Georgian Handwritten Character Recognition Using Deep Learning

Professor: Magda Tsintsadze

Tbilisi State University

Goals of the research:

- Test and modify best Machine Learning algorithm for handwritten recognition
- Create datasets for classification and correction
- Deploy the module in Web, Desktop, iOS

Definitions

- OCR - Optical Character Recognition
- Artificial neural networks - weighted directed graphs in which artificial neurons are nodes and directed edges with weights are connections between neuron outputs and neuron inputs
- Convolutional Neural Networks - deep, feed-forward artificial neural networks used for analyzing visual imagery
- Training/Validation sets - Training set is used for training ANN model, Test set consists of new data, which can be used to check generalization of the network

Importance

- Modified and improved existing CNN model
- Multiple NN models trained for Georgian Recognition
- Largest Georgian Handwritten character dataset
- High quality Georgian word dataset

Previous Work

For English Characters:

- Printed: 1920s - Emanuel Goldberg Statistical Machine
- 1974 - Ray Kurzweil omni-font OCR
- 2016 - Jian-Xia Wang 98.86 92% testing

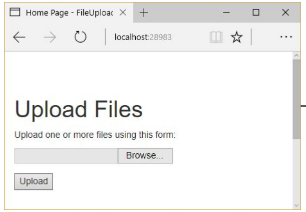
Alternative Methodology

- Matrix matching
- k-nearest neighbors algorithm
- Fourier Descriptors
- Feature extraction
- Support Vector Machines
- Neural networks

Process

Mobile

Web



Home Page - FileUploa

localhost:28983

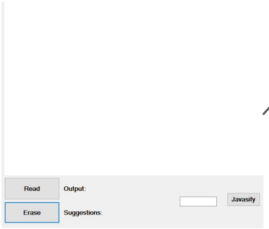
Upload Files

Upload one or more files using this form:

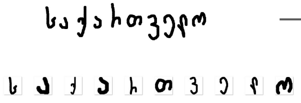
Browse...

Upload

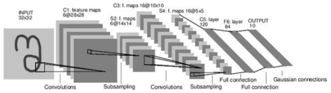
Windows



Segmentation



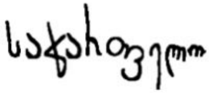
Recognition



Spelling checker

საქართველო -> საქართველო

Generator



Collecting training set

- TSU 2017 vefxistyaosani, parsed using our segmentation algorithm
- TSU 2006 vefxistyaosani, parsed using our segmentation algorithm
- School handwritten texts
- Digital handwritten text Collected via webpage
- Other donated handwritings

33 classes, over 60 000 letters

Dataset Composition

Training Set

33 classes

Unbalanced - AVR 2300 characters, Min1500 characters

Balanced using augmentation - 7000 characters each

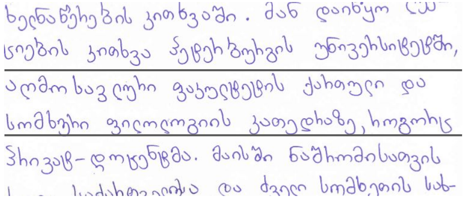
TestSet

33 classes

Unbalanced - AVR 920 characters, Min 600 characters

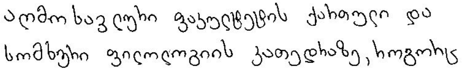
Balanced using augmentation - 2500 characters each

Segmentation and Preprocessing

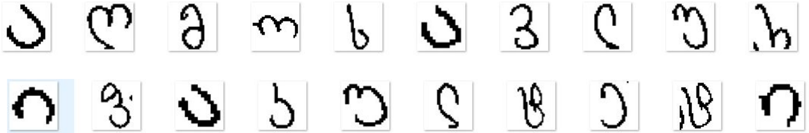


ხელნაწერების კითხვაში. მან დაიწყო ლა-
ციების კითხვა ჰეიდელბერგის უნივერსიტეტში,
აღმოსავლური ფაკულტეტის ქართული და
სომხური ფილოლოგიის კათედრაზე, როგორც
პრივატ-დოცენტმა. მაისში ნაშრომისათვის
სადისერტაციო და ძველი სომხურის სახ-

