Towards Visual Explanations for Document Table Detection using Coarse Localization Maps

Arnab Ghosh Chowdhury[1], Martin Atzmueller[1,2]

[1]Osnabrück University, Semantic Information Systems (SIS) Group, Osnabrück, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Osnabrück, Germany

Abstract

Computer-vision-based methods using deep neural networks offer considerable opportunities to extract tabular information from richly-structured documents. However, it is extremely challenging to build a unified framework for tabular information extraction, for example, due to a variety of document templates, as well as, diverse document table templates. Earlier, we proposed a transfer learning based table detection approach[1] and a supervised table detection framework initialized with pre-trained self-supervised image classification model weight for table detection [2] on domain specific document images. In this paper, we investigate different document table detection techniques with respect to explainability issues. These enable, e. g., diagnostics and method refinement, towards a complete tabular data extraction pipeline and tool. In particular, we present visual explanation approaches of earlier proposed table detection models on domain specific document images in order to enhance the explainability of the applied Convolutional Neural Network (CNN) based models. We discuss first experimental results for visual explanations of those models and outline several challenges in this context.

Keywords

Tabular Information Extraction, Barlow Twins, Grad-CAM, Grad-CAM++, Ablation-CAM

1. Introduction

Document tables commonly offer essential information in a systematic structured way. Document table detection is a critical task due to diverse layouts and formats of the document templates, as well as, table templates. In the age of digitalization, documents are often provided in digital form such as Portable Document Format (PDF) documents, LaTeX documents or scanned documents. The open-source tools such as Camelot[1] or Tabula[2] are not properly suitable yetx to support all possible PDF documents in order to extract tabular information, e. g., due to diverse layouts of PDF document templates. In this context, computer vision based object detection approaches have emerged for enabling document layout analysis and for table detection. Benchmark datasets such as TableBank [3], PubLayNet [4] along with pre-trained object detection models[3], for example, are readily available for document layout analysis. However, some domain specific information extraction tasks usually still suffer due to absence of manually annotated benchmark datasets, as well as, the pre-trained object detection models.

LWDA'22: Lernen, Wissen, Daten, Analysen. October 05–07, 2022, Hildesheim, Germany

arnab.ghosh.chowdhury@uos.de (A. Ghosh Chowdhury); martin.atzmueller@uos.de (M. Atzmueller)

[1] https://github.com/camelot-dev/camelot

[2] https://github.com/chezou/tabula-py

[3] https://github.com/Layout-Parser/layout-parser/blob/main/src/layoutparser/models/detectron2/catalog.py

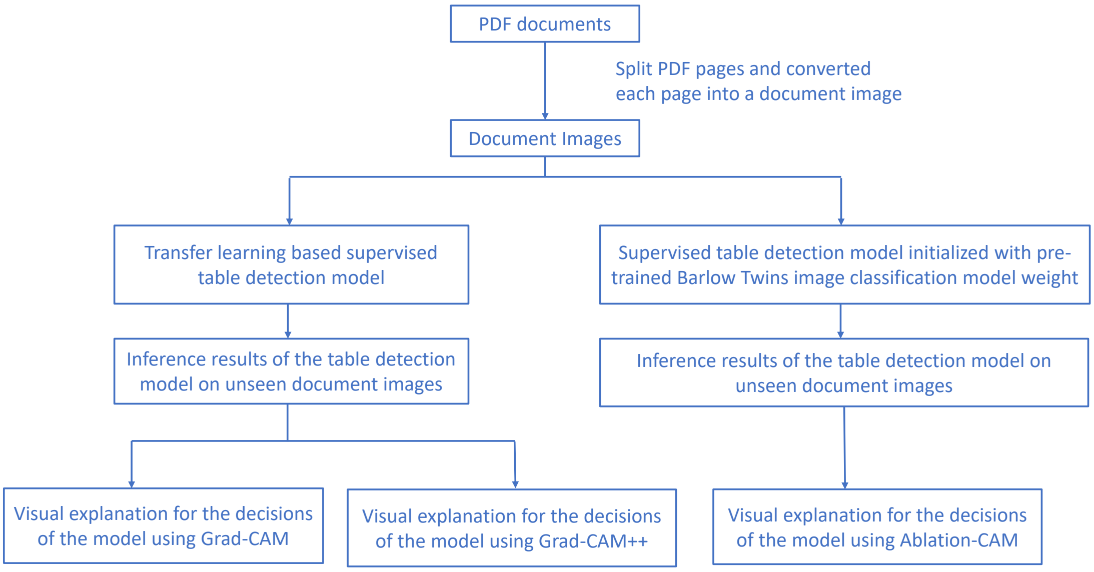

PDF documents
Split PDF pages and converted each page into a document image
Document Images
Transfer learning based supervised table detection model
Supervised table detection model initialized with pre-trained Barlow Twins image classification model weight
Inference results of the table detection model on unseen document images
Inference results of the table detection model on unseen document images
Visual explanation for the decisions of the model using Grad-CAM
Visual explanation for the decisions of the model using Grad-CAM++
Visual explanation for the decisions of the model using Ablation-CAM

Figure 1: Overview: Document table detection methods and explanation approaches.

