CPSC 532P / LING 530A: Deep Learning for Natural Language Processing (DL-NLP)

Muhammad Abdul-Mageed

muhammad.mageed@ubc.ca

Natural Language Processing Lab

The University of British Columbia

Table of Contents

Many of the current slides are a summary of an overview of deep learning in Goodfellow et al., 2016. More information can be found therein. Other sources are credited in the respective slides.

Creating Machines That Think

- Said to go back to at least the time of ancient Greece.
- Mythical figures like Pygmalion and Daedalus may be interpreted as legendary inventors.
- Artifacts like Galatea and Pandora may be regarded as artificial life. (e.g., Ovid & Martin, 2004 in Goodfellow et al., 2016).

Figure: Pygmalion and the Statue [Wikimedia commons].

Artificial Intelligence

- People wondered whether programmable machines might become intelligent over 100 years before one was built (Lovelace, 1842).

Figure: Ada Lovelace [Wikipedia].

Early AI Successes

- Early successes of AI took place in relatively sterile and formal environments and did not require computers to have much knowledge about the world.

IBM Deep Blue

IBM's Deep Blue chess-playing system defeat of world champion Garry Kasparov in 1997 (Hsu, 2002).

Deep Blue

- Chess is a very simple world (only 64 locations and 32 pieces that can move in only rigidly circumscribed ways), and can be completely described by a very brief list of completely formal rules.

Figure: Deep Blue defeats Garry Kasparov. [The Atlantic].

Common Sense Knowledge

- Humans posses immense amount of knowledge about the world.
- Much of this knowledge is **subjective** and **intuitive**; $\rightarrow$ (difficult to represent formally to computer).
- Representation challenge.
- **Knowledge base approach:** hard-code knowledge about the world in formal languages, with a computer trying to reason about statements in these formal languages automatically using logical inference rules.
- Still hard...

Machine Learning

- AI systems need the ability to **acquire their own knowledge**, by extracting patterns from raw data.
- This capability is known as **machine learning**.
- A simple ML algorithm called **logistic regression** can determine whether to recommend cesarean delivery (Mor-Yosef et al., 1990).
- A simple machine learning algorithm called **naive Bayes** can separate legitimate e-mail from spam e-mail.

Representation

- Simple ML algorithms such as logistic regression depends heavily on the **representation** of the data.
- For example, the doctor tells the system several pieces of relevant information, such as the presence or absence of a uterine scar.
- Each piece of information in the representation of the patient is known as a **feature**.
- Logistic regression learns how each of these features correlates with various outcomes, but cannot influence the way that the features are defined in any way.
- If logistic regression is given an MRI scan of the patient, it would not be able to make useful predictions based on it.

The Challenge of Representation

- Sometimes we do not know a priori what features are relevant to the task.
- Feature engineering is time consuming.
- One solution: use ML to discover not only the mapping from representation to output but also the representation itself. (**Representation learning**).
- Learned representations often yield much better performance and can also allow AI systems to rapidly adapt to new tasks. (**Transfer learning**).

