Konstrukcija konvolutivne neuronske mreže za klasifikaciju radiografskih snimaka grudnog koša

Seminarski rad D -
Napredni modeli neuronskih mreža
Univerzitet u Novom Sadu
Prirodno-matematički fakultet

Nikola Pujaz - 21m/19

3. februar 2021.

Sažetak

Rastuća popularnost neuronskih mreža otvorila je nove mogućnosti rešavanja različitih nelinearnih problema. Razumevanje sadržaja slike i njihova klasifikacija pomoću neuronskih mreža omogućavaju automatizovani proces visoke preciznosti koji može biti veoma koristan u medicinskoj dijagnostici različitih vrsta snimaka. U ovom radu prikazan je proces konstruisanja konvolutivne neuronske mreže za klasifikaciju radiografskih snimaka grudnog koša, kao i izazovi koje proces konstrukcije nosi sa sobom.

Sadržaj

1 Uvod

Upotreba neuronskih mreža, pre svega njihovih naprednih arhitektura u medicini i farmaciji je danas široko rasprostranjena, a od skoro i opšte prihvaćena tehnika koja se ne koristi samo u dijagnostici sa svrhom poboljšavanja i preciziranja različitih oboljenja pacijenata, već i za razumevanje uzročnika oboljenja i određivanje blagovremenog odgovora i adekvatne terapije.

U vremenu aktuelne pandemije *SARS-Cov-2* virusa kompleksna klinička slika *COVID19* oboljenja, prouzrokovanog prethodno pomenutim virusom, zahteva različite dijagnostičke metode kako bi se uspešno utvrdio stadijum bolesti i odgovarajuća terapija.

Radiografski snimci, kao i drugi tipovi snimaka, čine jednu od jednostavnih, neinvazivnih dijagnostičkih metoda koje služe za utvrđivanje stanja i promena unutrašnjih organa. Analizom oblika, veličine i lokacije "*senke*"na radiografskom snimku utvrđuje se sumnja na određenu dijagnozu koja se kasnije utvrđuje drugim, često invazivnim, dijagnostičkim metodama, čineći ovu metodu, u kombinaciji sa ostalim metodama dijagnostike, značajnim *pred-korakom* daljem utvrđivanju bolesti.

Uzimajući prethodno navedeno u obzir, precizna i tačna dijagnoza neinvazivnim metodama je od viskokog značaja i zahteva visok nivo sigurnosti u situacijama u kojima ova metoda služi za odlučivanje o daljem toku dijagnostike i lečenja, pogotovo u situacijama u kojima naredni korak uključije invazivne operativne zahvate. Međutim, metode koje podrazumevaju analizu snimaka u potpunosti zavise od kompetencija i iskustva radiologa koji ih analizira i greške u određivanju dijagnoze samo na osnovu jedne dijagnostičke metode nisu nepoznanica i predstavljaju visok rizik. Zbog toga se često u praksi vrši nekoliko različitih dijagnostičih metoda kako bi se upotpunila klinička slika i definisala jasna dijagnoza bolesti.

Koristeći tehnike mašinskog učenja, još preciznije rečeno - dubokog učenja (eng. *deep learning*), može se postići veća preciznost i tačnost pretpostavljene dijagnoze nego što je to slučaj kada dijagnozu uspostavlja radiolog.

U skladu sa naslovom ovog rada, u narednim sekcijama će biti prikazan proces konstrukcije konvolutivne neuronske mreže za klasifikaciju radiografskih snimaka grudnog koša, koja može da posluži kao bazični prikaz fundamentalnih koraka u razvoju ozbiljnijeg modela za uspešnu dijagnostičku podršku određivanja poremećaja i bolesti pluća, srca i drugih mekih tkiva unutar regije grudnog koša. Dodatno, kao posledica preduzetih akcija, zabeležena je kompleksna realizacija ovog procesa koja često podrazumeva i neuspešne modele. U narednoj sekciji je opisan skup podataka i korišćeni alati u razvoju.

