Deep Learning for Protein Structure Prediction

Badri Adhikari
adhikarib @ umsl.edu

Assistant Professor of CS
Department of Mathematics & Computer Science
University of Missouri-St. Louis

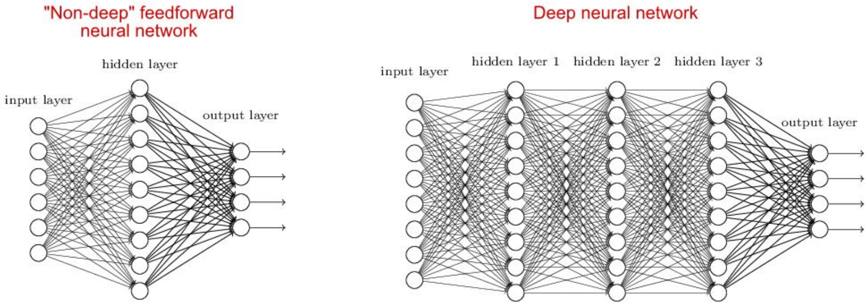
Topics

- Deep Learning, Trends, and Limitations
- DL Tool Chain
- DL for Protein Contact Prediction

Deep Learning (DL) - term coined in 2000

- DL is a subfield of ML

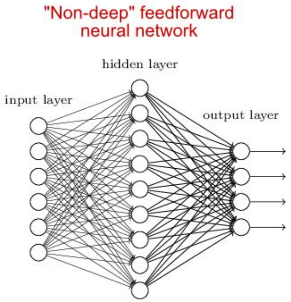
- DL is Large Neural Networks

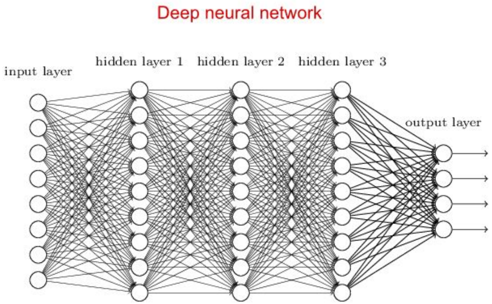


Deep Learning (DL) - term coined in 2000

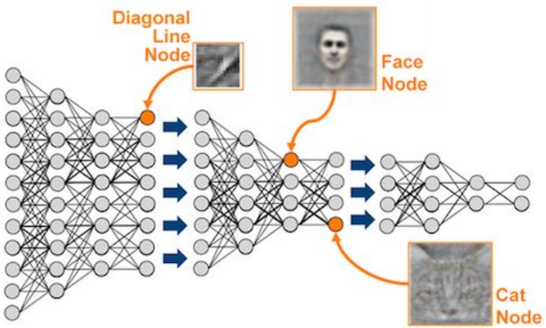
- DL is a subfield of ML

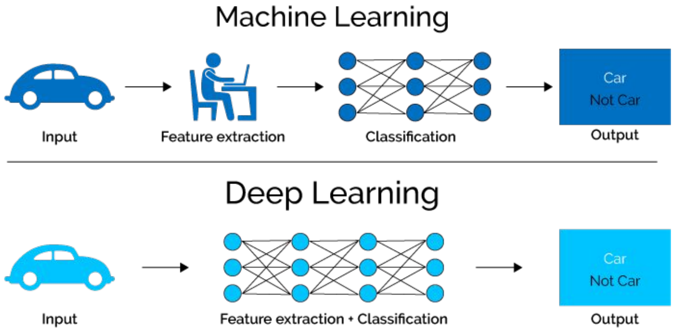
- DL is Large Neural Networks





- DL is Hierarchical Feature Learning





A Hidden Layer

x AND Y

x	y	f-and(x,y)
0	0	0
0	1	0
1	0	0
1	1	1

x OR y

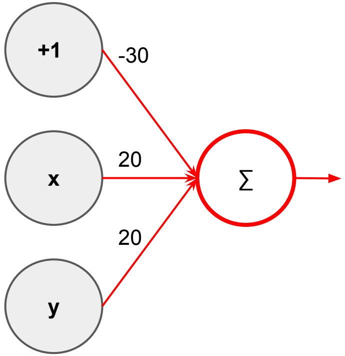
x	y	f-or(x,y)
0	0	0
0	1	1
1	0	1
1	1	1

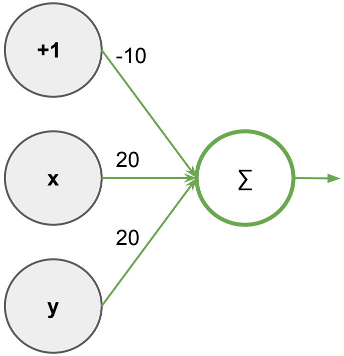
(!x) AND (!y)

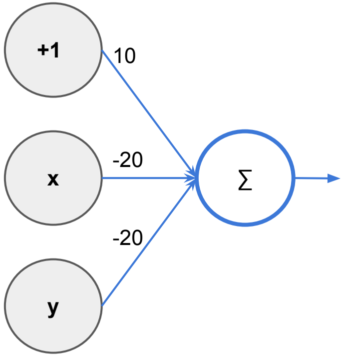
x	y	f-rev-and(x,y)
0	0	1
0	1	0
1	0	0
1	1	0

