Илија Витошевић IN 15/2021

Срђан Гарић IN 47/2021

Марко Коларски IN 60/2021

Примена експлоративне анализе и неуронских мрежа за класификацију религија на основу описа државе (Религиозно класификовање)

Извештај за практично истраживање

1. Увод

1.1. Предмет истраживања

Предмет овог истраживања је примена неуронских мрежа за класификацију религија држава на основу карактеристика њихових застава и основних геополитичких одлика држава.

1.2. Циљеви истраживања

Циљ је истражити да ли постоји веза између изгледа заставе и доминантне религије државе, као и које карактеристике су најзначајније за предвиђање, чији закључак ће бити и једним делом темељен на експлоративној анализи.

1.3. Задаци истраживања

Пронаћи и припремити скуп података за рад и упознати се са подацима. Урадити експлоративну анализу и извести неке закључке из ње. Имплементирати конволуциону неуронску мрежу (CNN). Прилагодити готову конволуциону мрежу (MobileNetV2). Након тога, потребно је упоредити успешност закључивања религије ове две мреже, а потом имплементирати вишеструку неуронску мрежу која ће обрадити нумеричке податке из CSV датотеке. Извести глобални закључак на крају.

1.4. Очекивани резултати истраживања

Очекивани резултати су идентификација најбитнијих карактеристика за класификацију религија, креирање модела са одређеном тачношћу предвиђања и стицање нових увида о везама између изгледа заставе и религије држава.

2. Методологија

2.1. Коришћени подаци

Користићемо скуп података који садржи информације о различитим државама и њиховим заставама. Подаци укључују основне геополитичке одлике, разноврсност боја, појаву разних облика и дизајн заставе државе. Скуп података се састоји од 177 држава и 30 колона које описују карактеристике као што су континент, површина, популација, језик, религија, број хоризонталних и вертикалних пруга на застави, боје заставе, доминантна нијанса, присуство симбола попут звезда, месеца, крстова итд. Такође смо саставили свој скуп података који садржи слике застава из 1986. године да одговара CSV подацима, детаљније о овом у методама истраживања.

2.2. Претходна истраживања других особа над коришћеним подацима

С обзиром да је овај скуп података са Универзитета Калифорније, Ервајн, а не са неког сајта као што је на пример Kaggle где се виде претходна истраживања других особа над коришћеним подацима, нисмо успели да нађемо ишта на сајту или интернету.

2.3. Методе истраживања

Наше истраживање укључује експлоративну анализу података, развој конволуционих неуронских мрежа и примену већ истренираног модела. У оквиру припреме података, први корак је био проналазак сајта са заставама држава из 1986. године. Неопходно је било осмислити метод за преузимање слика, за шта је развијена посебна Python скрипта.

Добијени скуп слика укључивао је бројне непотребне слике. Додатан посао је правило то што се називи датотека нису поклапали са називима држава у нашем CSV фајлу. Стога је било потребно прилагодити називе. Овај задатак смо реализовали кроз скрипту која је преименовала слике мењајући све називе у мала слова и брисањем свега што није назив државе.

Следећи корак је био развој нове скрипте која је упоређивала оба скупа података и уклонила непотребне слике. Направили смо на крају и тест који је проверавао да ли недостају неке слике. Иако скрипте нису биле савршене, значајно су убрзале процес обраде података. На крају, минималне грешке смо исправљали ручно. На пример, у случајевима када је једна држава имала више слика застава, било је потребно проверити која од њих одговара подацима у CSV фајлу, обрисати остатак и преименовати га да се идентично поклапају.

Након завршетка ових корака, преостало је мапирање назива датотека са називима држава. Подаци о сликама додати су као посебна колона у скупу података, што је омогућило потпуну интеграцију визуелних информација. На пример, у реду за Тунис треба да буде одговарајућа слика заставе Туниса заједно са свим осталим подацима о држави.

3. Резултати

3.1. Приказ резултата

Илија: Након извршене експлоративне анализе резултате сам приказивао графицима. Приказ сам поделио у три дела. Одлучио сам да не додајем графике у документ јер их има доста и тиме бих изгубио прегледност. Најбоље је покренути код који ће дати графике, а потом пратити текст о њима овде у документу. Усаглашен је редослед описа и графика са кодом.

