Using Multi-level Segmentation Features for Document Image Classification

Panagiotis Kaddas[1,2(✉)] and Basilis Gatos[1]

[1] Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research Demokritos, 153 10 Agia Paraskevi, Athens, Greece
{pkaddas,bgat}@iit.demokritos.gr
[2] Department of Informatics and Telecommunications, University of Athens, 157 84 Athens, Greece

Abstract. Document Image classification is a crucial step in the processing pipeline for many purposes (e.g. indexing, OCR, keyword spotting) and is being applied at early stages. At this point, textual information about the document (OCR) is usually not available and additional features are required in order to achieve higher recognition accuracy. On the other hand, one may have reliable segmentation information (e.g. text block, paragraph, line, word, symbol segmentation results), extracted also at pre-processing stages. In this paper, visual features are fused with segmentation analysis results in a novel integrated workflow and end-to-end training can be easily applied. Significant improvements on popular datasets (Tobacco-3482 and RVL-CDIP) are presented, when compared to state-of-the-art methodologies which consider visual features.

Keywords: Document image classification · Document image segmentation · Convolutional Neural Network · Deep Learning

1 Introduction

Digitization of Documents has already become a necessity in order to assist daily tasks and transactions. Moreover and throughout the globe, historical documents can be accessed on-line and information can be exploited by the community for any purpose. To this end, several methodologies are being applied on scanned document images in order to convert them to their digital twins. This leads to the application of image processing techniques such as Optical Character Recognition (OCR), automatic indexing and keyword-spotting for searching documents in huge databases.

Unfortunately, most of the techniques that have been mentioned do not always apply successfully due to the vast diversity of document types. For example, a historical handwritten document must be processed using a completely

S. Uchida et al. (Eds.): DAS 2022, LNCS 13237, pp. 702–712, 2022.
https://doi.org/10.1007/978-3-031-06555-2_47

different workflow when compared to processing a scanned invoice. So, it is clear that a prior processing step must be applied in order to: a) classify document images to their corresponding type (class) and b) select a suitable system for processing based on document class. Motivated by this observation, this paper addresses the problem of document classification, based on the visual features of document images.

Classic approaches [1–3] towards the document classification problem focus on extracting image features for defining structural similarity of documents. During the past ten years, the rise of Deep Learning has been proven suitable to match the need of classifying huge document image databases [4] with high intra-class and low inter-class variability by only using the document image as input and leveraging the advantages of transferring knowledge from similar domains [5,6]. Recent works [7–9] have shown that textual information can be combined with image features in order to improve classification accuracy.

The proposed work focuses on the Document Image Classification problem by combining a Convolutional Neural Network (CNN) architecture with multi-level information provided by image segmentation techniques (text block, paragraph, line, word, symbol segmentation results). Textual information is not considered in the proposed work, under the assumption that document image classification usually takes place in pre-processing stages where textual information (OCR) is not available.

The contributions of this paper are as follows: a) A novel integrated architecture is described and end-to-end training can be applied by only using a document image and one or more image masks that correspond to the segmentation levels that are mentioned above. b) An experimental study is being conducted in order to determine which segmentation levels should be used and decide whether multi-level segmentation features contribute to the task at hand. c) We present competitive results when compared to the state-of-the-art techniques, evaluated on commonly-used datasets (Tobacco-3482 [3], RVL-CDIP [4]). d) An additional proof-of-concept is presented for a new private dataset from The Library of the Piraeus Bank Group Cultural Foundation (PIOP)[1] used in the *CULDILE*[2] project.

The rest of the paper is organized as follows. Section 2 presents related works, Sect. 3 introduces the proposed architecture, Sect. 4 demonstrates experimental results and Sect. 5 presents the conclusion of this work.

2 Related Work

Several methodologies that leverage the advantages of Deep Neural Networks over document images have been proposed over the last decade for the document image classification problem. In [10], a basic CNN architecture is proposed in order to learn features from raw image pixels instead of relying on hand-crafted

[1] https://www.piop.gr/en/vivliothiki.aspx.
[2] http://culdile.bookscanner.gr.

features. In [4], it was shown that a CNN can extract robust features from different parts of the image (header, footer left and right body) and by training such different networks and using them in an ensemble scheme, significant classification or retrieval accuracy improvements are achieved. Moreover, reduction of feature space by using Principal Component Analysis (PCA) is important, while the performance is not affected significantly.