2 Analiza i pretprocesiranje podataka

Skup podataka koji je pružila grupacija američkih Nacionalnih instituta za zdravlje (eng. National Institutes of Health - NIH), a koji je opisan u pratećem radu [5], sadrži 112 120 radiografskih snimaka 1024x1024 rezolucije, od 30 805 pacijenata, podeljenih u sledećih 14 klasa prema tipu bolesti: *1.Atelektaza; 2.Kardiomegalija; 3.Izliv; 4.Infiltracija; 5.Masa; 6.Čvor; 7.Upala pluća; 8.Pneumotoraks; 9.Konsolidacija; 10.Edem; 11.Emfizem; 12.Fibroza; 13.Zadebljanje pleure; 14.Kila*, dok posebnu klasu čine snimci na kojima nije pronađena nikakva patološka promena. Uz skup snimaka pridružen je i skup meta-podataka koji sadrži informacije o godinama

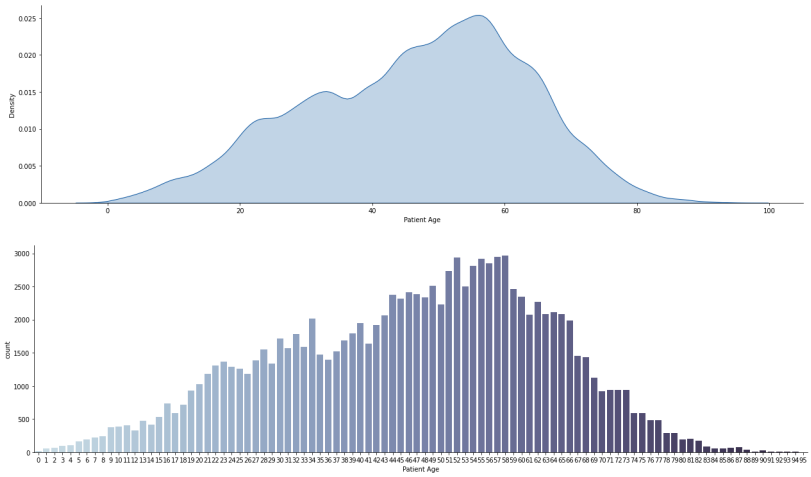
starosti pacijenata, polu pacijenata, regiji na kojoj je uočena "*senka*", i drugo. Neophodno je napomenuti da postoje ograničenja u čistoći podataka jer su algoritmi obrade prirodnih jezika korišćeni za određivanje dijagnoze pridružene radiografskim snimcima na osnovu evidencije bolesti, pa je procenjeno da je tačnost klasifikacije na osnovu obrade prirodnog jezika preko 90%.

Za realizaciju ovog rada korišćeno je *Google Colab* okruženje, biblioteke *Keras* i *Tensorflow*, programski jezik *Python* i njegove konstrukcije i druge biblioteke za analizu podataka i njihovu vizualizaciju. Implementacija rešenja se nalazi u *Jupyter Notebook* datotekama u prilogu ovog rada.

Kako bi skup podataka radioloških snimaka imao kontekst i dublji značaj, prethodno je izvršena analiza meta-podataka prikazana u narednoj sekciji.

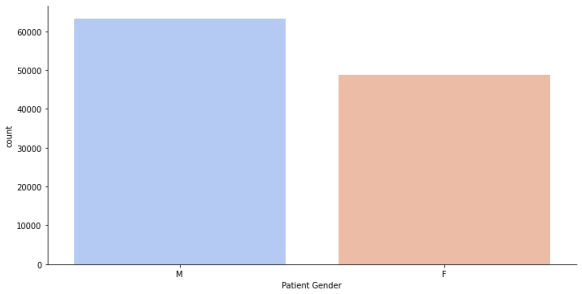
2.1 Analiza skupa meta-podataka

Na slici 1 može se primetiti da je najveći broj pacijenata, čiji su radiografski snimci u skupu podataka, u rasponu od 45 do 60 godina starosti. Prosek godina pacijenta je 46.63 godine, međutim, ovaj podatak je diskutabilan uzimajući u obzir da minimalan broj godina iznosi 0, što može da bude posledica korišćenja algoritama za obradu prirodnog jezika pri kreiranju skupa meta-podataka.



Slika 1: Pregled distribucije i broja pacijenata po godinama starosti

U odnosu na pripadajući pol izdvojeno je 63 340 muškarca i 48 780 žena (slika 2), čija je odgovarajuća distribucija u odnosu na godine starosti prikazana na slici 3.



Slika 2: Pregled broja pacijenata po polu

Slika 3: Pregled distribucije pacijenata po polu i godinama starosti

Veliki izazov prilikom pretprocesiranja podataka predstavlja sama činjenica da je 112 120 radiografskih snimaka preuzeto iz zdravstvenih dosijea 30 805 pacijenata, što može da implicira tome da su ovi pacijenti snimanje ponavljali više puta.

