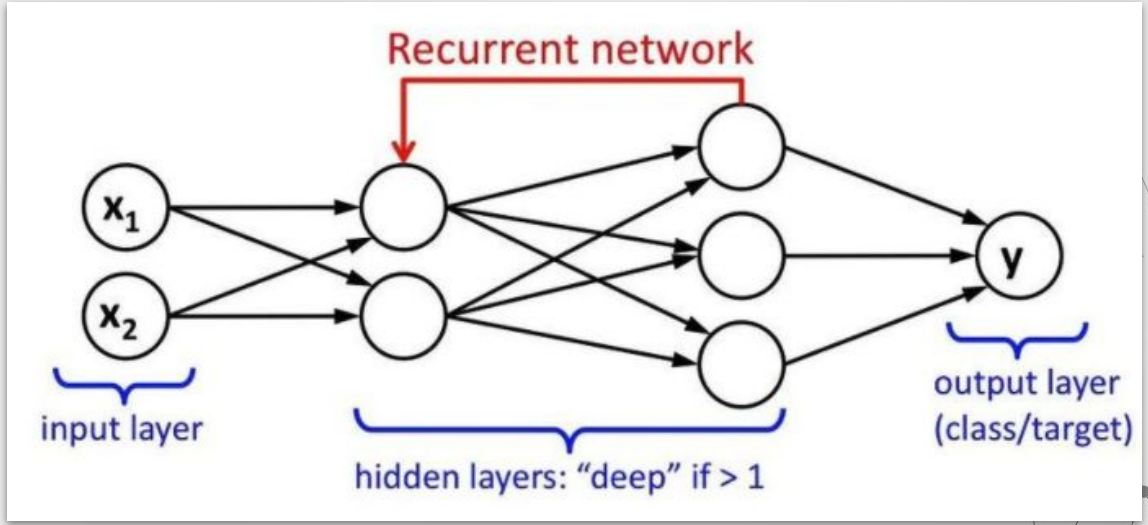
Βαθιά Συνελικτικά Νευρωνικά Δίκτυα

- P. Harar, R. Burget, and M. K. Dutta, “Speech emotion recognition with deep learning”, 2017
    - **EMO-DB: 96.97%**
- S. Zhang, S. Zhang, T. Huang, and W. Gao, “Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching”, 2018
    - Δίκτυο **AlexNet**, εκπαιδευμένο με σύνολο δεδομένων 1.2 εκατομμυρίων εικόνων
    - Fine-tuned με δείγματα από το σύνολο **EMO-DB: 83.53%**
    - **eNTERFACE: 70.25%**
    - **RML: 64.88%**
- Η αυτόματη επιλογή χαρακτηριστικών των DCNNs **μπορεί να αποδώσει καλύτερα από τη χειροκίνητη επιλογή τους στα απλά CNNs**
- Υψηλότερα ποσοστά επιτυχίας καθώς είναι ισχυρά στη μοντελοποίηση των μικρότερων μεταβολών του δείγματος
- Όμως, έχουν μεγάλο κόστος μνήμης και απαιτούν εκατομμύρια δείγματα για την εκπαίδευσή τους
- **Επιρρεπή σε υπερπροσαρμογή**

Recurrent Neural Networks RNNs

Ιδιότητες - Χαρακτηριστικά:

- Οι συνδέσεις μεταξύ κόμβων σχηματίζουν ένα κατευθυνόμενο ή μη κατευθυνόμενο γράφημα κατά μήκος μιας χρονικής ακολουθίας
- Χάρη στην ανάδραση, μπορούν να μαθαίνουν και να αντιδρούν χωρίς να αλλάζουν τα αργά διαμορφωμένα βάρη τους
- Κατά την εκπαίδευση του RNN μέσω οπισθοδιάδοσης στο χρόνο είναι δυνατό τα εσφαλμένα σήματα που τρέχουν ανάποδα στο χρόνο να είναι όλο και μεγαλύτερα → πιθανή ταλάντωση των βαρών → αργό δίκτυο ως προς την εκπαίδευση ή τη σύγκλιση



