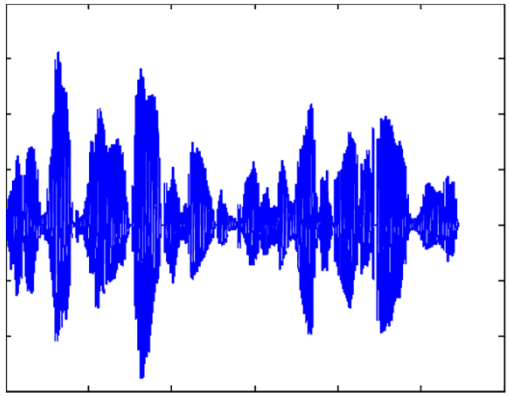
Deep Learning in sequencing data analysis

2019 Dragon Star Bioinformatics Course (Day 5)

Background

Machine Learning



Deep learning (also known as **deep structured learning** or **hierarchical learning**) is part of a broader family of machine learning methods based on artificial neural networks. Learning can be supervised, semi-supervised or unsupervised.[1][2][3]

Deep learning architectures such as deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases superior to human experts.[4][5][6]

Artificial Neural Networks (ANNs) were inspired by information processing and distributed communication nodes in biological systems. ANNs have various differences from biological brains. Specifically, neural networks tend to be static and symbolic, while the biological brain of most living organisms is dynamic (plastic) and analog.[7][8][9]

Label images

Speech recognition

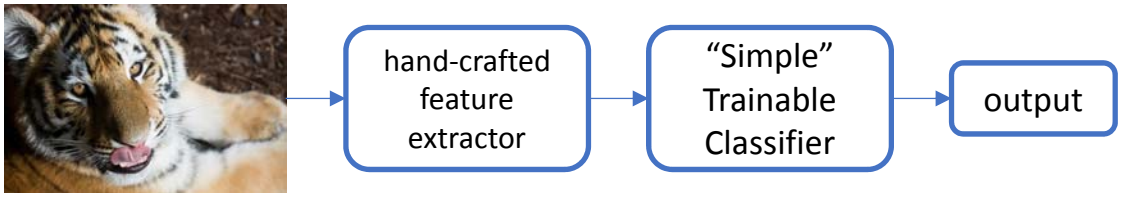
Text topics

Build a model based on known data
Rely on patterns and inference
Make predictions to perform tasks like human

Images are borrowed online

Traditional machine learning

- Traditional machine learning models use hand-crafted features and relatively simple trainable classifier.

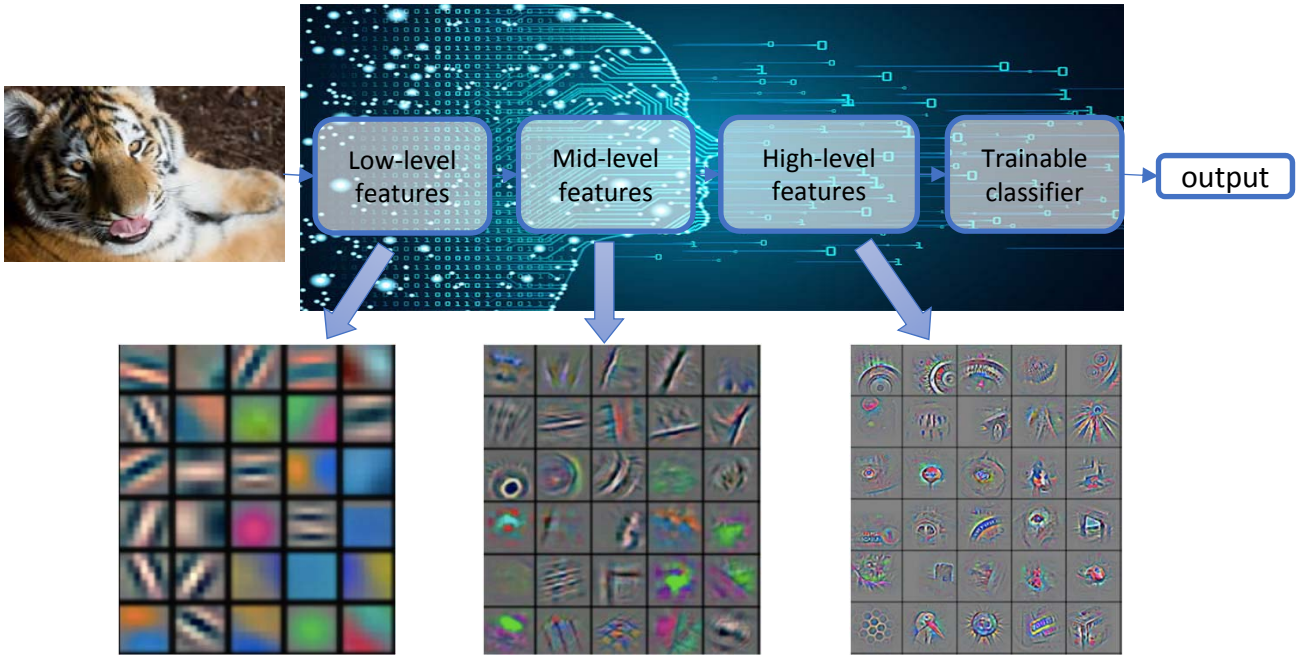


- This approach has the following limitations:
  - It is very tedious and costly to develop hand-crafted features
  - The hand-crafted features are usually highly dependents on one application, and cannot be transferred easily to other applications

