Универзитет у Београду
Електротехнички факултет

Детекција значајних тачака на лицу уз помоћ алгоритама машинског учења

Софтверско инжењерство великих база података

Ментори:
проф. др Мирослав Бојовић
ас. мс Стефан Тубић

Студент:
Матија Додовић,
2022/3053

Београд, јануар 2023.

САДРЖАЈ

1. Увод

Детекција кључних тачака лица је један од основних задатака компјутерске визије (*computer vision*) и има много практичних примена у различитим областима. Конкретно, откривање кључних тачака лица се широко користи у уређивању слика и видео записа, анимацији, препознавању лица, биометрији и интеракцији између човека и рачунара.

Могућност прецизног откривања и праћења кључних тачака лица омогућава креирање реалистичнијих и природнијих анимација, као и сигурније и прецизније системе за препознавање лица.

Кључне тачке на лицу су углови очију, зенице, врх носа и углови уста. Није искључено да се и центри образа убрајају у кључне тачке, али то за већину практичних примена не доноси боље резултате. Број кључних тачака везаних за очи и уста може варирати од неколико до веома великог броја, густо распоређених тако да веома прецизно дају контуре тих делова лица.

Кључне тачке пружају прецизну мапу црта лица, која се може користити у различите сврхе, као што су поравнавање лица, анализа израза лица и препознавање лица. Најчешћи приступ је коришћење учења надгледањем (*supervised learning*), где се велики скуп података слика са обележеним кључним тачкама користи за тренинг модела.

Неки од главних изазова везаних за податке у овој области јесту варијације у осветљењу, пози лица и изразима лица. Услови осветљења могу значајно утицати на изглед лица, што отежава моделима да прецизно открију кључне тачке. Слично томе, позе и изрази лица могу у великој мери да варирају, што резултира моделима који нису довољно робусни да поднесу ове варијације.

Важан изазов је проблем пристрасности података (*data bias problem*), где се скупови података који се користе за моделе обуке углавном састоје од слика људи са светлијим тоновима коже и неутралним изразима лица, што доводи до модела који нису довољно робусни на промене у тону коже или изразима лица.

Решавање ових изазова захтеваће развој модела који могу ефикасно да уче из различитих и разноврсних скупова података, као и интеграцију техника као што су повећање података и учење преноса.

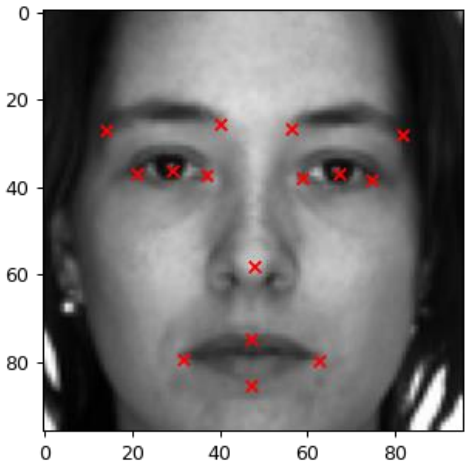
У овом раду су изложени неки модели неуралних мрежа и конволуционих неуралних мрежа који детектују основни скуп кључних тачака лица. Остатак рада је организован као што је наведено у наставку. Поглавље 2 (Подаци) даје приказ структуре података за тренинг и описује мотивацију и принцип препроцесирања података. Поглавље 3 (*Loss* и оптимизатори модела) описује могуће изборе функције *loss*, као и могуће, стандардне, оптимизације модела и наводи параметре коришћеног оптимизатора. Поглавље 4 (Модели неуралних мрежа и њихове перформансе) представља архитектуре свих коришћених модела, као и перформансе тих модела у погледу *loss*-а. Рад се завршава дискусијом резултата и закључком.

2. Подаци

Скуп података је преузет са Kaggle платформе [1]. Подаци представљају слике људских лица, резолуције 96x96 пиксела. Све слике су црно-беле. Подаци за тренирање се састоје од самих слика као и координата свих значајних тачака на лицу.

2.1. Значајне тачке лица

Пример изгледа једног податка из скупа података за тренирање (слике лица са одговарајућим значајним тачкама) приказан је на слици 1.



Слика 1. Пример једног податка из скупа података за тренинг са обележеним значајним тачкама лица.

За овај проблем је од интереса 15 значајних тачака лица, по 5 за два ока, 4 за уста и 1 за нос. Свака значајна тачка је репрезентована са две координате, при чему је почетак координатног система у горњем левом углу.

