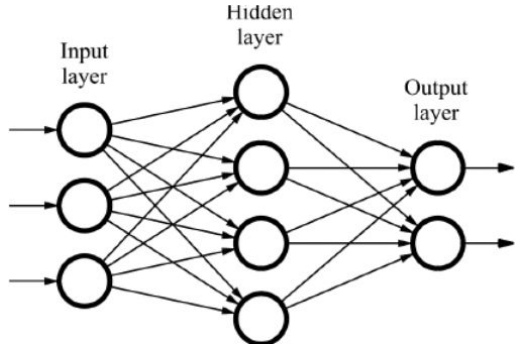
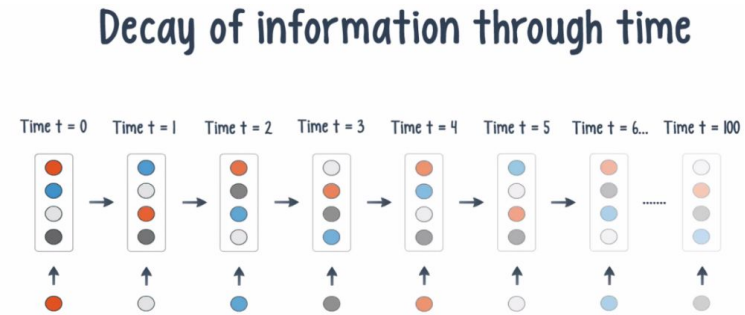
Recent developments in deep learning applied to protein structure prediction

Yaqin Si, Dillon Lloyd, Jorden Rabasco

1. Introduction

- Deep Neural Network (DNN) models have had a large impact on CASP experiments
- Able to approximate any arbitrary continuous function
- Training is achieved through random initialization of network parameters
- Generally organized in multiple layers
- Early problem → Vanishing gradient



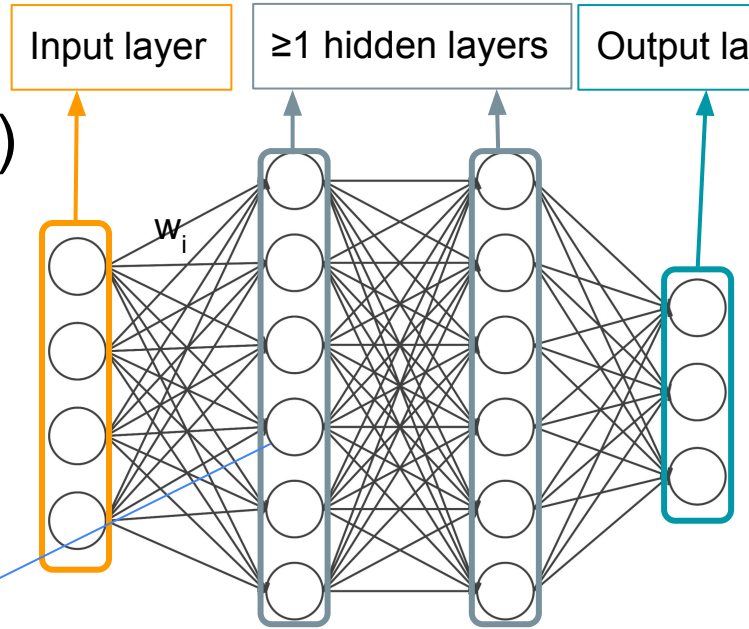


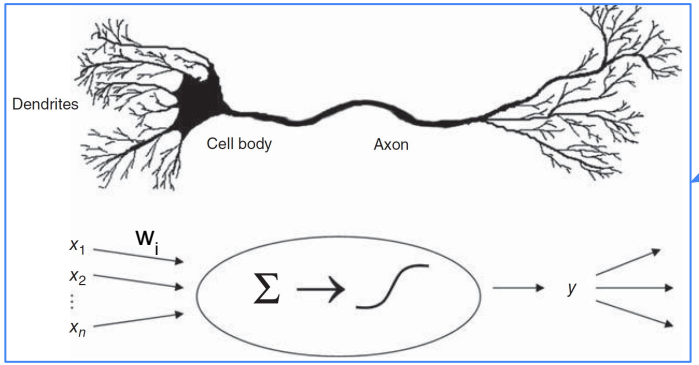
2. Methods

2.1 (Artificial) Neural Networks (NN)

