Razpoznavanje čustvenih stanj govorcev in razpoznavanje varnostno sumljivih zvokov

Praktični projekt pri predmetu Govorne tehnologije

Jernej Sabadin

Univerza v Ljubljani
Fakulteta za elektrotehniko

Mentorja: izr. prof. dr. Simon Dobrišek, univ. dipl. inž. el. in univ. dipl. inž. el.,
As. Marija Ivanovska, mag. inž. el.

Predmet: Govorne tehnologije
Datum: 1. December 2023

Kazalo vsebine

Ključne besede:

---

1 Uvod

V okviru praktičnega projekta iz Govornih tehnologij se osredotočamo na dve ključni področji: razpoznavanje čustvenih stanj govorcev in detekcijo varnostno sumljivih zvokov. Te naloge predstavljajo sofisticiran izziv, ki združuje področja strojnega učenja in obdelave zvočnih signalov. Za učinkovito klasifikacijo tako čustvenih stanj kot sumljivih zvokov uporabljamo hibridni model, ki kombinira dolgoročni kratkotrajni spomin (Long Short-Term Memory - LSTM) z konvolucijskimi nevronskimi mrežami (CNN). Ta pristop nam omogoča visoko natančnost pri razpoznavanju kompleksnih vzorcev. Dodatno, za povečanje robustnosti in generalizacije modela, razširimo učno zbirko z vključitvijo perturbiranih vzorcev zvoka, kar prispeva k izboljšanju natančnosti in zanesljivosti sistema pri razpoznavanju.

Razpoznavanje čustvenih stanj govorcev

Čustvena inteligenca umetnih sistemov je ključnega pomena za izboljšanje interakcije med človekom in strojem. Razpoznavanje čustev iz govora omogoča napravam, da se odzivajo na človeška čustvena stanja in s tem postanejo bolj empatični in učinkoviti komunikacijski partnerji.

Razpoznavanje varnostno sumljivih zvokov

Varnost je pomembna skrb v današnji družbi. Sposobnost avtomatskega razpoznavanja sumljivih zvokov v realnem času lahko pripomore k hitrejšemu in učinkovitejšemu odzivanju na morebitne nevarnosti.

Uporabljene tehnologije

V projektni nalogi uporabimo Mel-spektrograme, ki so učinkovite značilke pridobljene iz zvokov. V kombinaciji z naprednimi tehnologijami strojnega učenja, kot so konvolucijska nevronska omrežja (CNN) in dolgoročni kratkotrajni spomin (LSTM), razvijemo modele, ki lahko prepoznavajo subtilne vzorce v podatkih.

V naslednjih poglavjih bomo podrobneje opisali uporabljene metode in arhitekture ter prikazali rezultate našega raziskovanja.

2 Mel-spektrogrami

Mel-spektrogram je orodje za vizualizacijo spektra frekvenc zvočnega signala, pri čemer se uporablja Mel frekvenčna lestvica. Ta lestvica je zasnovana tako, da bolje ustreza človeškemu zaznavanju frekvenc, saj človeško uho ni enakomerno občutljivo na vse frekvence. Mel-spektrogram zato zagotavlja bolj intuitivno predstavitev zvočnih signalov glede na to,

kako jih zaznava človeško uho, in je zelo uporaben v različnih aplikacijah za obdelavo zvoka.

Nelinearno zaznavanje frekvenc: Človeško uho je bolj občutljivo na spremembe frekvence na nižjih frekvencah kot na višjih. Mel lestvica upošteva to nelinearno zaznavanje z uporabo logaritemske skale.

$$f_{\text{Mel}} = 2595 \log_{10}(1 + \frac{f}{700}), \tag{1}$$

kjer je f frekvenca v Hz in f_{Mel} je zaznana frekvenca na Mel lestvici.

Postopek izračuna Mel-spektrograma:

1. Zvočni signal razdelimo na kratke prekrivajoče se časovne segmente z oknenjem.

2. Za vsak segment signala, ki smo ga pridobili z oknenjem, izračunamo njegovo transformiranko. To storimo z DFT.

$$F(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-i \frac{2\pi n k}{N}}, \qquad 0 \leq k \leq N-1, \tag{2}$$

kjer je N število točk, ki jih uporabimo za izračun DFT.

3. Frekvence preslikamo na Mel lestvico z uporabo melodičnih filtrov.

4. Spekter vsakega segmenta nato predstavimo v barvni lestvici, kjer intenziteta barve predstavlja amplitudo določene frekvence.

Melodični filtri: Melodični filtri se uporabljajo za preslikavo frekvenc v spektru zvoka na Mel lestvico. Ti filtri so ključni za pridobivanje značilnosti zvoka, ki so relevantne za človeško zaznavanje.

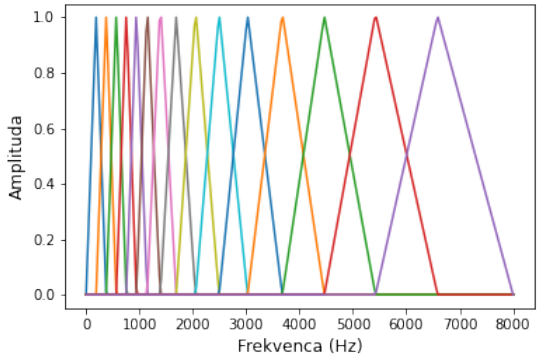

Fig. 1: Melodični filtri

Uporaba Mel-spektrogramov: Mel-spektrogrami se pogosto uporabljajo v sistemih za avtomatsko razpoznavanje govora in drugih aplikacijah za analizo zvoka, saj se dobro približajo slušni zaznavnosti človeških ušes.