Први део садржи анализу дистрибуције значајних варијабли. Дистрибуције су рађене на броју нивоа земаља поредећи језике, доминантну боју на застави, броја боја на застави, религије. Ту је и бинарна анализа фреквенција свих боја на заставама по њиховој учесталости појављивања на заставама. Дистрибуција језика показала је да највећи број држава спада у категорију осталих група језика. Језик који се говори као главни језик у врћини држава јесте енглески. Дистрибуције доминантне боје приказује црвену као најфреквентнију, док је браон боја најмање присутна на заставама. Потом следи дистрибуција броја боја на заставама из које се види да су заставе са три боје најучесталије, док су најмање присутне заставе са једном и осам боја, свега по једна држава. Однос религија по земљама говори да су остали хришћани у предности над свим осталим групама, а прате их католици и муслимани. Најмање има држава које припадају хиндуизму. Када је бинарна анализа учесталости у питању, графици свих боја (црвена, зелена, плава, жута, бела, црна, наранџаста) су представљни на једном графику. Видимо да се црвена и бела боја највише користи на заставама, а црне и наранџасте има најмање.

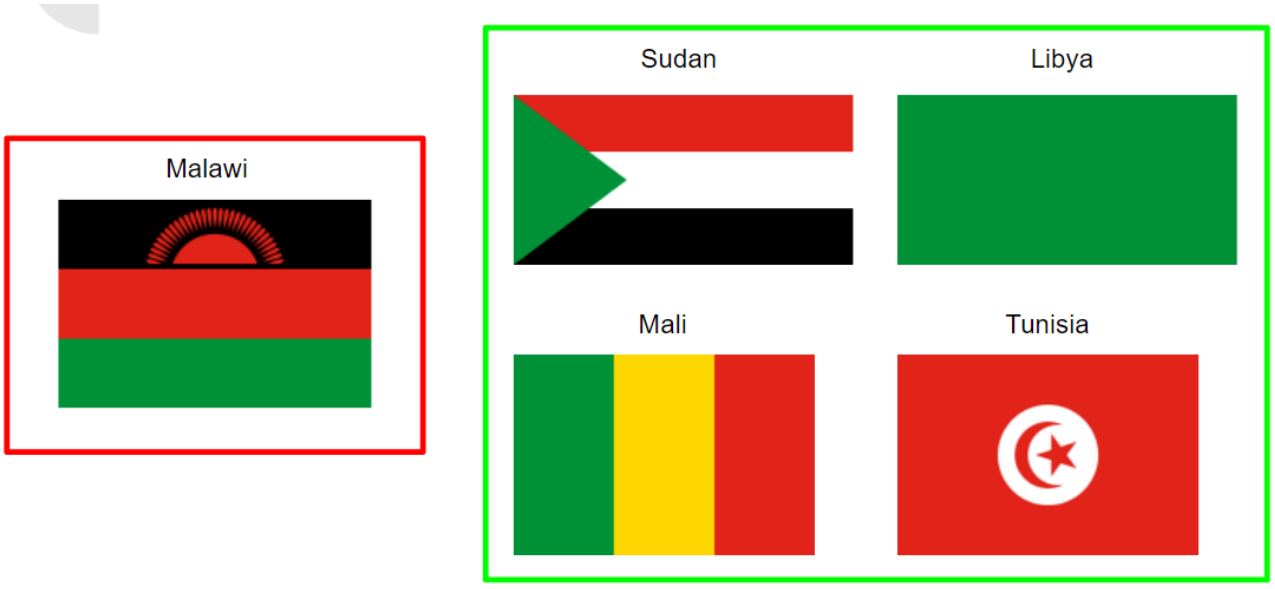
Други део обухвата поређење атрибута између класа. Ту желим да видим како постојање једне варијабле утиче на другу, однос какав је њихов однос. У првом сегменту поредио сам религију, оно што нам је био предмет истраживања са следећим компонентама: вертикалне пруге, хоризонталне пруге, број боја, број кругова, број крстова и број звезди и сунаца на застави заједно. У другом сегменту имам компоненте које поредим на нивоу континената и географских зона. То су доминантна боја, језик, религија. Потом следи поређење доминантне боје по религији и језику. Опис и тумачење свих ових графика ћу оставити за следећу секцију извештаја.

Трећи део представља корелациону анализу нумеричких варијабли у виду матрице. Ту налазим зависности између варијабли, како утичу једна на другу. Црвене нијансе означавају вероватноћу заједичког појављивања на истој застави неке две компоненте, док су плаве нијансе предвиђене за супротно значење. Такође ћу оставити тај део за тумачење.

