A Lightweight Solution to Recognition of Handwritten Chemical Structures

Chungkwong Chan *†

September 7, 2024

Abstract

Recognizing images of handwritten chemical structure is meaningful for educational and research purposes. A recognition system which make use of depthwise convolution, large kernel, instance normalization, counting loss, relaxed classification loss, model ensembling, customized representation of chemical structure, and formal grammar is proposed. Although the total number of parameters is only about $4 \times 1.3M$, the proposed solution has achieved a highly competitive accuracy in ICDAR 2024 Competition on Recognition of Chemical Structures.

Keywords Chemical structure recognition; handwriting recognition; optical character recognition

1 Introduction

Chemical structure is a kind of basic element in engineering and scientific documents, digitizing them would maximize the usability of these valuable information. Unfortunately, converting them to a machine-readable form can be cumbersome, graphical editors are inefficient, whereas markups like chemfig[1] and SMILES[2] are not easy to learn. An accurate recognition systems for chemical structures may enable a natural and efficient input method, many potential applications can also base upon that. For example, automatic marking of answers and retrieval of chemical information.

Recognition of chemical structures is a challenging task, which is even more difficult than recognition of mathematical expressions. Mathematical expressions are trees, whereas chemical structures are generic graphs. The diverse ways to draw the same molecule also pose difficulties. There are a number of competitions on handwritten mathematical expression recognition, including CROHME 2011, 2012, 2013, 2014, 2016, 2019 and 2023 [1], OffRaSHME 2020 [2], and MLHME 2023 [3]. However, up to our knowledge, ICDAR 2024 Competition on Recognition of Chemical Structures[3] (ICDAR 2024 CROCS) is the first competition devoted to handwritten chemical structure recognition.

In this technical report, a lightweight solution to the challenge is proposed. Standard encoder-decoder models are used to translate a bitmap image to a sequence of tokens in a language sightly different from SSML-norm [4], where type and angle (rounded to a multiple of 15) of a bond are represented with two tokens, whereas length is ignored. During preprocessing, all input images are first converted to grayscale, then rescaled and padded to the resolution of 1024x256. The model architecture is similar to the baseline DenseWAP [5], but we have modified the backbone. All convolution layers are replaced with pointwise and depthwise convolutions [6] to reduce the time complexity and the number of parameters. A larger kernel size (5x5 instead of 3x3) is used to enlarge the reception field. All batch normalization layers are replaced with instance normalization [7] layers to improve generalization ability and ensure that computation is the same across training and inference. A counting loss [8] is utilized during training in addition to a relaxed classification loss. Model ensembling and

*Sunia PTE LTD
†Email address: chan@chungkwong.cc
[1] https://ctan.org/pkg/chemfig
[2] http://opensmiles.org/opensmiles.html
[3] https://crocs-ifly-ustc.github.io/crocs/

LL(1) grammar parsing [9] are also employed during beam search to boost accuracy and avoid illegal output.

2 Related works

In the past, rule-based systems [10] are used for chemical structure recognition. They can recognize clear images of printed chemical structure reasonably well, but often fail badly on degraded images or handwritten chemical structures.

Recently, neural networks become popular. End-to-end approaches used string decoder [11] or graph decoder [12] to predict the representation of chemical structure directly. Detection based approaches use object detection to locate atoms and bonds, then graph neural network [13] or rules [14] are applied to reconstruct the structure. In theory, these methods can be applied to handwritten chemical structures, but they require a large amount of training data and labeling handwritten images is expensive. Therefore, a publicly available dataset is needed.

3 Methodology

Standard encoder-decoder models are used to translate a bitmap image to a sequence of tokens in a language sightly different from SSML-norm.

3.1 Data processing

3.1.1 Modified SSML-norm

Type and angle (rounded to a multiple of 15 as in [4]) of a bond are represented with two tokens, whereas length is ignored. Decoupling type and angle of a bond can reduce class imbalance. In fact, some combinations of type and angle are absent or extremely rare in the training data, they can still be predicted thanks to this language.