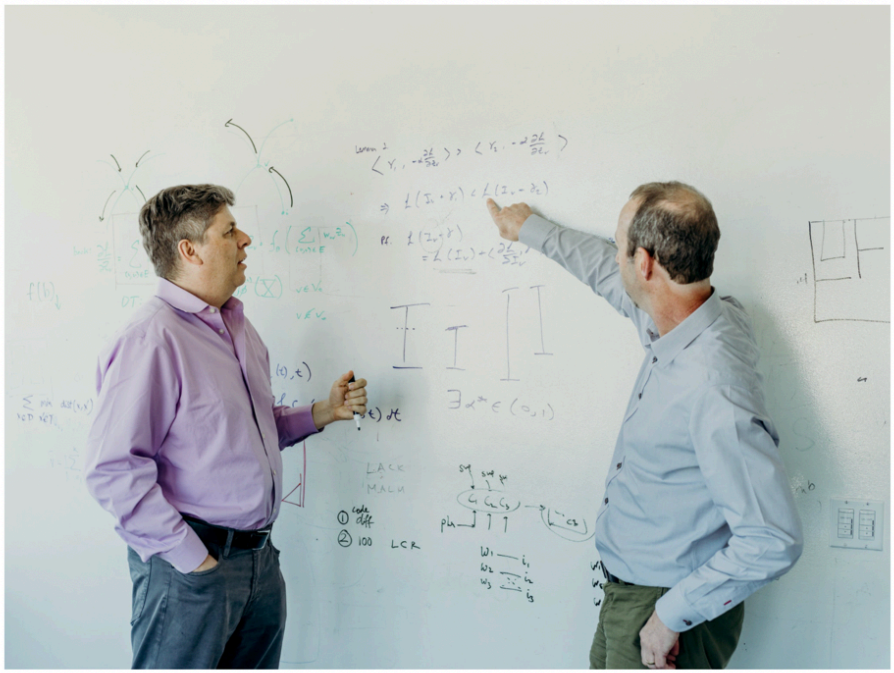
Passing an 8th-Grade Science Test

A Breakthrough for A.I. Technology: Passing an 8th-Grade Science Test

By Cade Metz

Sept. 4, 2019

The New York Times



Dr. Etzioni and Peter Clark, manager of the Aristo project, at the Allen Institute in Seattle.
Kyle Johnson for The New York Times

Deep Learning

- Deep learning is a solution to problems that **allows computers to learn from experience**, and **understand the world in terms of a hierarchy of concepts**.
- Each concept is defined in terms of its relation to **simpler concepts**.
- By gathering knowledge from experience, deep learning **avoids the need for humans to formally specify all of the knowledge** that the computer needs.
- The hierarchy of concepts allows the computer to **learn complicated concepts by building them out of simpler ones**.

Factors of Variation

- There are usually multiple "factors of variation" (Goodfellow et al., 2016) that can explain a given real-world phenomenon.
- These factors can be concepts or abstractions that help us make sense of the rich variability in the data.

Factors of Variation

- There are usually multiple "factors of variation" (Goodfellow et al., 2016) that can explain a given real-world phenomenon.
- These factors can be concepts or abstractions that help us make sense of the rich variability in the data.

Examples

Certain lexica (e.g., "good" and "bad"), emojis, etc. are factors of variation in sentiment and emotion expression.

Factors of Variation

- There are usually multiple "factors of variation" (Goodfellow et al., 2016) that can explain a given real-world phenomenon.
- These factors can be concepts or abstractions that help us make sense of the rich variability in the data.

Examples

Certain lexica (e.g., "good" and "bad"), emojis, etc. are factors of variation in sentiment and emotion expression.

- But even these cues have networks of factors, some of which might be hidden.

Factors of Variation

- There are usually multiple "factors of variation" (Goodfellow et al., 2016) that can explain a given real-world phenomenon.
- These factors can be concepts or abstractions that help us make sense of the rich variability in the data.

Examples

Certain lexica (e.g., "good" and "bad"), emojis, etc. are factors of variation in sentiment and emotion expression.

- But even these cues have networks of factors, some of which might be hidden.

Examples

Suppose a given emoji develops a very odd usage in a community unknown to us...

Factors of Variation

- There are usually multiple "factors of variation" (Goodfellow et al., 2016) that can explain a given real-world phenomenon.
- These factors can be concepts or abstractions that help us make sense of the rich variability in the data.

Examples

Certain lexica (e.g., "good" and "bad"), emojis, etc. are factors of variation in sentiment and emotion expression.

- But even these cues have networks of factors, some of which might be hidden.

Examples

Suppose a given emoji develops a very odd usage in a community unknown to us...

- To learn, we need to separate the **factors of variation** that explain the observed data.

Deep Learning as Representation Learning

- Challenge: In many real-world AI applications, many of the factors of variation influence every single piece of data we are able to observe.[1]

[1] (Goodfellow et al., 2016)

Deep Learning as Representation Learning

- Challenge: In many real-world AI applications, many of the factors of variation influence every single piece of data we are able to observe.[1]

Examples

A word like “disgusting" is used to describe extremely tasty food by some individuals. . .