Further experiments for document image classification where presented in [5,11,12], where many neural network architectures where compared (e.g. *AlexNet* [13], *VGG-16* [14], *ResNet50* [15], *GoogleNet* [16]) under different scenarios (transfer learning [5], data augmentation [17]). The advantages of learning using some kind of spatial information from parts of the document (holistic, header, footer, left and right body) and parallel VGG-16 based systems were presented in [6]. Furthermore, inter-domain and intra-domain transfer learning schemes are described and finally they present a comparison of possible meta-classifier techniques on the stacked output of the parallel sub-systems.

In addition to the methodologies described above which try to learn and make use of visual features of document images, there are recent techniques [7–9] that fuse visual and textual information in parallel systems. In [7], two classifiers are trained in parallel. The first is a classic visual-based CNN and the second takes as input text embeddings, extracted using open-source Tesseract OCR Engine[3] and FastText[4]. Similar approach is also presented in [8], where text embeddings proposed in [18] are considered as a textual feature selection scheme. Fernando et al. [9] proposed the use of *EfficientNet* models [19] as a lighter alternative to classic CNN architectures for the visual feature extraction and combined it with the well-known *BERT* model [20] as a textual transformer. Finally, Xu et al. [21] proposed the LayoutLM model, where layout and image embeddings extracted from *Faster R-CNN* [22] are integrated into the original BERT architecture and work together for feature extraction.

3 Proposed Method

As mentioned in Sect. 1, this work does not focus on using textual information, under the assumption that document image classification usually takes place in pre-processing stages where textual information (OCR) is not available or reliable. Instead of this, we try to embed segmentation features of multiple levels (e.g. text block, paragraph, line, word, symbol segmentation results), which are usually extracted during pre-processing stages (Fig. 1).

3.1 Integrated CNN Architecture

The proposed system relies on two kinds of input: At first, a classic CNN flow is considered, using *ResNet50* as a backbone architecture. We chose *ResNet50*,

[3] https://github.com/tesseract-ocr/tesseract.
[4] https://github.com/facebookresearch/fastText.