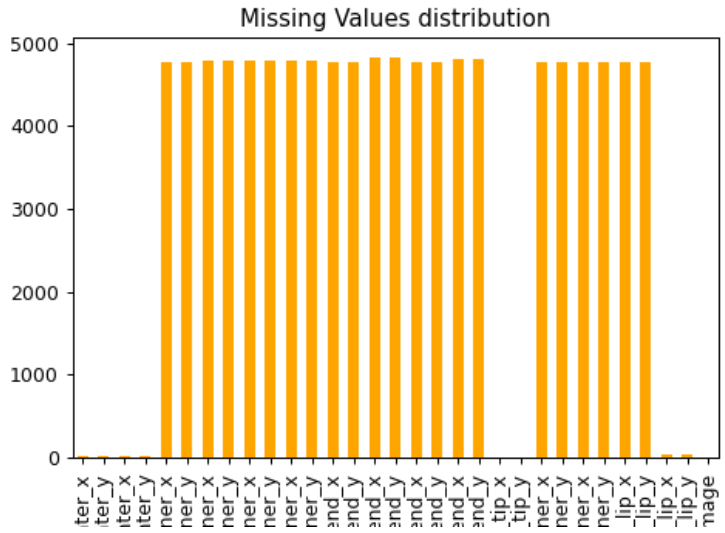
2.2. Препроцесирање

Почетни скуп података је такав да садржи 7049 слике. На слици 2 је приказана расподела недостајућих координата кључних тачака. За велики број улазних података, њих око 4800, важи да је бар једна од координата недостајућа. Одбацивање таквих улазних података би редуковало скуп података за тренинг за више од 68%, па се морају применити технике допуне података.

Овакав тип проблема представља вид континуалног проблема, зато што су координате реални бројеви у 2D координатном систему. Примењена је стандардна техника пропагирања вредности унапред (*forward fill*) која подразумева преписивање вредности из претходне ненедостајуће координате. Ова техника даје најбоље резултате услед природе самих

података, који су такви да је лице доминантно на фотографији, и исте кључне тачке на различитим сликама су релативно блиске једна другој.

Овом техником се, такође, одржава константан број улазних података, што је повољно за саме перформансе алгоритама.



Слика 2. Расподела недостајућих вредности.

2.3. Тренинг, валидациони и тест скуп података

Целокупни скуп података је подељен на део за тернирање, валидацију и тестирање перформанси. Скуп података за тренинг алгоритама чини 90% података (6344 податка). Над овим скупом сам алгоритам машинског учења учи тежине параметра. Валидациони и тест скуп су исте величине (оба по 5% почетног скупа података, што је по 352 податка). Валидациони скуп се посматра у току тренирања саме мреже, да би се на њему виделе перформансе у току самог извршавања. Тест скуп служи да се након завршеног процеса тренирања измере перформансе модела.

3. *LOSS* И ОПТИМИЗАТОРИ МОДЕЛА

loss функција описује колико направљени модел добро предвиђа вредности у односу на њихове стварне вредности. За овакав тип проблема су, стандардно, могућа два начина израчунавања *loss*-а. Један би био *MSE* (*Mean Squared Error*), а други *MAE* (*Mean Absolute Error*). MSE се рачуна као средње квадратно одступање између предвиђене и стварне вредности, док је MAE се рачуна као средња апсолутна разлика између предвиђене и стварне вредности. Главне разлике између ова два приступа су осетљивост на различите варијанте грешке. MSE је осетљивији за већа одступања предвиђене и стварне вредности, док је MAE једнако осетљив и на велика и на мала одступања.

