What is deep learning and how is it used in biology

Claire McWhite
@clairemcwhite
UT Austin, BCH394P/BCH364C
Princeton University
4/12/2022

What is

• How does deep learning work?
• What types of biological problems is deep learning useful for
What is machine learning?

Why use machine learning?

- Automate a manual decision process at larger scale
- Most deep learning doesn’t do anything that a human couldn’t do given enough time
  - Sort ripe from unripe tomatoes
  - Decide if a cat is in a photo
When do we use machine learning?

- When you have a set of examples with any label

- Classification
  - Data: Tumor biopsies label: malignant vs. benign
  - Data: Protein sequences label membrane vs soluble

- Regression
  - How much fuel will an airplane use?

Linear regression is machine learning

- Formula \( y = Ax + b \)
- \( A \) = slope, \( b \) = y intercept

- What are the values of \( A \) and \( b \) that minimize the errors between the line and the points

- Optimization: Find the values of variables by minimizing error to the expected result
Linear regression is machine learning

- A person could figure out the best values of $y = Ax + b$ by trying values and checking how large the error is between the line and the point.
- However, the process of doing can be done automatically by plotting the error as the variable changes.

- We can look at the error as the y intercept changes.
Linear regression is machine learning

• We can look at the error as we vary value for the y intercept

• With the relationship between y-intercept and error, we can find the value of y-intercept that minimizes the error
Importantly, both slope and intercept can be optimized simultaneously.

Imagine rolling a ball into this plane, seeing where it lands, and taking the values of slope and y-intercept at that point. This process isn’t limited to just two variable at a time, many variables can be optimized at one.


---

Deep learning is subset of machine learning

- Deep learning refers specifically to deep neural network architecture
- Architecture = how a model is set up

[DEEP LEARNING (DL) A deep neural network](https://quantib.com)
Imagery around deep learning is unhelpful

What does this mean?
What are the dots?
What are the line?
How does this predict?

Neural networks are a series of data transformations

- Training a neural network is figuring out the parameters for the different data transformations
What neural networks aren’t

- Not a brain-like or neuron-like structure
- Essentially they are a series of data transformations
- With parameters for the data transformations learned simultaneously
  - Like in the linear regression example

Contrived example

- A farm with blackberries and raspberries
- They would like a robot to automatically sort berries, but all they have is a scale
- They know that fruit weight varies drastically with rainfall levels

Options

- Make a rule for different rainfall levels
  - Ex. If there’s been 3-5 inches of rainfall, anything above 2 grams is a blackberry
- Find an equation that divides the two
- Use a neural network to find data transformations that allow you to cleanly separate between the two berries
Contrived example

- Transform x-axis: $\tan(x)$
- Rotate data
- Reduce to one dimension
- If output value > 0, blueberry
  - If < 0, raspberry

Weights are parameters of equations
- Ex. Linear equation
  - Weights can be positive or negative
• Neural networks begin with random weights
• Random weights create random output
• We can find good values from the weights because we can measure the error between the produced output and the expected output

Every variable has its own optimization relative to the error
Multiple parameters can be optimized simultaneously

Gradient descent allows the whole network to be optimized simultaneously

Finding the values for each of the data transformations that separate blueberries and raspberries

Source: https://reconsider.news/2018/05/09/ai-researchers-allege-machine-learning-alchemy/
Applications of neural networks in biology

- Something a human could do, but it takes too much time to do it more than a few times
- Particularly used in image analysis
  - A person can look at a few hundred images, a train classifier can look through many more

