Flowsheet Recognition using Deep Convolutional Neural Networks

Balhorn, Lukas Schulze; Gao, Qinghe; Goldstein, Dominik; Schweidtmanna, Artur M.

DOI
10.1016/B978-0-323-85159-6.50261-X

Publication date
2022

Document Version
Final published version

Published in
Computer Aided Chemical Engineering

Proceedings of the 14th International Symposium on Process Systems Engineering – PSE 2021+
June 19–23, 2022, Kyoto, Japan
http://dx.doi.org/10.1016/B978-0-323-85159-6.50261-X

Flowsheet Recognition using Deep Convolutional Neural Networks

Lukas Schulze Balhorn[a], Qinghe Gao[a], Dominik Goldstein[b], Artur M. Schweidtmann[a,*]

[a]*Department of Chemical Engineering, Delft University of Technology, Van der Maasweg 9, Delft 2629 HZ, The Netherlands*
[b]*Aachener Verfahrenstechnik - Process Systems Engineering, RWTH Aachen University, Aachen 52062, Germany*

a.schweidtmann@tudelft.nl

Abstract

Flowsheets are the most important building blocks to define and communicate the structure of chemical processes. Gaining access to large data sets of machine-readable chemical flowsheets could significantly enhance process synthesis through artificial intelligence. A large number of these flowsheets are publicly available in the scientific literature and patents but hidden among innumerable other figures. Therefore, an automatic program is needed to recognize flowsheets. In this paper, we present a deep convolutional neural network (CNN) that can identify flowsheets within images from literature. We use a transfer learning approach to initialize the CNN's parameter. The CNN reaches an accuracy of 97.9% on an independent test set. The presented algorithm can be combined with publication mining algorithms to enable an autonomous flowsheet mining. This will eventually result in big chemical process databases.

Keywords: Flowsheet, Data Mining, Image Classification, Deep Learning, Transfer Learning

1. Introduction

In recent years, machine learning (ML) has emerged as a popular method to solve complex problems in various domains. This popularity has predominantly been driven by (i) the increase of computational power, (ii) the improvement of ML algorithms, and (iii) the availability of big data (LeCun *et al.*, 2015). Chemical engineering has already seen many successful applications of ML (Schweidtmann *et al.*, 2021; Venkatasubramanian, 2018). However, literature on the structural synthesis of chemical processes through ML is scarce (c.f. (d'Anterroches & Gani, 2005; Zhang *et al.*, 2018; Oeing *et al.*, 2021)). While a variety of promising ML methods exist, big chemical process data is missing (Schweidtmann *et al.*, 2021; Weber *et al.*, 2021). We argue that this lack of structured chemical process data is hindering further progress of ML developments for chemical process synthesis.

The topological information about chemical processes is usually communicated through flowsheets. Flowsheets are technical drawings describing the unit operations connectivity of a process. There exists at least one flowsheet for every chemical process. Eventhough

most flowsheets are only available in internal company reports, a large number of flowsheets are also publicly available in scientific publications and patents. These flowsheets are mostly depicted on figures in PDF documents. However, searching for the flowsheet figures in scientific publications and patents can be as difficult as looking for a needle in a haystack. In particular, a manual search through the enormous amount of available literature would not only be a slow and labor-intense process, but it would also be prone to errors. Therefore, an algorithm is needed that autonomously recognizes flowsheet images.

In the previous literature, information extraction from scientific literature has mostly focused on text mining using natural language processing (Hong *et al.*, 2021; Nasar *et al.*, 2018). In the context of chemistry for example, Swain & Cole (2016) developed the ChemDataExtractor which extracts chemical identifiers, spectroscopic attributes, and chemical property attributes from scientific literature. Furthermore, information extraction from scholarly images has been performed in the past. The majority of research on the classification of scientific images has been conducted on biomedical literature pushed by the yearly ImageCLEF challenge (c.f. (Pelka *et al.*, 2020)). Furthermore, a few works exist in chemistry on information extraction from images. This works mostly focus on the recognition and digitization of structural formulas (Tharatipyakul *et al.*, 2012; Beard & Cole, 2020). Another example for chemical image analysis is the ImageDataExtractor which mines microscopy images to extract information about the particle sizes and shapes (Mukaddem *et al.*, 2019). However, to the best of our knowledge, image classification has not been applied to chemical process design literature and there exists no previous algorithm that identifies chemical flowsheet images.

