Glasovno upravljanje igrom Temple Run

SW-48/2018 Anastasija Đurić, SW-52/2018 Dina Petrov
Softversko inženjerstvo i informacione tehnologije, Fakultet tehničkih nauka u Novom Sadu
Soft kompjuting 2021/2022

Problem koji se rešava

Cilj odabranog projektnog zadatka bio je omogućiti upravljanje igricama pomoću glasovnih komandi umesto pritiskom na strelice tastature.
Na osnovu dataset-ova, koji se inicijalno sastoje od audio fajlova izgovorenih komandi za kretanje, generisani su mel spektrogrami korišćeni za obuku konvolucione neuronske mreže.
Obučeni model se potom koristi za klasifikaciju i mapiranje glasovnih komandi (koje korisnik zadaje putem mikrofona) na odgovarajuće tastere za upravljanje kretanjem u igricama poput Temple Run u realnom vremenu.

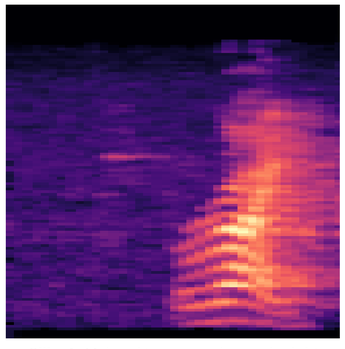
Skupovi podataka

Za rešavanje ovog problema korišćena su dva skupa podataka:

1. Dataset sa komandama na engleskom jeziku
Sadrži **15 219** audio fajlova u .wav formatu sa komandama na engleskom jeziku: **up**, **down**, **left**, **right**. Svaki je dužine 1s.

2. Dataset sa komandama na srpskom jeziku - skoči, dole, levo, desno
Dataset je ručno sakupljen uz pomoć prijatelja i kolega preko opcije glasovnih poruka na Telegram-u (.ogg format), WhatsApp-u (.opus format) i Viber-u (.mp4 format). Glasovne poruke su potom konvertovane u .wav format. Dataset originalno sadrži **1440** audio fajlova (**desno** - 359, **dole** - 369, **levo** - 357, **skoči** - 355).
Na svaki audio fajl primenjeno je 10 različitih audio efekata uz pomoć **spotify pedalboard** biblioteke (*chorus*, *phaser*, *lowpass filter*, *highpass filter*, *reverb*, *distortion*, *+0.2 pitch*, *+0.4 pitch*,*-0.2 pitch*, *-0.4 pitch*). Nakon primenjenih transformacija, ukupan broj audio fajlova u dataset-u broji **15 840**.

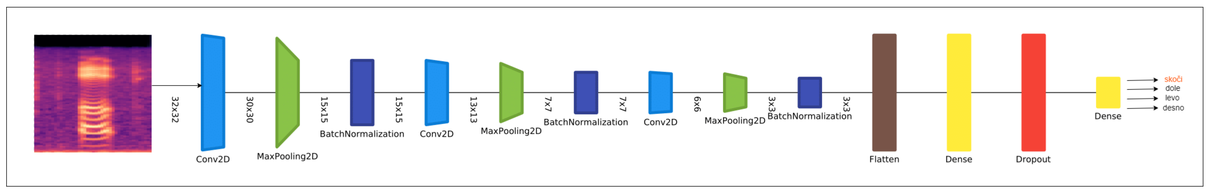
Za svaki audio fajl generisan je mel spektrogram na logaritamskoj skali čime su izraženi detalji zvučnog signala na način blizak ljudskoj percepciji zvuka.



Slika 1. *Mel spektrogram za komandu* ***right***

Arhitektura konvolucione neuronske mreže

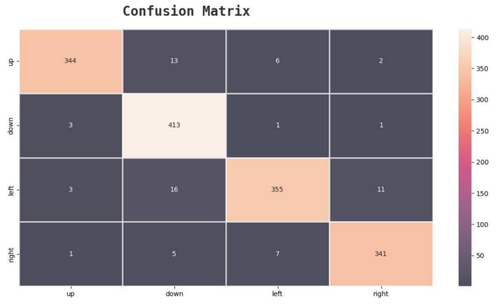
Upotrebom **keras** biblioteke kreiran je sekvencijalni model. On se sastoji od tri konvoluciona sloja. Prvi sloj sadriži 32 filtera, dimenzija 3x3. Aktivaciona funkcija koja se primenjuje na ovom sloju je relu. Zatim je dodat **Max Pooling** sloj sa pool dimenzijama 3x3 i strides 2x2. Finalno, dodat je i **Batch Normalization** sloj zadužen za normalizaciju težine neurona čime se ubrzava proces treniranja, i sam model postaje pouzdaniji. Drugi konvolucioni sloj u potpunosti je isti kao i prvi, dok su kod trećeg dimenzije jezgra i pool size 2x2. Izlaz iz konvolucionih slojeva je dvodimenzionalni niz koji se zatim pretvara u jednodimenzionalni prolaskom kroz **Flatten** sloj. Sledeći slojevi su **Dense** i **Dropout**. Izlazni sloj je **Dense** sloj sa 4 neurona od kojih svaki predstavlja po jednu glasovnu komandu, i softmax aktivacionom funkcijom. Dimenzije ulaza su 32x32x4.



