10/21/2023

# ECE/CS/ME 539 Introduction to Artificial Neural Networks

# Multi-Language Alphabetical Classifier

#### Connor Braun (cwbraun2@wisc.edu)

#### Elias Cassis (ecassis@wisc.edu)

#### Caleb Federman (cfederman@wisc.edu)

#### Tate Waugh (htwaugh@wisc.edu)

## Overview:

Our objective is to develop a machine learning framework adept at recognizing and classifying characters from diverse alphabets, subsequently identifying the associated language. By employing a convolutional neural network, the system will be trained on a comprehensive dataset of written characters, striving for proficient classification that can accurately discern the alphabet and language. Furthermore, the framework will encompass a component capable of analyzing words, where it leverages the certainty levels of individual character classifications, along with contextual cues like word length and dictionary references, to ascertain the language, and thus the individual characters, of the entire word.

## Background:

The challenge of distinguishing characters from diverse alphabets and pinpointing the corresponding language is a compelling venture within the sphere of machine learning and computer vision, opening avenues in areas like automated translation and multilingual text processing. There have been notable strides, particularly in Optical Character Recognition (OCR) projects, where the focus has been on classifying and identifying individual letters, laying a solid foundation for basic character analysis.

Our project aspires to move beyond the existing frameworks by not only recognizing and classifying individual characters accurately, but extending this recognition to validate the alphabet and language of entire words. This two-step verification aims to bolster the reliability of language identification, especially in scenarios blending alphabetic characters. We're looking to develop a framework employing modern neural network architectures and advanced training methodologies, all backboned by datasets loaded with a wide array of characters from various alphabets both electronic and handwritten.

With an eye on real-world applicability, meticulous data preparation and augmentation are essential to ensure our model's robustness, enabling it to adeptly navigate varied linguistic scenarios. The innovative aspect lies in the fusion of individual character classifications with contextual clues and (potentially) dictionary references to affirm the language of the entire word. This approach aims to contribute a significant stride towards more reliable and holistic multilingual text recognition and analysis.

#### Relevant works related to our implementation:

OCR Systems:

* [Comprehensive overview of OCR](https://ieeexplore.ieee.org/document/9151144#:~:text=Given%20the%20ubiquity%20of%20handwritten,intelligence%2Fmachine%20learning%20tools%20to)1
* [Survey on OCR systems](https://arxiv.org/pdf/1710.05703.pdf#:~:text=A%20Survey%20on%20Optical%20Character,the%20capability%20to%20very%20easily)2
* [OCR System Mechanism](https://link.springer.com/chapter/10.1007/978-3-319-50252-6_2#:~:text=Optical%20character%20recognition%20%28OCR%29%20,increasing%20attention%20in%20both)3
* [OCR for text recognition](https://ieeexplore.ieee.org/document/9935961/#:~:text=From%20this%20comes%20a%20need,accuracy)4

Cross-Language Character Recognition

* [Cross-Lingual Handwritten character recognition](https://www.sciencedirect.com/science/article/pii/S0167865522001490#:~:text=,31%2C%2034)5
* [Cross-Lingual Learning for Text Processing](https://www.sciencedirect.com/science/article/pii/S0957417420305893#:~:text=%23%20%E3%80%901%E2%80%A0Cross,it%20was%20not%20possible%20previously)6
* [Script Recognition in Multi-Script Documents](https://dl.acm.org/doi/10.1145/3396167#:~:text=This%20literature%20examines%20the%20Script,handwritten%2C%20Nandinagari%2C%20and%20Hebrew%2C%20wh)7
* [Cross-Lingual Text Image Recognition](https://www.researchgate.net/publication/361356717_Cross-Lingual_Text_Image_Recognition_via_Multi-Hierarchy_Cross-Modal_Mimic#:~:text=Cross,Institute%20of%20Automation%20of%20Chinese)8

