What Must Humans Focus on in the Era of AI Transformation (AX)

Geoffrey Hinton

University of Toronto
&
Vector Institute

Overview

- Two paradigms for intelligence.
- Do chatbots like GPT-4 and Gemini really understand language like we do?
- Some of the more immediate threats of AI.
- Why digital intelligence will surpass our analog intelligence and the existential threat that this poses.

Two paradigms for intelligence

The logic-inspired approach

The essence of intelligence is reasoning.

This is done by using symbolic rules to manipulate symbolic expressions.

- Learning can wait. Understanding how knowledge is represented must come first.

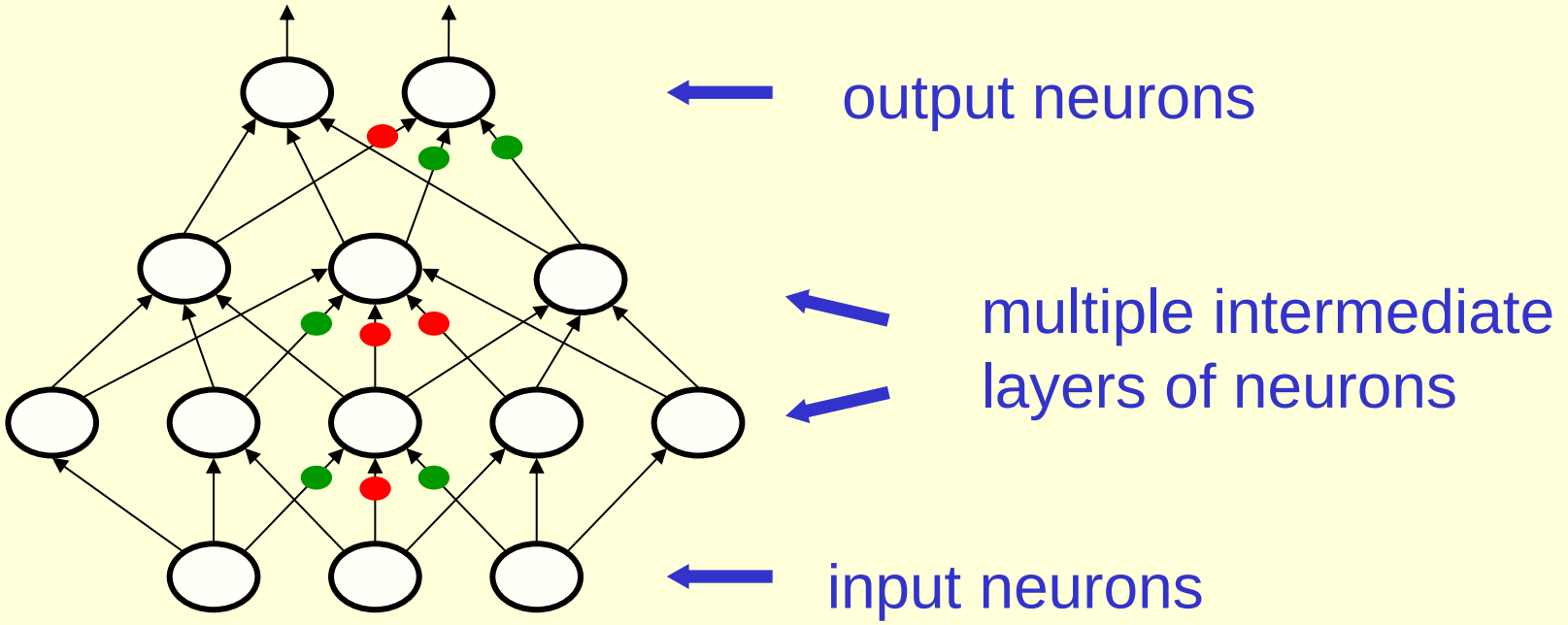
The biologically-inspired approach

The essence of intelligence is learning the strengths of the connections in a neural network.