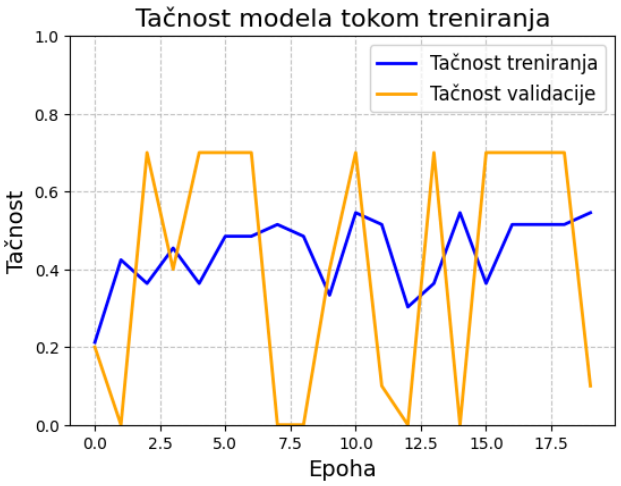
Марко: Помоћу конволуционе мреже предвиђао сам религију државе на основу њених слика заставе. Добијени су следећи резултати:

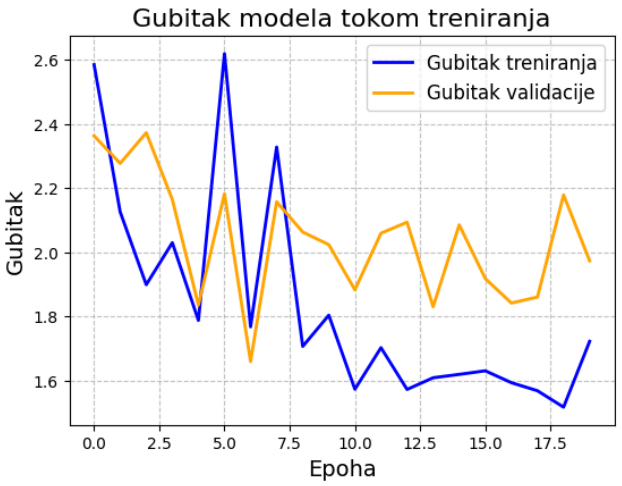
```
Test tačnost za "Africa": 0.80

    Test Uzorak |              Zemlja |     Stvarna Religija |  Predviđena Religija
-------------------------------------------------------------------------------------
              1 |               Mali |              Muslim |              Muslim
              2 |              Sudan |              Muslim |              Muslim
              3 |             Malawi |              Ethnic |              Muslim
              4 |            Tunisia |              Muslim |              Muslim
              5 |              Libya |              Muslim |              Muslim
```

Овде сам покушао да издвојим само један регион при подели података, јер сам хтео да видим да ли ће се повећати прецизност предвиђања уколико имамо заставе са већим бројем заједничких особина. За Африку смо имали највећи опсег слика као и разноврсне религије, али исто тако велике сличности. Тачност је 0.80 што је задовољавајуће али не и идеално. Важно је напоменути да у зависности од верзије Tensorflow-a добијају се различита решења. На графику је пример старије верзије са мањом тачношћу. При коришћењу нове верзија за овај специфичан случај добија се тачност 1.00.



Овде се јасно може видети зашто може доћи до грешке. Постоји јако велика сличност међу заставама иако се религије разликују.



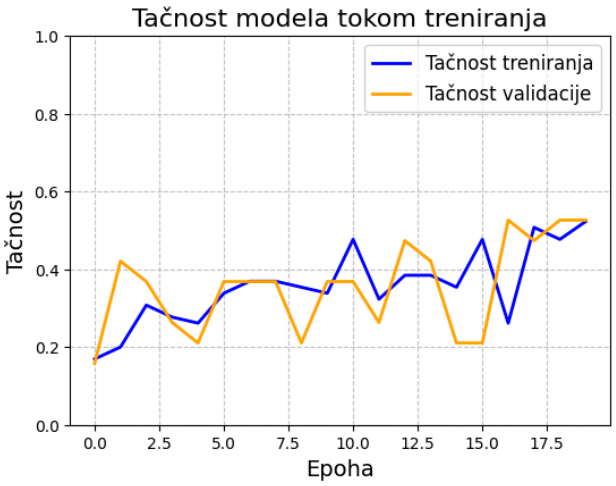


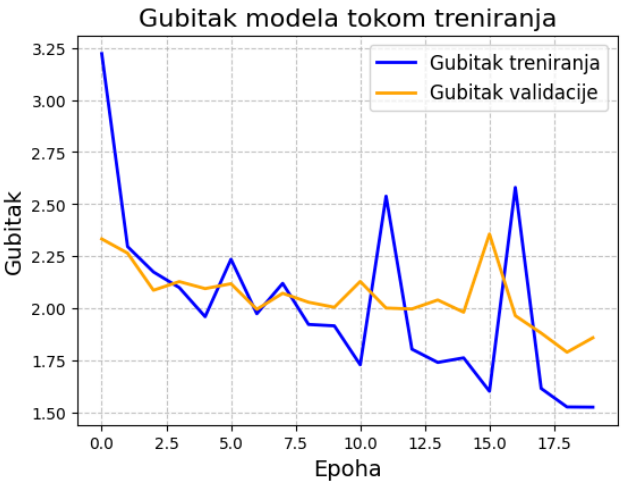
Са датог графикона можемо видети не тако учестале вредности за "тачност валидације", разлог тога је мали скуп података (свеукупно имамо 177 слика). Последица тога је велика разлика међу сликама због чега имамо тако велике скокове и падове.

```
Test tačnost za "Rest of the World": 0.50
```

