CTC (Connectionist Temporal Classification) este o tehnică utilizată în învățarea automată (machine learning) pentru a antrena modele de recunoaștere a vorbirii sau a textului. CTC este utilizat în special pentru a recunoaște și a transcrie cuvinte sau sunete care apar într-o secvență fără a fi cunoscute în prealabil limitele exacte ale cuvintelor sau sunetelor individuale.

RNN(rețea neuronală recurentă)

Language model (LM) = e un model care are rolul de a prezice ce caracter/cuvant apare dupa o anumita secventa, de a evalua probabilitatea apariției a unui șir de cuvinte sau caractere într-un anumit limbaj

n-gram= e un tip de LM

kenlm = o librarie care are 2 structuri de date, ajuta la predictia de cuvinte bazata pe alte cuvinte

statistical language model = model de limbaj statistic, poate fi folosita modalitatea n-gram

SRI-LM is a toolkit for building and applying statistical language models (LMs), primarily for use in speech recognition, statistical tagging and segmentation, and machine translation.

langchain = e un tool/framework care ne permite sa facem aplicatii 'normale' folosind large language models

and prompts

prompting = cand ii zicem la model sa genereze text conditionat de un alt input

large language model = Un large language model (LLM) este un tip de model de AI utilizat pentru a genera

și a analiza textul într-un mod semnificativ mai complex decât modelele mai vechi.

LLM-urile sunt antrenate pe cantități masive de date de text și sunt capabile să genereze

text nou într-un stil coerent și naturalețe. Aceste modele pot fi utilizate pentru o varietate

de sarcini, inclusiv generarea de text, traducerea de limbă, răspunsul la întrebări

și analiza textului. Exemple de LLM-uri sunt GPT-3 de la OpenAO

* HMM approach, primele tentative de speech recognition din lume
* Kaldi oferă resurse pentru normalizarea fișierelor audio. De exemplu, Kaldi include un script numit normalize\_audio.py, care utilizează Python Speech Recognition Library (PyDub) pentru a normaliza fisierul audio la un nivel constant de volum și a elimina zgomotul de fundal. De asemenea, Kaldi include și alte scripturi și unelte pentru preprocesarea și pregătirea datelor audio pentru antrenarea modelelor de recunoaștere a vorbirii.

Perceptual Linear Predictive Coding (PLPC) și Perceptual Cepstral Coding (PCC) sunt două tehnici de compresie a semnalului audio, utilizate în special pentru codarea vocii umane. Ambele metode au ca scop reducerea dimensiunii fișierelor audio prin eliminarea datelor redundante sau irelevante, fără a afecta calitatea percepută a sunetului de către ascultător.

Perceptual Linear Predictive Coding este o metodă de compresie care se bazează pe un model matematic al vocii umane, care estimează parametrii fizici ai vocii umane, cum ar fi frecvența fundamentală, formanta și energia, prin utilizarea unui filtru de predicție liniară. Acești parametri sunt apoi cuantificați și comprimați, reducând astfel dimensiunea fișierelor audio.

Perceptual Cepstral Coding, pe de altă parte, utilizează o transformare matematică numită transformare cepstrală, care convertește semnalul audio din domeniul timpului în domeniul frecvenței. Această transformare este aplicată astfel încât să reflecte modul în care urechea umană percepe semnalul audio, prin eliminarea componentelor de frecvență irelevante sau cu o pondere redusă în percepția umană.

În ambele cazuri, tehnologiile PLPC și PCC permit o compresie semnificativă a datelor audio, cu pierderi minime de calitate percepută a sunetului. Acestea sunt utilizate adesea în sistemele de telefonie mobilă, dispozitive audio portabile și alte aplicații de stocare și transfer de fișiere audio.

Analiza cepstrală este utilizată într-un pas cheie al recunoașterii vocale, numit extragerea caracteristicilor (feature extraction). Acest pas are loc înainte de aplicarea ferestrei Hamming și de trecerea caracteristicilor prin rețeaua neuronală.

Extragerea caracteristicilor are scopul de a transforma semnalul vocal într-o reprezentare numerică care capturează informațiile relevante pentru recunoașterea cuvintelor. Aici, analiza cepstrală joacă un rol important. Procesul constă în aplicarea unei transformate Fourier asupra logaritmului spectrului puterii semnalului vocal, ceea ce rezultă în coeficienții cepstrali. Acești coeficienți reflectă caracteristicile semnificative ale semnalului vocal, cum ar fi formanțele și energia.

În timpul analizei cepstrale, se elimină informațiile redundante sau irelevante, cum ar fi componentele de înaltă frecvență. Apoi, se aplică o fereastră Hamming pentru a atenua fenomenul de spectru de undă. Aceasta constă în multiplicarea semnalului cu o funcție fereastra Hamming, care reduce efectele secundare ale transformării Fourier, cum ar fi spectral leakage.

După aplicarea ferestrei Hamming și obținerea coeficienților cepstrali, aceștia sunt utilizati ca feature-uri de intrare pentru rețeaua neuronală. Rețeaua neuronală preia aceste caracteristici și învață să le asocieze cu cuvintele specifice în timpul fazei de antrenare. În timpul testării, rețeaua neuronală utilizează caracteristicile extrase pentru a face predicții cu privire la cuvintele din semnalul vocal de intrare.