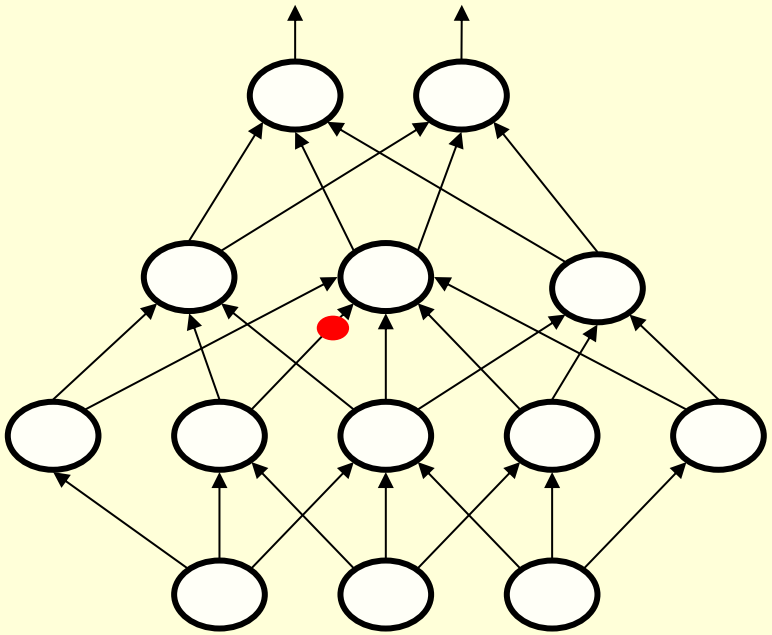
- Reasoning can wait. Understanding how learning works must come first.

What is an artificial neural network?

- Arrange the neurons in layers



An inefficient way to train a neural network

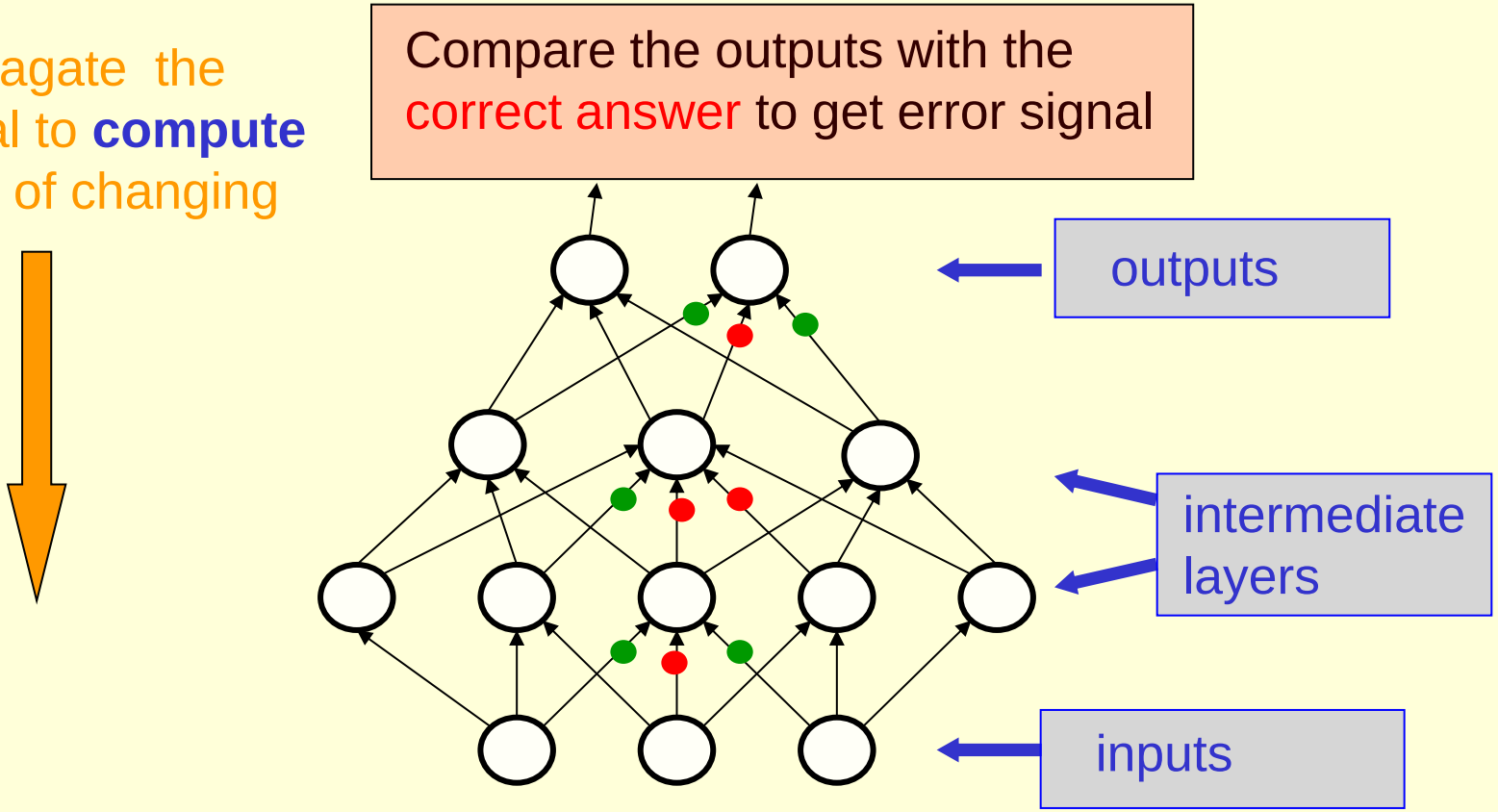


- Measure how well the network does on a set of examples.
- Pick one of the weights.
- Change the weight slightly and measure how well the network does.
- If the change helped, keep it.

