UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

SEMINAR

MinCall — MinION od raw signala do genoma

Neven Miculinić
Mentor: *izv. prof. dr. sc. Mile Sikić*

Zagreb, May 2017.

CONTENTS

1. Introduction **1**

2. Background **2**

2.1. Sequencing overview 2

2.2. Basecalling 2

3. Method **4**

3.1. Data preprocessing 4

3.2. Residual arhitecture 5

3.3. CTC 5

4. Results **6**

5. Conclusion and further work **7**

6. Acknowledgments **8**

7. Bibliography **9**

MinCall — MinION od raw signala do genoma

Sažetak

U ovoj studiji kreirali smo MinCall, cjelokupni model očitanja nukleotida za MinION, portabilni DNK sekvencer firme Oxford Nanopore Technology. Model je baziran na dubokom učenju i koristi konvolucijske neuronske mreže u svojoj implementaciji. Najperformantniji 270 slojeva dubok model postiže najbolje dosad performanse od 90.5% preciznosti na E.Coli skupu za učenje koristeći R9 kemiju i 1D čitanja.

Ključne riječi: Basecaller, MinION, R9, CNN, CTC, sekvenciniranje sljedece generacije

MinCall — MinION end2end deep learning basecaller

Abstract

In this study, we created MinCall, an end2end basecaller model for the MinION, an Oxford Nanopore Technology's portable DNA sequencing device. The model is based on deep learning and uses convolutional neural networks (CNN) in its implementation. The best performing 270 layers deep model achieves state-of-the-art 90.5% match rate on E.Coli dataset using R9 pore chemistry and 1D reads.

Keywords: Basecaller, MinION, R9, CNN, CTC, Next generation sequecing

1. Introduction

In this paper, we introduce the base caller for the MinION nanopore sequencing platform [1]. The MinION device by Oxford Nanopore, weighing only 90 grams, is currently the smallest high-throughput DNA sequencer. Thanks to its low capital costs, small size and the possibility of analyzing the data in real time as they are produced, MinION is very promising for clinical applications, such as monitoring infectious disease outbreaks [2][3], characterizing structural variants in cancer[4] and even full human genome assembly [5].

Although MinION is able to produce long reads, they have a high sequencing error rate. This has been somewhat alleviated with new R9 pore model, replacing older R7 ones. In this paper, we show that this error rate can be reduced by replacing the default base caller provided by the manufacturer with a properly trained neural network model.

In the MinION device, single-stranded DNA fragments move through nanopores, which causes drops in the electric current. The electric current is measured at each pore several thousand times per second. The electric current depends mostly on the context of several DNA bases passing through the pore at the time of measurement. As the DNA moves through the pore, the context shifts, and the electric current changes.

A MinION device typically yields reads several thousand bases long; reads as long as 100,000 bp have been reported. 1D read has usually error rate about 10%. (Check details!!, What is the error rate, how it's defined???) After parameter training, base calling can be performed by running the Viterbi algorithm, which will result in the sequence of states with the highest likelihood.

The exact error rate metric is unreliable since multiple pipline tools could be the issue. First the sample is prepared, hopefully uncontaminated and matching reference genome as close as possile, then sequenced using the MinION device obtaining raw data. Next our model (or other groups ones) come along, basecall the sequence. To evalute error rate metric basecalled read is aligned to reference using aligners with their own erros/biases, mostly commonly used BWA-MEM and Graphmap.

2. Background

2.1. Sequencing overview

Informally, the MinION sequencer works as follows. First, DNA is sheared into fragments of 8–20 kbp and adapters are ligated to either end of the fragments. The resulting DNA fragments pass through a protein embedded in a membrane via a nanometre-sized channel (this protein is the ‘nanopore’). A single strand of DNA passes through the pore; the optional use of a hairpin adapter at one end of the fragment allows the two strands of DNA to serially pass through the nanopore, allowing two measurements of the fragment. With hairpin, it’s called 2D read, which isn’t the topic of this paper. In here we focus only on 1D reads.