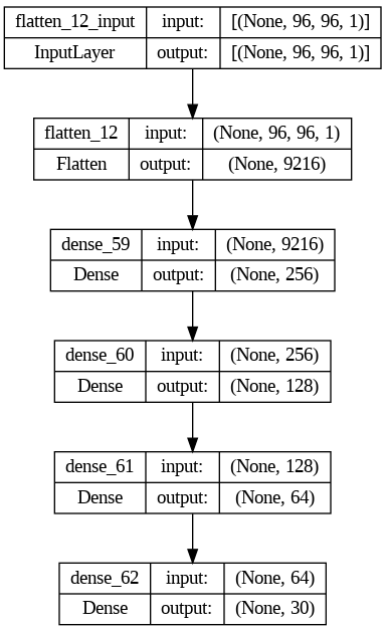
Сам процес тренинга се своди на промену тежина параметара у циљу минимизације *loss* функције. Основни алгоритам учења се ради методом градијентног спуста, чија се конвергенција може убрзавати адаптивним подешавањем хиперпараметра *learning rate*. Поред убрзавања самог процеса тренирања, потребно је обезбедити да модел не исконвергира ка неком локалном минимуму, и услед малог *learning rate*-а не успе да извуче најбоље перформансе. Стандардни оптимизатор који обједињује ова два приступа је *Adam* (*Adaptive Moment Estimation*), који је коришћен у свим алгоритмима. Он представља комбинацију два метода, један који представља адаптивни *learning rate* за сваки од параметара (*RMS prop*), и други који значајно убрзава процес тренирања (*Momentum*). Овај оптимизатор има почетну вредност *learning rate*-а која износи 0.001, а још су уведена и два хиперпараметра $\beta_1 = 0.9$ и $\beta_2 = 0.99$ (по један за сваки од алгоритама у *Adam*-у).

4. Модели неуралних мрежа и њихове перформансе

Тренирање свих модела је рађено у 100 епоха, а тренинг скуп је подељен на *batch*-eve величине 32 податка. У наставку су приказани различити модели неуралних мрежа и њихове перформансе. Излаз сваке од неуралних мрежа представља 30 параметара (по два за сваку од кључних тачака). За функцију *loss* је у свим неуралним мрежама коришћена функција *MAE*.

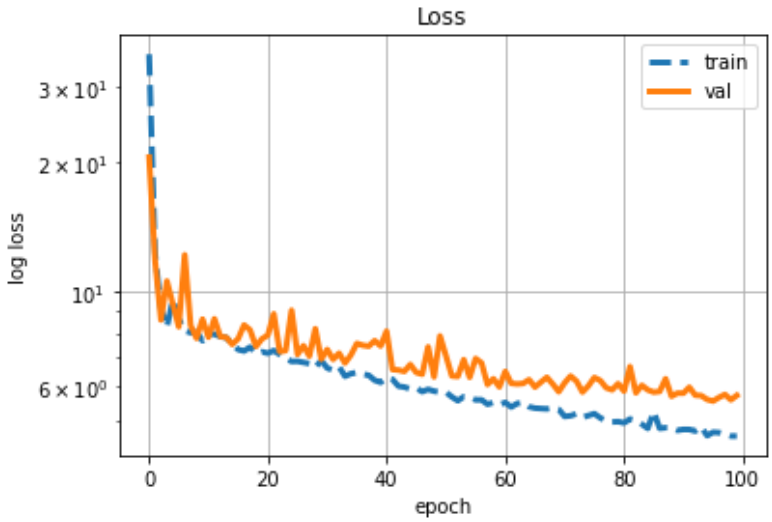
4.1. SmallNN

На слици 3 је приказана архитектура неуралне мреже под називом *SmallNN* (назив потиче од релативног броја скривених слојева у односу на остале разматране архитектуре мрежа). Улазни слој ове мреже представља слику 96x96 и након њега следи *Flatten* слој који линеаризује 2D слику у 1D вектор (са 9216 вредности). Три скривена слоја која следе су сва *Fully-Connected* и бројеви параметара су редом 256, 128 и 64. Излазни слој ове мреже је *Fully-Connected* слој са 30 параметара, који представља координате кључних тачака. Ова мрежа укупно има 2.402.654 параметара за тренирање.



Слика 3. Приказ архитектуре неуралне мреже под називом *SmallNN*.

На слици 4 је приказана функција *loss* за тренинг и валидациони скуп.

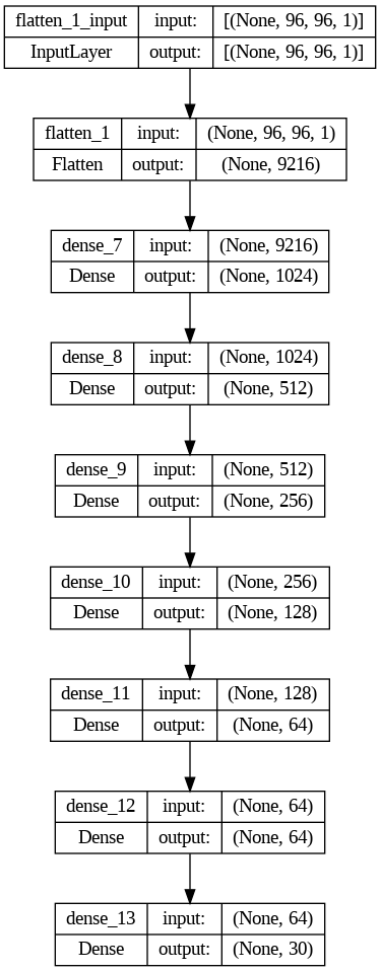


Слика 4. *loss* функција за тренинг и валидациони скуп података за неуралну мрежу под називом *SmallNN*.