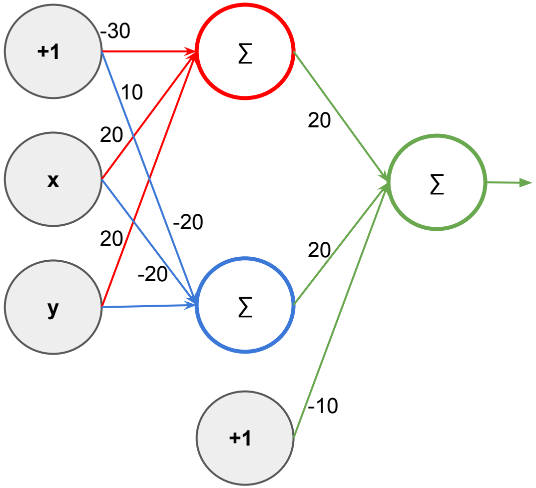
x XNOR y

x	y	f-xnor(x,y)
0	0	1
0	1	0
1	0	0
1	1	1









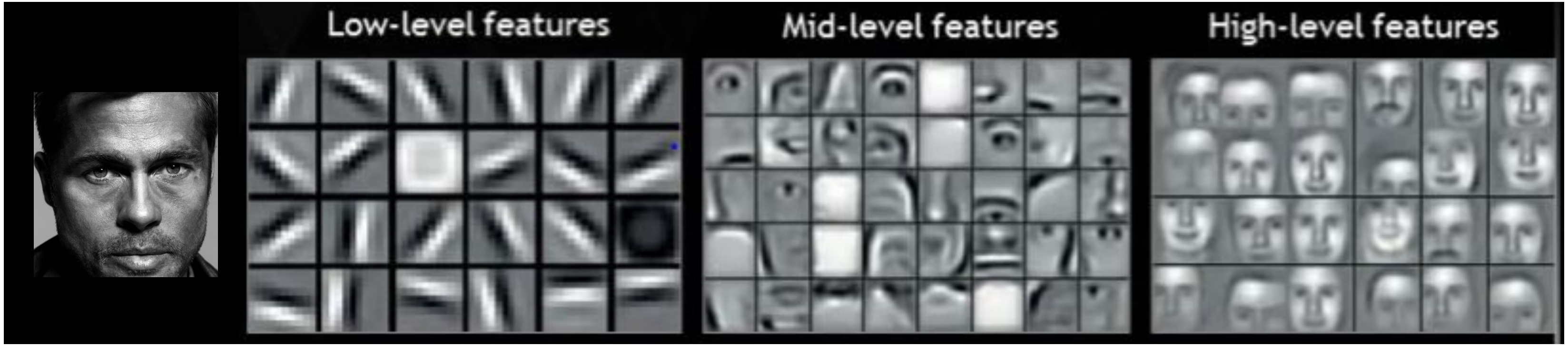
XNOR = (a AND b) OR (!a AND !b)

Many Hidden Layers

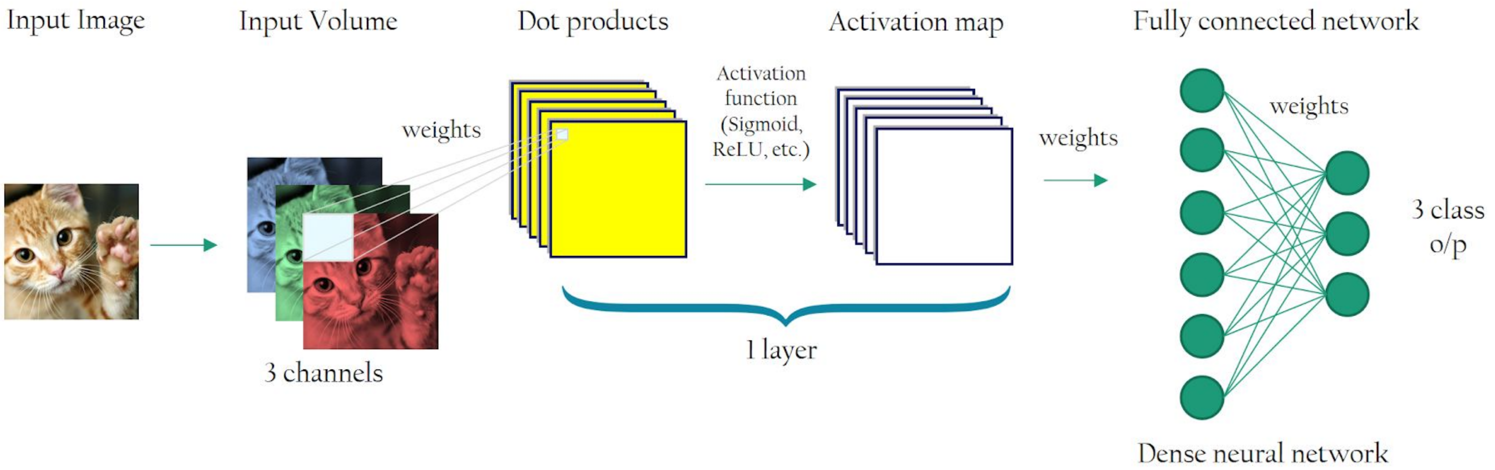
- A feed-forward network with a single hidden layer can approximate (any) continuous functions
  - Universal approximation theorem
  - ability to represent does not mean ability to learn

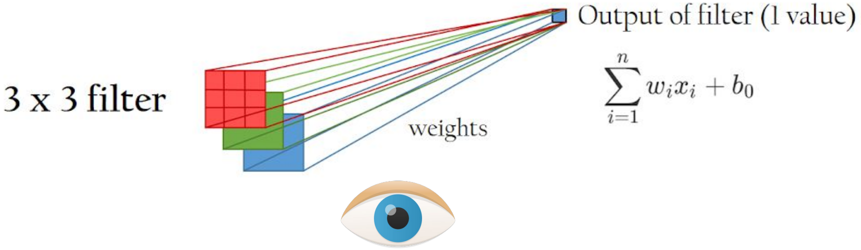
- "Deep" is useful when features need to be learned

"A Little Learning"
A little learning is a dangerous thing;
Drink deep, or taste not the Pierian spring:
There shallow draughts intoxicate the brain,
And drinking largely sobers us again.
- by Alexander Pope

