Klasifikacija Urbanih Zvukova

Branko Grbić, Željko Milovanović
Matematički Fakultet, Univerzitet u Beogradu

Motivacija

- Klasifikacija urbanih zvukova kao problem budućnosti
- Razlika između pretreniranog CNN-a i naših modela za klasifikaciju audio uzoraka
  - CNN
  - FFNN
- Zašto su pretrenirani modeli dominantniji u industriji
- Iskustvo u PyTorch-u
  - Korišćenje torch klasi za rad
  - Skalabilan softver za klasifikaciju zvukova

Skup Podataka

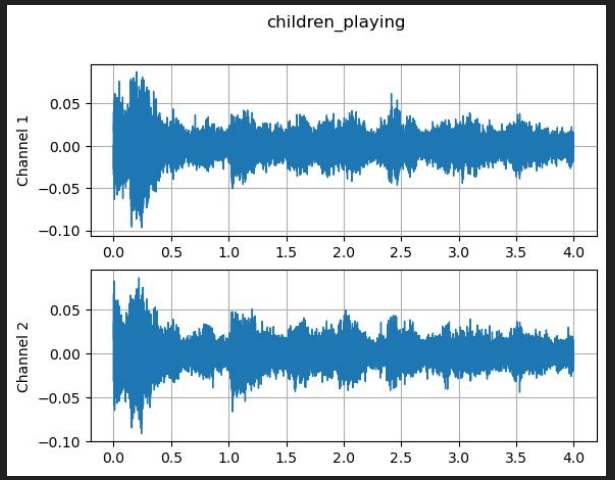
- UrbanSounds8K je skup audio podataka koji sadrži 8732 označenih uzoraka (<=4s)
- 10 klasa urbanih zvukova:
  - Klima, Truba automobila, Graja dece, Lajanje kucica, Busenje, Rad motora, Pucanj pištolja, Mehanički Čekić, sirena i ulična muzika
- Klase su skoro ravnomerno raspoređene

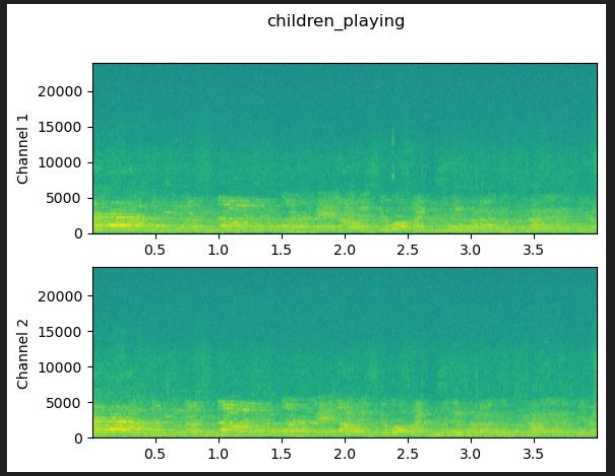
	index	jackhammer	dog_bark	children_playing	street_music	air_conditioner	drilling	engine_idling	siren	car_horn	gun_shot
0	fold1	120	100	100	100	100	100	96	86	36	35
1	fold2	120	100	100	100	100	100	100	91	42	35
2	fold3	120	100	100	100	100	100	107	119	43	36
3	fold4	120	100	100	100	100	100	107	166	59	38
4	fold5	120	100	100	100	100	100	107	71	98	40
5	fold6	68	100	100	100	100	100	107	74	28	46
6	fold7	76	100	100	100	100	100	106	77	28	51
7	fold8	78	100	100	100	100	100	88	80	30	30
8	fold9	82	100	100	100	100	100	89	82	32	31
9	fold10	96	100	100	100	100	100	93	83	33	32