Advanced Neural Network Architectures

* Our lecture notes from class

## Statement of Work:

#### Datasets:

* Our dataset will be a combination of datasets from the following:
* [English Alphabet Dataset](https://www.kaggle.com/datasets/mohneesh7/english-alphabets) (Latin Script)
  + Kaggle source: Handwritten characters from the English alphabet.
* [Greek Alphabet Dataset](https://www.kaggle.com/datasets/katianakontolati/classification-of-handwritten-greek-letters) (Greek Script)
  + Kaggle source: Greek characters of different resolutions.
* [Russian Alphabet Dataset](https://www.kaggle.com/datasets/tatianasnwrt/russian-handwritten-letters) (Cryllic Script)
  + Kaggle source: Handwritten Russian characters.
* [~~Kurdish Alphabet Dataset~~](https://www.kaggle.com/datasets/rebinma/central-kurdish-handwritten-characters) ~~(Arabic Script)~~
  + Kaggle source: Handwritten Kurdish characters.
* [Arabic Alphabet Dataset (Very extensive - good for training)](https://www.kaggle.com/datasets/insafbenlamari/arabic-letters) (Arabic Script)
  + Kaggle source: Handwritten Arabic characters.
* [Georgian Alphabet Dataset (Optional - not best format)](https://www.kaggle.com/datasets/alexandertropin/georgian-letters-photo-database) (Georgian Script)
* [Omniglot Extensive Alphabet Library](https://www.omniglot.com/index.htm)
  + Many character databases.

#### Method:

##### Process:

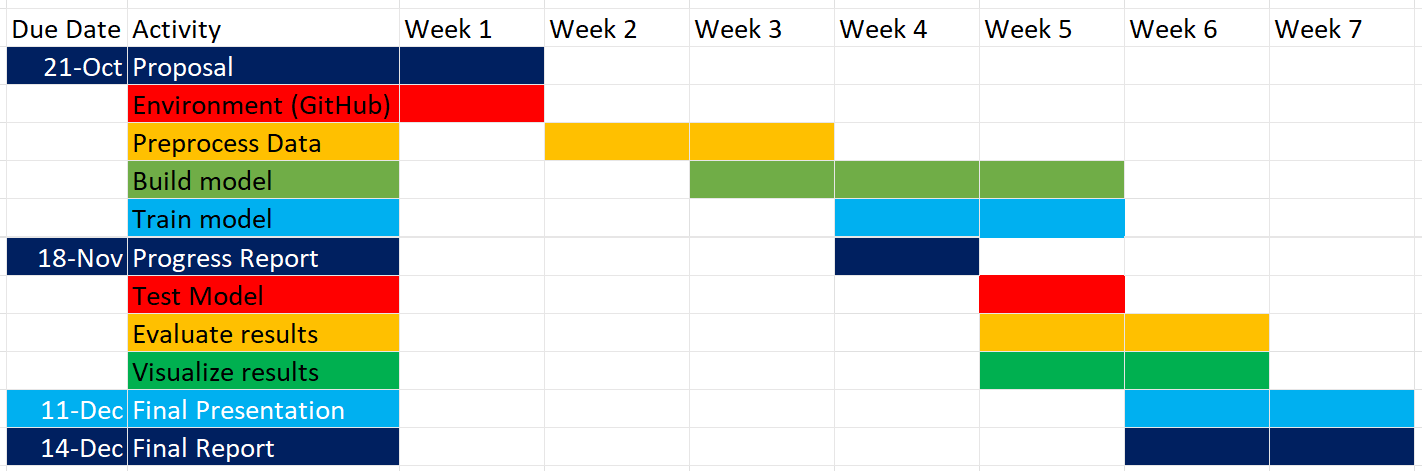
* Utilize a 1-hot neural network classifier or a KNN to obtain the probabilities/classifications that a character is any given character in our combined language dataset.
* Utilize these probabilities/classifications, and the language associated with each predicted character, in conjunction with each other to solidify the determination of each character. This will use various contextual clues, such as the probability a word is a given language, the predicted characters, the length of the word, as well as potentially dictionary references of known words for each language.
* This model could then potentially be scaled to attempt to classify an unfinished word based on what it can currently determine about the word as well as its stored list of words. (i.e. autocomplete)
* The goal of this project is to utilize machine learning in a two step process to create a model with the potential to be scaled up as a multi-language detection/prediction tool.