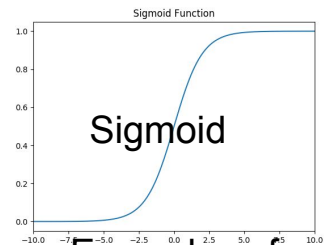
- Universal function approximator
- Structure and parameters
  - Deep Neural Network (DNN)

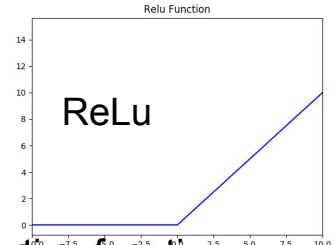
a. An example of a fully connected Neural Network





b. A real neuron structure and an artificial neuron model

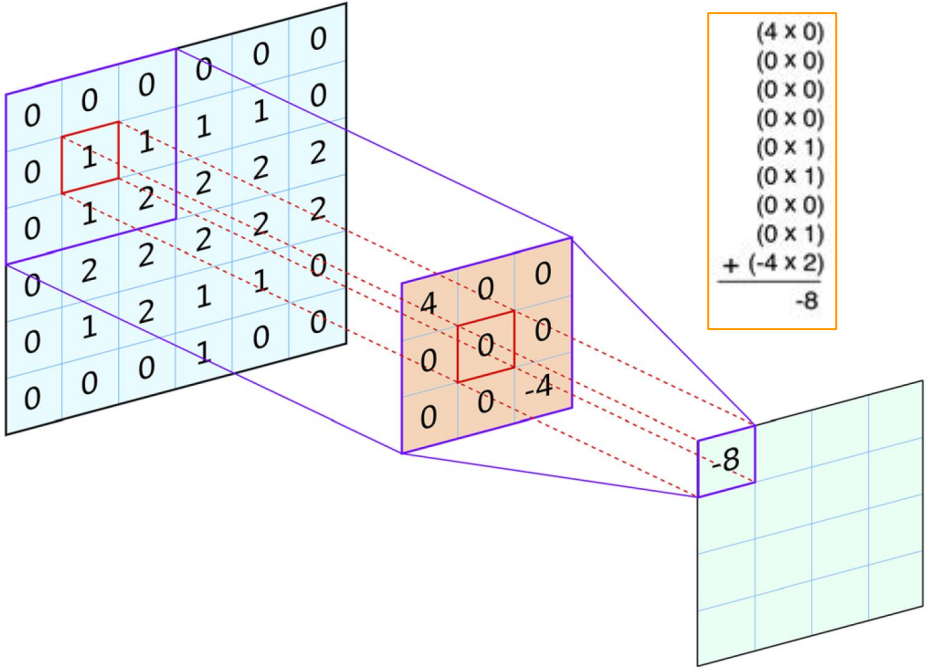




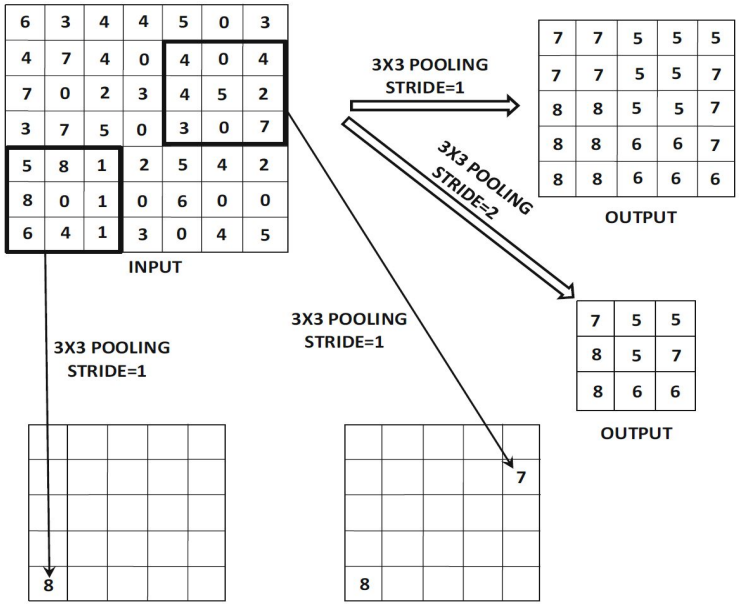
c. Examples of activation functions

