Зборник радова Факултета техничких наука, Нови Сад

UDK: 004.9
DOI: https://doi.org/10.24867/24BE04Curcic

ПРЕПОЗНАВАЊЕ ПИСАНИХ ЈАПАНСКИХ СЛОВА УПОТРЕБОМ НЕУРОНСКИХ МРЕЖА

RECOGNITION OF JAPANESE HANDWRITTEN CHARACTERS USING NEURAL NETWORKS

Марија Ђурчић, *Факултет техничких наука, Нови Сад*

Област – СОФТВЕРСКО ИНЖЕЊЕРСТВО И ИНФОРМАЦИОНЕ ТЕХНОЛОГИЈЕ

Кратак садржај – *У овом раду описано је решавање проблема детекције и класификације писаних јапанских слова употребом конволуционих неуронских мрежа. Решење се састоји из три модула: модул за обраду података, модул за детекцију и модул за класификацију. Коришћена је једна конволуциона неуронска мрежа за детекцију и једна за класификацију слова. Класификација је рађена на различитим групама јапанских слова.*

Кључне речи: *неурноске мреже, cnn, писана јапанска слова, детекција, класификација*

Abstract- *This paper describes solving the problem of detection and classification of written Japanese letters using convolutional neural networks. The solution consists of three modules: data processing module, detection module and classification module. One convolutional neural network was used for detection and one for letter classification. The classification was done on different groups of Japanese letters.*

Keywords: *neural natworks, cnn, handwritten japanese characters, detection, classification*

1. УВОД

Писани историјски документи представљају значајан део културног наслеђа једног народа. Аутоматска транскрипција оваквих докумената је од виталног значаја за стварање дигиталних библиотека, као и превођење тих докумената на савремени језик. Да бисмо слике писаних историјских докумената превели у дигиталне документе који би могли да постану доступни и разумљиви великом броју људи, и који би били погодни за претраживање, потребни су нам системи за транскрипцију од високе прецизности.

Кузушиђи (јап. 崩し字 – *kuzushiji*: „деформисани симболи“) је назив за курзивно јапанско писмо, односно ручно писана јапанска слова [1]. Овим писмом су писани многи историјски документи у Јапану, и оно се визуелно доста разликује од писма које се користи данас. С обзиром на то да мали број људи уме да чита кузушиђи, системи за превођење оваквих докумената на савремено јапанско писмо су од великог значаја.

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је био др Александар Ковачевић, ред. проф.

Решење представљено у овом раду анализира слике писаних јапанских текстова и ради детекцију сваког слова у тексту и класификацију појединачних слова употребом модела конволуционих неуронских мрежа. За евалуацију предложеног решења коришћене су мере тачности (*accuracy*) на обучавајућем и тестном скупу, као и F1 мера на тестном скупу података.

У поглављу 2. описани су скупови података који су коришћени у овом решењу. Архитектура предложеног решења описана је у поглављу 3. У поглављу 4. представљени су резултати. Поглавље 5. садржи закључак и сумаризацију целог рада.

2. СКУП ПОДАТАКА

Иницијални скуп података преузет је са адресе [2]. Овај скуп података је лабелиран и подељен на скупове за обучавање и тестирање. Садржи 3605 слика за обучавање и 1730 слика за тестирање. Свака слика представља скенирану папирну страница која садржи неки текст написан курзивним јапанским словима, а неке странице садрже и илустрације (Слика 1). Скуп података садржи и CSV фајлове са лабелама за сваку слику, које садрже назив слике, координате, висину, ширину и ознаку *Unicode* карактера за свако слово са слике.

Слика 1. *Примери слика из скупа података*

Овај скуп података је прошао обраду која је описана у поглављу за претпроцесирање. Од добијеног скупа су касније, уз додатне модификације креирани различити скупови података за сваки модул овог пројекта и за различите експерименте.

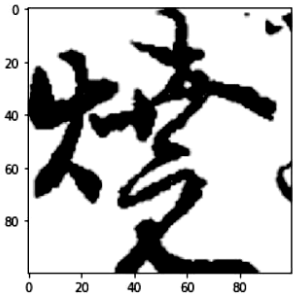
2.1. Скуп података за детекцију

За детекцију су коришћене слике целих страница из претходно описаног скупа, али су, због ограничења меморије при обучавању, додатно смањене са димензија 1933×1146 на димензије 1000×593. Ове слике су сачуване као нови скуп података, уз лабеле које су модификоване тако да одговарају новим димензијама слика. Улазне слике модела за детекцију изгледају као примери на Слика 2.

