Follow the Diagonals: Finding String Matches through Matrix Operations

Rome Duong, Professor Pavel Oleinikov
QAC Summer 2022, Wesleyan University

Introduction

- A common technique for extracting similarity between texts is a dot plot where pairwise comparisons would produce a diagonal line of matches. While it was originally used in homology for genetics similarity detection, its usage has been expanded to programs and languages detections (Helfman, 1996).
- Text patterns could be analyzed by breaking the text chunks into various tokens through dynamic programming approaches such as Global Alignment. This text manipulation approach can be run on either the CPUs or the GPU.
- An alternate approach is through image analysis where the tokenization of texts is used for "an image processing pipeline" (Abdul-Rahman et.al, 1996). The image processing pipeline is combined with the text alignment processes to create a pixel-based map for visual analysis.
- We are interested in creating a pixel-based model while utilizing image processing tools such Convolutional Neural Networks and Gaussian Filters to improve text similarity detections. In addition, this gives an opportunity to compare the runtime of pixel-based model with the Global Alignment approach on both CPUs and GPU.

Methods

Data

- Closed-caption transcripts of WSFB Connecticut News that were queried on Google BigQuery. The news coverage was from 2020-03-01 to 2020-03-05.

Procedures

- To prepare the data for Matrix operations, the text segments were split into chunks in form of arrays with the fixed length of shards (N=100) and utilizing bigrams.
- Tokens were converted into Hash codes.
- Input arrays were reshaped into Nx1 and 1xN matrices.
- After matrix multiplication, we select the pixels with matches and apply 2D convolution with Gaussian filter to remove single pixels
- Arrays of matches were extracted from the diagonals of matrices to construct data frames that showed the position of matches between the texts.
- All operations were programmed in TensorFlow to run as Keras models with batched data.

Results

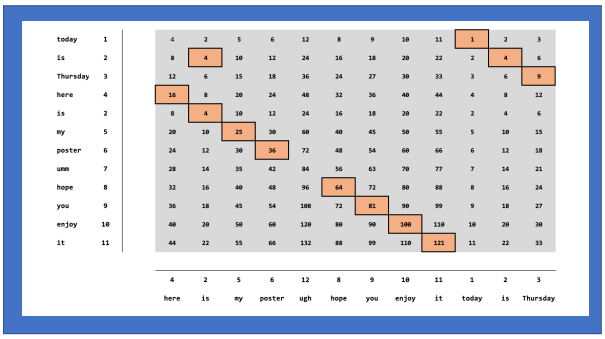

Figure 1. Matrix of matches is produced by multiplying Nx1 and 1xN matrices. Positions with matching tokens contain the square of a hash value from the 1xN matrix (or Nx1).

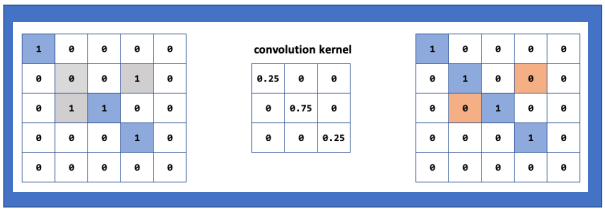

Figure 2. Image convolution followed by thresholding can eliminate isolated pixels and fill one-gap holes.

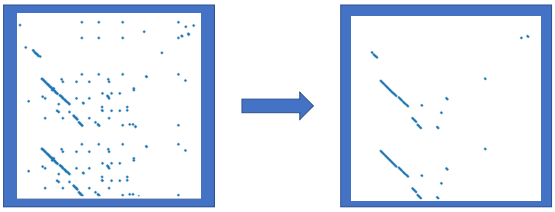

Figure 3. Dotplot of two TV news texts before and after the cleanup via convolutions. Each document is 500 tokens. One document is a broadcast from 5 AM, the other – from 5:30 AM. The diagonal lines are matching substrings – repeated phrases.

Comparisons between Diagonal Matrices Model vs Global Alignment

	Diagonal Matrices Model	Global Alignment
CPU Only	7.23189969	8.584862315
GPU Accelerated	5.975868338	8.38916621

Figure 4. Running the diagonal matrices model on the text segment that is tokenized with the token amount of 500, both the CPU and GPU acceleration tests illustrated faster time than Global Alignment approach.

Discussion

- Global alignment returns only one matching string, while our method returns all matching strings.
- The use of the diagonal matrices approaches in conjunction with Convolutional Neural Networks runs faster than Global alignment.
- The model is used to find news stories in the stream of closed caption data because local TV stations repeat the stories several times.
- The algorithm operates in an unsupervised manner.

Abdul-Rahman, A., et al. "Constructive Visual Analytics for Text Similarity Detection." Computer Graphics Forum, vol. 36, no. 1, 19 Feb. 2016, pp. 237–248, 10.1111/cgf.12798. Accessed 2 Oct. 2019.
Helfman, Jonathan. "Dotplot Patterns: A Literal Look at Pattern Languages." Theory and Practice of Object Systems, 1 Jan. 1996,