3000788 Intro to Comp Molec Biol

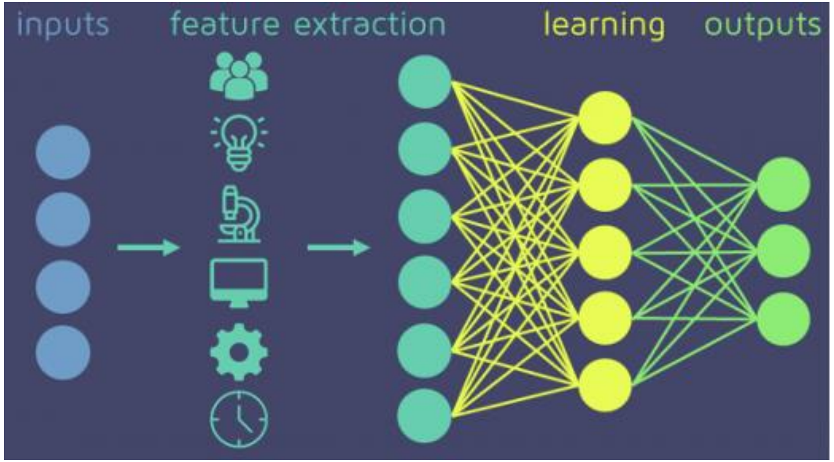
Lecture 29: Deep learning in life sciences

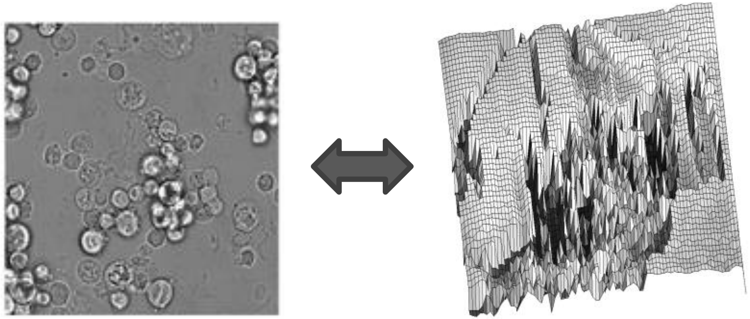
November 24, 2022

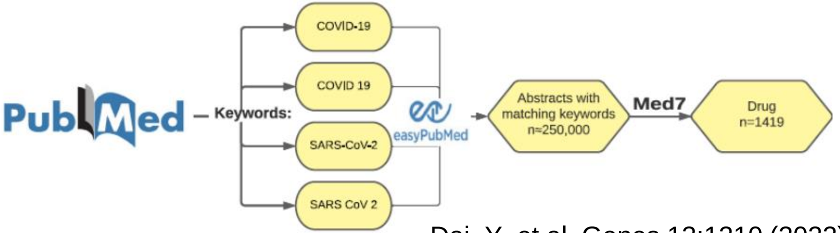
Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Limitation of classical (non-deep) learning



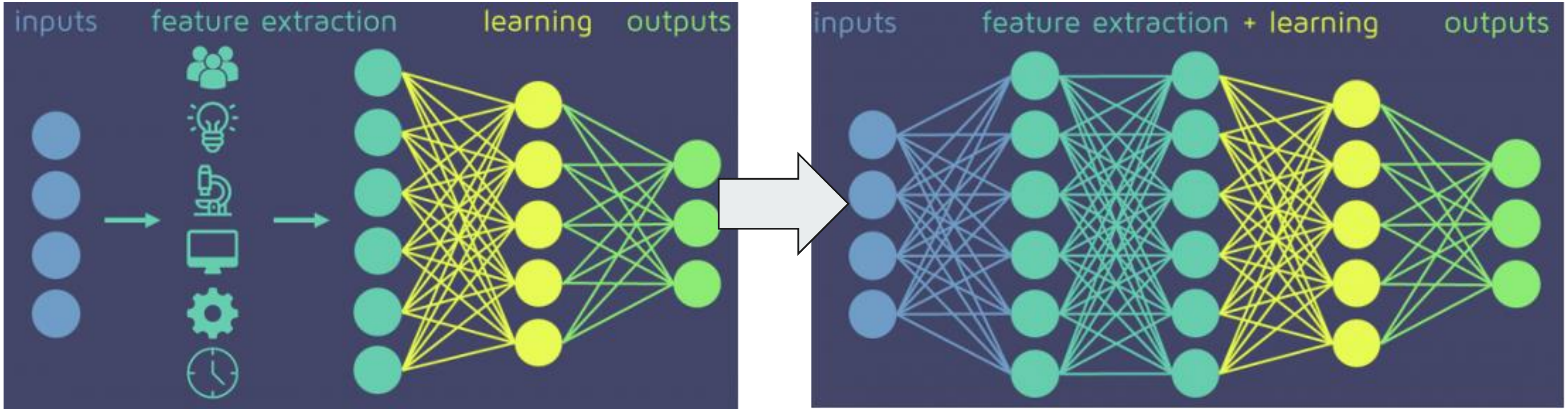




Dai, Y. et al. Genes 13:1210 (2022)

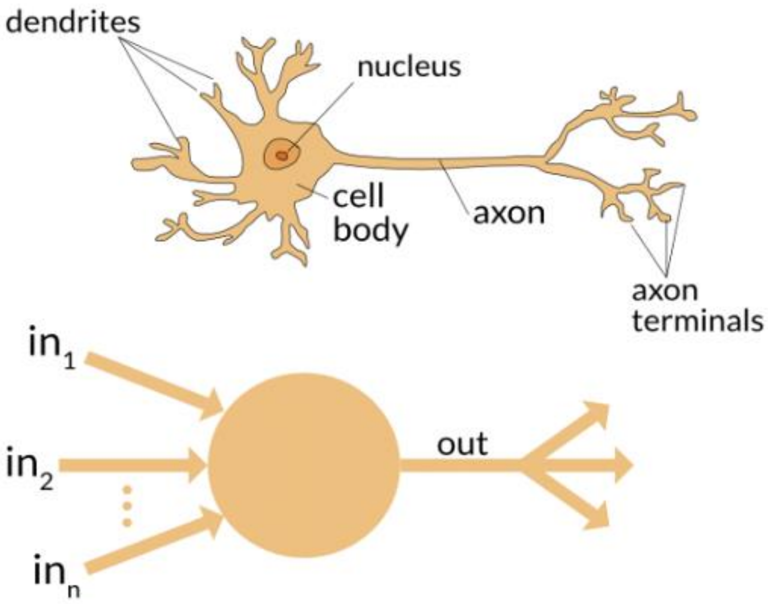
- Classical machine learning requires the input to be formatted and pre-processed by human

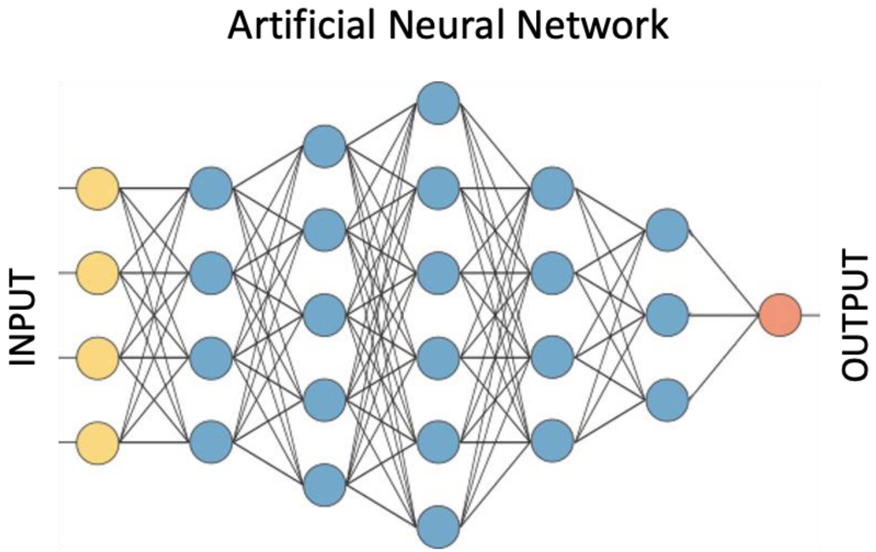
End-to-end learning



- Deep learning, via artificial neural network models, can learn to extract useful information from raw input directly
- The catch is a lot of data and supervision is needed

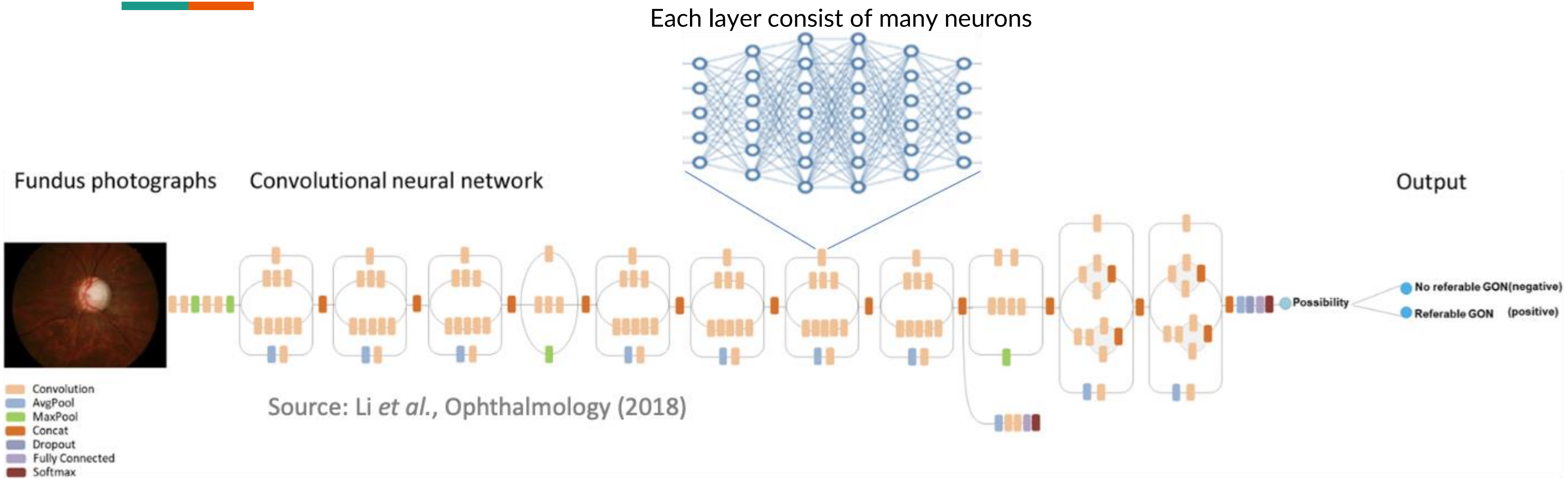
Artificial neural network





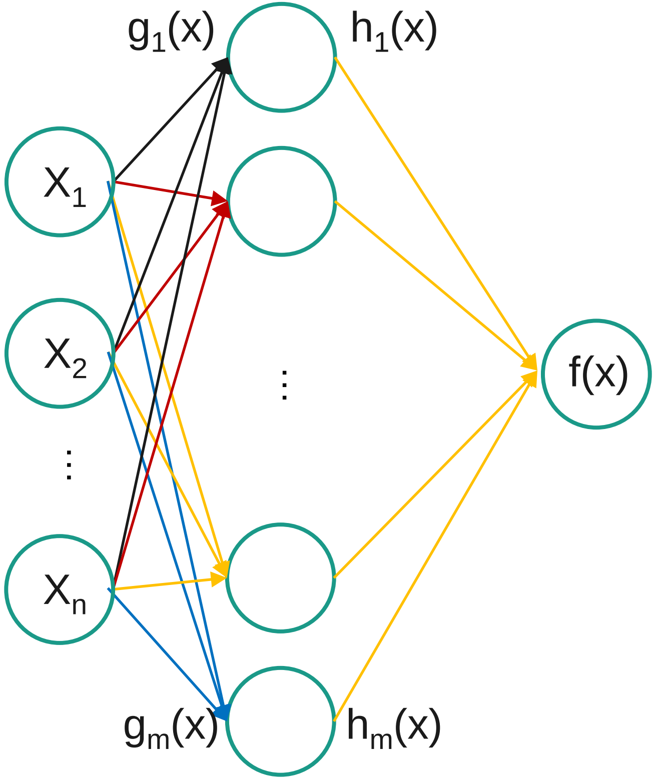
- Network of simple computation nodes: $out = f(w_1 in_1 + w_2 in_2 + \ldots + w_n in_n)$

Power from sheer number



- Deep learning = deep artificial neural network
- Hundreds of layers with >10M parameters

Calculations inside neural nentwork



Linear neuron input

- $g_1(x) = w_{1,1}x_1 + \cdots + w_{1,n}x_n$
- $g_m(x) = w_{m,1}x_1 + \cdots + w_{m,n}x_n$

Sigmoid activation

- $h_1(x) = \frac{1}{1+e^{-g_1(x)}}$
- $h_m(x) = \frac{1}{1+e^{-g_m(x)}}$

Linear aggregated output

- $f(x) = u_1 h_1(x) + \cdots + u_m h_m(x)$

Universal approximation theorem (Cybenko, 1989)

Universal Approximation Theorem: Fix a continuous function $\sigma : \mathbb{R} \to \mathbb{R}$ (activation function) and positive integers d, D. The function σ is not a polynomial if and only if, for every continuous function $f : \mathbb{R}^d \to \mathbb{R}^D$ (target function), every compact subset K of $\mathbb{R}^d$, and every $\epsilon > 0$ there exists a continuous function $f_\epsilon : \mathbb{R}^d \to \mathbb{R}^D$ (the layer output) with representation

$$f_\epsilon = \boxed{W_2 \circ \sigma \circ W_1},$$

where W_2, W_1 are composable affine maps and $\circ$ denotes component-wise composition, such that the approximation bound

$$\sup_{x \in K} \|f(x) - f_\epsilon(x)\| < \varepsilon$$

holds for any ϵ arbitrarily small (distance from f to f_ϵ can be infinitely small).