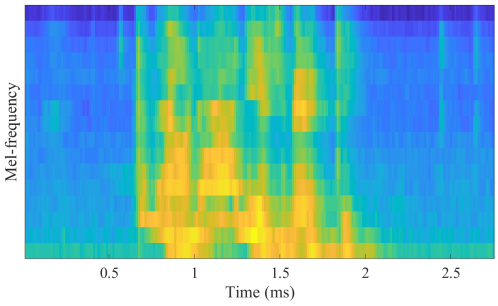

Fig. 2: Primer Mel-spektrograma iz vira [2]

3 Modeli strojnega učenja

V tem poglavju predstavimo modela strojnega učenja, ki ju uporabimo v našem raziskovalnem projektu: konvolucijska nevronska mreža (CNN) in dolgoročni kratkotrajni spomin (LSTM). Oba modela igrata pomembno vlogo pri klasifikaciji zvočnih signalov, pri čemer vsak pristopa k problemu na svoj način.

3.1 Konvolucijske Nevronske Mreže (CNN)

Konvolucijske nevronske mreže (CNN) so specializirane za obdelavo podatkov z izrazito prostorsko strukturo, kot so slike ali zvočni signali. Značilnost CNN je njihova sposobnost učenja prostorskih vzorcev neodvisno od njihove točne lokacije v vhodnem prostoru, kar dosežejo z uporabo konvolucijskih operacij.

Arhitektura CNN: CNN običajno vključujejo več ključnih plasti:

- *Konvolucijske plasti* zaznavajo lokalne vzorce z uporabo filtrov.
- *Aktivacijske funkcije*, kot je ReLU, uvedejo nelinearnosti v model.
- *Plasti združevanja* (pooling) zmanjšujejo prostorske dimenzije in povečujejo translacijsko invarianco.
- *Polno povezane plasti* izvajajo klasifikacijo ali regresijo na osnovi zbranih značilnosti.

Povezovanje Nevronov: V CNN so nevroni organizirani v plasti, kjer vsaka plast izvaja specifično funkcijo. V konvolucijskih plasteh so nevroni lokalno povezani s svojimi vhodnimi podatki, kar pomeni, da vsak nevron prejme vhod samo iz omejenega območja, imenovanega receptivno polje, prejšnje plasti ali vhodnih podatkov. V polno povezanih plasteh pa je vsak nevron v novi plasti povezan z vsemi nevroni v prejšnji plasti.

Delitev Uteži: Ključna značilnost konvolucijskih plasti je delitev uteži, kjer iste uteži (imenovane tudi filtri ali jedra) delujejo na različnih delih vhodnih podatkov. To omogoča zaznavanje istih vzorcev na različnih mestih vhodnega prostora in vodi do bistvenega zmanjšanja števila učljivih parametrov v primerjavi s polno povezanimi plastmi.

Na spodnji sliki prikažemo konceptualni pogled na CNN.

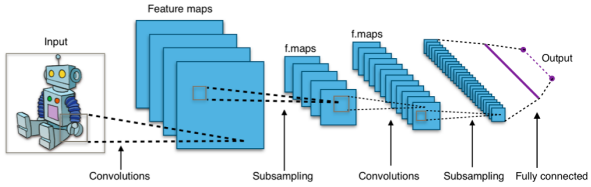

Fig. 3: Konceptualni pogled na CNN iz vira [4]

Konvolucijski Sloj: Konvolucijske operacije so opisane z naslednjo enačbo:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \quad (3)$$

kjer I predstavlja vhodno matriko, K je konvolucijski filter in $S(i,j)$ so aktivacijske vrednosti na izhodu.

Učenje v CNN: Proces učenja v CNN je osredotočen na optimizacijo uteži filtrov in uteži v polno povezanih plasteh. To dosežemo z minimizacijo kriterijske funkcije, ki ocenjuje razliko med napovedmi mreže in dejanskimi vrednostmi. Uporabljajo se različne kriterijske funkcije, kot so križna entropija ali srednja kvadratna napaka, odvisno od problema.

Optimizacija se običajno izvaja z algoritmom gradientnega spusta, ki je osnova za številne napredne metode optimizacije, kot so stohastični gradientni spust (SGD), Adam, Adagrad in RMSprop. Postopek posodabljanja uteži z gradientnim spustom je opisan z naslednjo enačbo:

$$W_{new} = W_{old} - \eta \frac{\partial \mathcal{L}}{\partial W} \quad (4)$$

kjer η predstavlja stopnjo učenja, $\mathcal{L}$ je kriterijska funkcija, in $\frac{\partial \mathcal{L}}{\partial W}$ je gradient kriterijske funkcije glede na uteži W.

Pomembno je omeniti, da je povratno širjenje napake (backpropagation) opisano zgoraj standardna metoda za učenje večslojnih nevronskih mrež. Ta metoda izračuna gradient funkcije napake glede na uteži mreže. Gradienti se nato uporabijo za posodobitev uteži mreže v smeri, ki zmanjšuje napako. Ta postopek se ponavlja, dokler se napaka ne zmanjša na sprejemljivo raven ali dokler se ne izpolnijo drugi pogoji za ustavitev.

CNN so zaradi lokalne povezanosti in delitve uteži še posebej učinkovite pri zaznavanju vzorcev in redukciji dimenzij, kar omogoča učinkovito obdelavo kompleksnih prostorskih podatkov.

3.2 Dolgoročni Kratkoročni Pomnilnik (LSTM)

Dolgoročni kratkoročni pomnilnik (LSTM) je specializirana vrsta rekurentnih nevronskih mrež (RNN), namenjena obdelavi dolgih sekvenčnih podatkov. Te mreže so oblikovane za reševanje problema izginjajočega in eksplozivno naraščajočega gradienta, ki se pojavi pri standardnih RNN, in so učinkovite pri učenju odnosov v dolgih zaporedjih.

Struktura LSTM celice: Vsaka LSTM celica vsebuje tri regulativna "vrata" – vhodna, izhodna in vrata za pozabo. Ta vrata omogočajo celici, da modulira tok informacij skozi čas, kar povečuje njeno sposobnost ohranjanja pomembnih informacij in pozabljanja nepomembnih.

- **Vhodna vrata (Input gate):** Določajo, katere nove informacije se bodo dodale v celično stanje.
- **Pozabljiva vrata (Forget gate):** Določajo, katere informacije iz preteklega celičnega stanja se bodo zavrgle ali ohranile.
- **Izhodna vrata (Output gate):** Nadzorujejo, katere informacije iz celičnega stanja se bodo uporabile za ustvarjanje izhoda mreže.