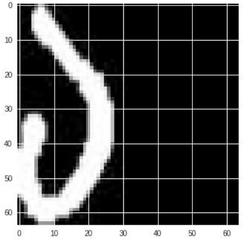
Otsu's method



აღმოსავლური ფაკულტეტის ქართული და
სომხური ფილოლოგიის კათედრაზე, როგორც

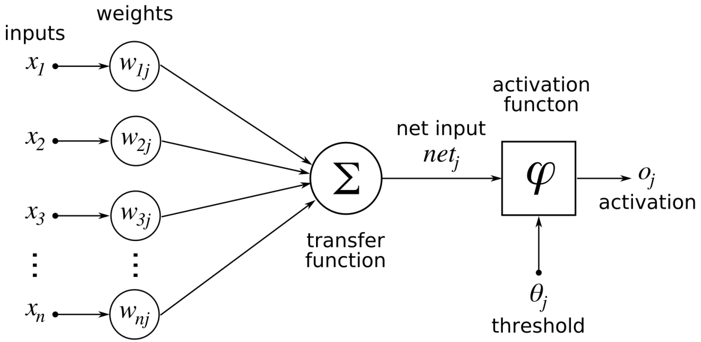


Inverted
standardized
Range [0 -1]

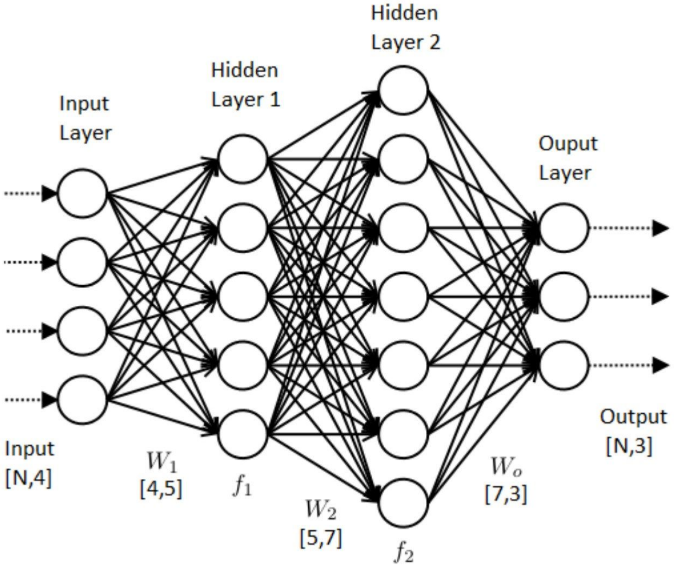


Recognition

Neuron



Neural Network

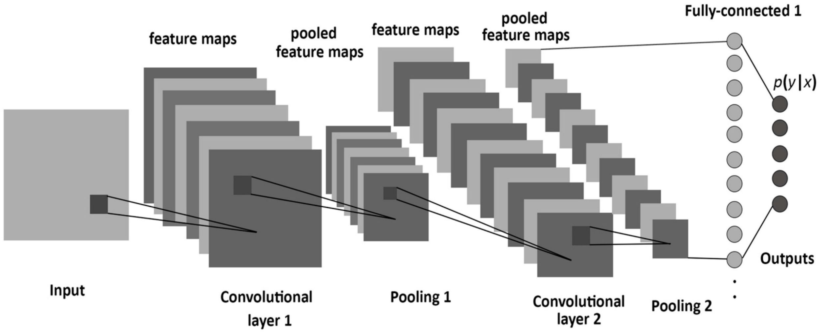


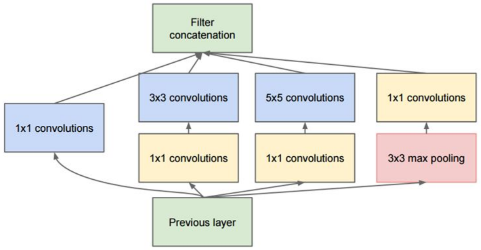
Recognition

Resnet

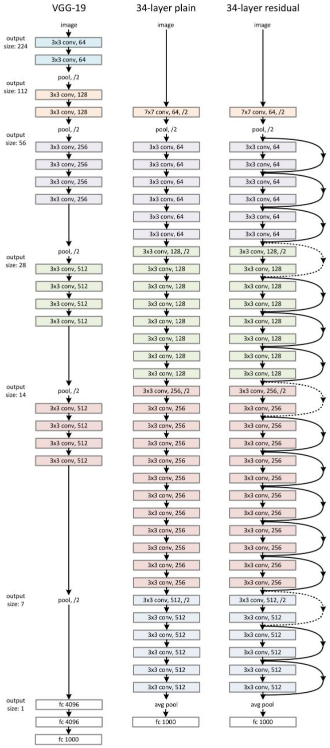
Convolutional Neural Network

Model	Test Accuracy
Inception (Imagenet pretrained)	63
VGG16 (Cifar100 pretrained)	54
VGG16(Chinese characters)	74
VGG16	89
ResNet	95
SNN	93