In this work, we propose an algorithm that recognizes flowsheet images from chemical engineering journal articles. The proposed algorithm will contribute to our long-term vision to build a database of chemical processes. In Section 2., we provide a brief background on Convolutional Neural Networks (CNNs). In Section 3., we present our methods, data set, and pre-processing. In Section 4., we evaluate the performance of the proposed flowsheet image classification model and discuss the results. Finally, we conclude our findings in Section 5.

2. Deep Convolutional Neural Networks

Inspired by the biological visual system (O'Shea & Nash, 2015), deep CNNs have been proposed as a computational method to bridge the gap between the capabilities of humans and machines for high-level tasks such as image classification, text recognition, and speech recognition (LeCun *et al.*, 2015). The powerful performance of deep CNNs in advanced tasks is achieved through the layout of the framework, which generally consists of three parts: Convolutional layers, pooling layers, and fully-connected layers. Convolutional layers contain a set of learnable filters that will convolve over the inputs to extract the underlying features. Intuitively, simple features such as edges, corners, and blotches will be detected in the early convolutional layers. Ultimately, more complex patterns such as 'unit operations' will appear with further layers. Pooling layers are usually periodically inserted between two convolutional layers to reduce the spatial dimension and the number of parameters. Average pooling and max pooling are the most common choices. Fully-connected neural network layers play the role of mapping the learned "distributed feature representation" to the sample label space, namely, making a classification. Additionally,

to introduce nonlinearity into the output, activation functions such as sigmoid, ReLu, and hyperbolic tangent are usually included after convolutional or fully-connected layers. Furthermore, the size of the training data is an important factor for the performance of the deep CNNs and data-labeling is often expansive. Therefore, the concept of transfer learning emerged in recent years. In transfer learning, the CNN is first trained with a sufficiently big data set from one domain of interest. Afterward, the data set of the classification task from another domain of interest is used to fine-tune the CNN.

3. Method

The flowsheet recognition algorithm aims to identify flowsheets among a large number of images. We train a deep CNN for the recognition algorithm based on manually labeled images mainly from scientific journal articles.

3.1. Data Set

At present, no public data set of flowsheet images exists. To create a training data set, we automatically mine figures of scientific journal articles. First, we retrieve a list of all DOIs for a given journal ISSN from the crossref API. Then, the PDFs of the corresponding journal articles are downloaded through publisher APIs. Subsequently, all figures are extracted from the PDFs using the Python package PyMuPDF. The describe procedure is applied to the journals "Theoretical Foundations of Chemical Engineering" and "Frontiers of Chemical Science and Engineering" to generate an initial dataset. Subsequently, the extracted images are manually reviewed and labeled as being a flowsheet or not. In addition to the figures from scientific journal articles, we also add flowsheet images retrieved from a google search to our data set. In total, our data set contains about 1,000 flowsheet images and about 13,000 other images from scientific publications.

3.2. Data Augmentation and Oversampling

As a result of the data mining from journal articles, the data set is imbalanced. In particular, there exist far fewer flowsheet images than other images. This imbalance can cause the classifier to develop a bias towards the majority class. To overcome this issue, oversampling has been used in previous studies (Johnson & Khoshgoftaar, 2019). We oversample the flowsheet images by a factor of 13 to balance the data set. As this large oversampling factor can cause overfitting, we also employ a data augmentation technique (Shorten & Khoshgoftaar, 2019). Each copy of a flowsheet image is augmented by stretching it along the horizontal and vertical axis independently by a random factor between 0.7 and 1.2. Other common data augmentation techniques such as shifting, rotation, and shearing were dismissed because they are expected to destroy some key features of flowsheet images. For example, flowsheets usually include horizontal and vertical lines making image rotation pointless. The images of the negative "other" class are not augmented because of abundant data availability.