3.1.2 Image preprocessing

During preprocessing, all input images are first converted to grayscale, then rescaled and padded to the resolution of 1024x256, where the median color is used for padding so that no mask is needed.

3.1.3 Data augmentation

Some image transformations are applied on-the-fly during training to improve the robustness of the models. Here are some of them:

- Affine transforms
- Blur and sharpen
- Adjusting brightness and contrast
- Addition of noise

3.2 Model architecture

The model architecture is similar to the baseline DenseWAP [5], but we implement it ourselves with TensorFlow[4] and Keras[5] for training, and ONNX Runtime[6] for inference. Compared with the baseline, the backbone is another variant of DenseNet [15]. The major modifications are listed in the following.

3.2.1 Depthwise convolution

All convolution layers except the first one are replaced with pointwise and depthwise convolutions as in MobileNetv2 [6] to reduce the time complexity and the number of parameters. The use of depthwise convolutions also make them insensitive to the kernel size.

3.2.2 Larger kernel size

A larger kernel size (5x5 instead of 3x3) is used to enlarge the reception field, which is a common practice adopted by many modern convolutional networks [16].

3.2.3 Instance Normalization

All batch normalization [17] layers are replaced with instance normalization [7] layers to improve generalization ability and ensure that computation is the same across training and inference. Unlike batch normalization, instance normalization cannot be folded with the previous convolutional or fully-connected layers before inference, so it results in small overhead during inference.

[4] https://www.tensorflow.org/
[5] https://keras.io/
[6] https://onnxruntime.ai/

3.3 Training

Adadelta [18] with gradient clipping is applied to optimize the loss. Early stopping and epsilon decay are also used to prevent overfitting and numerical instability.

3.3.1 Counting loss

In addition to the classification loss, a counting loss [8] is added as a kind of weak supervision.

3.3.2 Relaxed loss

At the final stage of training, we modified the cross entropy classification loss so that it pay less attention to minor recognition errors of bond angle, so that the model can focus on something more important.

3.4 Inference

The beam search algorithm is used for decoding. The beam width is set to 4. Model ensembling and formal grammar parsing are also employed during beam search to boost accuracy and avoid illegal output.

3.4.1 Model ensembling

Model ensembling is implemented by simply averaging the predictions of different models during beam search. Those models are trained on the same dataset independently with different initialization.

3.4.2 Grammar constraint

We have designed a LL(1) context free grammar [9] and enforce the conformance of the outputs. The grammar prevented common errors like mismatched braces from being generated. During the decoding process, a candidate is discarded if it can not be a prefix of a valid output. It is important for the consumers of recognition results because they may not be robust to malformed input.

4 Experiment

In this section, some design decisions are justified. We split the training data into train set, validation set, and test set. The validation set containing 1000 samples is used for early stopping and the test set containing 1000 samples is used for evaluation.

Table 1: The effect of resolution

Resolution	EM	Structure EM
1024 × 256	**66.30%**	**76.10%**
512 × 512	61.80%	73.20%

Table 2: The effect of kernel size

Normalization	EM	Structure EM
3 × 3	65.80%	76.00%
5 × 5	**66.30%**	**76.10%**

4.1 Resolution

Table 1 shows that the input resolution of model is quite important. Although $1024 \times 256 = 512 \times 512$ and therefore the inference time is similar, the model consuming images of 1024×256 performed much better. We believe that the reason is that for most of the images in the dataset, the width is higher than the height, so more details are lost when they are resized to 512×512.

4.2 Kernel size

Table 2 shows that using a larger convolutional kernel resulted in a higher accuracy.

4.3 Normalization

Table 3 shows that both instance normalization (IN) and layer normalization (LN) [19] outperformed batch normalization (BN).