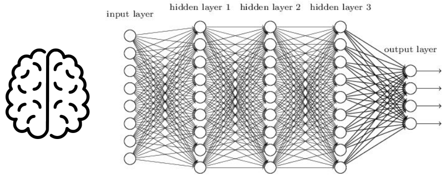


Convolutional Neural Networks for Image Classification



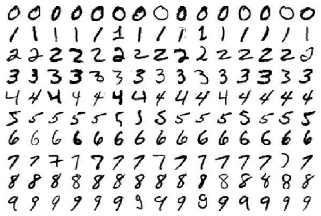


$$\sum_{i=1}^{n} w_i x_i + b_0$$

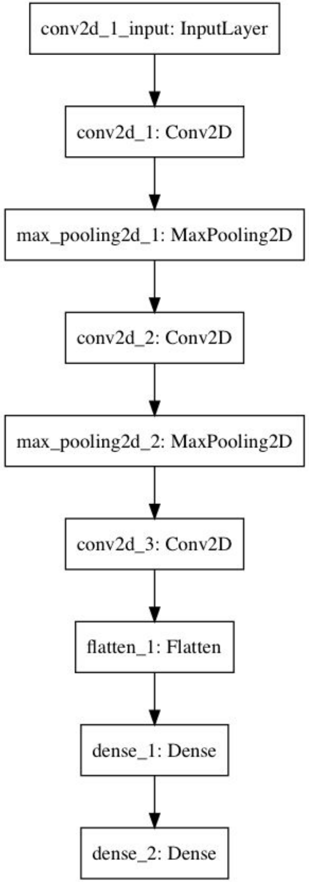


GPUs for Deep CNN Learning

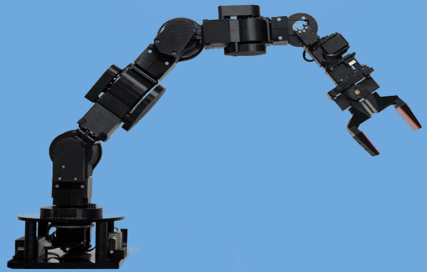
- The MNIST dataset of classifying images
  - contains 60,000 training images and 10,000 testing images



```
with tf.device('/device:GPU:0'):
    model = models.Sequential()
    model.add(layers.Conv2D(32, (3, 3), activation='relu', input_shape=(28, 28, 1)))
    model.add(layers.MaxPooling2D((2, 2)))
    model.add(layers.Conv2D(64, (3, 3), activation='relu'))
    model.add(layers.MaxPooling2D((2, 2)))
    model.add(layers.Conv2D(64, (3, 3), activation='relu'))
    model.add(layers.Flatten())
    model.add(layers.Dense(64, activation='relu'))
    model.add(layers.Dense(10, activation='softmax'))
    model.compile(optimizer='rmsprop', loss='categorical_crossentropy', metrics=['accuracy'])
    model.fit(train_images, train_labels, epochs=8, batch_size=64)
```



AI vs ML vs DL

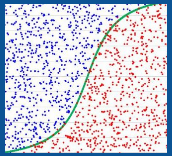


ARTIFICIAL INTELLIGENCE

a very broad field

including algorithms such as DFS, A* search

1950s



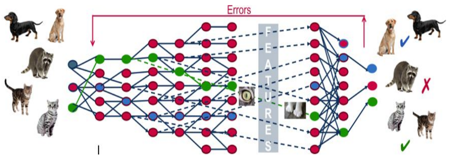
MACHINE LEARNING

“learning from data”

1980s

Deep Learning

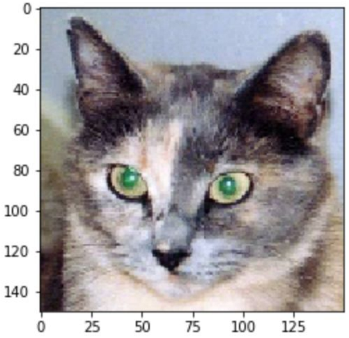
Trending ML methods

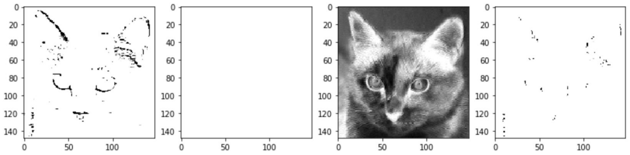


2010s

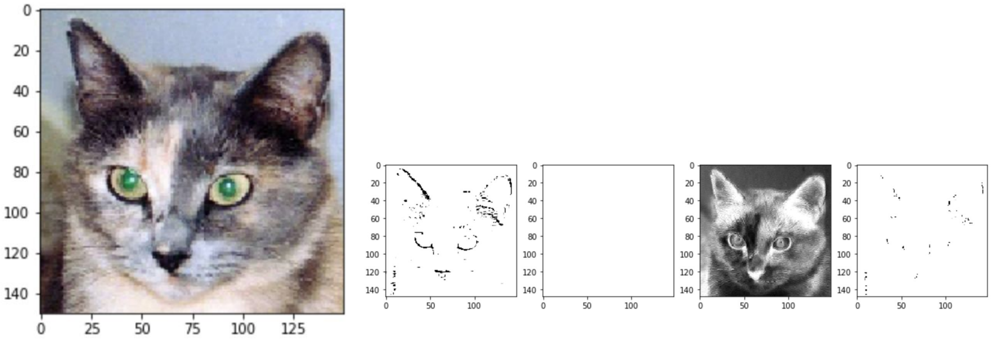
Deep Learning Models are NOT Black Boxes

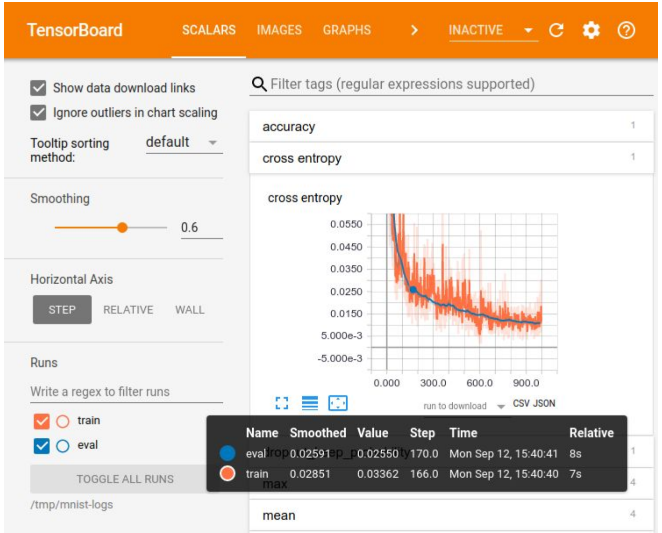
Deep Learning Models are NOT Black Boxes