Long Short - Term Memory Neural Networks (LSTMs)

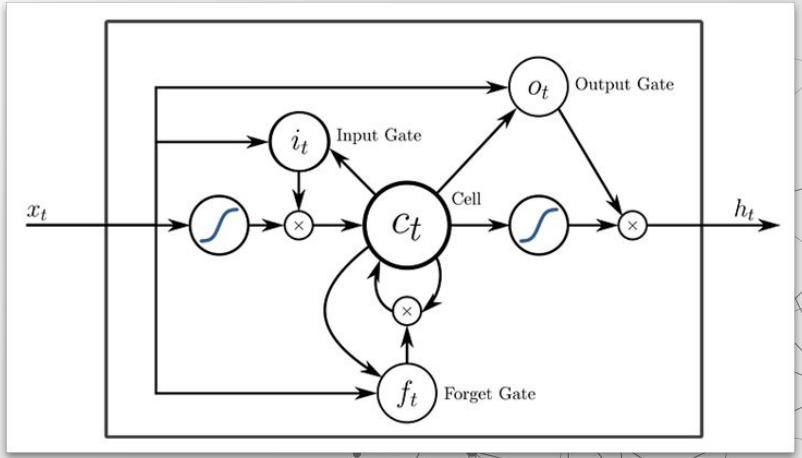
Λόγος Δημιουργίας:
Προκειμένου να αντιμετωπιστούν τα προβλήματα καθυστέρησης των RNNs, οι Hochreiter και Schmidhuber παρουσίασαν την αρχιτεκτονική του LSTM το 1997.

Βασική Ιδιότητα:
Τα δίκτυα LSTMs μπορούν να «γεφυρώσουν» χρονικά κενά, ακόμα και όταν η ακολουθία εισόδου είναι ασυμπίεστη και θορυβώδης. Ουσιαστικά, με έναν αλγόριθμο κλίσης κόβονται οι υπολογισμοί κλίσης σε ένα καθορισμένο σημείο, χωρίς να επηρεαστούν οι Long-Term ενέργειες.

Πιο πρόσφατες έρευνες:

Speech Emotion Recognition using deep 1D & 2D CNN LSTM networks, Zhao et al., 2019
Υλοποίηση: DCNN, LSTM , Χαρακτηριστικά: PCM, Log-Mel Spectrogram
Αποτελέσματα: EMO-DB: 95.33%, IEMOCAP: 86.16%

When Old Meets New: Emotion Recognition from Speech Signals, Araño et al., 2021
Υλοποίηση: SVM, LSTM , Χαρακτηριστικά: MFCC, χαρακτηριστικά εξαγόμενα από φασματογραφήματα μέσω DCNN
Αποτελέσματα: RAVDESS: 73.5%

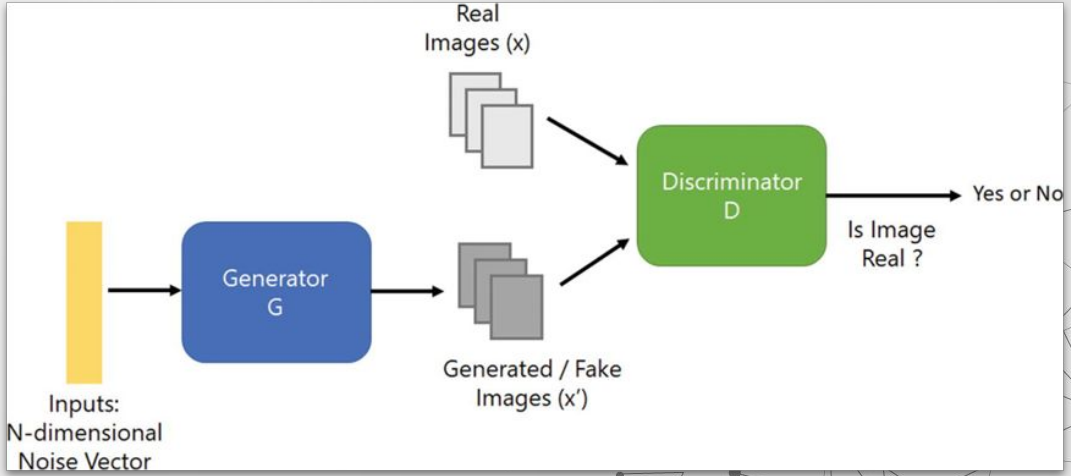


Generative Adversarial Networks (GANs)