Na spodnji sliki prikažemo konceptualni pogled na LSTM.

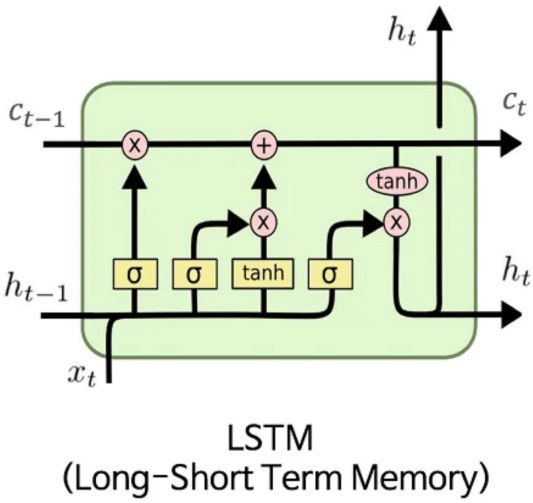

Fig. 4: Konceptualni pogled na LSTm iz vira [5]

Celično Stanje (Cell state): Je kot "nosilec" informacij skozi celotno zaporedje, ki omogoča ohranjanje informacij na dolgi rok. Celično stanje je ključno za zmožnost LSTM, da ohranja in prenaša informacije skozi dolga zaporedja brez problema izginjajočega ali eksplodirajočega gradienta.

Povezovanje celic: Celice v LSTM mreži so povezane zaporedno, kjer izhod ene celice vhod za naslednjo. To zaporedno povezovanje omogoča mreži, da prenaša informacije skozi čas in razvija notranji "spomin", ki lahko zajema kompleksne časovne odvisnosti. Ključna lastnost LSTM je, da vse celice v mreži uporabljajo isti nabor uteži (deljenje uteži), kar pomeni, da se iste uteži uporabljajo na vsakem časovnem koraku.

Matematične enačbe LSTM: Delovanje posamezne LSTM celice je opredeljeno z naslednjimi enačbami:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (5)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (6)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (7)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (8)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (9)$$

$$h_t = o_t \odot \tanh(C_t) \quad (10)$$

kjer σ je sigmoidna funkcija, tanh je hiperbolični tangens, W in b predstavljajo uteži in pragove, f_t, i_t, in o_t so vrata za pozabo, vhodna in izhodna vrata, C_t je stanje celice, h_t je izhodni vektor, in x_t je vhod na časovnem koraku t.

Velikost skrite plasti (Hidden Size): Velikost skrite plasti pri LSTM določa dimenzijo izhodnega vektorja h_t in vektora celičnega stanja C_t za vsako celico. Ta parameter je ključen za določanje zmogljivosti modela pri shranjevanju in obdelavi informacij. Večji "hidden size" omogoča modelu, da se uči bolj kompleksnih vzorcev, vendar pa to hkrati povečuje število modelnih parametrov in računsko zahtevnost.

Postopek učenja: Učenje v LSTM mrežah vključuje prilagajanje uteži in pragov v vratih in povezavah celic za optimizacijo določene kriterijske funkcije. To se običajno

doseže z uporabo algoritma povratnega širjenja skozi čas (BPTT), ki je posebej prilagojen za obdelavo časovnih zaporedij. BPTT upošteva zaporedno naravo podatkov in posodablja uteži tako, da minimizira napako v celotnem zaporedju. Ta postopek omogoča LSTM mrežam, da se učinkovito prilagajajo in izboljšujejo pri modeliranju dolgoročnih odvisnosti v podatkih.

Dinamika mreže: Dinamika mreže LSTM omogoča celicam, da ohranjajo relevantne informacije skozi daljše časovne razpone in učinkovito obdelujejo zaporedne podatke, kar je ključno pri naprednih nalogah, kot so jezikovno modeliranje, generiranje besedila, strojno prevajanje in prepoznavanje govora.

4 Mere za vrednotenje sistemov razpoznavanja

Vrednotenje učinkovitosti sistemov za razpoznavanje temelji na merah, kot so natančnost, priklic, F1-score in točnost. Natančnost (Precision) meri delež pravilno identificiranih pozitivnih primerov med vsemi identificiranimi pozitivnimi primeri. Ta mera nam pove, kolikšen delež identificiranih primerov kot pozitivnih je dejansko pozitiven. To je ključno v situacijah, kjer so posledice lažno pozitivnih rezultatov visoke. Na primer, v medicinski diagnostiki bi lažno pozitiven rezultat lahko pomenil nepotrebno zdravljenje za pacienta, kar bi lahko bilo škodljivo. Priklic (Recall) meri delež pravilno identificiranih pozitivnih primerov med vsemi dejanskimi pozitivnimi primeri v podatkih. Ta mera nam pove, kolikšen delež dejanskih pozitivnih primerov je sistem pravilno identificiral. To je pomembno, ko so posledice lažno negativnih rezultatov pomembne. Na primer, v medicinski diagnostiki bi lažno negativen rezultat lahko pomenil opustitev zdravljenja za pacienta, kar bi lahko bilo usodno. F1-score je harmonično povprečje natančnosti in priklica, ki zagotavlja ravnovesje med obema merama in je še posebej koristen v primerih, kjer sta oba aspekta - natančnost in priklic - enako pomembna. Visok F1-score pomeni, da sistem uspešno združuje visoko natančnost in priklic, kar je pogosto zaželeno v mnogih aplikacijah. Točnost (Accuracy) meri delež vseh pravilno klasificiranih primerov (tako pozitivnih kot negativnih) glede na vse primere.
Matematično so te mere opredeljene kot:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{14}$$

5 Opis podatkovne zbirke

Prvi eksperiment izvedemo na zbirki posnetkov čustvenega govora ***emoDB***. Zbirka *emoDB* (Berlin Database of Emotional Speech) vsebuje 535 posnetkov čustvenega govora, izvedenih s strani 10 govorcev (5 moških in 5 žensk). Posnetki zajemajo sedem različnih čustvenih stanj: veselje, žalost, jeza, strah, gnus, presenečenje in nevtralno stanje. Vsak posnetek je bil skrbno izbran in ocenjen s strani strokovnjakov za zagotovitev visoke kvalitete in reprezentativnosti čustvenih stanj.