- Neural network with one hidden layer can mimic any mathematical function

Gradient of a neural network

Neuron input: $g_i(x) = w_{i,1}x_1 + \cdots + w_{i,n}x_n$

Sigmoid activation: $h_i(x) = \frac{1}{1+e^{-g_i(x)}}$

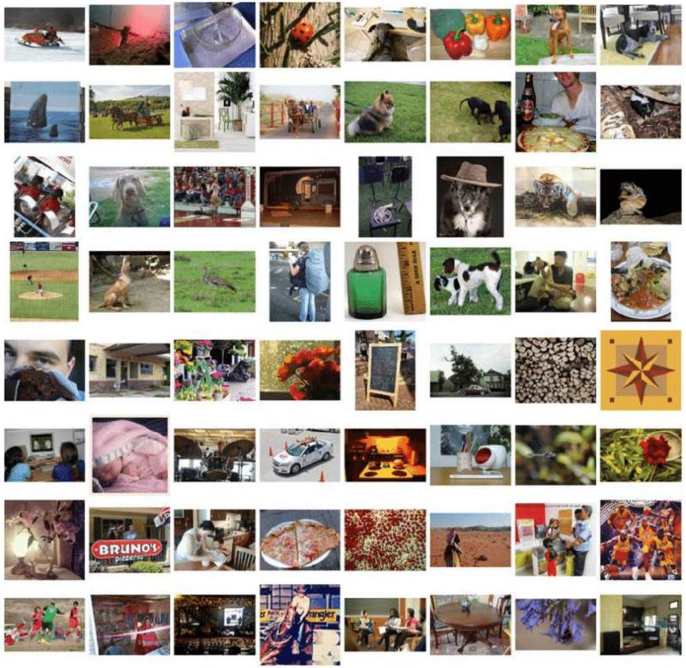
Linear output: $f(x) = u_1 h_1(x) + \cdots + u_m h_m(x)$

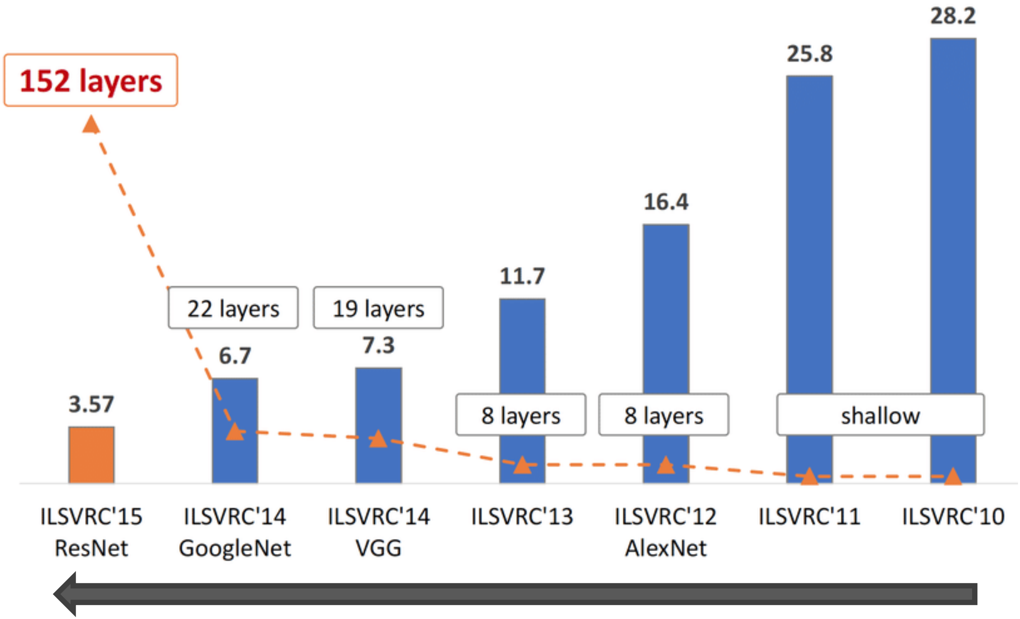
MSE loss: $L(f(x), y) = \frac{1}{2}\|f(x) - y\|^2$

Gradient: $\frac{\delta L}{\delta w_{i,j}} = \frac{\delta L}{\delta f}\frac{\delta f}{\delta h_i}\frac{\delta h_i}{\delta g_i}\frac{\delta g_i}{\delta w_{i,j}} = (f(x) - y) \cdot u_i \cdot g_i(\mathrm{x})(1 - g_i(\mathrm{x}))\, x_j$

We can update $w_{i,j}$ by following the gradient!

ImageNet: The rise of deep artificial neural network





Graphical processing unit (GPU)

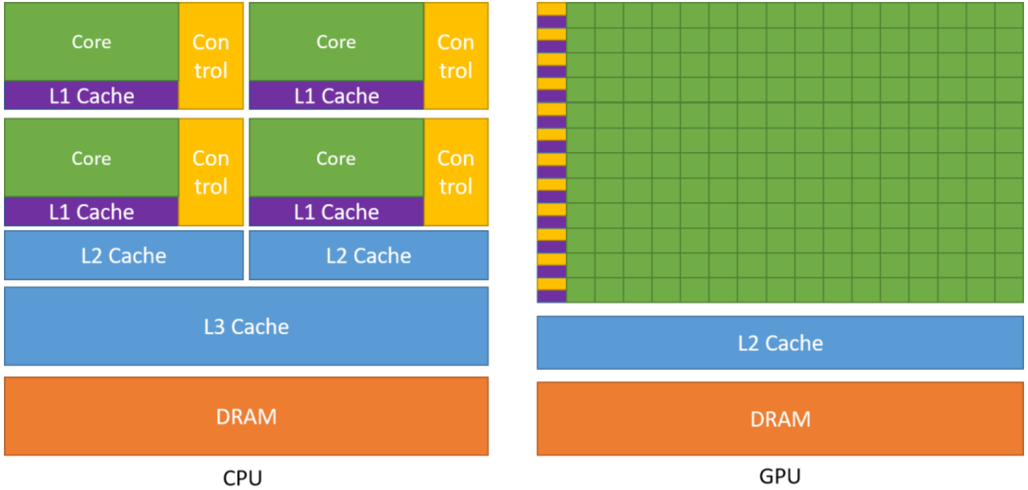


Image from analyticsvidhya.com

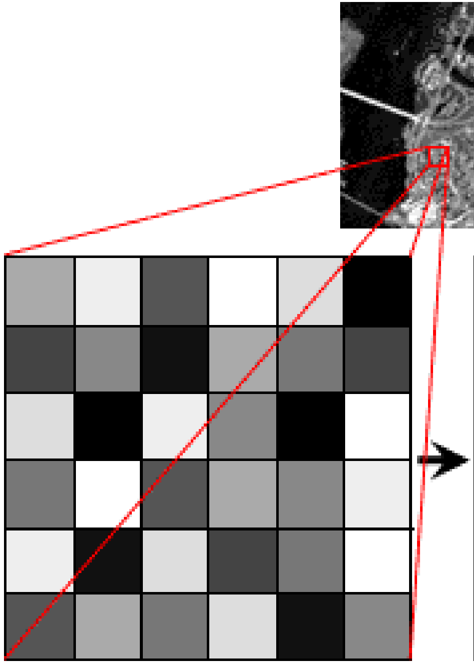
- Calculation of gradient for an ANN requires millions of simple operations that can be performed in parallel → Similar to the calculation of graphics

Representation learning

Naïve representations

	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

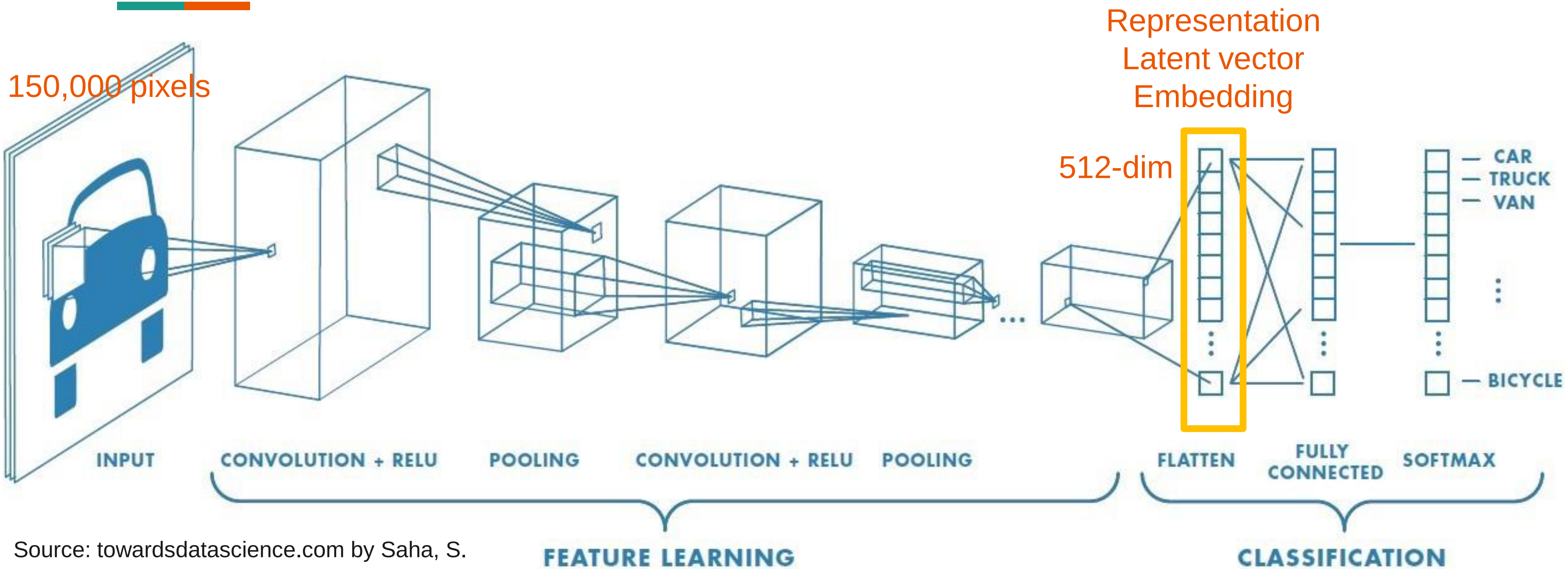
Image from hackermoon.com



170	238	85	255	221	0
68	136	17	170	119	68
221	0	238	136	0	255
119	255	85	170	136	238
238	17	221	68	119	255
85	170	119	221	17	136

Image from naushardsblog.wordpress.com

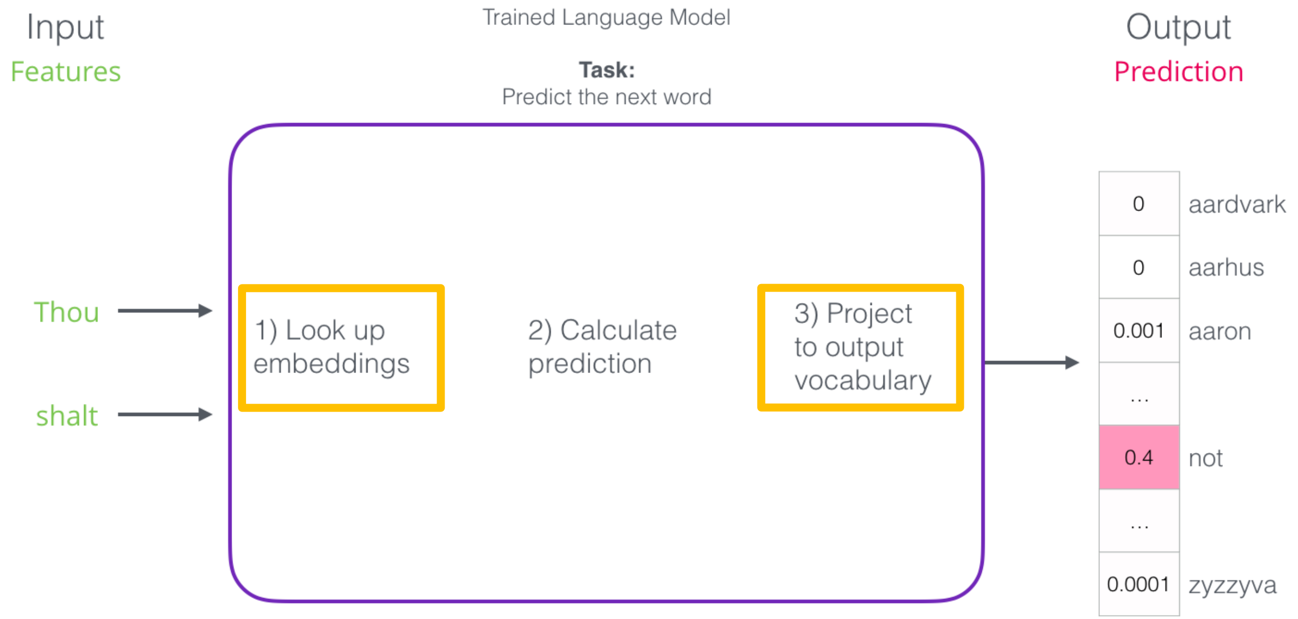
Encoder-Decoder view of neural network



Source: towardsdatascience.com by Saha, S.

