DeepNano-blitz: A Fast Base Caller for MinION Nanopore Sequencers

Vladimír Boža*, **Peter Perešíni***, **Broňa Brejová**, **Tomáš Vinař**

*First two authors contributed equally to the study.

Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava
http://compbio.fmph.uniba.sk/

Introduction

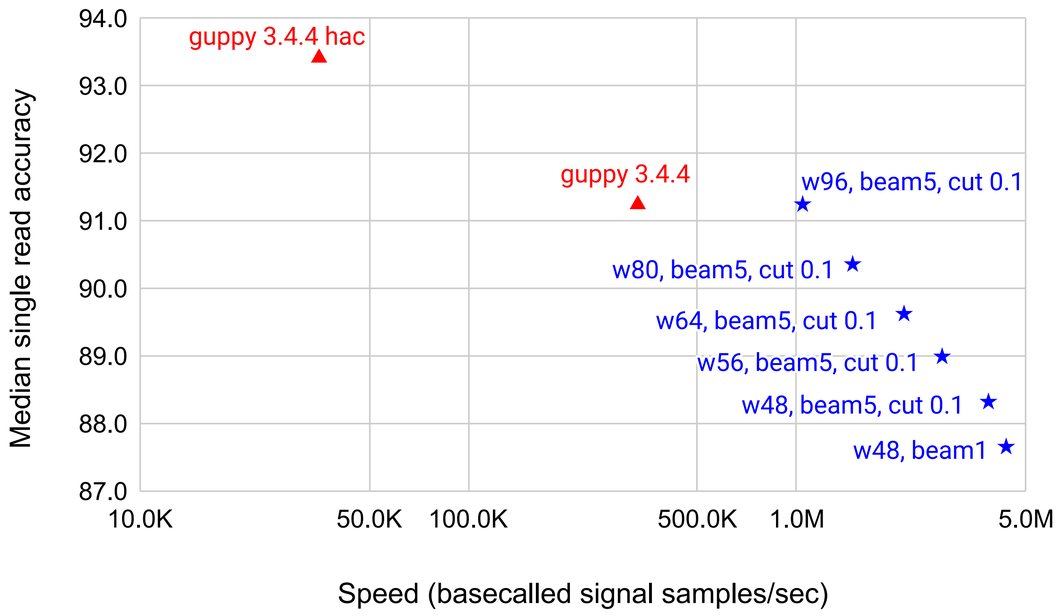
Base calling nanopore sequence data is a computationally intensive process and as such requires expensive hardware (e.g. GPUs or computational clusters), otherwise it takes too much time to be practical. This problem is exacerbated during data acquisition when access to such hardware is often limited (e.g., MinION sequencing in the field).

Our goal is to develop a very fast base caller capable of keeping up with the MinION device even on a consumer laptop device.

Benchmarks

To assess the capabilities of Deepnano Blitz, we benchmark it on a common laptop CPU (i7-7700HQ) and compare to Guppy 3.4.4 on a benchmark set of native R9.4 *K. pneumoniae* reads [Wick et al., 2019].

Our results suggest that on a CPU, Deepnano Blitz can rival accuracy of standard Guppy while maintaining 3.7-times its performance. With its fastest setting, Deepnano Blitz can keep up with two MinION devices producing real-time data (single MinION device can produce about 2M samples/sec).

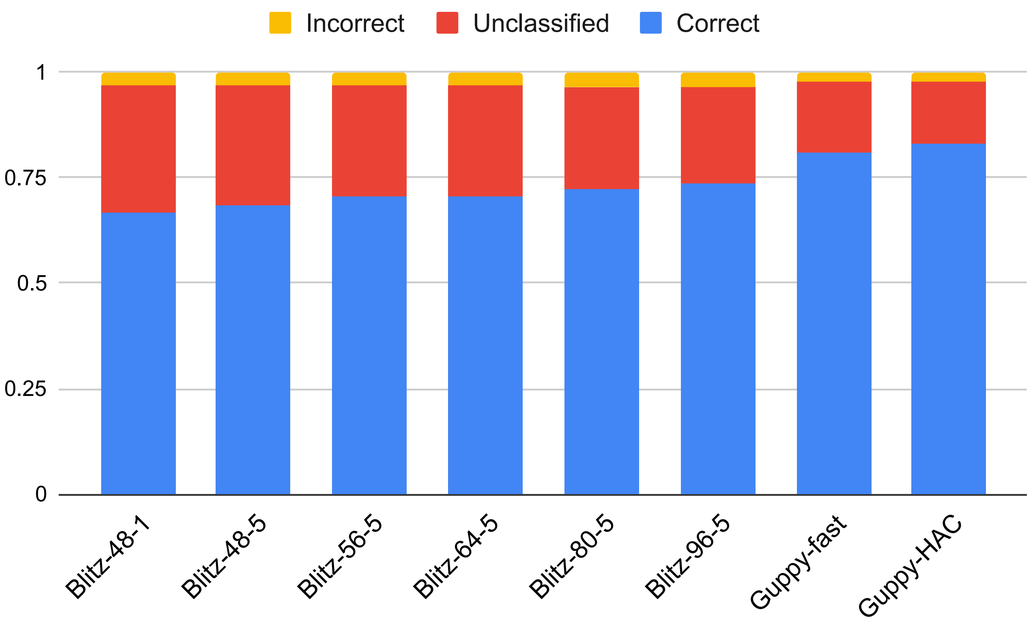


Use cases

We envision usage of Deepnano Blitz in circumstances where trading speed for slightly lower base calling accuracy is preferred:

- real-time monitoring (i.e. checking sample composition, barcode balance, quick identification of species) and give experimenters an instant feedback on the state of the data acquisition.
- pre-filtering data for more detailed analysis (e.g. filtering out human DNA from human–pathogen runs)

To this end, we test barcode composition precision and recall of various settings of Deepnano Blitz by barcoding 4000 reads from the data set of 12 barcoded bacterial samples PRJEB28450 [Wick et al., 2018].



Building a fast base caller

Deepnano Blitz is specifically designed to run fast on conventional laptop CPUs. To achieve this goal, we use multiple techniques

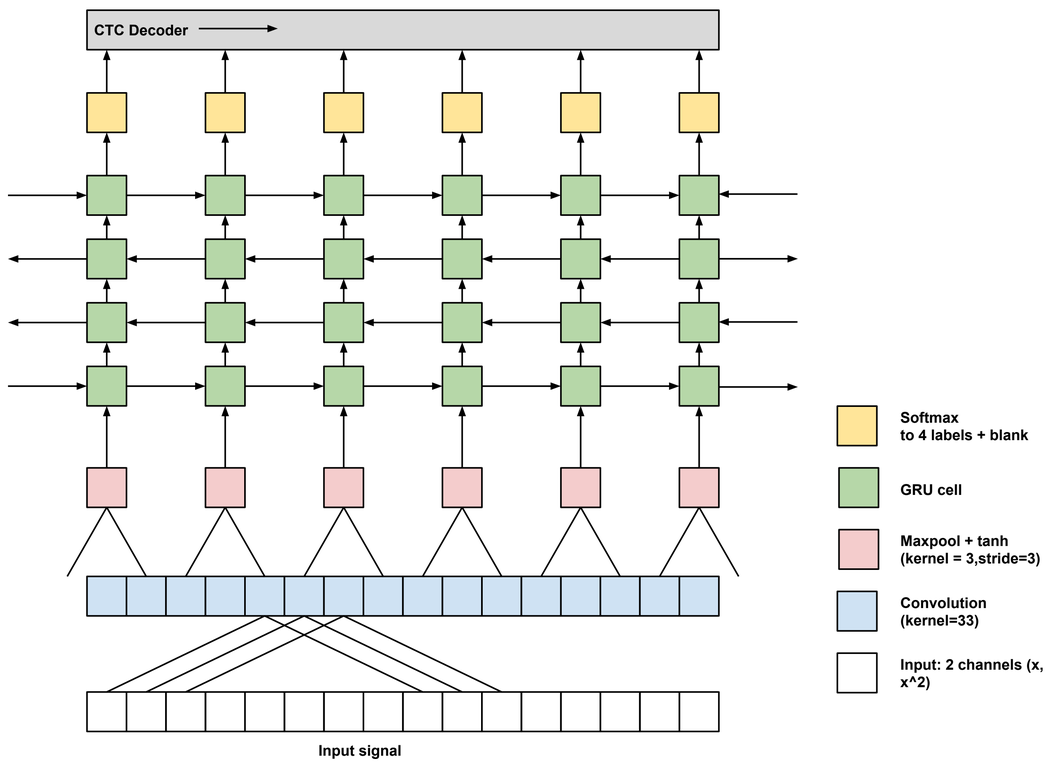
- neural network highly optimized for computational complexity
- using Rust language and hand-optimization of the critical path, including details such as exchanging slow math operations (e.g. $\tanh$) with their fast approximations
- use of Intel MKL library to perform all matrix multiplications and batch matrices whenever possible

Neural network architecture

Deepnano Blitz is based on a greatly optimized and simplified neural network. The basic architecture consists of

- preprocessing stage – convolution over 2 channels
- four layers of GRU RNN with alternating directions – we find that alternating RNN outperforms bi-directional RNNs with the same computational budget.
- postprocessing stage – softmax & CTC decoding with beam search and cutoff for unpromising beams

The size of hidden input for GRU cells greatly affects both performance and accuracy and is thus customizable. We provide pre-trained weights for sizes 48,56,64,80 and 96.



Availability and implementation

- The software has been implemented and benchmarked for Intel processors on Linux.
- Download at https://github.com/fmfi-compbio/deepnano-blitz (MIT licence).
  We provide both a standalone command as well as a Python library for easy integration.
- Learn more in our preprint: https://www.biorxiv.org/content/10.1101/2020.02.11.944223v1

References

[Wick et al., 2018] Wick, R. R., Judd, L. M., and Holt, K. E. (2018). Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLoS Computational Biology*, 14(11):e1006583.

[Wick et al., 2019] Wick, R. R., Judd, L. M., and Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, 20(1):129.

Acknowledgements:
This research was supported in part by grants from the Slovak Research and Development Agency (APVV-18-0239) and VEGA (1/0458/18 to TV and 1/0463/20 to BB).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 872539