Table 3: The effect of normalization

Normalization	EM	Structure EM
BN	61.50%	72.30%
LN	64.20%	74.10%
IN	**66.30%**	**76.10%**

Table 4: The effect of classification loss

Loss function	EM	Structure EM
Standard	66.30%	76.10%
Relaxed	**68.40%**	**79.70%**

Table 5: The effect of model ensembling

Model count	EM	Structure EM
1	66.30%	76.10%
2	70.50%	80.70%
3	71.00%	81.70%
4	**72.30%**	**82.50%**

4.4 Classification loss

Table 4 shows that trained model benefited from the relaxed classification loss function.

4.5 Model ensembling

Table 5 shows that the use of model ensembling is a reliable way to boost the accuracy, if inference time is not a concern.

4.6 Comparison with state of the art

Table 6 shows that the proposed solution is highly competitive among the participants of ICDAR 2024 CROCS. Although it is not as good as the one by CNKI AI at the end of the competition, the gap is small (less than one percent point). Meanwhile, it outperformed the third place finisher by a large margin (over 8 percent points for EM). It should be noted that the total number of parameters is only about 4×1.3 millions, which is quite small compared with other recent models for chemical structure recognition.

5 Conclusion

With a set of standard tricks, competitive accuracy for handwritten chemical structure recognition can be achieved, even if the number of total parameters is much lower than most of the modern neural network based systems reported in the literature.

Table 6: Comparison with other teams

Team	EM	Structure EM
CNKI AI	**70.66%**	**80.12%**
Sunia	70.05%	79.44%
imucs	62.03%	72.72%
shenweiping	56.76%	67.56%
TSNUK	50.11%	67.16%
cuong.nt2	19.60%	29.29%

References

[1] Y. Xie, H. Mouchère, F. S. Liwicki, S. Rakesh, R. Saini, M. Nakagawa, C. T. Nguyen, T. Truong, ICDAR 2023 CROHME: competition on recognition of handwritten mathematical expressions, in: G. A. Fink, R. Jain, K. Kise, R. Zanibbi (Eds.), Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part II, Vol. 14188 of Lecture Notes in Computer Science, Springer, 2023, pp. 553–565. https://doi.org/10.1007/978-3-031-41679-8_33. https://doi.org/10.1007/978-3-031-41679-8_33

[2] D. Wang, F. Yin, J. Wu, Y. Yan, Z. Huang, G. Chen, Y. Wang, C. Liu, ICFHR 2020 competition on offline recognition and spotting of handwritten mathematical expressions - offrashme, in: 17th International Conference on Frontiers in Handwriting Recognition, ICFHR 2020, Dortmund, Germany, September 8-10, 2020, IEEE, 2020, pp. 211–215. https://doi.org/10.1109/ICFHR2020.2020.00047. https://doi.org/10.1109/ICFHR2020.2020.00047

[3] C. Gao, Y. Liu, S. Yao, J. Bai, X. Bai, L. Jin, C. Liu, ICDAR 2023 competition on recognition of multi-line handwritten mathematical expressions, in: G. A. Fink, R. Jain, K. Kise, R. Zanibbi (Eds.), Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part II, Vol. 14188 of Lecture

Notes in Computer Science, Springer, 2023, pp. 566–576. https://doi.org/10.1007/978-3-031-41679-8_34. https://doi.org/10.1007/978-3-031-41679-8_34

[4] J. Hu, H. Wu, M. Chen, C. Liu, J. Wu, S. Yin, B. Yin, B. Yin, C. Liu, J. Du, L. Dai, Handwritten chemical structure image to structure-specific markup using random conditional guided decoder, in: A. El-Saddik, T. Mei, R. Cucchiara, M. Bertini, D. P. T. Vallejo, P. K. Atrey, M. S. Hossain (Eds.), Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023, ACM, 2023, pp. 8114–8124. https://doi.org/10.1145/3581783.3612573. https://doi.org/10.1145/3581783.3612573