- Encode raw data into useful features → decode features for prediction

Supervised representation learning



- Learn word representation by predicting next word in a sentence

Word2Vec

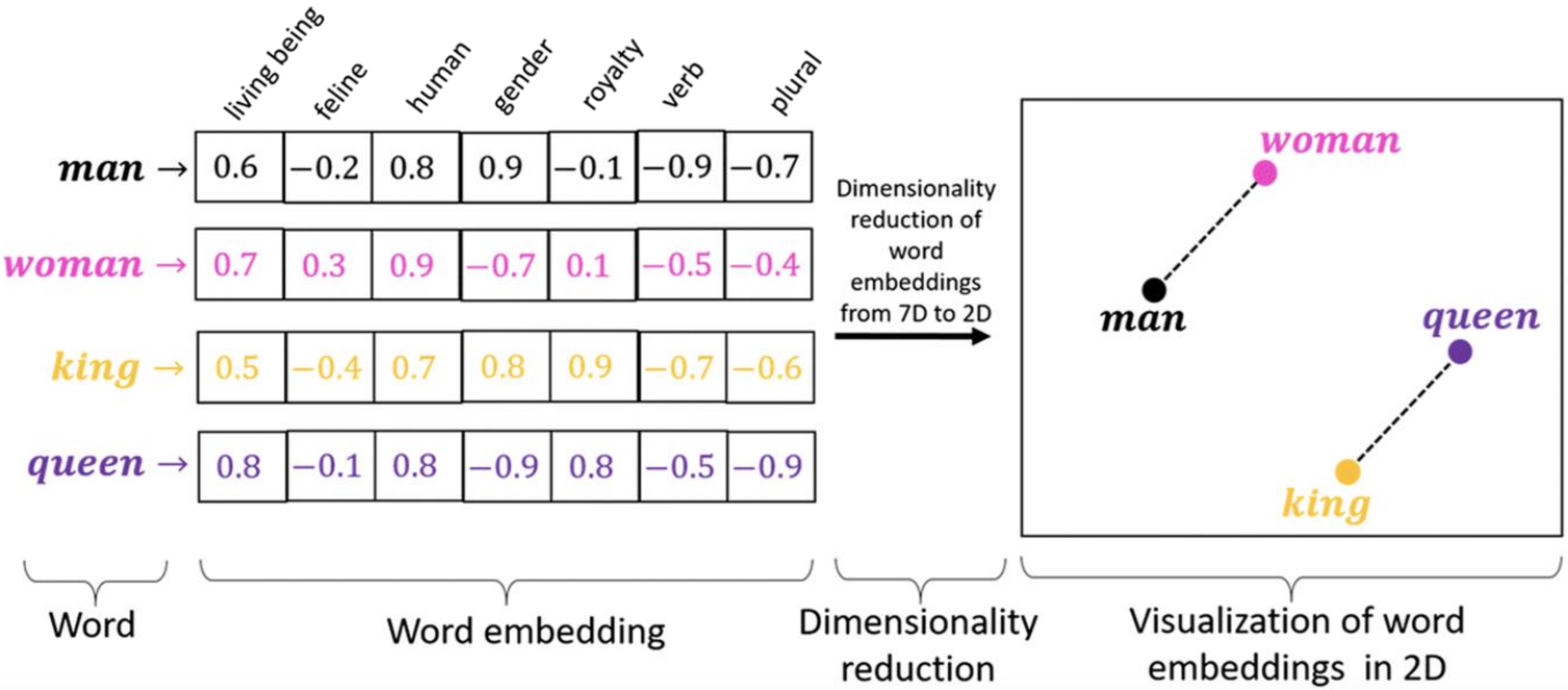
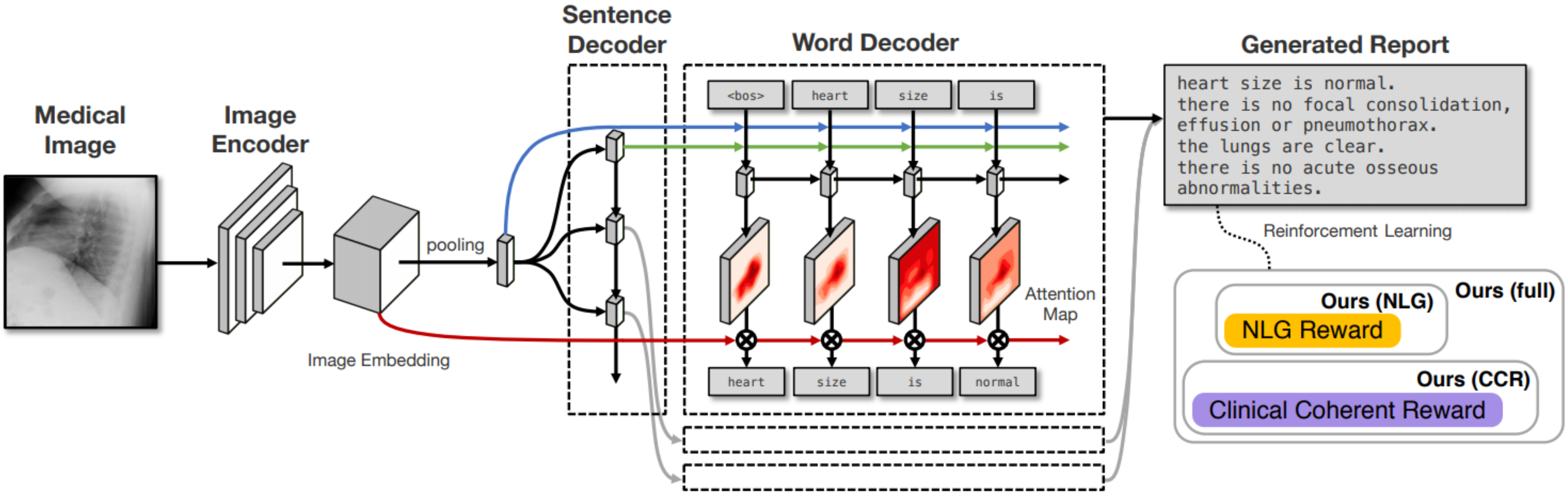


Image from medium.com

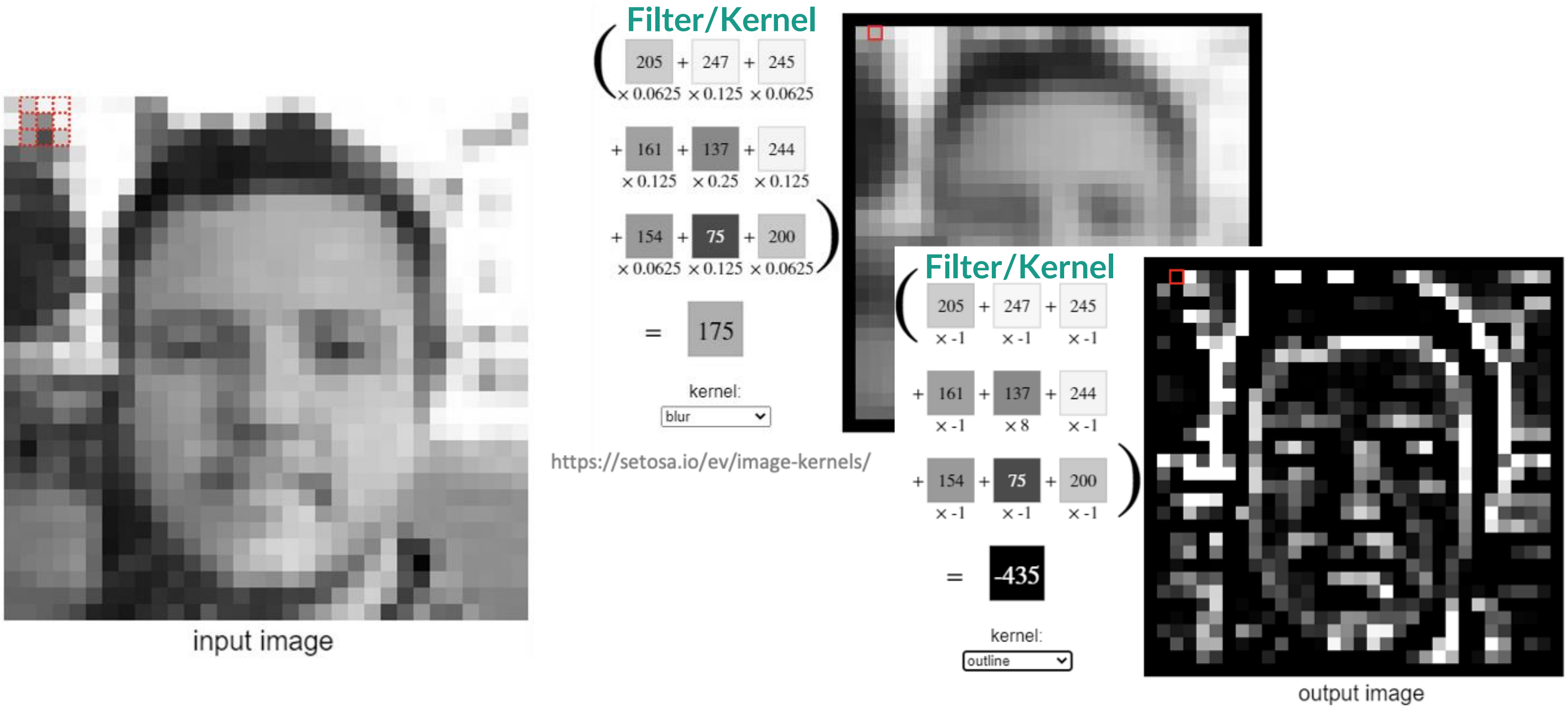
Combining image and word embeddings



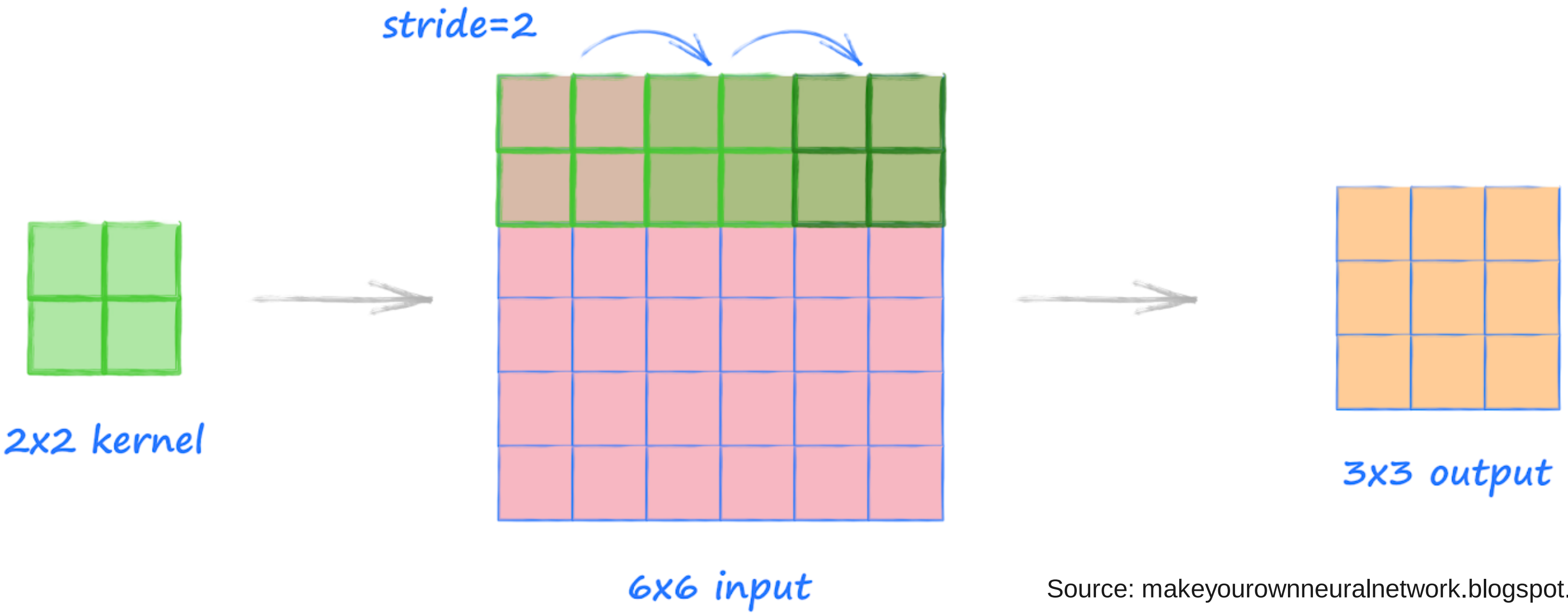
Liu et al. Clinically Accurate Chest X-Ray Report Generation (2019)

Convolution

Extracting contextual pattern with filter



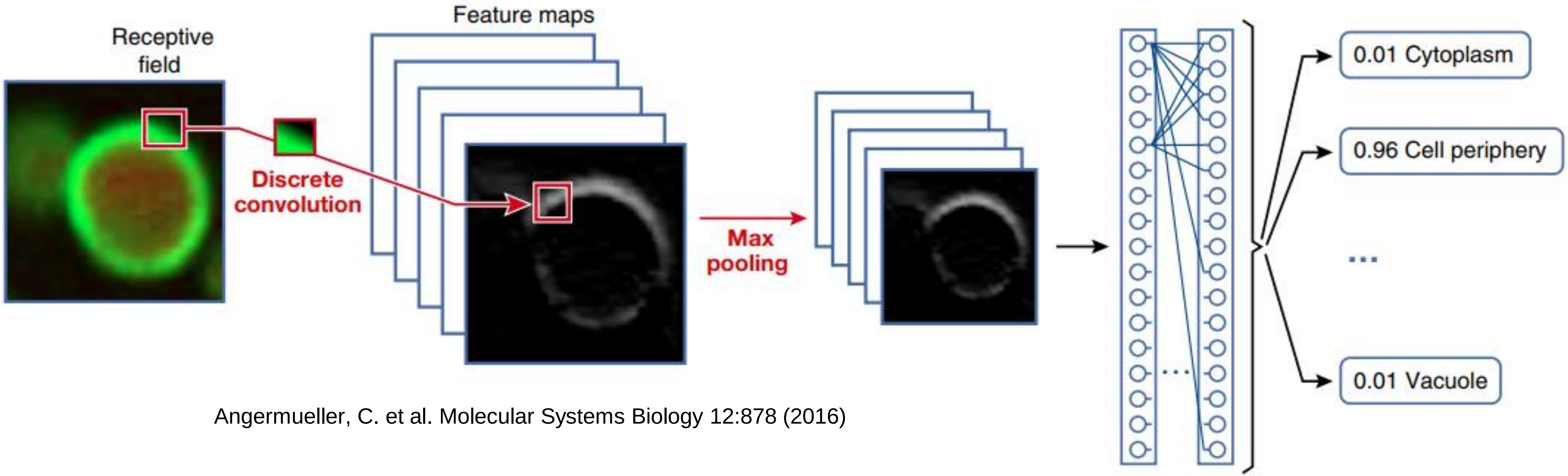
Convolutional operation



Source: makeyourownneuralnetwork.blogspot.com

- Linear combination of values in nearby pixels – applied throughout

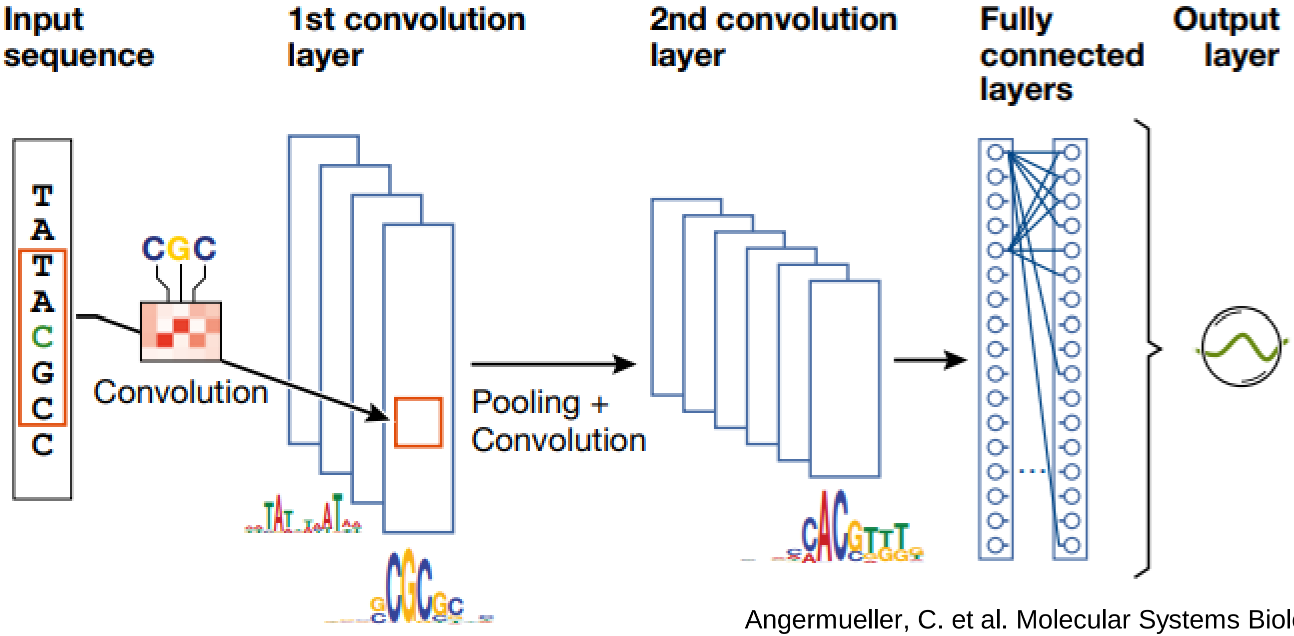
Convolutional neural network (CNN)