Posnetki v zbirki emoDB so bili posneti v kontroliranem studijskem okolju z uporabo visokokakovostne avdio opreme. To zagotavlja jasnost in enotnost zvočnih zapisov. Poleg tega so bili posnetki normalizirani glede na glasnost, da se zmanjšajo variacije v glasnosti med posnetki.

Zbirka emoDB je bila uporabljena v številnih raziskavah na področju avtomatskega prepoznavanja čustev, razvoja čustveno odzivnih sistemov in v študijah o človeški paralingvistiki. Njen prispevek k znanstveni skupnosti je pomemben, saj omogoča raziskovalcem, da preizkusijo in primerjajo različne pristope in tehnike na standardiziranem in dobro dokumentiranem naboru podatkov.

Drugi eksperiment izvedemo na zbirki ***varnostno sumljivih zvokov***. Zbirka je sestavljena iz 2524 varnostno sumljivih zvokov, ki so v povprečju dolgi okoli 5 sekund in razdeljeni v 7 razredov. Razredi vključujejo alarm, pasji lajež, eksplozijo, lomljenje stekla, kričanje, streljanje in sireno. Vsak razred predstavlja specifičen nabor zvokov, ki so pomembni za prepoznavanje potencialnih varnostnih groženj ali izrednih dogodkov.

Zvoki v zbirki so bili zbrani iz različnih virov, predvsem iz spletnega arhiva Freesound, ki je znan po svoji obsežni zbirki avdio posnetkov.

6 Metoda 2D CNN in LSTM omrežja

Arhitektura uporabljena v tem delu združuje 2D konvolucijske nevronske mreže (CNN) z dolgoročnim kratkoročnim pomnilnikom (LSTM) za naloge prepoznavanja čustev iz govora ali razpoznavanje varnostno sumljivih zvokov. Metoda se zgleduje po članku J. Zhaa et al. [3]. Ključne komponente vključujejo:

Blok za učenje lokalnih značilnosti (LFLB): Ta modul je ključnega pomena v strukturah globokih nevronskih mrež, zlasti v konvolucijskih nevronskih mrežah, in je namenjen izvlečenju vitalnih lokalnih značilnosti iz vhodnih podatkov. Sestavljen je iz:

1. **Konvolucijskih plasti:** Uporabljajo konvolucijo za detekcijo vzorcev, kot so robovi in teksture v podatkih.
2. **Normalizacijskih plasti:** Te plasti sledijo konvolucijskim plastem in normalizirajo njihove izhodne podatke. Ta postopek pomaga stabilizirati proces učenja in pospeši konvergenco modela med usposabljanjem.
3. **Eksponencialnih linearnih enot (ELU):** To so napredne funkcije aktivacije, ki sledijo konvolucijskim in normalizacijskim plastem. ELU omogoča modelu, da se nauči bolj kompleksnih vzorcev v podatkih in izboljša splošno učinkovitost učenja.
4. **Sloji za maksimalno združevanje (Max pooling):** Ti sloji zmanjšujejo dimenzionalnost vhodnih podatkov, s čimer optimizirajo računalniške zahteve in pomagajo pri preprečevanju prekomernega prilagajanja na usposabljanje podatkov.

Skupaj ti elementi omogočajo LFLB, da efektivno identificira in ekstrahira pomembne lokalne značilnosti iz vhodnih podatkov, kar je bistvenega pomena za različne aplikacije v globokem učenju.

Globalno učenje značilnosti preko LSTM: Po bloku LFLB se z uporabo plasti LSTM uči dolgoročne odvisnosti iz zaporedja lokalnih značilnosti.

Obdelava mel-spektrogramskih slik: V članku [3] je omenjeno, da je omrežje, ki združuje 2D konvolucijske nevronske mreže (CNN) in dolgoročni kratkoročni pomnilnik (LSTM), učinkovito pri zajemanju tako lokalnih kot globalnih značilnosti iz log-mel spektrogramskih slik. CNN se osredotoča na izluščenje lokalnih prostorskih značilnosti, kot so teksture in vzorci, medtem ko LSTM obdeluje časovne zaporedne podatke, kar omogoča razumevanje globalnih kontekstualnih informacij v časovni dimenziji.

Spodaj prikažemo postopek izluščanja značilk.

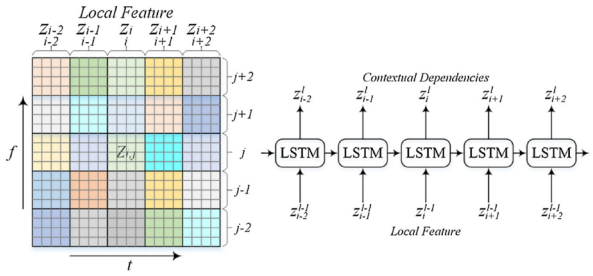

Fig. 5: Slika je vzeta iz vira [3]. Na prikazanem mel-spektrogramu vsak segment z_i predstavlja lokalno značilnost v določenem času t in frekvenci f. Te značilnosti so zaporedno vnesene v omrežje LSTM, kar modelu omogoča učenje časovnih odnosov znotraj podatkov, kot je omenjeno v [3].

Spodnja tabela predstavlja arhitekturo omrežja.

Tab. 1: Arhitektura Modela iz [3]

Layer	Type	Specification
1	Conv2d	64/k3/s1/p1
	BatchNorm2d	64
	ReLU	-
	MaxPool2d	k2/s2
2	Conv2d	64/k3/s1/p1
	BatchNorm2d	64
	ReLU	-
	MaxPool2d	k4/s4
3	Conv2d	128/k3/s1/p1
	BatchNorm2d	128
	ReLU	-
	MaxPool2d	k4/s4
4	Conv2d	128/k3/s1/p1
	BatchNorm2d	128
	ReLU	-
	MaxPool2d	k4/s4
5	LSTM	256 (hidden size)
	Linear	num_classes
	LogSoftmax	-

Za metodo optimizacije izberemo optimizacijsko metodo Adam, ki jo na kratko opišemo v nadaljevanju.