(1) Layer structures

Sparse connection / local feature



d. Convolutional layer -- a 2D filter (orange) applied to input layer to obtain the value for an output layer.

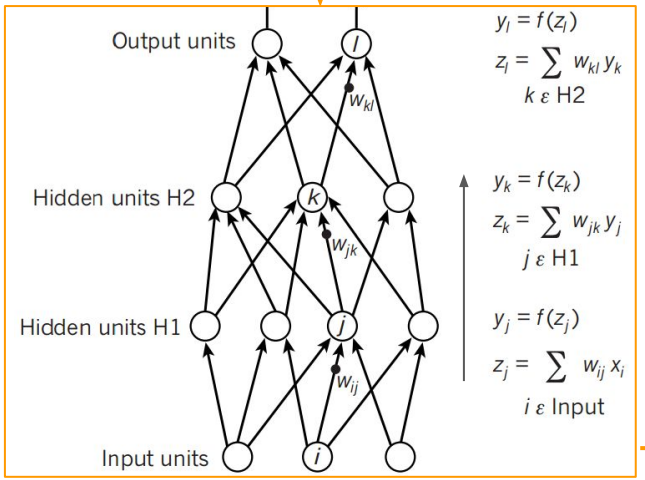


e. Max pooling layer

(2) Parameter training

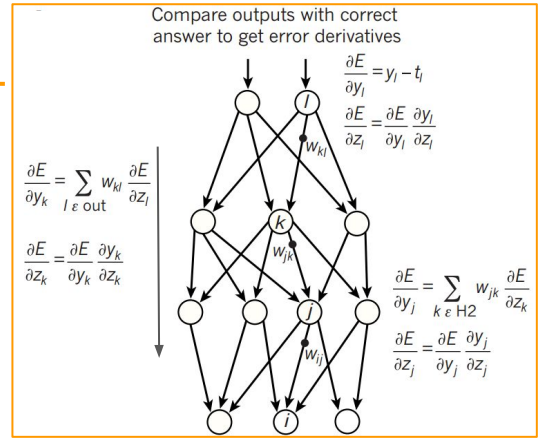
- Supervised training--Training dataset with paired inputs and output

1.Weight Initialization(adjustment)