Angermueller, C. et al. Molecular Systems Biology 12:878 (2016)

- Instead of using human-define filters to extract contextual pattern, CNN learns the best filters from the data

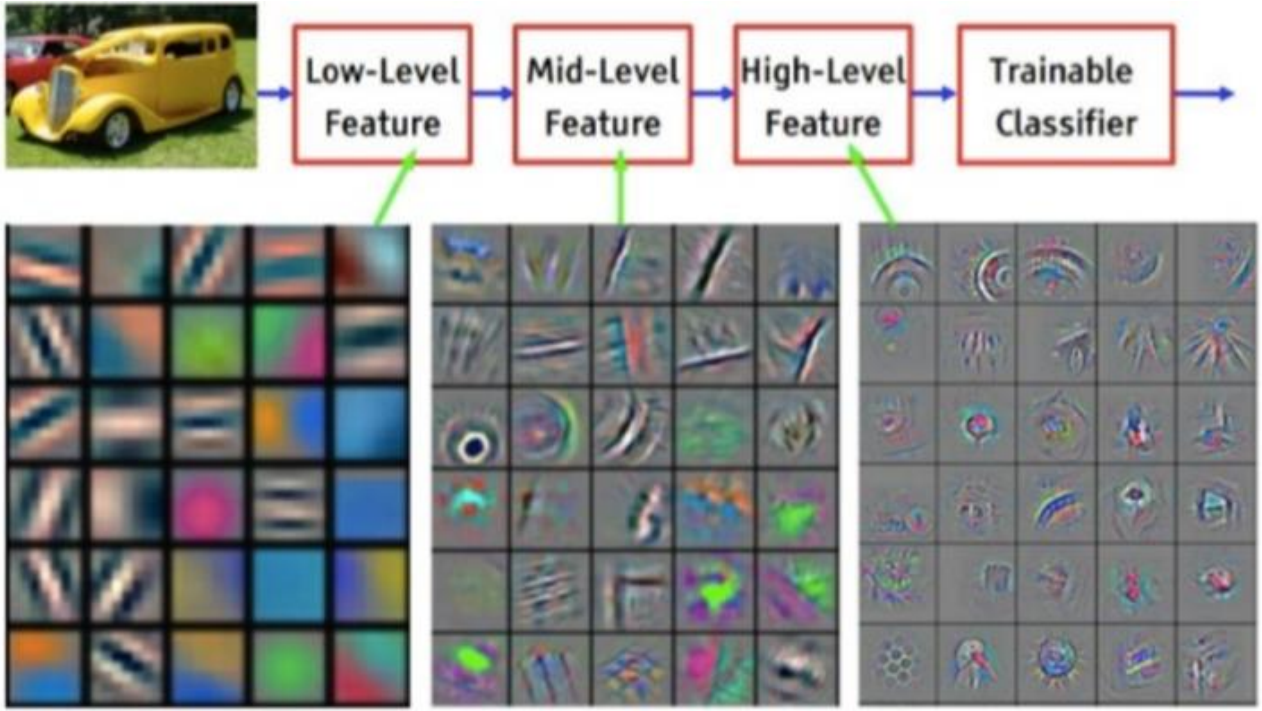
Convolution for DNA sequences



Angermueller, C. et al. Molecular Systems Biology 12:878 (2016)

- Motif = contextual pattern on DNA sequence

Hierarchical feature assembly inside CNN



Source: Zeiler and Fergus (2013)

Some CNN designs

Vanishing and exploding gradient problems

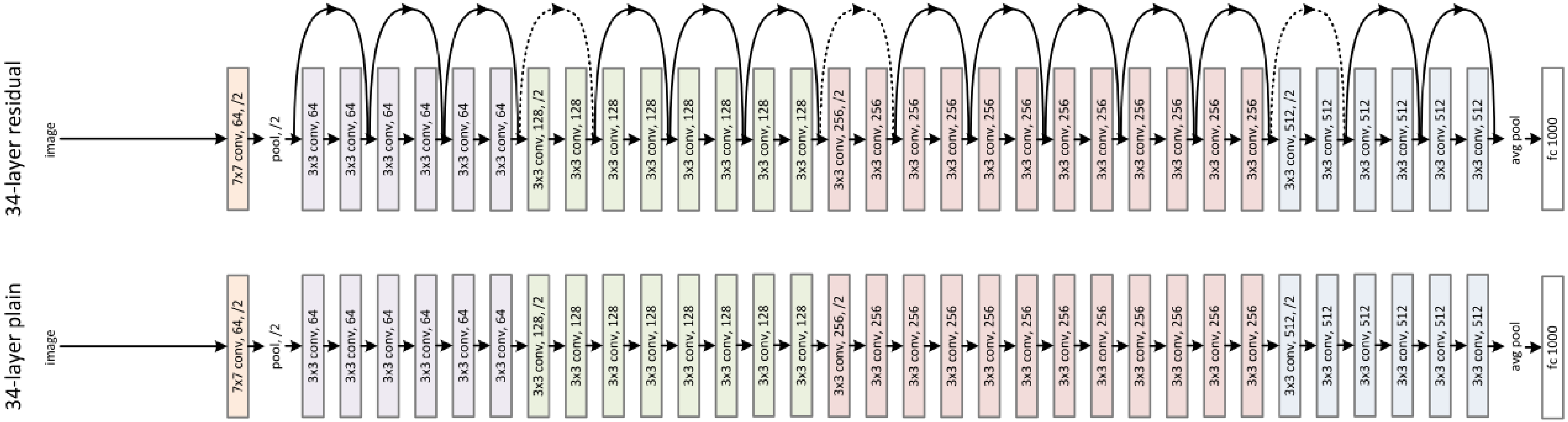
Gradient: $\frac{\delta L}{\delta w_{i,j}} = \frac{\delta L}{\delta f}\frac{\delta f}{\delta h_i}\frac{\delta h_i}{\delta g_i}\frac{\delta g_i}{\delta w_{i,j}} = (f(x) - y) \cdot u_i \cdot g_i(\mathrm{x})(1 - g_i(\mathrm{x}))\ x_j$

The number of multiplicative terms scales with the number of layers

What would happen if all values are << 1 or >> 1?

- Gradient became very small → No weight update
- Gradient became very large → Unstable

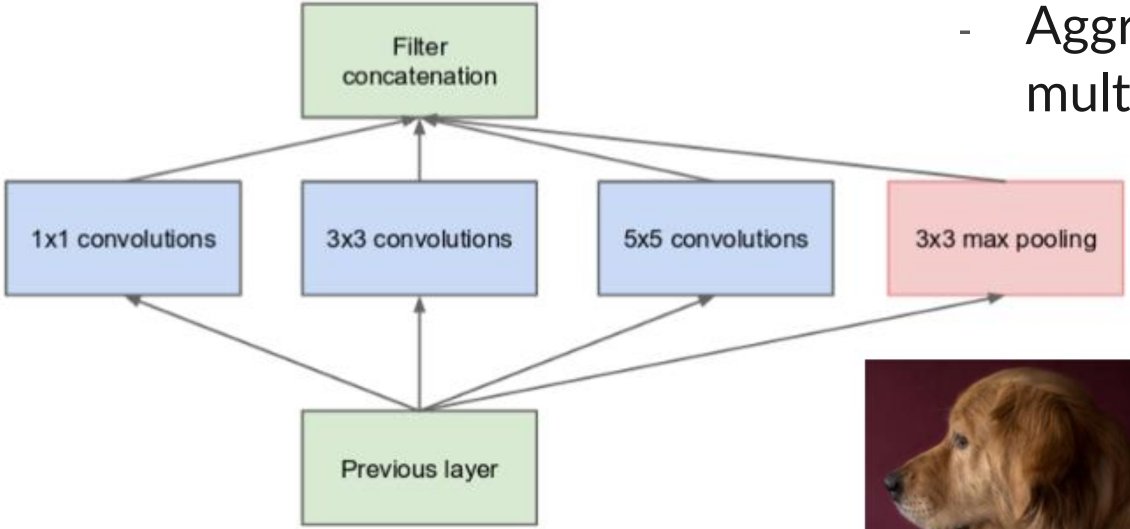
Residual network (ResNet)



Source: medium.com

- Adding skip connections jumping over blocks of convolutional layers
- Reduce the number of terms in gradient of early weights

Inception = multi-resolution layer



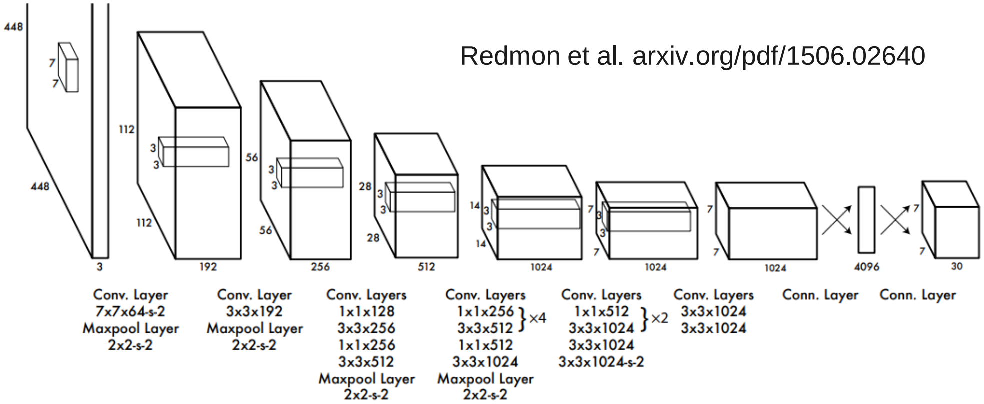
- Aggregate information from multiple resolutions together

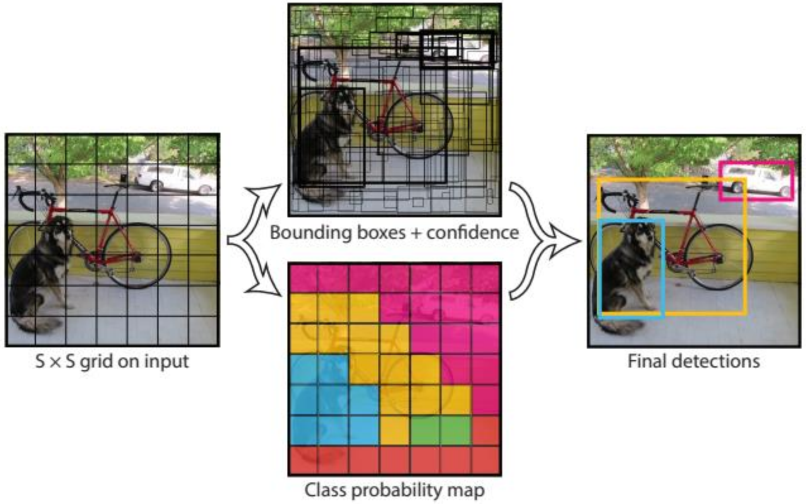
From left: A dog occupying most of the image, a dog occupying a part of it, and a dog occupying very little space (Images obtained from Unsplash).

- Convolution with multiple filter sizes per layer

YOLO (You Only Look Once)

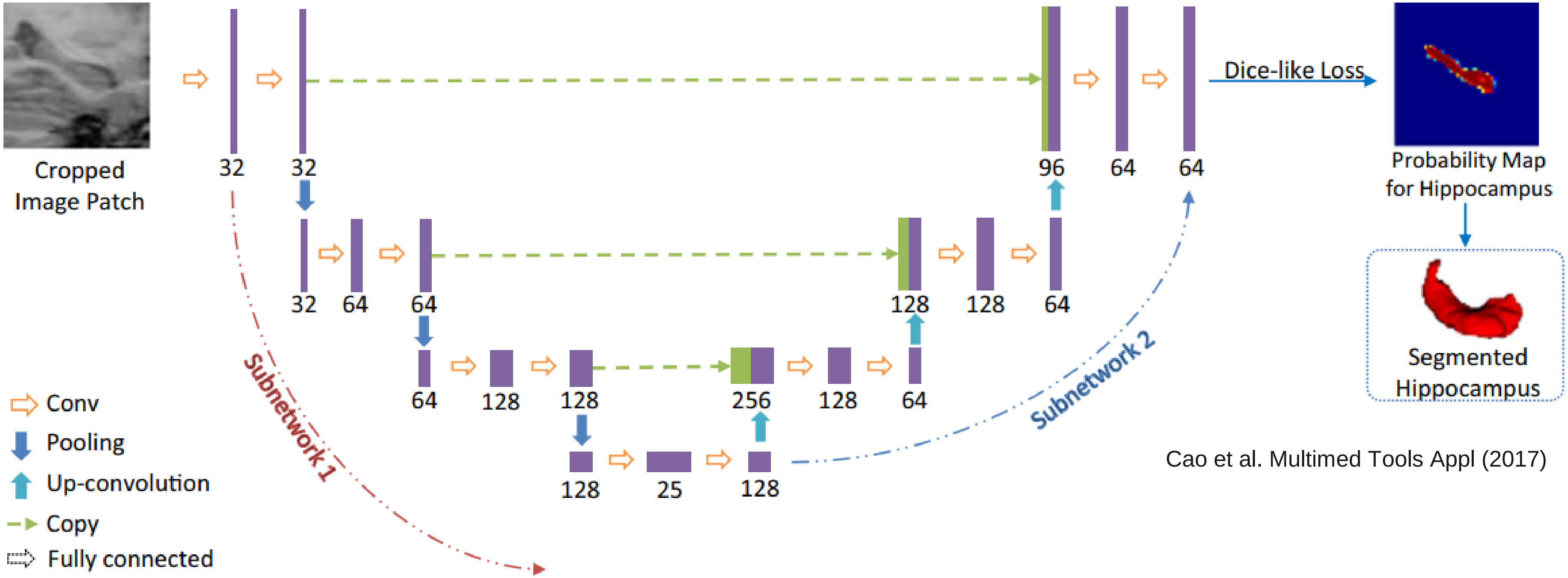
Redmon et al. arxiv.org/pdf/1506.02640





- Process each input image only once
- Simultaneously predict locations of objects and class probabilities
- Applicable to real-time video input (45-155 frames per second)