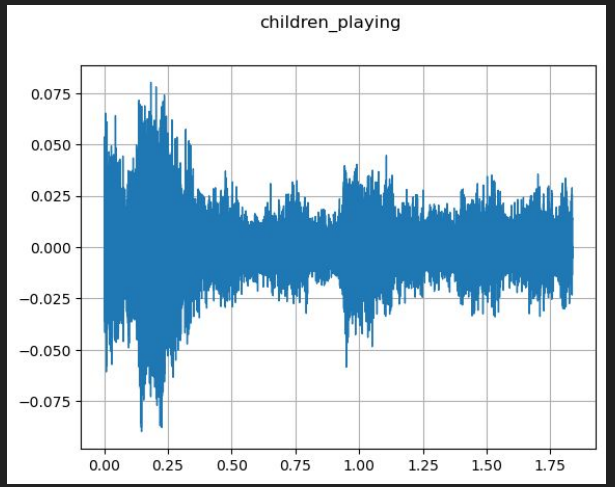
```
dog_bark          0.114521
children_playing  0.114521
air_conditioner   0.114521
street_music      0.114521
engine_idling     0.114521
jackhammer        0.114521
drilling          0.114521
siren             0.106390
car_horn          0.049130
gun_shot          0.042831
```

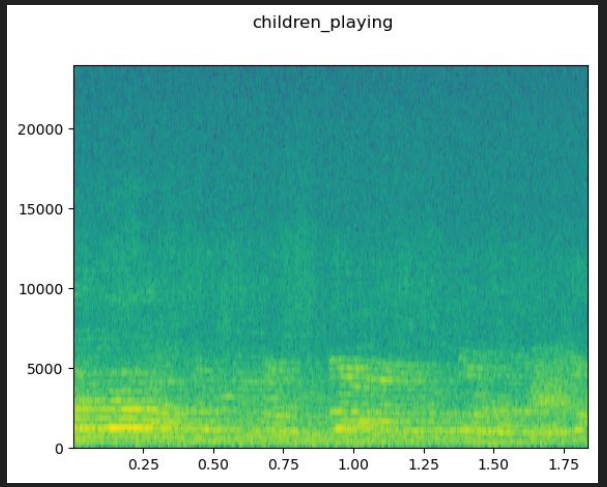
Preprocesiranje podataka

- Ponovo uzorkovanje na 44.1 kHz (resampling)
- Konvertovanje zvuka u mono
- Skraćivanje uzorka na 2s
- Dopunjavanje sa desne strane
- Izvlačenje atributa
- Konvertovanje u tenzor i prosleđivanje DataLoader klasi
- 9 foldova za trening, 1 za test
    - Moramo napraviti 10 modela i usrednjiti tačnost po foldu za test


children_playing
Channel 1
Channel 2
0.05
0.00
−0.05
−0.10
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

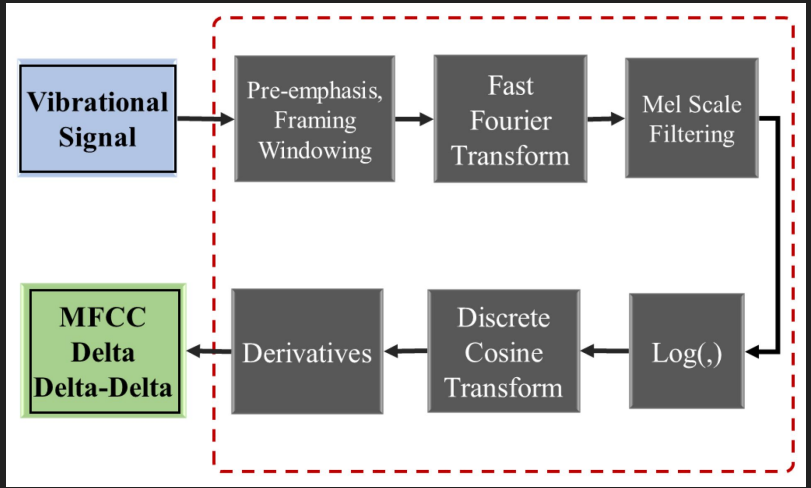

children_playing
Channel 1
Channel 2
20000
15000
10000
5000
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


children_playing
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


children_playing
20000
15000
10000
5000
0
0.25
0.50
0.75
1.00
1.25
1.50
1.75

Atributi

- FFNN
  - RMS
  - Spektralni centroid
  - Spektralna širina opsega (spectral bandwidth)
  - Spektralno opadanje (spectral falloff)
  - Stopa nultih prelaza (zero crossing rate)
  - MFCC (usrednjen)
- CNN i VGG
  - MFCC