Вредност функције *loss* (*MAE*) је над валидационим скупом, на крају тренирања, била 4.612, док је та вредност са тренинг скупом била 5.88.

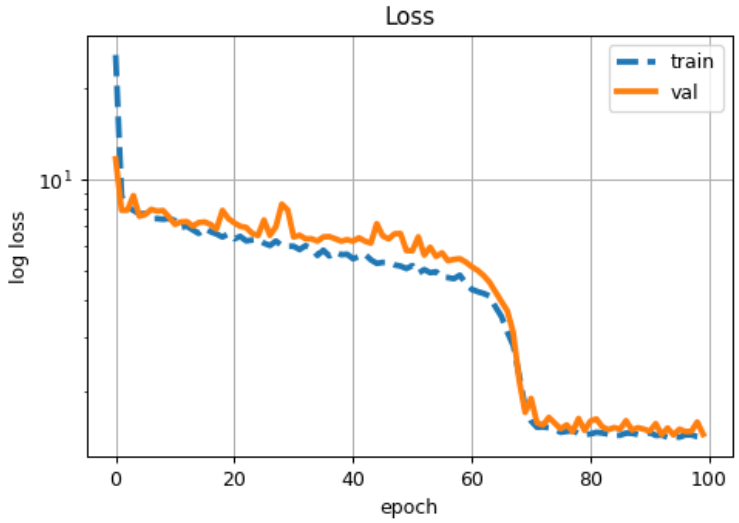
4.2. DeepNN

На слици 5 је приказана архитектура неуралне мреже под називом *DeepNN* (назив потиче од значајно већег броја скривених слојева у односу на неуралну мрежу *SmallNN*). Улазни слој ове мреже представља слику 96x96 и након њега следи *Flatten* слој који линеаризује 2D слику у 1D вектор (са 9216 вредности). Шест скривених слојева која следе су сви *Fully-Connected* и бројеви параметара су редом 1024, 512, 256, 128, 64 и 64. Излазни слој ове мреже је *Fully-Connected* слој са 30 параметара, који представља координате кључних тачака. Ова мрежа укупно има 10.141.598 параметара за тренирање.



Слика 5. Приказ архитектуре неуралне мреже под називом *DeepNN*.

На слици 6 је приказана функција *loss* за тренинг и валидациони скуп.

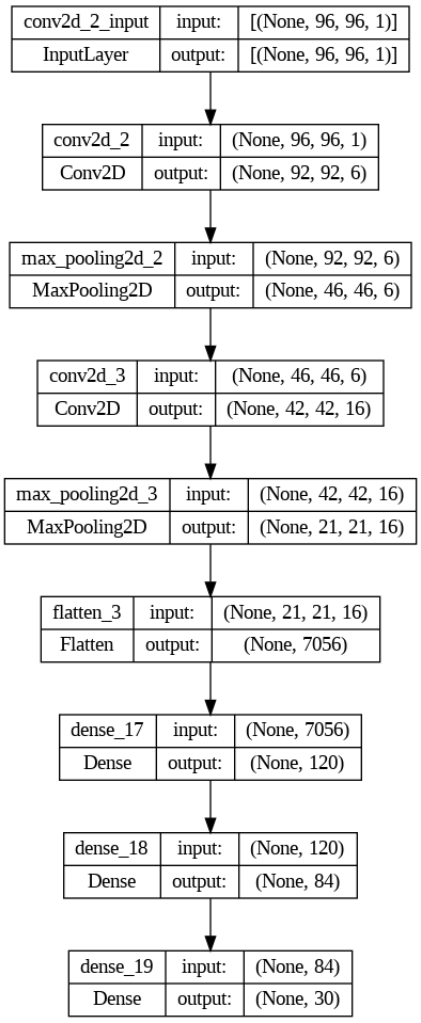


Слика 6. *loss* функција за тренинг и валидациони скуп података за неуралну мрежу под називом *DeepNN*.