Previously, we extracted tabular information directly from document images using Optical Character Recognition (OCR) [1]. In this paper, we specifically investigate these approaches towards their explainability, which can then be applied for assessment, diagnostics, and method refinement. Then, ultimately, such approaches can the enable flexible tabular data extraction, i. e., by mapping the bounding box coordinates of predicted tables of document images to PDF document pages in order to identify the table region or region of interest (ROI) on the respective document page. We discuss this in the context of a prototypical implementation using the ROI information along with the Camelot tool to extract tabular data from PDF documents to facilitate knowledge management, e. g., [5, 6, 7, 8, 9]. Regarding explainability, we apply coarse localization maps to offer visual explanations for the decisions of the two table detection models on the domain specific (so-called) Di-Plast dataset leveraging the respective Grad-CAM [10], Grad-CAM++ [11] and Ablation-CAM [12] methods. Figure 1 depicts the methods and structure.

In previous work, we proposed a transfer learning based table detection approach [1] on the given domain specific Di-Plast dataset[4] on the basis of a pre-trained TableBank model, which is trained on the TableBank dataset [1]. Such pre-trained TableBank or PubLayNet models generally follow the supervised object detection framework, for example, Faster Region based Convolutional Neural Network (Faster R-CNN) [3, 4] which is primarily initialized with pre-trained ImageNet supervised image classification model weight. In another previous experiment [2], we utilized a pre-trained self-supervised image classification model weight instead of the pre-trained ImageNet supervised image classification model weight, and obtained a comparatively substantial table detection results compared to the transfer learning approach.

[4] https://github.com/cslab-hub/MatrixDataExtractor/tree/main/tabledetection

We leveraged the architecture of Barlow Twins, a redundancy-reduction based self-supervised learning method [13] to build an image classifier, which is trained on subsets of PubTabNet [14] and DVQA [15] datasets. Subsequently, we exploited the supervised Faster R-CNN object detection framework, which is primarily initialized with this Barlow Twins image classification model weight [2]. However, the transfer learning based table detection model achieves better performance than this model on our domain specific Di-Plast dataset. Therefore, it is quite beneficial to obtain the visual explanations for the decisions made by our transfer learning based table detection model, and the respective supervised Faster R-CNN table detection model. In general, this can enhance the transparency of the respective models, which also provides important options for assessment, tuning, and further analysis.

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 summarizes the applied table detection methods, with an outlook on tabular data extraction. Section 4 presents visual explanations for the decisions of our table detection models using Grad-CAM, Grad-CAM++ and Ablation-CAM. Finally, Section 5 concludes the paper with a summary and outlines interesting directions for future work.

2. Related Work

Below, we briefly discuss related work concerning document layout analysis, before we summarize visual explanation methods for deep Convolutional Neural Network (CNN) based models.

2.1. Barlow Twins in Document Layout Analysis

Barlow Twins is a redundancy-reduction based self-supervised learning method. It works on a joint embedding of two augmented views for all images of a batch sampled from a training dataset, learning representations that are invariant under different image augmentations. It estimates the cross-correlation between the embeddings of two identical networks incorporating augmented views for all images of a batch of samples, aiming to make the cross-correlation matrix close to the identity matrix [13]. [16], for example, conduct a document image classification task using Barlow Twins on the RVL-CDIP [17] and the Tobacco-3482 [18] datasets.

2.2. Class Activation Mapping (CAM)

To build trustworthy CNN models, it is important to explain their decisions, for example, why these table detection models predict what they predict. This transparency helps to comprehend the failure scenarios and to debug CNN based models along with identifying and eliminating potential biases in training data [10]. [19] demonstrates that the convolutional units of different layers of Convolutional Neural Network (CNN) act as object detectors in spite of no supervision on the location of the object was offered. Such ability to localize objects is lost when fully-connected (FC) layers are used for image classification. The class activation mapping (CAM) for CNN is proposed with global average pooling to empower the classification-trained CNN to learn to perform object localization without utilizing bounding box annotations. Towards object localization, class activation maps facilitate to visualize the predicted class scores on the image by highlighting the discriminative object parts detected by the CNN [20].