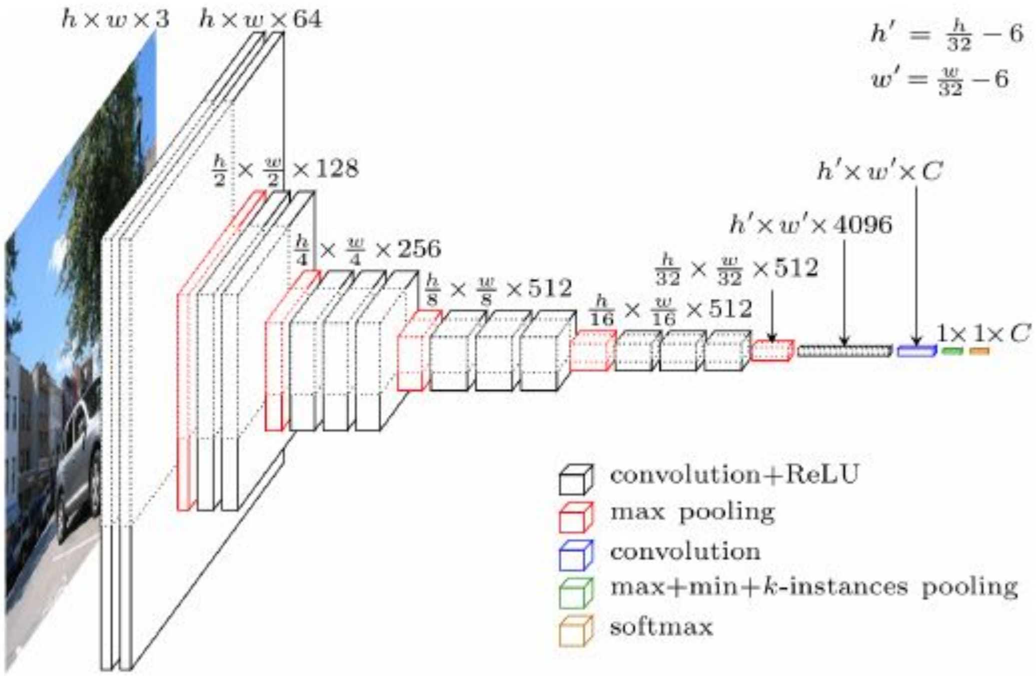


Deep Learning Models are NOT Black Boxes





Transfer Learning



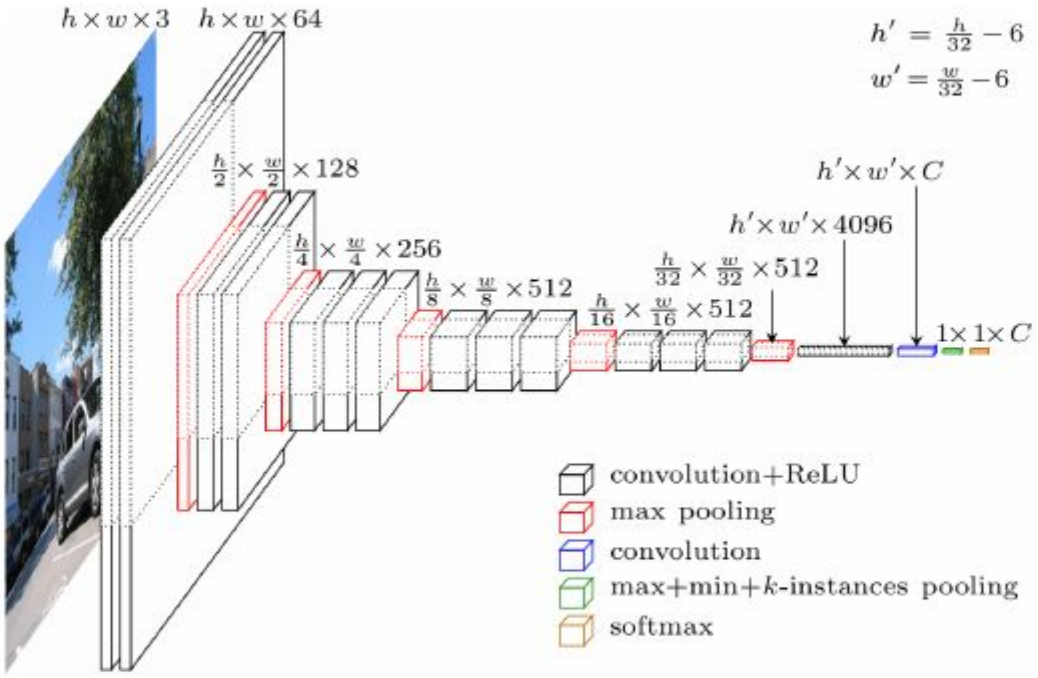
The VGG-16 Architecture

- A deep convolutional network for object recognition developed and trained by Oxford's renowned Visual Geometry Group (VGG)
- VGGNet performed very well in the Image Net Large Scale Visual Recognition Challenge (ILSVRC) in 2014

Current Practice:

- Use pretrained models such as VGG16, Inception-v3 (by Google), etc.
- Most of them are independent of image size (the convolutional layers)

Transfer Learning



The VGG-16 Architecture

- A deep convolutional network for object recognition developed and trained by Oxford's renowned Visual Geometry Group (VGG)
- VGGNet performed very well in the Image Net Large Scale Visual Recognition Challenge (ILSVRC) in 2014

Current Practice:

- Use pretrained models such as VGG16, Inception-v3 (by Google), etc.
- Most of them are independent of image size (the convolutional layers)

Example:
You want to build your own face recognizer to unlock your door

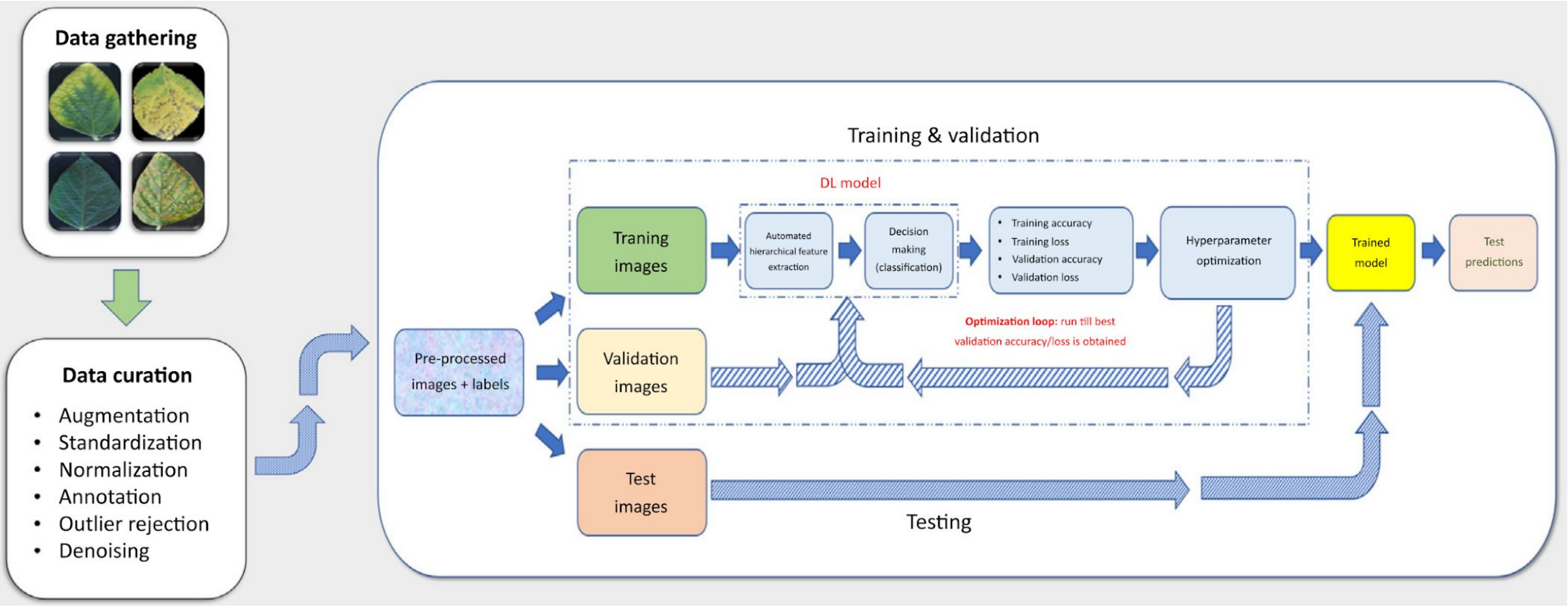
Limitations of DL

- Deep learning model is **just a chain of simple continuous geometric transformations** mapping one vector space into another