2. Forward pass to produce model output

3.Loss function→ errors



4.Backpropagation to partition error to each parameter

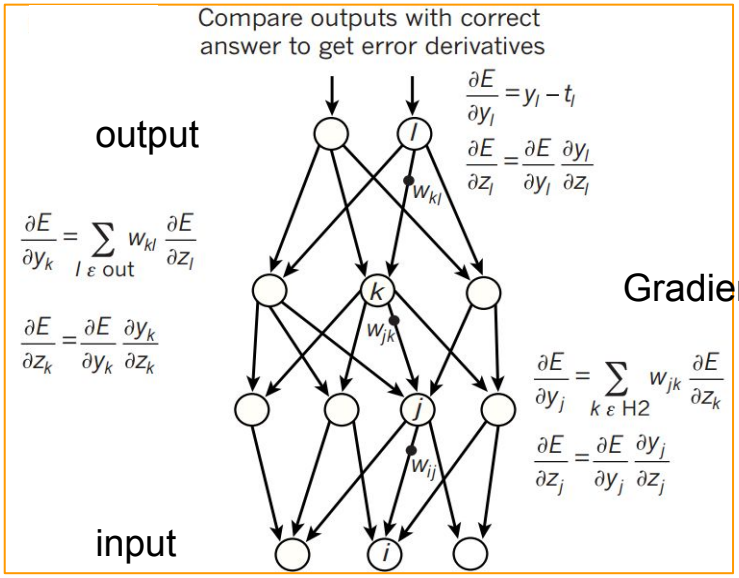
Gradient Derivative

f. Training process

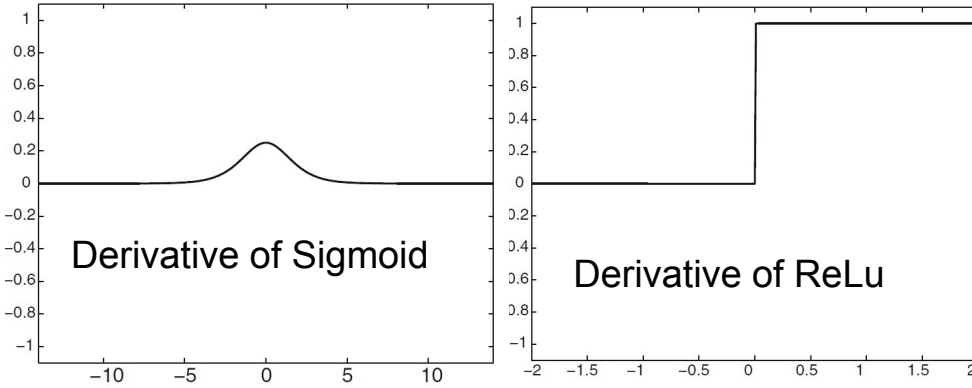
2.2 Model improvements

(1) “Deeper” layers

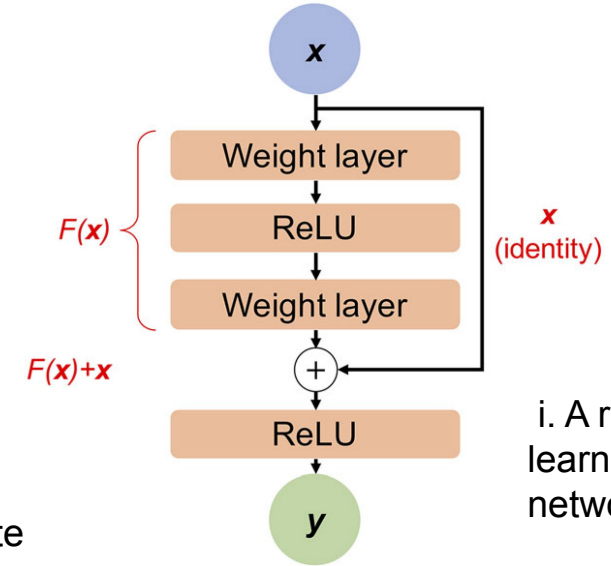
Difficult to train --Gradient vanishing



g. Backpropagation to partition error to each paramete



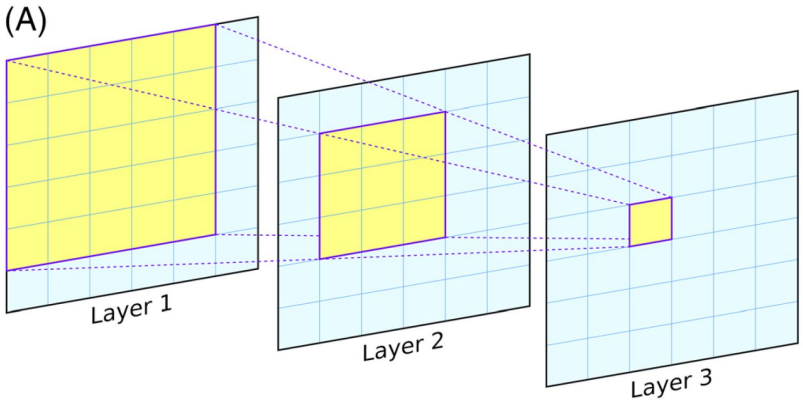
h. Derivative function of Sigmoid/ReLu activation function