Test Uzorak	Zemlja	Stvarna Religija	Predviđena Religija
1	Italy	Catholic	Catholic
2	Czechoslovakia	Marxist	Marxist
3	Mongolia	Marxist	Catholic
4	San-Marino	Catholic	Other Christian
5	Afghanistan	Muslim	Muslim
6	Finland	Other Christian	Other Christian
7	Israel	Others	Other Christian
8	Philippines	Catholic	Marxist
9	Bhutan	Buddhist	Catholic
10	Kuwait	Muslim	Muslim

Тачност за остатак света је 0.50, што наравно није довољно за неку поуздану предикцију. Разлог оваквих грешака је због превелике разноврсности међу заставама и њиховим религијама.



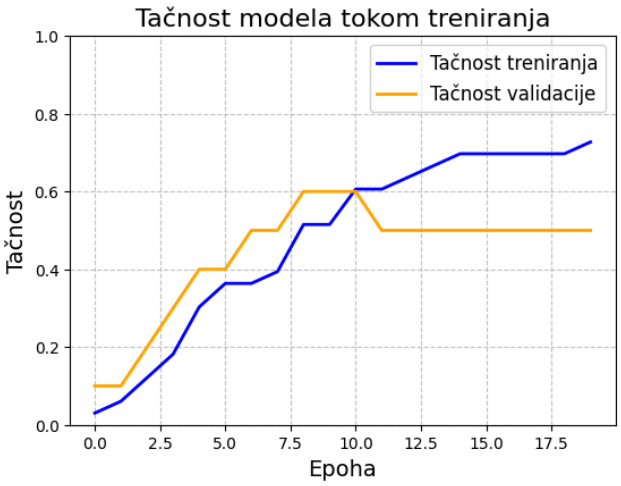


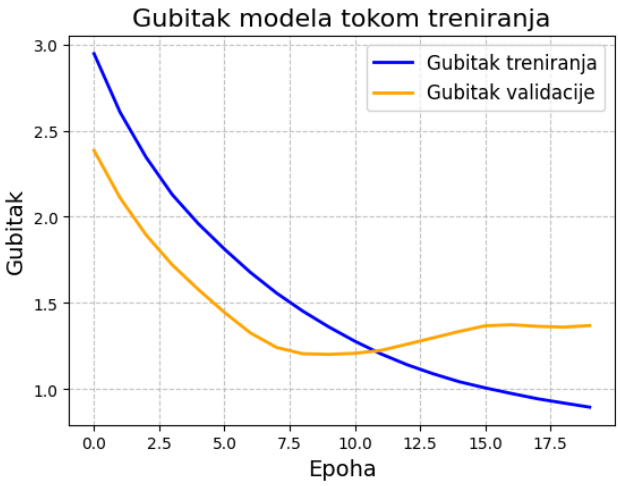
Овде можемо видети много нормалнији тренд за “тачност валидације” као и свеукупно мањи губитак модела током времена. Разлог томе је већи скуп података у односу на само регију (нпр. Африку).

Срђан: У оквиру пројекта, покушао сам да пронађем готову конволуциону неуронску мрежу која би се могла прилагодити нашем проблему и покушала да изврши предикцију, упоређујући своје резултате са резултатима колеге Марка. Међутим, резултати нису били најбољи.

```
Test tačnost za "Africa": 0.20

  Test Uzorak |              Zemlja |     Stvarna Religija |  Predviđena Religija
-------------------------------------------------------------------------------------
            1 |               Mali |              Muslim |              Ethnic
            2 |              Sudan |              Muslim |              Ethnic
            3 |             Malawi |              Ethnic |              Ethnic
            4 |            Tunisia |              Muslim |              Ethnic
            5 |              Libya |              Muslim |              Ethnic
```

Готова конволуциона неуронска мрежа коју сам користио, MobileNetV2, постигла је тест тачност од само 20% за регион Африке. Мрежа је давала предвиђања да све земље имају етничку религију, што није било тачно. Једино је била у праву за Малави, док је све остале државе помешала.