In ONT terminology, the first strand going through the nanopore is the template, and the second is the complement. As a DNA strand passes through the pore it partially blocks the flow of electric current through the pore. The flow of current is sampled over time which is the observable output of the system. The central idea is that the single-stranded DNA product present in the nanopore affects the current in a way that is strong enough to enable decoding the electric signal data into a DNA sequence. This process, called basecalling, takes as input a list of current measurements and produces as output a list of DNA bases most likely to have generated those currents.

The nanopore is 6 nucleotides wide, and eariler HMM-based models model it as such.

2.2. Basecalling

The core of the decoding process is the basecalling step. Official basecaller is Metrichor, previously performed in the cloud before being open-sourced under name Nanonet.

Earlier models were (Hidden Markov model) HMM-based where hidden state modeled DNA sequence of length 6 (6-mer) in the nanopore. Pore models were used in computing emission probabilities. [6, 7, 8, 9] and the recent open source HMM-based basecaller Nanocall [10].

Recent models opted using (Recurrent neural network) RNN, notably DeepNano [11] and recently open sourced official Nanonets from ONT.

3. Method

Instead of opting for the traditional path using HMM or RNN we tried using CNN (Convolutional neural networks) [12], that is their residual version [13]. We opted out for gated residual network variant [14]. For loss we used CTC (Connectionist temporal classification) [15] between basecallled and the target sequence. The implementation used is open source warp-ctc [16]. Main computation framework is tensorflow [17]

3.1. Data preprocessing

Dataset was obtained from (Ask somebody, Marko doesn't know!). In this research, models were trained on the ecoli K-12 strand. (Should I cite someone for strand). Data was split into train and test subset, such that aligned reads map to different reference genome parts. For initial model data bootstrapping metrichorn was used, version (Ask Someone?).

The fast5 input training files were further split into smaller training blocks, consisting of fixed block size on raw signal. For each block target sequence, basecalled data is used in following way.

We're using basecalled knowledge which tells us on each raw read part which 6-mer were currently in the nanopore. Using this data, we get the intermediate target sequence for each block. To correct for model errors, we use aligned information, that is reference genome and alignment data(cigar string) to correct that information, yielding finalized target sequence. For further performance we skip first and last training block, since most error aggregate on edges.

Adjacent matching bases are separated with surrogate one, for example AAA -> AA'A for reasons described in the following section.

3.2. Residual arhitecture

We use gated residual network, that is layer function is like this: $f(x) = kg(x) + x$ Where x is input layer and k constant learnt during training. Specifically, g used is multiple, 1–3 times in our experiments, Relu-BatchNorm-CNN layers.

3.3. CTC

This section gives a brief overview over CTC, for further detail and in depth explanation we recommend original paper [15] or searching for contemporary blog posts.

In CTC we have target sequecne $\mathbf{l}$ consisting from symbols of alphabet Σ. Our models outputs $\mathbf{x}$, and CTC loss maximizes $p(\mathbf{l}|\mathbf{x})$.

$\mathbf{x}$ is consistend of n independent discrete random variables, X_i over domain $\Sigma \cup$ Blank. Path p through $\mathbf{x}$ is assigning for each X_i single value. It's log probabity is $\log P(p) = \sum_i \log P(X_i = x_i)$ where x_is define the path.

We further have merging operator on path, merge$(x_1, x_2, \ldots, x_n)$ which merges together adjacent equivalent elements. For example merge$(AAAGC) = AGC$.

Then $p(\mathbf{l}|\mathbf{x}) = \sum_{p \in \text{paths}(\mathbf{x}, \text{merged}(p) = \mathbf{l}} P(p)$ or less formally, $P(\mathbf{l}|\mathbf{x})$ is sum of all path probabilites which when merged give $\mathbf{l}$

In out concrete x application $\Sigma = \{$A G T C A' G' T' A'$\}$

For decoding, we use beam search decoder, with beam width 100.