U Net

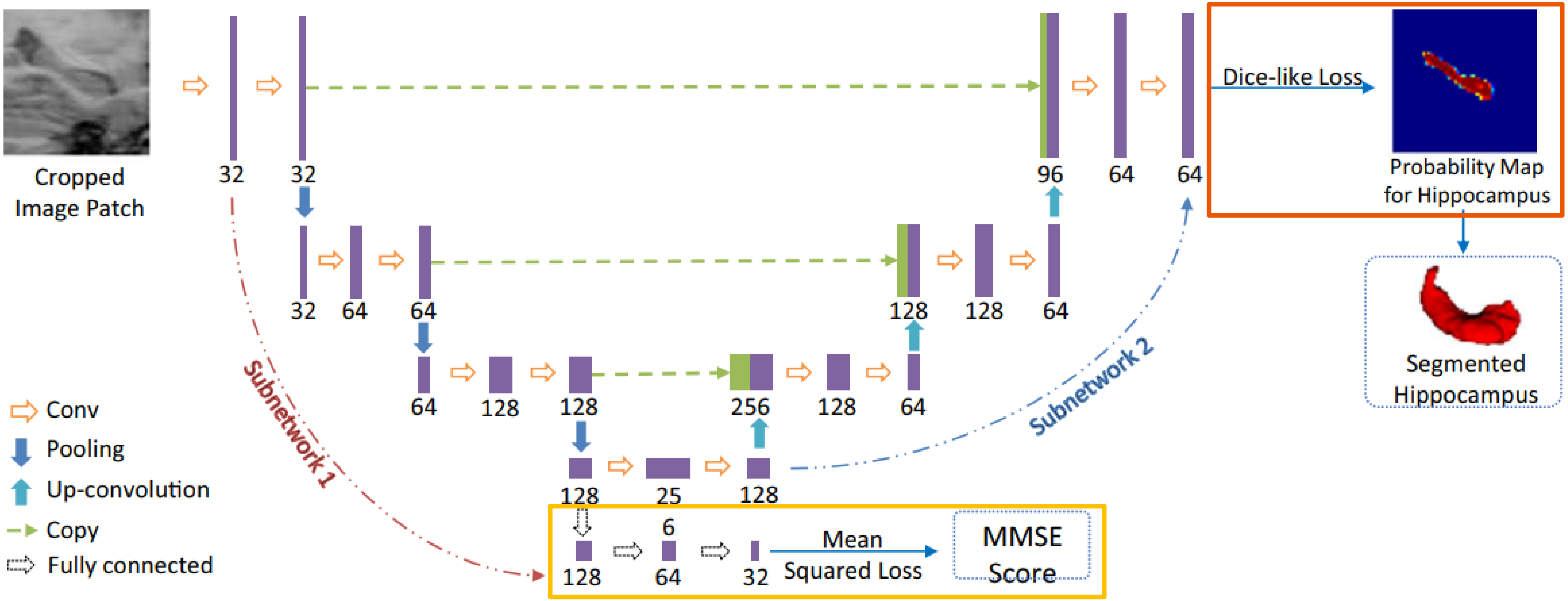


Cao et al. Multimed Tools Appl (2017)

- Make prediction for every pixel → output size = input size

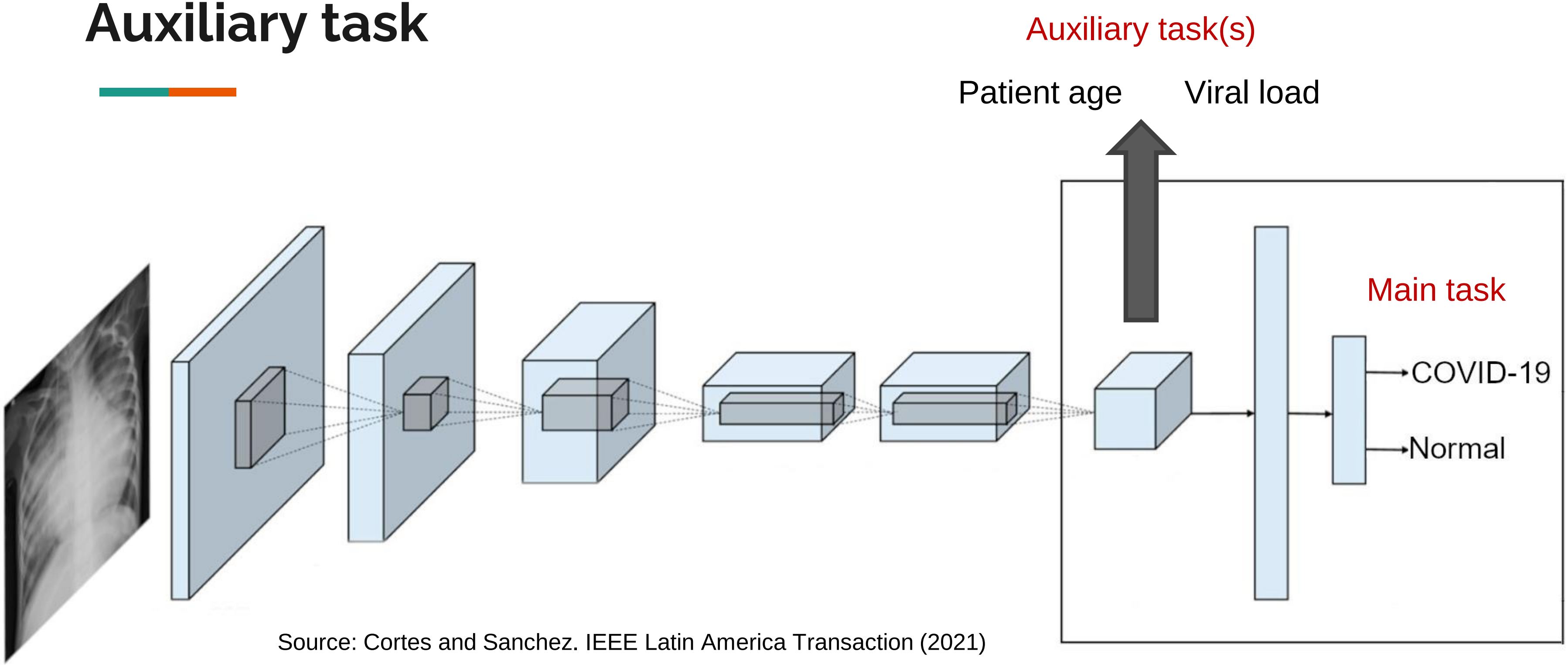
Multitasking

Simultaneous segmentation & classification



- Combine gradients from both tasks

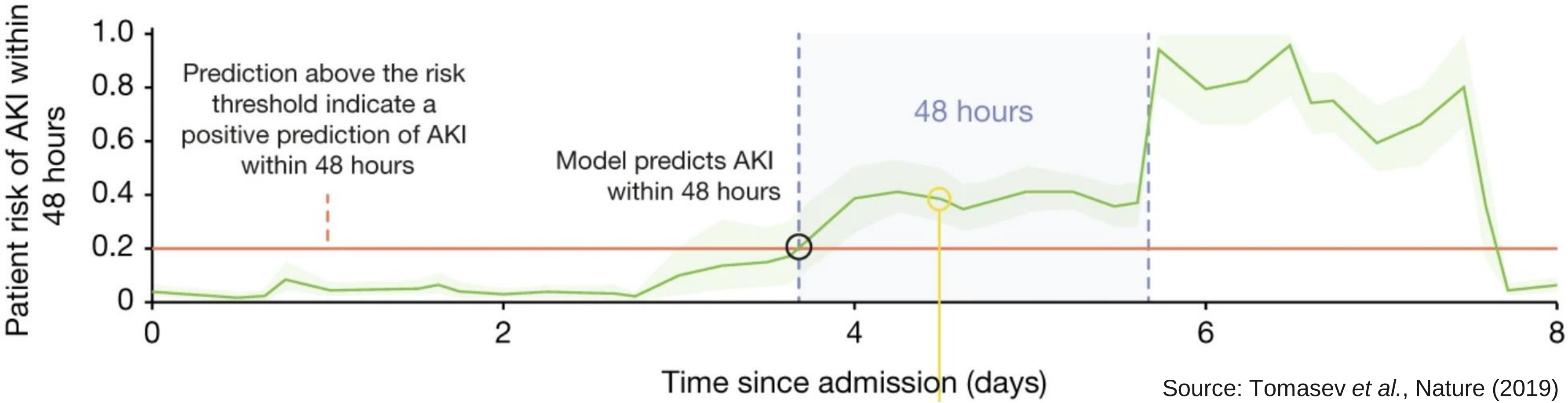
Auxiliary task



Source: Cortes and Sanchez. IEEE Latin America Transaction (2021)

- Encourage the learned representation to include more information

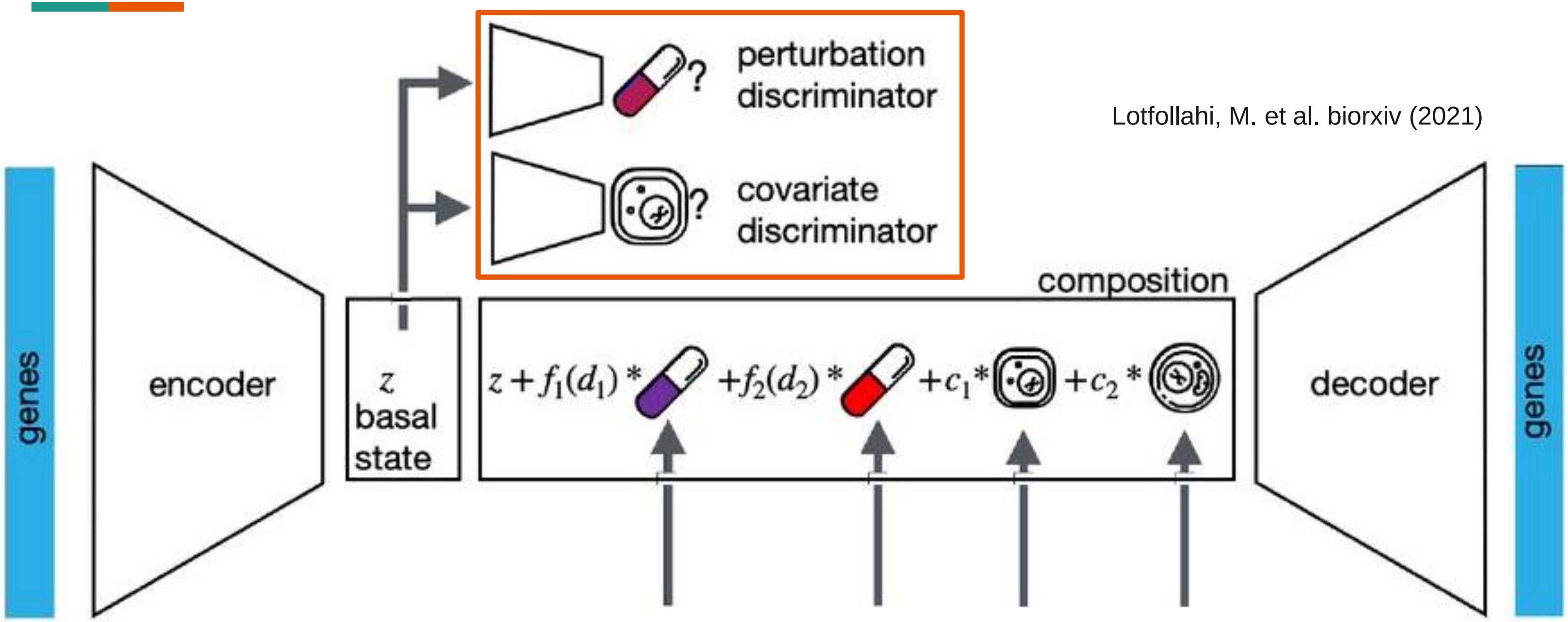
Acute kidney injury prediction



Source: Tomasev *et al.*, Nature (2019)

- **Main task**: Occurrence of acute kidney injury within 48 hours
- **Auxiliary tasks**: Maximal values of 7 key lab tests within 48 hours
  - Provide more feedback on what the model gets wrong

Decoupling

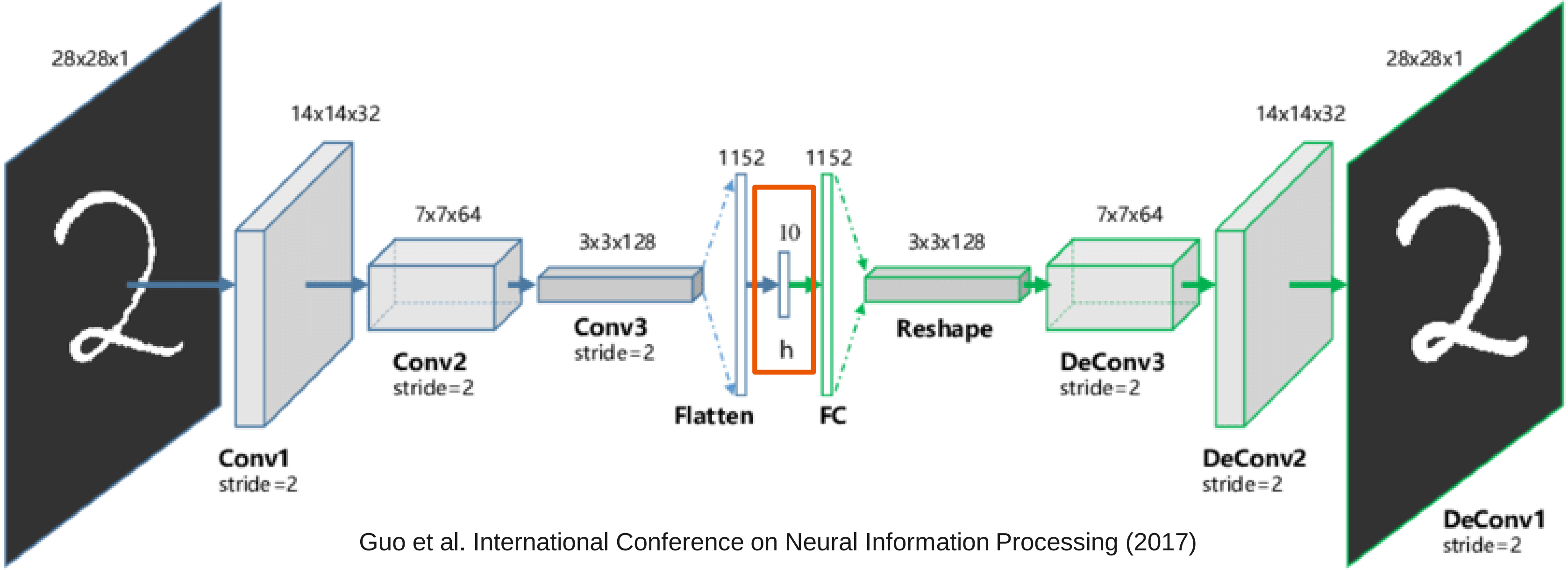


Lotfollahi, M. et al. biorxiv (2021)