[1] (Goodfellow et al., 2016)

Deep Learning as Representation Learning

- Challenge: In many real-world AI applications, many of the factors of variation influence every single piece of data we are able to observe.[1]

Examples

A word like "disgusting" is used to describe extremely tasty food by some individuals...

- Thus, most applications require us to disentangle the factors of variation and discard the ones that we do not care about.
- Deep learning solves this problem in representation learning by introducing representations that are expressed in terms of other, simpler representations.
- It allows the computer to build complex concepts out of simpler concepts.

[1] (Goodfellow et al., 2016)

Biological Inspiration I

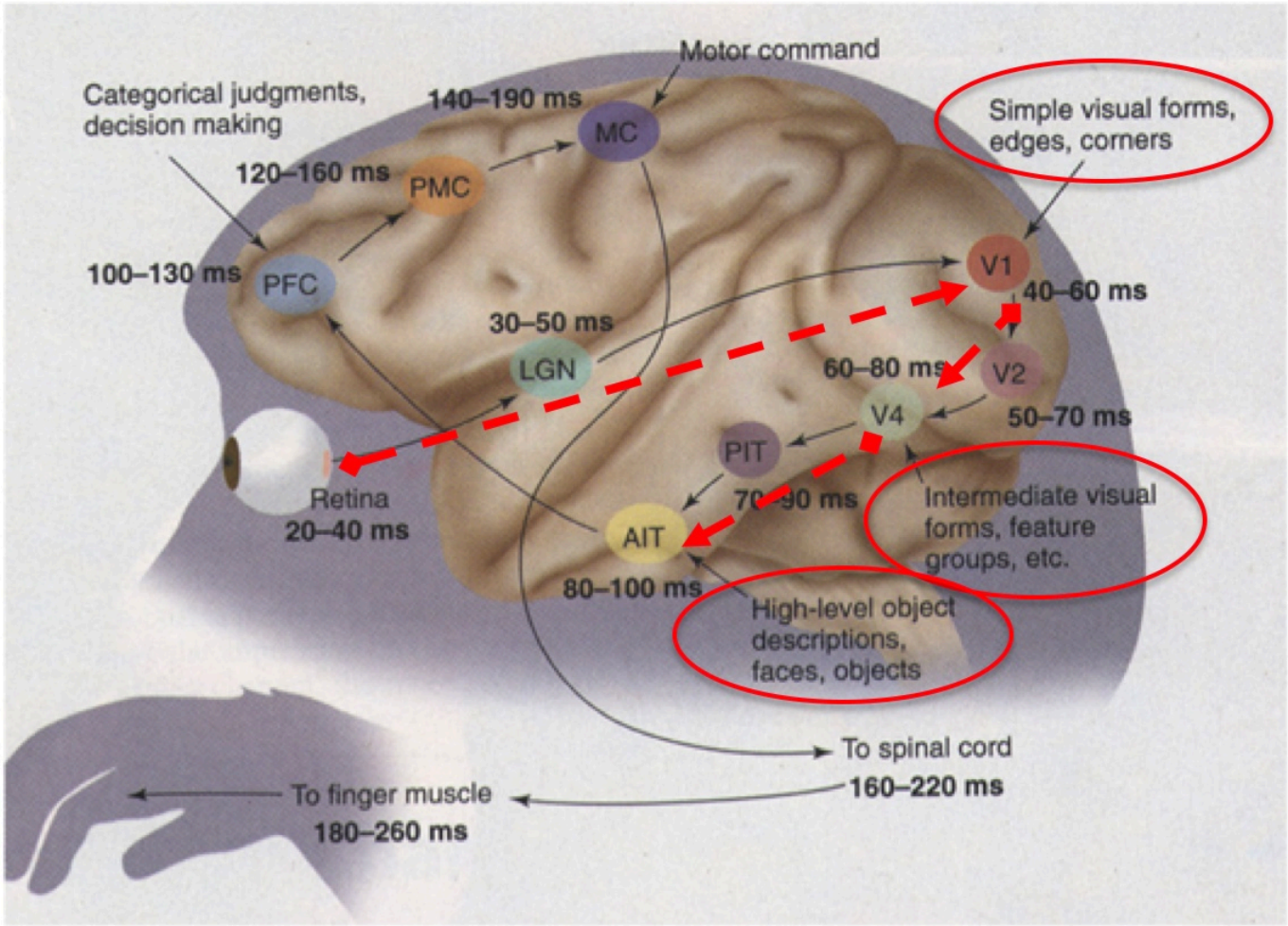


Figure: Visual cortex. [Modified from Simon Thorpe].