- A deep learning model can be interpreted as a kind of program; **but inversely most programs can't be expressed as deep learning models**
    - algorithm ≠ deep learning model

- Extreme generalization vs Local generalization
    - Extreme generalization: an ability to adapt to novel, never-before-experienced situations using little data or even no new data at all (abstraction and reasoning)
    - Local generalization: mapping from inputs to outputs

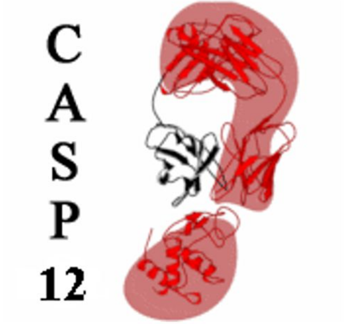
DL Tool Chain: From Gathering Data to Decision Making



Deep Learning for Plant Stress Phenotyping: Trends and Future Perspectives

Asheesh Kumar Singh,[1] Baskar Ganapathysubramanian,[2] Soumik Sarkar,[2,*] and Arti Singh[1,*]

How Accurately Can We Predict Protein Structures Today?



World-wide competition held every two years (3 months long)

Competition: CASP12 (2016)				
Predictor: Baker-Rosetta (UW)				
Protein		**RMSD**		
Type	**Count**	**Best**	**Median**	**Worst**
Template-based	57	0.69	4.7	24.2
Template-free	58	2.04	12.9	22.8

most recent competition

a top participant

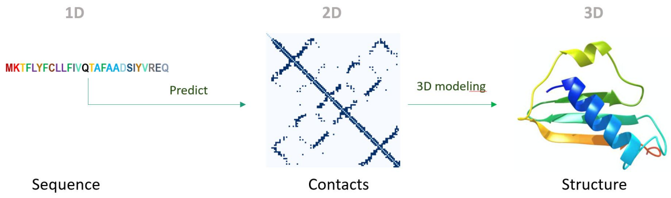
root mean square deviation

dataset

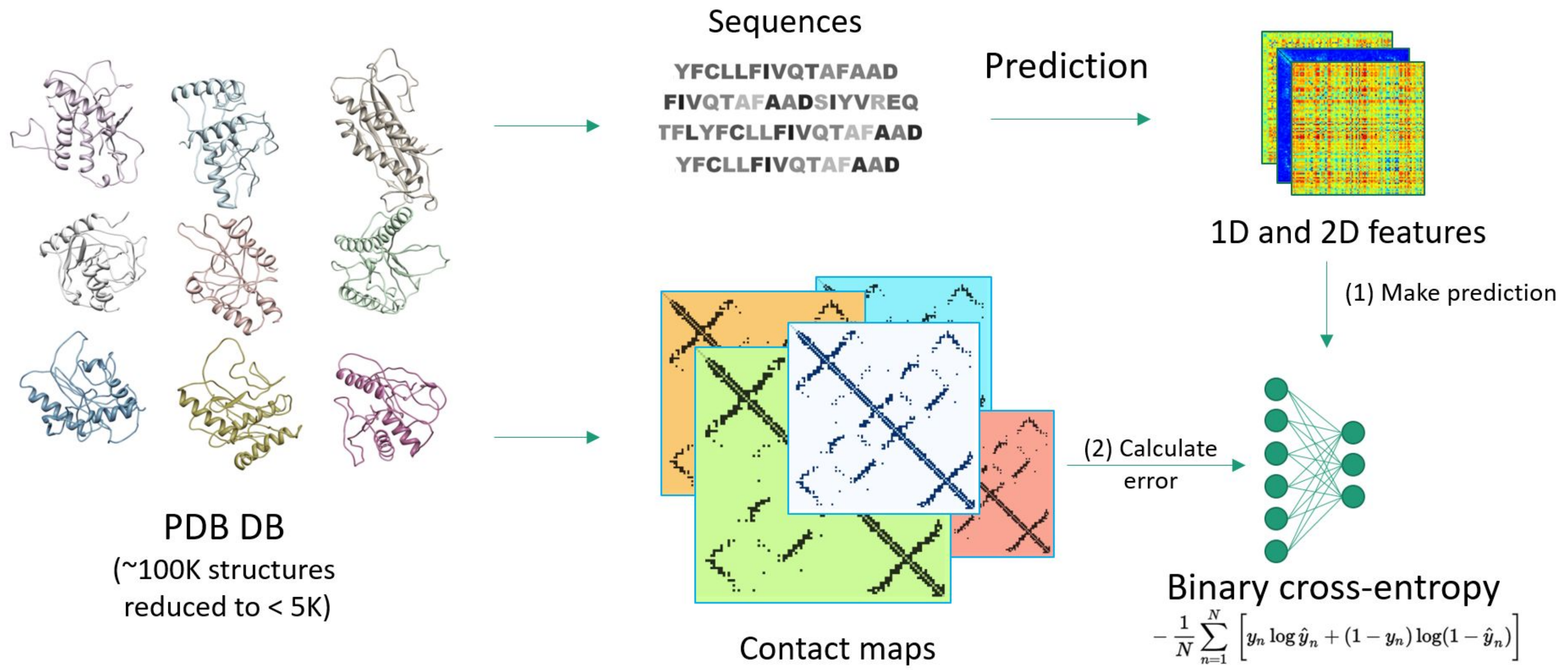
99% similarity (experimental biologists' are happy)

