Using CNNs to Detect Graphical Representations of Structural Equation Models in IS Papers

Tobias Genz, Burkhardt Funk

Leuphana University Lüneburg, Institute of Information Systems, Lüneburg, Germany
tobias.genz@stud.leuphana.de, funk@uni.leuphana.de

Abstract. Literature reviews are an essential but time-consuming part of every research endeavor and play an important role in the quality of the research findings. Traditional tools and literature databases only make use of the textual information and do not consider graphical representations like figures of structural equation models (SEMs). These models are often used in empirical studies to visualize theoretical models and key results. We design and implement an application for image recognition to simplify the search for relevant papers, by automatically recognizing SEM figures in scientific papers stored as PDF files. To classify whether a page in a paper contains an SEM figure we make use of convolutional neural networks and achieve an F_1 score of 98,7% together with a recall of 100% for the SEM class. We further describe how we intend to automatically extract information from these SEM figures.

Keywords: Structural equation models, deep neural networks, information extraction, literature review

1 Introduction

The number of scientific publications grows steadily. The plethora of valuable scientific literature bears a huge base of knowledge. However, synthesizing and utilizing this knowledge is one of the greatest challenges of science. Bong et al. [1] highlight the huge potential that mining this knowledge with machine learning techniques can have.

Structural equation models (SEMs) are often used when documenting knowledge from empirical studies and building theories within scientific fields [2]. They are a hypothesis-driven statistical method to describe correlations and dependencies between theoretical constructs. One explicit challenge when employing SEMs is the so-called jingle and jangle fallacy [3]. The two fallacies describe unrecognized construct overlaps, where either different names describe the same latent construct (jangle), or the same name is used for different theoretical constructs (jingle). This poses a great challenge as it contributes to the complexity of performing literature reviews and thereby hampers theory development. We argue that technical solutions could support researchers in reviewing existing literature and developing new theories. Following the design science paradigm [4] we propose the following artifact:

15th International Conference on Wirtschaftsinformatik,
March 08-11, 2020, Potsdam, Germany

https://doi.org/10.30844/wi_2020_a8-genz

We design and implement a software application, based on convolutional neural networks (CNNs), that is able to recognize figures in scientific papers that represent SEMs. We evaluate the model on papers from the basket of eight journals. As a next step, we plan to localize the position of an SEM figure within one page, to obtain accurate images of the figures. Eventually, we intend to use those SEM images to automatically extract construct relations and path coefficients from SEM figures. That way not only semi-automated meta-reviews could be created. Instead, by connecting various studies, statistics for paths that were found insignificant in individual studies could be aggregated to identify significant relationships between constructs, thereby fostering theory development.

The paper is structured as follows: We provide a short overview of related research in the area of knowledge discovery and approaches to simplify literature review for researchers. Then we describe the methodology of our approach and apply it to a set of papers from the IS field. Finally, we discuss the limitations of the proposed approach and give an outlook on the ongoing work.

2 Related Research

We identified two main branches of research, that try to automatically infer knowledge from scientific literature. First, a variety of natural language processing (NLP) approaches has been proposed [3, 5–8]. Second, computer vision systems have been developed which extract information from figures and graphics [9, 10].

An NLP-based tool that deals directly with SEMs is presented by Bong et al. [1]. They create a network of constructs named ConstructNet, discovering relationships between constructs. It makes use of machine learning to calculate a similarity score for two constructs. This unsupervised approach allows exploring relations between constructs that have not been studied. Larsen and Bong [3] create a tool to address the construct identity fallacy (CIF) and thereby help researchers with literature reviews, support them in conducting a meta-analysis and evaluating the validity of constructs.

Considering approaches of knowledge extraction from graphical representations in research papers we find a variety of established systems like ParsCit or OCR++ [11, 12]. These tools are capable of extracting captions, references and other literature meta information; however, they cannot recognize and extract whole figures or tables from a paper. Other researchers use handcrafted features or heuristics to segment different parts of a PDF file and leverage the information contained in figures and tables [10, 13]. More recent approaches try to utilize deep learning techniques like CNNs and pixel-wise segmentation for this task [9, 14]. CNNs have proven to work well on image data, as has been shown on various large datasets [9, 15]. To the best of our knowledge, there is no tool available that allows to explicitly extract knowledge from graphical representations of SEMs. Our contribution aims at this research gap.

https://doi.org/10.30844/wi_2020_a8-genz

3 Method and Application

To identify scientific papers that include graphical representations of SEMs, we propose the following process (see Figure 1). First, convert all pages from the papers into images. Second, classify images from the step 1 and return a confidence score for each image to contain an SEM figure. Third, localize SEM figures on pages which were identified to contain at least one such figure, in order to return a clean cropped figure. The final fourth step is to extract the constructs, their relations and path coefficients and store this information in a database. In this paper we focus on the first two steps and present the results of the classification task. Additionally, we give an outlook on how we intend to proceed with the localization problem.

