Klasifikacija audio snimaka prema jeziku govora pomoću konvolucione neuronske mreže

Jovan Najdovski SV30/2020

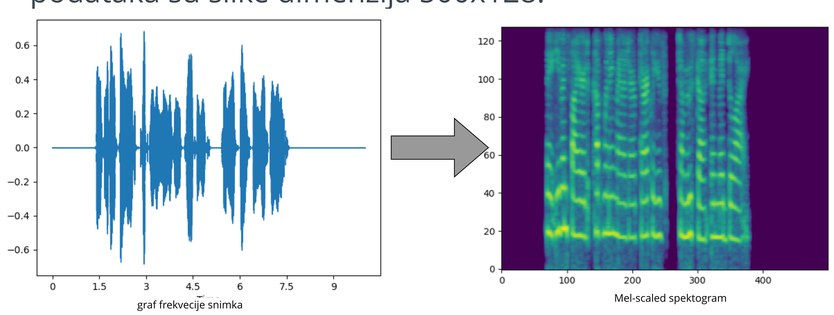
UVOD

Cilj projekta je izvršiti klasifikaciju audio snimaka na osnovu jezika govora. Klase koje postoje su srpski, engleski i špasnki jezik. Analizira se spektogram kratkih audio snimaka. Koristićemo sledeće korake:

1. prikupljanje i priprema skupa podataka
2. augmentacije podataka
3. izdvajanje Mel-scaled spektograma (preprocessing)
4. treniranje modela
5. evaluacija modela

SKUP PODATAKA

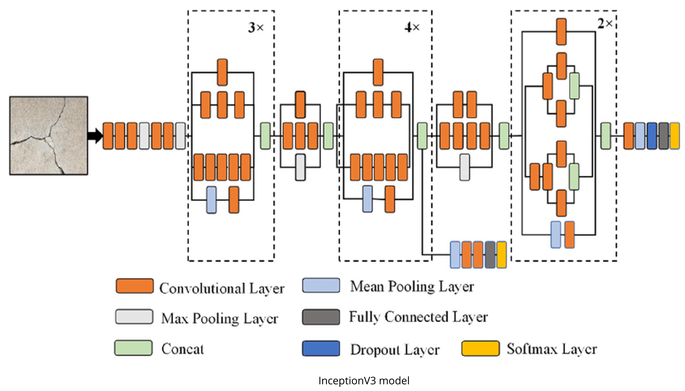
Koristiće se dataset audio snimaka sa Mozilla Common Voice za srpski, engleski i španski jezik. Izdvojiće se validirani audio snimci dužina od 7.5 do 10 sekundi. Za srpski jezik zbog nedostatka podataka korsite se i isečci iz audio knjiga. Koristimo 10000 audio snimaka po jeziku. Augmentaciju podataka vršimo dodavanjem belog šuma dobijenog normalnom raspodelom, čime dobijamo bolju robusnost modela. Ulazni skup podataka će se podeiliti na trening (70%), validacioni (10%) i test (20%) skup. Spektogrami dobijeni preprocesiranjem podataka su slike dimenzija 500x128.



KORIŠĆENE METODOLOGIJE

Rešenje je implemntirano u python progrmaskom jeziku uz pomoć biblioteka tensorflow, keras, numpy, librosa, matplotlib, soundfile, pydub. Svi audio snimci su prethodno obradjeni da bi se izdvojio Mel-scaled spektogram (smanjenje na 8 kHz, produžavanje do 10s, izdvajanje spektograma pomoću librose i normalizacija). Za klasifikaciju su korišćena dva CNN tipa modela:

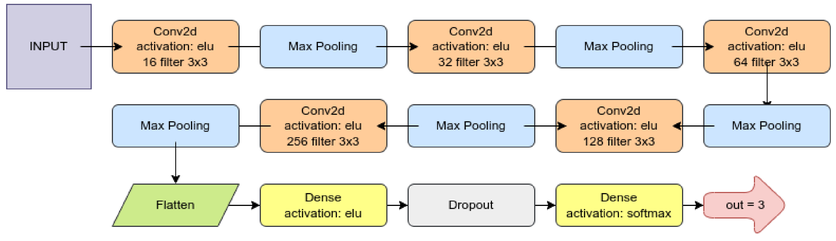
- kompleksni InceptionV3 (iz biblioteke keras)
- 5x-Conv-MaxPool (redjanjem slojeva)



InceptionV3 model

TRENIRANJE MODELA

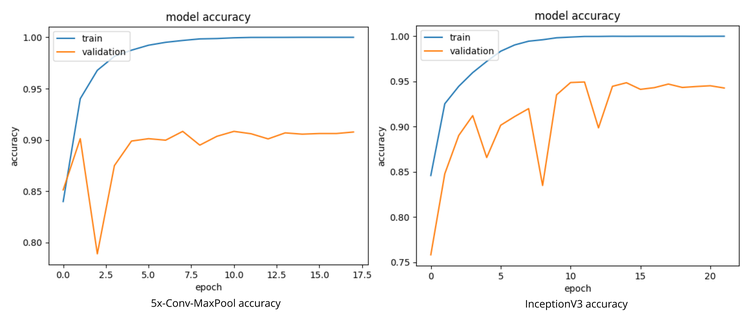
- Za treniranje InceptionV3 korišćen je **batch size 16**, dok je za 5x-Conv-MaxPool korišćen **batch size 32.**
- Korišćena loss funkcija je **categorical crossentropy**.
- Kod 5x-Conv-MaxPool za skrivene slojeve korišćena je **elu** aktivaciona funkcija, dok je za poslednji sloj korišćena **softmax.**
- Korišćena metrika je preciznost (**accuracy**).
- Modeli su trenirani sa optimizer-ima algoritama RMSprop i Nadam, ali se **Nadam** dosta bolje pokazao.
- Modeli su trenirani u **60 epoha**, ali je korišćen **early stopping** kako bi se izbegao veći overfitting.



5x-Conv-MaxPool model

REZULTATI

5x-Conv-MaxPool dostiže preciznost od 90.8%, dok InceptionV3 dostiže 94.9% na validacionom skupu pri treniranju. Tačnost na skupu za testiranje je **92.4%** za 5x-Conv-MaxPool i **96.3%** za InceptionV3.



5x-Conv-MaxPool accuracy

InceptionV3 accuracy

ZAKLJUČAK

Postignuta je dovoljno velika preciznost za ovaj slučaj korišćenja. InceptionV3 model se pokazao bolje nad skupom validacionih podataka. Medjutim razlog može biti veličina InceptionV3 modela koja može izazvati overfitting. Modeli se mogu poboljšati treniranjem nad većim skupom podataka, kao i korišćenjem boljih tehnika augmentacije.