Word Spotting Using Convolutional Siamese Network

Berat Kurar Barakat, Reem Alasam, Jihad El-Sana
Department of Computer Science
Ben-Gurion University of the Negev

Introduction

Spotting words from a possibly large range of vocabulary is challenging. Training a discriminative model requires many positive samples per word. We present a word spotting method for historical document images using convolutional siamese network. This method can spot

1. Words with varying writing styles and backgrounds
2. Out of vocabulary words

Method

The method works on segmented word images using query-by-example. It is based on training a discriminative model to rank the similarity of two input images regardless of being in or out of vocabulary. On the other hand, previous methods for the same problem are based on using word embedding.

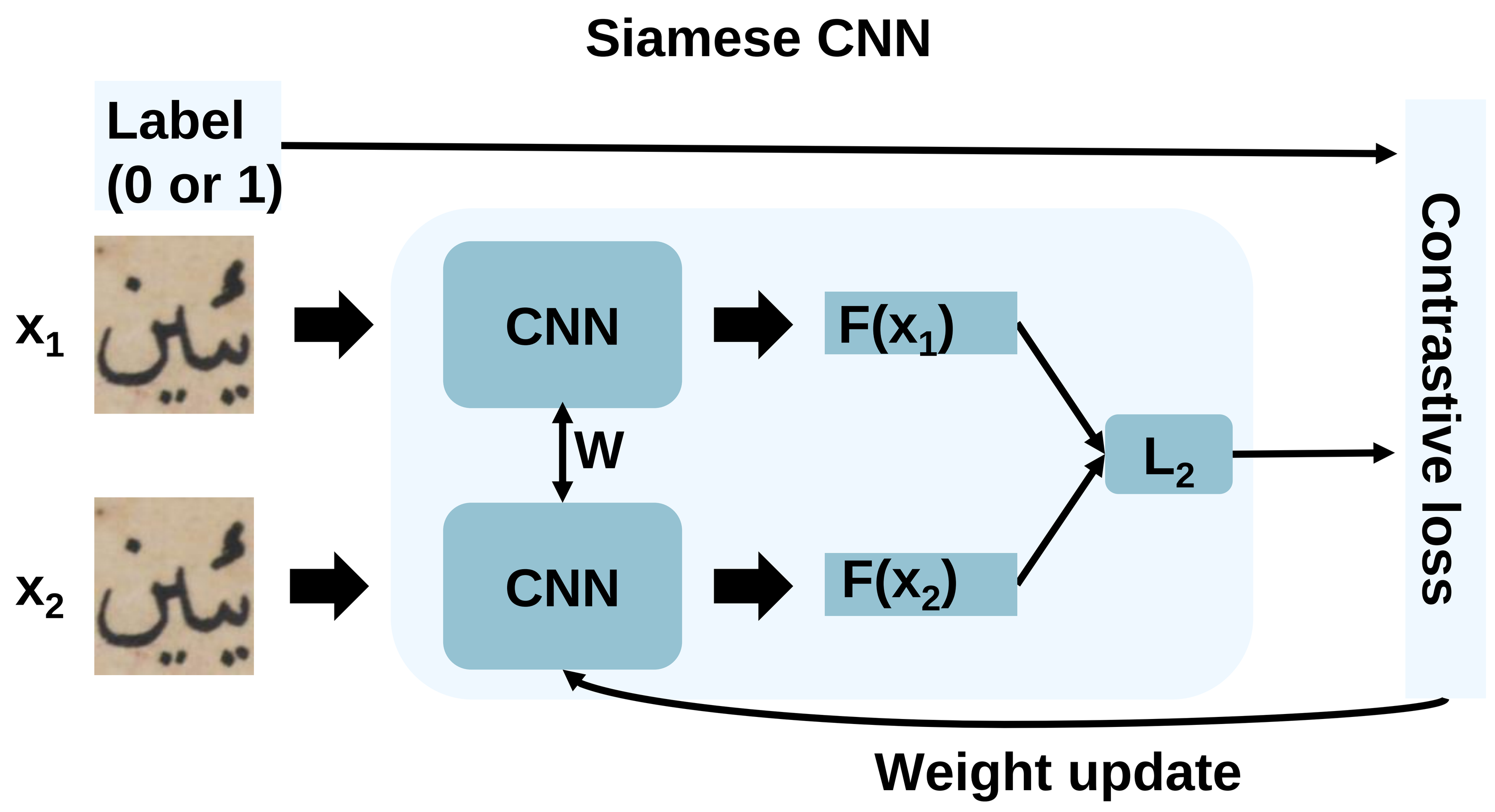