Slika 2. *Arhitektura mreže*

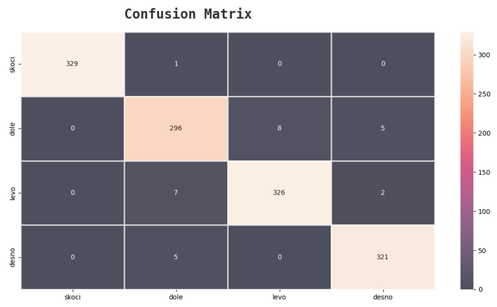
Analiza rezultata

Tensorflow dataset sa komandama na engleskom jeziku
Accuracy vrednost nakon 20 epoha:
- Trening set: 0.9812905192 ~ 98%
- Validacioni set: 0.9667882919 ~ 97%
- Test set: 0.9638633132 ~ 96%



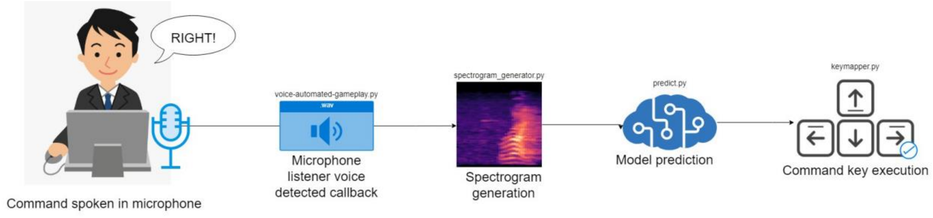
Slika 3. *Matrica konfuzije za dataset na engleskom*

Custom dataset sa komandama na srpskom jeziku
Accuracy vrednost nakon 30 epoha:
- Trening set: 0.9823746085 ~ 98%
- Validacioni set: 0.9687938094 ~ 97%
- Test set: 0.9797979593 ~ 98%



Slika 4. *Matrica konfuzije za dataset na srpskom*

Tok aplikacije



Slika 5. *Tok aplikacije*

Uočeni problemi i dalji rad

Dalji rad na projektu obuhvatio bi dodavanje novih komandi kako bi se podržalo upravljanje i drugim vrstama igrica. Na primer, trenutna verzija pritiska taster i potom ga pušta, dok mnoge igrice zavise od kontinualnog držanja pritisnutog tastera i kasnijeg otpuštanja.
Ipak, najbitniji bi bio dalji rad na optimizaciji. Kako od momenta izgovora komande, preko detektovanja i procesuiranja glasa preko mikrofona, generisanja mel spektrograma, predikcije komande na osnovu ovog spektrograma, do konačnog iniciranja pritiska tastera postoji čitav tok procesa, gotovo da je nemoguće dobiti trenutačni pritisak finalnog tastera odmah po izgovorenoj komandi. Samim tim, igranje igrica koje zahtevaju “brze” reakcije (kao što je Temple Run) je otežano. U inicijalnoj verziji projekta, detektovani glas i generisani spektrogram čuvali su se u fajl sistemu. Dosadašnji implementirani koraci optimizacije potpuno eliminišu rad sa diskom. Prelaskom na **in-memory** rad sa detektovanim glasom i generisanim spektrogramom, došlo je do ubrzanja.
Merenjem vremena operacija utvrđeno je da je usko grlo rada zapravo u maksimalnoj dužini snimanja glasa (parametar **max_phrase_length**). Inicijalno podesivši ovaj parametar na 2 sekunde, tačnost predikcije bila je u skladu sa tačnošću izmerenom nad test skupom. Međutim, dve sekunde je previše dugačak period da bi se igra poput Temple Run mogla igrati, usled kašnjenja izvršavanja komande. Podesivši ovaj parametar na neki od brojeva iz intervala (0.5, 1.0) uočeno je da je igru moguće igrati, jer je vreme odziva pritiska dugmića gotovo trenutačno. Tu se međutim, javlja drugi problem: Skraćivanjem maksimalnog dozvoljenog vremena snimanja komandi dolazi do situacije da se izgovorena komanda preseče pre nego što je izgovorena u celosti, te model nije u stanju da ispravno izvrši predikciju. Neophodno je imati višu maksimalnu dužinu snimanja, kako bi se komanda tačno procesuirala i klasifikovala. Dolazimo do zaključka da je odabrana igra Temple Run nepovoljna za ovakav tip aplikacije, te da bi je trebalo primeniti nad nekim drugim tipom igre, kao što su Tastatour.co i slično.

Zaključak

Mel spektrogrami pokazali su se kao dobar reprezent zvučnog signala korišćenog za problem klasifikacije izgovorenih reči.
Istrenirani modeli dali su odlične rezultate i tačnost, kako nad izdvojenim test skupom, tako i nad izgovorenim komandama preko mikrofona u realnom vremenu. Posebno treba naglasiti performanse modela treniranog nad ručno sakupljenim dataset-om na srpskom jeziku. U sakupljanju ovog dataset-a učestvovalo je ukupno 37 osoba koje su “posudile” svoj glas slanjem brojnih glasovnih poruka na mrežama za dopisivanje. Učesnici su podsticani da u svakoj glasovnoj poruci uvedu varijacije u vidu visine glasa, brzine i jasnoće izgovora, pozicije mikrofona i slično. Ovi faktori uticali su na veliki obim i raznovrsnost podataka u dataset-u.

Reference

https://viscom.net2vis.uni-ulm.de/
https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53
https://librosa.org/doc/latest/index.html
https://github.com/spotify/pedalboard
https://github.com/Uberi/speech_recognition
https://github.com/moses-palmer/pynput