Borrowed from Prof. Qiang Yang's slides

Deep learning

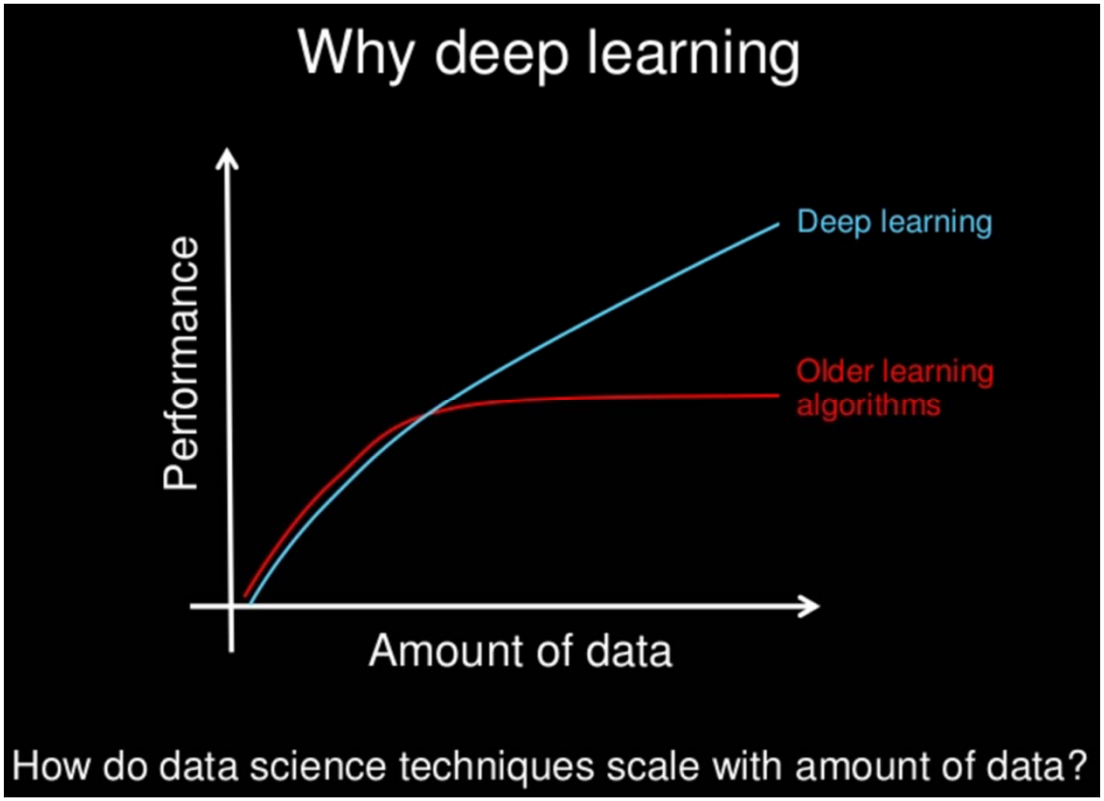
- Deep learning (a.k.a. representation learning) seeks to learn rich hierarchical representations (i.e. features) automatically through multiple stage of feature learning process.



Feature visualization of convolutional net trained on ImageNet (Zeiler and Fergus, 2013)

Borrowed from Prof. Qiang Yang's slides

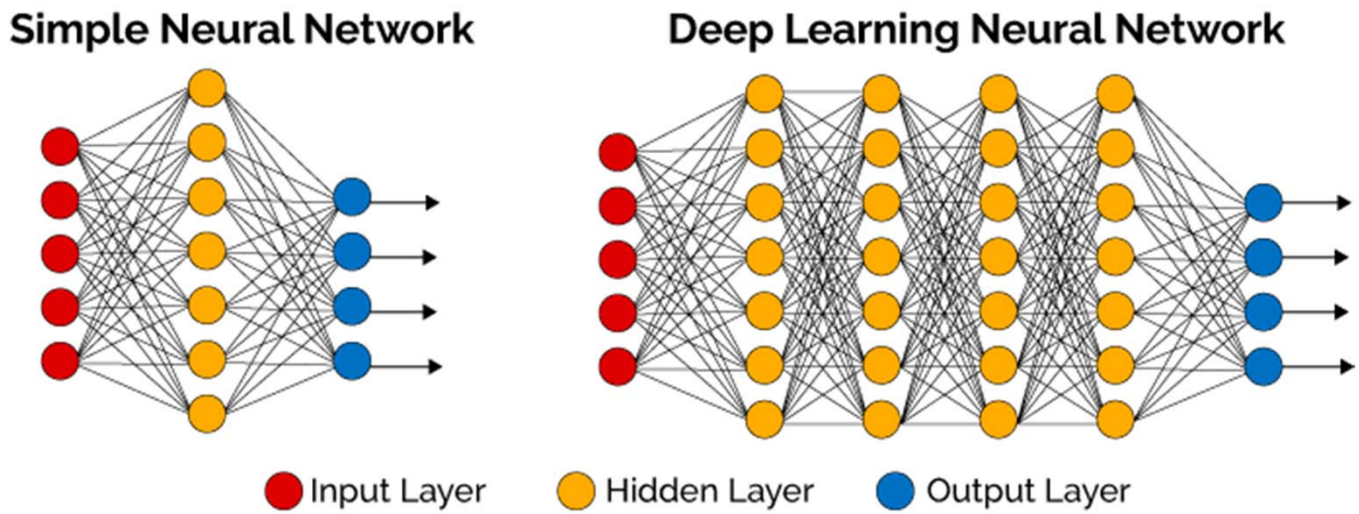
Background – deep learning



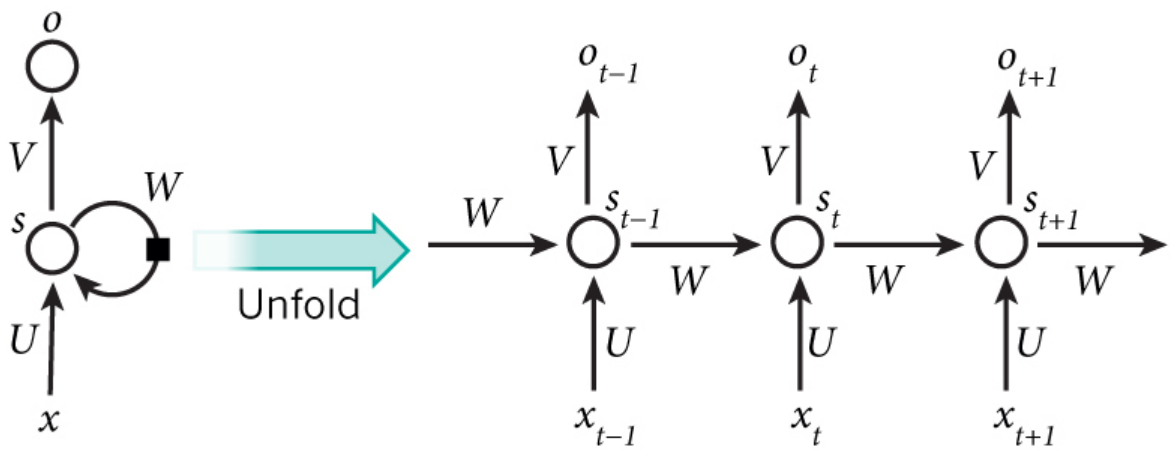
Deep learning

- Deep learning
  - Supervised, semi-supervised, and unsupervised
  - Types of deep learning frameworks:
    - deep neural networks, deep belief networks, recurrent neural networks and so on
  - Use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation
    - Each successive layer uses the output from the previous layer as input.
  - Learn multiple levels of representations
    - Correspond to different levels of abstraction
    - The levels form a hierarchy of concepts.