Figure 1. A siamese network consists of two Convolutional Neural Network (CNN) with same architecture and weights (W)

CNN is trained to map pixels into linear space in which $\mathbf{L_2}$ distance of the two feature vectors is:

- close if the inputs are same and
- far if the inputs are different

How does contrastive loss work?

Contrastive loss pulls the similar points together, pushes the different points apart.

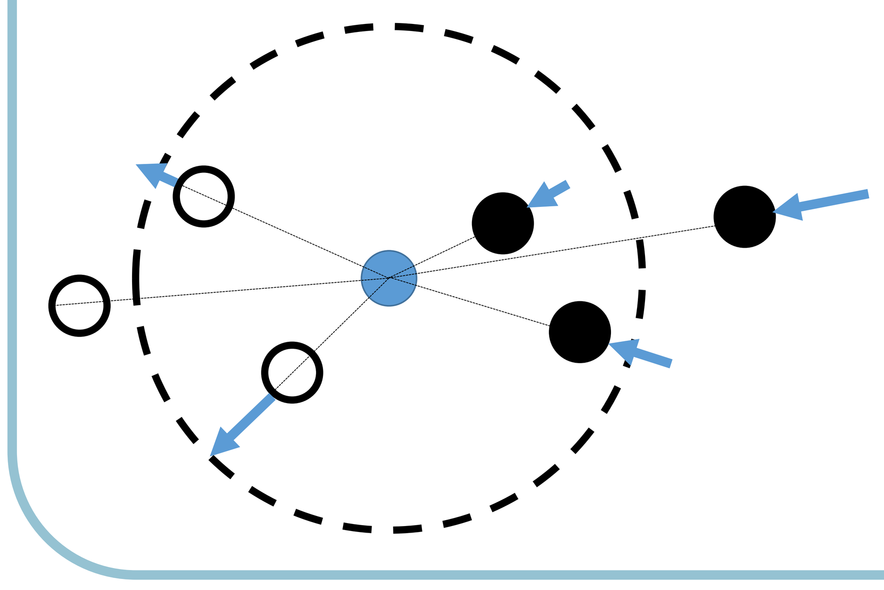

Figure 2. Solid circles represents similar, hollow circles represents dissimilar points to the point in the center.
Forces acting on the points are shown in blue arrows with length gives the strength of the force.

Results

George Washington (GW) dataset [1] contains 929 classes with 2372 words. The model was trained using 5 fold cross-validation and tested on completely OOV. It achieves mAP value of 0.49 whereas the work of Rodriguez and Perronin [2] achieves **0.53**.