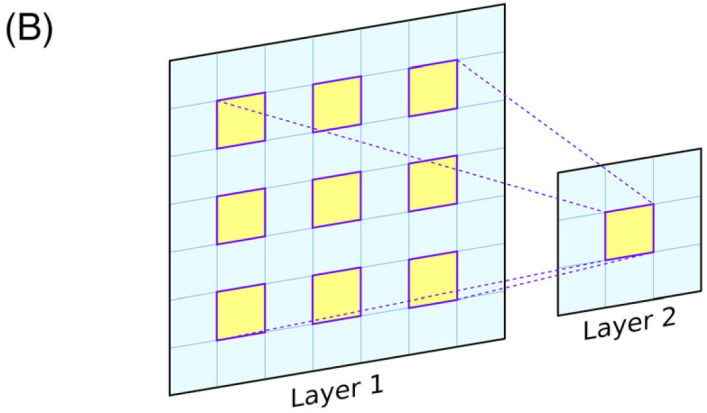


i. A residual block for learning very deep neural networks

(2) Increase Receptive Field of CNN

The input features can be “seen” from one output neuron





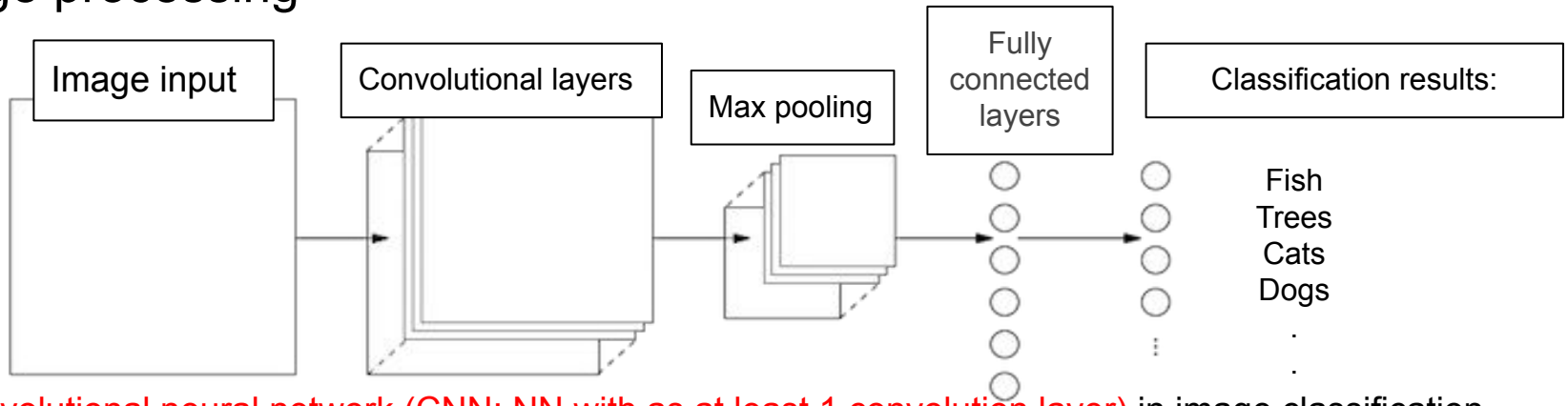
j. Receptive field and dilation rate

A. 3 by 3 convolutional filter
9 parameters for each layer.
3 by 3 receptive field layer 2, 5 by 5 layer 1

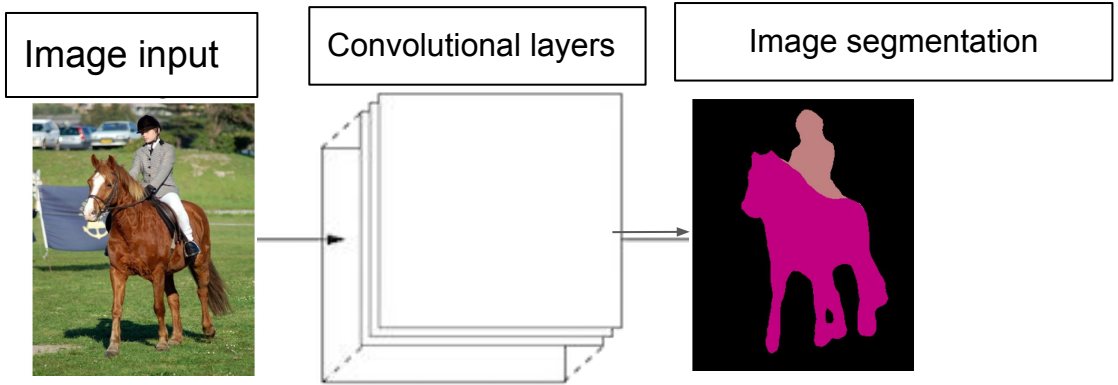
B. 3 by 3 convolutional filter with dilation rate of 2
9 parameters for each layer
5 by 5 receptive field