random prediction

Significance of Contact prediction

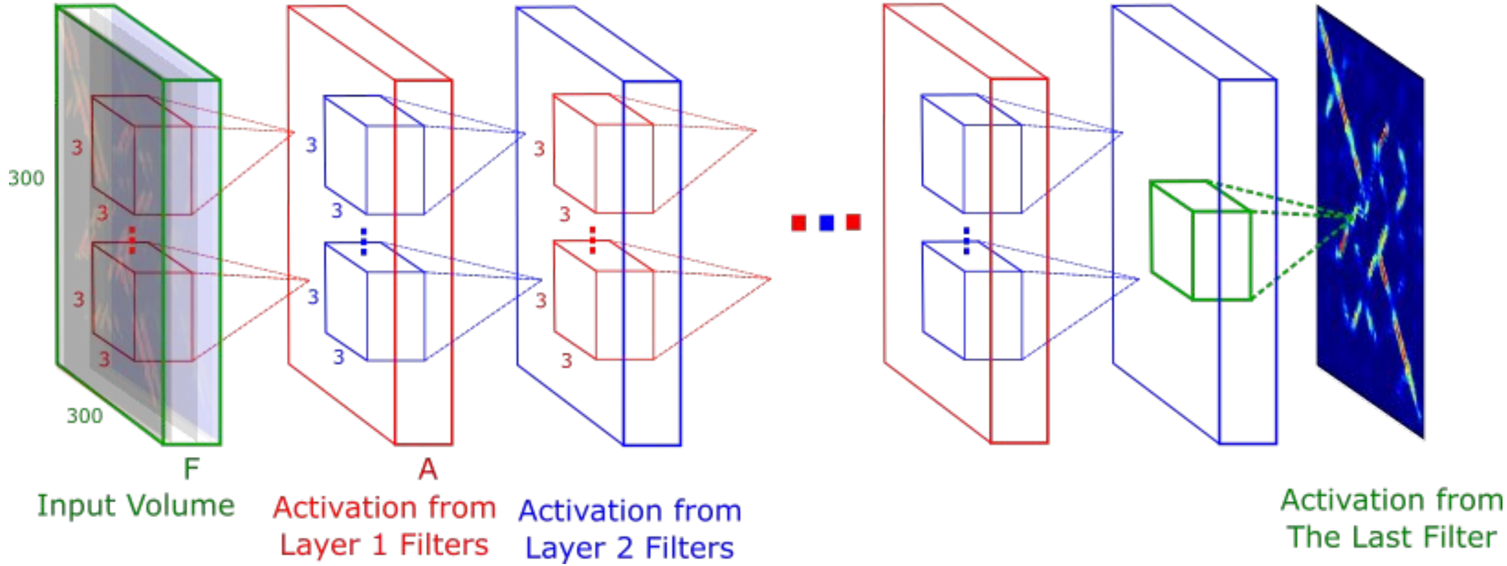


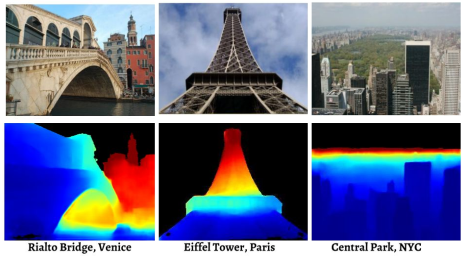
Protein Contact Prediction as a Machine Learning Problem



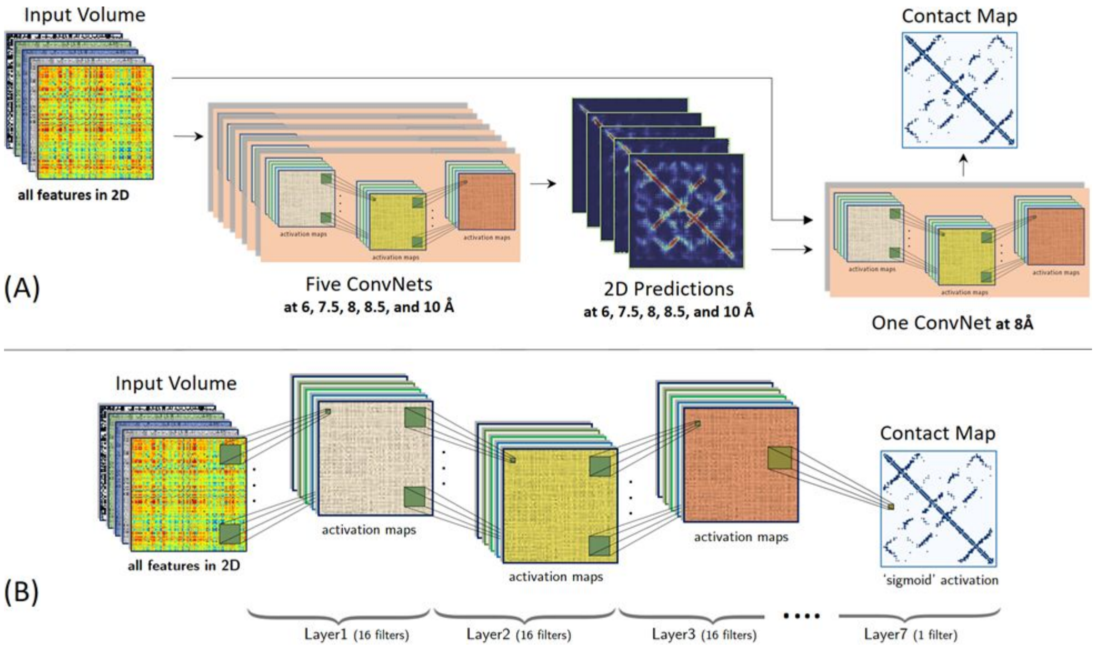
$$-\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)\right]$$