Ob danih parametrih (utežeh) modela θ, optimizator Adam posodablja te parametre θ na vsakem koraku iteracije t:

Izračun gradienta glede na kriterijsko funkcijo J_t v časovnem koraku t:

$$g_t = \nabla_\theta J_t(\theta_{t-1}) \tag{15}$$

Posodobitev pristranske ocene prvega momenta:

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \tag{16}$$

Posodobitev pristranske ocene drugega surovega momenta:

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \tag{17}$$

Izračun pristransko korigirane ocene prvega momenta:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{18}$$

Izračun pristransko korigirane ocene drugega surovega momenta:

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{19}$$

Posodobitev parametrov:

$$\theta_t = \theta_{t-1} - \frac{\text{lr} \cdot \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{20}$$

kjer:

- $f_t(\theta)$ je kriterijska funkcija, ki jo optimiziramo.
- g_t je gradient kriterijske funkcije v časovnem koraku t.
- m_t in v_t sta oceni prvega in drugega momenta gradientov.
- $\hat{m}_t$ in $\hat{v}_t$ sta pristransko korigirani verziji m_t in v_t.
- β_1 in β_2 sta stopnji eksponentnega upadanja teh ocen momentov.
- lr je učna stopnja, nastavljena na 0.0006.
- ϵ je majhna skalarna vrednost, ki preprečuje deljenje z ničlo.
- θ_t so posodobljeni parametri vektorja na časovnem koraku t.

Pri učenju modela uporabljamo L2 regularizacijo, znano tudi kot utežna dekadenca (weight decay), ki deluje kot mehanizem za preprečevanje prekomernega prilagajanja (overfitting) v modelu. Vključitev L2 regularizacije v kriterijsko funkcijo je definirana z naslednjo enačbo:

$$J(\theta) = L(\theta) + \lambda \sum_{i=1}^{n} \theta_i^2 \tag{21}$$

kjer:

- $J(\theta)$ predstavlja skupno kriterijsko funkcijo.
- $L(\theta)$ je originalna kriterijska funkcija modela. Za logaritemsko softmax funkcijo je definirana kot:

$$f(\theta_i) = \log\left(\frac{e^{\theta_i}}{\sum_j e^{\theta_j}}\right) \tag{22}$$

  kjer θ_i predstavlja i-to komponento vhodnega vektorja θ, in vsota v imenovalcu se izvaja čez vse komponente tega vektorja.
- λ je regularizacijski parameter, ki uravnava moč L2 regularizacije.
- $\sum_{i=1}^{n} \theta_i^2$ je L2 norma kvadratov parametrov modela, kar pomeni, da se kaznujejo večje vrednosti uteži.

L2 regularizacija deluje tako, da dodaja kazenski izraz k skupni izgubi, ki je sorazmeren s kvadratom velikosti uteži. To vodi k temu, da model preferira manjše in bolj regularne uteži, kar pomaga preprečiti prekomerno prilagajanje na podatke. Velike vrednosti parametrov so pogosto znak, da se model preveč specifično prilagaja učnim podatkom, izgublja pa sposobnost generalizacije na nove, nevidene podatke. Zato L2 regularizacija pomaga ohranjati model bolj splošen in manj nagnjen k prekomernemu prilagajanju.

7 Eksperiment razpoznavanja čustvenih stanj govorcev

V naši metodi, ki sledi modelu, opisanem v članku [3], pridobivamo mel-spektrogramske slike iz zvočnih zapisov iz zbirke emoDB. Te slike nato uporabimo kot vhod v model, ki združuje 2D konvolucijske nevronske mreže (CNN) in dolgoročno kratkoročno pomnilniško mrežo (LSTM). Spektrogramske slike dobimo tako, da posnetke bodisi skrajšamo bodisi podaljšamo s tišino, da dosežemo enotno dolžino 8 sekund, pri čemer je frekvenca vzorčenja 16 kHz. Uporabljamo dolžino okna 2048 za FFT in dolžino skoka 512. To nam da log-mel spektrogram s 251 časovnimi okni in 128 mel frekvenčnimi bini za vsako časovno okno. Podatke tudi normaliziramo s povprečno vrednostjo in standarno deviacijo (ang. Z-score normalisation).

Podatke razdelimo na učne (4/5 vseh podatkov), validacijske (1/10 vseh podatkov) in testne (1/10 vseh podatkov). Učni podatki se uporabljajo za treniranje modela, kjer se model "uči" iz podatkov. Validacijski podatki služijo za fino nastavitev parametrov modela in preprečevanje prenaučenosti, to je, da bi se model preveč prilagodil samo na učne podatke in ne bi bil generalno uporaben. Testni podatki so ločeni od učnih in validacijskih in se uporabljajo za končno oceno uspešnosti modela, saj predstavljajo neznane podatke, na katerih se preveri, kako dobro model deluje na primerih, ki jih med učenjem ni videl.