Слика 2. *Примери слика након претпроцесирања*

2.2. Скуп података за класификацију

Из скупа података који је прошао кроз претпроцесирање описано у поглављу издвајају се слике појединачних слова исецањем слика страница на основу података за граничне оквире у фајлу са анотацијама. Све исечене слике се пребацују на величину од 100×100 (Слика 3).



Слика 3. *Пример улазне слике за класификатор*

На овај начин је добијен нови скуп података који садржи слике појединачних слова и лабеле које садрже ознаке по једног *Unicode* карактера за сваку слику.

У овом скупу постоји 1400 различитих слова и симбола. Због превелике количине података, меморије и времена које је потребно за обучавање, модел за класификацију није обучаван на целом скупу. Уместо тога је издвојено четири подскупа над којима је обучаван и евалуиран модел. Један од ових подскупова добијен је издвајањем слова која су најзаступљенија у скупу података. Остала три подскупа добијена су издвајањем слова која припадају различитим групама слова у јапанском језику. Издвајање ових подскупова и обучавање модела детаљније су описани у експериментима 2, 3, 4 и 5.

3. МЕТОДОЛОГИЈА

Систем је имплементиран у три модула: модул за претпроцесирање, модул за детекцију и модул за класификацију.

3.1. Претпроцесирање

Пре употребе скупа података примењени су одређени кораци претпроцесирања. За обраду слике коришћена је OpenCV библиотека. Свака слика је прво конвертована из BGR у RGB формат. Висина и ширина су промењене тако да све слике буду истих димензија. Приликом промене димензија слика, измењене су вредности за граничне оквире у фајлу са анотацијама тако да одговарају новим димензијама слика. Затим је одрађено конвертовање слике у нијансе сиве, односно *grayscale*. На сваку слику *threshold*, коришћењем *Otsu* методе за проналажење прага. На Слика 1 су примери слика из почетног скупа података, а на Слика 2 може се видети како изгледају исте те слике након што су прошле кроз претпроцесирање.

3.2. Модул за детекцију

Задатак модула за детекцију јесте да нађе сва јапанска слова која се налазе на некој улазној слици, односно да одреди њихове граничне оквире. Због великог броја различитих слова која се могу појавити на слици, модул за детекцију је направљен тако да препознаје само да ли се на слици налази слово и да одреди граничне оквире, али не и врсту тог слова. Због тога за овај модел постоји само једна класа на излазу.

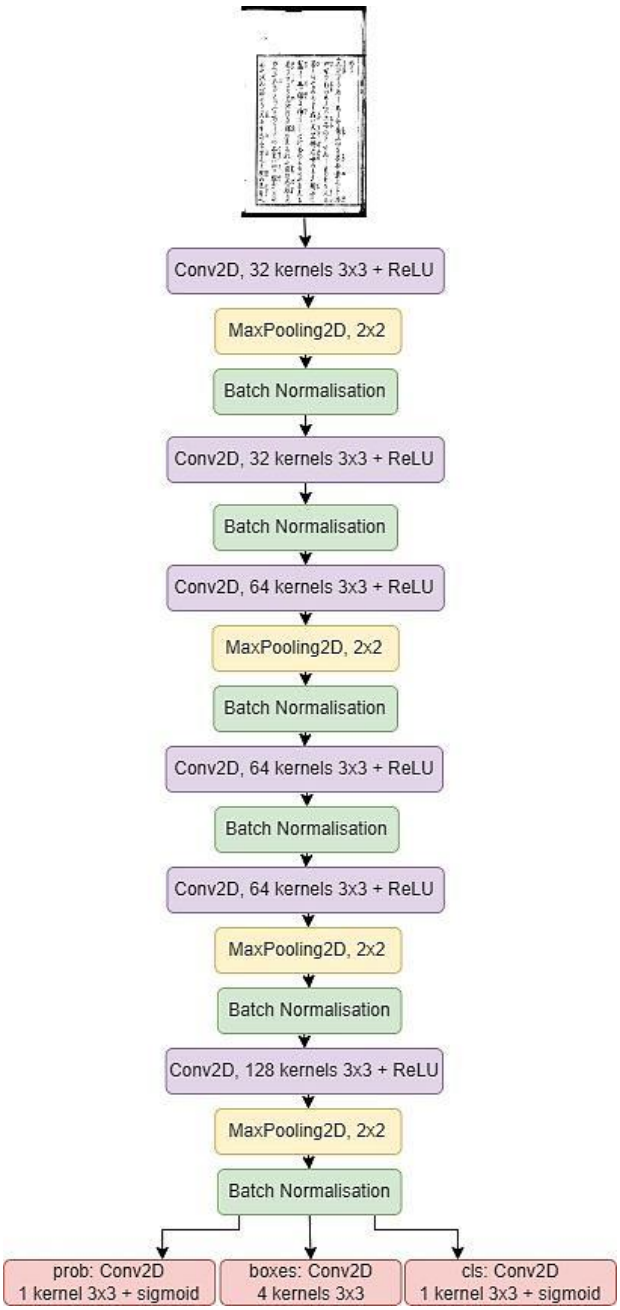
За детекцију је коришћен модел конволуционе неуронске мреже, чија је архитектура приказана на Слика 4. Ова архитектура је креирана по угледу на модел за детекцију писаних цифара [3], уз неке измене.

