A thousand word images are worth a word

Berat Kurar Barakat, Tan Lu, and Ann Dooms

Vrije University of Brussels

25.11.2021

1 Rationale and positioning with regard to the state-of-the-art

1.1 Handwritten document image analysis is not accurate

The digital era has made available large quantities of handwritten document images through the Internet. However, the plenty of information conveyed by the text in these document images remains mostly inaccessible. End users envisage to search images of handwritten documents through ASCII text queries, but present document image analysis technology is still far from providing accurate transcripts for handwritten document images.

1.2 What is handwritten document image analysis?

Handwritten document image analysis consists of page segmentation, text line segmentation, word segmentation and text recognition stages. Page segmentation is the process of detecting homogeneous regions of handwritten document images. The detected regions are labeled as either text regions or non-text regions. Page segmentation of handwritten documents poses problems such as arbitrarily shaped complex layout, multiple font types and sizes in a single document image. Text line segmentation is the process of detecting the regions that contain a single text line. Text line segmentation of handwritten documents is particularly difficult because of multiple skew directions, touching and overlapping text lines, multiple font types and sizes, crowded diacritics, and cramped text lines. Word segmentation is the process of detecting the regions that contain a single word. Word segmentation of handwritten documents is challenging because of alternating gap sizes among and within the words. Text recognition is the process of converting the handwritten text present in a segmented word image to the ASCII text. Text recognition suffers from image degradation due to stains and inks, multiple font types and sizes, and cursive writing.

1.3 Related work

In the early days of handwritten document image analysis, the researchers use hand crafted features to solve handwritten document structure and text recognition problems. Since then the AlexNet [11] achieved a top-5 test error rate almost halved the error rate of the second best entry that uses hand crafted features, numerous deep learning based methods have been proposed for handwritten document image analysis tasks. In terms of deep learning for page segmentation the most effective methods are based on unsupervised Convolutional Neural Network (CNN) [3,34,6], supervised CNN [2,18] and Fully Convolutional Network (FCN) [35,23,16]. Apart from page segmentation, state of the art results for text line segmentation has been achieved using unsupervised CNN [13,15] and FCN [33,25,14,20,7,12]. Deep learning methods are not so common for word segmentation yet an interesting direction. It has started to be explored using CNN, Long Short Term Memory (LSTM) and Connectionist Temporal Classification (CTC) [22] and YOLO [5]. Given the segmented word images, CNN feature extractors have been observed to work properly for text recognition [27,9,29,30,31].

1.4 Rationale

A deep learning method is made of data and model. Recent literature on the handwritten document image analysis has developed a lot of progress over improving the model. These works frequently bring into attention that it would be useful to improve the data. It is therefore a significant part of the works included data augmentation [33,29,23,31,7], synthetic data generation [26,9,10,24,8,21] or transfer learning [1,16,27,10,28,20] to improve the data. The principle of improving data using pseudo methods was observed to improve the performance. However the significant performance improvement [32] prospected from a large scale handwritten document image dataset could not yet be seen on handwritten document image analysis tasks. Hence a thousand word images are worth a word!

2 Scientific research objectives

2.1 Hypothesis

The hypothesis of the proposed research is to create a large scale handwritten document image dataset for significantly improving the accuracy of deep learning algorithms on document image analysis tasks. We propose to work on data to improve the performance although there is value to improve the model as well. But lots of document image analysis literature have already explored to improve the model rather than the data whereas the performance of deep learning algorithms increases logarithmically based on the size of the data [32,17].

2.2 Objectives

A lot of literature in handwritten document image analysis have explored different deep learning models. The researchers now know to achieve baseline results using the publicly available latest deep learning models. They also know to perform improvements by hyperparameter tuning. But the size of the handwritten document image analysis datasets has remained constant. The work proposed here has the objective to perform bigger improvements by constructing a large scale handwritten document image analysis dataset. Two important aspects of this objective are constructing the dataset and demonstrating the performance improvement.