3.3. Model Training

The CNN architecture for the flowsheet recognition is based on the VGG16 network by Simonyan & Zisserman (2014). The network includes 13 convolutional layers, 5 max

pooling layers, and 3 fully-connected layers. Since our data set is limited, we use a transfer learning approach. In particular, we use the publicly available VGG16 network that has been pre-trained on the ImageNet data set including tens of millions of images and 1,000 categories. To adapt the network to the use case of this work, we reduced the number of nodes in the output layer to two. The training is conducted using the PyTorch framework which is built on the Torch library. The model takes in images with a resolution of 224 × 224 pixels. We randomly divide our data set into training (70%), validation (15%), and test (15%) data set. The model is trained on the training data in batches of 150 images. The validation set was used to validate the training progress and tune the hyperparameters of the model. The independent test data set is used for the final performance evaluation. Notably, the test set is truly independent as it does not contain any augmented images from the training or validation sets.

4. Results and Discussions

The most important performance metrics for classifiers is the accuracy as defined in Eq. 1. In the light of class imbalance, we also evaluate the precision (Eq. 2) and recall (Eq. 3):

$$Accuracy = \frac{TN + TP}{TP + FP + TN + FN}, \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

where TP denotes true positive, TN denotes true negative, FP denotes false positive and FN denotes false negative. The training history is shown in Fig. 1. The classifier reaches a satisfying accuracy already after the first epoch. This good initial performance can be explained by the use of a pre-trained model. After the second epoch of training, the classifier shows a validation accuracy of over 98%. The training process was ended after 10 epochs. In training runs with more epochs no further improvement was experienced. The final training accuracy after 10 epochs is 98.1% while the validation accuracy is 98.2%. Notably, we do not observe any overfitting behavior in the training process.

Overall, the flowsheet recognition algorithm shows a promising performance on the independent test set. The confusion matrix on the test set is shown in Table 1. Of all predictions on the test set, 97.9% were correct. Furthermore, the precision is 80.7% and lower than the recall with 94.4%. The high recall shows that almost all flowsheet images are retrieved while the number of false negative flowsheets is very low. Furthermore, the fairly low precision could be explained by the class imbalance. The data set contains about thirteen times more images of the class “other”. If only a small fraction of the class “other” is misclassified, these images already make up a great share of the flowsheet predictions.

Finally, the runtime of the image classification is investigated. The evaluation of an image by the trained CNN takes about 7 milliseconds on average on a personal computer. This short evaluation time allows for an online application that autonomously mines flowsheets from literature.

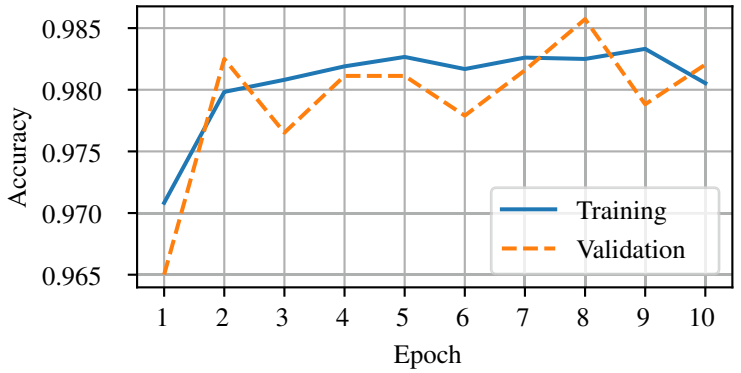

Figure 1: Training history of the CNN.

Table 1: Confusion matrix for the flowsheet recognition algorithm on the test set.

