“Handwritten English word recognition using deep learning based object detection architecture”

Background/Motivation:

Although as a society we have increasingly grown our communications online using typed methods there are still many areas where handwritten text is leveraged. With the increased usage of technology handwritten text documents are subjected to new standards such as becoming digitally searchable or being translated into digital text.

Most popular method of converting handwritten text into a machine editable form using a three-step process:

1) Image Segmentation: Segmentation from document images to word images
2) Handwritten word recognition: Handwritten words are recognized and converted into word images
3) Assembly: Words are assembled to understand entire document image

Traditional approaches to handwritten word recognition include segmentation and identification of characters then conversion into recognized words. Handwritten words are complex due to their interconnectedness between characters; interconnectedness occurs in cursive but also occurs commonly in non-cursive writing. Segmentation ambiguity tends to be a main contributing factor for this approach and corrections for slant and skew can reduce the under-/over-segmentation.

Due to this complexity most solutions skip this step by using one of the following methods:

- Hidden Markov Models (HMMs)
- Recurrent neural networks (RNNs) commonly paired with
    - Bi-directional long short-term memory (BLSTM) or Multi-dimensional long short-term memory (MDLSTMs)
    - Connectionist temporal classification (CTC) loss

These models are lexicon driven where the recognized string must form a valid word from a defined dictionary within the network model, which requires a significant amount of training to have acceptable accuracy. Producing these samples would be significantly tricky due to English having 50 K words and considering case and when tenses this number increases.

The sampling cost would be significantly decreased if character-based recognition was leveraged over word.

Methods:

1) Object Detection: You Only Look Once (YOLOv3)
    a. Character spotting and classification occurs simultaneously
2) Dataset: IAM dataset for Latin Characters

YOLOv3 model learns the character boundaries and the associated classification during the training phase. When preparing the training data, they assured all images were 416 X 416 and any augmentation was applied prior. The model is fed these images and learning proceeds by finding the loss between the predicted and ground truth. The model weight updates as more batches are processed. Additionally, they added validation steps to avoid overfitting the model. The validation step calculates the accuracy of model during model training.

Training data included 700 randomly selected samples that were writer independent from IAM dataset and another 800 writing samples from an in-house dataset to improve alphabet coverage. The dataset was then manually bound using AlexeyAB GUI. The validation set was randomly selected from the 1,500, therefore their model was trained using 1,200-word images.

Performance metrics:

- Word Error Rate (WER): Percentage of total test set word samples that were incorrectly recognized (NumberOfWordsIncorrectlyRecognized / TotalNumberOfWordsPresentInTheTestSet)
- Character Error Rate (CER): Percentage of total characters from entire test set that were incorrectly recognized (NumberOfErroneouslyRecognizedCharacters / TotalNumberOfCharactersPresentInTheTestSetWords)

Connection to Other Work/Significance of Work:

Their method was inspired by a Visual Geometry Group (VGG16) Faster Region-based Convolutional Neural Network (R-CCN) used to identify Bangla which is a connected script. This helped them compensate for the under-/over-segmentation.

To test their performance, they used methods from previous literatures, one using BLSTM and another using the same BLSTM with fuzzy membership function and used the same data sets to train, test those models as well as run it against the IAM dataset in its entirety. Their results were the best when it comes to error rates.

Relevance to Capstone Project:

My capstone project will need to be able to detect handwritten language. I need to determine which methodologies I need to leverage for this detection. This introduced me to two different approaches breaking down a document image into word images vs going a step further to break them down by characters. Additionally, it helped me identify tools I may need to perform my capstone project. Lastly, it referenced many other approaches that I can further research.

Works Cited

Mondal, R., Malakar, S., Barney Smith, E.H. *et al.* Handwritten English word recognition using a deep learning based object detection architecture. *Multimed Tools Appl* **81**, 975–1000 (2022). https://doi.org/10.1007/s11042-021-11425-7