Biological Inspiration II

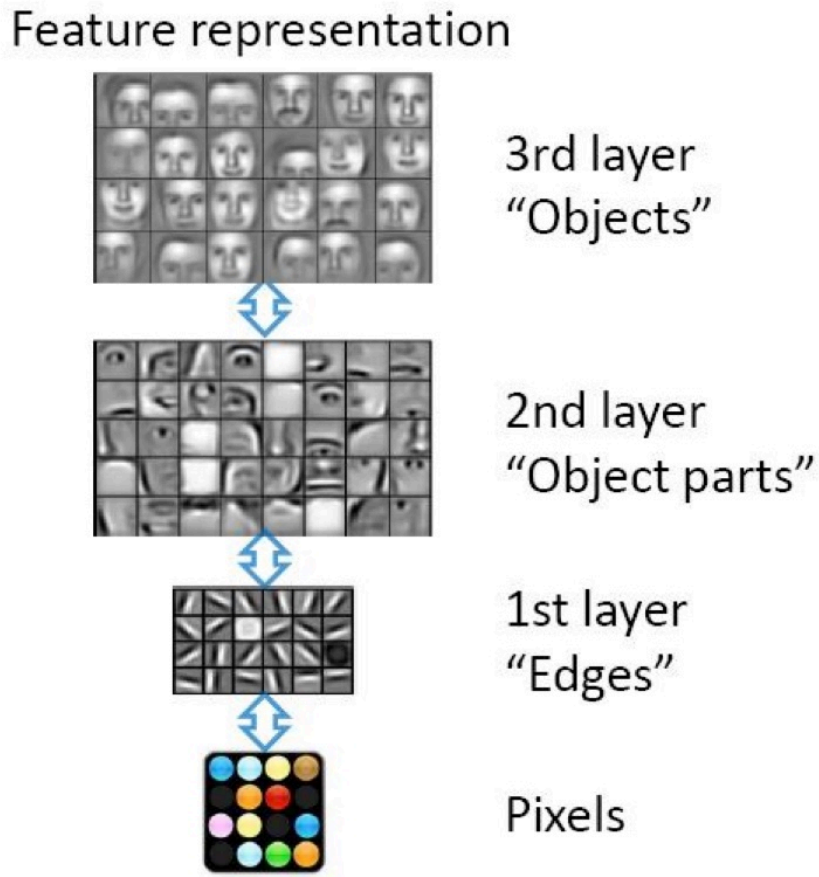


Figure: Visual cortex. [Modified from Simon Thorpe].

Information processing in a neuron

- **Neurons:** Are nerve cells that carry the main information processing in the brain. Estimated between 10^{10} and 10^{11}.
- **Axons:** Each neuron has a very long formation called an axon that connects it to other cells.
- Axons have a biomedical machinery to transmit signals called action potentials (or spikes).
- An **action potential** is an electrical impulse that travels via the axon of a neuron.

Information processing in a neuron I

- **Neurons:** Are nerve cells that carry the main information processing in the brain. Estimated between 1010 and 1011.
- **Axons:** Each neuron has a very long formation called an axon that connects it to other cells.
- Axons have a biomedical machinery to transmit signals called action potentials (or spikes).
- An **action potential** is an electrical impulse that travels via the axon of a neuron. [From Hyvarinen, Hurri, and Hoyer, 2009]

Information processing in a neuron II

- **Dendrites:** formations of a neuron that receive signals from another neuron's axons. (input 'wires' vs. axons which are output 'wires')
- **Synapse:** The site where an axon meets a dendrite.
- **Soma:** Is the main cell body
- **Firing rate:** The number of spikes emitted by a neuron per second. [From Hyvarinen, Hurri, and Hoyer, 2009]