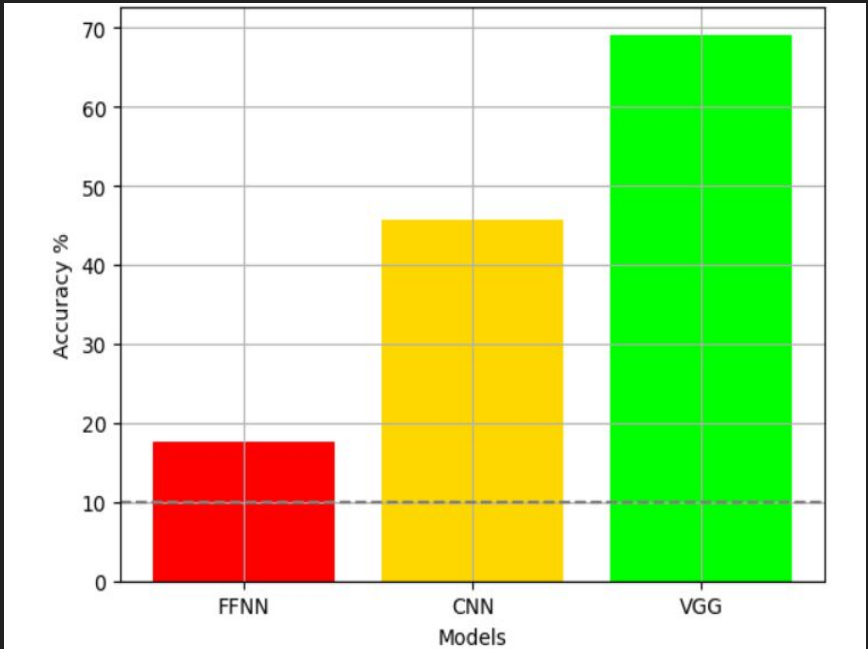
Modeli

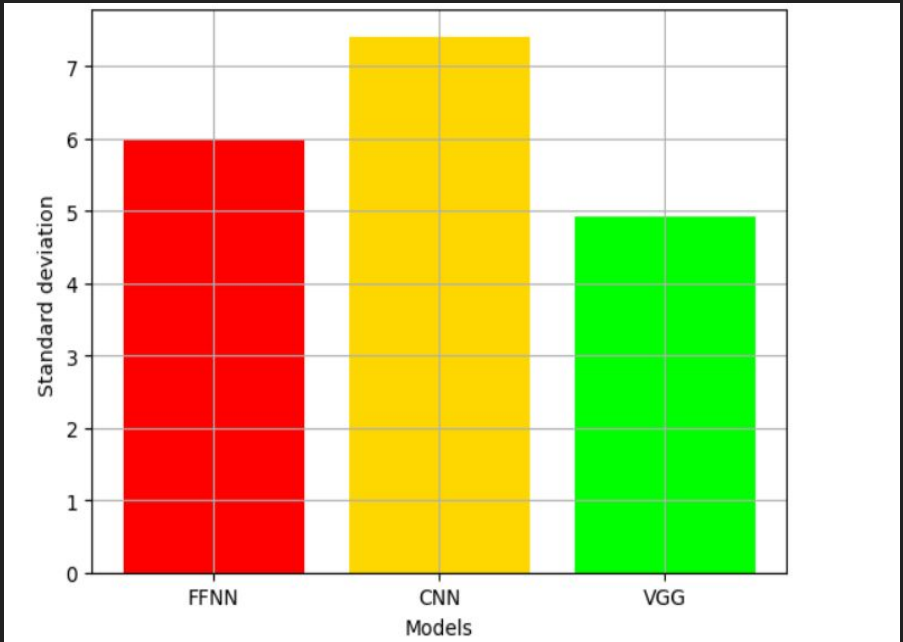
- Potpuno povezana neuronska mreža (FFNN)
  - 2 skrivena sloja veličine 25 i 20
- Konvolucijska neuronska mreža (CNN)
  - Struktuirana slično kao VGG - sa manje konvolucionih slojeva
  - 6 konvoluciona sloja na koje se nadovezuju 2 skrivena sloja
- Pre-trenirani VGG-11 sa batch normalizacijom (treniran na ImageNet-u)