- Deconvolute cell basal state from perturbation and covariate
- Update weights in the opposite direction of gradient

Autoencoder

Representation learning via self-reconstruction



Guo et al. International Conference on Neural Information Processing (2017)

- Similar to dimensionality reduction

Denoising autoencoder

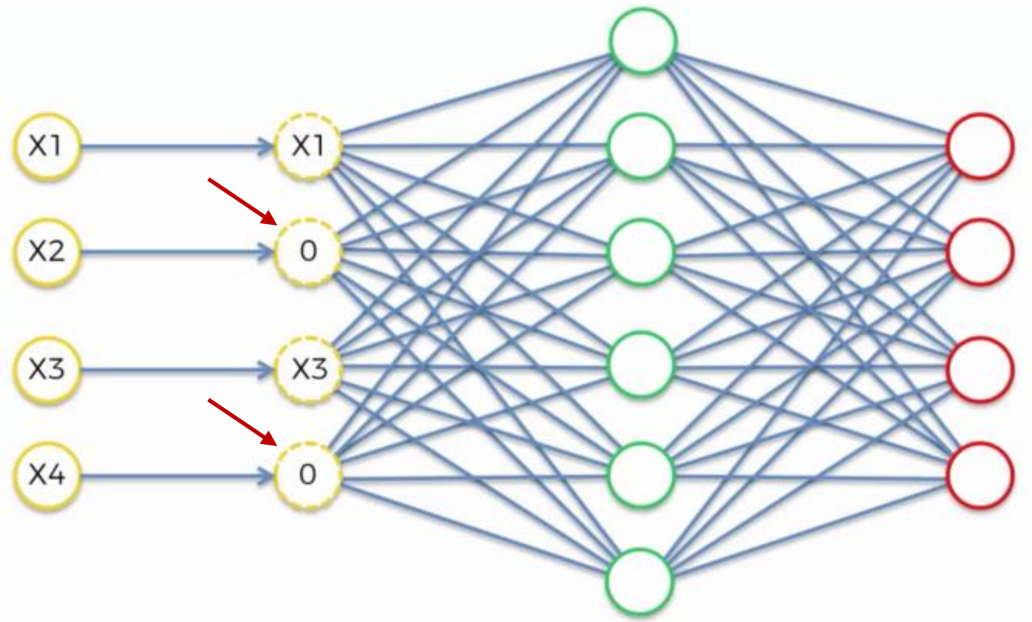


Image from towardsdatascience.com/denoising-autoencoders-explained-dbb82467fc2

- Randomly set some inputs to zero → robust representation

Variational autoencoder (VAE)

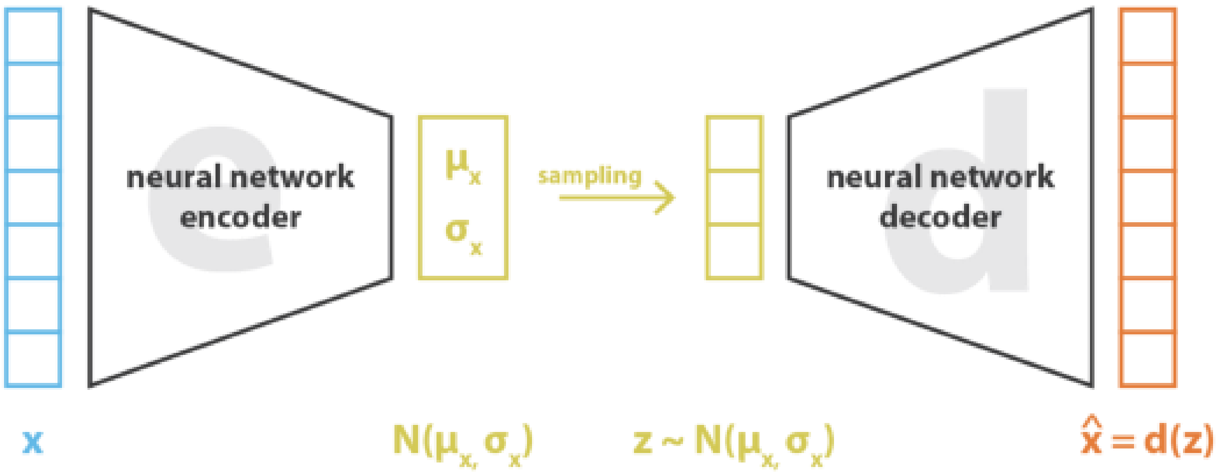
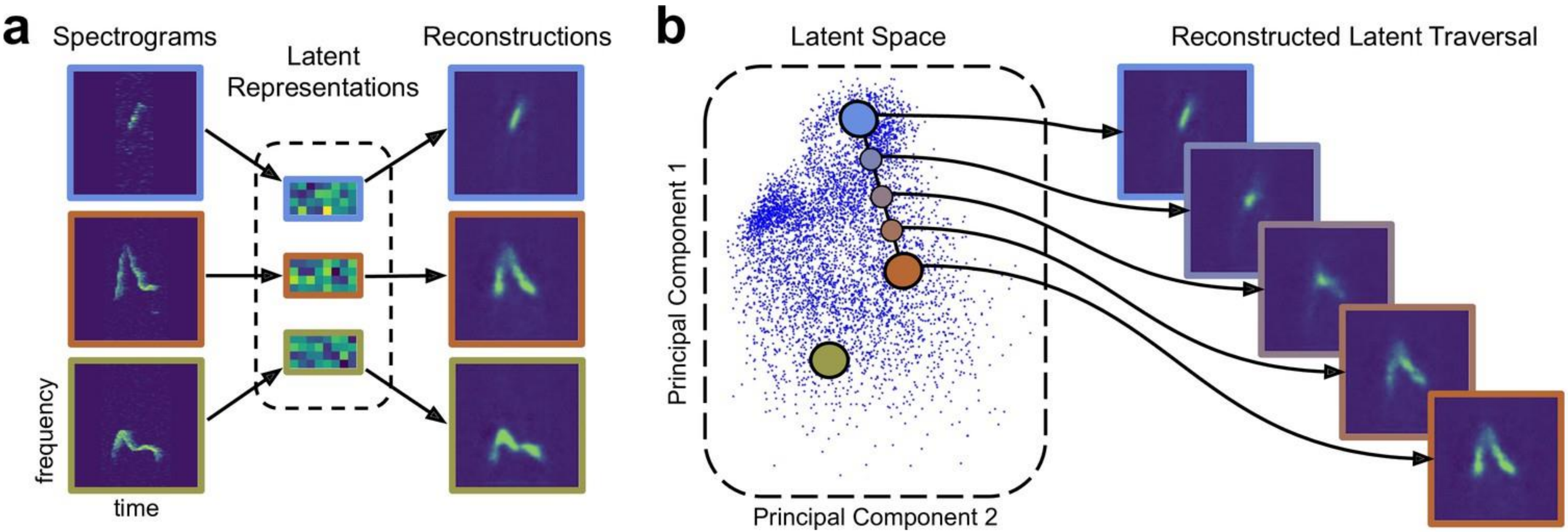


Image from www.jeremyjordan.me/variational-autoencoders/

- Learned representation = parameters for distribution
- Decoder is robust to small changes in the representation
  - Smooth representation space

VAE generates smoother representation space



Doffinet, J. et al. eLife 67855 (2021)

- VAE learn representation distribution, not just individual vectors

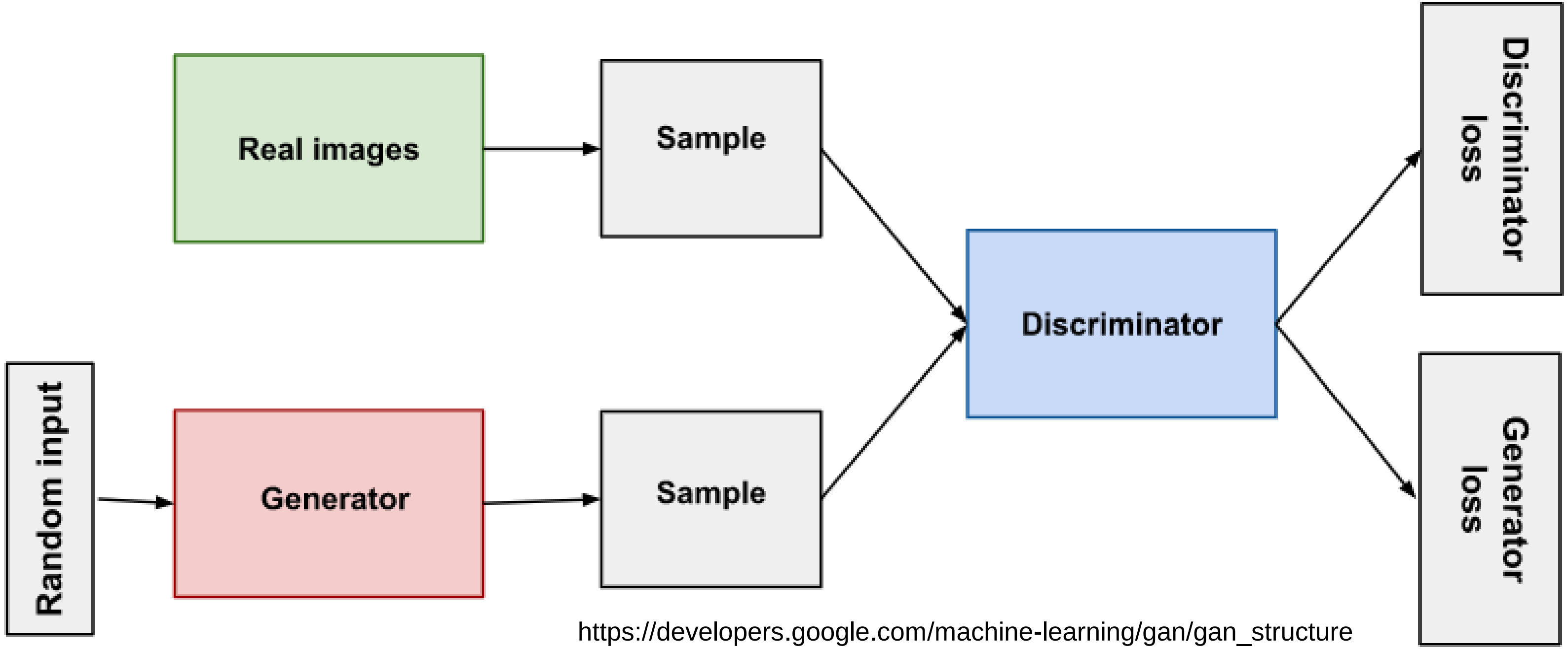
Generative model

Why generative model?

https://developers.google.com/machine-learning/gan/gan_structure

- Models that generate realistic data can tell us about the underlying mechanisms of the system

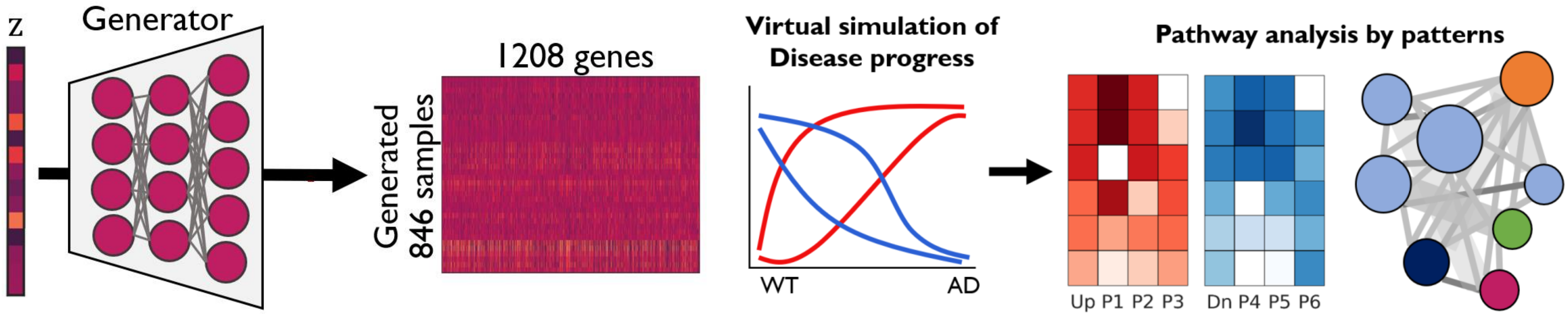
Generative adversarial network (GAN)



https://developers.google.com/machine-learning/gan/gan_structure

- Simultaneous training of generator and discriminator

Knowledge from simulated data

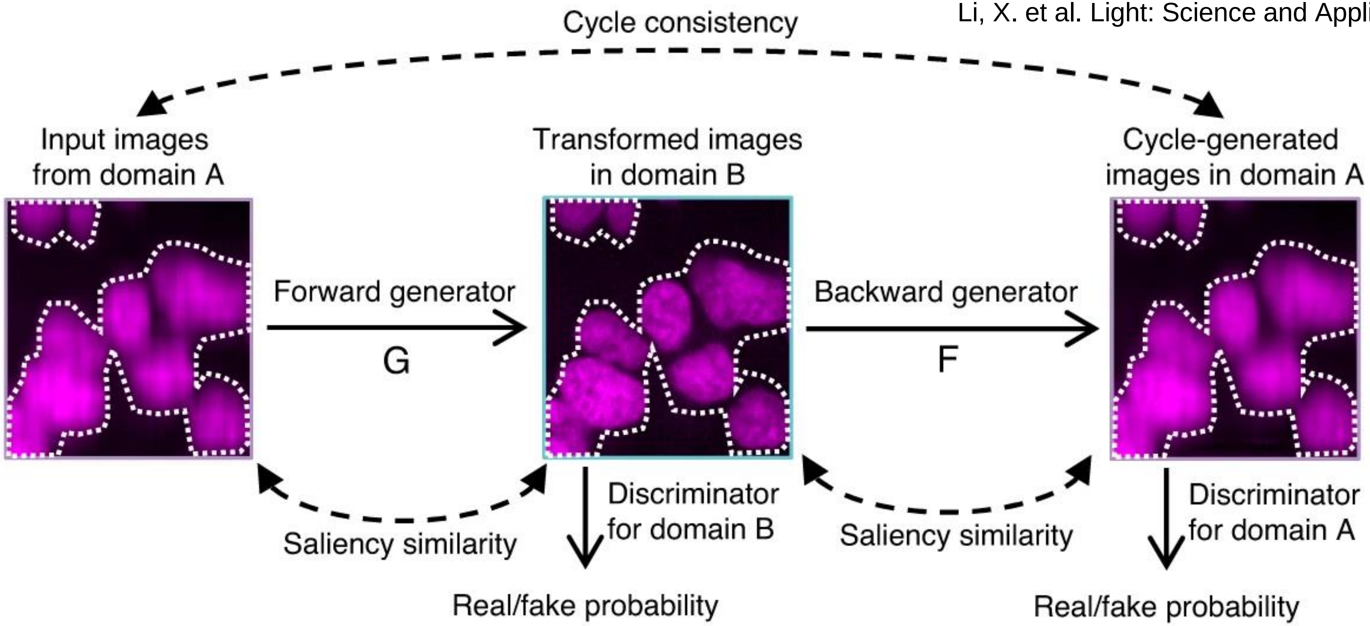


Park, J. et al. PLoS Computational Biology 16:e1008099 (2020)