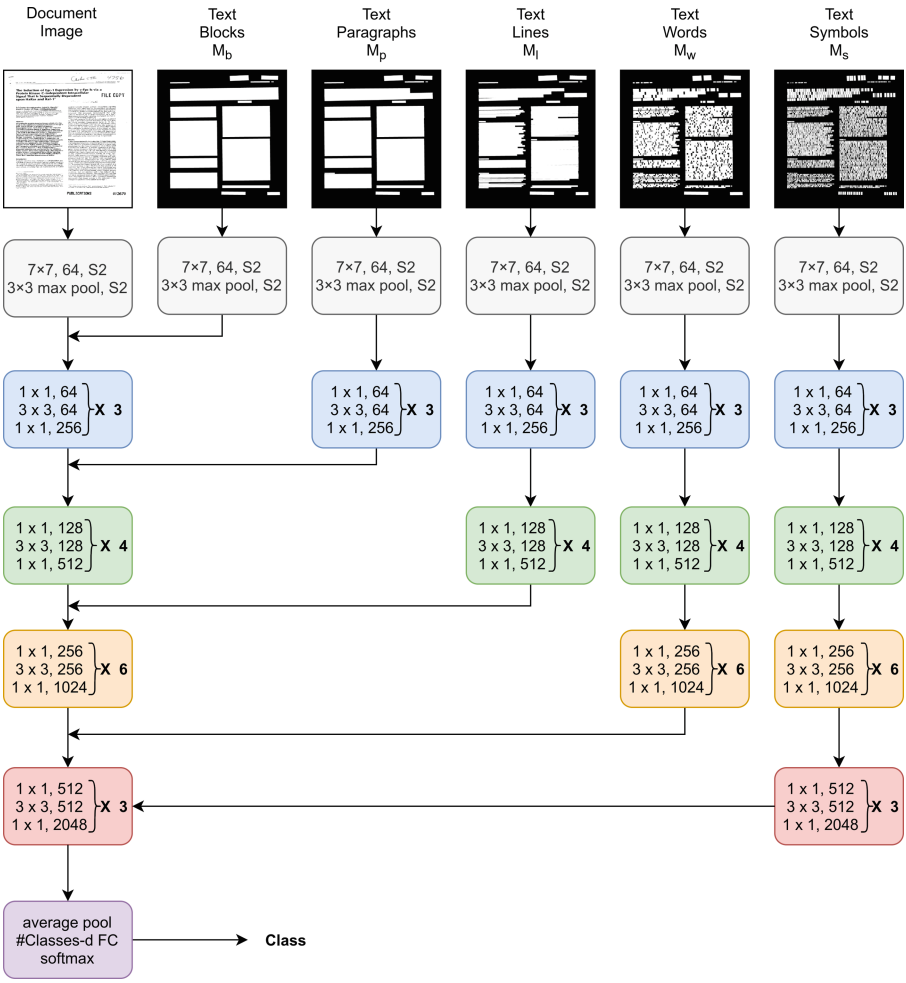

Fig. 1. Overview of the proposed network. Document image and segmentation masks are forwarded in parallel network streams using *ResNet50* as backbone network. Each segmentation stream is "deeper" than the previous one in order to be able to learn higher level of information. Each output stream is added to the corresponding layer of the backbone (left branch) and finally a Fully Connected layer yields class probabilities. (Color figure online)

over other backbones (e.g. *VGG-16*, *GoogleNet*) for reasons such as: advantages of residual connections [15], simplicity in architecture, number of weight parameters. We do not conduct experiments using other CNN backbones and this is out of the scope of this paper. In addition, our main goal is to demonstrate accuracy improvements of our proposed system over similar state-of-the-art techniques mainly based on *ResNet50*.

The second kind of inputs to our system are document image binary masks that represent segmentation information at various text levels, namely M_b : *block*, M_p : *paragraph*, M_l : *line*, M_w : *word*, M_s : *symbol*. A pixel of each mask has an "on" value if it is contained in a detected polygon of this level. So, we consider binary masks that correspond to the detected polygons *(x,y coordinates)* of multiple segmentation levels as inputs. As described in our experiments (Sect. 4), not all segmentation masks are required in order to achieve accuracy improvements in the classification task. Moreover, segmentation results may not be 100% accurate and can be used exactly as extracted by any segmentation tool. We claim (Sect. 4) that the proposed system is robust even when using noisy segmentation results.

The overall architecture is provided in Fig. 1. The left branch is a *ResNet50*, where all blocks are included (a convolution block followed be 4 stacked residual building blocks which are repeated *3*, *4*, *6* and *3* times respectively). The left branch is considered as the backbone of the proposed system. When forwarding to the next type of a residual block (illustrated with different colors in the left branch), output features are fused with those extracted from a segmentation mask.

The other five branches take as input the segmentation masks. Each branch is "deeper" compared to the previous one and is forwarded through an extra stack of residual building blocks, following the scheme shown in Fig. 1. This scheme is inspired by the fact that convolution layers applied at early stages learn abstract layouts and shapes (such as spatial position of text blocks in our case, which are considered the higher level of information), in contrast to deeper layers which can handle more complex visual elements and details (like positions of text lines, words and even symbols, the lowest level of information). So, we handle higher levels of information with less layers and we increase the depth of a branch considering the level of details in the input segmentation mask. The proposed architecture was also verified after trying many alternative schemes (e.g. forward high level segmentation masks through "deeper" network branches) which yielded less accurate results.

3.2 Implementation Details

In general, our TensorFlow[5] implementation of the proposed model follows [15]. We use "bottleneck" residual blocks for convolutions, followed by Batch Normalization (BN) and $ReLU$ activation layers. The main differences from [15] are that, during training, we use input images of size 256×256 as long as cyclic learning rate [23] with Stochastic Gradient Descent (SGD), with values ranging in $[0.0001, 0.1]$. Before the final Fully Connected (FC) layer, we apply *Dropout* with skipping ratio of 0.5. Finally, our inputs are augmented using random cropping (80% of the original size at most) and mirroring over $y - axis$. Weight initialization comes from ImageNet weights [24].

[5] https://www.tensorflow.org.