Trening

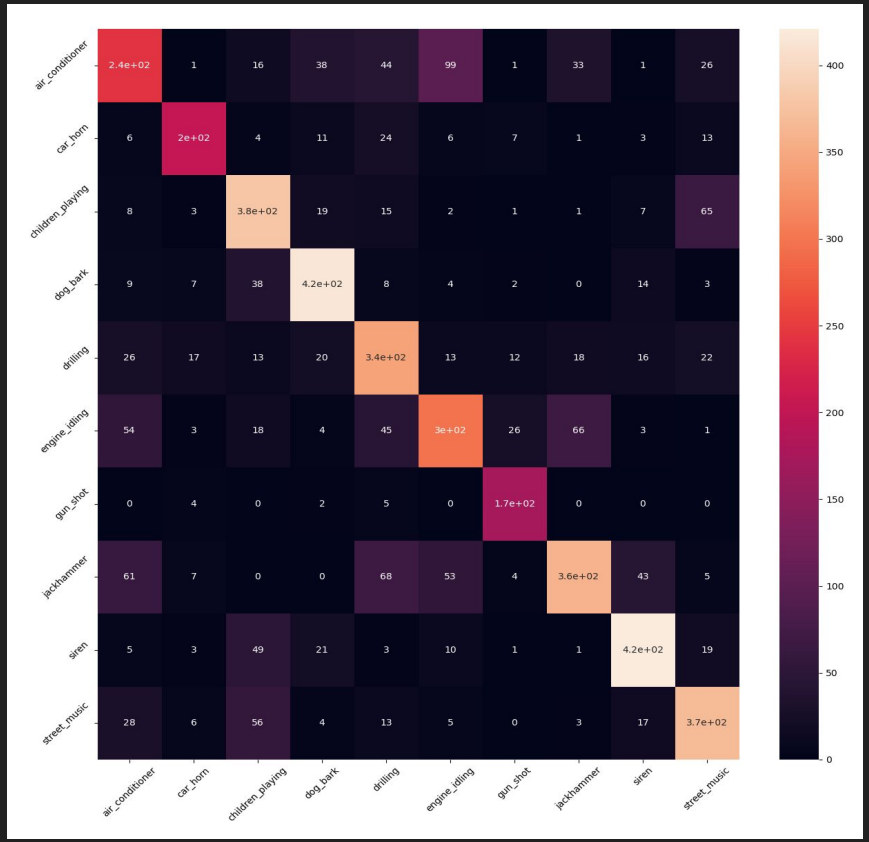
- Funkcija gubitka - Kategorička unakrsna entropija
- Optimizator - Adam
- Stopa učenja - 1e-4
- Korak stope učenja - 5 (gamma 0.1)
- Batch size - 64
- Broj epoha
  - 9 - VGG, CNN
  - 13 - FFNN