Deep neural network



Recurrent neural network (RNN)

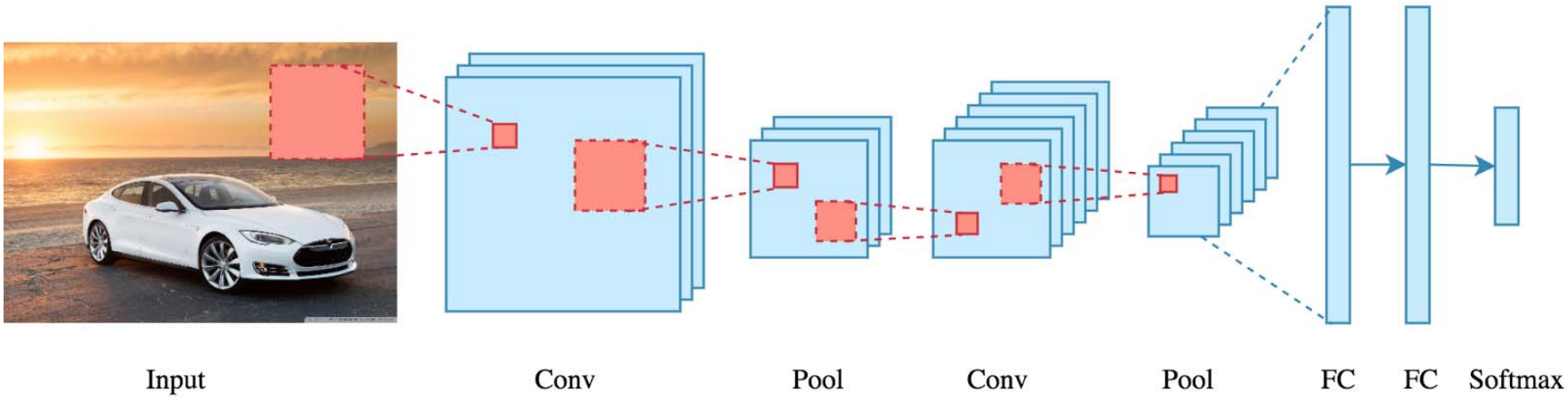


Recurrent neural network (RNN)

- Types of RNN
  - Forward RNN
  - Bidirectional RNN

- Cell types in RNN
  - LSTM --- long short-term memory cell
  - GRU --- gated recurrent unit

Convolutional neural network (CNN)



Convolutional Neural Networks

- Convolution

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Input

1	0	1
0	1	0
1	0	1

Filter / Kernel

Convolution

https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Convolutional Neural Networks