Вредност функције *loss* (*MAE*) је над валидационим скупом, на крају тренирања, била 0.712, док је та вредност са тренинг скупом била 1.41.

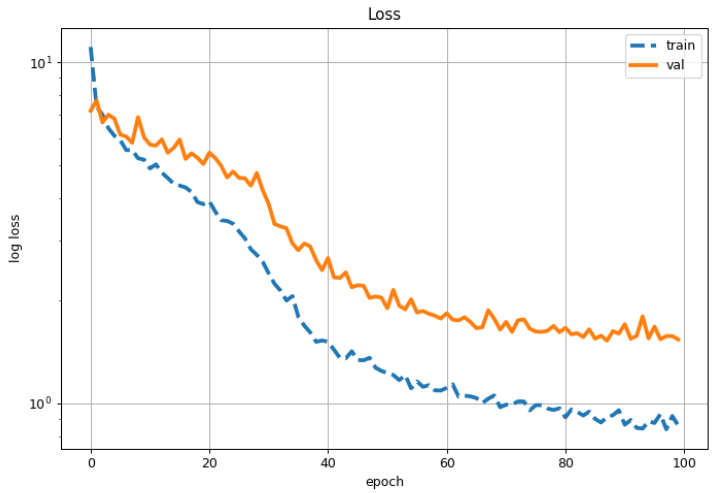
4.3. LeNet-5

На слици 7 је приказана архитектура конволуционе неуралне мреже под називом *LeNet-5* (назив потиче од архитектуралне сличности са конволуционом неуралном мрежом са истим називом, описаном у раду [2]). Улазни слој ове мреже представља слику 96x96 и након њега следе два конволуциона слоја, сваки праћен са по једним *MaxPool* слојем. Величине кернела оба конволуциона слоја су 5x5, док су димензије матрице за *MaxPool* слојеве биле 2x2. Први конволуциони слој је димензија 96x96 и има 6 филтера, док је други конволуциони слој димензија 46x46 са 16 филтера. Након конволуционог дела ове мреже следи *Flatten* слој који линеаризује 2D улазни податак (чије су димензије након примена кнволуционих и pooling слојева 21x21x16). Последња два скривена слоја су *Fully-Connected* са бројевима параметара редом 120 и 84. Излазни слој ове мреже је *Fully-Connected* слој са 30 параметара, који представља координате кључних тачака. Ова мрежа укупно има 862.126 параметара за тренирање.



Слика 7. Приказ архитектуре неуралне мреже под називом *LeNet-5*.

На слици 8 је приказана функција *loss* за тренинг и валидациони скуп.

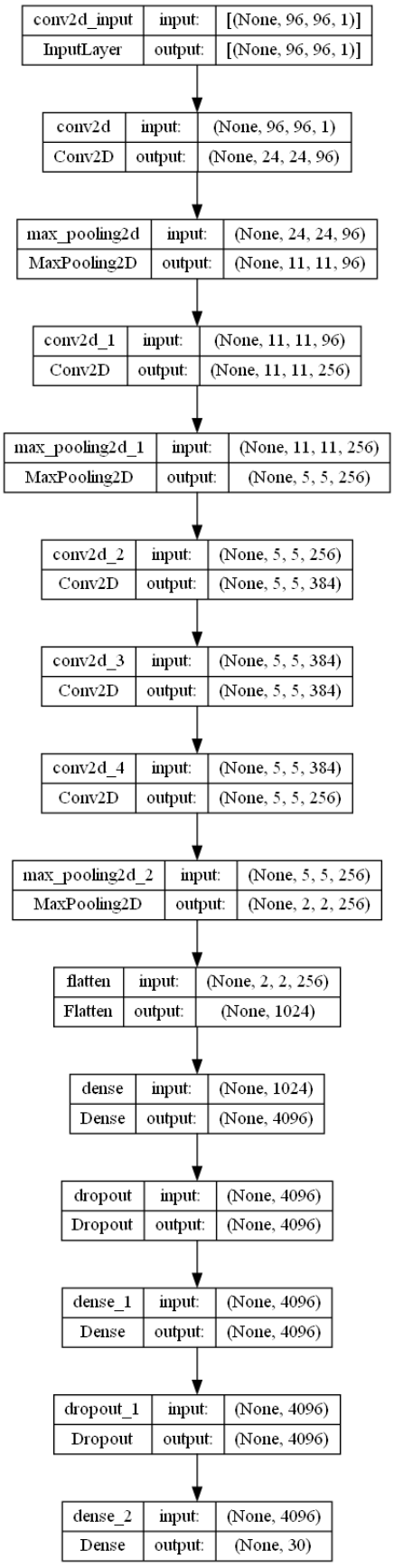


Слика 8. *loss* функција за тренинг и валидациони скуп података за неуралну мрежу под називом *LeNet-5*.