Rezultati

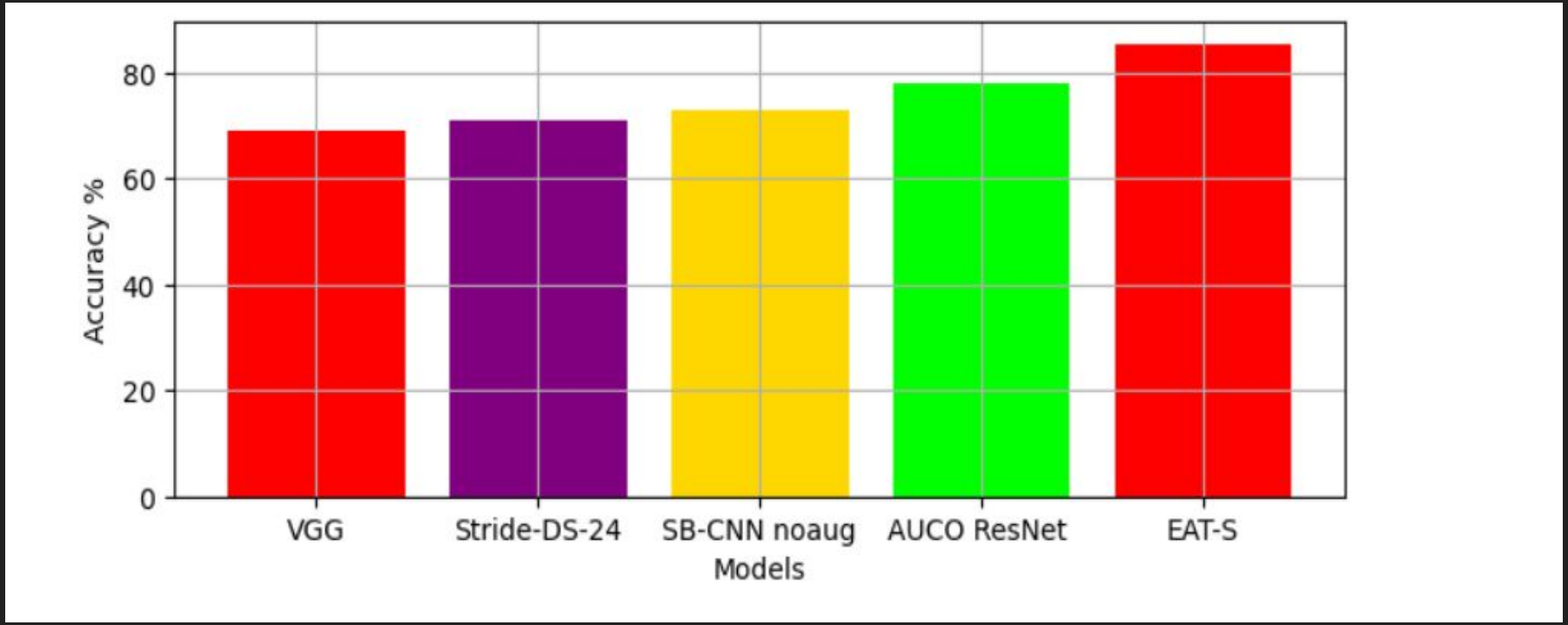

70
60
50
40
30
20
10
0
Accuracy %
FFNN
CNN
VGG
Models

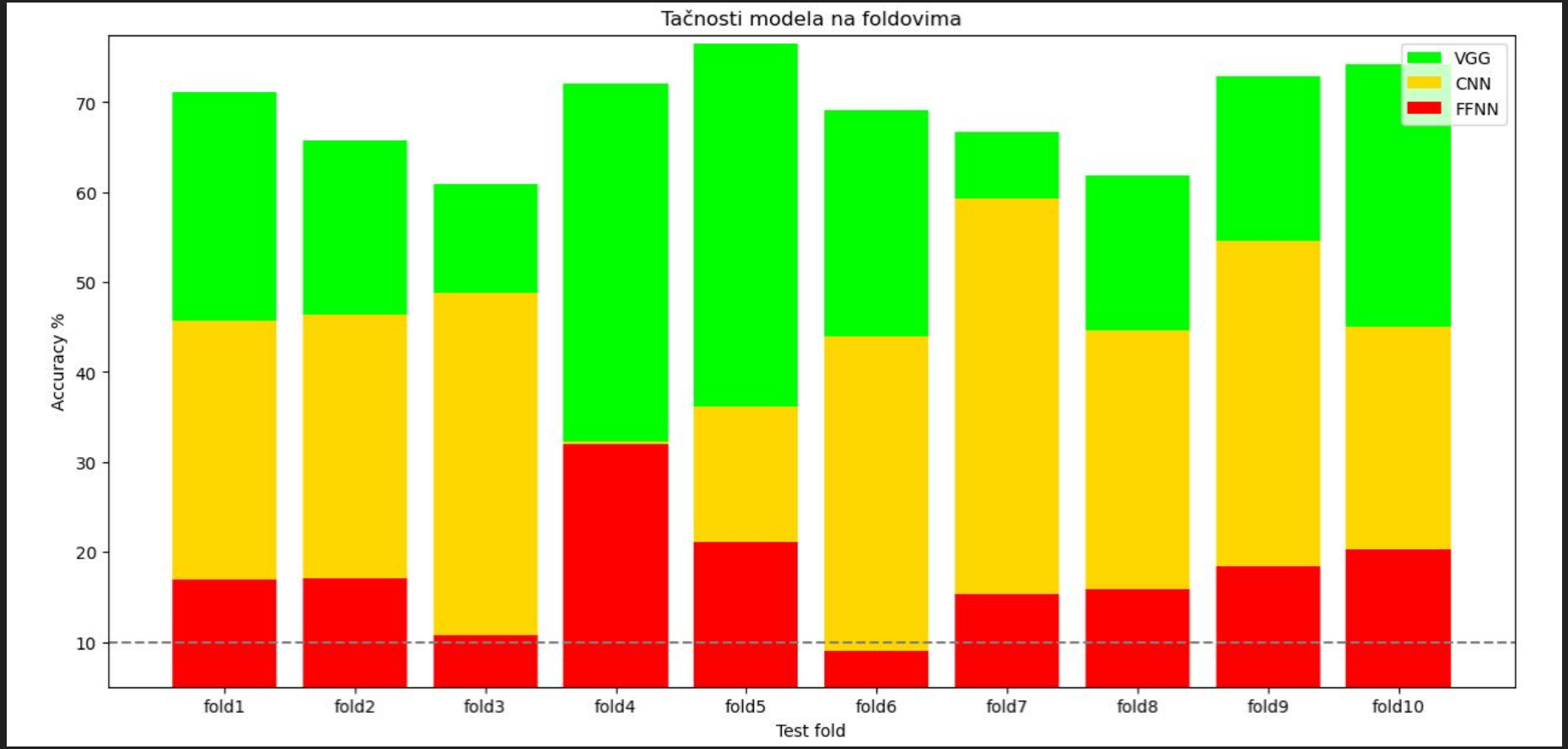

7
6
5
4
3
2
1
0
Standard deviation
FFNN
CNN
VGG
Models

Matrica konfuzije za jedan fold VGG-a



Poređenje sa dostupnim modelima




Tačnosti modela na foldovima
VGG
CNN
FFNN
Accuracy %
70
60
50
40
30
20
10
fold1
fold2
fold3
fold4
fold5
fold6
fold7
fold8
fold9
fold10
Test fold

Diskusija

- Sva 3 modela su bolja od nasumičnog nagađanja - naš model uči
- FFNN očekivano najgori - preveliko pojednostavljenje uzorka
- CNN u rangu sa netreniranim VGG-om
- Pretrenirani VGG je ubedljivo najbolji model - 69.14%
    - Isproban je i netrenirani VGG - lošiji rezultati
- Pobednik: VGG
    - Brži trening - bolja tačnost

Dodaci

- Custom test opcija
- Čuvanje modela
- Pokazivanje i čuvanje rezultata (tačnost, loss itd)
- Argumenti komandne linije radi lakšeg pokretanja
- Hiperparametrizovan kod

Hvala na pažnji

[1] Very Deep Convolutional Networks for Large-Scale Image Recognition, Karen Simonyan, Andrew Zisserman

[2] Neural network based recognition of speech using MFCC features, Pialy Barua,Kanij Ahmad, Ainul Anam Shahjamal Khan, Muhammad Sanaullah (2014)

[3] Justin Salamon, Christopher Jacoby and Juan Pablo Bello, Music and Audio Research Laboratory (MARL), New York University, Center for Urban Science and Progress (CUSP), New York University

[4] Douglas O'Shaughnessy (1987). Speech communication: human and machine. Addison-Wesley. str. 150.

[5] Adam: A Method for Stochastic Optimization, Diederik P. Kingma, Jimmy Ba

[6] End-to-End Audio Strikes Back: Boosting Augmentations Towards An Efficient Audio Classification Network, Avi Gazneli, Gadi Zimerman, Tal Ridnik, Gilad Sharir, Asaf Noy

[7] AUCO ResNet: an end-to-end network for Covid-19 pre-screening from cough and breath, Vincenzo Dentamaro, Paolo Giglio, Donato Impedovo, Luigi Moretti, Giuseppe Pirlo

[8] Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification, Justin Salamon, Juan Pablo Bello

[9] Environmental Sound Classification on Microcontrollers using Convolutional Neural Networks, Jon Nordby

[10] Min Xu; et al. (2004). "HMM-based audio keyword generation"

[11] J. Salamon, C. Jacoby and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research", 22nd ACM International Conference on Multimedia, Orlando USA, Nov. 2014.