Manuelnim pregledom određenih instanci skupa meta-podataka utvrđeno je da se pojedine dijagnoze razlikuju za iste pacijente što može biti posledica korišćenja algoritama obrade prirodnih jezika sa svrhom sakupljanja informacija skladištenih u skupu meta-podataka.

Uz prethodno navedeno, utvrđeno je da 20 796 pacijenata ima više dijagnoza pridruženih istom snimku. Kako potpuna čistoća podataka nije zagarantovana, nekoliko mogućih scenarija pripreme skupa podataka se stavlja na razmatranje.

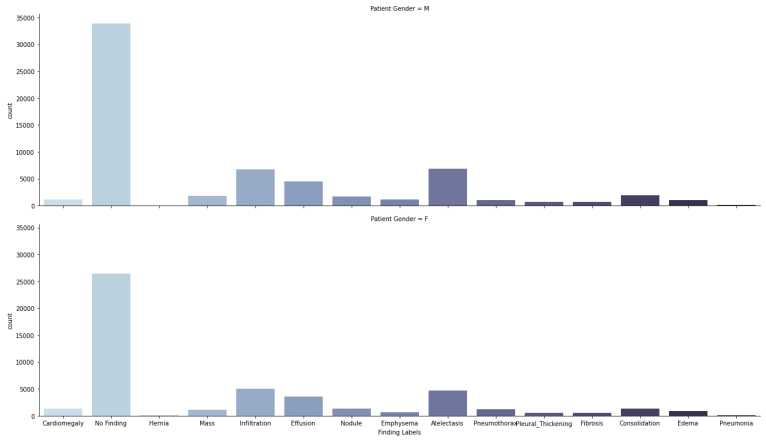
Prvi scenario podrazumeva proširivanje skupa meta-podataka tako da određene instance skupa budu duplirane pri čemu se podatak o pridruženoj dijagnozi razlikuje za svaku instancu. Na ovaj način, skup meta-podataka, a samim tim i skup podataka, povećava se na 141 537 instanci – radioloških snimaka.

Međutim, ovaj scenario implicitno nosi opasnost od preprilagođavanja (eng. *overfitting*) jer se iste ili slične fotografije mogu naći i u trening-validacionom i u test skupu. Dodatni izazov ovom pristupu predstavlja činjenica da su određeni pacijenti snimanje izvršili preko 150 puta, zbog čega se postavlja pitanje verodostojnosti podataka u slučaju ovih pacijenata, što ide u prilog prethodno spomenutom problemu čistoće zbog korišćenja algoritama obrade prirodnih jezika. Rešenje koje se intuitivno nameće u ovom slučaju jeste da se problem klasifikacije proširi na taj način da umesto jedne klase bude dozvoljeno da određeni snimak ima više klasa (eng. *multi-label classification*). Ipak, zbog razloga objašnjenih u nastavku, drugi scenariji su uzeti u razmatranje.

Drugi pristup podrazumeva opredeljenje za samo jednu dijagnozu iz liste dijagnoza pridruženih radiografskom snimku pri čemu se zadržava orginalni broj snimaka u skupu podataka. Iako je u odvojenom skupu pružena informacija o okvirima koji obuhvataju „senku“ koja upućuje na određenu dijagnozu, autor ovog rada nije domenski ekspert, stoga dalje razumevanje i prioritizacija jedne od dijagnoza iz liste dijagnoza nije bila moguća, pa je za pridruženu dijagnozu uzeta prva dijagnoza iz liste.

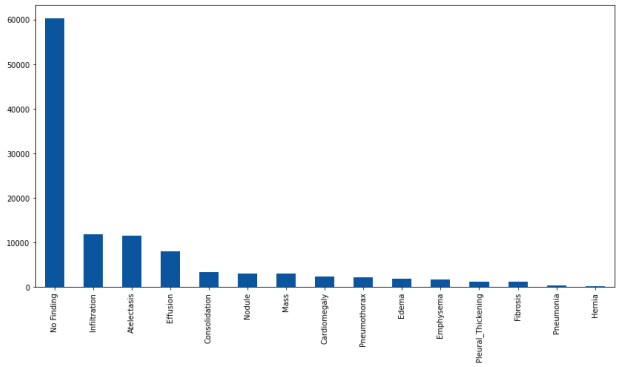
Na ovaj način mogu se iskoristiti dodatna dva skupa meta-podataka koja dolaze u paketu sa celokupnim skupom koja izdvajaju podatke (snimke) za trening-validacioni i test skup, pritom pazeći na to da se snimci jednog pacijenta nalaze samo u jednom od ovih skupova. U ovom koraku izdvojeni su i atributi i podaci koji su od važnosti za naredni korak pretprocesiranja skupa podataka (*Image Index* i *Finding Labels* preimenovani u *image_path* i *labels*, u navedenom redosledu).

