Prirodoslovno matematički fakultet u Zagrebu

Prepoznavanje govora koristeći TensorFlow

Projektni prijedlog za kolegij Strojno učenje

Karmen Kapov, Petra Međimurec, Siniša Novak, Petra Vitez
30. travnja, 2018.

Sadržaj

Uvodni opis problema

Prepoznavanje glasa ili prepoznavanje govora (engl. *speech recognition*) je transformiranje ljudskog govora u tekst, odnosno sposobnost stroja ili programa da prepoznaje riječi i fraze i pretvori ih u strojno čitljiv format. Problem koji pokušavamo riješiti je problem predstavljen na natječaju na web stranici www.kaggle.com u kojem je izazov bio upotrijebiti dataset za naredbe govora za izradu algoritma koji razumije jednostavne izgovorene naredbe. Natječaj se otvorio 15.11.2017.godine, a zatvorio 17.01.2018.

Skup podataka koji koristimo su datoteke također sa stranice Kaggle:

train- Sadrži nekoliko informativnih datoteka i mapu audio datoteka.

Naredbe koje ćemo predvidjeti u testu su *yes, no, up, down, left, right, on, off, stop, go.* Sve ostalo treba smatrati *silence* ili *unknown*. Datoteke su imenovane tako da je prvi element subjekt id osobe koja je dala glasovnu naredbu.

Ponavljaju se naredbe kada subjekt ponavlja istu riječ više puta.

Mogu se očekivati neke nedosljednosti u svojstvima podataka train-a. (npr. duljina zvuka)

test – Sadrži audio mapu s 150.000 datoteka.

Cilj i hipoteze istraživanja problema

Cilj istraživanja je korištenjem metoda strojnog učenja, na temelju training seta, prepoznati navedene glasovne naredbe i pretvoriti ih u tekst, primjenjujući algoritme i metode opisane u idućim točkama. Obrađene podatke treba usporediti s dosadašnjim. Poboljšanjem točnosti prepoznavanja alata za glasovno sučelje, može se poboljšati učinkovitost proizvoda i njihova dostupnost.

Pregled dosadašnjih istraživanja

Problem koji rješavamo je natječaj objavljen na web stranici www.kaggle.com. U natjecanju je sudjelovalo 1315 timova. (Kaggle, 2018.)

Najuspješniji timovi su za obradu audio datoteka (odabir i vizualizaciju značajki audio signala za prepoznavanje govornih obilježja) iz train seta koristili razne spektograme. Spektogram je vizualna prezentacija spektra frekvencija zvuka ili signala. Za analizu spektograma korišten je MFCC (Mel Frequency Cepstral Coefficient), VAD (Voice Activity Detection) za izbacivanje tišine te FFT (Fast Fourier Transform) za smanjivanje dimenzionalnosti tako da se za svaku snimku računa FFT te se analizira Gaussovom metodom. (DavidS, 2018; Ostyakov, 2018; CherKeng, 2018)

Rješenja su bazirana na neuronskim mrežama: Convolutional Neural Network (CNN) te Recurrent Neural Network (RNN). Eksperimentiralo se s različim brojem slojeva i filtera. (Ostyakov, 2018; CherKeng, 2018)

Najbolji ostvareni rezultati imaju točnost od oko 91% u kategorizaciji audio datoteka s ispravnom oznakom (Kaggle, 2018).

Materijali, metodologija i plan istraživanja

Na koji način ćemo pokušati riješiti problem?

Rekurentna neuronska mreža (RNN) je mreža koja se može koristiti kada naše podatke možemo tretirati kao neki slijed. Ulaz u RNN u svakom koraku je trenutna vrijednost kao i vektor stanja koji pamti što je RNN vidjela u prijašnjim koracima (Lipton i sur., 2015). Ulazni podatak biti će sekundu dug audio isječak, i za svaki će se probati prepoznati riječ koja odgovara onome što je izgovoreno. Gradit ćemo rekurentnu neuronsku mrežu pomoću TFLearn – Tensorflow biblioteke i trenirat ćemo je na *train dataset* , a zatim testirati na *test dataset* . Za audio isječak u testnom setu mora se predvidjeti točna oznaka. Datoteka u kojoj to čuvamo će imati otprilike sljedeći format:

filename, label

clip1.waw, silence

clip2.waw, left

Kako ćemo prikupiti podatke?

Podatke ćemo prikupiti sa Kaggle stranice natjecanja.

Ovo je skup sekunde dugih .wav audio datoteka, od kojih svaka sadrži izgovorenu jednu englesku riječ. Ove su riječi iz malog skupa zapovijedi, izgovoreni od strane različitih govornika. Audio datoteke su organizirane u mape na temelju riječi koju sadrže. Audio datoteke prikupljene su pomoću *crowdsourcinga*. Snimljeno je dvadeset glavnih riječi, pri čemu ih je većina govornika izgovorila pet puta. Glavne riječi su: *Yes, No, Up, Down, Left, Right, On, Off, Stop, Go, Zero, One, Two, Three, Four, Five, Six, Seven, Eight, Nine*. Da bi se razlikovale riječi koje nisu prepoznate tu je i deset pomoćnih riječi koje su govornici izgovorili jednom: *Bed, Bird, Cat, Dog, Happy, House, Marvin, Sheila, Tree, and Wow.* (Kaggle, 2018)

Što ćemo koristiti?

TensorFlow- softverska biblioteka otvorenog koda, koristi se i za aplikacije strojnog učenja kao što su neuronske mreže.

TFLearn – Python modul za aplikacije strojnog učenja unutar TensorFlow-a

Python 3.6

Kako mislimo ocijeniti uspješnost rezultata svog projektai?

Radit ćemo matricu konfuzije; točnost (*accurancy*) će zapravo biti omjer točno prepoznatih glasovnih naredbi u odnosu na ukupan broj u testnom skupu primjera (Confusion matrix, n.d.).

Očekivani rezultati

Očekujemo da će naš algoritam moći točno prepoznati odnosno dodijeliti točnu oznaku: *yes*, *no*, *up*, *down*, *left*, *right*, *on*, *off*, *stop*, *go*, *silence ili unknown* za 80% do 90% audio dateoteka (.wav) iz testnog seta.

Literatura

CherKeng, H. (2018) My Tricks and Solution) [Online post]. Objavljeno na https://www.kaggle.com/c/tensorflow-speech-recognition-challenge/discussion/46945

DavidS (2018) Speech representation and data exploration [Online Kernel]. Objavljeno na https://www.kaggle.com/davids1992/speech-representation-and-data-exploration

Confusion matrix (n.d.). In Wikipedia. https://en.wikipedia.org/wiki/Confusion_matrix. Pristupljeno 30.4.2018.

Kaggle (2018) TensorFlow Speech Recognition Challenge https://www.kaggle.com/c/tensorflow-speech-recognition-challenge/data. Pristupljeno 30.4.2018.

Lipton, Z. C., Berkowitz, J., Elkan C. (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning. *eprint arXiv*:1506.00019, https://arxiv.org/pdf/1506.00019.pdf. Pristupljeno 30.4.2018.

Ostyakov, P. (2018) Our approach (4th place) [Online post]. Objavljeno na https://www.kaggle.com/c/tensorflow-speech-recognition-challenge/discussion/47674