Training can be easily applied over the integrated network. If information is not available at a certain segmentation level (e.g. paragraphs), the corresponding branch can be discarded. We used an NVIDIA GTX 1080 Ti 11 GB GPU with batch size 16 at most cases.

For segmenting the documents, the Google Cloud Vision API[6] is considered.

4 Experiments

4.1 Datasets

For our experiments, we use three datasets. The smallest one is Tobacco-3482 [3] which consists of ten document classes. As there is no official split in train-val-test subsets, we average over five random splits, following the same logic as in other state-of-the-art methods [4]. We use this dataset in order to evaluate the performance of our proposed network and to investigate the contribution of segmentation information for the document classification task.

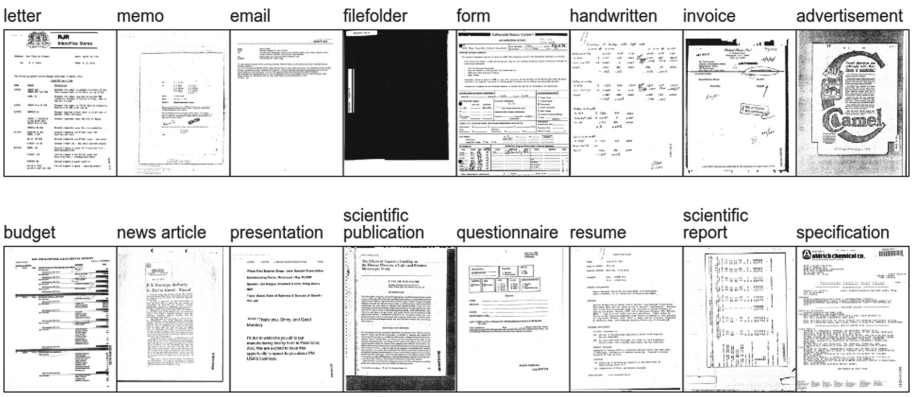

Fig. 2. Example of the RVL-CDIP dataset, consisting of 16 document classes.

Secondly, we use the large-scale Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP) dataset [4]. It consists of 320,000 training images and a validation and test dataset with 40,000 images each. This dataset has 16 document classes (see Fig. 2) and is considered the most challenging dataset for Document Image Classification.

Finally, we introduce a new dataset, obtained from The Library of the Piraeus Bank Group Cultural Foundation (PIOP)[7] and used in the *CULDILE*[8] project. We selected pages belonging to four classes (Mail, Contracts, Financial,

[6] https://cloud.google.com/vision.
[7] https://www.piop.gr/en/vivliothiki.aspx.
[8] http://culdile.bookscanner.gr.

Architecture Plans) and each class has about 9,000 images. We split the dataset randomly (60% train, 20% validation and 20% test). Exemplar images of this dataset are given in Fig. 3.

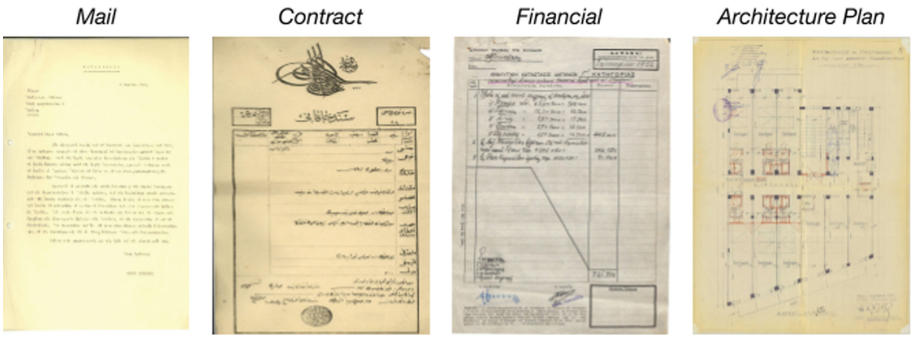

Fig. 3. Exemplar images of the PIOP dataset, consisting of 4 document classes.