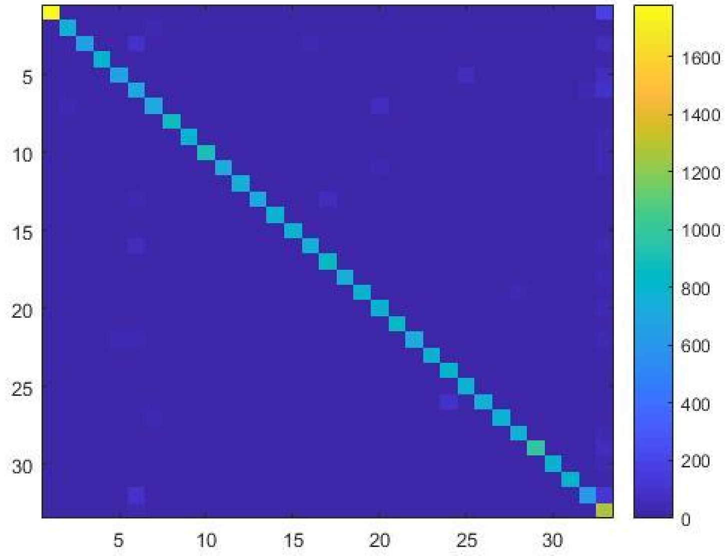
Full Inception module



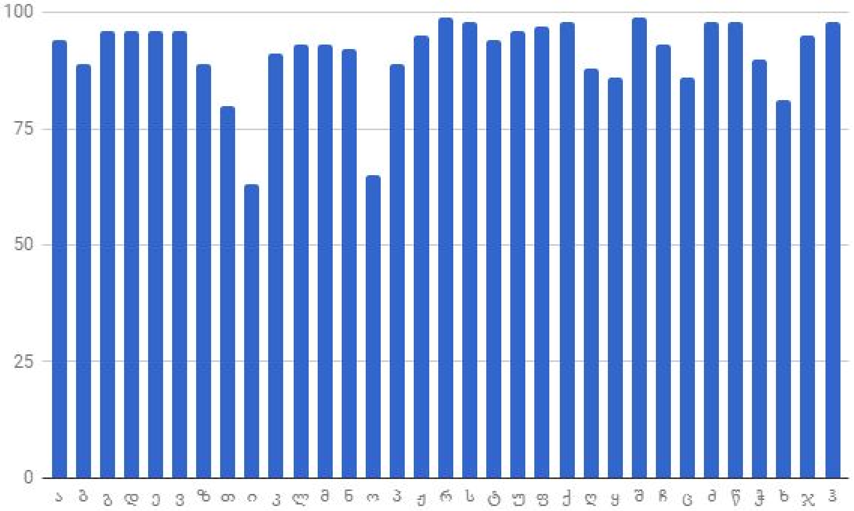
Results

Measured over 66 000 Test samples

Confusion Matrix



Recall

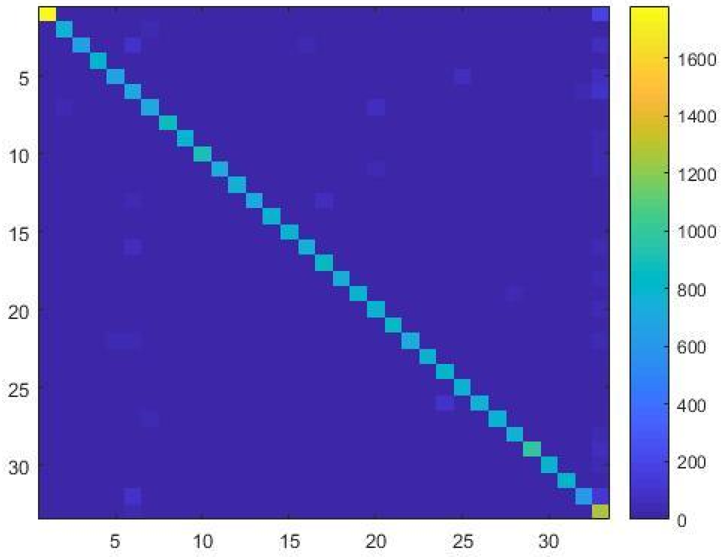


loss: 0.0678 - acc: 0.983 Validation Accuracy - 0.934

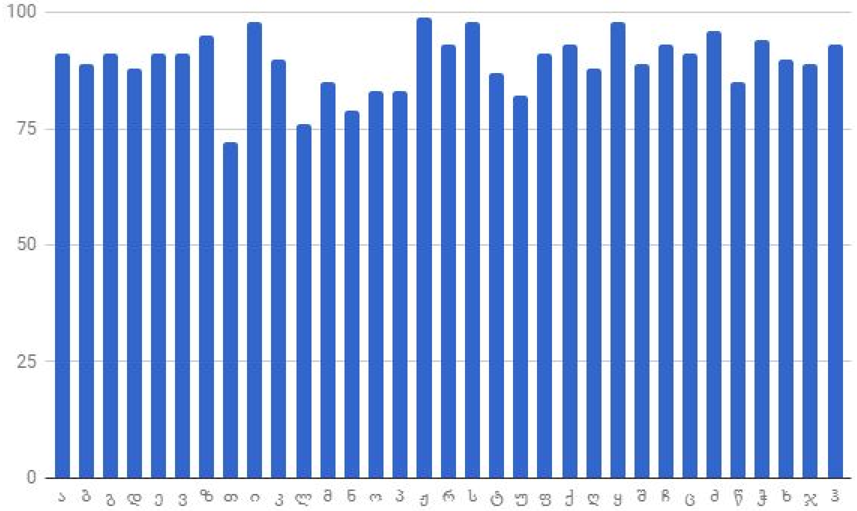
Results

Measured over 66 000 Test samples

Confusion Matrix



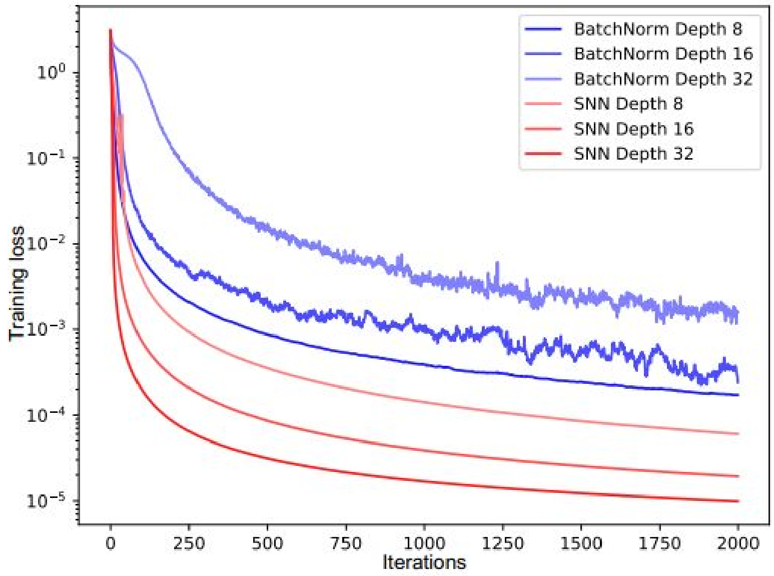
Recall

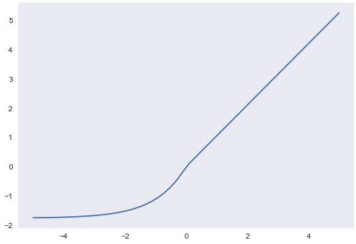


loss: 0.0678 - acc: 0.983 Validation Accuracy - 0.934

Self-normalizing Neural Networks (SNNs)

- Robust to perturbations.