Вредност функције *loss* (*MAE*) је над валидационим скупом, на крају тренирања, била 0.744, док је та вредност са тренинг скупом била 1.178.

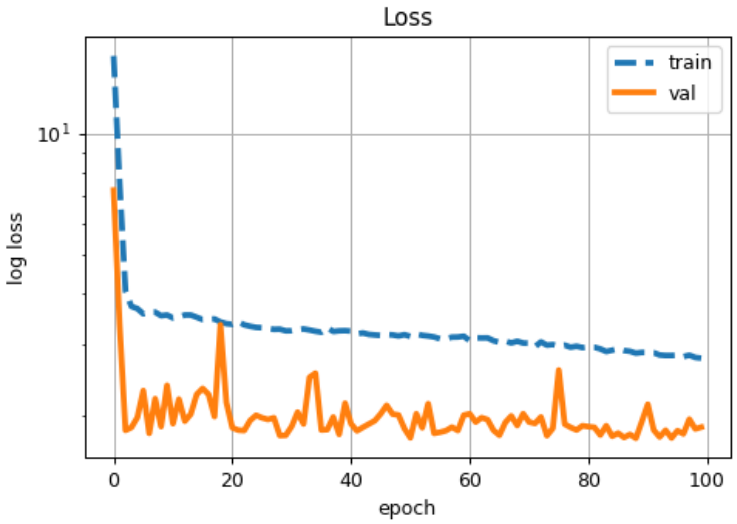
4.4. AlexNet

На слици 9 је приказана архитектура конволуционе неуралне мреже под називом *AlexNet* (назив потиче од архитектуралне сличности са конволуционом неуралном мрежом са истим називом, описаном у раду [3]). Улазни слој ове мреже представља слику 96x96 и након њега следе два конволуциона слоја, сваки праћен са по једним *MaxPool* слојем. Први конволуциони слој има кернел величине 11x11 и садржи 96 филтера, док други конволуциони слој има 256 филтера и кернел величине 5x5. Димензије матрице за *MaxPool* слојеве су 3x3 у оба случаја. Након ове групе, следи група од три сукцесивна конволуциона слоја, са редом 384, 384 и 256 параметара, док су величине кернела у свим случајевима 3x3. Један *MaxPool* слој величине 3x3 долази након ове три сукцесивне конволуције, да би након њега следео *Flatten* слој који линеаризује 2D улазни податак (чије су димензије након примена конволуционих и *pooling* слојева 2x2x256). Последња два скривена слоја су *Fully-Connected*, оба са 4096 параметара, а након сваког постоји и *Dropout* слој који за сваки *batch* гаси по 50% неурона, у оба случаја. Излазни слој ове мреже је *Fully-Connected* слој са 30 параметара, који представља координате кључних тачака. Ова мрежа укупно има 24.826.590 параметара за тренирање.



Слика 9. Приказ архитектуре неуралне мреже под називом *AlexNet*.

На слици 10 је приказана функција *loss* за тренинг и валидациони скуп.

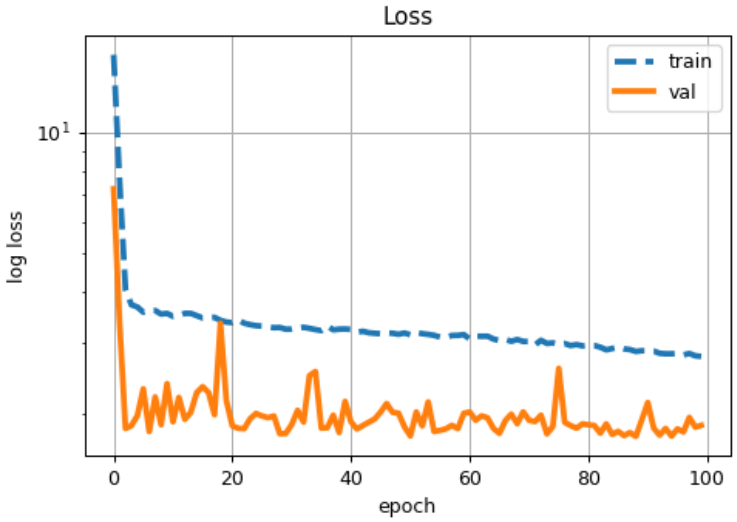


Слика 10. *loss* функција за тренинг и валидациони скуп података за неуралну мрежу под називом *AlexNet*.

