Afro-MNIST

Synthetic generation of MNIST-style datasets for low-resource languages

Daniel J. Wu[1] **Andrew C. Yang**[1] Vinay Prabhu[2]

[1]Stanford University

[2]UnifyID Inc.

Introduction

- Classifying Hindu-Arabic numerals in the MNIST dataset[1] has become the "Hello world" challenge in the machine learning community.
- This task has excited a large number of prospective machine learning scientists and has led to practical advancements in OCR.

MNIST dataset[2]

[1] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[2] Wikimedia Commons.

Introduction

- Work in ML and AI focuses almost exclusively on high-resource languages, which use the Hindu-Arabic numeral system.
- Of over 7,000 languages in the world,[3] the vast majority are not represented in the ML research community.
- There are many alternative numeral systems for which an MNIST-style dataset is not available.

[3] David M. Eberhard, Gary F Simons, and Charles D Fennig. *Languages of the World*. 2019. URL: http://www.ethnologue.com/.

Introduction

- Much of the world's linguistic diversity comes from languages spoken in developing nations.
- There is a wealth of linguistic diversity in the languages of Africa, many of which have dedicated orthographies and numeral systems.
- One notable example is the Ge'ez (Ethiopic) script, which is used to transcribe languages such as Amharic and Tigrinya, spoken by some 30 million people.[4]

ግዕዝ

Ge'ez written in the Ge'ez script

[4] *Ibid.*

The numeral system is the most endangered aspect of any language[5]

[5] Emmanuel Mfanafuthi Mgqwashu. "Academic literacy in the mother tongue: A pre-requisite for epistemological access". In: *Diversity, Transformation and Student Experience in Higher Education Teaching and Learning* (2011), p. 159.

Proposal

- Large amounts of training data for African languages such as these are not readily available.
- But effective neural networks can be trained on highly perturbed versions of just a single image of each class.[6]
- We experiment with creating synthetic numerals that mimic the likeness of hand-written numerals in those writing systems.

[6] Alexey Dosovitskiy et al. "Discriminative unsupervised feature learning with exemplar convolutional neural networks". In: *IEEE transactions on pattern analysis and machine intelligence* 38.9 (2015), pp. 1734–1747.

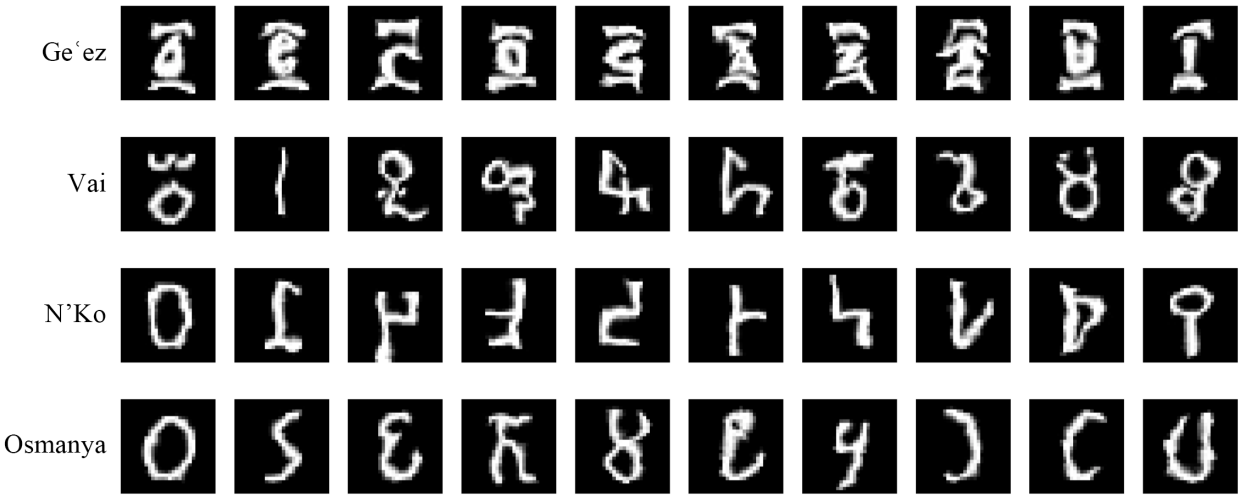
Contributions

- We release synthetic MNIST-style datasets for four scripts used to write Afro-Asiatic or Niger-Congo languages: Ge'ez, Vai, Osmanya, N'Ko[7], which serve as **drop-in replacements** for the MNIST dataset.
- We describe a general framework for resource-light syntheses of MNIST-style datasets.
- These datasets can be found at `https://github.com/daniel-wu/AfroMNIST`.

[7] The Vai, Osmanya, and N'Ko scripts are not in wide use, but nonetheless they can be synthesized using the methods we present.

Methodology

- Generate an exemplar seed dataset for each numeral system from the corresponding Unicode characters.
- Apply series of elastic deformations and corruptions.[8]



[8] We note that, in cases where a limited amount of handwritten data is available, deformations and corruptions can be applied to those examples instead of Unicode exemplars.

Experiments

- We begin by training LeNet-5, the network architecture first used on the original MNIST dataset[9], for numeral classification.