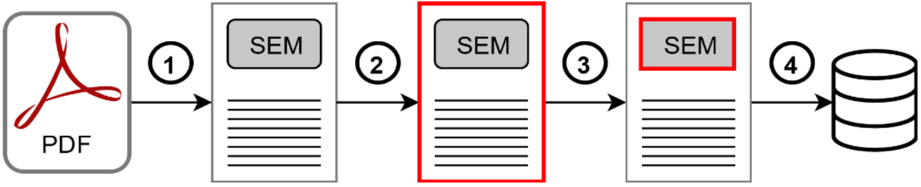

Figure 1. Process Model

For our analysis we use a set of 203 scientific papers (which we share upon request) that we obtain by systematically applying a keyword-based search on the eight leading IS journals, the basket of eight. We manually screen the papers and mark the pages containing SEM figures. The PDF files are transformed to get one image per page. Overall, we obtain 4437 images of which 491 contain at least one SEM figure. We use a hold-out sample of 42 papers (842 images) for the final test set to evaluate the out-of-sample performance of the learned model. Of the remaining papers we use 25% as a validation set to perform hyperparameter tuning. The remaining 2696 images are used to train a CNN classifier that we implement with Keras [16] using the Tensorflow backend [17]. We evaluated three network architectures (Xception, ResNet, Mobile-Net [18–20]) and made use of transfer learning. This is, we used these architectures and weights which where pretrained on more than one million images of the ImageNet dataset [15]. To adjust the architectures to our problem setting we added a global average pooling layer, as well as a final dense layer with one unit and sigmoid activation for our binary classification. We found that global average pooling is less prone to overfitting when compared to fully connected dense layers.

The model is trained for different image-sizes (we tried 128^2, 256^2 and 512^2 pixels). We also make use of different data-augmentation techniques, as it can help to reduce overfitting [21]. To improve learning for our highly imbalanced dataset with only about 11% of the images containing an SEM figure, we apply class weights to balance how much a sample contributes to the loss. For training we initially freeze all layers of the backbone architecture and only train our added classification layers for 10 epochs. Afterwards we unfreeze all but the first few layers (depending on the architecture) and finetune the model with a ten times smaller learning rate for another 20 epochs. This way we make sure that we do not unlearn relevant features, that are

https://doi.org/10.30844/wi_2020_a8-genz

contained within the pretrained model. For the evaluation on the final test set, we use the model with the best validation score for the above training settings. We obtained the best results for images resized to 512^2 pixels with using the Xception architecture and augmenting the data by using mirroring and rotation. Further augmentation like random crops or zooms, as well as using different aspect ratios to keep the original DIN A4 layout, led to slightly worse performance and were not used for the final model.

4 Results and Discussion

To evaluate the results on the test set we mainly use recall and the F_1 score. We especially care for the recall of classifying a page to contain an SEM. To let the model favor for recall we use a threshold parameter, which indicates the minimum confidence score for the model to predict an SEM. For a baseline, we use a model that always predicts no SEM. The majority class, being images that contain no SEM figures, accounts for about 89% of the test samples which corresponds to a weighted F_1 score of 84,1% for the baseline model. Our basic model without augmented data achieves a recall for the SEM class of 95,6% and a weighted F_1 score of 98,8% which exceeds the baseline F_1 score by more than 14 percentage points. While using the augmented dataset yields a slightly worse F_1 score of 98,7%, it robustly achieved a recall of 100% for the SEM class. This shows that given enough training data, the model is certainly able to identify figures within the papers. Examples of false positives in the test set were exclusively related to other figure types that included boxes and/or arrow like shapes. A potential pitfall of our learned model can be associated with the data used for training. As we have only a small sample size of few selected journals it is possible that the model specifically matches formats used in these journals and performs worse on unseen papers with different layouts. We test the model on two further scenarios. First, we fully omit one specific journal in the training process and instead use this journal as a test set later. Second, we test the model on about 30 new documents with non-standardized layouts, mainly unpublished papers or tutorials on SEMs. In both scenarios the model maintained the recall of 100% with each time about 0.7% points worse F_1 score. Overall the results show that we can identify SEM figures accurately with a very high recall.