Gradient-weighted Class Activation Mapping (Grad-CAM) is proposed to offer visual explanations for the decisions from a large class of CNN based models to make those models more transparent and explainable. In image classification, it uses the gradients of target concepts, for instance, class labels, flowing into the final convolutional layer to generate a coarse localization map highlighting the important regions in the image to predict the concept or the class label [10]. Grad-CAM methods have some limitations, such as that the performance falls when multiple occurrences of the same class are localized. Also, Grad-CAM heatmaps commonly do not capture the entire object in completeness for single object images. Here, Grad-CAM++ method is proposed to offer better visual explanations of the CNN model predictions [11]. Furthermore, Grad-CAM suffers from the gradient saturation problem; this induces the backpropagating gradients to diminish, adversely affects the quality of visualizations, and suffers from detecting multiple occurrences of the same object in an image. Unlike *gradient based* methods (e. g., Grad-CAM, Grad-CAM++), a *gradient free* visualization method known as Ablation-CAM is proposed to produce visual explanations for interpreting CNN models, that avoids the use of gradients, and simultaneously offers high quality class-discriminative localization maps [12].

In this context, [21] proposes a CNN architecture for handwritten Chinese character recognition and leverages CAM method for visual explanations. [22] studies numerous weakly supervised object localization and detection approaches along with CAM methods. [23] investigates the pseudo-label-based semi-supervised object detection system and applies Grad-CAM to give further evidence for their proposed Mix/UnMix (MUM) data augmentation method. A novel weakly supervised object detection approach is introduced that uses both the proposal-level relationship and the semantic-level relationship, and generates object proposals based on the heatmaps extracted by Grad-CAM [24]. [25] leverages Grad-CAM generating and selecting high-quality proposals for weakly supervised object detection.

3. Methods

In this section, we first summarize our earlier proposed table detection methods before presenting first results of our experimentation below. We also sketch a prototypical tabular data extraction approach by mapping the bounding box coordinates of the predicted tables of document images to PDF document pages, then leveraging the Camelot tool for data extraction.

3.1. Barlow Twins based Table Detection Model

The encoder of the Barlow Twins model consists of a ResNet50 network (without the final classification layer, and using 2048 output units) which is followed by a projector network. Such a projector network consists of three linear layers, each with 8192 output units. The first two layers of the projector are followed by a batch normalization layer and rectified linear units. The output of the encoder is denoted as the *representation*, while the output of the projector is denoted as the *embedding* [13]. We exploited the learned representations for table detection; the obtained embeddings are fed to the loss function of the Barlow Twins model. After image classification model training, the encoder of the ResNet-50 network (without the final classification layer and with 2048 output units) is fed into the ResNet-FPN (Feature Pyramid Network) architecture, as the backbone of the Faster R-CNN table detection framework [2].

Table 1
Train and test dataset distribution for image classification.

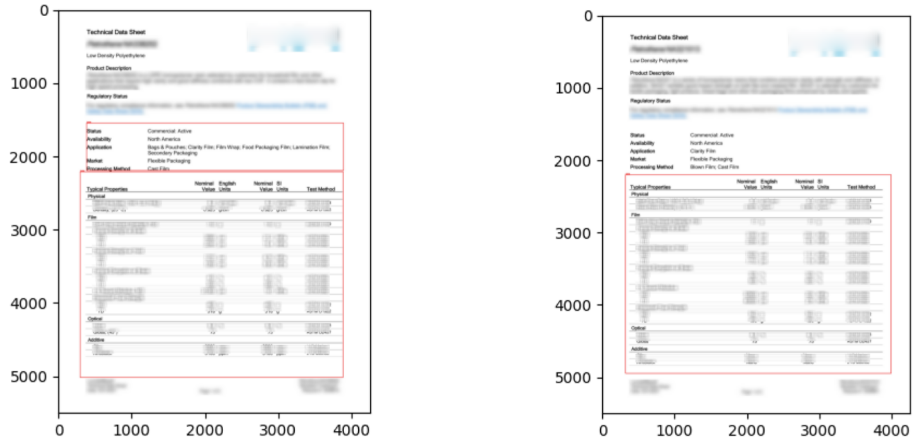
Dataset (randomly selected)	**Train dataset**	**Test dataset**
PubTabNet	5000 table images	100 table images
DVQA	5000 bar chart images	100 bar chart images

Table 2
The consolidated result of table detection models using transfer learning and Barlow Twins architecture on Di-Plast validation dataset – with different Intersection over Union (column *IoU*) thresholds w.r.t. overlap/intersection and union of the respective regions, c. f., [1, 2].