[5] J. Zhang, J. Du, L. Dai, Multi-scale attention with dense encoder for handwritten mathematical expression recognition, CoRR abs/1801.03530. arXiv:1801.03530. http://arxiv.org/abs/1801.03530

[6] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, L. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474. http://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html

[7] D. Ulyanov, A. Vedaldi, V. S. Lempitsky, Instance normalization: The missing ingredient for fast stylization, CoRR abs/1607.08022. arXiv:1607.08022. http://arxiv.org/abs/1607.08022

[8] B. Li, Y. Yuan, D. Liang, X. Liu, Z. Ji, J. Bai, W. Liu, X. Bai, When counting meets HMER: counting-aware network for handwritten mathematical expression recognition, CoRR abs/2207.11463. arXiv:2207.11463, https://doi.org/10.48550/ARXIV.2207.11463. https://doi.org/10.48550/arXiv.2207.11463

[9] A. V. Aho, M. S. Lam, R. Sethi, J. D. Ullman, Compilers: Principles, Techniques, and Tools (2nd Edition), Addison-Wesley, Boston, 2006.

[10] I. V. Filippov, M. C. Nicklaus, Optical structure recognition software to recover chemical information: Osra, an open source solution, Journal of chemical information and modeling 49 3 (2009) 740–3. https://api.semanticscholar.org/CorpusID:23797072

[11] D.-A. Clevert, T. Le, R. Winter, F. Montanari, Img2mol - accurate smiles recognition from molecular graphical depictions (2021). https://doi.org/10.26434/chemrxiv.14320907.v1. https://doi.org/10.26434/chemrxiv.14320907.v1

[12] Y. Qian, J. Guo, Z. Tu, Z. Li, C. W. Coley, R. Barzilay, Molscribe: Robust molecular structure recognition with image-to-graph generation, J. Chem. Inf. Model. 63 (7) (2023) 1925–1934. https://doi.org/10.1021/ACS.JCIM.2C01480. https://doi.org/10.1021/acs.jcim.2c01480

[13] L. Morin, M. Danelljan, M. I. Agea, A. S. Nassar, V. Weber, I. Meijer, P. W. J. Staar, F. Yu, Molgrapher: Graph-based visual recognition of chemical structures, CoRR abs/2308.12234. arXiv:2308.12234, https://doi.org/10.48550/ARXIV.2308.12234. https://doi.org/10.48550/arXiv.2308.12234

[14] Y. Xu, J. Xiao, C. Chou, J. Zhang, J. Zhu, Q. Hu, H. Li, N. Han, B. Liu, S. Zhang, J. Han, Z. Zhang, S. Zhang, W. Zhang, L. Lai, J. Pei, Molminer: You only look once for chemical structure recognition, CoRR abs/2205.11016. arXiv:2205.11016, https://doi.org/10.48550/ARXIV.2205.11016. https://doi.org/10.48550/arXiv.2205.11016

[15] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July

21-26, 2017, IEEE Computer Society, 2017, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243. https://doi.org/10.1109/CVPR.2017.243

[16] X. Ding, X. Zhang, Y. Zhou, J. Han, G. Ding, J. Sun, Scaling up your kernels to 31x31: Revisiting large kernel design in cnns, CoRR abs/2203.06717. arXiv:2203.06717, https://doi.org/10.48550/ARXIV.2203.06717. https://doi.org/10.48550/arXiv.2203.06717

[17] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, CoRR abs/1502.03167. arXiv:1502.03167. http://arxiv.org/abs/1502.03167

[18] M. D. Zeiler, ADADELTA: an adaptive learning rate method, CoRR abs/1212.5701. arXiv:1212.5701. http://arxiv.org/abs/1212.5701

[19] L. J. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, CoRR abs/1607.06450. arXiv:1607.06450. http://arxiv.org/abs/1607.06450