Improved Architecture of Artificial Neural Network for Secondary Structure Analysis

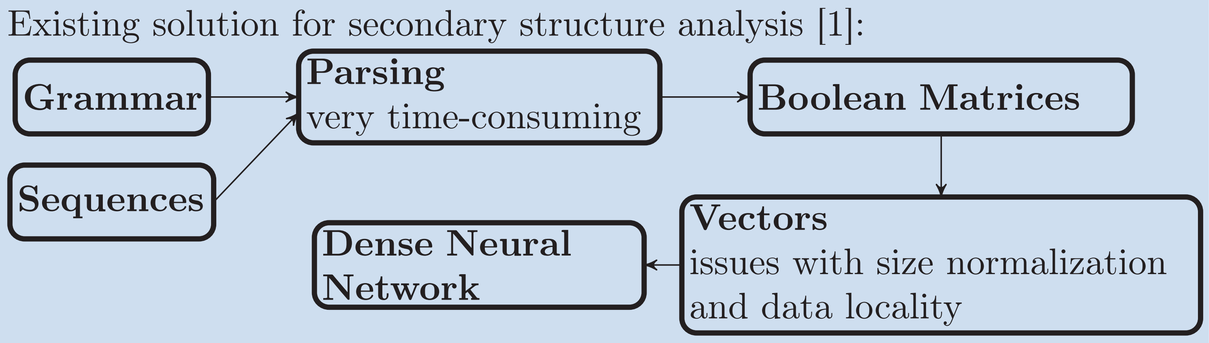
Polina Lunina[1], Semyon Grigorev[1]

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

E-mail: *lunina_polina@mail.ru, semyon.grigorev@jetbrains.com*

Motivation

Existing solution for secondary structure analysis [1]:



Questions:

- **Is it possible to move parsing to network training step?**
- **Is it possible to use convolutional neural networks for parsing result processing?**

Parsing elimination

We solve this problem by using two-staged learning.

1. Train a network which takes parsed data as an input (**base model**).
2. Extend trained network with a number of input layers that convert the nucleotide sequence into parsing result (**extended model**).

Parsing is required only for the network training.

Results: tRNA classification

- 2 classes: eukaryotes and prokaryotes (EP).
- 4 classes: archaea, bacteria, fungi and plants (ABFP).

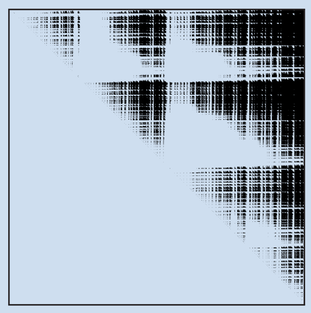
Classifier	EP		ABFP	
Approach	Vectors	Images	Vectors	Images
Base model accuracy	94.1%	96.2%	86.7%	93.3%
Extended model accuracy	97.5%	97.8%	96.2%	95.7%
Total samples (train:valid:test)	20000:5000:10000		8000:1000:3000	

Sequences from open databases [2, 3].

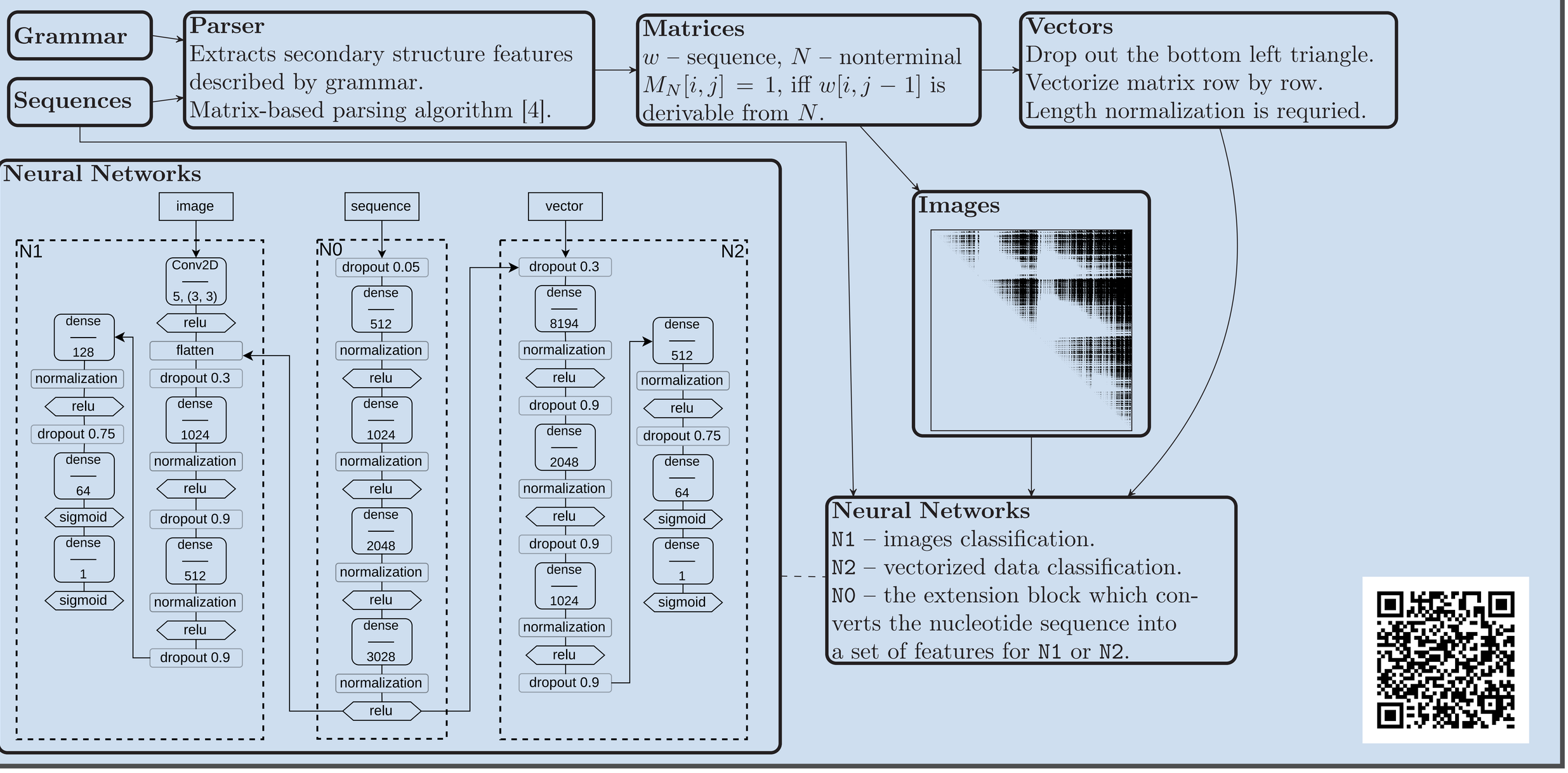
Data format

Matrices can be treated as bitmaps.

- We can use convolutional networks.
- Images can be easely resized.
- Data locality is saved.



Solution Overview



Future Research

- 16s rRNA processing and chimeric sequences filtration.
- Proteomic sequences processing, proteins functions prediction.
- Generative networks for sequences secondary structure prediction.

Acknowledgments

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

Information

Trained mofels and other materials are pablished at 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. Web page. URL: `http://gtrnadb.ucsc.edu/`. Last accessed 05.06.2019.

[3] tRNADB-CE. Web page. URL: `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.