Metric	IoU	area	maxDets	ValueTL	ValueBT
AP	[0.50:0.95]	all	100	0.900	0.770
AP	0.50	all	100	1.000	0.987
AP	0.75	all	100	1.000	0.933
AP	[0.50:0.95]	small	100	-1.000	-1.000
AP	[0.50:0.95]	medium	100	-1.000	-1.000
AP	[0.50:0.95]	large	100	0.900	0.770
AR	[0.50:0.95]	all	1	0.542	0.447
AR	[0.50:0.95]	all	10	0.909	0.811
AR	[0.50:0.95]	all	100	0.909	0.811
AR	[0.50:0.95]	small	100	-1.000	-1.000
AR	[0.50:0.95]	medium	100	-1.000	-1.000
AR	[0.50:0.95]	large	100	0.909	0.811

In our experimentation in [2], we considered 5,100 random samples each from the PubTabNet [14] and DVQA [15] datasets to create train and test datasets for the Barlow Twins image classification model, c. f., Table 1. We trained the Barlow Twins model on a training dataset consists of 5,000 table images and 5,000 bar chart images. As there is no label in self-supervised learning, we considered the encoder (ResNet-50) of the Barlow Twins model and froze the model weight. For the evaluation of the model, we added the final classification layer on top of the encoder and trained only that final classification layer on the same training dataset (i. e., on 5,000 table images and 5,000 bar chart images). We evaluated the image classification model and obtained 93% accuracy on the test dataset, which consists of 100 table images and 100 bar chart images, c. f., [2] for a detailed discussion.

Subsequently, we performed supervised Faster R-CNN based table detection initialized with the pre-trained Barlow Twins image classification model weight on the Di-Plast training dataset. We evaluated table detection model on Di-Plast validation dataset and obtained nearly 77% mAP (mean average precision) of IoU (Intersection over Union) as presented in Table 2 in the *ValueBT* column [2]. The *ValueTL* column presents evaluation result of our transfer learning based table detection method [1]. Commonly three typical table detection errors are observed, such as, *partial-detection, un-detection, and mis-detection.* In partial-detection, only some part of the ground-truth table is predicted and some information is missing. Entire ground-truth table is not predicted in un-detection problem. Other components such as text blocks, figures or bar charts on document images are predicted as tables in mis-detection problem [3].



(a)

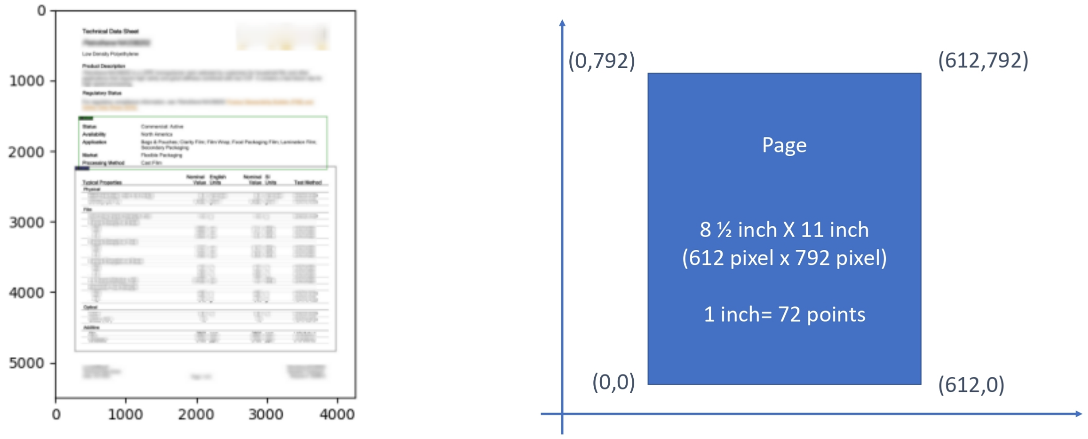
(b)

Figure 2: Inference results (with red-colored bounding box) of Faster R-CNN based table detection model with pre-trained Barlow Twins image classification model weight on two document images.