Construct a large scale document image dataset We aim at constructing a dataset at the scale of 1M segmented word images and their labels. The size is intuited by the ImageNet dataset. The dataset is intended to contain ground truth for all stages of the handwritten document image analysis that are page segmentation, text line segmentation, word segmentation and text recognition. To our knowledge this will be the largest handwritten document image dataset in terms of the number of images and ground truth levels. The present handwritten document image analysis datasets are at a scale of tens of thousands [19] besides none of them provides all the ground truth levels together for page segmentation, text line segmentation, word segmentation and text recognition. Two important aspects of dataset construction are collecting the raw handwritten document images and labeling the ground truths.

1. We want to collect the raw handwritten document images from the national scientific library of Belgium (KBR). The number of required document images will be calculated approximately to gather the 1M segmented word images. We will collect the documents that are written in Latin alphabet because the prospective labelers are likely to be able to read the Latin alphabet.

2. We want to use a systematic and repeatable way to label the dataset. Accordingly data is split into a number of sets. At the beginning, human labelers label the first set of data and produce its ground truth. We train the model on the first set of data using its ground truth. The trained model is tested on the second set of data to produce its predicted truth. The predicted truth is validated by the human labelers to produce the ground truth of the second set of data. The ground truth of the second set of data is aggregated with the ground truth of the first set of data. This cycle continues until the ground truth for all the sets of data is produced. The predicted truth reduces the workload on the human labelers and shortens the labeling time.

Demonstrate performance improvement on handwritten document image analysis We aim at demonstrating performance improvement for all stages of the handwritten document image analysis that are page segmentation, text line segmentation, word segmentation and text recognition. Performance improvement is to be demonstrated by graphing the relationship between the dataset size and performance, presenting new state of the art results on existing benchmarks and organizing an international challenge.

1. We want to figure out the relationship between the dataset size and the performance. This relationship has been investigated on computer vision tasks [32] but not on handwritten document image analysis tasks because of the limited dataset sizes.

2. We want to present new state of the art results for all stages of the handwritten document image analysis that are page segmentation, text line segmentation, word segmentation and text recognition on the present handwritten document image analysis benchmarks using the models learned from the large scale handwritten document image dataset. Usually ImageNet [4], a dataset of 1M labeled natural scene images is used to pretrain models for handwritten document image analysis tasks. Since the domain of ImageNet is very different than handwritten text domain only the early layers of the pretrained model are useful for handwritten document image analysis tasks. This is because the earlier features of a model contain more generic features that are useful to many tasks, but later layers of the model becomes specific to the dataset.

3. The final objective is to organize an international challenge on the large scale handwritten document image analysis dataset for advancing handwritten document image analysis research by a collective effort. This will also increase reputation of the dataset and make it a popular benchmark for handwritten document image analysis tasks.

3 Research methodology and work plan

Work package 1: Collect handwritten English document images (3 months) Objectives: To collect the raw handwritten English document images from libraries. The number of required document images will be calculated approximately to gather the 1M segmented word images. We will collect the documents that are written in English because the prospective labelers are likely to be able to read English.

Tasks: Task 1.1 - Identify the approximate number of handwritten document images required to extract 1M segmented word images.

Task 1.2 - Identify the libraries that curate English digital collections with open access. - Download the required number of document images.

Risks: - The required number of English handwritten document images with open access is not available.

Contingency plan: - Identify the libraries that curate English handwritten document images with copyright legislation. Propose the authorized people to collaborate the research by giving access rights to the copyrighted English handwritten document images. - Identify the libraries that curate open access digital collections written in Latin alphabet.

Deliverable: A raw dataset of handwritten English document images that contain approximately 1M words.

Work package 2: Label the dataset at a single ground truth level (4 months) Objectives: To label the dataset at a single ground truth label. Dataset is split into a number of sets. Human labelers label the ground truth or validate the predicted truth of a set of data. A model is trained on the present ground truth and used to produce the predicted truth.

Tasks: Task 1.1 - Identify the human labelers.

Task 1.2 - Identify the number of pages for each cycle so the accuracy of predicted truths will be maximum while the validation time of human labelers will be minimum per cycle.

Task 1.3 - Identify the open access annotation tool to be used. - Define the labeling convention. - Train the labelers about the annotation tool and labeling convention. - Labelers label the dataset using the annotation tool and according to the labeling convention. - Download the label files.

Task 1.4 - Implement the model that will be used for producing the predicted truth. - Use the trained model to produce the predicted truth.