- Convolution

1x1	1x0	1x1	0	0
0x0	1x1	1x0	1	0
0x1	0x0	1x1	1	1
0	0	1	1	0
0	1	1	0	0

4		

Convolution

Convolutional Neural Networks

- Convolution

1	1x1	1x0	0x1	0
0	1x0	1x1	1x0	0
0	0x1	1x0	1x1	1
0	0	1	1	0
0	1	1	0	0

4	3	

Convolution

Convolutional Neural Networks

- Convolution

1	1	1x1	0x0	0x1
0	1	1x0	1x1	0x0
0	0	1x1	1x0	1x1
0	0	1	1	0
0	1	1	0	0

4	3	4

Convolution

Convolutional Neural Networks

- Convolution

1	1	1	0	0
0x1	1x0	1x1	1	0
0x0	0x1	1x0	1	1
0x1	0x0	1x1	1	0
0	1	1	0	0

4	3	4
2		

Convolution

Convolutional Neural Networks

- Convolution

1	1	1	0	0
0	1x1	1x0	1x1	0
0	0x0	1x1	1x0	1
0	0x1	1x0	1x1	0
0	1	1	0	0

4	3	4
2	4	

Convolution

Convolutional Neural Networks

- Convolution

1	1	1	0	0
0	1	1x1	1x0	0x1
0	0	1x0	1x1	1x0
0	0	1x1	1x0	0x1
0	1	1	0	0

4	3	4
2	4	3

Convolution

Convolutional Neural Networks

- Convolution

1	1	1	0	0
0	1	1	1	0
0x1	0x0	1x1	1	1
0x0	0x1	1x0	1	0
0x1	1x0	1x1	0	0

4	3	4
2	4	3
2		

Convolution

Convolutional Neural Networks

- Convolution

1	1	1	0	0
0	1	1	1	0
0	0x1	1x0	1x1	1
0	0x0	1x1	1x0	0
0	1x1	1x0	0x1	0

4	3	4
2	4	3
2	3	

Convolution

Convolutional Neural Networks

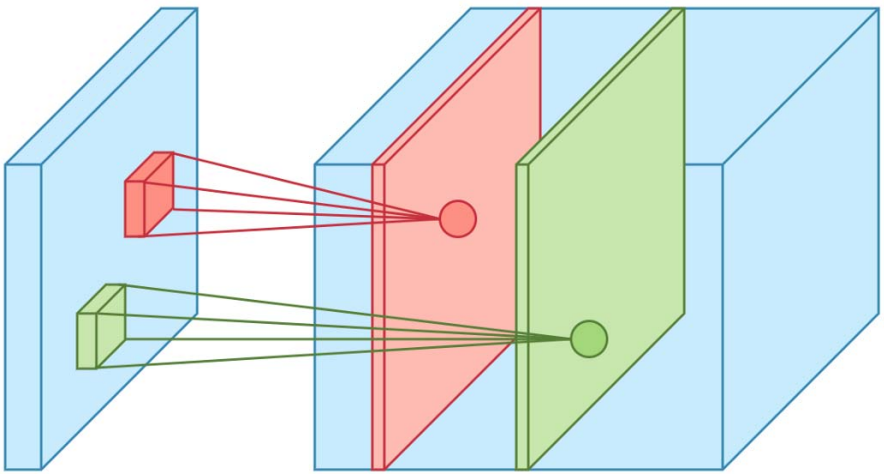
- Convolution

1	1	1	0	0
0	1	1	1	0
0	0	1x1	1x0	1x1
0	0	1x0	1x1	0x0
0	1	1x1	0x0	0x1

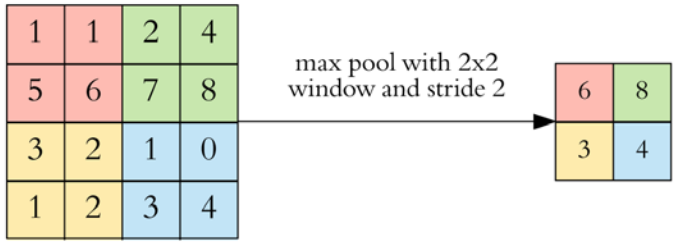
4	3	4
2	4	3
2	3	4

Convolution

Convolutional Neural Networks



Convolution with multiple filters

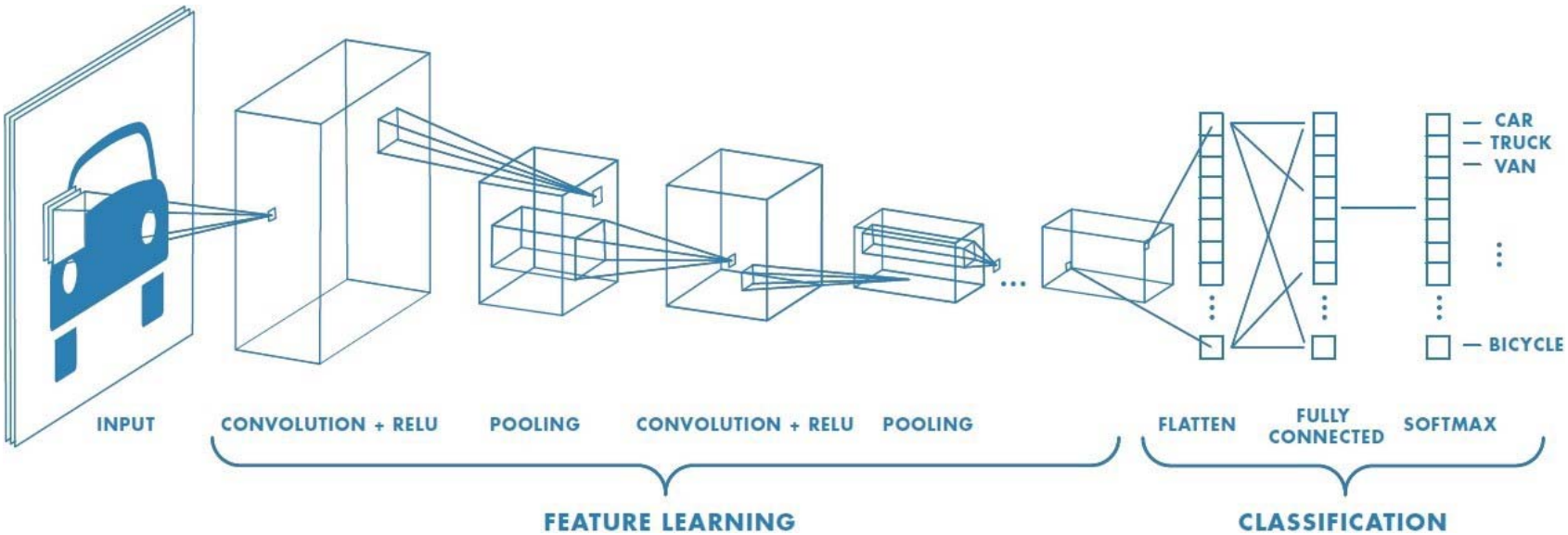


Maxpooling

https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Convolutional Neural Networks

- An example

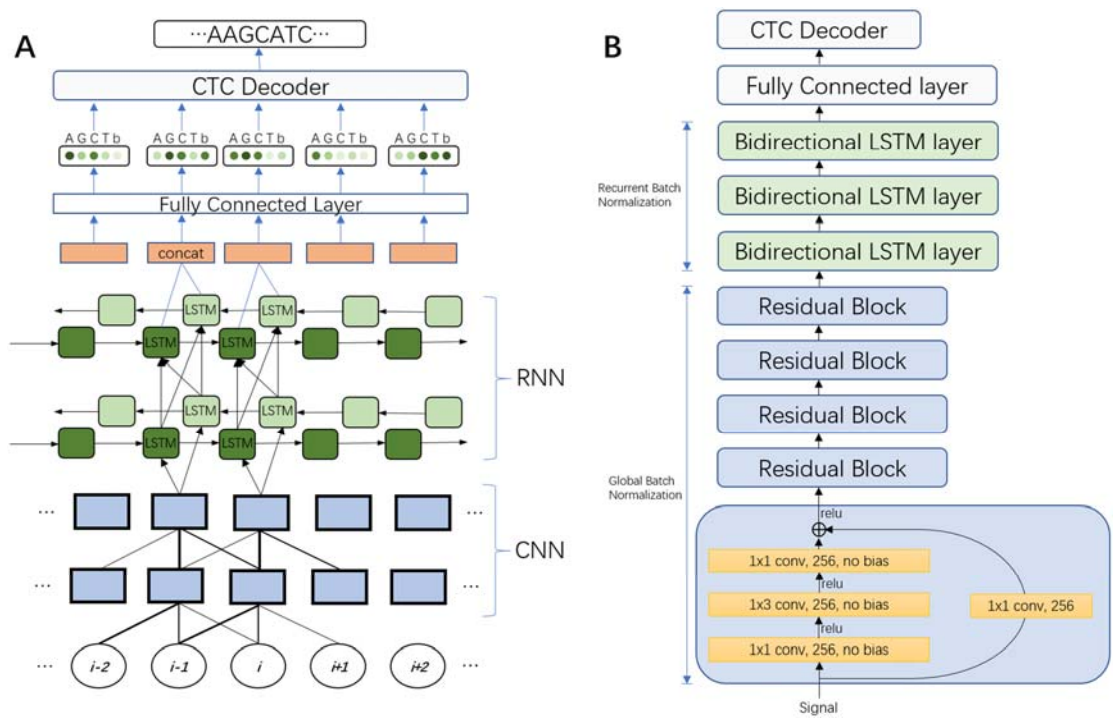


Typical CNN model

https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Chiron for Nanopore base calling