Information processing in a neuron III

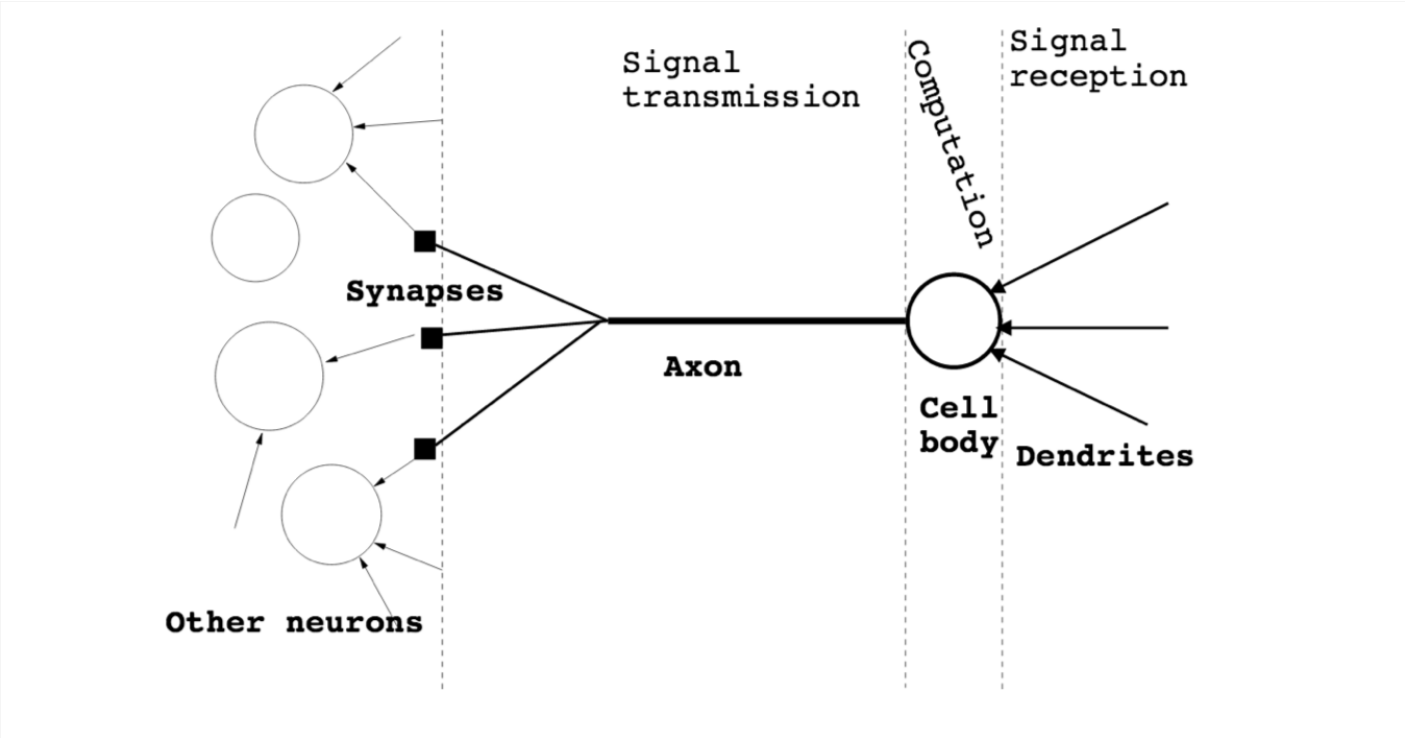


Figure: Information processing in a neuron. [Hyvarinen, Hurri, & Hoyer, 2009]

Information processing in a neuron IV

Computation by the neuron How is information processed, i.e. how are the incoming signals integrated in the soma to form the outcoming signal? This question is extremely complex and we can only give an extremely simplified exposition here.