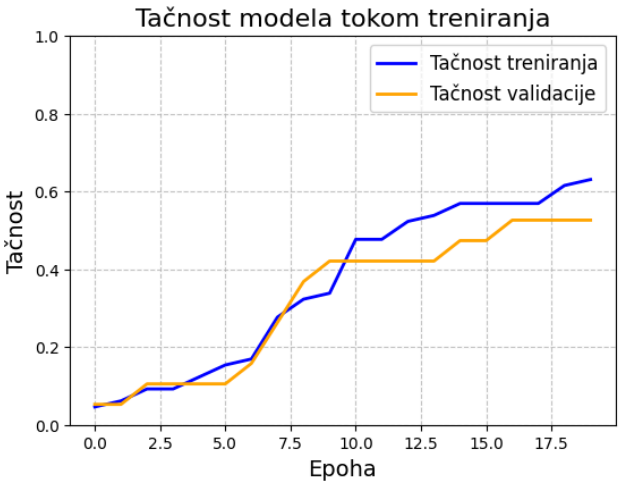


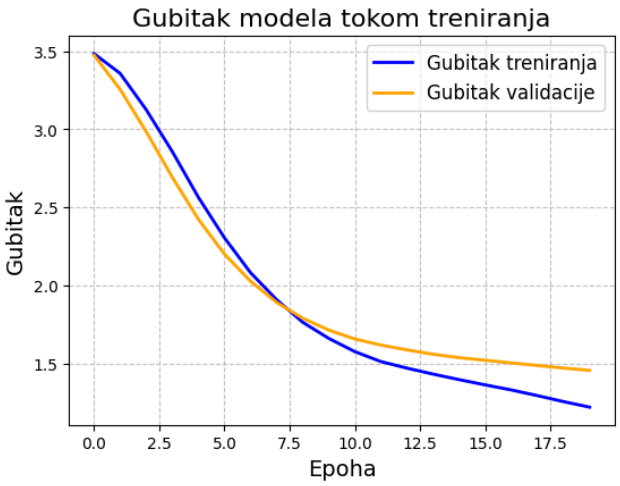
На графикону се види да је модел почео да се прекомерно прилагођава на тренинг податке, али то није значајно утицало на тачност предикције, јер су предвиђања већински била нетачна, без обзира на модел.

```
Test tačnost za "Rest of the World": 0.20
```

Test Uzorak	Zemlja	Stvarna Religija	Predviđena Religija
1	Italy	Catholic	Catholic
2	Czechoslovakia	Marxist	Catholic
3	Mongolia	Marxist	Catholic
4	San-Marino	Catholic	Muslim
5	Afghanistan	Muslim	Catholic
6	Finland	Other Christian	Catholic
7	Israel	Others	Other Christian
8	Philippines	Catholic	Other Christian
9	Bhutan	Buddhist	Other Christian
10	Kuwait	Muslim	Muslim

Што се тиче остатка света, тачност је такође била 20%. Мрежа је успела да тачно предвиди религије Италије и Кувајта, али је погрешила за све остале земље.





Са графикона се види да је модел био нешто бољи него код Африке, и да није показивао знаке прекомерног прилагођавања, али ово није довело до значајног побољшања у тачности предикције.

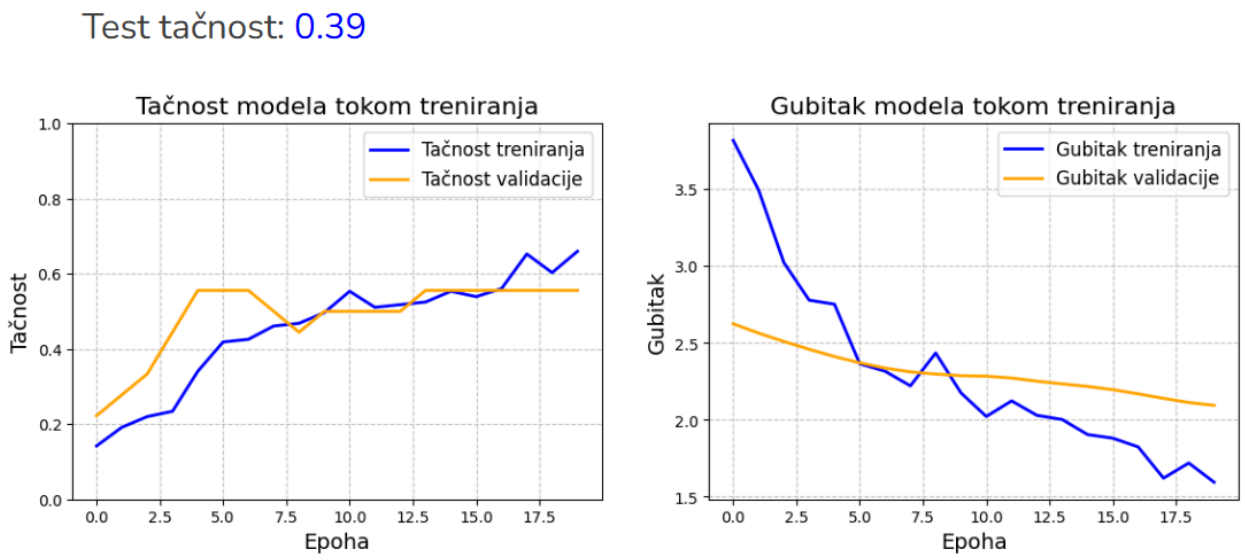
```
Test tačnost za "Africa": 0.80 (CNN) i 0.20 (MobileNetV2)

    Zemlja |     Stvarna Religija |  CNN Predviđanje | MobileNetV2 Predviđanje
-------------------------------------------------------------------------------
      Mali |               Muslim |            Muslim |                  Ethnic
     Sudan |               Muslim |            Muslim |                  Ethnic
    Malawi |               Ethnic |            Muslim |                  Ethnic
   Tunisia |               Muslim |            Muslim |                  Ethnic
     Libya |               Muslim |            Muslim |                  Ethnic
```