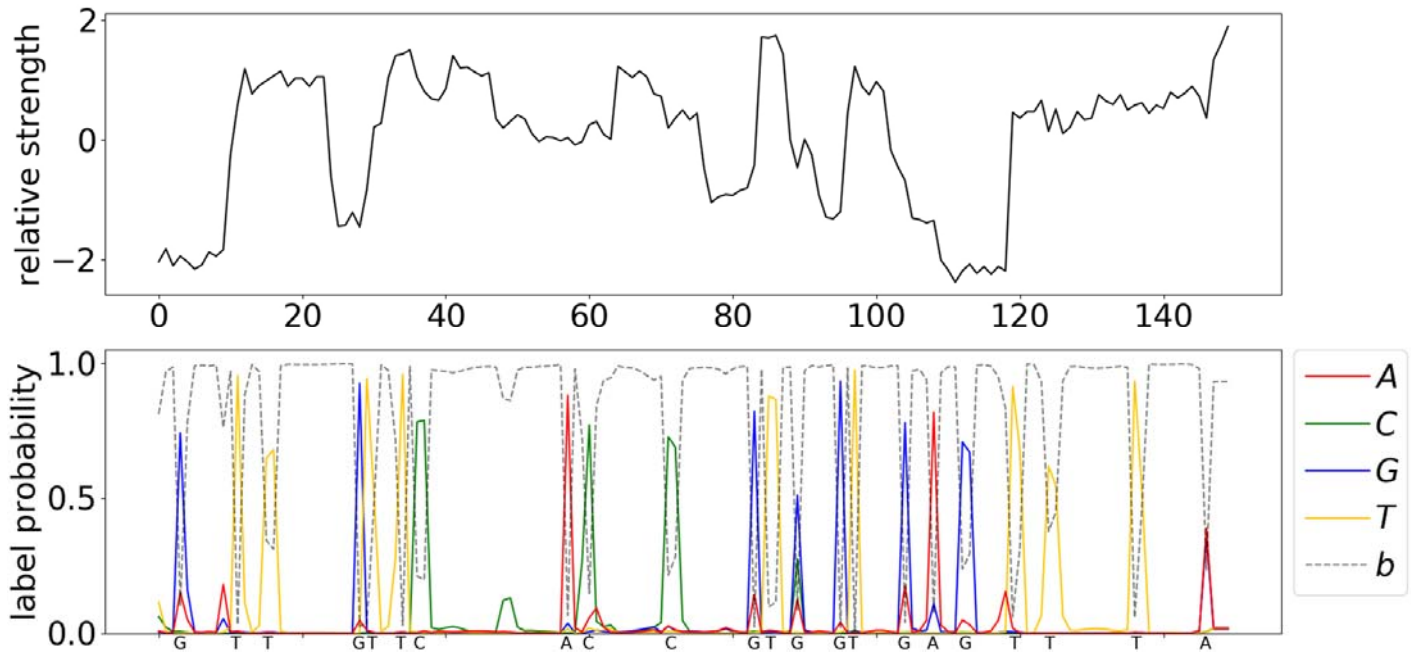
- Translating nanopore raw signal directly into nucleotide sequence using deep learning
- A novel architecture that couples a convolutional neural network (CNN) with an RNN



Teng, Giga Science, 2018

Examples of Chiron base calling

- Visualization of the predicted probability of bases and the readout sequence



Teng, Giga Science, 2018

DeepVariant for variant calling

- A deep convolutional neural network to call genetic variation in aligned sequencing reads.
- Learns statistical relationships (likelihoods) between images of read pileups around variant sites and ground-truth genotype calls.

- Learned model generalizes across genome builds and mammalian species.
- Call variants in a variety of sequencing technologies and experimental designs, from deep whole genomes to Ion Ampliseq exomes

Poplin et al. Nature Biotechnology, 2018

DeepVariant: Workflow

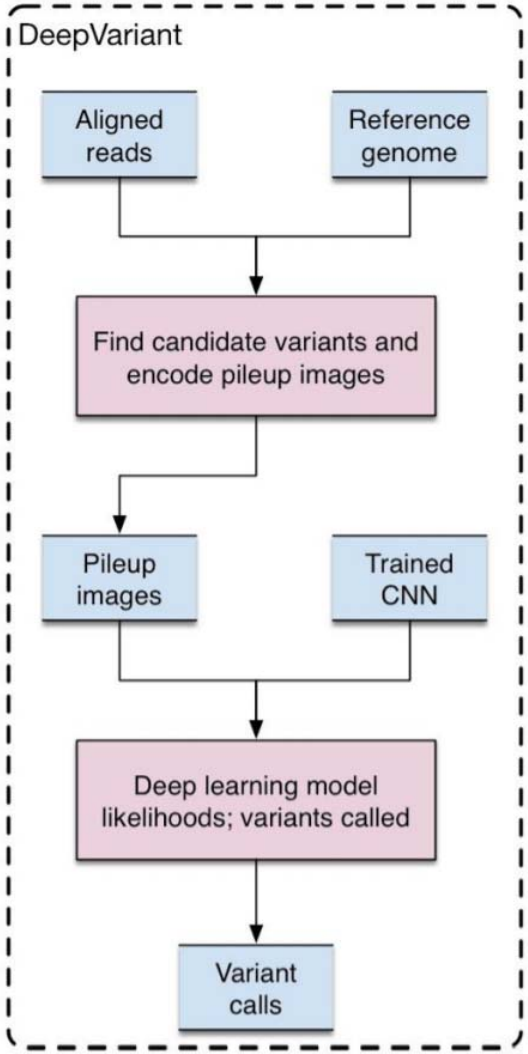
- 1. DeepVariant begins by finding SNPs and indels in reads aligned to the reference genome.
  - Candidate sites chosen with high sensitivity but low specificity.
  - Criteria includes read depth, alternative allele frequencies and base qualities.
- 2. Pileup image of reference and read data around each candidate site is created.
- 3. Using adaptation of Inception_v2 network architecture
- 4. Pileup images and emits probabilities for each of the three diploid genotypes (hom-ref, hom-alt, het) at the candidate site

- The model is trained using labeled true genotypes and is saved for future application to novel samples.