4.2 Experimental Results

As a first experiment, we investigate the contribution of using multi-level segmentation features to the document classification problem. For this reason, we use the Tobacco-3482 and train using different schemes. At scenario $Baseline_A$, we train a single $ResNet50$. At $Baseline_B$, we train several classic $ResNet50$ models using only segmentation information (and not the original document image), by stacking masks in a single image of depth n, where n is the number of the stacked masks. We do this for every possible combination over the segmentation masks. At $Baseline_C$, we just concatenate the output probabilities of already trained models of $Baseline_A$ and $Baseline_B$ for every combination of the latter. Finally, at $Baseline_D$, we train $Baseline_A$ and $Baseline_B$ models (all combinations again) in a parallel scheme, where we concatenate the convolution outputs for both models and we use an FC layer of 1024 neurons before the output FC layer. This can be considered as a simple ensemble scheme.

The investigation mentioned above help us to decide which levels of segmentation to keep in the proposed architecture. Table 1 demonstrates the best results for each baseline scenario. The best combination uses the initial document image, line, word and symbol segmentation masks. For completeness, we summarize the results of previous works for the Tobacco-3482 dataset that use visual features (we do not include methods that use textual features). As shown in Table 1, our proposed method outperforms (80.64%) all ResNet50-based models ($Baseline_A$ and [5]), as long as other architectures that use AlexNet, VGG-16 or GoogleNet as backbones [4,5]. We note that we do not compare with methods that use weight initialization from models trained on the much larger RVL-CDIP dataset.

Table 1. Accuracy of combinations over multi-level segmentation masks for document image classification using F-Measure for Tobacco-3482 (%).

Method	Image	Block	Line	Word	Symbol	Accuracy (%)
$Baseline_A$	✓					68.78
$Baseline_B$			✓	✓	✓	75.40
$Baseline_C$	✓		✓	✓	✓	79.86
$Baseline_D$	✓	✓	✓	✓		78.63
Proposed method	✓		✓	✓	✓	**80.64**
Harley et al. - Ensemble of regions [4]	✓					79.90
Afzal et al. - VGG-16 [11]	✓					77.60
Afzal et al. - ResNet50 [5]	✓					67.93
Audebert et al. - MobileNetV2 [7]	✓					**84.50**
Fernando et al. - EfficientNet [9]	✓					**85.99**

Furthermore, from all baseline experiments that were conducted, it was clear that segmentation information contributes significantly to classification tasks (almost an additional 11% in accuracy for $Baseline_C$). In fact, even when using only segmentation masks instead of the original document image, accuracy increases remarkably ($Baseline_B$). We found out that Line and Word and Symbol segmentation masks play the most important role in most combinations and yielded better results (not included in Table 1, for convenience), in contrast to Paragraph and Block masks that are of less importance. Finally, our proposed method does not outperform methods that depend on more recent backbones (MobileNetV2 [7] and EfficientNet [9]). We believe that applying our proposed scheme in a future work, using such backbones, will improve accuracy even more, when compared to [7] and [9].

Our second experiment concerns the evaluation of our proposed architecture over the RVL-CDIP and PIOP datasets. Again, we do not consider textual methods in our comparison (Table 2). As in our first experiment, we use the best input combination (document image, line, word and symbol segmentation masks) as the proposed system. We observe that, concerning the RVL-CDIP dataset, the proposed method outperforms all other techniques no matter the backbone architecture that is used (92.95%). Finally, the PIOP dataset is another proof that our proposed scheme can improve accuracy results when applied on a ResNet50 backbone.

Table 2. Accuracy of combinations over multi-level segmentation masks for document image classification using F-Measure for RVL-CDIP and PIOP datasets (%).

Method	Accuracy on RVL-CDIP (%)	Accuracy on PIOP (%)
Harley et al. - Ensemble of regions [4]	89.80	–
Csurka et al. - GoogleNet [12]	90.70	–
Afzal et al. - ResNet50 [5]	90.40	–
Afzal et al. - VGG-16 [5]	90.97	–
Das et al. - Ensemble of VGG-16 models [6]	92.21	–
Fernando et al. - EfficientNet [9]	92.31	–
$Baseline_A$	90.55	84.28
Proposed method	**92.95**	**86.31**

5 Conclusion

In this paper we proposed a novel integrated architecture in which document images and multi-level segmentation features can be fused for document classification. We showed in our experiments that segmentation is considered to be useful for improving image-based classification methods, even when we use noisy multi-level masks. This also introduces accuracy improvements for the RVL-CDIP dataset, as presented in our experiments. Moreover, we conducted an investigation on Tobacco-3482 in order to define which segmentation levels can yield better results and found that more detailed segmentation (at line, word and symbol) is of greater importance rather than segmentation of higher levels (blocks, paragraphs).