Rezultate učenja prikažemo na slednji sliki, ki prikazuje kako so se skozi iteracije spreminja vrednost kriterijske funkcije (ang. loss) in pa točnost razpoznavanja (ang. accuracy)

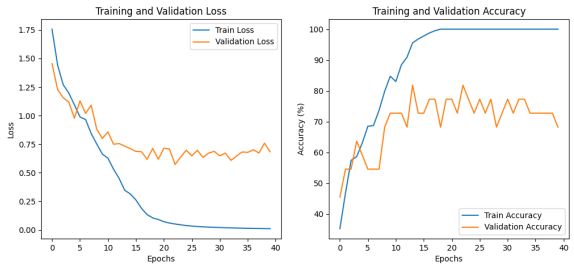

Fig. 6: Vrednost kriterijske funkcije levo in stopnje razpoznavnja desno skozi iteracije učenja

Opazna je prenaučenost modela na učnih podatkih, ki se kaže v stopnji razpoznavanja 100% na učnih podatkih, medtem pa je generalizacija na validacijske podatke veliko slabša. Znaša kar 20 % manj. Po učenju izrišemo konfuzijsko matriko testnih podatkov:

Stopnja pravilnega razpoznavanja na testnih podatkih znaša 74%. Natančnost znaša 0.79, priklic 0.78 in F1 score 0.77. Rezultati testiranja modela kažejo na razmeroma dobro delovanje, s stopnjo pravilnega razpoznavanja 74%. To pomeni, da model učinkovito obravnava kompleksne naloge, čeprav še vedno obstaja prostor za izboljšave. Natančnost 0.79 in priklic 0.78 skupaj kažeta, da model ne le dobro loči med različnimi kategorijami, temveč tudi uspešno identificira relevantne primere. F1 score 0.77 dodatno potrjuje uravnoteženost med natančnostjo in priklicem, kar je ključno za aplikacije, kjer sta oba aspekta pomembna. Kljub temu ti rezultati nakazujejo na možnosti za nadaljnje izboljšave modela, zlasti v smislu zmanjševanja števila napačno klasificiranih primerov.

Ker je model prenaučen (100% stopnja razpoznavanja na

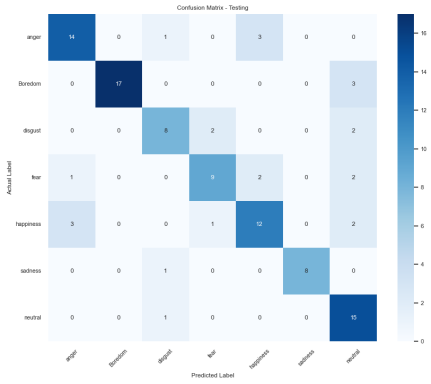

Fig. 7: Konfuzijska matrika na y osi prikazuje dejanska, na x osi pa razpoznana čustvena stanja.

učni in 74% na validacijski zbirki), povečamo zbirko učnih podatkov tako, da originalnim učnim podatkom dodamo šum, spremenimo osnovno frekvenco ali pa jih raztegnemo po časovni osi. Parametre perturbacij izbiramo naključno znotraj določenih intervalov vrednosti. Za učenje uporabimo poleg originalnih tudi perturbirane učne podatke. Novo nastala zbirka je tako za faktor 4 večja od prvotne. Testne in validacijske zbirke ne spreminjamo.

Dodajanje Šuma Naj bo $x(t)$ originalni signal. Šum $n(t)$ lahko dodamo signalu tako, da ustvarimo perturbiran signal $y(t) = x(t) + \alpha \cdot n(t)$, kjer je α faktor šuma, ki se naključno izbere iz določenega intervala.

Spreminjanje Osnovne Frekvence Če želimo spremeniti osnovno frekvenco signala, lahko uporabimo tehniko, kot je modulacija frekvence. Če je f originalna frekvenca, potem novo frekvenco f' dobimo kot $f' = f + \delta$, kjer je δ naključno izbrana vrednost iz določenega intervala.

Raztegovanje po Časovni Osi To dosežemo s spremenitvijo hitrosti reprodukcije signala. Če je $x(t)$ originalni signal, potem raztegnjeni signal $y(t)$ dobimo kot $y(t) = x(\beta \cdot t)$, kjer je β faktor raztegovanja, ki se prav tako naključno izbere iz določenega intervala.

Po ponovnem učenju dosežemo naslednje rezultate

Fig. 8: Vrednost kriterijske funkcije levo in stopnje razpoznavnja desno skozi iteracije učenja

Po učenju na večji množici podatkov, je model sposoben bolje razvrščati tudi validacijske vzorce in sicer z stopnjo razpoznavanja 100%. Natančnost, priklic in F1-score znašajo 1.

Po učenju izrišemo konfuzijsko matriko testnih podatkov:

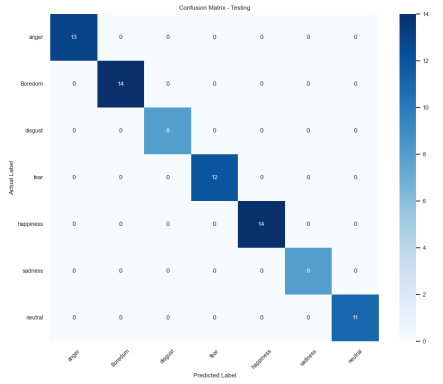

Fig. 9: Konfuzijska matrika na y osi prikazuje dejanska, na x osi pa razpoznana čustvena stanja.

8 Eksperiment razpoznavanja varnostno sumljivih zvokov

Podatke iz zbirke varnostno sumljivih zvokov razdelimo na učne (4/5 vseh podatkov), validacijske (1/10 vseh podatkov) in testne (1/10 vseh podatkov). Učno množico razširimo z prej omenjenimi perturbacijami.

Iz posnetkov izluščimo mel-spektogramske slike tako kot smo to storili v poglavju 7 (slike tudi normaliziramo z mean-std normalizacijo tako kot prej) in jih podamo modelu 2D CNN v kombinaciji z LSTM.

Po učenju dosežemo naslednje rezultate

Fig. 10: Vrednost kriterijske funkcije levo in stopnje razpoznavnja desno skozi iteracije učenja

Model sposoben razvrščati tudi validacijske vzorce z 100% stopnjo razpoznavanja.

Po učenju izrišemo konfuzijsko matriko testnih podatkov:

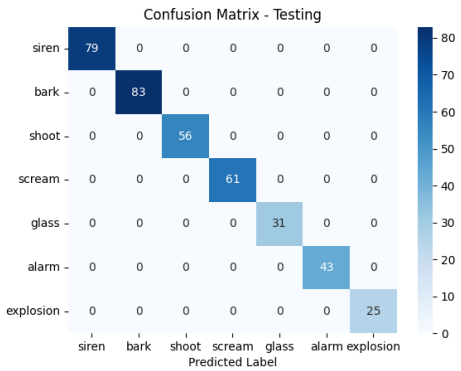

Fig. 11: Konfuzijska matrika na y osi prikazuje dejanske, na x osi pa razpoznane varnostno sumljive zvoke.