Након проналаска готове неуронске мреже и испитивања њених перформанси, упоредио сам је са моделом од колеге Марка. Ту можемо да видимо да је његова мрежа била много успешнија. Занимљиво је што су предикције за Африку донеле "обрнуте" резултате. Моја MobileNetV2 неуронска мрежа је рекла да су све државе етничке религије, док је Маркова рекла да су све муслиманске. Он је био убедљиво тачнији са 80% тест тачности, док сам ја имао 20%.

```
Test tačnost za "Rest of the World": 0.50 (CNN) i 0.20 (MobileNetV2)

          Zemlja |     Stvarna Religija |  CNN predviđanje | MobileNetV2 predviđanje
-------------------------------------------------------------------------------------
           Italy |             Catholic |         Catholic |               Catholic
    Czechoslovakia |              Marxist |          Marxist |               Catholic
        Mongolia |              Marxist |         Catholic |               Catholic
      San-Marino |             Catholic |  Other Christian |                Muslim
     Afghanistan |              Muslim |           Muslim |               Catholic
         Finland |  Other Christian |  Other Christian |               Catholic
          Israel |              Others |  Other Christian |  Other Christian
     Philippines |             Catholic |          Marxist |  Other Christian
          Bhutan |            Buddhist |         Catholic |  Other Christian
          Kuwait |              Muslim |           Muslim |               Muslim
```

Даље, за остатак света можемо видети да је прича иста. Моја готова мрежа са 20% тест тачности је бледела у поређењу са његових 50%. Обојица смо добили добре резултате за Италију и Кувајт, али он је такође имао добре резултате и за Чехословачку, Авганистан и Финску. Његова свеже имплементирана неуронска мрежа је била доста боља од моје готове, што није било чудно, као што ће бити објашњено касније у тумачењу.

Коначно, након што сам завршио са обрадом слика, изградио сам обичну неуронску мрежу која ће, на основу доступних нумеричких података (CSV), одредити успешност закључивања религије. У овом случају нисам делио државе по регијама, већ сам гледао целокупан скуп података да бих видео како ће се модел снаћи. Нажалост, модел опет није био толико успешан.



Тест тачност је била 39%, а максимум који сам достигао је био 56%, али то је било када је модел био превише прилагођен, што је била пука срећа.

3.2. Тумачење резултата

Илија: Тумачење ми ја дало бољи увид у податке са којима радим. Поновоћу дискутовати по деловима.

Први део ми доноси једноставан увид у бројеве, односно појављивања одређених ставки по државама. Пошто су ти описи прости, уврстио сам их у секцију приказ резултати, јер они то заиста јесу, па ту више не бих ништа коментарисао, јер једноставност приказаног искључује потребу за тумачењем.