For each predicted table, we compute the IoU w.r.t. each ground-truth bounding box of *table* class on document images. Thereafter, Average Precision (AP) and Average Recall (AR) values are averaged over multiple IoU values. AP is averaged over all categories (here is only one category, e. g., *table* class), which is referred as mean average precision (mAP). No distinction are made between AP and mAP, and similarly between AR and mean average recall (mAR) in COCO (Common Objects in Context) evaluation metrics[5]. Values of -1.000 indicate that the metric cannot be computed, since no predictions are performed for small (with an area less than 32×32 pixels) and medium objects (with an area between 32×32 pixels and 96×96 pixels), because the area of each table is larger than 96×96 pixels in our Di-Plast dataset [1].

To exemplify our analysis and the respective predictions of Faster R-CNN based table detection model initialized with the pre-trained Barlow Twins image classification model weight [2], some exemplary instances are shown in Figure 2. Two tables are predicted as shown in Figure 2(a). We observe, that only one table is predicted and the second table is not predicted as shown in Figure 2(b). In the shown exemplary instances, we sketch the structure of the documents contained in the domain-specific Di-Plast dataset. We notice that Barlow Twins based table detection model suffers from partial-detection and un-detection problem on the Di-Plast test dataset. As partial table detection is observed in Figure 2(a), where some left part of the top table is not accurately predicted, hence textual information of the table could be missed during tabular data extraction. On the other hand, Figure 2(b) denotes un-detection problem, where the top table is not predicted at all.

[5] https://cocodataset.org/#detection-eval



(a) Document image coordinate system

(b) PDF page coordinate system

Figure 3: Overview of coordinate systems of a document image and a PDF page.

3.2. Application: From Document Table Detection to Tabular Data Extraction

In general, after extracting the bounding box information, we can apply tabular data extraction in order to enable information extraction afterwards, e. g., using knowledge-based approaches [26, 27, 28]. This then also enables ultimately knowledge management since then we can apply standardized representations for the extracted information.

For tabular data extraction based on transfer learning based document table detection approach, we have implemented a prototypical approach using the Camelot open-source tool to extract tabular data from PDF documents. Below, we sketch the basic idea of exploiting the region of interest information, being used for enabling data and information extraction. Essentially, we create a mapping between the bounding box pixel values of the document images and the relevant coordinate values of the PDF document pages. For the mapping, we have to align the coordinate systems of a PDF document page and of a document image. In the 2-dimensional coordinate system of a PDF page, the coordinate value (0,0) of a PDF page starts at the bottom-left corner.[6] In contrast, the pixel value (0,0) of a document image starts at the top-left corner. Figure 3 depicts the respective coordinate systems. The number of printed dots contained within 1 inch of an image printed by a printer is measured by dots per inch (dpi).[7]

During pre-processing, when we convert a PDF document to (potentially a set of) document images, we consider a dpi value of 72 dpi. Afterwards, the transfer learning based table detection model is used to infer the unseen document images to predict table images on those document images. We obtain the bounding box coordinate values of the predicted tables of the document images similar to Figure 3(a). Here, we also need to take into account that the dimension of each document image used in our tabular data extraction process is given as: width of 612 pixels and height of 792 pixels. In contrast, the the dimension of the image shown in Figure 3(a), has a width of 4250 pixels and height of 5500 pixels.

[6] https://www.pdfscripting.com/public/PDF-Page-Coordinates.cfm

[7] https://www.sony.com/electronics/support/articles/00027623

4. Results: Visual Explanations

In this section, we focus on different explanatory visualization methods in our context of coarse localization maps by leveraging Grad-CAM, Grad-CAM++ and Ablation-CAM. In Figure 4 and Figure 5, we focus on the structure of the documents and the localization maps. In general, Deep residual networks such as ResNets exhibit state-of-the-art performance in several challenging tasks in computer vision, which makes these models difficult to interpret. CAM is proposed to identify discriminative regions used by a restricted class of image classification CNN models which do not contain any fully-connected layer [20]. On the other hand, Grad-CAM offers existing state-of-the-art deep neural network models interpretable without altering their architecture. A good visual explanation for image classification model is considered as any target category or class label should be (1) class-discriminative, i. e., localize the category in the image, and (2) high-resolution, i. e., capture fine-grained details [10].