- Train a generator with data from small-scale experiment
- Simulate time-course gene expression profiles
- Perform usual bioinformatics analyses to infer biological knowledge

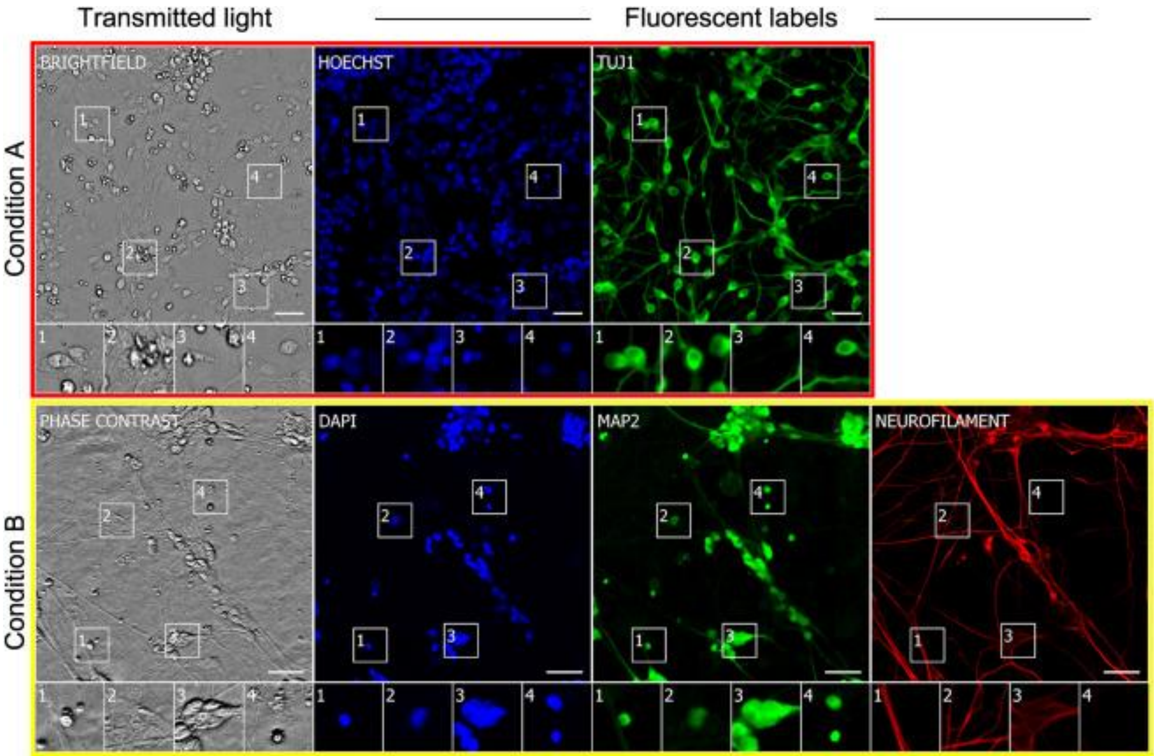
Cycle GAN for transforming image

Li, X. et al. Light: Science and Applications10:44 (2021)

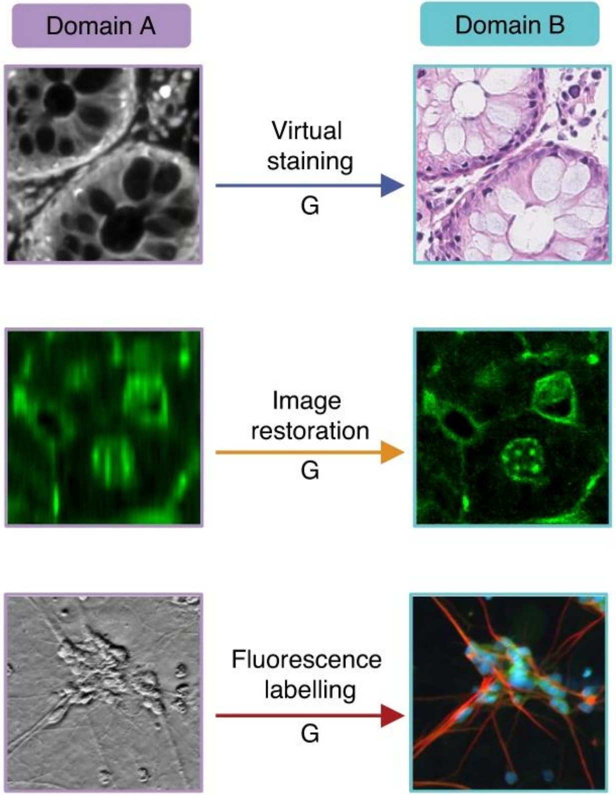


- Generate sharpened image from blurry image and back

Virtual staining



Christiansen, E.M. et al. Cell 173:792-803.e19 (2018)

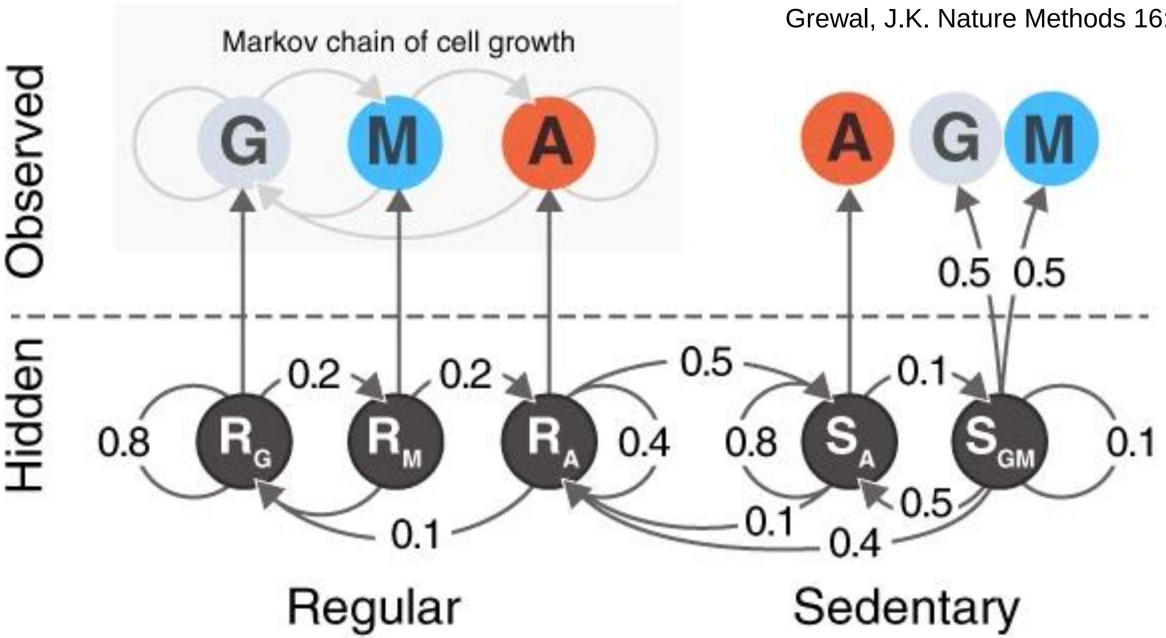


Li, X. et al. Light: Science and Applications10:44 (2021)

Recurrent neural network

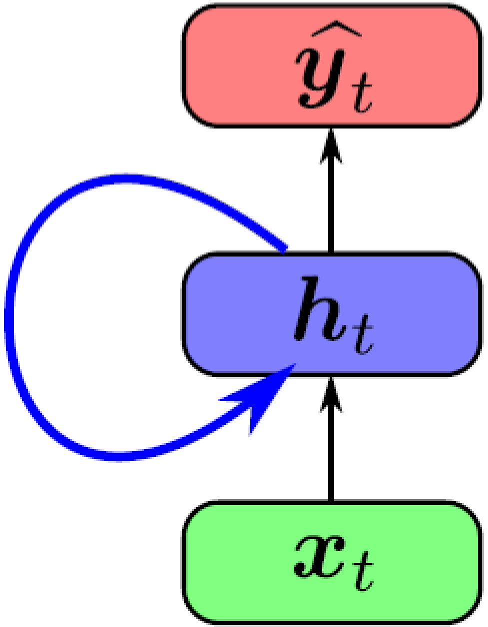
Hidden Markov Model

Grewal, J.K. Nature Methods 16:795-796 (2019)



- Sequence of observations, each generated from a model

Recurrent neural network



Output of a hidden layer

Variable-length input

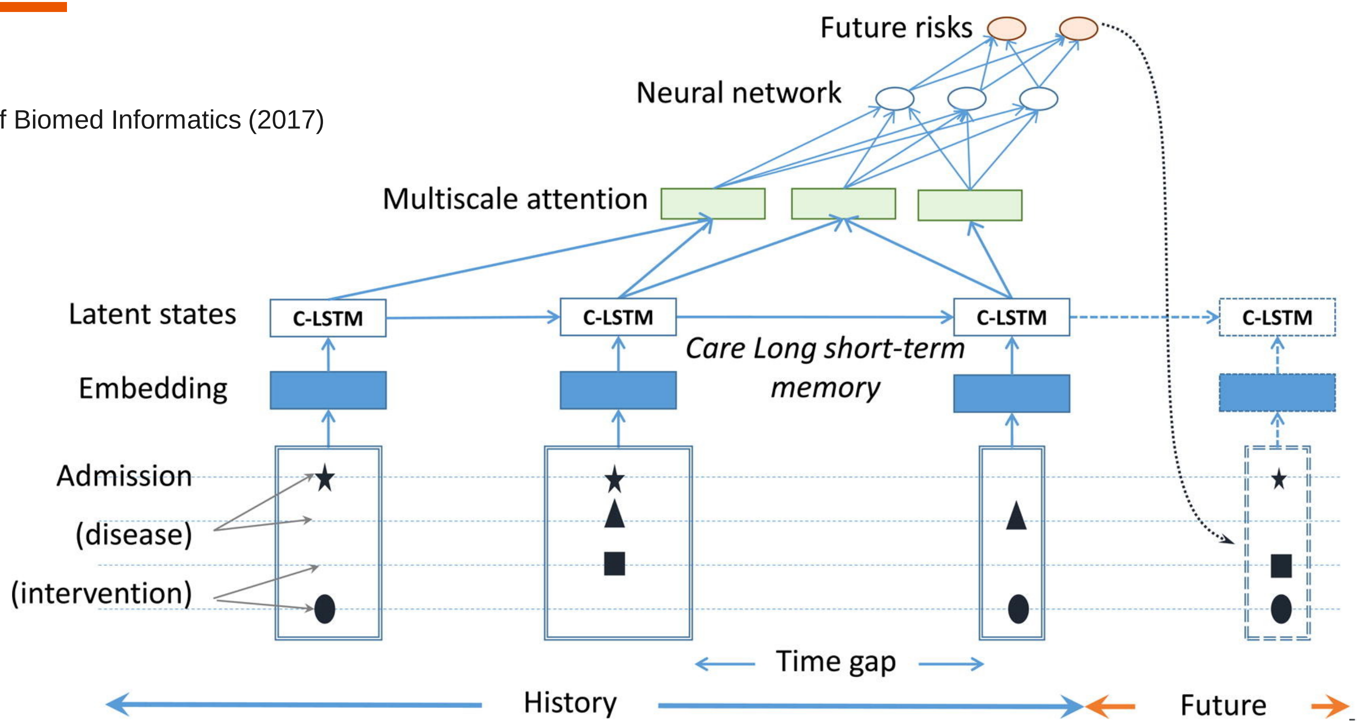
Shared weights!

$$
\begin{aligned}
h_1 &= f(\boldsymbol{u} \cdot x_1 + \boldsymbol{v} \cdot h_0 + c) \\
h_2 &= f(\boldsymbol{u} \cdot x_2 + \boldsymbol{v} \cdot h_1 + c) \\
&\cdots \\
h_t &= f(\boldsymbol{u} \cdot x_t + \boldsymbol{v} \cdot h_{t-1} + c) \\
\hat{y}_t &= \boldsymbol{w} \cdot h_t + b
\end{aligned}
$$

- Reuse a single layer (weights) over time with different input

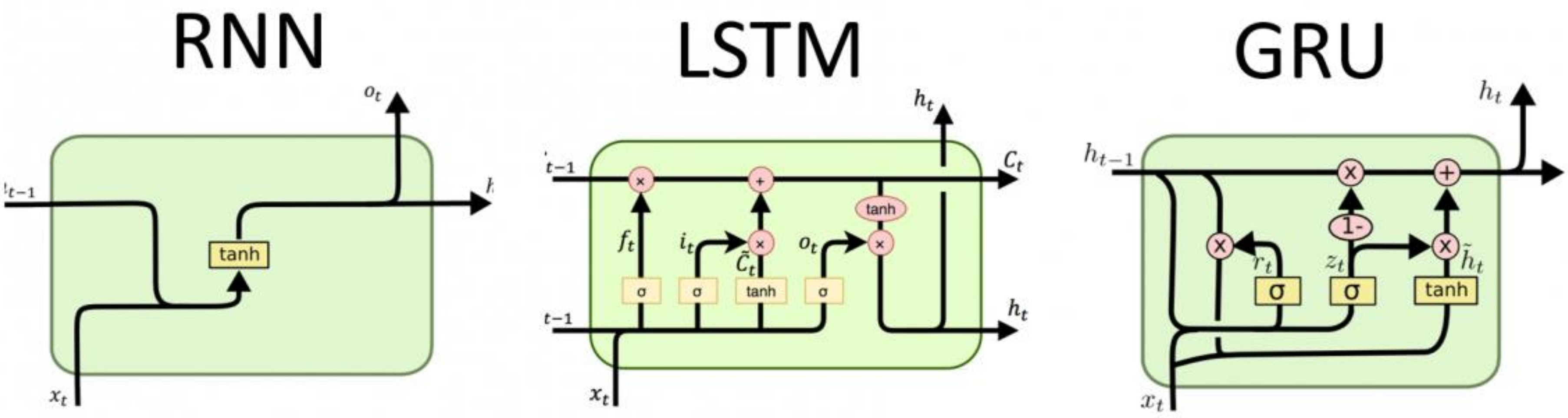
RNN on medical history

Pham et al. J of Biomed Informatics (2017)