Вредност функције *loss* (*MAE*) је над валидационим скупом, на крају тренирања, била 2.48, док је та вредност са тренинг скупом била 1.69.

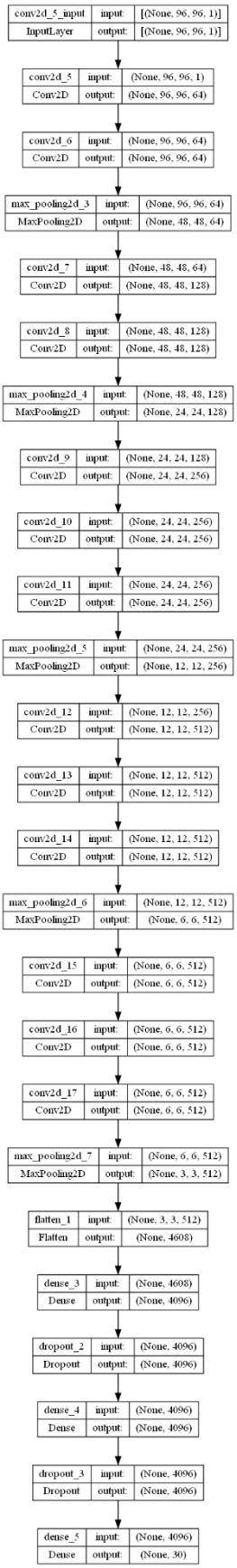
4.5. VGG-16

На слици 12 је приказана архитектура конволуционе неуралне мреже под називом *VGG-16* (ова неуронална мрежа је потпуно иста као из референтног рада [4]). На слици 11 је приказана функција *loss* за тренинг и валидациони скуп. Ова мрежа укупно има 50.496.222 параметара за тренирање.



Слика 11. *loss* функција за тренинг и валидациони скуп података за неуралну мрежу под називом *VGG-16*.

Вредност функције *loss* (*MAE*) је над валидационим скупом, на крају тренирања, била 3.2832, док је та вредност са тренинг скупом била 1.73.



Слика 12. Приказ архитектуре неуралне мреже под називом *VGG-16*.

5. ДИСКУСИЈА И ЗАКЉУЧАК

У табели 1 су приказани резултати перформанси свих коришћених алгоритама.

Табела 5.1. Перформансе алгоритама.

АЛГОРИТАМ	Број скривених слојева	Број параметара за тренирање	Loss над тест скупом [px]
SmallNet	3	2.402.654	5.88
DeepNet	6	10.141.598	1.41
LeNet-5	4	862.126	1.178
AlexNet	7	24.826.590	1.69
VGG-16	15	50.496.222	1.73

Графици тренда опадања вредности *loss* функције код алгоритама *AlexNet* и *VGG-16* показују да и тренинг и валидациони *loss* опадају кроз епохе. Тренд опадања ове функције има нагли пад у првим епохама. Услед чињенице да валидациони *loss* има мању вредност од вредности над тренинг скупом, закључује се да је дошло до појаве *Underfitting*-а. Ова два модела су значајно комплекснија (у погледу броја скривених слојева и у погледу броја параметара за тренирање) од осталих модела. Појава *Underfitting*-а се може јавити у ситуацији када модели нису довољно комплексни (што овде не представља случај) или у случају да величина и квалитет тренинг скупа нису довољно добри. Величине тренинг скупова над којима су ова два модела успешно тренирана су милион или десетине милиона слика (различите верзије *ImageNet* скупа података). За овај, релативно мали скуп података, ове две мреже су изразито комплексне и не дају добре резултате.

Перформансе мреже *SmallNet* су лошије у поређењу са мрежом *DeepNet*. Обе мреже имају архитектуру где је сваки слој *Fully-Connected*. Мрежа *SmallNet* има 5 пута мање параметара за тренирање и дупло мање скривених слојева. Ово указује на то да је за овај проблем и овај скуп података потребно имати комплекснију мрежу од представљене архитектуром мреже *SmallNet*.