Модел садржи 9 конволуционих слојева, од којих су први, трећи, пети и шести праћени по једним *MaxPooling* слојем, а између сваког блока је коришћена *batch* нормализација. Последња три конволуциона слоја служе за одређивање вероватноће, класе и граничних оквира, и њиховом конкатенацијом се добија излаз из мреже. Конволуциони слојеви за вероватноћу и класу користе *Sigmoid* активациону функцију, док сви унутрашњи конволуциони слојеви користе *ReLU* активациону функцију.

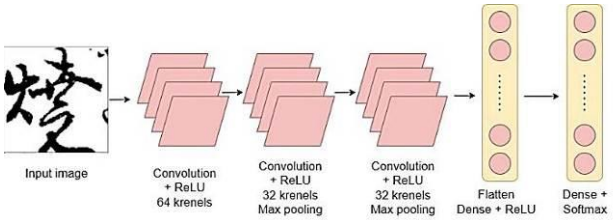
Функција губитка *binary cross-entropy* је примењена за одређивање вероватноће, а *mean squared error* за одређивање граничних оквира.

3.3. Модул за класификацију

Задатак модула за класификацију је да свакој слици на додели једну класу која представља неко јапанско слово. Улазни податак представља слика димензија 100×100 на којој се налази једно писано јапанско слово. Као класификатор коришћена је конволуциона неуронска мрежа чија је архитектура приказана на Слика 5.



Слика 4. *Структура модела CNN за детекцију*



Слика 5. *Структура модела CNN за класификацију*

Први слој мреже је конволуциони слој са 64 кернела димензија 3×3. После њега следе два блока који се састоје из једног конволуционог слоја са 32 кернела, сажимајућег слоја који примењује *max pooling* технику са кернелом димензије 2×2, и *dropout* регуларизације од 0.4. Сви конволуциони слојеви примењују *ReLU* активациону функцију. Између ових слојева је примењивана *Batch* нормализација. Затим следи слој за поравнање и потпуно повезани слој са *ReLU* активационом функцијом. Излазни слој је *softmax*, а број класа се мењао у зависности од експеримента и скупа података. Функција губитка помоћу које је модел оптимизован је *categorical cross-entropy*.

4. РЕЗУЛТАТИ И ДИСКУСИЈА

Модел за детекцију обучаван је у 30 епоха. Обучавајући скуп је чинио 80%, а валидациони скуп 10% целокупног скупа података за детекцију. За валидацију модела коришћена је метрика тачност (*accuracy*). Број епоха и подела скупа података утврђени су експериментално. Након обучавања, модел је евалуиран на тестном скупу, који је креиран применом *random* узорковања и чини 10% целокупног скупа података.

Модел за класификацију обучаван је на четири различита подскупа скупа података за класификацију. Први подскуп су чиниле слике слова која се појављују 300 или више пута. Други подскуп чине само слова која припадају хирагани, а трећи слова која припадају катакани. Четврти поскуп чине канци симболи. Сва четири подскупа за класификацију су подељена на исти начин: 80% скупа је коришћено за обучавање, 10% за валидацију и 10% за тестирање. Скуп за тестирање креиран је применом *stratified* узорковања. Модел за класификацију је на свим скуповима обучаван у 20 епоха.

Табела 1 представља резултате тачности на скуповима за обучавање и тестирање за детекцију и класификацију.

Табела 1. *Резултати*

Експеримент	Обучавање	Тест
Детекција	0.97	0.96
Класификација најзаступљенијих	0.99	0.94
Класификација хирагане	0.96	0.93
Класификација катакане	0.99	0.94
Класификација канџија	0.99	0.95

За експерименте класификације на изолованим групама слова (катакана, хирагана и канџи) одређене су F1 мере по класама на тестном скупу и макро F мера. У

Табела 2У Табели 2. приказане су неке од класа хирагане и њихове F1 мере, као и заступљеност тих класа у тестном скупу. Из Табеле 2. можемо приметити да су подаци прилично неизбалансирани, што се одражава и на њихове вредности за F1 меру. Макро F1 мера за хирагану износила је 0,86.