Na slici 4 prikazan je broj pacijenata po pripadajućim klasama dijagnoze, podeljeni prema polu.



Slika 4: Pregled pacijenata prema pripadajućim klasama

Sa prethodne slike se moglo naslutiti da najveći broj snimaka je onaj na kojima nije pronađena nijedna od ponuđenih patologija, što se vidi i na slici 5 koja prikazuje broj snimaka po pripadajućim klasama. Već u ovom trenutku se može pretpostaviti da se u ovom slučaju javlja problem neuravnoteženih klasa koji će biti adresiran u sekciji 2.3.



Slika 5: Pregled broja snimaka prema pripadajućim klasama

2.2 Pretprocesiranje skupa podataka

Kako skup podataka zauzima preko 42GB prostora, pojavio se problem učitavanja skupa u *Google Colab* okruženje. Ovo okruženje nudi nekoliko različitih načina za preuzimanje i učitavanje podataka. Prva opcija je podrazumevala učitavanje uz pomoć *Google Drive* servisa, međutim, zbog ograničenog prostora od 15GB koje *Google Drive* pruža bez novčane naknade, od ove opcije se odmah odustalo. Drugi način za učitavanje podataka je *Kaggle* servis koji sadrži stariju verziju skupa podataka iz 2017. godine, dok je treći način direktno preuzimanje uz pomoć *TensorFlow* biblioteke i *Google Cloud* servisa koji korisnicima sa nalogom pružaju ovaj skup podataka u saradnji sa već pomenutom grupacijom američkih Nacionalnih instituta za zdravlje (u nastavku *NIH*). Ipak, izabrana je četvrta mogućnost - direktno preuzimanje skupa podataka iz delova u vidu *batch* paketa uz modifikaciju skripte koju pruža *NIH*. Modifikacija podrazumeva da se preuzeta datoteka otpakuje i obriše, što nije bilo moguće izvesti na drugi način, kako bi se uštedeo ograničeni slobodan memorijski prostor od oko približno 77GB koje *Google Colab* okruženje pruža. Modifikovana skripta je prikazana u nastavku, pri čemu su elementi liste linkova izuzeti u ovom prikazu.

In [1]:
```
import urllib.request
import tarfile
import os

links = [ ] # links go here

for idx, link in enumerate(links):
    fname = 'images_%02d.tar.gz' % (idx+1)
    print('Downloading: ' + fname + ' ...')
    urllib.request.urlretrieve(link, fname)
    tar = tarfile.open(fname, "r:gz")
    tar.extractall()
    tar.close()
    print('Extracted: ' + fname)
    os.remove(fname)
    print('Removed: ' + fname)
    print('----------------------------')

print('Download complete.')
```

Kako bi snimci bili pripremljeni po specifikacijama ulaznog sloja neuronske mreže opisane u sekciji 3, a ujedno i redukovani u veličini memorijskog prostora koji zauzimaju, pomoću biblioteke *Pillow* smanjeni su na 224x224 rezoluciju. Ovako modifikovani skup podataka je sačuvan u arhivu i skladišten na *Google Drive* servisu. Operacija smanjivanja veličine snimaka je jedina operacija koja će biti izvršena u fazi pretprocesiranja ovog skupa podataka. Dodatna augmentacija snimaka nije izvršena zbog razloga opisanih u sekciji 2.3.

