Classifying images in chemical patent documents using Convolutional Neural Networks

Fatah Fattah

July 2021

1 Abstract

Images that are contained in chemical patent documents are a key information source that can be used to extract valuable information for chemistry database systems. Chemistry experts can already visually infer such information (i.e. chemical composition) from these images, however, this is not a viable way of processing the vast amount of information available. Deep learning approaches have become increasingly popular to tackle the image retrieval from chemistry images, and with this work we explore contribute to this by tackling a novel classification task. We observe that a simple Convolutional Neural Network has the ability to solve the classification task with high accuracy for a single patent office dataset, but does not generalize well to other patent offices by-default. The results are promising and make a sound argument to further pursue the possibilities for applying deep-learning on chemistry patent images.

2 Introduction

Chemical patent documents aim to legally protect an invention in the chemical or pharmaceuticals industry and are the prime source of information regarding new compounds/drugs. The documents are however not written from a scientific perspective but rather as legal documents that try to cover as many aspects of the new information as possible so that they cover all their bases from a legal standpoint. Because of this legality characteristic of the documents, there are many chemical compounds described in the papers that are (often loosely) related to the main compound(s) that the patent is meant to protect [1]. Therefore searching through the text can be a time consuming process in which one will have to read through many related compounds to get to the actual compound(s) that are being patented [1].

The always increasing volume of patent information has become a key challenge for knowledge management systems which try to discover and utilize this information. This has lead to tasks such as patent search and analysis to become more important than ever. Analyzing these patents is important for organisations for numerous purposes such as [2]:

- Determining novelty in patents.
- Analyzing patent trends.
- Forecasting technological developments in a particular domain.
- Extracting the information from patents for identifying the infringements.
- Identifying the promising patents.
- Identifying technological competitors.

Furthermore, the patents are key information sources for commercial chemical substance databases and chemistry literature search databases such as Reaxys [3], which is a solution developed by the host organi-

zation Elsevier. Its goal is to make information from chemistry literature available to the field of research so that one does not have to delve through the vast amount of information that is out there.

Scientific literature often uses images to provide information about certain topics. In some cases because words do not suffice and often because visuals are a great way to present large amounts of data in an easily understandable way with a relatively small amount of document space. Images in chemical patents are especially interesting because, if they depict one or more chemical structures, they are often closely related to the main compound(s) or they could even be the key compound itself. Whereas the textual descriptions often contain huge amounts of information that are mainly there for legal reasons [1]. Because of this, images are a key input signal for searching through chemical patents and make the above mentioned purposes easier to realize [1]. See figure 1 for an example of such a chemical structure.

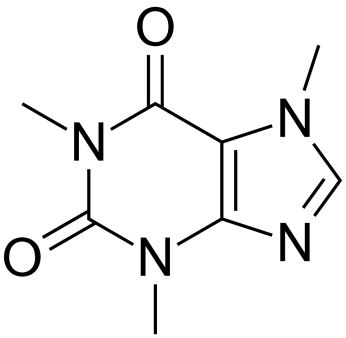

Figure 1: Example of a 2d chemical structure (caffeine).

The task of recognizing chemical structures within images is called Optical Chemical Structure Recognition (OCSR) and has historically mainly been tackled using rule-based approaches where the input image is vectorized after which the vectors and nodes are interpreted as bonds and atoms [4]. These solutions are sometimes open sourced (*i.e.*, OSRA [5]) and often closed sourced (*i.e.*, Chemgrapher [6]). More recently successful efforts have been made to tackle OCSR by applying deep learning which has had promising results. Examples of this are Chemgrapher [6], DECIMER [7] and MSE-DUDL [8]. However, these approaches mostly focus on extracting information, given an image for which it is certain that it contains a chemical structure depiction. The step which identifies which images contain chemical structures and which do not, is missing in these solutions, possibly increasing the chances of false positives or irrelevant outputs.

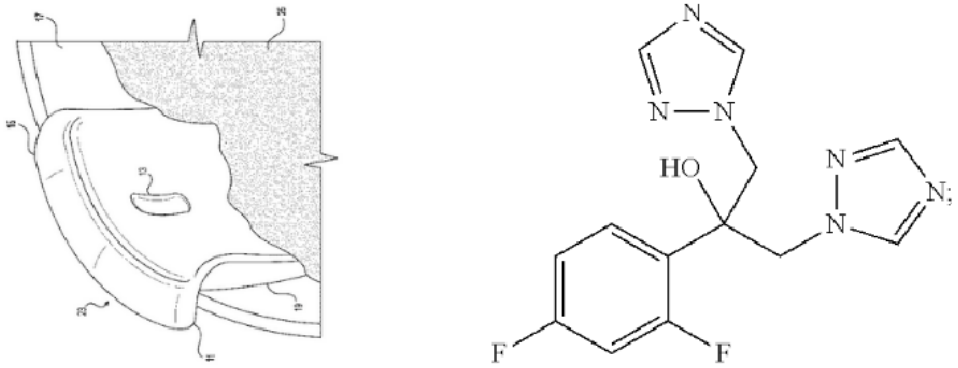

Figure 2: Example images that unrelated/related to chemistry

We have defined our classification task as recognizing whether an image does or does not contain chemical structures. After one knows this information, they can apply additional information extraction techniques