Poplin et al. Nature Biotechnology, 2018

DeepVariant: Workflow

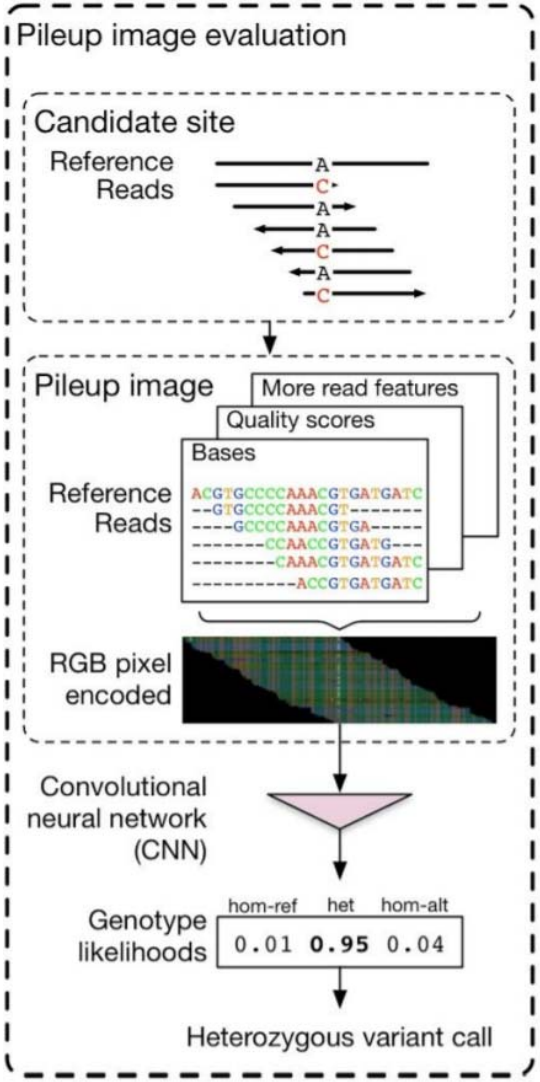
- The input
  - A BAM file after alignment
  - A reference genome in a FASTA file
- The output
  - Variant calls saved in a VCF file format.



Poplin et al. Nature Biotechnology, 2018

DeepVariant: Workflow

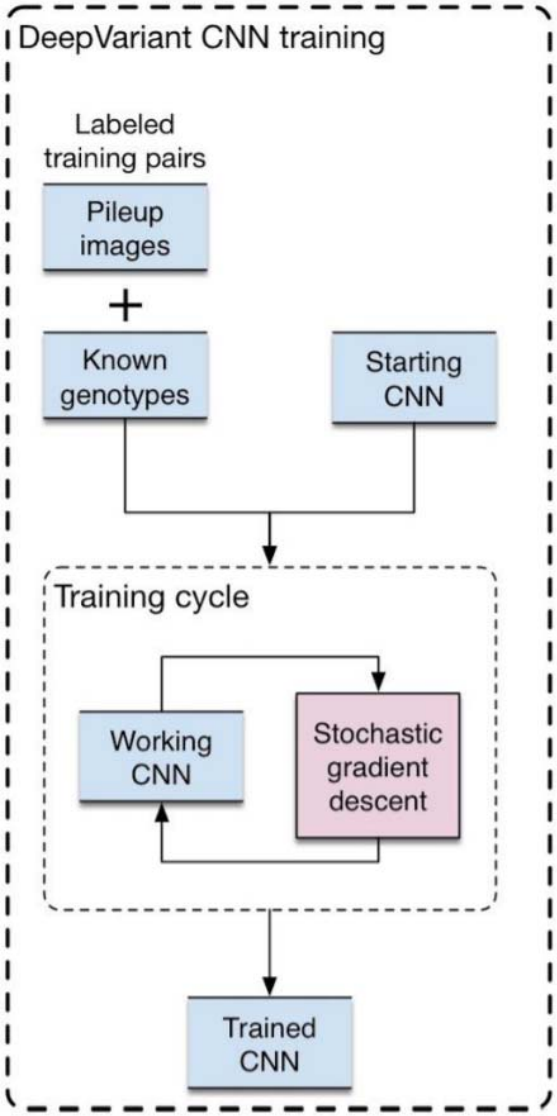
- How to create the pileup images.
  - Each pileup image is an RGB channel image.
  - First five row stores reference sequence information
  - Following rows store information about each individual read.
  - Red channel encodes each base by a different color.
  - Green channel encodes the base quality. Bases in reference rows have 60 (maximum) value by default.
  - Blue channel encodes if the read is on positive strand or not. Reference rows are positive by default.