	Actual flowsheet	Actual other
Predicted flowsheet	151	36
Predicted other	9	1,976

5. Conclusions

We propose an image classification algorithm that can recognize flowsheet images. The algorithm consists of a deep CNN which classifies images with a high accuracy of 97.9%. In order to train the CNN, we mined about 1,000 flowsheet images from scientific literature and online search engines. Moreover, the transfer learning improved the prediction accuracy. The proposed tool can be used to automatically identify flowsheet images from scientific literature or other sources within a few milliseconds. In a preliminary study we applied our mining algorithm to the journal "Computers & Chemical Engineering" and identified more than 1500 flowsheets. Future work will digitize the flowsheet images to identify process topologies. This will eventually result in an open-source knowledge graph database providing chemical processes in a structured format. We believe that this database has a tremendous value for future process design because it allows the search and optimization over existing processes. In addition, our database will eventually serve as a training database for advanced ML algorithms able to design novel processes.

References

Beard, Edward J., & Cole, Jacqueline M. 2020. ChemSchematicResolver: A Toolkit to Decode 2D Chemical Diagrams with Labels and R-Groups into Annotated Chemical Named Entities. *Journal of Chemical Information and Modeling*, **60**(4), 2059–2072.

d'Anterroches, Loïc, & Gani, Rafiqul. 2005. Group contribution based process flowsheet synthesis, design and modelling. *Fluid Phase Equilibria*, **228-229**(Feb.), 141–146.

Hong, Zhi, Ward, Logan, Chard, Kyle, Blaiszik, Ben, & Foster, Ian. 2021. Challenges and Advances in Information Extraction from Scientific Literature: a Review. *JOM*, Oct.

Johnson, Justin M., & Khoshgoftaar, Taghi M. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, **6**(27).

LeCun, Yann, Bengio, Yoshua, & Hinton, Geoffrey. 2015. Deep learning. *Nature*, **521**(May), 436–444.

Mukaddem, Karim T., Beard, Edward J., Yildirim, Batuhan, & Cole, Jacqueline M. 2019. ImageDataExtractor: A Tool To Extract and Quantify Data from Microscopy Images. *Journal of Chemical Information and Modeling*, **60**(5), 2492–2509.

Nasar, Zara, Jaffry, Syed Waqar, & Malik, Muhammad Kamran. 2018. Information extraction from scientific articles: a survey. *Scientometrics*, **117**(3), 1931–1990.

Oeing, Jonas, Henke, Fabian, & Kockmann, Norbert. 2021. Machine Learning Based Suggestions of Separation Units for Process Synthesis in Process Simulation. *Chemie Ingenieur Technik*, Sept.

O'Shea, Keiron, & Nash, Ryan. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Pelka, Obioma, Friedrich, Christoph M, García Seco de Herrera, Alba, & Müller, Henning. 2020 (Sept.). Overview of the ImageCLEFmed 2020 Concept Prediction Task: Medical Image Understanding. *In: Proceedings of the CLEF 2020-Conference and labs of the evaluation forum*.

Schweidtmann, Artur M., Esche, Erik, Fischer, Asja, Kloft, Marius, Repke, Jens-Uwe, Sager, Sebastian, & Mitsos, Alexander. 2021. Machine Learning in Chemical Engineering: A Perspective. *Chemie Ingenieur Technik*, Oct.

Shorten, Connor, & Khoshgoftaar, Taghi M. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, **6**(60).

Simonyan, Karen, & Zisserman, Andrew. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Swain, Matthew C., & Cole, Jacqueline M. 2016. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling*, **56**(10), 1894–1904.

Tharatipyakul, Atima, Numnark, Somrak, Wichadakul, Duangdao, & Ingsriswang, Supawadee. 2012. ChemEx: information extraction system for chemical data curation. *BMC Bioinformatics*, **13**(17).

Venkatasubramanian, Venkat. 2018. The promise of artificial intelligence in chemical engineering: Is it here, finally? *AIChE Journal*, **65**(2), 466–478.

Weber, Jana M., Guo, Zhen, Zhang, Chonghuan, Schweidtmann, Artur M., & Lapkin, Alexei A. 2021. Chemical data intelligence for sustainable chemistry. *Chemical Society Reviews*.

Zhang, Tong, Sahinidis, Nikolaos V., & Siirola, Jeffrey J. 2018. Pattern recognition in chemical process flowsheets. *AIChE Journal*, **65**(2), 592–603.