- Learn faster.
- Neuron activations automatically converge towards zero mean and unit variance.
- Do not suffer from high variance.





A neural network is self-normalizing if it possesses a mapping g : Ω → Ω for each
activation y that maps mean and variance from one layer to the next.

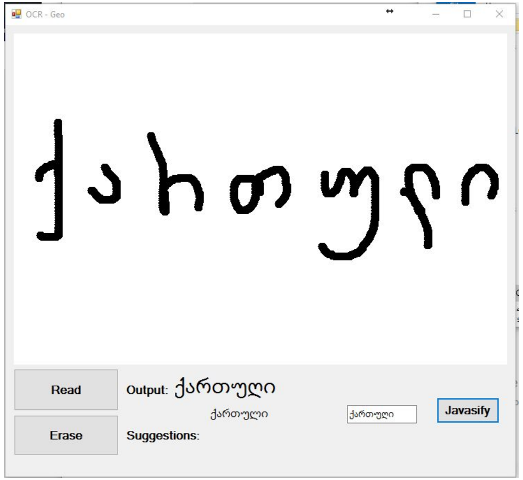
Approximate String Matching

- Merged largest Georgian words dataset with clean and obtained a new dataset with 363202 Unique Georgian Words.
- Symmetric Delete spelling correction algorithm allows for correction of a word with 2 edit distance within 33μs.