Poplin et al. Nature Biotechnology, 2018

DeepVariant: Workflow

- In a training process:
  - Inputs are pileup images and known genotypes
  - Model is updated after each cycle using stochastic gradient descent



Poplin et al. Nature Biotechnology, 2018

DeepVariant: Performance

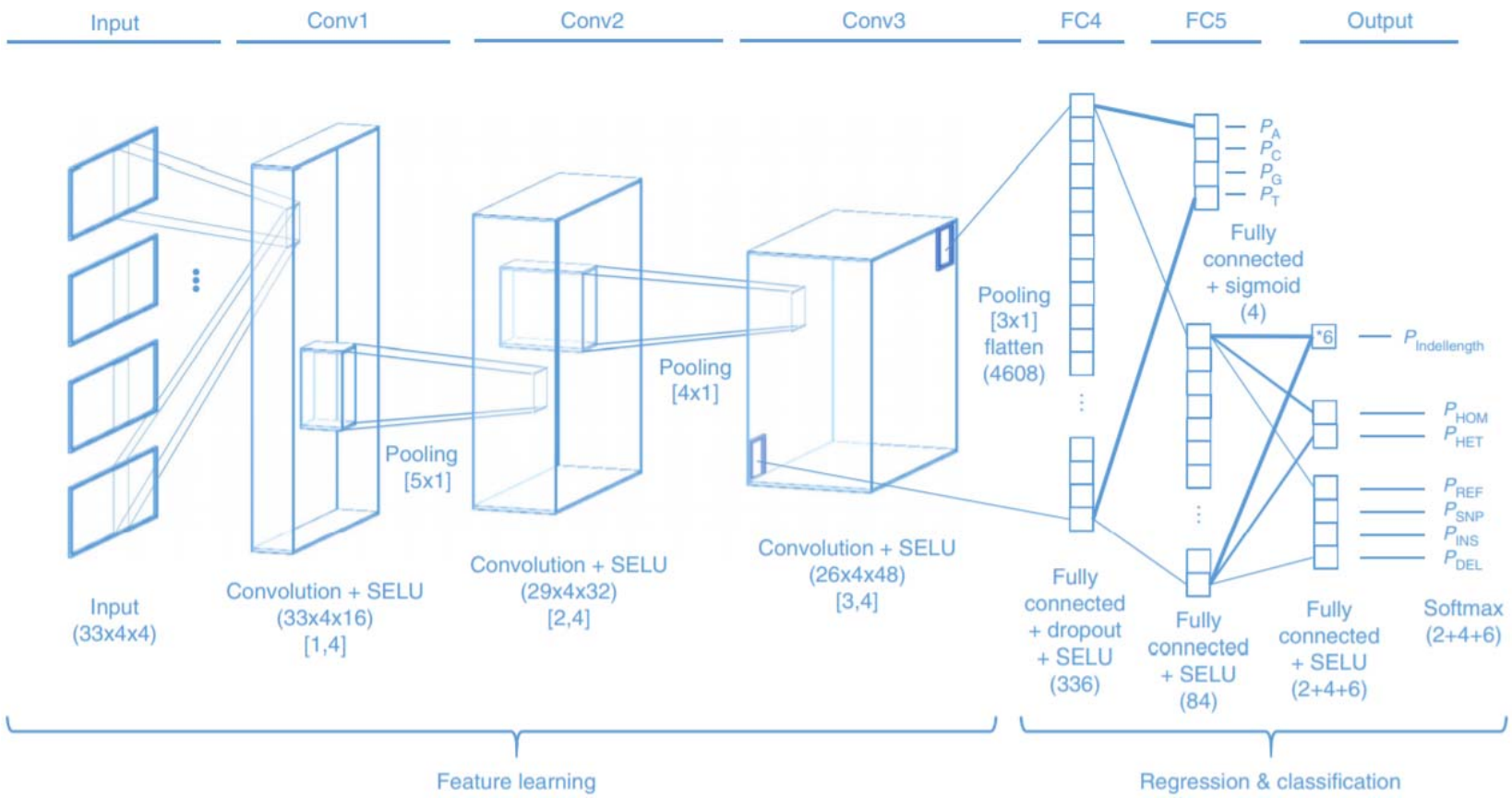
Table 1 Evaluation of several bioinformatics methods on the high-coverage, whole-genome sample NA24385

Method	Type	F1	Recall	Precision	TP	FN	FP	FP.gt	FP.al	
DeepVariant (live GitHub)	Indel	0.99507	0.99347	0.99666	357,641	2350	1,198	217	840	La
GATK (raw)	Indel	0.99366	0.99219	0.99512	357,181	2810	1,752	377	995	3
Strelka	Indel	0.99227	0.98829	0.99628	355,777	4214	1,329	221	855	2
DeepVariant (pFDA)	Indel	0.99112	0.98776	0.99450	355,586	4405	1,968	846	1,027	pl
GATK (VQSR)	Indel	0.99010	0.98454	0.99573	354,425	5566	1,522	343	909	3
GATK (flt)	Indel	0.98229	0.96881	0.99615	348,764	11227	1,349	370	916	3
FreeBayes	Indel	0.94091	0.91917	0.96372	330,891	29,100	12,569	9,149	3,347	v1
16GT	Indel	0.92732	0.91102	0.94422	327,960	32,031	19,364	10,700	7,745	v1
SAMtools	Indel	0.87951	0.83369	0.93066	300,120	59,871	22,682	2,302	20,282	1
DeepVariant (live GitHub)	SNP	0.99982	0.99975	0.99989	3,054,552	754	350	157	38	La
DeepVariant (pFDA)	SNP	0.99958	0.99944	0.99973	3,053,579	1,727	837	409	78	pl
Strelka	SNP	0.99935	0.99893	0.99976	3,052,050	3,256	732	87	136	2
GATK (raw)	SNP	0.99914	0.99973	0.99854	3,054,494	812	4,469	176	257	3
16GT	SNP	0.99583	0.99850	0.99318	3,050,725	4,581	20,947	3,476	3,899	v1
GATK (VQSR)	SNP	0.99436	0.98940	0.99937	3,022,917	32,389	1,920	80	170	3
FreeBayes	SNP	0.99124	0.98342	0.99919	3,004,641	50,665	2,434	351	1,232	v1
SAMtools	SNP	0.99021	0.98114	0.99945	2,997,677	57,629	1,651	1,040	200	1
GATK (flt)	SNP	0.98958	0.97953	0.99983	2,992,764	62,542	509	168	26	3