for that specific image, *e.g.*, translation to machine readable formats. To our knowledge, the classification task has not been tackled at the moment.

It is noteworthy that this work is carried out as part of a master thesis that covered broader machine learning topics and acted as a tool to explore the main thesis research. All code developed for this body of work is open and can be found on GitHub [1].

3 Method

Our approach is to train a CNN[2] on a baseline dataset and then measure how well the network generalizes to data from other patent office. This will give an idea on (1) how well we can tackle the classification task for a single patent offices and (2) future design decisions on whether we need to train a model for each patent offices or a single model that tries to "learn it all". Furthermore we run experiments, such as, adding additional classification classes to the task, so that we can explore additional solutions.

4 Datasets

To train and experiment, we collected a number of datasets which are described in this section. First we collected a baseline dataset that is used to train the neural network with, and secondly we received a number of validation sets from the host-organisation.

4.1 Baseline (USPTO) dataset

The baseline dataset is curated from the United States Patent and Trademark Office (USPTO[3]), using "Patent Grant Full Text Data with Embedded TIFF Images" from the first four weeks in the year 2021. Hereby all images with a "C" suffix in the image name are labeled as a chemical image and all remaining are labeled as non-chemical. This leaves us with a dataset of approx 140k images that either contain no chemicals or at least one chemical structure. Finally we take subsets of 10k datapoints of each class so that the dataset becomes balanced and to make the experimentation-cycle faster.

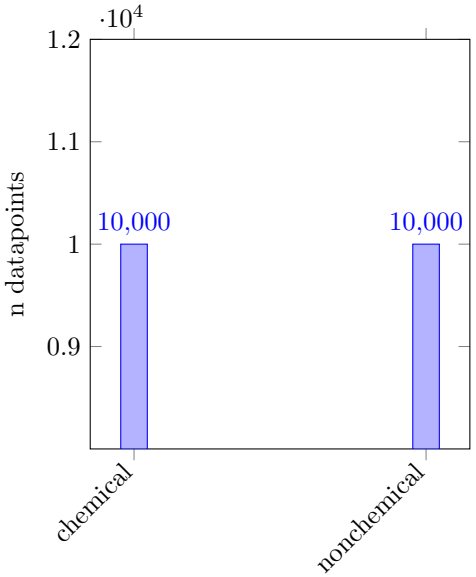

Figure 3: Dataset distribution of chemical and nonchemical images

[1] Chemical image recognition - https://github.com/fatahfattah/chemical-image-recognition
[2] Convolutional Neural Network (CNN)
[3] United States Patent and Trademark Office bulkdata `https://bulkdata.uspto.gov/`

See table 3 for the final distribution over the two classes.

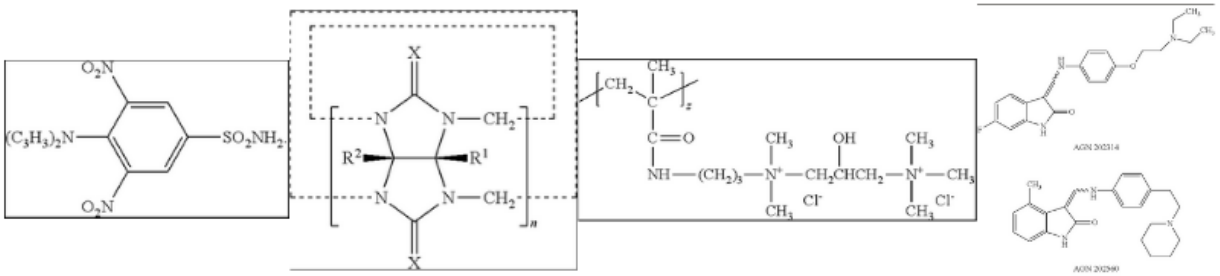

Figure 4: Example images for `chemical` class

See figure 4 for example images that are contained in the `chemical` class subset.