Risks: - Labels are not consistent. - Some document images do not contain handwritten English text.

Contingency plan: - Train the labelers second time about the labeling convention. - Gather more handwritten English document images from the libraries.

Deliverable: A labeled dataset of handwritten English document images that contain approximately 1M words.

Work package 3: Figure the incremental performance by the handwritten document image dataset. (1 months) Objectives: To figure out the results of model performance with the increase in the size of the dataset in each iteration.

Tasks: Task 1.1 - Implement the model, tune its hyperparameters. - Make the experiments to figure the graph of performance by increased dataset size.

Risks: - The performance does not increase by dataset size.

Contingency plan: - Review the labels by a consensus check. Consensus check label a sample according to the majority of the labelers.

Deliverable: A figure that shows the incremental performance by the size of handwritten document image dataset.

Work package 4: New state of the art results on the present handwritten document image analysis benchmarks. (3 months) Objectives: To present new state of the art results on the present handwritten document image analysis benchmarks .

Tasks: Task 1.1 - Implement the model. - Train the model on the large scale handwritten document image dataset. - Fine tune the model on the present benchmark. - Compare the results on the present benchmark with the results of the literature.

Risks: - The results do not over perform the literature.

Contingency plan: - Review the labels by a consensus check. Consensus check labels a sample according to majority of the labelers.

Deliverable: A table that compares the results of the pretrained model and the results of the literature on the present benchmarks.

Work package 5: Organize an international competition (3 months) Objectives: To organize an international competition on the large scale handwritten document image analysis dataset for advancing handwritten document image analysis research by a collective effort.

Tasks: Task 1.1 - Identify a premier international conference to propose the competition. - Identify the evaluation metrics for each stage of the handwritten document image analysis. - Identify the official data splits for training, testing and blind testing. - Make experiments to get the baseline results. - Identify the submission format. - Prepare the competition proposal. - Execute the submissions locally and evaluate them on the blind test set. - Compare the submission results.

Risks: - No one participates the competition.

Contingency plan: - Execute the publicly available methods and compare them with the baseline results.

Deliverable: A paper that presents the dataset, evaluation metrics, baseline results and the comparison of the submitted method's results.

References

1. Afzal, M.Z., Capobianco, S., Malik, M.I., Marinai, S., Breuel, T.M., Dengel, A., Liwicki, M.: Deepdocclassifier: Document classification with deep convolutional neural network. In: 2015 13th international conference on document analysis and recognition (ICDAR). pp. 1111–1115. IEEE (2015)
2. Chen, K., Seuret, M.: Convolutional neural networks for page segmentation of historical document images. arXiv preprint arXiv:1704.01474 pp. – (2017)
3. Chen, K., Seuret, M., Liwicki, M., Hennebert, J., Ingold, R.: Page segmentation of historical document images with convolutional autoencoders. In: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. pp. 1011–1015. IEEE (2015)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Divya, B., Goswami, M.M., Mitra, S.: Dnn based approaches for segmentation of handwritten gujarati text. In: 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC). pp. 1–6. IEEE (2020)
6. Droby, A., Barakat, B.K., Madi, B., Alaasam, R., El-Sana, J.: Unsupervised deep learning for handwritten page segmentation. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 240–245. IEEE (2020)
7. Grüning, T., Leifert, G., Strauß, T., Michael, J., Labahn, R.: A two-stage method for text line detection in historical documents. International Journal on Document Analysis and Recognition (IJDAR) **22**(3), 285–302 (2019)
8. Karpinski, R., Belaïd, A.: Semi-synthetic data augmentation of scanned historical documents. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 268–273. IEEE (2019)
9. Krishnan, P., Dutta, K., Jawahar, C.: Deep feature embedding for accurate recognition and retrieval of handwritten text. In: Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on. pp. 289–294. IEEE (2016)
10. Krishnan, P., Jawahar, C.: Hwnet v2: An efficient word image representation for handwritten documents. International Journal on Document Analysis and Recognition (IJDAR) **22**(4), 387–405 (2019)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
12. Kurar Barakat, B., Droby, A., Alasam, R., Madi, B., Rabaev, I., El-Sana, J.: Text line extraction using fully convolutional network and energy minimization. International Workshop on Pattern Recognition for Cultural Heritage (2020)