4. Results

TODO

See https://github.com/nmiculinic/minion-basecaller/blob/master/notebook/Lambda_nanonet_vs_m270.ipynbforpreliminaryresults.

5. Conclusion and further work

This model used advance state-of-the-art gated residual convolutional neural network, with top models having 270 layers and over 3M parameters, yet improvements over Metrichorn baseline are marginal. As the conclusion, it might be that we've reached Bayesian error rate for R9 chemistry. Furthermore, R9.5 and $1D\hat{2}$ reads are under development which shall yield this paper's result obsolete quite soon, yet underlying code developed could easily be adjusted and trained on new data.

Unlike Nanonet which uses custom OpenCL kernels or Albacore — a novel ONT basecaller as of May 2017 lacking GPU support, this work used world-class computational framework tensorflow with highly optimized kernels and large development community. Therefore resulting paper's effect is showcasing Residual CNN approach or pure CNN approach with CTC loss is marginally better than already established basecaller and providing code in the contemporary framework.

6. Acknowledgments

TODO.

7. Bibliography

[1] Alexander S Mikheyev i Mandy MY Tin. A first look at the oxford nanopore minion sequencer. *Molecular ecology resources*, 14(6):1097–1102, 2014.

[2] Kim Judge, Simon R Harris, Sandra Reuter, Julian Parkhill, i Sharon J Peacock. Early insights into the potential of the oxford nanopore minion for the detection of antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, 70 (10):2775–2778, 2015.

[3] Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, et al. Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228–232, 2016.

[4] Alexis L Norris, Rachael E Workman, Yunfan Fan, James R Eshleman, i Winston Timp. Nanopore sequencing detects structural variants in cancer. *Cancer biology & therapy*, 17(3):246–253, 2016.

[5] Miten Jain, Sergey Koren, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *bioRxiv*, stranica 128835, 2017.

[6] Nicholas J Loman, Joshua Quick, i Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods*, 12(8):733–735, 2015.

[7] Jacob Schreiber i Kevin Karplus. Analysis of nanopore data using hidden markov models. *Bioinformatics*, stranica btv046, 2015.

[8] Tamas Szalay i Jene A Golovchenko. De novo sequencing and variant calling with nanopores using poreseq. *Nature biotechnology*, 33(10):1087–1091, 2015.

[9] Winston Timp, Jeffrey Comer, i Aleksei Aksimentiev. Dna base-calling from a nanopore using a viterbi algorithm. *Biophysical journal*, 102(10):L37–L39, 2012.

[10] Matei David, Lewis Jonathan Dursi, Delia Yao, Paul C Boutros, i Jared T Simpson. Nanocall: an open source basecaller for oxford nanopore sequencing data. *Bioinformatics*, stranica btw569, 2016.

[11] V. Boža, B. Brejová, i T. Vinař. DeepNano: Deep Recurrent Neural Networks for Base Calling in MinION Nanopore Reads. *ArXiv e-prints*, Ožujak 2016.

[12] Y. LeCun, L. Bottou, Y. Bengio, i P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, i Jian Sun. Deep residual learning for image recognition. U *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, stranice 770–778, 2016.

[14] Pedro HP Savarese. Learning identity mappings with residual gates. *arXiv preprint arXiv:1611.01260*, 2016.

[15] Alex Graves, Santiago Fernández, Faustino Gomez, i Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. U *Proceedings of the 23rd international conference on Machine learning*, stranice 369–376. ACM, 2006.

[16] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Y. Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Y. Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, i Zhenyao Zhu. Deep speech 2: End-to-end speech recognition in english and mandarin. *CoRR*, abs/1512.02595, 2015. URL `http://arxiv.org/abs/1512.02595`.

[17] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard,

Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, i Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `http://tensorflow.org/`. Software available from tensorflow.org.