CNNs for Protein Contact Prediction





The DNCON2 Method for Protein Contact Prediction

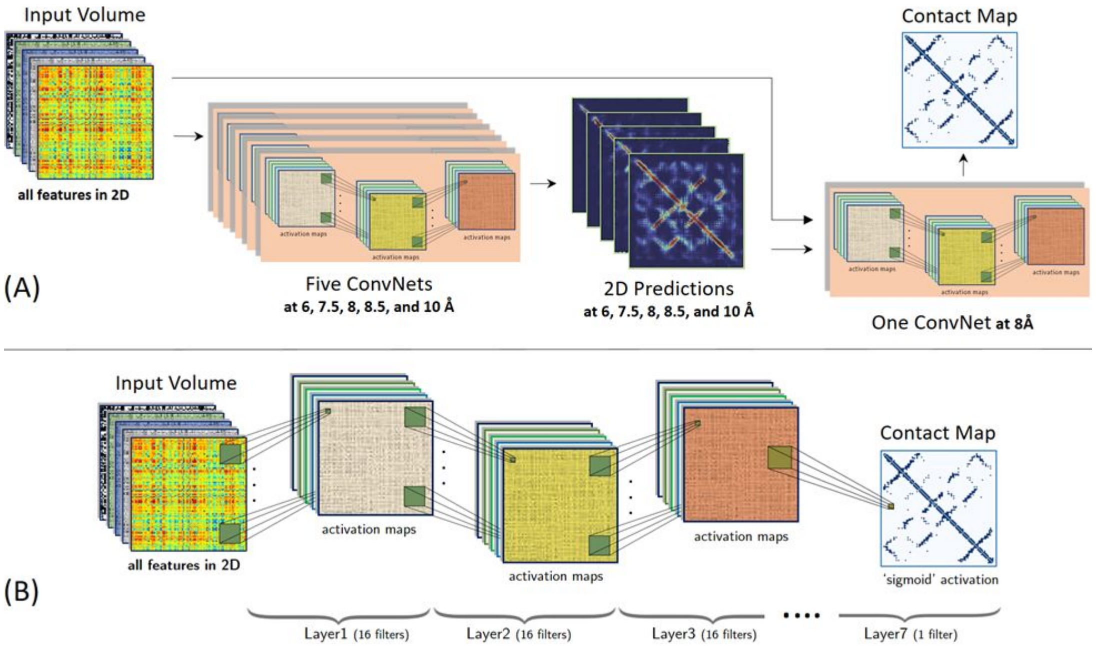


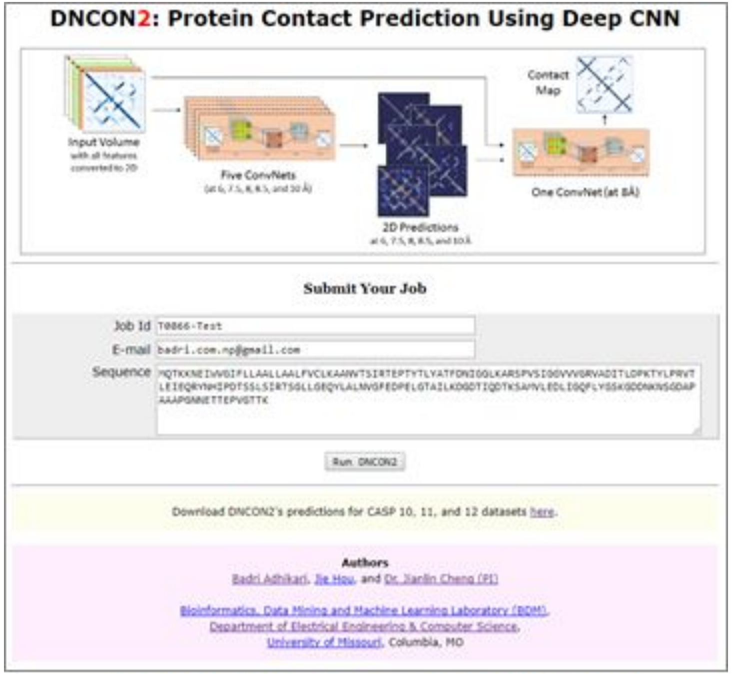
Structural Bioinformatics

DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks

Badri Adhikari, Jie Hou, and Jianlin Cheng

The DNCON2 Method for Protein Contact Prediction





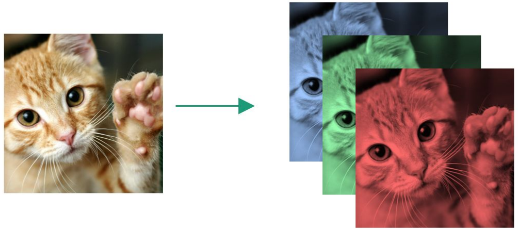
http://sysbio.rnet.missouri.edu/dncon2/

Structural Bioinformatics

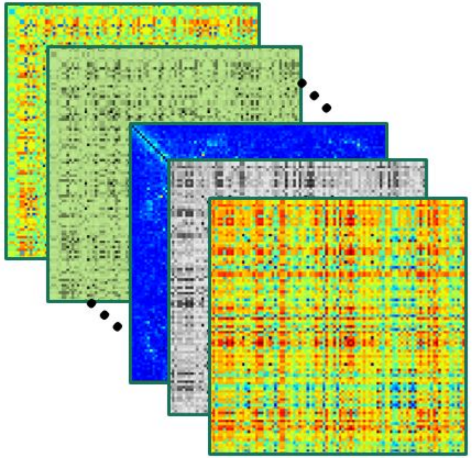
DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks

Badri Adhikari, Jie Hou, and Jianlin Cheng

Number of Features (Channels) in Bioinformatics Problems



3 channels

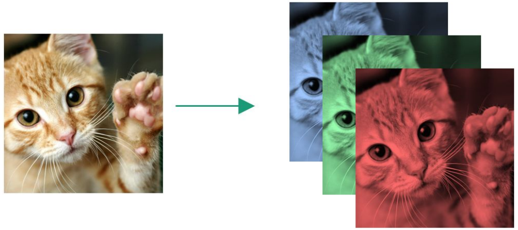


around 100 channels

Object Recognition

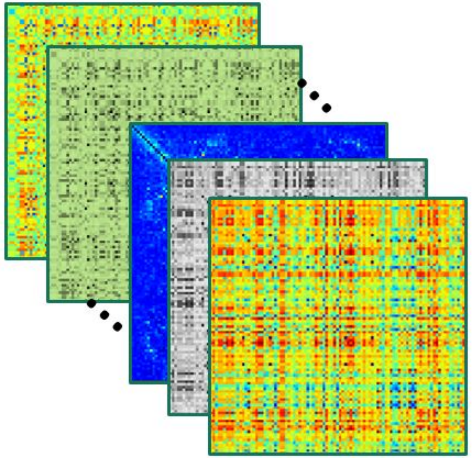
Protein Structure Prediction

Number of Features (Channels) in Bioinformatics Problems



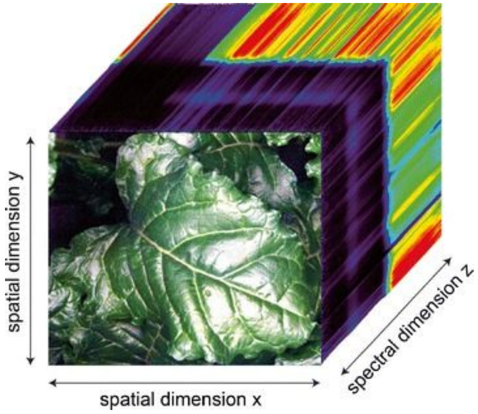
3 channels

Object Recognition



around 100 channels

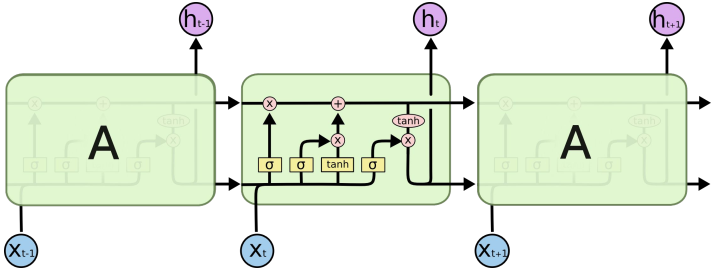
Protein Structure Prediction

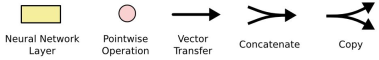


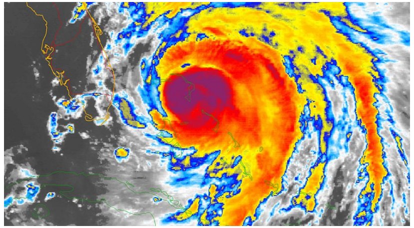
Hyperspectral imaging
at Donald Danforth Plant Science Center