4.1. Transfer Learning based Model: Grad-CAM and Grad-CAM++

A table detection model does not contain only label information alike image classification, but also contains bounding box and score information. We leverage the Grad-CAM and Grad-CAM++ methods[8] to visualize coarse localization maps for the decisions made by our transfer learning based table detection model on Di-Plast test dataset. Our transfer learning based table detection model follows the Faster R-CNN architecture initialized with pre-trained TableBank table detection model referred in [1], which uses PyTorch Detectron2 library[9]. Figure 4(a) and Figure 4(b) exhibit coarse localization maps produced by Grad-CAM/Grad-CAM++ for our transfer learning based table detection model on the same document image. Red color indicates the areas in which the heatmaps have a higher intensity[10], which are expected to match with the location of the objects corresponding to the *table* class [29, 30, 31].

Our first visualization results indicate that the Grad-CAM++ method seems to perform better than the Grad-CAM method to visually explain the decisions of our transfer learning based table detection model. Most of the red region of the localization maps produced by Grad-CAM++ perfectly remain within the predicted bounding box of *table* class in the Di-Plast test dataset rather than for the Grad-CAM method. We use the JET colormap in Matplotlib [32] for Grad-CAM and Grad-CAM++ methods during visualization[11].

4.2. Barlow Twins based Model: Ablation-CAM

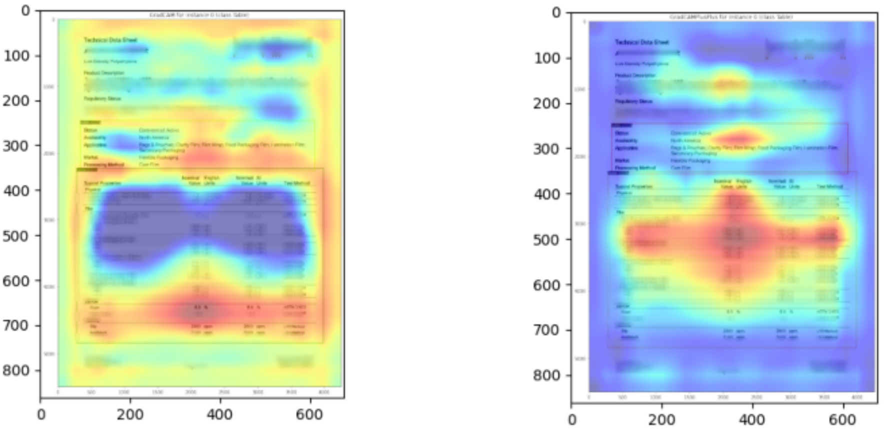
The Grad-CAM and Grad-CAM++ methods rely on gradients backpropagating from the output class nodes during visualization. Grad-CAM fails to offer reliable visual explanations for highly confident decisions due to gradient saturation. Several times, Grad-CAM highlights relatively incomplete and imperfect regions to detect multiple occurrences of same object in an image that may not be sufficient for the trustworthiness of CNN based models. Ablation-CAM, a *gradient free* method is proposed to visualize class-discriminative localization maps for

[8] https://github.com/alexriedel1/detectron2-GradCAM

[9] https://github.com/facebookresearch/detectron2

[10] https://www.oreilly.com/library/view/python-data-science/9781491912126/ch04.html

[11] https://matplotlib.org/stable/tutorials/colors/colormaps.html



(a) Table detection with Grad-CAM

(b) Table detection with Grad-CAM++

Figure 4: Visualization of coarse localization maps of transfer learning based table detection model on same document image

explaining decisions of CNN based models [12]. We use the Ablation-CAM method[12] to get a visual explanation of the decisions made by Faster R-CNN table detection model initialized with pre-trained Barlow Twins image classification model weight on Di-Plast test dataset, where the model uses PyTorch TorchVision library[13]. Figure 5(a) and Figure 5(b) show the table detection model inference result (with red colored bounding box) and corresponding coarse localization map produced by Ablation-CAM. The red colored region on localization map indicates the heatmap has a higher intensity, which are expected to match with the location of the objects corresponding to the *table* class [30, 33].

Here, it appears that the visual explanation for the table detection model produced by Ablation-CAM is not satisfactory. The reason might be that this table detection model obtained nearly 77% mAP of IoU compare to transfer learning based table detection model, which obtained nearly 90% mAP of IoU referred in [1]. We use the JET colormap in OpenCV [34] for the Ablation-CAM method during visualization[14]