Табела 2. *F1 мера по класама за хирагану*

Класа	F1 мера	Заступљеност
い	0.96	415
が	0.92	272
ぼ	0.83	27
ぱ	0.25	4
を	0.98	762

У Табели 3. приказане су неке од класа катакане, њихове F1 мере и заступљеност у тестном скупу.

Док су и овде подаци неизбалансирани, скуп података у овом експерименту био је знатно мањи од скупа за хирагану, а самим тим су и разлике у заступљености различитих класа у тестном скупу мање. Макро F1 мера је износила само 0,63, што је знатно лошији резултат у односу на хирагану. Ово је последица тога што су многе класе за F1 меру имале вредност 0.

Табела 3. *F1 мера по класама за катакану*

Класа	F1 мера	Заступљеност
ア	0.98	72
イ	0.93	24
シ	0.66	1
ツ	0.92	19
リ	0.75	4

У Табели 4. приказане су неке од класа из канџи скупа, њихове F1 мере и заступљеност у тестном скупу. Без обзира на многобројност и већу комплексност канџија у односу на хирагану и катакану, овај експеримент је имао најбољу тачност на тестирању од свих експеримената класификације. За разлику од хирагане и катакане, скуп канџија је био боље избалансиран, што се може видети по заступљености података у тестном скупу. Такође можемо приметити да све класе имају приближно високе вредности за F1 мере и нема великих одступања. Макро F1 мера износила је 0,93.

Табела 4. *F1 мера по класама за канџи*

Класа	F1 мера	Заступљеност
三	0.95	64
上	0.94	59
大	0.92	65
太	0.86	25
屋	0.98	27

Задатак детекције слова у овом раду био је олакшан, тако да се одређује само позиција слова, односно гранични оквири, без одређивања врсте слова које је детектовано. Због овога је било могуће користити једноставнију архитектуру неуронске мреже, која није претходно обучавана.

Проблем код класификације хирагане и катакане била је неизбалансираност података, али су, без обзира на то, постигнути добри резултати код тачности у свим експериментима.

5. ЗАКЉУЧАК

У овом раду представљене су методе детекције и класификације писаних јапанских слова. Решење је подељено на модул за обраду слика, модул за детекцију и модул за класификацију слова и знакова.

Модел за детекцију обучаван је на сликама целих страница исписаних курзивним јапанским писмом, а модел за класификацију обучаван је на сликама изолованих слова које су добијене исецањем страница из првобитног скупа. Модел за класификацију обучаван је на четири различита подскупа од датог скупа података и анализирани су резултати класификације за различите групе јапанских слова.

У овом раду класификатор је обучаван на различитим групама класа (слова и знакова) одвојено. На скупу са хираганом је обучен за 71 слово, са катаканом је обучен за 50 слова, а на канџи скупу је обучен за 103 слова. На мешовитом скупу, односно скупу у којем се појављују све врсте слова, класификатор је обучен тако да препознаје 59 слова. Скуп података који је представљен у раду довољно је обиман да се ово решење прошири на већи број класа.

На основу добијених резултата, показало се да се уз довољан број примерака у скупу података може постићи задовољавајућа прецизност у препознавању различитих врста слова и карактера у курзивном јапанском писму. Проблем неизбалансираности података, који је био присутан код класификације, могао би се лако превазићи проширивањем скупа исецањем још изолованих слова из слика страница које нису биле искоришћене у скупу података или аугментацијом података.

Модул за детекцију би се могао унапредити тако да осим детекције слова, препознаје и класу, тј врсту тог слова. Ово би се могло постићи коришћењем сложенијих и претренираних модела за детекцију, као што су YOLO и R-CNN.

6. ЛИТЕРАТУРА

[1] Introduction to kuzushiji 崩し字: http://naruhodo.weebly.com/blog/introduction-to-kuzushiji, приступљено у марту 2023.

[2] Kuzushiji Recognition: https://www.kaggle.com/competitions/kuzushiji-recognition/data, приступљено у марту 2023.

[3] https://tree.rocks/a-simple-way-to-understand-and-implement-object-detection-from-scratch-by-pure-cnn-36cc28143ca8, приступљено у марту 2023.

Кратка биографија:

Марија Ђурчић рођена је у Новом Саду 1999. године. Дипломирала је на Факултету техничких наука у Новом Саду 2021. године и исте године уписала мастер студије на Факултету техничких наука, смер Софтверско инжењерство и информационе технологије – интелигентни системи.