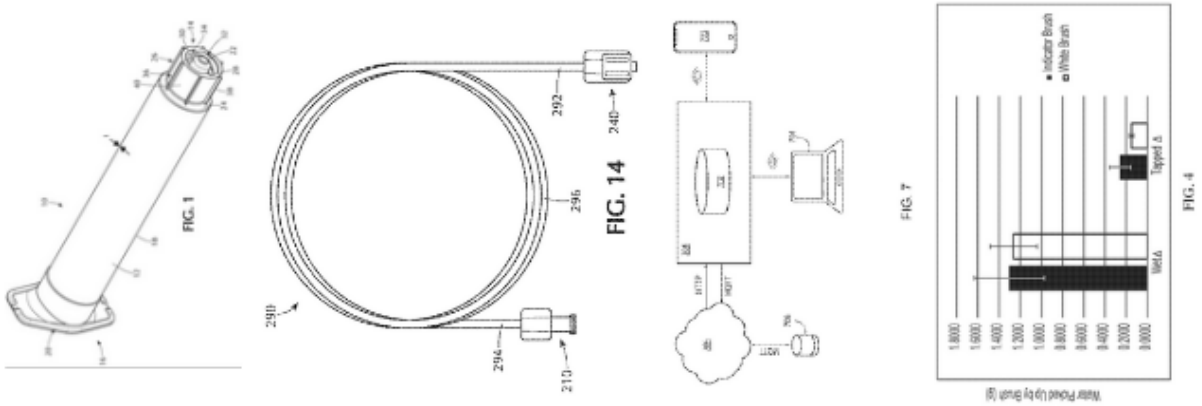

Figure 5: Example images for `nonchemical` class

See figure 5 for example images that are contained in the `nonchemical` class subset.

4.2 Additional validation datasets

In order to test generalization of the trained model and validate, using patents from other patent offices, we have curated a number of datasets that contains images from patents of the Chinese, Taiwanese, Korean, Japanese, European and WO[4] offices. The data for these offices are shared (and owned) by the host-organization (Elsevier) and is non-public data. The datasets ranged in sizes from 500 - 30k datapoints, containing non-labeled patent images.

5 Training configuration

The architecture of all models is based on the inception v3 network from Szegedy et al.[9], in which the last fully connected layer is connected to the output layer with n nodes, where n = number of output classes, and finally a softmax loss function. All models are initialized with random weights. Although the architecture was introduced five years ago, at the moment of writing, it still outperforms many of the top performing approaches in the field. Additionally, the decision to use the inception architecture is related to the wide support in ML-frameworks such as PyTorch [10] and also the fast speed in experimentation due to the relatively low number of parameters in the network [9].

[4]WIPO - World Intellectual Property Organisation

All training and experimentation is performed on a single Nvidia GTX 1070 with a batch size of 8 for 10 epochs. For training, we used Stochastic Gradient Descent (SGD) [11] and momentum with a decay of 0.9, a learning rate of 0.001 and Nesterov acceleration [12]. This configuration is decided through pointers in literature [11, 12] and through experimenting.

Finally we apply some image augmentation in an attempt to increase generalization of the network. In this we apply (1) random horizontal image flips (2) random vertical image flips and (3) random sharpness adjustments.

6 Performance metrics

Throughout the document we describe various results for which we use the following performance benchmarking metrics to compare the performance of the classifiers.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision tells us how often we are actually correct when positively classifying something.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall tells us how well we can find all the datapoints that belong to a certain class.

$$\text{F1-score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

F1-score is a mean of both precision and recall in a way that it emphasizes the lowest value.

$$\text{Macro F1-score} = \sum f1_n * \frac{1}{n}$$

Macro F-1 score is the average of all F1-scores.

7 Results

This section describes the results of our trained network. We first present the baseline performance on our USPTO dataset and then we run validation on patents from other patent offices.

The baseline performance is as depicted in table 1:

Class	Precision	Recall	F1-score	Support
chemical	1.00	0.94	0.97	2000
nonchemical	0.94	1.00	0.97	2000
Accuracy			0.97	4000
Macro avg	0.97	0.97	0.97	4000
Weighted avg	0.97	0.97	0.97	4000

Table 1: Baseline performance for chemical/nonchemical task on USPTO data

Next we provide all raw results for completeness of results on all patent offices.