5 Conclusion and Outlook

We devise and implement a software application that helps to identify SEMs in scientific literature. Our trained model allows performing page-wise prediction for the existence of SEMs in PDF files with an F_1 score of 98,7% and a recall of 100% for the SEM class. It thereby already eases literature review when searching for visual representations of SEMs and could support the screening of massive amounts of papers from other journals and proceedings.

As described for the process flow, the next step will be to additionally predict the location of SEM figures. There are several successful architectures for object

https://doi.org/10.30844/wi_2020_a8-genz

detection [22]. We mainly care about the best prediction quality and have no high demands for fast detection speed, as it is needed in other fields like autonomous driving. As a promising approach for these demands, we investigate the R-CNN architecture, which is a region proposal-based method [22]. Learning the location of SEM figures in scientific literature will facilitate to extract information of these figures, precisely to identify and extract constructs and path coefficients. We believe that a simple and systematical access to these condensed research insights will not only ease literature reviews, but also allows to make a first step into semi-automated theory development.

References

1. Bong, C.H., Larsen, K.R., James, M.: A large scale knowledge integration leading to human decision making. In: 2012 IEEE Symposium on Computers Informatics (ISCI), pp. 22–27 (2012)
2. Urbach, N., Ahlemann, F.: Structural equation modeling in information systems research using partial least squares. Journal of Information technology theory and application 11, 5–40 (2010)
3. Larsen, K.R., Bong, C.H.: A Tool for Addressing Construct Identity in Literature Reviews and Meta-Analyses. MIS Quarterly 40, 529–551 (2016)
4. Gregor, S., Hevner, A.R.: Positioning and Presenting Design Science Research for Maximum Impact. MIS Q 37, 337–355 (2013)
5. Mueller, R., Abdullaev, S.: DeepCause: Hypothesis Extraction from Information Systems Papers with Deep Learning for Theory Ontology Learning. Proceedings of the 52nd Hawaii International Conference on System Sciences (2019)
6. Sturm, B., Sunyaev, A.: You Can't Make Bricks Without Straw: Designing Systematic Literature Search Systems. In: ICIS 2017: Proceedings of the International Conference on Information Systems (2017)
7. Mueller, R.M., Huettemann, S.: Extracting Causal Claims from Information Systems Papers with Natural Language Processing for Theory Ontology Learning. Proceedings of the 51st Hawaii International Conference on System Sciences (2018)
8. Bosco, F., Steel, P., Oswald, F., Uggerslev, K., Field, J.: Cloud-based Meta-analysis to Bridge Science and Practice: Welcome to metaBUS. PAD 1 (2015)
9. Siegel, N., Lourie, N., Power, R., Ammar, W.: Extracting Scientific Figures with Distantly Supervised Neural Networks. Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL 18, 223–232 (2018)
10. Clark, C., Divvala, S.: PDFFigures 2.0. Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL 16, 143–152 (2016)
11. Singh, M., Barua, B., Palod, P., Garg, M., Satapathy, S., Bushi, S., Ayush, K., Sai Rohith, K., Gamidi, T., Goyal, P., et al.: OCR++: A Robust Framework For Information Extraction from Scholarly Articles. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 3390–3400 (2016)

12. Councill, I.G., Giles, C.L., Kan, M.-Y.: ParsCit: an Open-source CRF Reference String Parsing Package. In: LREC (2008)
13. Ray Choudhury, S., Mitra, P., Giles, C.L.: Automatic Extraction of Figures from Scholarly Documents. Proceedings of the 2015 ACM Symposium on Document Engineering - DocEng 15, 47–50 (2015)
14. Stahl, C.G., Young, S.R., Herrmannova, D., Patton, R.M., Wells, J.C.: DeepPDF: A Deep Learning Approach to Extracting Text from PDFs (2018)
15. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vision 115, 211–252 (2015)
16. Chollet, F. and others: Keras (2015)
17. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: TensorFlow: A System for Large-scale Machine Learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, pp. 265–283. USENIX Association, Berkeley, CA, USA (2016)
18. Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto and Hartwig Adam: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications (2017)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
20. Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807 (2017)
21. Wong, S.C., Gatt, A., Stamatescu, V., McDonnell, M.D.: Understanding Data Augmentation for Classification: When to Warp? 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 1–6 (2016)
22. Zhao, Z.-Q., Zheng, P., Xu, S.-T., Wu, X.: Object Detection With Deep Learning: A Review. IEEE transactions on neural networks and learning systems (2019)

https://doi.org/10.30844/wi_2020_a8-genz