Stopnja razpoznavnja na testnih podatkih znaša 100%. Natančnost, priklic in F1-score znašajo prav tako 1.

Dobljeni rezultati na zbirki **varnostno sumljivih zvokov** so boljši od uporabljenega CNN-ja z drugačno arhitekturo na spektogramksih slikah izluščenih iz zvočnih signalov, kjer dosežemo stopnjo razpoznavanja testnih podatkov 94% ter od MLP-ja s stopnjo razpoznavanja testnih podatkov 88 % kjer so kot značilke uporabljene povprečne MFCC značilke pridobljene s povprečenjem po času.

9 Ablacijska študija

Pri prejšnih metodah smo uporabili CNN v kombinaciji z LSTM. Morda iz prej navedenega ni bila najbolje razvidna izhodna dimenzija podatkov iz CNN omrežja oziroma vhodna dimenzija podatkov za LSTM omrežje.

Table 2. The layer parameters of the 2D CNN LSTM network. The output dimension is represented as height × width × number. M × N is the size of the low-level features. The kernel size K of 2 F is the number of the emotions. 2C1 and 2P1 are the convolutional layer and the max-pooling layer of 2 LFLB1, and so on.

Name		Output Dim	Kernel Size	Stride
2 LFLB1	2C1	M × N×64	3 × 3	1 × 1
	2P1	M/2 × N/2 × 64	2 × 2	2 × 2
2 LFLB2	2C2	M/2 × N/2 × 64	3 × 3	1 × 1
	2P2	M/8 × N/8 × 64	4 × 4	4 × 4
2 LFLB3	2C3	M/8 × N/8 × 128	3 × 3	1 × 1
	2P3	M/32 × N/32 × 128	4 × 4	4 × 4
2 LFLB4	2C4	M/32 × N/32 × 128	3 × 3	1 × 1
	2P4	M/128 × N/128 × 128	4 × 4	4 × 4
2 L		–	256	–
2 F		–	K	–

Fig. 12: Dimenzije izhodnih podatkov po vsaki plasti omrežja. Slika je vzeta iz vira [3].

Mel-spektogramske slike so dimenzij 128×251, kjer 251 predstavlja število časovnih oken, 128 pa število mel-frekvenčnih binov za vsako okno. Preračunane izhodne dimenzije podatkov CNN omrežja znašajo $(1 \times 1 \times 128)$, kar je razvidno iz slike 12. To pomeni, da je vhod v LSTM omrežje vektor 128 značilk z dolžino sekvence 1. To je razvidno tudi iz kode na github implementaciji omrežja opisanega v članku 3. Glede na to, da LSTM omrežje uporabljamo za sekvenčne podatke, pri nas pa sekvence ni (dolžina niza na vhodu v LSTM znaša 1, število značilk pa 128), ta plast ni potrebna. Po izhodu iz CNN lahko izluščene značilke podamo polno povezanim plastem. Rezultati, ki jih s tem dobimo, so enaki prejšnim.

Razlog za takšno arhitekturo CNN-ja, ki ne izkorišča dobrih lastnoti LSTM-ja (učenje konteksta v sekvencah podatkov) je morda v uporabljeni optimizaciji hiper-parametrov, ki jo v članku [3] opravijo z Bayesovo optimizacijsko metodo.

Spodaj prikažemo rezultate s perturbiranimi učnimi vzorci, kjer smo LSTM omrežje odstranili.

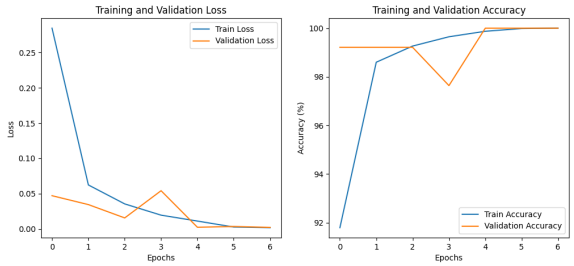

Fig. 13: Vrednost kriterijske funkcije levo in stopnje razpoznavnja desno skozi iteracije učenja na zbirki **emoDB**. LSTM omrežje ni uporabljeno.

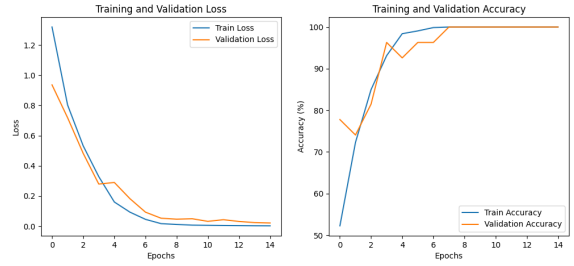

Fig. 14: Vrednost kriterijske funkcije levo in stopnje razpoznavnja desno skozi iteracije učenja na zbirki **varnostno sumljivih zvokov**. LSTM omrežje ni uporabljeno.

Konfuzijske matrike za testne podatke so enake kot na slikah 11 in 9, prav tako tudi metrike kot so natančnost, priklic, točnost in F1-score.

Z ne razširjeno učno množico dobimo enake rezultate kakor z uporabo LSTM omrežja; točnost na testni zbirki znaša 74 %.

