Mihailo Vukorep (IN 40/2021)

Препознавање ручно написаног текста  
(Book Archiver)

Извештај за практично истраживање

# Увод

## Предмет истраживања

Предмет овог пројекта се заснива на испитивање шта је све потребно да би раучар препознао ручно написан текст без коришћења неких већ готових библиотека које своје решење нуде у пар линија кода.

## Циљеви истраживања

Сам циљ пројекта јесте архивирање ручно написаних књига у неки текстуални документ зарад презервације истих.

## Задаци истраживања

Можемо задатак за имплементирање оваквог програма поделити у две обимне целине.

Првa целина представља само тренирање довољно добре конволуционе неуронске мреже за препознавање ручно написаног тескста, док другу целину представља издвајање индивидуалних карактера са слике и просеђивање тих слика истренираној мрежи.

Очекивани резултати истраживања

Због недостатка рачунарске снаге које поседујемо не очекујемо да ћемо имати неке довољно успешне резултате.

# Методологија

## Коришћени подаци

Користимо један од познатијих тренинг скупова под именом EMNIST (Extended MNIST). Његов претходник MNIST је био сачињен само од ручно написаних цифара, док је EMNIST сачињен и од малих и великих слова. Слике су црно-беле и у формату 28\*28 пиксела.

Сами подаци су подељени по категеоријама: ByClass, ByMerge, Balanced, Letters, Digits, MNIST

Такође је свака категорија подељена на тренинг и тест скупове.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Назив категорије** | **Укупан број слика** | **Тренинр (train)** | **Тест (test)** | **Број класа (карактера)** |
| EMNIST ByClass | 814 255 | 697 932 | 116 323 | 62 |
| EMNIST ByMerge | 814 255 | 697 932 | 116 323 | 47 |
| EMNIST Balanced | 131 600 | 112 800 | 18 800 | 47 |
| EMNIST Letters | 145 600 | 124 800 | 20 800 | 26 |
| EMNIST Digits | 280 000 | 240 000 | 40 000 | 10 |
| EMNIST MNIST | 70 000 | 60 000 | 10 000 | 10 |

Пре коришћења самог скупа проверамамо да ли постоје дупликати.

Дупликате пролазимо тако што израчунавамо MD5 хеш за сваку индивидуалну слику затим мапирамо хеш на индек слике у сету који касније поредимо са следећим хешом који рачунамо. Ако израчунамо хеш који смо претходно израчунали наишли смо на дупликат.

Aнализом сваке од категорија, утврђено је да постоје дупликати у тренинг сетовима:

|  |  |
| --- | --- |
| **Назив категорије** | **Број дупликата** |
| EMNIST ByClass | 9 |
| EMNIST ByMerge | 9 |
| EMNIST Letters | 5 |

Даљом аналзизом је утврђено да категорије попут ByClass и ByMerge имају велики број слика који међусобно користе, на пример ByClass тренинг (train) скуп садржи слике које ByMerge користи у свом (test) скупу.

Консултовати следећи графикон:

Тако да је одлучено да

Претходна истраживања других особа над коришћеним подацима

## Методе истраживања

...

# Резултати

## Приказ резултата

...

## Тумачење резултата

...

# Закључак

## Анализа испуњења циљева истраживања

...

## Анализа остварења очекиваних резултата истраживања

...

## Могућности за примену истраживања у пракси

...

## Идеје за побољшање и разраду истраживања

...

# Литература

## ...

## ...

## ...