Convolution is a common step in image processing neural networks

In a convolutional neural networks, the original image is modified with multiple filters to create richer inputs for the model to learn from

https://medium.com/analytics-vidhya/understanding-convolution-operations-in-cnn-1914045816d4
Convolution moves a filter over an input image to create a new image

https://medium.com/@bdhuma/6-basic-things-to-know-about-convolution-daef5e1bc411

Different types of filters

<table>
<thead>
<tr>
<th>Original</th>
<th>Gaussian Blur</th>
<th>Sharpen</th>
<th>Edge Detection</th>
</tr>
</thead>
</table>
| \[
| 0 & 0 & 0 \\
| 1 & 0 & 0 \\
| 0 & 0 & 0 \\
| \] | \[
| 1 & 2 & 1 \\
| 2 & 4 & 2 \\
| 1 & 2 & 1 \\
| \] | \[
| 0 & -1 & 0 \\
| -1 & 5 & -1 \\
| 0 & -1 & 0 \\
| \] | \[
| -1 & -1 & -1 \\
| -1 & 8 & -1 \\
| -1 & -1 & -1 \\
| \] |

https://medium.com/@bdhuma/6-basic-things-to-know-about-convolution-daef5e1bc411
Classifying protein localization

Aiming to separate images of live and dead neurons, convolutional neural networks create richer representations for the fully connected layers.

Classifying living/dead neurons

- Network to separate images of live and dead neurons
- Convolutional neural networks create richer representations for the fully connected layers
Natural language processing

• By just looking at all the text in Wikipedia, a natural language model can accurately describe English, even without prior knowledge about grammar/parts of speech/etc.
  • Many differently styles of computer model for doing this
• Give these same models every protein sequence, and you get a model that describes “protein language”

Why model protein sequence as a language?

• Sequences of tokens, where token meaning depends on context

• Language: “I drank a cold strawberry smoothie”
• Protein: “M E C E C A H N S H Q”

* To format protein sequences for language models, all you need to do is put a space between each amino acid
Why model protein sequence as a language?

- There are specific relationships between tokens
  - subject-verb modifier
  - “I drank a cold strawberry smoothie”
  - C2H2 zinc-finger motif
  - “M E C E C A H N S H Q”

- A sentence can be represented as a long numeric vector
- The difference between any two sentences can be measured by comparing their vectors
Why model protein sequence as a language?

• Some positions in the sequence can be changed without changing the meaning of the sentence too much

“I drank a cold strawberry smoothie”
“I sipped a chilled strawberry smoothie”

“M E C E C A H N S H E”
“M E C E C A H N S H Q”

Highly similar

“I drank a cold computer smoothie”

“M E C E C A H N S H E”
“M E C E C A H N S H P”

Change in meaning

Change in meaning
Each amino acid has its own meaning vector

*The sentence vector is the mean of all word vectors

"I drank a cold strawberry smoothie"

Vectors are composed of many smaller vectors, which capture different relationships

"I drank a cold strawberry smoothie"

"M E C E C A H N S H E"
A calculation between word vectors overlays a network onto the sequences

“\text{I drank a cold strawberry smoothie}”

“\text{M E C E C A H N S H E}”

Recap

- Each word/amino acid has a long vector that captures its identity
- Averaging these word vectors creates a sentence vector
- Vectors are composed of smaller vectors which capture different properties
Connecting back to neural networks

- Weighted average steps alternated with neural networks with trainable weights.
- The weighted average operation is called “attention.”

Each of the twenty amino acids begins with one of twenty vectors.
By mixing information between them, output vectors that represent each amino acid in context are produced.
Semantic shift variant effect prediction

Deep mutational scanning dataset

Measure the similarity between a mutant sequence and wild-type

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Functional readout</th>
<th>Semantic shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>M E C E C A H N S H E</td>
<td>Wild-type</td>
<td>0</td>
</tr>
<tr>
<td>M E C Q C A H N S H E</td>
<td>Neutral</td>
<td>0.1</td>
</tr>
<tr>
<td>M E C Q P A H N S H E</td>
<td>Deleterious</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Using models trained for sequence similarity should improve performance

Modern language models trained on large numbers of sequences can be fine-tuned to produce embeddings that are better suited for particular tasks

Don’t need to create a whole new neural for every problem

Jay Alammar – The illustrated transformer