Acknowledgements. This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the RESEARCH-CREATE-INNOVATE call (project code: T1EDK-03785 and acronym: CULDILE) as well as by the program of Industrial Scholarships of Stavros Niarchos Foundation[9].

References

1. Shin, C.K., Doermann, D.S.: Document image retrieval based on layout structural similarity. In: Proceedings of the 2006 International Conference on Image Processing, Computer Vision, & Pattern Recognition (IPCV), Las Vegas, Nevada, USA, pp. 606–612 (2006)

[9] https://www.snf.org/en/.

2. Chen, S., He, Y., Sun, J., Naoi, S.: Structured document classification by matching local salient features. In: 21st International Conference on Pattern Recognition (ICPR), Tsukuba Science City, Japan, pp. 1558–1561 (2012)
3. Kumar, J., Ye, P., Doermann, D.: Structural similarity for document image classification and retrieval. Pattern Recogn. Lett. **43**, 119–126 (2016)
4. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: 13th International Conference on Document Analysis and Recognition (ICDAR), Nancy, France, pp. 991–995 (2015)
5. Afzal, M.Z., Kölsch, A., Liwicki, S.A.M.: Cutting the error by half: investigation of very deep CNN and advanced training strategies for document image classification. In: 14th International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, pp. 883–890 (2017)
6. Das, A., Roy, S., Bhattacharya, U., Parui, S.K.: Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. In: 24th International Conference on Pattern Recognition (ICPR), Beijing, China, pp. 3180–3185 (2018)
7. Audebert, N., Herold, C., Slimani, K., Vidal, C.: Multimodal deep networks for text and image-based document classification. In: Cellier, P., Driessens, K. (eds.) ECML PKDD 2019. CCIS, vol. 1167, pp. 427–443. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-43823-4_35
8. Asim, M.N., Khan, M.U.G., Malik, M.I., Razzaque, K., Dengel, A., Ahmed, S.: Two stream deep network for document image classification. In: 15th International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, pp. 1410–1416 (2019)
9. Ferrando, J., et al.: Improving accuracy and speeding up document image classification through parallel systems. In: Krzhizhanovskaya, V.V., et al. (eds.) ICCS 2020. LNCS, vol. 12138, pp. 387–400. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50417-5_29
10. Kang, L., Kumar, J., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for document image classification. In: 22th International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, pp. 3168–3172 (2014)
11. Afzal, M.Z., et al.: DeepDocClassifier: document classification with deep convolutional neural network. In: 13th International Conference on Document Analysis and Recognition (ICDAR), Nancy, France, pp. 1273–1278 (2015)
12. Csurka, G., Larlus, D., Gordo, A., Almazan, J.: What is the right way to represent document images?. arXiv preprint arXiv:1603.01076 (2016)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: 26th Conference on Neural Information Processing Systems (NIPS), Harrah's Lake Tahoe, USA, pp. 1097–1105 (2012)
14. Simonyan, K., Zisserman, A: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, pp. 770–778 (2016)
16. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, pp. 1–9 (2015)
17. Tensmeyer, C., Martinez, T.: Analysis of convolutional neural networks for document image classification. arXiv preprint arXiv:1708.03273 (2017)

18. Noce, L., Gallo, I., Zamberletti, A., Calefati A.: Embedded textual content for document image classification with convolutional neural networks. In: Proceedings of the ACM Symposium on Document Engineering (DocEng), Vienna, Austria, pp. 165–173 (2016)
19. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (PMLR), Long Beach, California, pp. 6105–6114 (2019)
20. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistic (NAACL), Mineapolis, Minesota, USA, pp. 4171–4186 (2019)
21. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery (SIGKDD), pp. 1192–1200 (2020)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
23. Smith, L.N.: Cyclical learning rates for training neural networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, California, USA, pp. 464–472 (2017)
24. Deng, J., Dong, W., Socher, R., Li, L.J., Li K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Computer Vision and Pattern Recognition (CVPR), Miami, Florida, USA, pp. 248–255 (2009)