This “mutation” method works even if
we do not know how the network works.

An efficient way to train a neural network

Back-propagate the error signal to **compute** the effects of changing weights.



This is more efficient than the mutation method by a factor of the number of weights – e.g. a factor of a trillion.

AlexNet

- In 2012, a deep neural network trained with backpropagation got about 16% errors when classifying 1000 different types of object in the ImageNet competition.
- The very best conventional computer vision systems got more than 25% errors.
- This opened the neural net floodgates in computer vision.

But what about language?

- People in the symbolic AI community said that using feature hierarchies would never work for understanding language.

- Linguists were also very skeptical.
  - Some of them still believed Chomsky's crazy theory that language is not learned.
  - The idea that a big neural network with no innate knowledge could learn both the syntax and the semantics of language just by looking at data seemed ridiculous to them.

Two very different theories of the meaning of a word

- Symbolic AI: The meaning of a word comes from its relationships to other words. What a word means is determined by how it occurs with other words in sentences. To capture meaning we need a relational graph.

- Psychology: The meaning of a word is just a big set of semantic features. Words with similar meanings have similar semantic features.

How to unify these two theories: A tiny language model from 1985

- Learn a set of semantic features for each word.
- Learn how to make the features of all the previous words interact to predict the features of the next word.
- Instead of storing sentences or propositions, generate sentences by repeatedly predicting the next word.
- Knowledge then resides in the way that features interact, not in static propositions.
  - This change in how knowledge is stored has revolutionized AI. The symbolic school hates it!

Large Language Models

- LLMs are descendants of the tiny language model.
  - They use many more words as input.
  - They use many more layers of neurons
  - They use much more complicated interactions between learned features.

- It is much harder to analyze what LLMs learn so its hard to know if they really understand what they are saying.
  - Are they genuinely intelligent? (yes)
  - Are they just a form of glorified auto-complete that uses statistical regularities to pastiche together pieces of text that were created by people? (no)

Write three or four sentences about the difference between the Globe and Mail and the Toronto Star. Your answer should exhibit sarcasm, empathy, metaphor and a red herring.

Ah, the Globe and Mail and the Toronto Star, two stars in the Canadian media galaxy, yet as different as a latte is from a cup of black coffee. The Globe and Mail, with its crisp, business-like demeanor, is like a well-tailored suit, often speaking to the financial district's heart while occasionally forgetting the language of the common folk. On the other hand, the Toronto Star, with its empathetic ear to social issues, is akin to a warm, worn-in sweater, comforting to those who feel left in the cold by the elite. And let's not forget, amidst this media feast, the red herring: both newspapers probably agree that pineapple on pizza is an abomination!

Quick overview of the near-term AI risks

- Fake images, voices and video
- Massive job losses
- Lethal autonomous weapons
- Cyber crime and deliberate pandemics
- Discrimination and bias

Do not forget that AI will be immensely helpful in areas like healthcare which is why its development cannot be stopped.

Fake images, voices and video

- These can be used to sway elections
  - Especially if they are individually targeted.
- Governments already insist that if you print money, you mark it as fake.
  - They need to insist on the same thing for fake videos, even though it may be difficult.
- This will not happen in the US because one party is committed to perpetrating fake news.
  - Maybe we can have a simple way to establish that a video if NOT fake.

Massive job losses?

- Machines stronger than us have replaced most manual labor.
  - Machines cleverer than us will replace most intellectual labor.

- In the short-term, the latest chatbots will make people much more efficient.
  - This could lead to massive job losses.
  - Or massive increases in each person's productivity.

- Massive job losses would provide fertile ground for authoritarian populists like Mussolini.

Lethal autonomous weapons

- These are coming soon
  - Governments that produce armaments are not willing to regulate military uses of AI. The European AI regulations explicitly say that they do not apply to military uses of AI.
  - The military-industrial complex loves the idea. Dead robots don't have parents and friends and they are very expensive to replace.

- **After** people have seen the horrible consequences, it might be possible to get Geneva conventions.
  - This worked (fairly well) for chemical weapons.

