Improved Architecture of Artificial Neural Network for Secondary Structure Analysis

Semyon Grigorev[1], Polina Lunina[1]

[1] *Saint Petersburg State University, JetBrains Reserach, St. Petersburg, Russia*

E-mail: *semyon.grigorev@jetbrains.com*

Motivation

An approach for biological sequences processing by combination of formal grammars and neural networks is proposed in the work [1]. While classical way is to model secondary structure of the full sequence by using grammar, the proposed approach utilizes it only for primitive secondary structure features description. These features can be extracted by parsing algorithm and processed by using artificial neural network. It is shown that this approach is applicable for real-world data processing and some questions are formulated for future research. In this work we provide answers to some of them.

Questions

Is it possible to use convolutional neural networks for parsing result processing? The result of parsing algorithm is a set of upper triangular boolean matrices. The original idea is to vectorize these matrices row by row and use DNNs for these vectors processing. Matrices can be also treated as bitmaps, where the false bits of matrix correspond to white pixels and the true bits to black ones. To handle these images we use network with a small number of convolutional layers, linearization and then the same structure as for vectorized data.

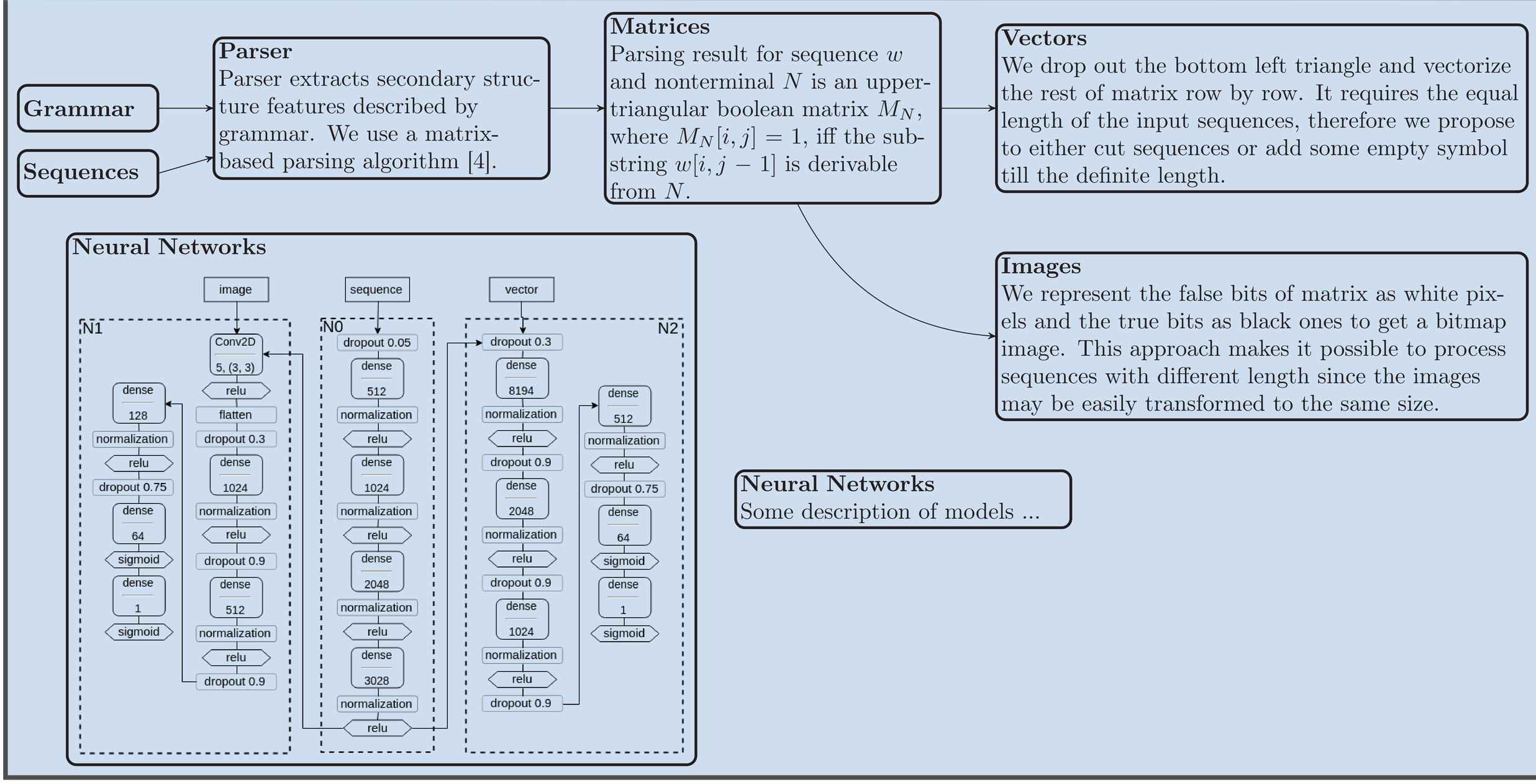
Is it possible to move parsing to network training step? Parsing is the most time-consuming operation of the proposed solution. We solve this problem by using two-staged learning. At the first step, we prepare a neural network (vector- or image-based) for our task which takes parsed data as an input. After that we extend trained network with a number of input layers that should convert original nucleotide sequence into parsing result. This way we create a network which can handle sequences, not parsing result. So, parsing is required only for training the first network.

Results

We use the proposed improvements to create neural networks for tRNA sequences analysis problems: classification of tRNA into 2 classes: eukaryotes and prokaryotes (EP) and 4 classes: archaea, bacteria, plants and fungi (ABFP). We use sequences from databases [2, 3]. Results for both image- and vector-based classifiers are presented in the table, where base model means network which handles parsing result and extended model handles sequences and is based on the corresponding base model.

Classifier	EP		ABFP	
Approach	Vectors	Images	Vectors	Images
Base model accuracy	94.1%	96.2%	86.7%	93.3%
Extended model accuracy	97.5%	97.8%	96.2%	95.7%
Total samples (train:valid:test)	20000:5000:10000		8000:1000:3000	

Solution Overview



Future Research

- Chimeric sequences filtration
- Secondary structure prediction
- Proteins functions prediction

Acknowledgments

The research was supported by the Russian Science Foundation grant 18-11-00100 and a grant from JetBrains Research.

Information

All materials are available on GitHub:
https://github.com/LuninaPolina/SecondaryStructureAnalyzer

References

[1] Semyon Grigorev. and Polina Lunina. The composition of dense neural networks and formal grammars for secondary structure analysis. In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3: BIOINFORMATICS*,, pages 234–241. INSTICC, SciTePress, 2019.

[2] Genomic tRNA Database. http://gtrnadb.ucsc.edu/. Last accessed 05.06.2019.

[3] tRNADB-CE. http://trna.ie.niigata-u.ac.jp/cgi-bin/trnadb/index.cgi. Last accessed 05.06.2019.

[4] Rustam Azimov and Semyon Grigorev. Context-free path querying by matrix multiplication. In *Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, page 5. ACM, 2018.