#### Outcome and Performance Evaluation:

* The result of our project should be able to accurately predict the characters, language, and potentially word of a given handwritten/typed input. Utilizing contextual clues from words, it should be able to do this even if a given character is very similar or essentially the same as another from a different language. Possibly, the project should also be able to predict what a completed word would be given a subset of that word.
* The accuracy of this model will be computed via F-score and Cross Entropy Loss to directly compute the accuracy of predicted letters and possibly predicted words.

## Project Plan:

1. Environment
   * For this project, we will be writing our code in iPython notebooks hosted on Google Colab. The utilization of Google Colab will allow us to work collaboratively, whereas a local or virtual environment would make it more difficult to work together. Colab files will eventually be uploaded to GitHub for version control. GitHub will also host our backlog of tasks and roadmap.
   * You can access the GitHub repository [here](https://github.com/calebfederman/handwritten-alphabet-language-classifier).
2. Timeline
   * The proposed timeline of our implementation is outlined in the Gantt chart below.



## References:

1 J. Memon, M. Sami, R. A. Khan and M. Uddin, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," in *IEEE Access*, vol. 8, pp. 142642-142668, 2020, doi: 10.1109/ACCESS.2020.3012542.

2 Islam, Noman, et al. “A Survey on Optical Character Recognition System.” Rnal of Information & Communication Technology-JICT, vol. 10, no. 2, Jan. 2016, pp. 4–7. jms.ilmauniversity.edu.pk/index.php/JICT/article/download/641/362.

3 Chaudhuri, A., Mandaviya, K., Badelia, P., Ghosh, S.K. (2017). Optical Character Recognition Systems. In: Optical Character Recognition Systems for Different Languages with Soft Computing. Studies in Fuzziness and Soft Computing, vol 352. Springer, Cham. <https://doi.org/10.1007/978-3-319-50252-6_2>

4 R. Avyodri, S. Lukas and H. Tjahyadi, "Optical Character Recognition (OCR) for Text Recognition and its Post-Processing Method: A Literature Review," *2022 1st International Conference on Technology Innovation and Its Applications (ICTIIA)*, Tangerang, Indonesia, 2022, pp. 1-6, doi: 10.1109/ICTIIA54654.2022.9935961.

5 Nirmala S Guptha, V. Balamurugan, Geetha Megharaj, Khalid Nazim Abdul Sattar, J. Dhiviya Rose,

Cross lingual handwritten character recognition using long short term memory network with aid of elephant herding optimization algorithm,

Pattern Recognition Letters,

Volume 159,

2022,

Pages 16-22,

ISSN 0167-8655,

https://doi.org/10.1016/j.patrec.2022.04.038.

(<https://www.sciencedirect.com/science/article/pii/S0167865522001490>)

6 Dhivya S and Usha Devi G. 2021. Study on Automated Approach to Recognize Characters for Handwritten and Historical Document. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 20, 3, Article 37 (May 2021), 24 pages. https://doi.org/10.1145/3396167

7 Dhivya S and Usha Devi G. 2021. Study on Automated Approach to Recognize Characters for Handwritten and Historical Document. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 20, 3, Article 37 (May 2021), 24 pages. https://doi.org/10.1145/3396167

8 Chen, Zhuo & Yin, Fei & Yang, Qing & Liu, Cheng-Lin. (2022). Cross-Lingual Text Image Recognition via Multi-Hierarchy Cross-Modal Mimic. IEEE Transactions on Multimedia. PP. 1-13. 10.1109/TMM.2022.3183386.