- Aggregate information across time to make prediction

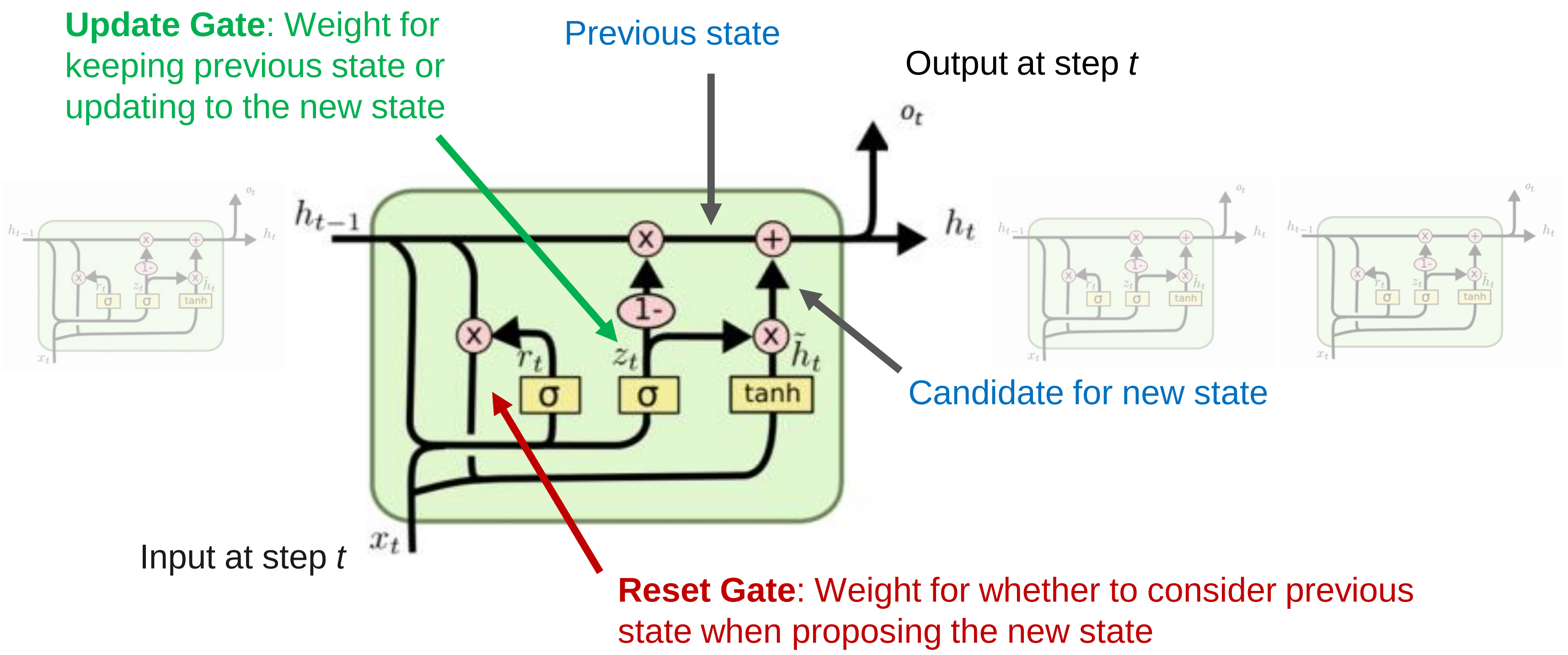
RNN architecture



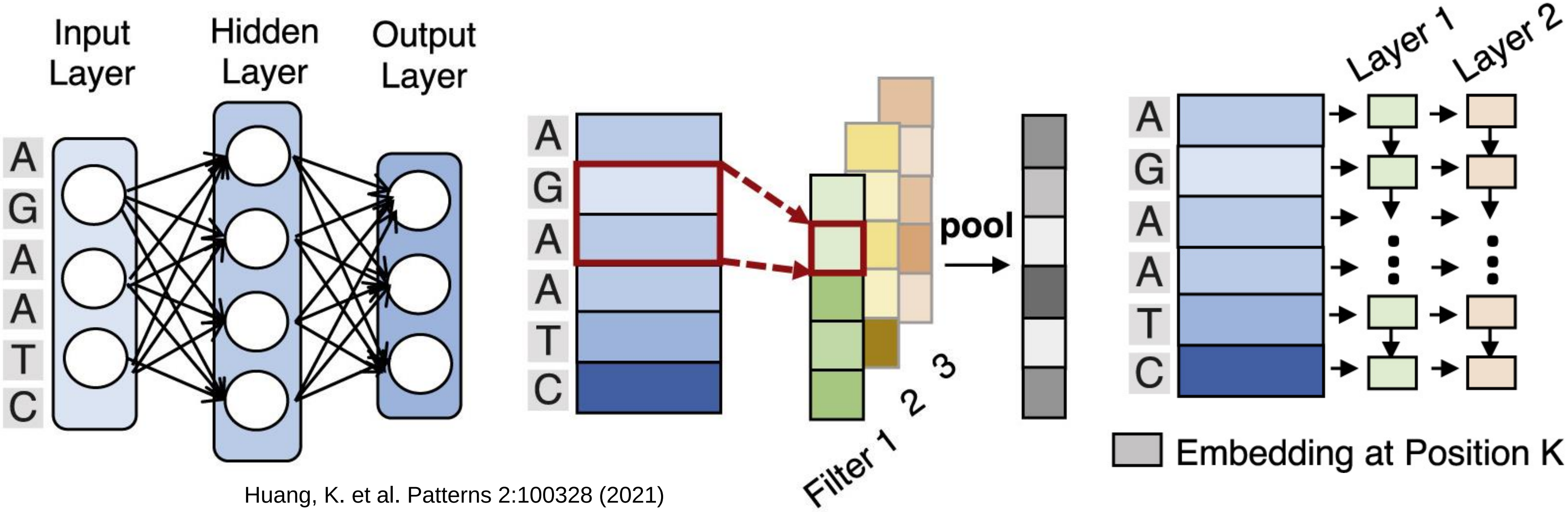
Source: www.linkedin.com/pulse/recurrent-neural-networks-rnn-gated-units-gru-long-short-robin-kalia

- Allow the model to retain / forget information from earlier time points
- Include shortcuts for gradient calculation – similar to ResNet

Gated recurrent unit (GRU)



ANN on DNA sequences

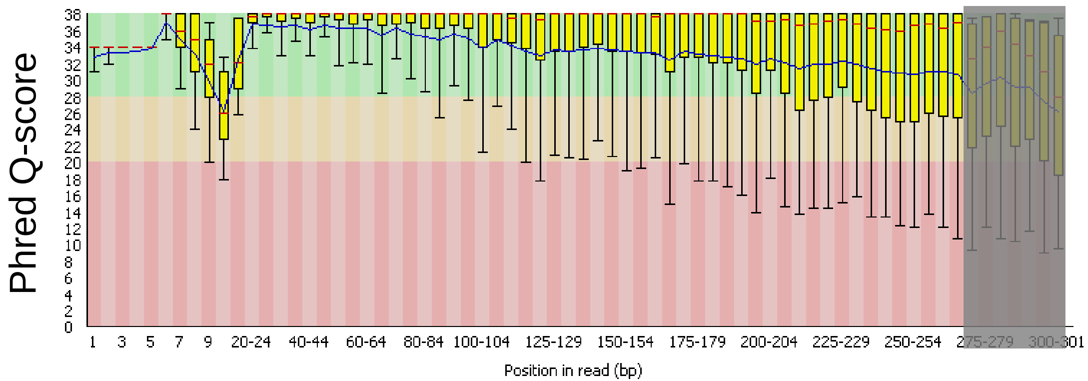


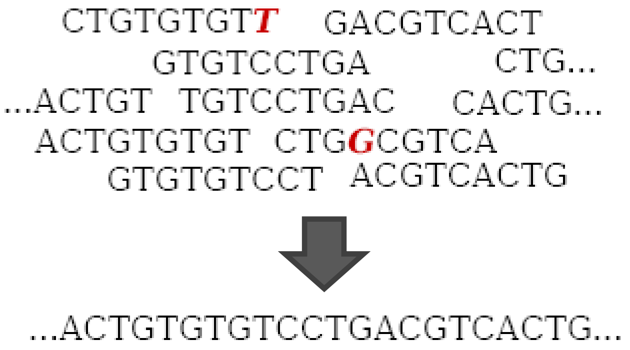
Huang, K. et al. Patterns 2:100328 (2021)

- Choosing the "right" model depends on the interpretation of the task and the underlying mechanisms – require domain knowledge

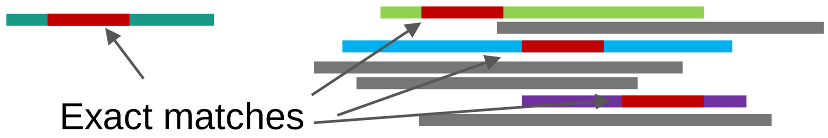
Enhanced bioinformatics

Bioinformatics relies on statistics and scoring





Chandra Varma Bogaraju, S. Int J Embed Syst 9:74 (2017)



Grow alignment

- Instead of applying hand-made scoring + cutoffs, ANN model can be trained to predict the outcome directly

Enhancing bioinformatics with deep learning

Published: 24 September 2018

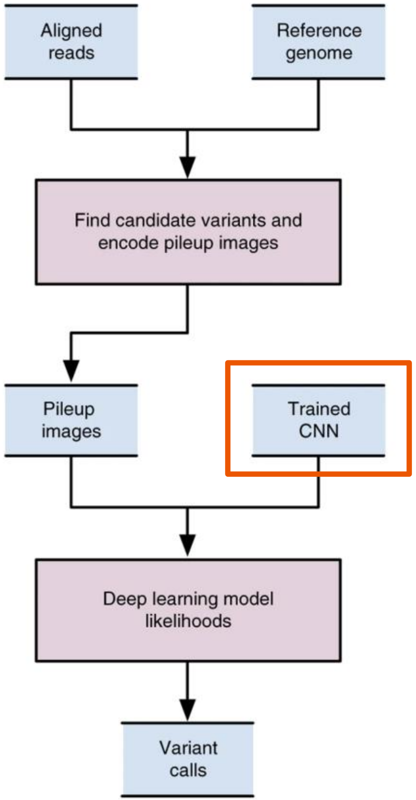
A universal SNP and small-indel variant caller using deep neural networks

Published: 27 July 2015

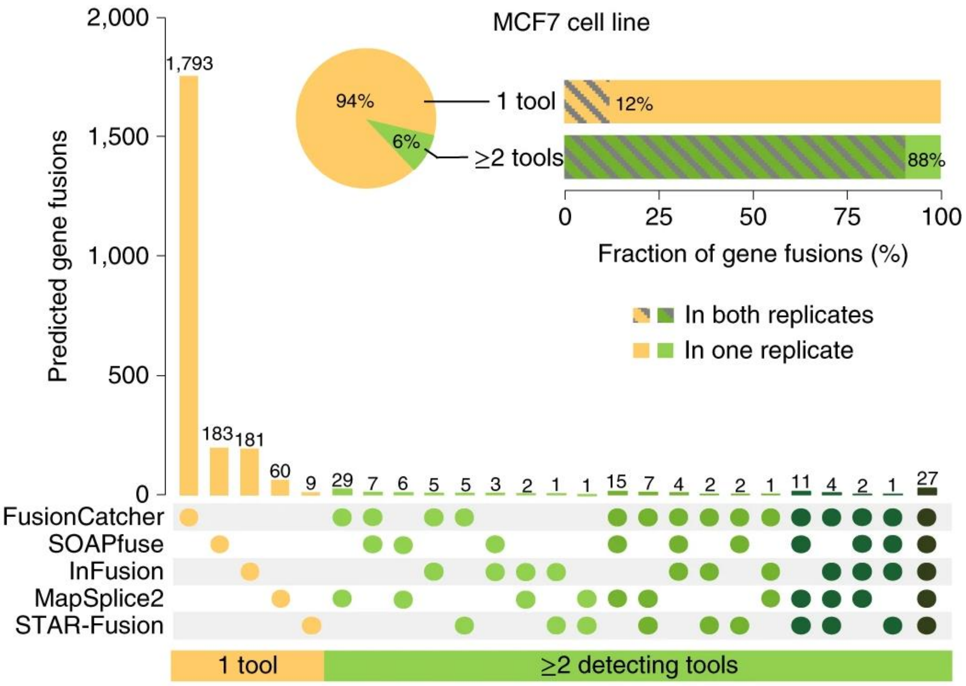
Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

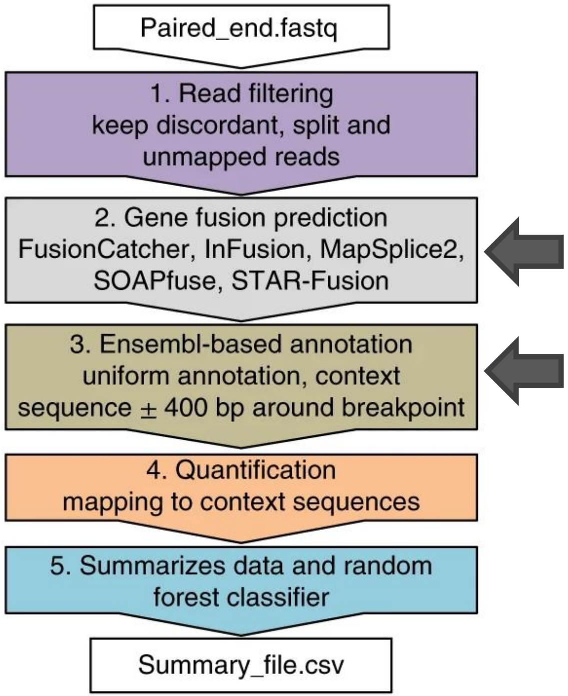
Article | Open Access | Published: 19 May 2022

Prediction of protein–protein interaction using graph neural networks



Aggregate scores from multiple tools





Weber, D. et al. Nature Biotechnology 40:1276-1284 (2022)

Summary

- Deep learning = machine learning with artificial neural network
- Powerful but requires a lot of data and supervision

- Representation learning
- Encoder-Decoder view of ANN
- Autoencoder

- Convolutional and recurrent architecture designs

- Generative models

- DL-enhanced bioinformatics

Any question?

- See you next week on November 29th 9-10:30am