Long Short Term Memory networks (may) have a lot of potential for Problems in Bioinformatics

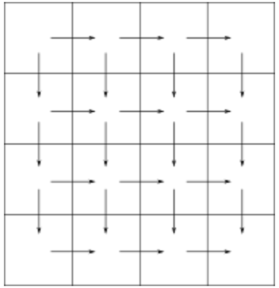
1D LSTM

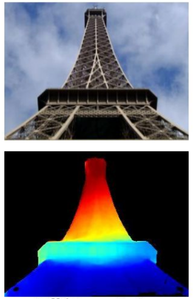




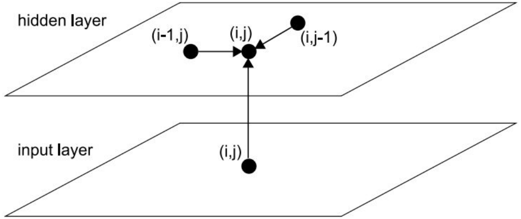


2D LSTM

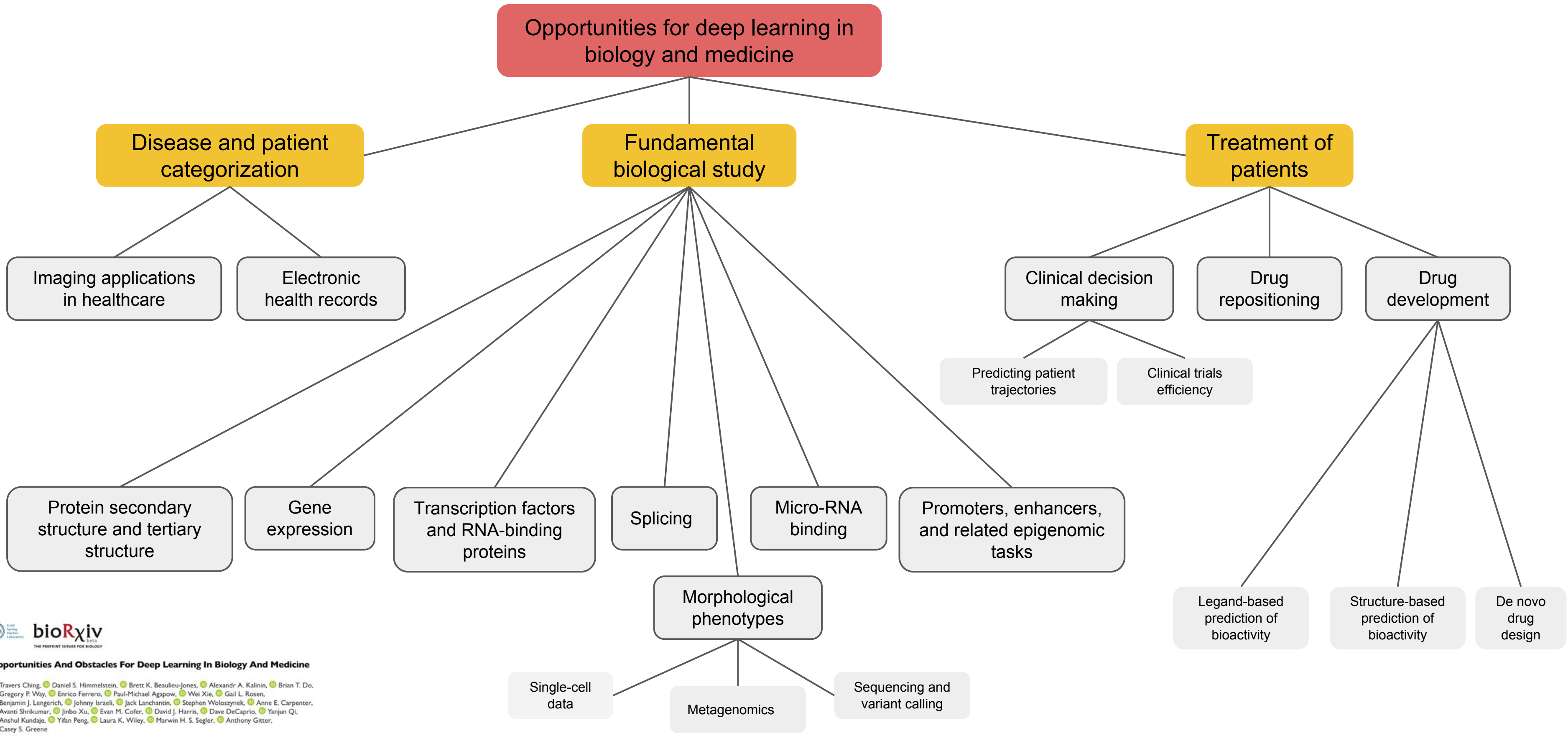




Eiffel Tower, Paris



Deep Learning for Biology and Medicine



bioRχiv beta — THE PREPRINT SERVER FOR BIOLOGY

Opportunities And Obstacles For Deep Learning In Biology And Medicine

Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H. S. Segler, Anthony Gitter, Casey S. Greene

doi: https://doi.org/10.1101/142760

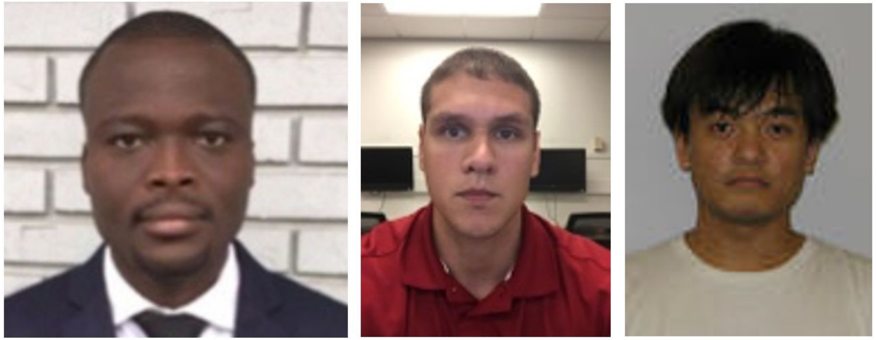
Conclusion

- Deep learning is models are not a black boxes but deep learning does have limitations

- Convolutional neural networks (and its variants) have a huge potential to more accurately solve many problems in bioinformatics

- CNNs have dramatically improved the accuracy of protein contact prediction, just like they have for many other problems

Acknowledgements



From left - Anthony Ackah-Nyanzu, Cody Hawkins, and Pak Kong

Thank You !!