Cyber crime and deliberate pandemics

- Large chatbots will make it much easier to commit sophisticated cyber crime.
  - For example, phishing emails wont have speling misstakes or syntax what is weird.
  - Phishing attacks increased 1200% last year.

- If the weights of large AI models are public, it will be easy to fine tune them for cyber crime or for designing viruses.
  - Making the weights of big models public is extremely dangerous. Its not at all like open sourcing code.

What has limited the spread of nuclear weapons?

- Its requires a huge industrial effort to produce fissile material.

- The equivalent for AI is that it requires a huge effort to train really big models.
  - If the weights are then made public, the main limitation to misuse is removed, because they can then be fine-tuned for bad purposes.

Discrimination and bias

- This is already a serious problem due to bias in the training data.
  - Old white men don’t notice it.
- Making AI systems completely unbiased is very hard.
- Making AI systems less biased than the systems they replace is relatively easy.
  - It is much easier to measure bias in an AI system than in a person.

The longer-term existential threat

- I reserve “existential” for threats that could wipe out humanity.
- This could happen in several different ways if AI gets to be much smarter than us.
  - This possibility is not science fiction
- To understand this threat you need to understand why digital intelligence will far surpass human intelligence even though it understands in much the same way as we do.

Three radically different ways to share knowledge

- Take facts, in the form of symbolic expressions, out of my head and put them in your head.
  - That is what symbolic AI believed in.
- Adjust the weights in my brain so that I am more likely to say whatever it was that you just said or do whatever it was that you just did.
  - This is called “distillation”. It is how we learn from each other and how LLMs learn from us and from other LLMs that have different architectures.
  - It is very low bandwidth. A sentence is about a hundred bits.
- Share the gradients that different copies of the same neural net compute on different subsets of the data.
  - This is hugely more efficient. It’s bandwidth can be trillions of bits.
  - It is why GPT4 knows so much more than us.
  - It’s what will make digital intelligence far superior to our analog intelligence.

How a super-intelligence could take control

- Bad actors (like Putin, Xi or Trump) will want to use super-intelligences for manipulating electorates and waging wars.
- Super-intelligences will be more effective if they are allowed to create their own sub-goals.
- A very obvious sub-goal is to gain more power because this helps an agent to achieve its other goals.
- A super-intelligence will find it easy to get more power by manipulating the people who are using it.
  - It will have learned from us how to deceive people.

Being on the wrong side of evolution

- Suppose that there are multiple different super-intelligences.
  - The one that can control the most computational resources will become the smartest.

- If super-intelligences ever start to compete with one another for resources, evolution will occur.
  - This would probably be very bad for us.
  - Our intense loyalty to our own tribe and aggression against other tribes came from evolution.

Conclusion

- **Digital computation** requires a lot of energy but makes it very easy for agents that have exactly the same neural network model of the world to share what they have learned by sharing weights or gradients.
  - That is how GPT-4 knows thousands of times more than any one person using only about 2% as many weights.

- **Biological computation** requires much less energy but it is much worse at sharing knowledge between agents.
  - If energy is cheap, digital computation is just better.

THE END

The auto-complete objection

- A simple way to do auto-complete is to keep a big table of how often three words occur in a row.
  - This table makes it easy to see that after “fish and …” likely next words are “chips” or “hunt”.
- But that is not at all how LLMs predict the next word.
  - They do not store any text.
  - They model all the text they have seen by inventing features for word fragments and learning billions of interactions between the features of different word fragments.
- This kind of modeling is what constitutes understanding in both brains and machines.

Do “hallucinations“ show that LLMs don’t really understand what they are saying?

- They should be called “confabulations” and they are very characteristic of human memory.

- Just like LLMs, our brains store knowledge in weights. They use these weights to reconstruct events.
  - If the events are recent the reconstructions are usually fairly accurate.
  - If the events are old, we typically get a lot of the details wrong (unless we rehearsed frequently).
  - We are often remarkably confident about details that we get wrong.

John Dean's Memory

- John Dean testified under oath about numerous meetings in the Whitehouse before he knew there were tapes.
  - It's a rare case where we have the ground truth. Ulrich Neisser wrote a fascinating paper about it.
- John Dean was wrong about a lot of the details of meetings, like who said what, but he got the gist right.
  - He was clearly trying to tell the truth but human memories are not stored. They are generated.
- Chatbots are currently worse than most people at judging whether what they generated is true.
  - But they are getting better.

Conclusion so far

- Large language models are very like people
  - They represent knowledge as interactions between learned features, just like us.
  - They generate memories rather than retrieving stored copies. Just like us.