Други део је сложенији у односу на први и треба му посветити пажњу као што сам обећао у секцији изнад. Кренућу са првим сегментом. Поређење броја вертикалних пруга на застави и религија доноси увид у то да вертикалне пруге у нису тако популарне међу државама, невезано за религију, јер су највиши стубићи на графику управо присутни код броја нула за вертикалне пруге. Једино што бих издвојио јесте да су три вертикалне пруге најкарактеристичније са католичке државе. Поређење броја хоризонталних пруга и религије говори да је најчешће видети такву заставу са три боје на њој. Главни носиоци овога су религије попут католичанства и ислама. Приметан је велики број осталих хришћанских земаља које немају хоризонталне пруге на својим заставама. Поређење броја боја и религија већ има везу са претходним бројем вертикалних и хоризонталних пруга, јер су управо најбројније заставе са три боје. Прате их заставе са две и четири боје. Може се рећи да је углавном обележје обе класе хришћанства три боје на застави, док је за ислам то или две или четири боје. Потом следи поређење броја кругова на застави и религије. Кругови су нешто што није популарно на заставама према графику, али се истиче постојање једног круга на заставама у ретким случајевима код свих религија. Поређење броја крстова на заставама са религијом апсолутно говори о томе да ако држава има један или више крстова на застави, увек смо сигурни да је у питању држава хришћанског света, углавном то нису католици. Преостало је још поређење збира звезди и сунаца на застави са религијом. Ову збирну комбинацију сматрам незахвалном за поређење узимајући у обзир моје познавање географије и застава, али хајде да видимо шта график каже. Хришћанске и муслиманске државе предњаче у томе да немају ништа од наведеног на застави, но, ако имају једно обележје углавном су то муслиманске земље. Сада следи други сегмент. Доминантна боја заставе по континентима казује да је црвена најкоришћенија, што смо већ сазнали, али даљим увидом сада знамо да је то баш изражено у Азији, Европи, Африци и Јужној Америци. Зелена боја ипак предњачи на афричком континенту, док је плава главна за Северну Америку и Аустралију. Што се тиче географских зона, на северо-истоку далеко испред осталих је црвена, како североисток обухвата Европу и Азију увиђамо правилност у делу поређења са претходним погледом на нивоу континената. И даље се види да црвена боја предњачи, невезано за регион. Језик по континентима и зонама након дужег рада над подацима не сматрам тако битним параметром, па ћу кратко рећи да се само погледа график који код избацује и све ће бити јасно. Исто важи и за доминантну боју на застави по језику. Сада се враћам на озбиљнији део. Религија по континентима говори да је Европа претежно насељена хришћанима, као и Аустралија, Северна и Јужна Америка. У Европи су значајно још присутни и они који су посвећени марксистичом учењу. У Африци преовладавају етничке племенске религије. У Азији се истичу ислам, будизам и марксизам. Следи сада увид у график доминантне боје и религије. У свим групама хришћана доминирају црвена, плава и бела боја. Марксисте, односно социјалистиче државе одликује црвена боја, док су црвена и зелена ознаке муслиманских земаља. Сада се направио леп круг поређења континената, религије и доминантне боје.

У трећем делу изнећу закључке до којих сам дошао посматрањем мартице корелације. Површина и популација укључују једна другу, односно што је већа површина, то ће бити већа популација државе. Када смо код величине државе, што је већа, имаће већу могућност појаве звезди или сунаца на застави. У први мах, то ми је било чудно, али сам провером дошао до тога да то заиста јесте тачно, јер 1986. ту имамо СССР, САД, Кину и Бразил као примере. Са друге стране очекивано је да ће појава хоризнталних пруга искључивати појаву вертикалних, важи и обрнуто. Појава вертикалних пруга искључиће појаву беле боје. Када је у питању број боја на застави, што их је више, пре очекујемо да су међу тим бојама црвена, зелена, жута и наранџаста, као и слике и текст. Што се тиче боја приметно је неслагање зелене и плаве, плаве и црне, као и жуте и беле, док се плава и бела очекују, као и жута и слика на застави. Још једна од очекиваних ставри јесте да ако нађемо крст на застави, на њој у мањој мери можемо очекивати да нађемо и Андрејин крст, који изгледа као ћирилично слово х. Велика Британија на својој застави има ову комбинацију, где је Андрејин крст преузет као шкотско обележје.

Марко: Након завршетка рада са конволуционом мрежом, могу да кажем да су се моје претпоставке делимично испуниле. Уколико државе поделимо по смисленим регијама тако да се повећа број заједничких карактеристика, повећаће се и прецизност предикције. У мом примеру тачност за Африку је била 0.80, а за остатак света 0.50, што указује на мој закључак.

Међутим, ови резултати нису довољно поуздани за предвиђање религије. Постоје бројни случајеви где се прецизност значајно мења. Због тога, ако користимо само слике застава као податке за одређивање религије, то неће бити довољно за наше потребе. Неопходно је интегрисати додатне информације и контекстуалне податке како бисмо побољшали тачност и поузданост наших предвиђања.

Срђан: Што се тиче резултата MobileNetV2 готове конволуционе неуронске мреже, она се показала лоше из више разлога. Главни проблем је тај што је обучена на Imagenet скупу података, који има милионе слика - међутим, наше заставе су из 1986. године, а Imagenet има новије верзије застава. Неке ни нема јер су се у међувремену државе распале (нпр. Совјетски Савез) и та неусклађеност између података се показала критичном. Такође, у обзир долазе и мале димензије наших слика које су са потенцијално недовољним детаљима угрозиле предвиђања.

Ако ћемо конкретно, Африка је јако изазован регион јер има доста различитих застава и религија, културолошки је измешана, и модел није успео да се снађе. Ово је било очекивано, зато смо је и задали. За остатак света се види исти проблем као и код вишеслојне мреже која је покушала да обради цео свет одједном - једноставно су заставе превише разноврсне да би се категорисале у конкретну религију, што ће бити додатно потврђено у закључку.