Το GAN εφευρέθηκε το 2014 και βασίζεται στη λογική της αντιπαραθετικής μάθησης.

Βασική Ιδιότητα:
Αποτελείται από δύο νευρωνικά δίκτυα, τον «παραγωγό» (G) και τον «διαχωριστή» (D) που διαγωνίζονται σε ένα παίγνιο. Ουσιαστικά, ο G δημιουργεί υποψήφια δεδομένα και ο D τα αξιολογεί, καθορίζοντας ποια δεδομένα εισόδου είναι πραγματικά και ποια όχι. Τελικά, τα δίκτυα φτάνουν σε ισορροπία.

Σημαντικότερη Έρευνα:
Data Augmentation Using GANs for Speech Emotion Recognition, Chatziagapi et al., 2019
Υλοποίηση:DCNN, GAN Χαρακτηριστικά: 128 MFCCs
Αποτελέσματα: IEMOCAP: 53.6%, Feel-25k: 54.6%



7

Συμπεράσματα

Προκλήσεις

- Διαθεσιμότητα συνόλων δεδομένων σχεδιασμένα για βαθιά μάθηση
- Αληθοφανή σύνολα δεδομένων (πραγματικά ή όχι)
- Ανεξαρτητοποίηση σε ένα βαθμό των μοντέλων από τον πολιτισμό και τη γλώσσα, αντικειμενικότητα της ετικετοποίησης
- Συνένωση συνόλων δεδομένων

Μελλοντικές Κατευθύνσεις

- Κατασκευή ενός εξ' ολοκλήρου συνθετικού συνόλου δεδομένων αξιοποιώντας παραγωγικές μεθόδους, όπως τα GANs
- Ανάπτυξη μοντέλου που θα επιτυγχάνει ικανοποιητικές επιδόσεις ανεξάρτητα από το γλωσσικό περιβάλλον, πιθανώς με τη χρήση transformers
- Προσθήκη θορύβου στα δείγματα για ενίσχυση αληθοφάνειας
- Επιτυχής ταξινόμηση συναισθημάτων σε κατάσταση συνεχούς ομιλίας, ίσως με χρήση κάποιου κυλιόμενου παραθύρου για διαχωρισμό του σήματος και εξέταση του συναισθηματικού περιεχομένου κάθε τμήματος που προκύπτει

Τελικά Συμπεράσματα

- Μεταξύ όλων των βάσεων δεδομένων που συγκρίθηκαν, οι πιο σύγχρονες είχαν μεγαλύτερο πλήθος δεδομένων

- Στην εργασία χρησιμοποιήθηκε ως κύριο μέτρο απόδοσης και αξιολόγησης των ερευνών η τιμή ορθότητας που κατάφεραν να πετύχουν, αλλά στην πραγματικότητα δε μπορεί μόνο αυτή η τιμή να καθορίσει την επιτυχία ή αποτυχία μιας έρευνας

- Η μεγάλη εκτόξευση της έρευνας στον τομέα των CNNs δείχνει ότι τα αναπτυσσόμενα συστήματα βελτιώνουν την ικανότητά τους στο να εντοπίζουν μικροδιαφορές. Η ενσωμάτωση και των αρχιτεκτονικών βαθιά συνελικτικών LSTM έδωσε στο δίκτυο μακροπρόθεσμη μνήμη, και άρα τη δυνατότητα να μπορεί να ταυτοποιήσει τα μακροπρόθεσμα παραγλωσσικά σχήματα.

Ευχαριστούμε για την προσοχή σας!