13. Kurar Barakat, B., Droby, A., Alasam, R., Madi, B., Rabaev, I., Shammes, R., El-Sana, J.: Unsupervised deep learning for text line segmentation. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 3651–3656. IEEE (2020)
14. Kurar Barakat, B., Droby, A., Kassis, M., El-Sana, J.: Text line segmentation for challenging handwritten document images using fully convolutional network. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 374–379. IEEE (2018)
15. Kurar Barakat, B., Droby, A., Saabni, R., El-Sana, J.: Unsupervised learning of text line segmentation by differentiating coarse patterns. In: 2021 16th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1355–1360. IEEE (2021)
16. Kurar Barakat, B., El-Sana, J.: Binarization free layout analysis for arabic historical documents using fully convolutional networks. International Workshop on Arabic Script Analysis and Recognition (2018)
17. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
18. Lee, J., Hayashi, H., Ohyama, W., Uchida, S.: Page segmentation using a convolutional neural network with trainable co-occurrence features. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1023–1028. IEEE (2019)
19. Marinai, S.: Introduction to document analysis and recognition. In: Machine learning in document analysis and recognition, pp. 1–20. Springer (2008)
20. Mechi, O., Mehri, M., Ingold, R., Amara, N.E.B.: Text line segmentation in historical document images using an adaptive U-Net architecture. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 369–374. IEEE (2019)
21. Monnier, T., Aubry, M.: docextractor: An off-the-shelf historical document element extraction. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 91–96. IEEE (2020)
22. Neche, C., Belaïd, A., Kacem-Echi, A.: Arabic handwritten documents segmentation into text-lines and words using deep learning. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 6, pp. 19–24. IEEE (2019)
23. Oliveira, S.A., Seguin, B., Kaplan, F.: dhsegment: A generic deep-learning approach for document segmentation. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 7–12. IEEE (2018)
24. Pondenkandath, V., Alberti, M., Diatta, M., Ingold, R., Liwicki, M.: Historical document synthesis with generative adversarial networks. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 5, pp. 146–151. IEEE (2019)
25. Renton, G., Soullard, Y., Chatelain, C., Adam, S., Kermorvant, C., Paquet, T.: Fully convolutional network with dilated convolutions for handwritten text line segmentation. International Journal on Document Analysis and Recognition (IJDAR) **21**(3), 177–186 (2018)
26. Seuret, M., Chen, K., Eichenbergery, N., Liwicki, M., Ingold, R.: Gradient-domain degradations for improving historical documents images layout analysis. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1006–1010. IEEE (2015)

27. Sfikas, G., Retsinas, G., Gatos, B.: Zoning aggregated hypercolumns for keyword spotting. In: Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on. pp. 283–288. IEEE (2016)
28. Studer, L., Alberti, M., Pondenkandath, V., Goktepe, P., Kolonko, T., Fischer, A., Liwicki, M., Ingold, R.: A comprehensive study of imagenet pre-training for historical document image analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 720–725. IEEE (2019)
29. Sudholt, S., Fink, G.A.: Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 277–282. IEEE (2016)
30. Sudholt, S., Fink, G.A.: Evaluating word string embeddings and loss functions for cnn-based word spotting. In: 2017 14th iapr international conference on document analysis and recognition (icdar). vol. 1, pp. 493–498. IEEE (2017)
31. Sudholt, S., Fink, G.A.: Attribute cnns for word spotting in handwritten documents. International journal on document analysis and recognition (ijdar) **21**(3), 199–218 (2018)
32. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision. pp. 843–852 (2017)
33. Vo, Q.N., Kim, S.H., Yang, H.J., Lee, G.S.: Text line segmentation using a fully convolutional network in handwritten document images. IET Image Processing **12**(3), 438–446 (2017)
34. Wei, H., Seuret, M., Chen, K., Fischer, A., Liwicki, M., Ingold, R.: Selecting autoencoder features for layout analysis of historical documents. In: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing. pp. 55–62. ACM (2015)
35. Xu, Y., He, W., Yin, F., Liu, C.L.: Page segmentation for historical handwritten documents using fully convolutional networks. In: Document Analysis and Recognition (ICDAR), 2017 15th International Conference. IEEE (2017)