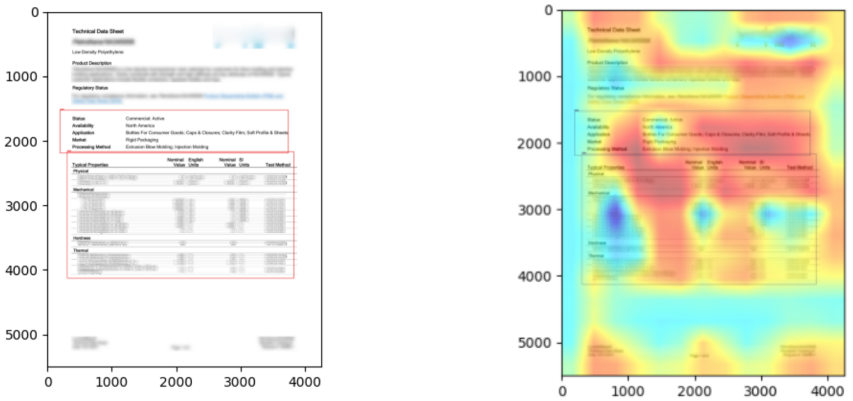
5. Conclusions

In this paper, we focused on techniques for tabular data extraction from PDF documents with the help of computer vision based deep neural networks. In particular, we investigated those approaches towards their explainability, which can then be applied for assessment, diagnostics, method refinement and tuning. Regarding the methods, we discussed inference results of Faster R-CNN based table detection model on document images, which is initialized with pre-trained Barlow Twins image classification model weight, building on our previous research [1, 2].

[12] https://github.com/jacobgil/pytorch-grad-cam

[13] https://pytorch.org/vision/stable/index.html

[14] https://docs.opencv.org/4.x/d3/d50/group__imgproc__colormap.html



(a) Table detection: bounding box

(b) Table detection: Ablation-CAM

Figure 5: Exemplary inference results (with red-colored bounding box) and coarse localization maps visualization for Faster R-CNN table detection model with pre-trained Barlow Twins model weight

We explored the visual explanations for the decisions made by the Barlow Twins based table detection model along with previously analyzed transfer learning based table detection model using Grad-CAM, Grad-CAM++ and Ablation-CAM methods. For Faster R-CNN based table detection model initialized with pre-trained Barlow Twins image classification model weight, we applied the gradient free Ablation-CAM method to visualize coarse localization map. In our first experiments, we observe the trend that here visual explanations of the decisions of this model is not adequately possible due to lower mAP of IoU value, as well as, for partial-detection and un-detection problems in our context. For transfer learning based table detection model, we utilized gradient based Grad-CAM and Grad-CAM++ methods. The red colored region on localization map indicates the heatmap with higher intensity value to match the location of the objects corresponding to *table* class. Grad-CAM++ method seem to offer better visual explanation of the decisions made by the transfer learning based table detection model compared to the Grad-CAM method.

For future work, we aim to explore semi-supervised and active learning based document table detection approaches for minimizing the manual image annotation effort on domain specific datasets. We also aim to concentrate on further methods for providing visual explanations of the decisions of such CNN based models for table detection. Furthermore, we intend to explore further benchmark datasets, e. g., PubLayNet [4] and TableBank [3] datasets. In addition, combining the tabular data extraction approach with further knowledge-based techniques, e. g., [26, 27, 28] indicates promising directions for future research.

Acknowledgments

This work has been funded by the Interreg North-West Europe program (Interreg NWE), project Di-Plast - Digital Circular Economy for the Plastics Industry (NWE729).

References

[1] A. G. Chowdhury, N. Schut, M. Atzmueller, A hybrid information extraction approach using transfer learning on richly-structured documents, in: T. Seidl, M. Fromm, S. Obermeier (Eds.), Proc. LWDA 2021 Workshops: FGWM, KDML, FGWI-BIA, and FGIR, volume 2993 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 13–25.

[2] A. Ghosh Chowdhury, M. b. Ahmed, M. Atzmueller, Towards Tabular Data Extraction From Richly-Structured Documents Using Supervised and Weakly-Supervised Learning, in: Proc. IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), IEEE, 2022.

[3] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, Z. Li, Tablebank: A benchmark dataset for table detection and recognition, arXiv preprint arXiv:1903.01949 (2019).

[4] X. Zhong, J. Tang, A. J. Yepes, Publaynet: largest dataset ever for document layout analysis, in: Proc. International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019, pp. 1015–1022.

[5] M. Alavi, D. E. Leidner, Knowledge management and knowledge management systems: Conceptual foundations and research issues, MIS quarterly (2001) 107–136.

[6] Q. Zhu, X. Cheng, The opportunities and challenges of information extraction, in: Proc. International Symposium on Intelligent Information Technology Application Workshops, IEEE, 2008, pp. 597–600.