Class	Precision	Recall	F1-score	Support
chemical	0.87	0.58	0.70	214
nonchemical	0.53	0.85	0.65	286
Macro avg	0.70	0.71	0.67	500

Table 2: Validation performance for chemical/nonchemical task on CN data

Class	Precision	Recall	F1-score	Support
chemical	0.97	0.53	0.69	196
nonchemical	0.45	0.96	0.61	304
Macro avg	0.71	0.75	0.65	500

Table 3: Validation performance for chemical/nonchemical task on WO data

Class	Precision	Recall	F1-score	Support
chemical	0.98	0.69	0.81	222
nonchemical	0.65	0.98	0.78	278
Macro avg	0.81	0.83	0.79	500

Table 4: Validation performance for chemical/nonchemical task on TW data

Class	Precision	Recall	F1-score	Support
chemical	0.40	0.67	0.50	50
nonchemical	0.98	0.94	0.96	454
Macro avg	0.69	0.80	0.73	504

Table 5: Validation performance for chemical/nonchemical task on KR data

Class	Precision	Recall	F1-score	Support
chemical	0.96	0.41	0.58	105
nonchemical	0.64	0.98	0.77	395
Macro avg	0.80	0.70	0.67	500

Table 6: Validation performance for chemical/nonchemical task on JP data

Class	Precision	Recall	F1-score	Support
chemical	0.98	0.92	0.95	235
nonchemical	0.92	0.98	0.95	265
Macro avg	0.95	0.95	0.95	500

Table 7: Validation performance for chemical/nonchemical task on EP data

Finally we can compare the results from all patent offices against each other.

Patent office	chemical F1-score	nonchemical F1-score	F1-score macro avg.
US	0.97	0.97	0.97
CN	0.70	0.71	0.67
WO	0.69	0.61	0.65
TW	0.81	0.78	0.79
KR	0.50	0.96	0.73
JP	0.58	0.77	0.67
EP	0.95	0.95	0.95

Table 8: F1-score performance for all patent offices

We are able to observe that the baseline performance of the USPTO validation, does not generalize well to the data from other patent offices. Even though the performance is not *as high*, overall network F1-scores on these datasets range from 65% up to 95%, indicating that the network still has some ability to classify the images from these offices.

7.1 Threats to validity

The bench-marked performances of table 8 are slightly unfair in that we do not consider weighted performance scores. This is an issue in that a high performance on one class with only few support points, can lead to a overall good performing network.

8 Experiment - distinguish images that contain many chemical structures

We ran an experiment to distinguish another image class, namely ones that contain more than one chemical structure. We call this class the `manychemical` class, and figure 6 shows example images from all classes side-by-side. It is possible that one image contains multiple chemical structure, due to incorrect image extraction or because they are simply just two related structures. Therefore it can be useful to recognize such instances and take actions accordingly.

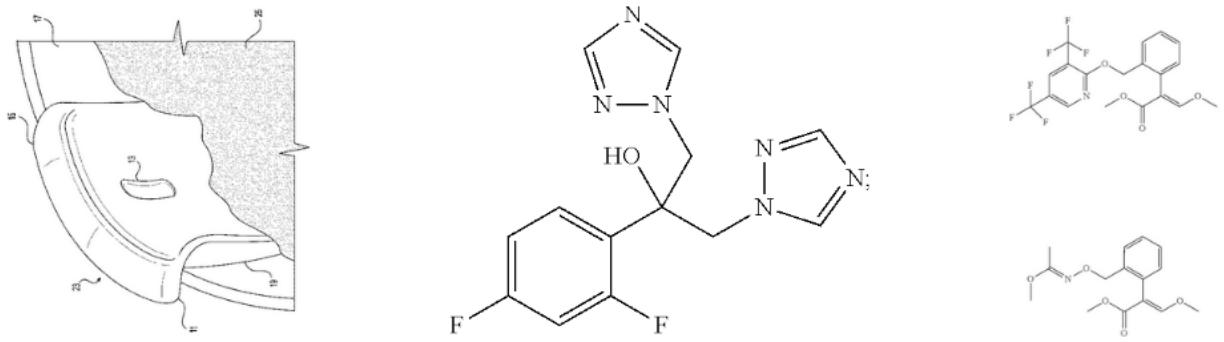

Figure 6: Example images containing none, one and many chemical structures

Dataset From our baseline model of chemical/nonchemical we have acquired a dataset of approx 140k images that either contain no chemicals or at least one chemical structure, from the USPTO. We run a tool called PraLine [13] over the chemical datapoints, to estimate the number of chemical structures in each of the images. We then split the chemical images into the onechemical and manychemical classes whenever they contain one or more than one chemical structures respectively, according to Praline. Finally, we take a

subset of 7k images for each class. The reason that the final class amounts are significantly lower, compared to the chemical/nonchemical dataset, is that there are a lower amount of manychemical datapoints. So the distribution has come to be, so that we get a balanced dataset.