4. Закључак

4.1. Анализа испуњења циљева истраживања

Илија: Да ли постоји веза између изгледа заставе и доминанте религије државе? У некој мери постоји. Увидели смо да су на пример искључиво хришћанске државе имају крст или више њих на застави. За остале државе и њихове религије сазнали смо носице боја застава по религији и континенту, али то не би било довољно да са сигурношћу можемо да тврдимо које је религија држава чија је застава дата, свакако то су неке од најбитнијих карактеристика за посматрање.

Марко: Изградњом конволуционе мреже и тренирање исте са сликама застава држава закључујем да је то премали скуп података и да ће константно долазити до грешака поготово уколико узмемо скуп који није издељен на неке регије. Видели смо побољшање када смо је поделили на Африку и остатак света али ни то није идеално и грешке су очекиване. Велики разлог томе су сличности међу заставама а различите религије за исте. Можемо видети ограничења ма како год ми поделили наш скуп. Сматрам да се религија овако не може предвиђати и да треба да се нађе неко боље алтернативно решење.

Срђан: Када говоримо о испуњењу циљева готове конволуционе неуронске мреже, она није заблистала. Делом је то због неадекватног скупа података на којем је била тренирана, али и зато што на религију неке државе утиче мноштво комплексних фактора. Поред баналних распореда боја и присуства одређених дезена, ту су и историјски, културолошки, демографски, социоекономски и географски фактори. Овај проблем се највише видео код вишеслојне неуронске мреже која је покушала да донесе закључак на глобалном нивоу, при чему је изразито пала. Закључити религију само на основу заставе и пар додатних информација је тежак, скоро немогућ посао, због постојања великог броја аномалија које се крше са трендовима изгледа неких застава.

4.2. Анализа остварења очекиваних резултата истраживања

Илија: Нисам очекивао да ћу моћи да по било којој ставки будем у могућности да одгонетнем неку религију државе на основу карактеристика заставе, јер је екплоративна анализа требала да ми пружи добар увид у податке са којима радимо, не нужно и неке закључке које бих очекивао од неуронске мреже, те сам задовољан добијеним резултатима, како увидом у податке, тако и чињеницеом да само хришћанске земље имају крст на застави. То сада можда звучи тривијално, али на почетку нисам одмах имао ту идеју у мислима.

Марко: Моје неко првобитно мишљење је било да очекујем релативно лошије резултате за измешан већи скуп података (нпр. остатак света), али оно што нисам могао предвидети је тачност по регијама. Очекивао сам да ће то бити поуздано и сигурно. На основу резултата видим да то није случај и да сам се преварио. Разлога за грешку има пуно, али опет то није нешто што може да се исправи, односно барем не у мом случају.

Срђан: Лично сам искрено мислио да ће бити лако направити неуронску мрежу која ће са тачношћу од чак 80% или 90% закључити религију, али сам се грдно преварио јер при почетној изради нисам дубље размишљао о комплекснијим факторима који утичу на религију. Очекивања су ми се потпуно оповргнула, али сам у међувремену схватио изазовну реалност овог истраживања, који фактори највише утичу на религију и добио позамашно знање о неуронским мрежама.

4.3. Могућности за примену истраживања у пракси

Ово истраживање и цео овај модел би се најбоље уклопили у некој потенцијалној видео игри која би припадала катеогорији стратегије. Ту би на основу дате заставе могли да утврдимо као играчи у игри са ким нам је најлакше да склопимо савез, пошто своју религију знамо. Овим бисмо обавезали да постоји правило игре које каже да је склапање савеза лакше ако државе деле исту религију. Такође, могло би да помогне у неком креативном стваралаштву фикције где се измишљају нове државе, па би као помоћ овај пројекат дошао у смислу одређивање културе и религије те државе.

4.4. Идеје за побољшање и разраду истраживања

Велика побољшања једино можемо видети у моделима неуронских мрежа, у смислу коришћења јачих модела, дужег тренирања, мењања броја слојева, активационих функција, итд. Ове промене би довеле до минималног побољшања због претходно споменуте природе проблема. Поред тога, коришћење ажурнијих података и слика застава већих димензија довело би до већег успеха готових неуронских мрежа, мада је и даље пожељно правити свој модел од нуле, јер тренутно не постоји ниједан готов модел који је специјализован за предвиђање религије на основу заставе државе. Последња опција би била ручна подела података по регијама, где би се бирали представници религија који имају најупечатљивији и најчешћи изглед, па би се мрежа кроз тренинг угледала на то, али и даље би постојали изузеци код неких држава око чега се мало шта може урадити. Места за напредак има, али побољшање не би било велико.

5. Литература

[1] Скуп података: https://archive.ics.uci.edu/dataset/40/flags

[2] Слике застава: https://flaglog.com/1986