2.3 Application of Neural Networks

(1) Image processing

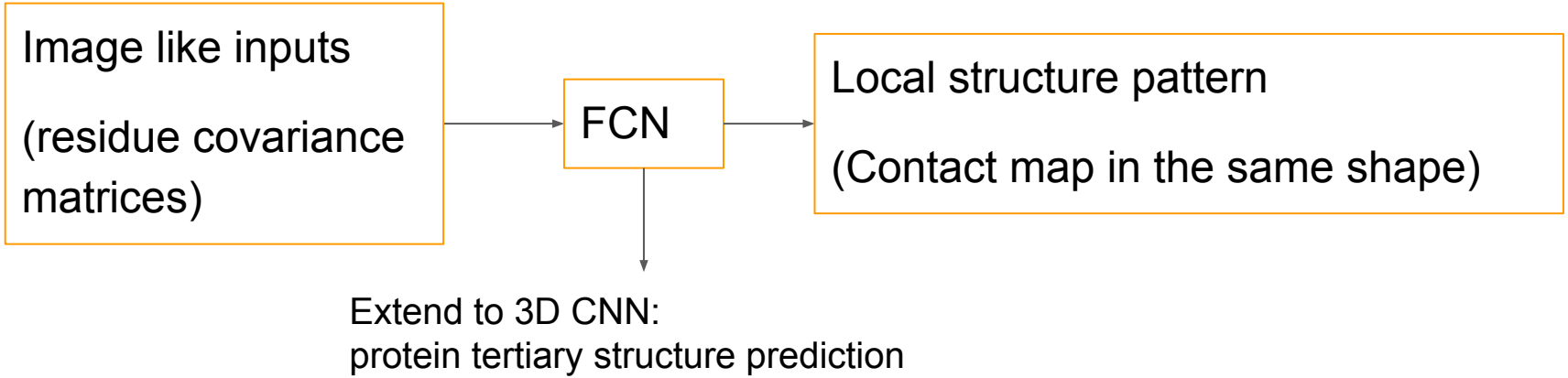


g. Convolutional neural network (CNN: NN with as at least 1 convolution layer) in image classification



h. Fully convolutional networks (FCN) in image segmentation

(2) Contact prediction



More effective than global statistical models in contact prediction:

Use only local subsets of data to recognize local structure patterns

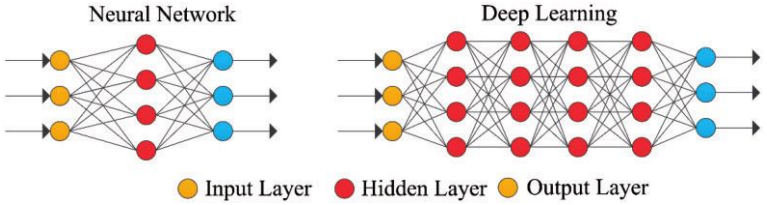
Increase receptive fields (use more covariance data)