Modelul acustic utilizează aceste caracteristici extrase, inclusiv MFCCs, împreună cu o arhitectură de rețea neuronală sau alte metode de învățare automată, pentru a învăța să asocieze caracteristicile acustice cu unitățile fonetice și cuvintele. Astfel, modelul acustic reprezintă partea componentă a sistemului de ASR care este responsabilă pentru decodificarea semnalului audio într-o secvență de cuvinte sau de unități de limbă recunoscute.

În rezumat, un model acustic în recunoașterea vorbirii este responsabil pentru transformarea semnalului audio într-o reprezentare abstractă și pentru asocierea caracteristicilor acustice, inclusiv coeficienții cepstrali de frecvență Mel (MFCC), cu unități fonetice sau cuvinte recunoscute.

Kaldi = RSC 2.79%, ssc-eval 16.63% pe cuvant, sunt diferite, citit si vorbit spontan = medie de 9.71

Robin TOWARDS A ROMANIAN END-TO-END AUTOMATIC SPEECH RECOGNITION [1]

* 230 de ore si a reusit o eroare per cuvant de 9.91% si eroare per caracter de 2.81%
* Inainte de robin era facuta alta arhitectura cu Kaldix, ce e robin acum e cu mai putine layere pentru a aduce eficienta de timp
* Se zice de kaldix

ROMANIAN SPEECH RECOGNITION EXPERIMENTS FROM THE ROBIN PROJECT

* Aici zice de kaldi ca e mai lent datorita sistemului pipeline
* Scopul proiectului este ROBIN Dialog, de a imbunatati viteza si precizia in microworlds
* Asr putin diferit de cel robin
* 2019a -> Kaldi y

Folosind modelul ASR și modelul lingvistic combinat (cu parametrii optimi), rezultatul general a fost de 9,91% WER pe setul de testare, îmbunătățind WER al sistemului ASR brut cu 5,66%. (din experiments)

În cercetarea anterioară, a fost implementat un sistem de recunoaștere automată a vorbirii (ASR) pentru limba română, folosind arhitectura DeepSpeech2. A fost realizat un proces de prelucrare a datelor, în care coeficienții cepstrali de frecvență Mel (MFCC) au fost calculați pe ferestre audio de 20 ms. Spectrograma rezultată a fost utilizată ca intrare într-un model de rețea neuronală profundă. Modelul folosit a constat în 8 straturi neuronale, incluzând 2 straturi convoluționale 2D, 4 straturi BiLSTM cu memorie bidirecțională pe termen scurt, 1 strat de convoluție de tip lookahead și un strat final complet conectat (FC).

În scopul optimizării instruirii modelului, a fost aplicată normalizarea pe loturi (BN) după fiecare strat, cu excepția ultimului strat. De asemenea, s-a folosit funcția de activare HardTanh înainte de fiecare strat de normalizare pe loturi (BN), ca o alternativă eficientă din punct de vedere computațional la funcția tanh.

Kaldi-based DNN Architectures for Speech Recognition in Romanian x

Georgescu et al. (2019) au investigat utilizarea rețelelor neuronale în sistemele de recunoaștere automată a vorbirii (ASR) în limba română, folosind platforma de dezvoltare Kaldi3. În studiul lor, ei au evaluat performanța modelului lor pe două corpusuri: RSC-eval și SSC-eval, obținând rezultate ale Ratei de Erori a Cuvintelor (WER) de 2,79%, respectiv 16,63%. Deși cele două corpusuri reprezintă tipuri diferite de vorbire (citită și spontană), pentru implementarea noastră ne dorim un sistem ASR general, care să funcționeze independent de tipul de vorbire. Prin urmare, în scopul comparației, am luat în considerare media WER a ambelor corpusuri de evaluare, rezultând o valoare medie a WER de 9,71%. (din experiments)

Making Pepper Understand and Respond in Romanian kaldi y

* 25% WER
* De aici e prezentarea robin

Am încercat să evaluăm performanțele modulului ASR,

dar rezultatele preliminare sunt destul de modeste: 25% WER

(Word Error Rate) și 75% SER (Sentence Error Rate). Site-ul

evaluare a fost realizată în principal pe înregistrări radiofonice de discurs liber.

Cu toate acestea, atunci când am evaluat înregistrări de vorbire fără zgomot, rezultatele

recunoașterea a fost aproape perfectă (două recunoașteri ușor greșite

cuvinte din 49). Deoarece interacțiunea vocală adecvată cu

robotului este foarte probabil să fie afectată de zgomot, considerăm că

actuala versiune ASR necesită îmbunătățiri serioase.

A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language

* 2006, e cu hmm

Diferente dintre cele 2 ASR-uri: robin si experiments

MFCC, PLP:

Ambele tehnici sunt eficiente în extragerea caracteristicilor semnalelor audio, dar pot fi utilizate în funcție de necesitățile specifice ale aplicației. PLP este mai potrivit pentru aplicațiile care necesită o reprezentare mai fidelă a semnalelor audio și o mai bună capacitate de distingere a vorbirii de zgomot. Pe de altă parte, MFCC este mai eficient în extragerea caracteristicilor semnalelor audio pentru aplicații de recunoaștere a vorbirii și identificare a vorbitorilor.

* Performanta lui mfcc in prezenta zgomotelor de font este mai mica decat la celelalte

perceptual linear predictive coding = LPC(linear predictive coding )