Анализа тренда опадања вредности функције *loss* за мрежу *DeepNet* указује на то да се процес тренинга одвијао на сличан начин као код мреже *SmallNet* до ~60-те епохе, када се примећује значајан пад вредности *loss*-а. Разлог за овакав пад, односно проналажење неког бољег локалног минимума функције, лежи у већој комплексности ове мреже. У додатним експериментима са ове две, и додатним мрежама сачињеним само од *Fully-Connected* слојева, утврђено је да ће овај сценарио проналаска бољег локалног минимума функције бити присутан само уколико су бројеви неурона у првим скривеним слојевима бар два реда величине већи у односу на последњи слој, који генерише 30 координата кључних тачака. Конкретно се овај ефекат видео тек приликом постављања броја неурона првог слоја на 1024,

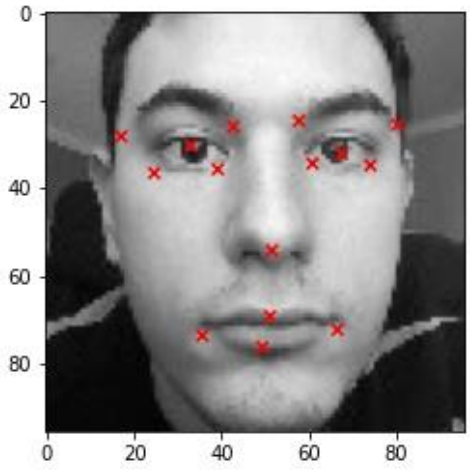
док је за веће вредности (2048 и 4096) утицај на перформансе био занемарив, док се време извршавања и број параметара значајно повећавао.

Квалитет мреже *DeepNet* се види у вредности функције *loss* над тест скупом, где она износи 1.41. То значи овај модел у просеку греши 1.41 пиксела приликом предвиђања значајних тачака лица, што представља значајно високу прецизност.

Високу прецизност у предвиђању кључних тачака постиже и конволуциона мрежа *LeNet-5*. Ова мрежа има вредност функције *loss* над тест скупом 1.2, што је боље од мреже *DeepNet*. Број параметара за тренирање које конволуциона мрежа *LeNet-5* има је више од 10 пута мањи од броја параметара *DeepNet*-а, док су перформансе ове две мреже релативно блиске. Објашњење за ову разлику лежи у чињеници да је проблем проналажења кључних тачака лица превасходно из домена рада са сликама, у коме доминантно најбоље резултате дају конволуционе неуралне мреже. Оне постижу значајно боље перформансе, а да притом имају значајно мањи број параметара за тренирање и да се те перформансе остварују над мањим скуповима података за тренирање.

Утицај на перформансе свих неуралних мрежа је имала и чињеница да је више од 60% скупа података за тренирање било препроцесирано и допуњено вредностима које не одговарају тачној позицији одговарајуће кључне тачке.

LeNet-5 је мрежа која има најбоље перформансе, а уз то и има најмање параметара за тренирање, и сходно томе најкраће тренирање. Она је коришћена у финалном генерисању резултата за *Kaggle* такмичење и у наставку анализе. На слици 13 је приказана слика која није из оригиналног скупа података заједно са кључним тачкама које су пронађене од стране мреже *LeNet-5*.



Слика 13. Тестирање мреже *LeNet-5* над сликом ван оригиналног скупа за тестирање.

Ова слика је урађена тако да буде блиска по структури сликама из оригиналног скупа података. Све кључне тачке су релативно добро предвиђене на овој слици, са одређеним одступањима у генерисању централне кључне тачке за оба ока. Предвиђене кључне тачке код слика које више одступају у односу на структуру слика из оригиналног скупа података су, код свих наведених алгоритама, у значајнијој мери лошије него код предвиђања на слици 13. То указује на чињеницу да је оригинални скуп података веома униформан по изгледу слика.

Литература

[1] https://www.kaggle.com/competitions/facial-keypoints-detection/data (посећен 25.01.2023)

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. “Gradientbased learning applied to document recognition“. *Proceedings of the IEEE*, 86 (11): 2278 – 2324, 1998.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks“. *Advances in neural information processing systems*, 1097–1105, 2012.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

Списак слика

Списак табела