2.3 Neuravnoteženost klasa

Iz sekcije 2.1 se može videti evidentna neuravnoteženost između klasa gde klasa *No Finding* sadrži najveći broj instanci. Postoji nekoliko tehnika balansiranja skupa podataka kako bi svaka od pripadajućih klasa imala jednak broj instanci, a samim tim i pružala jednak uticaj na proces učenja. Jedna od osnovnih tehnika podrazumeva augmentaciju klasa koje imaju manji broj instanci, pri čemu se generišu nove instance, u ovom slučaju snimci, od postojećih. Ova tehnika podrazumeva opsežno korišćenje afinih transformacija u ravni na veoma malom uzorku instanci, čime se na veštački način pokušava povećanje varijabilnosti podataka za određenu klasu. Druga tehnika se ogleda u obrnutom pristupu od onog koji je prethodno definisan, tako što se podskupovi skupa podataka koji pripadaju određenoj klasi umanjuju kako bi se postigla uspešna uravnoteženost. Ova tehnika sa sobom nosi opasnost od gubitka dragocenih podataka koji čine podatke u određenoj klasi visoko varijabilnim. Neke tehnike podrazumevaju dodeljivanje različitih težina klasama kako bi se uspostavila uspešna uravnoteženost. Spektar različitih tehnika za rešavanje problema neuravnoteženosti klasa se može naći u [6].

Pored prethodno navedenih ograničenja skupa podataka prateća dokumentacija ipak navodi da distribucija podataka teži približavanju podacima koji se mogu naći u realnom slučaju. Postoje uspešni primeri realizacije preciznih modela koji sadrže originalnu distribuciju podataka bez ikakvog balansiranja broja instanci po pripadajućim klasama, kao u [2], ali zbog samog ograničenja skupa podataka, ipak je odlučeno da se podese težine klasama kako bi se na ovaj način uspostavila uravnoteženost.

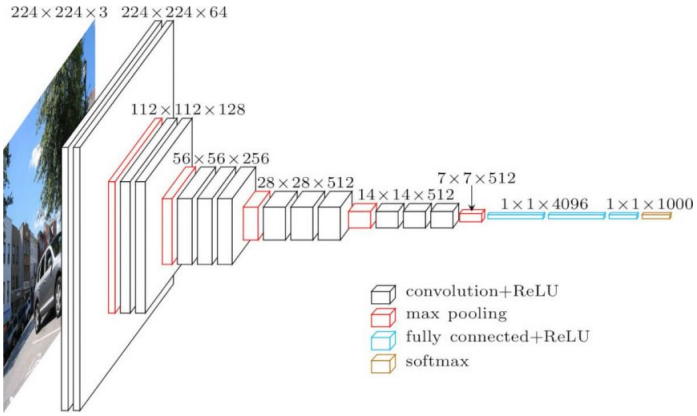
Korišćenjem klase *ImageDataGenerator* koju pruža *Keras API* na osnovu ranije podeljenih trening i test skupova kreirani su generatori koji će biti korišćeni u procesu treniranja validacije i evaluacije modela.

3 VGG-16 Arhitektura

Nakon nekoliko neuspešnih pokušaja kreiranja veoma jednostavne mreže koja daje zadovoljavajuće rezultate, gde je kao osnova za arhitekturu činila *LeNet-5* mreža koju je predložio *Yann LeCunn* u [3], prikazana u dodatku *Jupyter Notebook*-a, razvoj je otišao u smeru primene tehnika modifikacije već postojećih modela (eng. *Transfer Learning with Fine Tuning*).

Zbog svoje jednostavnosti i odličnih rezultata, kao osnova arhitekture neuronske mreže u ovom radu iskorišćena je VGG-16 arhitektura predložena u [4] i prikazana na slici 6. Generisani model ove arhitekture ima veoma visoku uspešnost (92.7% preciznosti) na *ImageNet* skupu podataka koji sadrži preko 14 miliona fotografija podeljenih u 1000 pripadajućih klasa. Originalna arhitektura je podeljena u 5 konvolutivnih blokova sa

ukupno 13 konvolutivnih slojeva, pri čemu nakon svakog sledi sloj maksimalnog udruživanja (eng. *max pooling*), nakon čega sledi blok maksimalne povezanosti (eng. *fully-connected layer*), pri čemu je izvršena modifikacija uz fino podešavanje ovih slojeva kako bi se arhitektura prilagodila problemu koji se rešava u ovom radu. Ono što dodatno odlikuje ovu arhitekturu su male dimenzije konvolutivnih filtera - 3 x 3.