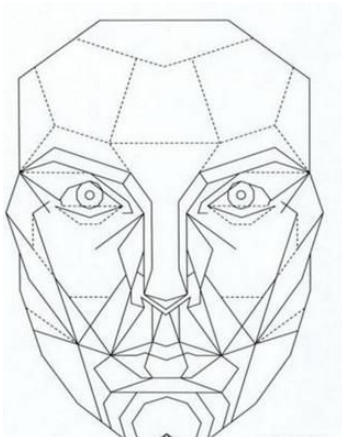
When Receptive Files reach 15, little or no gain in model precision (contact prediction)

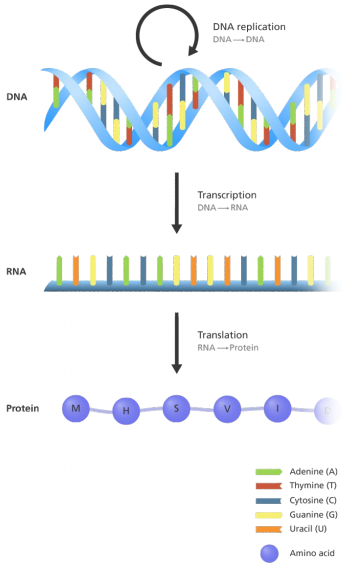
3.WHY IS DEEP LEARNING EFFECTIVE?

3.1 Learning hierarchical representations

- Each layer of a Deep Neural Network composes and interprets the outputs from the previous layer.
- Allows for recognition of increasingly complex patterns
- DNNs not effective at modeling unstructured data
- Very effective in a very narrow range of problems and shouldn't be used as a one size fits all approach.

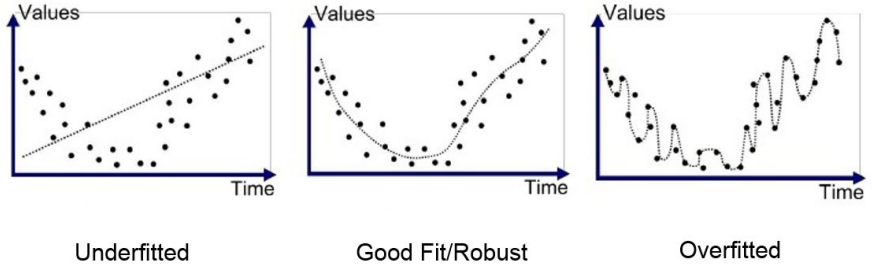


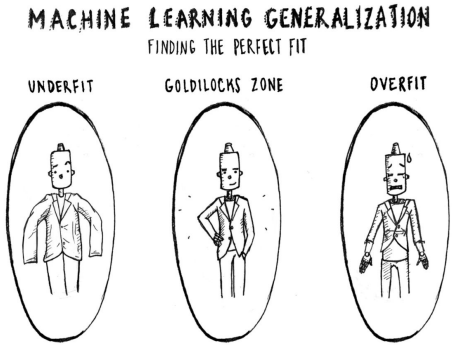




3.2 Deep learning as a neighborhood density estimation method

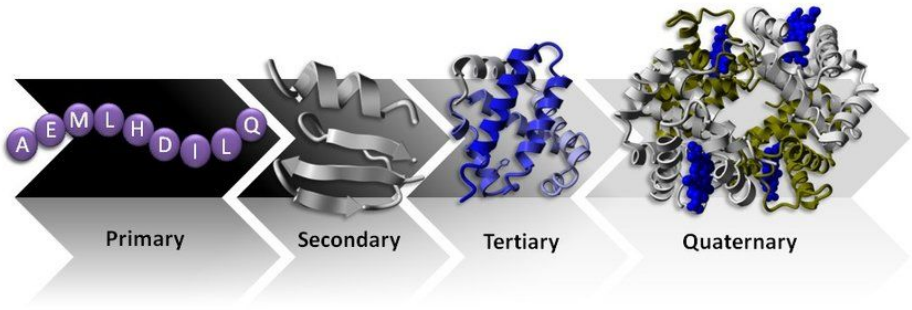
- Neural Network can act as a highly sophisticated look-up table
- Regularization techniques used to avoid overfitting
- Collection of more data or reducing the network complexity can also avoid overfitting scenarios
- Compressing data to the point where the original input is lost is called Underfitting