Our validation performance after training a network on our dataset for 10 epochs, looks as follows:

Class	Precision	Recall	F1-score	Support
manychemical	0.90	0.98	0.94	1400
nonchemical	0.99	0.97	0.98	1400
onechemical	0.96	0.90	0.93	1400
Accuracy			0.95	4200
Macro avg	0.95	0.95	0.95	4200
Weighted avg	0.95	0.95	0.95	4200

The performance over all classes seems near perfect, on our unseen USPTO validation dataset. We investigate this performance further by looking at how well it generalizes on a completely unseen, manually curated, dataset from various patent offices. A total of around 200 datapoints were manually validated which showed the following performance[5]:

Class	Precision	Recall	F1-score
manychemical	0.02	0.50	0.38
nonchemical	0.74	0.74	0.74
onechemical	0.84	0.65	0.74

We observe a significant decrease in performance over the baseline, with drops in F1-score over all classes.

9 Future work

As this work was part of a more general machine learning thesis, there were some temporal constraints that limited the depth of the scope. We have however observed some promising results that are interesting to further pursue.

9.1 Fine-tuning on other patent office datasets

We observed that a network that was trained on a single patent office, did not generalize well to data from other patent offices. One approach that might be interesting is to apply transfer learning using (smaller) labeled datasets from these patent offices and test how well the network then performs. Transfer learning generally allows a network to converge on a new dataset, using less datapoints. Considering that the baseline network is already trained on a visually similar task, it is intuitively an appropriate model to use for this fine-tuning.

9.2 Re-training on all patent office datasets

Ideally one has labeled datasets for all patent offices, however, creation of these datasets is a labor and time intensive. If one does indeed have such datasets, they can (1) train separate neural networks on each of these datasets and (2) train one network using all of these datasets.

[5]Not all considered metrics are the same as the baseline, reason being that we do not have all the data due to the manual validation process

10 Conclusion

In this work we explored the possibilities of Convolutional Neural Networks to determine presence of chemical structures in chemistry patent document images. We observed that a simple neural network has the ability to perform well (97% accuracy) for a single patent office, but this performance does not generalize well to data from other patent offices. We additionally experimented with an attempt to determine the presence of single or multiple chemical structures in an image, which seemed to perform well (95% accuracy) on a single patent office, however manual validation on an unseen dataset from other patent offices did not reproduce such results, reaching only 65% accuracy. In conclusion, the opportunities of image classification in this context are promising according to the results on single patent office data and further exploration would be an interesting and logical next step.

References

[1] S. Akhondi, H. Rey, M. Schwörer, M. Maier, J. Toomey, H. Nau, G. Ilchmann, M. Sheehan, M. Irmer, C. Bobach, M. A. Doornenbal, M. Gregory, and J. Kors, "Automatic identification of relevant chemical compounds from patents," *Database: The Journal of Biological Databases and Curation*, vol. 2019, 2019.

[2] A. Abbas, L. Zhang, and S. Khan, "A literature review on the state-of-the-art in patent analysis," *World Patent Information*, vol. 37, Jun. 2014. DOI: 10.1016/j.wpi.2013.12.006.

[3] RELXgroup, *Reaxys*. [Online]. Available: https://www.reaxys.com/.

[4] K. Rajan, H. Brinkhaus, C. Steinbeck, and A. Zielesny, "A review of optical chemical structure recognition tools," *Journal of Cheminformatics*, vol. 12, Oct. 2020. DOI: 10.1186/s13321-020-00465-0.

[5] I. Filippov and M. Nicklaus, "Optical structure recognition software to recover chemical information: Osra, an open source solution," *Journal of chemical information and modeling*, vol. 49, pp. 740–3, Apr. 2009. DOI: 10.1021/ci800067r.

[6] M. Oldenhof, A. Arany, Y. Moreau, and J. Simm, "Chemgrapher: Optical graph recognition of chemical compounds by deep learning," *Journal of chemical information and modeling*, vol. 60, Sep. 2020. DOI: 10.1021/acs.jcim.0c00459.

[7] K. Rajan, A. Zielesny, and C. Steinbeck, "Decimer: Towards deep learning for chemical image recognition," *Journal of Cheminformatics*, vol. 12, no. 1, pp. 1–9, 2020.

[8] J. Staker, K. Marshall, R. Abel, and C. McQuaw, *Molecular structure extraction from documents using deep learning*, 2018. arXiv: 1802.04903 [cs.LG].

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[10] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[11] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, Springer, 2010, pp. 177–186.

[12] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[13] V. A. Simossis and J. Heringa, "Praline: A multiple sequence alignment toolbox that integrates homology-extended and secondary structure information," *Nucleic acids research*, vol. 33, no. suppl_2, W289–W294, 2005.