A fundamental principle of neural computation is that the reception of a spike at a dendrite can either excite (increase the firing rate) of the receiving neuron, or inhibit it (decrease the firing rate), depending on the neuron from which the signal came. Furthermore, depending on the dendrite and the synapse, some incoming signals have a stronger tendency to excite or inhibit the neuron. Thus, a neuron can be thought of as an elementary pattern-matching device: its firing rates is large when it receives input from those neurons which excite it (strongly), and no input from those neurons which inhibit it. A basic mathematical model for such an action is to consider the firing rate as a linear combination of incoming signals; we will consider linear models below.

Figure: Processing in a neuron. [Hyvarinen et al., 2009]

Information processing in a neuron V

- **Firing rates** of different neurons influence the firing rates of others (either increase/excite or decrease/inhibit them) . . . approximated by the **activation function**.
- **Weights** between neurons model whether they excite or inhibit each other
- **Thresholded** behavior of the action potentials modeled by the **activation** function and **bias**.

Deep Learning in AI

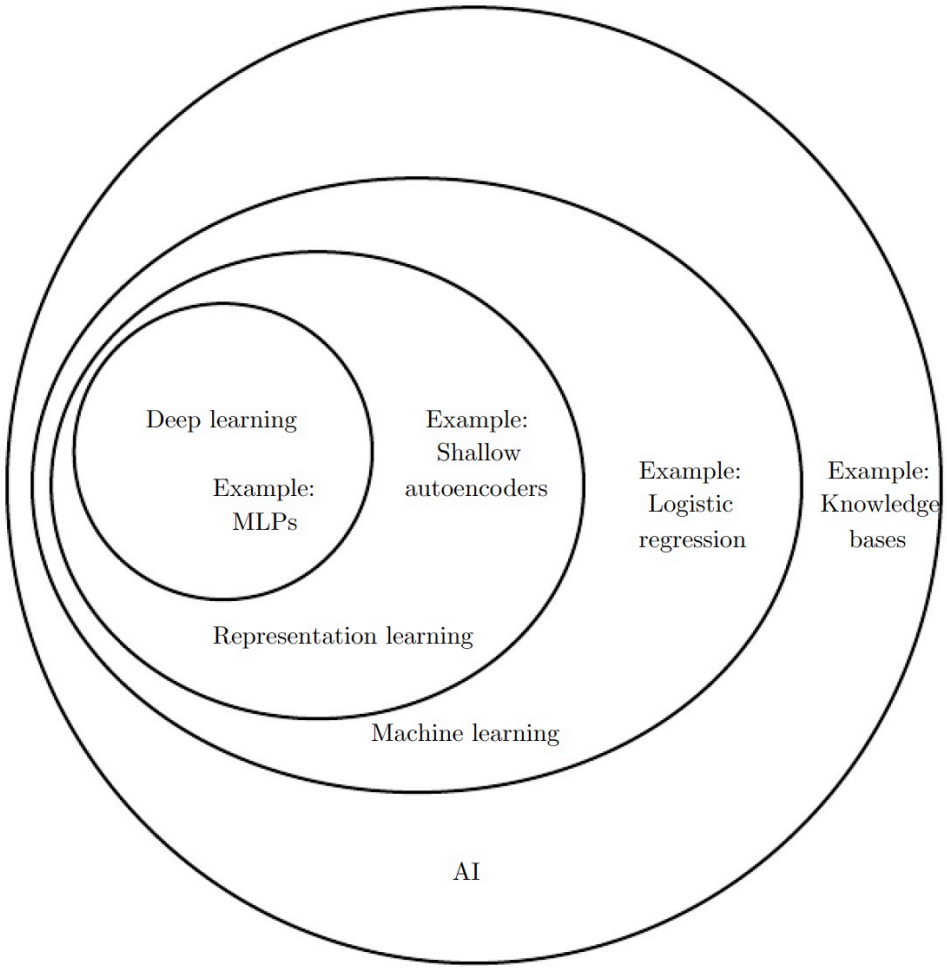


Figure: A Venn diagram showing DL as representation learning, which in turn is a type of ML (which is one approach to AI) [Goodfellow et al., 2016].