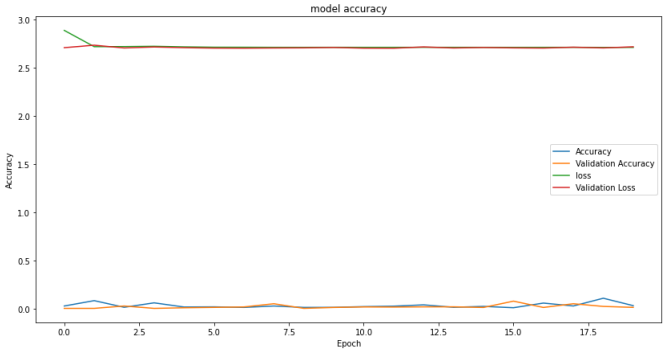
Slika 6: VGG-16 arhitektura (autor dijagrama: Davi Frossard[1])

Glavni nedostaci ove arhitekture uključuju veoma sporo treniranje modela i veliku memorijsku zahtevnost. Ipak, njena jednostavnost i visoko zadovoljavajući rezultati je čini dobrim osnovom za inicijalno razmišljanje o daljem razvoju ili eventualnim promenama u arhitekturi mreže.

U prvom krugu treniranja, svi slojevi mreže su bili *zamrznuti*, što znači da nije bilo moguće učiti ove slojeve, već njihove težine su preuzete iz gotovog modela datog u okviru *Keras API*. Drugi krug treniranja je obuhvatio fino podešavanje poslednjih slojeva mreže (4. i 5. konvolutivni blok uz izmenu sloja maksimalne povezanosti) uz promenu u stopi učenja u okviru Adam optimizatora.

4 Validacija i evaluacija modela

Na slici 7 je prikazana uspešnost istreniranog modela nakon 50 epoha na validacionom skupu, pri čemu su korišćene preciznost klasifikacije i mera gubitka kao mere evaluacije.



Slika 7: Rezultati treniranja modela

Tačnost evaluacije istreniranog modela na test skupu podataka iznosi 2.99%. Sa priloženog dijagrama se može primetiti da je preciznost modela veoma nezadovoljavajuća. Potencijalni koraci ka rešavanju problema loše preciznosti ovog modela diskutovana su u narednoj sekciji.

5 Diskusija

Iako rezultati nisu bili ni približno na zadovoljavajućem nivou, u ovom radu su uspešno prikazani koraci u razvoju jednog modela konvolutivne neuronske mreže sa realnim rezultatima na skupu podataka koji sa velikom verodostojnošću prikazuje podatke iz realnog slučaja. Na mnogim mestima u ovom radu predložene su različite tehnike daljeg poboljšanja ovog modela, kao i kreiranja modela koji daje značajno bolje rezultate. Kao početni korak u rešavanju problema uspešnosti modela mogao bi ponovo biti započet u koraku pretprocesiranja izborom druge tehnike rešavanja problema neuravnoteženosti klasa, npr. augmentacijom podataka. Takođe, neophodno je proveriti parametre klasa i metoda jer je zbog velikog broja podešavanja velika verovatnoća da se loši rezultati javljaju zbog pogrešno definisanih parametra.

Izazov treniranja modela u distribuiranom okruženju se ogleda u korišćenju TPU hardvera u *Google Colab* okruženju. Zbog korišćenja *ImageDataGenerator* klase prilikom pretprocesiranja podataka bilo je neophodno *obaviti* generatore *Tensorflow Dataset* klasom, međutim, pokušaj treniranja u ovom okruženju nije bio uspešan. Uvidom u sekciju sa problemima *Tensorflow Github* repozitorijuma, kao i nekoliko odgovora pronađenih na *StackOverflow* stranici, utvrđeno je da postoji aktuelni problem koji se javlja mnogim korisnicima i koji nije uspešno rešen do kraja realizacije ovog rada. Umesto TPU hardvera korišćen je GPU hardver uz pretpostavku boljih performansi od standardnog CPU hardvera koje *Google Colab* pruža.

Uobičajeni pristup razvijanja veoma duboke neuronske mreže, pogotovo u komercijalne svrhe, podrazumeva korišćenje već postojećih modela kao osnova za dalji razvoj i modifikaciju mreže. Ova tehnika često zahteva hardver visokih performansi za veće probleme i poznata je pod nazivom *Transfer Learning with Fine Tuning*.

Literatura

[1] Davi Frossard. VGG in TensorFlow, 2016. na stranici: https://www.cs.toronto.edu/~frossard/post/vgg16/.

[2] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402, December 2016.

[3] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, Contour and Grouping in Computer Vision*, pages 319–345. Springer Berlin Heidelberg, 1999.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[5] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

[6] Chuanhai Zhang. Medical image classification under class imbalance. *Graduate Theses and Dissertations*, 2019.