Results

- Mobile, Web, Desktop applications for character recognition
- Self Normalizing VGG network with 7% higher accuracy than standard VGG
- Model with ~95% accuracy for single character prediction
- Largest Georgian Handwritten character dataset
- High quality Georgian word dataset

Javasify

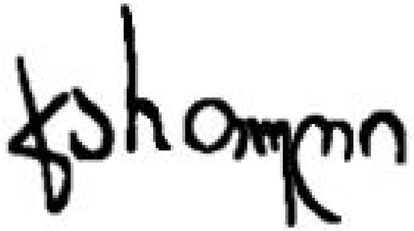


Output: ქართული

ქართული

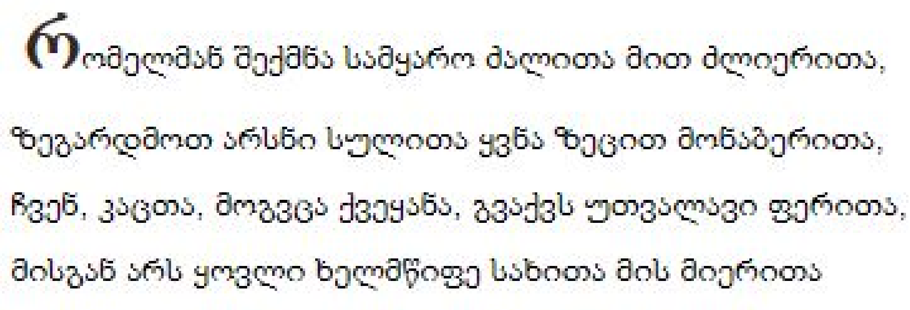
Suggestions:

ქართული



- Generates Georgian handwriting from predicted word.
- Currently it generates handwriting of Ivane Javakhishvili using font created by averaging and filtering bitmaps of existing handwritten characters, generative adversarial network to match handwriting of arbitrary person is in development.

Javasify



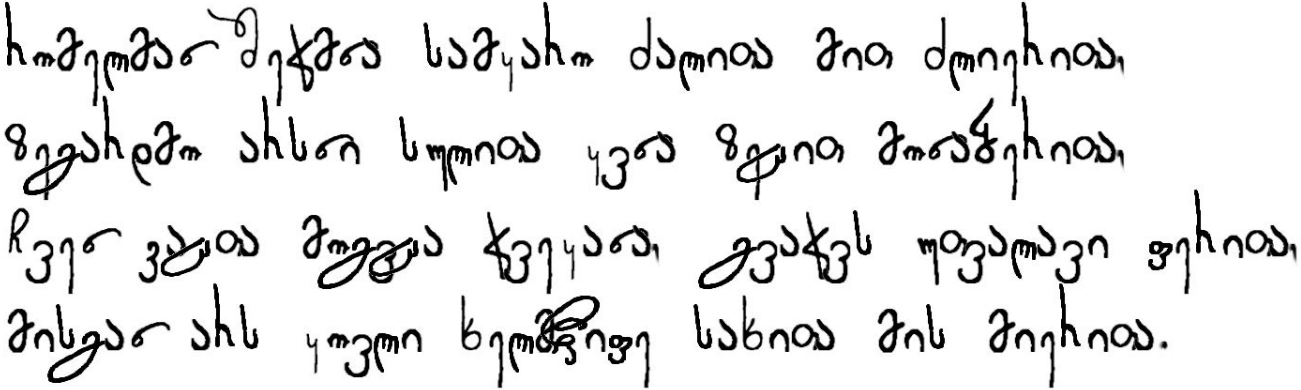
რომელმან შექმნა სამყარო ძალითა მით ძლიერითა,

ზეგარდმოთ არსნი სულითა ყვნა ზეცით მონაბერითა,

ჩვენ, კაცთა, მოგვცა ქვეყანა, გვაქვს უთვალავი ფერითა,

მისგან არს ყოვლი ხელმწიფე სახითა მის მიერითა

↓



რომელმან შექმნა სამყარო ძალითა მით ძლიერითა,

ზეგარდმოთ არსნი სულითა ყვნა ზეცით მონაბერითა,

ჩვენ კაცთა მოგვცა ქვეყანა, გვაქვს უთვალავი ფერითა,

მისგან არს ყოვლი ხელმწიფე სახითა მის მიერითა.

Future Work

- Increase Dataset Size
- Add Recurrent Neural Networks
- Add symbols, and old georgian handwriting

მადლობა ყურადღებისთვის

geohandwrittenOCR@gmail.com

Davit Soselia, Irakli Koberidze,
Sandro jijavadze, Shota
Amashukeli, Giorgi Gigauri,
Levan Shughliashvili