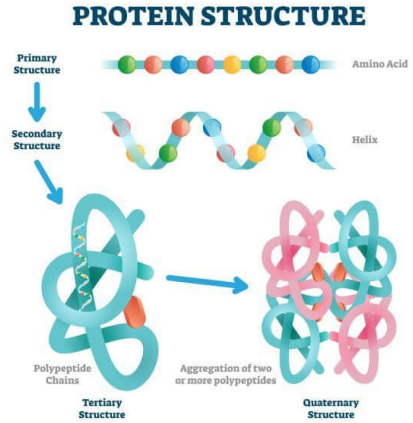




3.2 Deep learning as a neighborhood density estimation method

- Models employ features derived from input sequences
- Sequence features are at a higher organizational level than the raw sequences themselves
- Learn about statistical features of the family to which the target sequence belongs
- Can only learn structure of an average member of the family
- Can lead to inaccurate predictions and biases





Robustness, Pitfalls and Conclusions

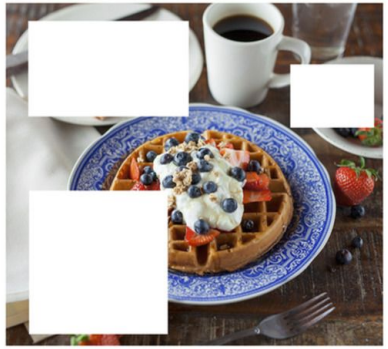
3.3 Robustness to missing or noisy inputs

- One of the many advantages of Deep Neural Networks is how robust they are to missing data
- Since missing data is a huge aspect of protein prediction, properly training a model to account for this problem is extremely important.
- As shown in the picture in the following slide, DNNs can perform admirably when data is missing
- For protein prediction, sequence and structure missingness is common and thus needs to be worked around

(A)

Concept	Probability
no person	0.997
breakfast	0.981
food	0.979
delicious	0.977
homemade	0.973
lunch	0.956
traditional	0.937
dinner	0.936
brunch	0.930
plate	0.930

(B)



Concept	Probability
no person	0.993
food	0.986
breakfast	0.975
plate	0.962
dawn	0.957
delicious	0.933
traditional	0.932
milk	0.927
lunch	0.913
table	0.913

(C)

Concept	Probability
no person	0.991
coffee	0.987
cup	0.971
paper	0.962
empty	0.957
dawn	0.954
breakfast	0.944
tea	0.939
food	0.934
table	0.931

(D)

Concept	Probability
no person	0.992
coffee	0.982
paper	0.964
dawn	0.934
cup	0.926
breakfast	0.919
business	0.901
milk	0.870
retro	0.852
indoors	0.846

3.4 Potential pitfalls

- Image data and biological data are incredibly different, where biological databases have less data points so less training data.
- This lack of data points can lead to overtraining on the training set and makes the model less applicable to other applications
- One way around this is to make sure that proteins in the validation set do not share similar 3D structure to those in the testing and training sets.
- Knowing how to make this split based on sequence and structure similarity is important to correctly assessing models

3.4 Potential pitfalls

- However, proteins with the same structure can have different sequences thus it can be difficult to differentiate
- In these cases, the model will just repeat what it has done in training, and not make an actual new prediction
- Overall, a split on sequence similarity is not sufficient and thus an area of improvement would be to figure out the optimal way to split data

4. Conclusions and Outlook

- Recurrent NN architectures that map sequences of data to other sequences are used and effective in prediction protein secondary structure and is now being used to predict protein interactions and tertiary structure
- Deep Learning methods are being popularized due to them being able to account for missing data, raw feature and ability to make predictions from data not in the training or testing set
- Future CASPs will largely focus on Deep Learning due to the outlined advantages

Questions?