10 Metoda 1D CNN in LSTM omrežja ter eksperimenta na zbirki emoDB in zbirki varnostno sumljivih zvokov

V nadaljevanju preizkusimo tudi lastno omrežje 1D CNN in LSTM z naslednjo arhitekturo, kjer pa uporabimo MFCC značilke:

Tab. 2: Arhitektura 1D CNN in LSTM Modela

Layer	Type	Specification
1	Conv1d BatchNorm1d ReLU	(39, 251) → (512, 251)/k5/s1/p2 (512, 251) -
2	Conv1d BatchNorm1d ReLU	(512, 251) → (256, 251)/k5/s1/p2 (256, 251) -
3	Conv1d BatchNorm1d ReLU	(256, 251) → (256, 251)/k5/s1/p2 (256, 251) -
4	Conv1d BatchNorm1d ReLU	(256, 251) → (128, 251)/k3/s1/p1 (128, 251) -
5	LSTM Dropout	(128, 251) → (50, 251) 0.5
6	Linear LogSoftmax	(50, 251) → (7) -

MFCC značilke imajo dimenzije (39, 251), kjer 39 predstavlja število MFCC značilk (13) in njihovih prvih (13) ter drugo stopenjskih (13) odvodov, 251 pa število časovnih oken. Pred izluščenjem MFCC značilk smo originalne posnetke tako kot prej skrajšali oz. podaljšači s tišino. Tako so vsi posentki dolgi 8 sekund. MFCC značilke smo tudi standardizirali z Z-score normalizacijo.

Postopek določanja značilk MFCC:

1. Za vsak segment signala, ki smo ga pridobili z oknenjem, izračunamo njegovo transformiranko. To storimo z DFT.

$$F(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)e^{-i\frac{2\pi nk}{N}}, \qquad 0 \le k \le N-1, \tag{23}$$

kjer je N število točk, ki jih uporabimo za izračun DFT.

2. Kepster signala definiramo kot

$$\{c(n)\} = \mathcal{F}^{-1}\{\log F(k)\}. \tag{24}$$

Ker uho ni občutljivo na fazne zamike med frekvenčnimi komponentami, uporabimo le močnostni spekter signala.

$$\{c(n)\} = \mathcal{F}^{-1}\{\log |F(k)|^2\} \tag{25}$$

3. Za nadaljnje izboljšanje kepstralne reprezentacije v enačbo {c(n)} vključimo več informacij o slušnem zaznavanju. Log-spekter že upošteva zaznavno občutljivost po amplitudni osi, saj je občutljivost ušesa po amplitudi logaritemska. Slušno zaznavanje pa je tudi po frekvenčni osi porazdeljeno nelinearno.

$$f_{\text{Mel}} = 2595 \log_{10}(1 + \frac{f}{700}), \tag{26}$$

kjer je f frekvenca v Hz f_{Mel}, pa zaznana frekvenca.

Zato se za izračun kepstra uporablja logaritme povprečnih moči frekvenčnih območij, razporejenih po melodični delitvi. Tako utežen kepster imenujemo melodični kepster. Uteženo povprečje močnostnega spektra oknenjega zvočnega odseka izračunamo kot:

$$s(m) = \sum_{k=0}^{N-1} [|F(k)|^2 H_{\text{m}}(k)], \qquad 0 \le m \le M-1, \tag{27}$$

kjer je M število vseh trikotnih melodičnih filtrov. Vsak filter je neničeln le na določenih frekvencah. $H_{\text{m}}(k)$ je utežna funkcija, vezana na k-ti vzorec spektra. Izražena je na sliki 1.

Izračun koeficientov melodičnega kepstra s pomočjo Inverzne diskretne kosinusne transformacije:

$$c_{\text{Mel}}(n) = \sum_{m=0}^{M-1} (\log s(m)) \cos\left(\frac{\pi n(m-0.5)}{M}\right); \tag{28}$$

kjer je $n = 0, 1, 2, ..., C-1$, C število koeficientov MFCC. Izračunani koeficienti predstavljajo vektor značilk, ki pripada posameznemu odseku signala, pridobljenega z oknenjem.

Delta (prvi odvod) in delta-delta (drugi odvod) značilke MFCC, prinašajo informacije o hitrosti in pospešku spreminjanja osnovnih MFCC značilk skozi čas. S tem se zagotovi bolj dinamičen vpogled v spremembe zvočnega signala, kar je še posebej pomembno v aplikacijah, kot je prepoznavanje govora.

Brez augmentiranih učnih podatkov dosežemo na testnih podatkih zbirke emoDB stopnjo razpoznavanja 70%, kar

je nekoliko manj kakor z metodo 2D CNN in LSTM. Z augmentacijo pa dosežemo stopnjo razpoznavanja na testni zbirki 100%.

Na testnih podatkih zbirke varnostno sumljivih zvokov prav tako dosežemo stopnjo razpoznavanja 100%.

11 Zaključek

V zaključku lahko povzamemo, da predstavljeni model, ki združuje konvolucijsko nevronsko mrežo in LSTM, kaže obetavne rezultate v razpoznavanju čustvenih stanj govorcev in varnostno sumljivih zvokov. Modelova sposobnost prilagajanja različnim aplikacijam poudarja njegovo prenosljivost in potencial za široko uporabo. Dodatno, raziskava pokaže, da količina in kakovost učnih podatkov ključno vplivata na natančnost in zanesljivost modela, kar nakazuje potrebo po obsežnih in raznolikih podatkovnih zbirkah za nadaljnje izboljšanje modelov.

V ablacijski študiji pokažemo odvečnost LSTM omrežja glede na predhodno arhitekturo CNN omrežja in dimenzije vhodnih podatkov. To je razvidno iz enakega delovanja omrežja z ali brez dodane LSTM plasti.

Na koncu preizkusimo še lastno 1D CNN in LSTM omrežje na MFCC značilkah in dosežemo primerljive rezultate kot z prej omenjeno metodo.

12 Reference

1 D. M. Low, K. H. Bentley, S.S. Gosh, Automated assessment of psychiatric disorders using speech: A systematic review, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7042657/. (Datum ogleda: 20. 10. 2023)

2 T. Bäckström, Cepstrum and MFCC, https://wiki.aalto.fi/display/ITSP/Cepstrum+and+MFCC

3 J. Zhao a, X. Mao a, L. Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, https://www.sciencedirect.com/science/article/pii/S1746809418302337. (Datum ogleda: 15. 12. 2023)

4 Aphex34 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=45679374(Datum ogleda: 15. 12. 2023)

5 Rahuljha, LSTM Gradients, https://towardsdatascience.com/lstm-gradients-b3996e6a0296 (Datum ogleda: 15. 12. 2023)