Prepoznavanje muzičkih simbola i generisanje melodije

Autori: Dragan Ćulibrk i Danijel Radulović
Fakultet Tehničkih Nauka, Univerzitet u Novom Sadu

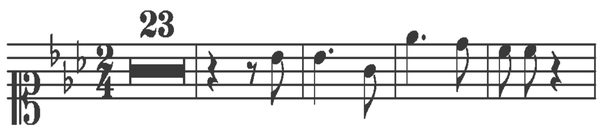
MOTIVACIJA I DEFINICIJA PROBLEMA

Optičko prepoznavanje muzičkih simbola (*Optical music recognition*) predstavlja oblast istraživanja, čiji je cilj softversko čitanje notnih zapisa, prepoznavanje muzičkih simbola i kreiranje mašinski čitljivih verzija notnog zapisa, koje se mogu čuvati u formatima, kao što su MIDI (*Musical Instrument Digital Interface*) i MusicXML. Ova oblast uključuje i druga područja istraživanja, kao što su computer vision, analiza dokumenata i teorija muzike.

Cilj ovog projekta je kreiranje softvera, koji bi trebalo da čita muzičke zapise, prepoznaje muzičke simbole, generiše njihovu tekstualnu predstavu i na osnovu nje generiše melodiju. Korišćen je *End-to-end Optical Music Recognition* algoritam, koji radi na principu prepoznavanja muzičkih simbola jednog reda notnog zapisa.

SKUP PODATAKA

Za trening i validacioni skup podataka korišćen je PrIMuS (*The Printed Images of Music Staves)*. PrIMuS se sastoji od 87678 slika, na kojima je je prikazan po jedan red notnog zapisa i dokumenata, koji predstavljaju tekstualnu predstavu muzičkih simbola sa odgovarajućih notnih zapisa.



clef-C1 keySignature-EbM timeSignature-2/4 multirest-23 barline rest-quarter
rest-eighth note-Bb4_eighth barline note-Bb4_quarter.
note-G4_eighth barline note-Eb5_quarter. note-D5_eighth barline
note-C5_eighth note-C5_eighth rest-quarter barline

ARHITEKTURA REŠENJA

Arhitekturu rešenja predstavlja konvolutivna rekurentna neuronska mreža (*CRNN*), tačnije neuronska mreža, koja se sastoji od 4 konvolutivna i 2 rekurentna bloka. Pre puštanja slika iz skupa podataka u mrežu, slike su skalirane, na osnovu parametra visine slike, koja je iznosila 32px i normalizovane. Model je obučavan na serijama (*batch*) od 8 slika.

Konvolutivni blok

- konvolutivni sloj – inicijalni broj filtera 32, veličina kernela 3x3
- batch normalizacija
- leaky relu aktivaciona funkcija
- pooling sloj – veličina 2x2

Rekurentni blok

- bidirekcioni LSTM (*Long Short Term Memory*) sloj sa 256 jedinica
- dropout – koeficijent 0.5

Posle 4 konvolutivna i 2 rekurentna bloka, sledi potpuno povezani sloj i (*fully connected layer*) i softmax aktivacija. Broj neurona u potpuno povezanom sloju, jednak je broju mogućih tekstualnih predstava muzičkih simbola u rečniku.

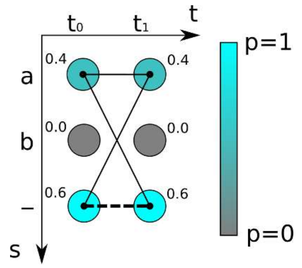
Za klasifikaciju i računanje *loss*-a, korišćen je CTC (*Connectionist temporal classification*), koji se koristi za treniranje rekurentnih neuronskih mreža, za rešavanje *sequence-to-sequence* problema gde postoji vremenska zavisnost između svakog dela ulazne sekvence.

Connectionist temporal classification

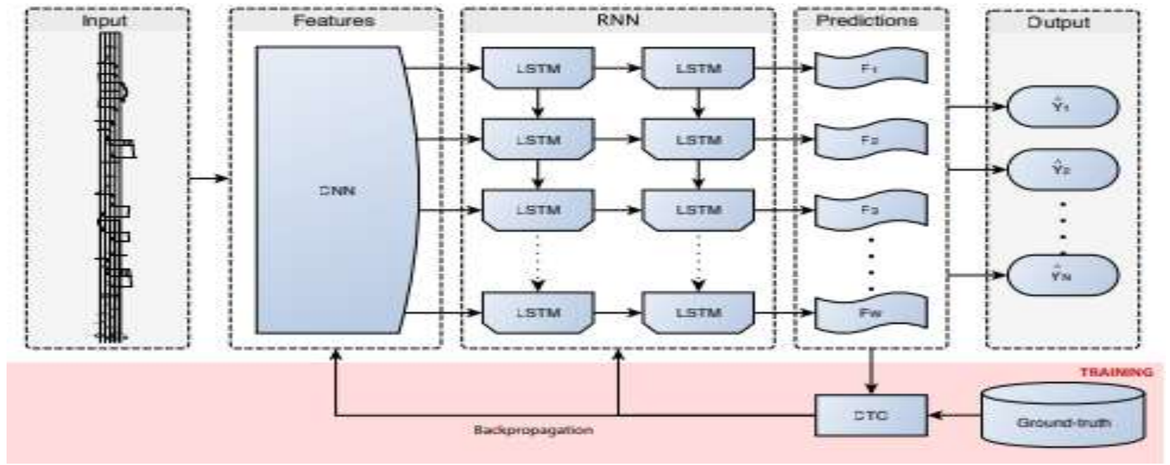
CTC funkcioniše na principu anotacije svake horizontalne pozicije na slici (*time step*), zbog čega se uvodi termin vremenske zavisnosti ulazne sekvence. Kako bi se rešio problem duplih karaktera (npr. reč *too*), uveden je specijalni karakter „-“, zbog čega je broj neurona u potpuno povezanom sloju povećan za jedan.