[7] P. Năstase, D. Stoica, F. Mihai, A. Stanciu, From document management to knowledge management, Annales Universitatis Apulensis Series Oeconomica 11 (2009) 325–334.

[8] S. Furth, J. Baumeister, Semantification of large corpora of technical documentation, in: Enterprise Big Data Engineering, Analytics, and Management, IGI Global, 2016, pp. 171–200.

[9] S. A. Siddiqui, M. I. Malik, S. Agne, A. Dengel, S. Ahmed, Decnt: Deep deformable cnn for table detection, IEEE access 6 (2018) 74151–74161.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proc. IEEE international conference on computer vision, 2017, pp. 618–626.

[11] A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: Proc. IEEE winter conference on applications of computer vision (WACV), IEEE, 2018, pp. 839–847.

[12] H. G. Ramaswamy, et al., Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization, in: Proc. IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 983–991.

[13] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, in: Proc. International Conference on Machine Learning, PMLR, 2021, pp. 12310–12320.

[14] X. Zhong, E. ShafieiBavani, A. Jimeno Yepes, Image-based table recognition: data, model, and evaluation, in: European Conference on Computer Vision, Springer, 2020, pp. 564–580.

[15] K. Kafle, B. Price, S. Cohen, C. Kanan, Dvqa: Understanding data visualizations via question answering, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5648–5656.

[16] S. A. Siddiqui, A. Dengel, S. Ahmed, Self-supervised representation learning for document image classification, IEEE Access 9 (2021) 164358–164367.

[17] A. W. Harley, A. Ufkes, K. G. Derpanis, Evaluation of deep convolutional nets for document image classification and retrieval, in: Proc. International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2015, pp. 991–995.

[18] M. Z. Afzal, A. Kölsch, S. Ahmed, M. Liwicki, Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification, in: Proc. IAPR International Conference on Document Analysis and Recognition (ICDAR), volume 1, IEEE, 2017, pp. 883–888.

[19] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object detectors emerge in deep scene cnns, arXiv preprint arXiv:1412.6856 (2014).

[20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[21] P. Melnyk, Z. You, K. Li, A high-performance cnn method for offline handwritten chinese character recognition and visualization, soft computing 24 (2020) 7977–7987.

[22] D. Zhang, J. Han, G. Cheng, M.-H. Yang, Weakly supervised object localization and detection: A survey, IEEE transactions on pattern analysis and machine intelligence (2021).

[23] J. Kim, J. Jang, S. Seo, J. Jeong, J. Na, N. Kwak, Mum: Mix image tiles and unmix feature tiles for semi-supervised object detection, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14512–14521.

[24] D. Zhang, W. Zeng, J. Yao, J. Han, Weakly supervised object detection using proposal- and semantic-level relationships, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[25] G. Cheng, J. Yang, D. Gao, L. Guo, J. Han, High-quality proposals for weakly supervised object detection, IEEE Transactions on Image Processing 29 (2020) 5794–5804.

[26] M. Atzmueller, P. Kluegl, F. Puppe, Rule-Based Information Extraction for Structured Data Acquisition using TextMarker, in: Proc. LWA 2008 (KDML Track), University of Wuerzburg, Wuerzburg, Germany, 2008, pp. 1–7.

[27] P. Kluegl, M. Atzmueller, F. Puppe, Meta-Level Information Extraction, in: The 32nd Annual Conference on Artificial Intelligence, Springer, Berlin, 2009. (233–240).

[28] P. Kluegl, M. Atzmueller, F. Puppe, TextMarker: A Tool for Rule-Based Information Extraction, in: Proc. Unstructured Information Management Architecture (UIMA) 2nd UIMA@GSCL Workshop, Conference of the GSCL, 2009.

[29] M. Lerma, M. Lucas, Grad-cam++ is equivalent to grad-cam with positive gradients, arXiv preprint arXiv:2205.10838 (2022).

[30] C. Molnar, Interpretable machine learning, Lulu. com, 2020.

[31] J. VanderPlas, Python data science handbook: Essential tools for working with data, " O'Reilly Media, Inc.", 2016.

[32] N. Rougier, Matplotlib tutorial, Ph.D. thesis, INRIA, 2012.

[33] I. Culjak, D. Abram, T. Pribanic, H. Dzapo, M. Cifrek, A brief introduction to opencv, in: 2012 Proc. 35th international convention MIPRO, IEEE, 2012, pp. 1725–1730.

[34] T. T. Santos, Scipy and opencv as an interactive computing environment for computer vision., Revista de Informática Teórica e Aplicada 1 (2015) 154–189.