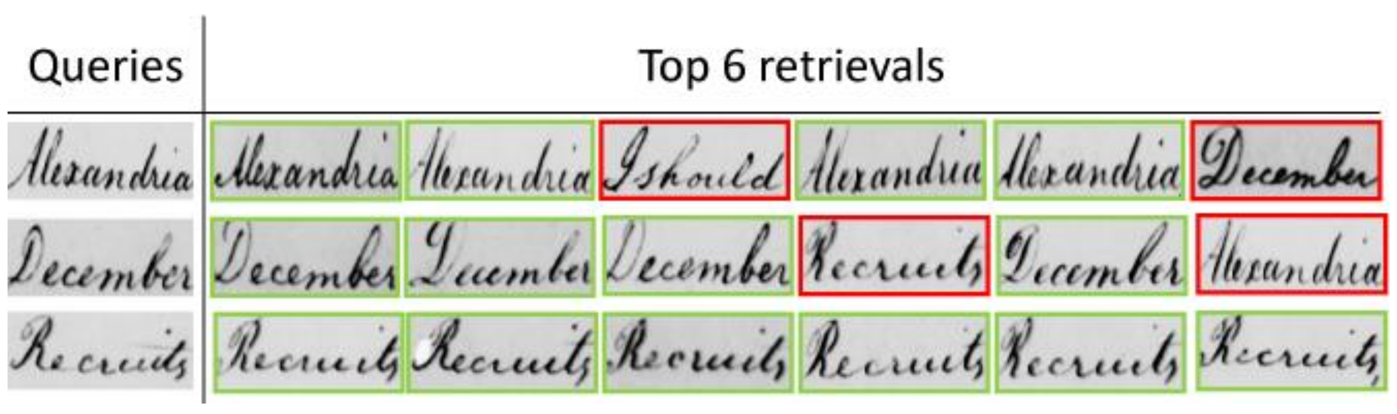

Figure 3. Qualitative results on GW dataset.

Results

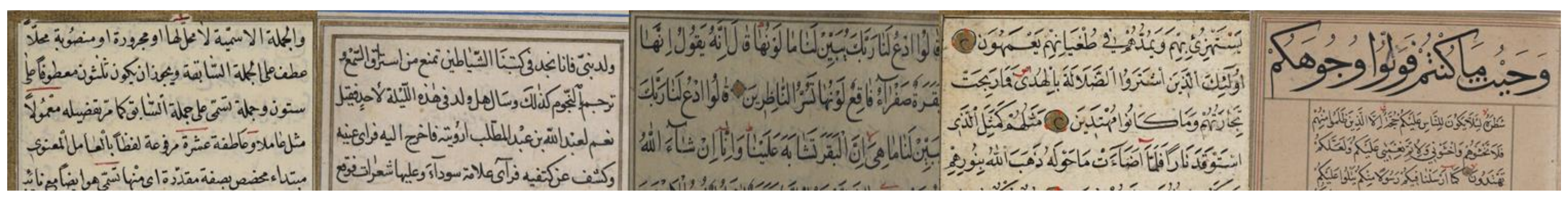

Figure 4. VML dataset [3] consists of 5 books with varying backgrounds and writing styles.

Training and testing

The model was trained only on book1. Train set contains 100 classes each with 3 samples. Validation set contains 20 classes each with 10 samples.

Trained model was tested on 5 books individually and on a set of their combination. Each individual test set contains 21 classes each with 100 samples. Combined test set contains 10500 words with 58 classes.

Figure 5. Vertical direction shows inter-book variation, horizontal direction shows in-book variation.

Figure 6. Comparison of proposed method with the work of Kassis and El-Sana [4]. Green bars show their results, blue bars show our results.

Table 1. mAP and P@k values of the model trained on 500 words from book1 and tested on 5 books of VML dataset.

Book#	mAP	P@1	P@2	P@3	#Queries	#Labels	#OOV
Book1	0.91	1.0	1.0	1.0	2100	21	21
Book2	0.75	1.0	1.0	1.0	2100	21	14
Book3	0.82	1.0	1.0	0.98	2100	21	16
Book4	0.75	1.0	1.0	0.98	2100	21	15
Book5	0.80	0.95	0.95	0.95	2100	21	16
All	0.62	0.95	0.95	0.93	10500	58	47

Conclusion

The method is general and can be applied to historical documents in different languages. Convolutional network makes the method robust to varying writing styles and backgrounds. Discriminative model can be used to spot OOV words. As the number of OOV words increases performance decreases. That's why results on VML are more successful than results on GW dataset.

References

[1] https://ciir.cs.umass.edu/download
[2] J. A. Rodriguez-Serrano and F. Perronnin, "A modelbased sequence similarity with application to handwritten word spotting," IEEE transactions on pattern analysis and machine intelligence, 2012.
[3] https://www.cs.bgu.ac.il/vml/
[4] M. Kassis and J. El-Sana, "Word spotting using radial descriptor graph," in Frontiers in Handwriting Recognition (ICFHR), 15th International Conference on. IEEE, 2016.