Na primer:

- „to“ -> „---ttttttooo“ ili „-t-o“ ili „to“
- „too“ -> „---tttttto-o“ ili „-t-o-o-“ ili „to-o“, ali ne može „too“



Da bi se izračunao *loss* u slučaju, koji je dat na slici levo, potrebno je prvo izračunati verovatnoću *ground truth* teksta. Na primer, za „a“ verovatnoća se dobija sumiranjem svih putanja, koje daju slovo „a“. U ovom slučaju, verovatnoća iznosi 0.64 (moguće putanje su „aa“, „a-“ i „-a“. *Loss* predstavlja negativan logaritam od verovatnoće. Isti postupak, korišćen je i u slučaju muzičkih simbola.

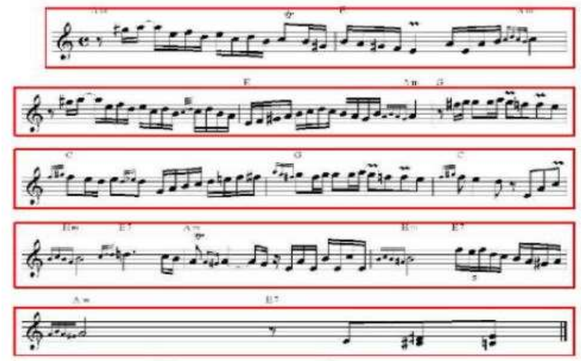


PRIMENA REŠENJA

S obzirom da je ideja bila da je moguće čitati cele notne zapise, a korišćeni algoritam radi samo na jednom redu notnog zapisa, pre primene rešenja, potrebno je ulazni notni zapis, podeliti na pojedinačne redove. To je postignuto korišćenjem tehnika *computer vision*-a.

Nakon učitavanja slike, radi se negativ i pretvaranje u *grayscale*, zatim sledi operacija dilacije (dodaje piksele na ivice objekata) i zatvaranje (kombinacija dilacije i erozije (uklanja piksele sa ivice objekata). Nakon toga se traže konture i filtriraju po dimenzija, kako bi izdvojili samo redove notnog zapisa.

Redovi notnog zapisa se zatim, jedan po jedan, propuštaju kroz obučenu mrežu i tako dobijene tekstualne predstave muzičkih simbola, se konkateniraju i konvertuju u MIDI fajl.



ZAKLJUČAK

U dosadašnjim radovima i istraživanjima na temu OMR-a, koristio se tradicionalni pristup, zasnovan na više faza. Prvo je bilo potrebno inicijalno pretprocesiranje slika notnih zapisa, koje uključuje binarizaciju slike, detekciju redova, podela redova po taktnim crtama ili odvajanje teksta od melodije.

Dosta se pažnje obraćalo na uklanjanje notnih linija, jer njihovo prisustvo sprečava izolaciju muzičkih simbola. U narednim fazama je sledila klasifikacija izolovanih muzičkih simbola, pomoću *k-Nearest Neighbors* ili *Support Vector Machines* i na kraju, interpretacija i davanje muzičkog značenja, tako izolovanim i klasifikovanim simbolima. *End-to-end Optical Music Recognition* algoritam, koji je korišćen pri realizaciji ovog rešenja, posmatra jedan red notnog zapisa kao jedinicu, umesto sekvencu izolovanih elemenata i na taj način, uzima u obzir i muzički kontekst. Duboke neuronske mreže (*Deep Neural Networks*) su se pokazale dobro u sličnim problemima, kao što su prepoznavanje teksta i govora, stoga je ovakav pristup primenjen uspešno i u OMR-u.

Za dalji razvoj primene datog algoritma u ovoj oblasti, trebalo bi uzeti u razmatranje i notne zapise slikane pod različitim okolnostima, kao što su loše osvetljenje i geometrijske distorzije. Takođe, tu spada i uključivanje znakova za muzičku artikulaciju, dinamiku i tempo, kao i polifonije, u obliku akorada ili više instrumenata.