Dataset	Accuracy (%)
MNIST	99.65
Ge'ez-MNIST	99.92
Vai-MNIST	100
Osmanya-MNIST	99.99
N'Ko-MNIST	100

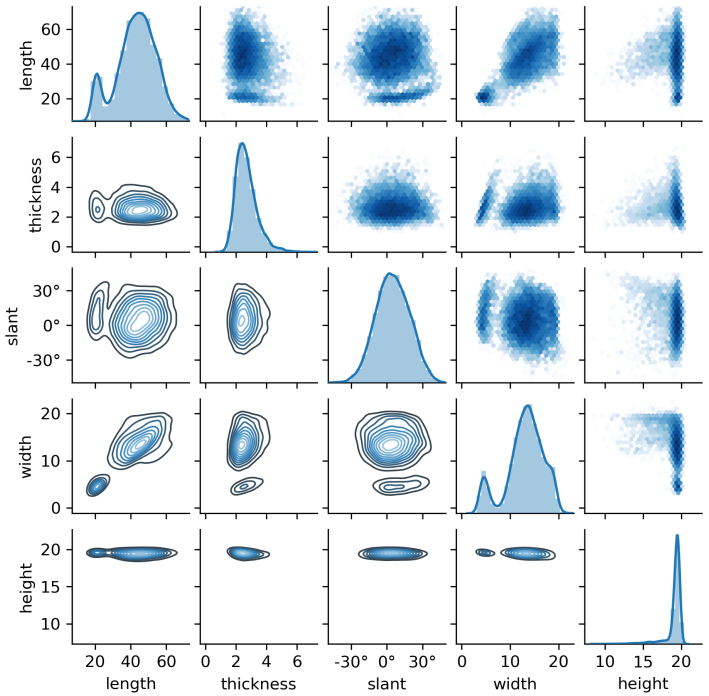
- After testing a LeNet-5 trained on Ge'ez-MNIST on a small dataset of handwritten Ge'ez numerals[10], we found the model achieved only an accuracy of 30.30%.

[9] LeCun et al., "Gradient-based learning applied to document recognition".

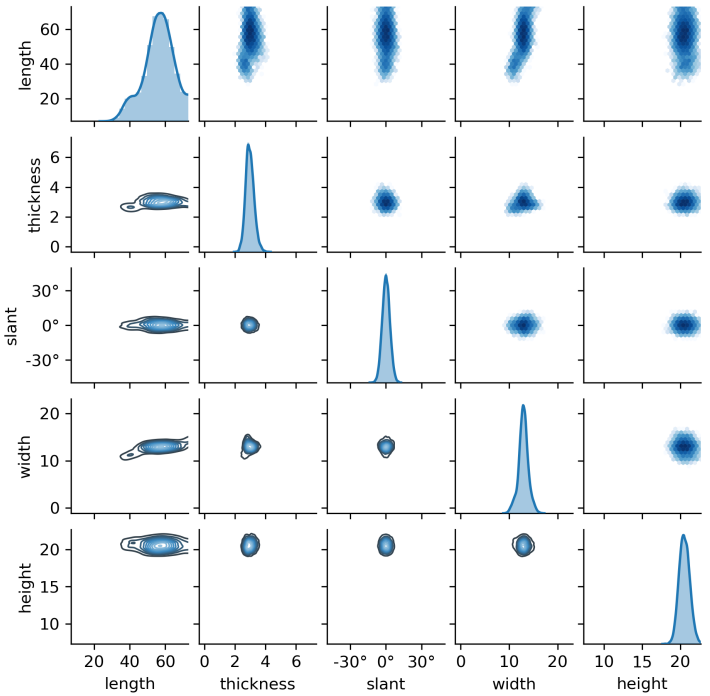
[10] Tesfamichael Molla. *Ethiopian-MNIST*. https://github.com/Tesfamichael1074/Ethiopian-MNIST. 2019.

Experiments

- Morphological comparisons[11] between MNIST and Ge'ez-MNIST show clear differences in variance.



MNIST



Ge'ez-MNIST

[11] Vinay Uday Prabhu. "Kannada-mnist: A new handwritten digits dataset for the kannada language". In: *arXiv preprint arXiv:1908.01242* (2019).

Summary

- Many writing systems, especially those used in developing nations, are underrepresented in the ML community.
- Elastic deformations and corruptions show promise in generating synthetic numeral data, but other methods of creating synthetic digits more similar to handwritten digits ought to be explored as well.
- We expect this benchmark to be a fertile starting point for exploring augmentation and transfer learning strategies for low-resource languages.
- We hope that endeavors such as these help encourage the next generation of diverse ML practitioners to be part of the broader research community.

References

Dosovitskiy, Alexey et al. "Discriminative unsupervised feature learning with exemplar convolutional neural networks". In: *IEEE transactions on pattern analysis and machine intelligence* 38.9 (2015), pp. 1734–1747.

Eberhard, David M., Gary F Simons, and Charles D Fennig. *Languages of the World*. 2019. URL: http://www.ethnologue.com/.

LeCun, Yann et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

Mgqwashu, Emmanuel Mfanafuthi. "Academic literacy in the mother tongue: A pre-requisite for epistemological access". In: *Diversity, Transformation and Student Experience in Higher Education Teaching and Learning* (2011), p. 159.

Molla, Tesfamichael. *Ethiopian-MNIST*. https://github.com/Tesfamichael1074/Ethiopian-MNIST. 2019.

Prabhu, Vinay Uday. "Kannada-mnist: A new handwritten digits dataset for the kannada language". In: *arXiv preprint arXiv:1908.01242* (2019).