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

#### Truth Tables

<table>
<thead>
<tr>
<th>x AND Y</th>
<th>x OR y</th>
<th>(x) AND (y)</th>
<th>x XNOR y</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>f-and(x,y)</td>
<td>x</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

#### Circuit Diagrams

![Circuit Diagrams](image)

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

Input Image  

Input Volume  

Dot products  

Activation map  

Fully connected network  

3 x 3 filter  

Output of filter (1 value)  

\[ \sum_{i=1}^{n} w_i x_i + b_0 \]
- The MNIST dataset of classifying images
  - contains 60,000 training images and 10,000 testing images

```python
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

1950s

Artificial Intelligence
a very broad field
including algorithms such as DFS, A* search

1980s

Machine Learning
“learning from data”

2010s

Deep Learning
Trending ML methods
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
Deep Learning for Plant Stress Phenotyping: Trends and Future Perspectives

Data gathering:
- Augmentation
- Standardization
- Normalization
- Annotation
- Outlier rejection
- Denoising

Data curation:

Training & validation:
- DL model
  - Automated hierarchical feature extraction
  - Decision making (classification)
  - Hyperparameter optimization
- Optimization loop: run till best validation accuracy/loss is obtained

Testing:
- Trained model
- Test predictions
How Accurately Can We Predict Protein Structures Today?

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Type</td>
<td>Count</td>
</tr>
<tr>
<td>Template-based</td>
<td>57</td>
</tr>
<tr>
<td>Template-free</td>
<td>58</td>
</tr>
</tbody>
</table>

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

99% similarity (experimental biologists’ are happy)

Random prediction

Most recent competition

Root mean square deviation

Significance of Contact prediction

Sequence → Contacts → Structure
Protein Contact Prediction as a Machine Learning Problem

Sequences
YFCLLLFFIVQTAFAAD
FIVQTAFAADSYYVREQ
TFLYFCLLLFFIVQTAFAAD
YFCLLLFFIVQTAFAAD

Prediction

1D and 2D features
(1) Make prediction

Contact maps
(2) Calculate error

Binary cross-entropy
\[- \frac{1}{N} \sum_{n=1}^{N} y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n)\]

PDB DB
(~100K structures reduced to < 5K)
CNNs for Protein Contact Prediction
The DNCON2 Method for Protein Contact Prediction

(A) Input Volume → Five ConvNets at 6, 7.5, 8, 8.5, and 10 Å → 2D Predictions at 6, 7.5, 8, 8.5, and 10 Å → One ConvNet at 8 Å

(B) Input Volume → Activation maps → Layer1 (16 filters) → Layer2 (16 filters) → Layer3 (16 filters) → … → Layer7 (1 filter) → Contact Map

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

Badri Adhikari, Jia 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, Jia Hou, and Jianli Cheng
Number of Features (Channels) in Bioinformatics Problems

Object Recognition

Protein Structure Prediction

3 channels

around 100 channels
Number of Features (Channels) in Bioinformatics Problems

Object Recognition

Protein Structure Prediction

Hyperspectral imaging at Donald Danforth Plant Science Center

3 channels

around 100 channels

Hyperspectral imaging at Donald Danforth Plant Science Center
Long Short Term Memory networks (may) have a lot of potential for Problems in Bioinformatics
Deep Learning for Biology and Medicine

Opportunities for deep learning in biology and medicine

Disease and patient categorization
- Imaging applications in healthcare
- Disease and patient categorization
- Electronic health records
- Fundamental biological study
- Splicing
- Protein secondary structure and tertiary structure
- Gene expression
- Transcription factors and RNA-binding proteins
- Micro-RNA binding
- Promoters, enhancers, and related epigenomic tasks
- Morphological phenotypes
- Single-cell data
- Metagenomics
- Sequencing and variant calling

Treatment of patients
- Clinical decision making
- Drug repositioning
- Drug development
- Predicting patient trajectories
- Clinical trials efficiency
- Treatment of patients
- Imaging applications in healthcare
- Clinical decision making
- Drug repositioning
- Drug development
- Predicting patient trajectories
- Clinical trials efficiency
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-Nyansu, Cody Hawkins, and Pak Kong
Thank You !!