Poplin et al, Nature Biotechnology, 2018

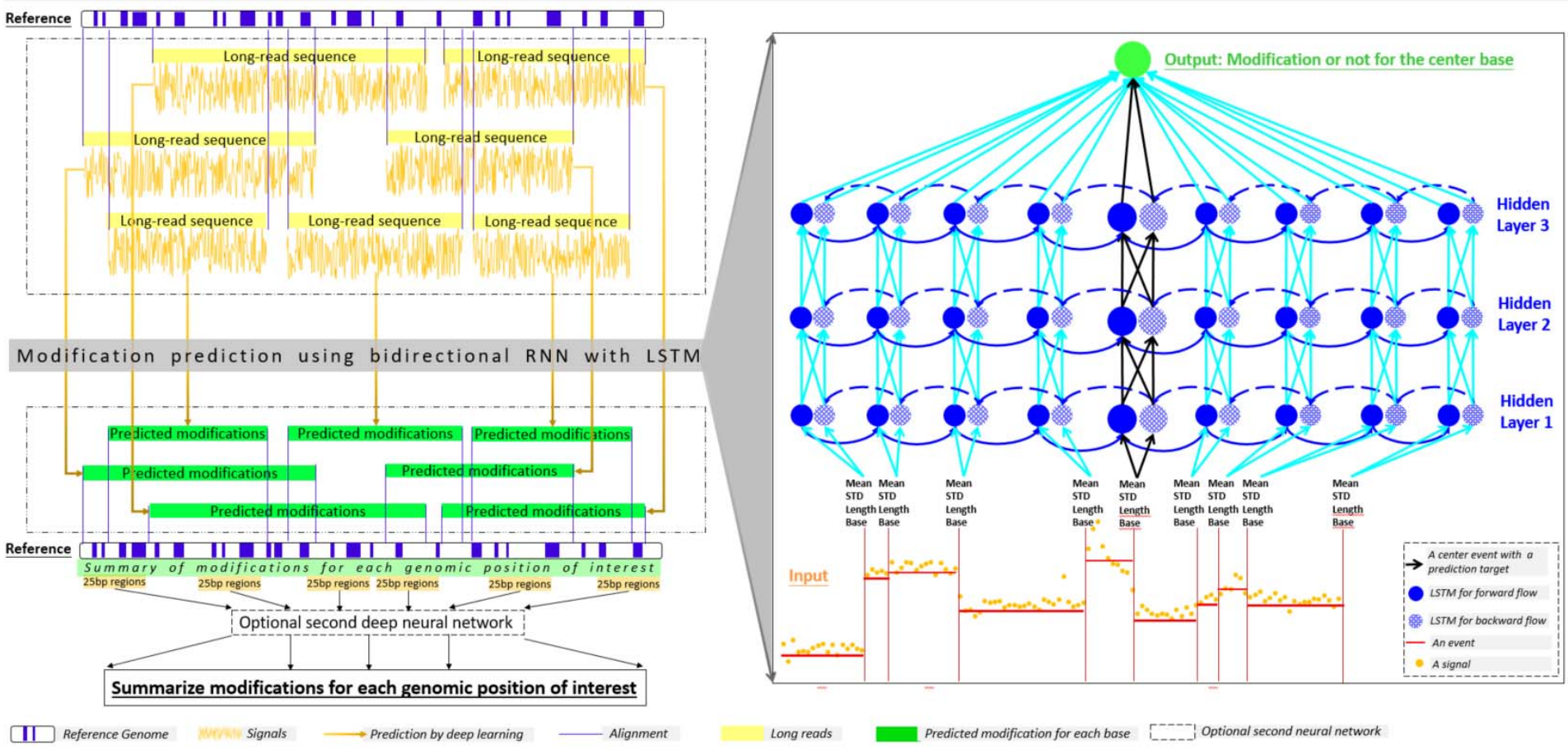
Other Long-read Variant Callers

- Clairvoyante



Luo et al. Nat Commun, 2019

DeepMod: detecting DNA methylation by LSTM RNN



Liu et al. Nat Commun, 2019

Many applications in variant interpretation already

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

Babak Alipanahi[1,2,6], Andrew Delong[1,6], Matthew T Weirauch[3–5] & Brendan J Frey[1–3]

Nat Methods. 2015 October ; 12(10): 931–934. doi:10.1038/nmeth.3547.

DANN: a deep learning approach for annotating the pathogenicity of genetic variants

Daniel Quang[1,2,†], Yifei Chen[1,†] and Xiaohui Xie[1,2,*]

[1]Department of Computer Science and [2]Center for Complex Biological Systems, University of California, Irvine, CA

nion, the first two authors should be regarded as joint First Authors.

ccepted on October 19, 2014

Predicting effects of noncoding variants with deep learning–based sequence model

Jian Zhou[1,2] and **Olga G Troyanskaya**[1,3,4]

[1]Lewis-Sigler Institute for Integrative Genomics, Princeton Univ
USA

Convolutional neural network architectures for predicting DNA–protein binding

Haoyang Zeng, Matthew D. Edwards, Ge Liu, David K. Gifford

Bioinformatics, Volume 32, Issue 12, 15 June 2016, Pages i121–i127,
https://doi.org/10.1093/bioinformatics/btw255
Published: 11 June 2016

Many more published methods for sequence data analysis recently

Title: A deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure

Tuan Trieu[1,2,3,*], Ekta Khurana[1,2,3,4,*]

1. Meyer Cancer Center, Weill Cornell Medicine, New York, New York 10065,
2. Department of Physiology and Biophysics, Weill Cornell Medicine, New Y 10065, USA.

Article | Published: 16 July 2018

Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk

Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong & Olga G. Troyanskaya

Nature Genetics **50**, 1171–1179 (2018) | Download Citation

Article | Published: 27 May 2019

Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk

Jian Zhou, Christopher Y. Park, Chandra L. Theesfeld, Aaron K. Wong, Yua
John J. Fak, Julien Funk, Kevin Yao, Yoko Tajima, Alan Packer, Robert B. D
Troyanskaya

Nature Genetics **51**, 973–980 (2019) | Download Citation

A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential

Steven T Hill, Rachael Kuintzle, Amy Teegarden, Erich Merrill, III, Padideh Danaee, David A Hendrix

Nucleic Acids Research, Volume 46, Issue 16, 19 September 2018, Pages 8105–8113, https://doi.org/10.1093/nar/gky567

A good collection of references can be found at https://github.com/hussius/deeplearning-biology

Concluding remarks

- Traditional data mining and machine learning approaches require feature engineering from data, and then build prediction models.
- Deep learning provides new paradigms to extract learning models directly from complex data.
- Deep learning approaches will play important roles in human genomics research, such as sequencing data analysis, variant interpretation and precision health.