Deep Learning: history and modernity

Andrey Kutuzov
Language Technology Group
University of Oslo

IN3050/4050
11 April 2024

Contents

1 Deep learning: why are we here

- 1. Increased compute
- 2. Increased data
- 3. Better architectures: transformers

2 AI hype

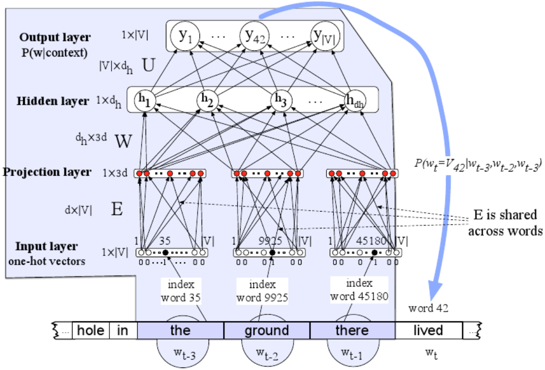
Deep learning: why are we here

Most of modern machine learning is done with multi-layered artificial neural networks.

Deep learning: why are we here

Most of modern machine learning is done with multi-layered artificial neural networks.

- First artificial neural networks: 1950s
- First really working neural language model in [Bengio et al., 2003]
  - feed-forward neural network architecture



(image from Jurafsky and Martin, 2023)

Deep learning: why are we here

- The same Yoshua Bengio who received the 2018 ACM A.M. Turing Award 'for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing'.
- Together with Geoffrey Hinton and Yann LeCun
- This signalled the end to the long 'AI Winter'.
- `https://awards.acm.org/about/2018-turing`

Deep learning: why are we here

- The same Yoshua Bengio who received the 2018 ACM A.M. Turing Award 'for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing'.
- Together with Geoffrey Hinton and Yann LeCun
- This signalled the end to the long 'AI Winter'.
- `https://awards.acm.org/about/2018-turing`

So, what made deep learning efficient in real-world applications?

1. Increased compute

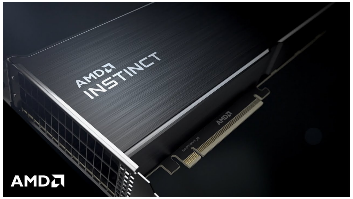
- Hardware capabilities are growing: graphic processing units (GPUs) and Tensor Processing Units (TPUs). They excel in parallelized matrix multiplication.

1. Increased compute

- Hardware capabilities are growing: graphic processing units (GPUs) and Tensor Processing Units (TPUs). They excel in parallelized matrix multiplication.
- But note compute divide: not everyone can afford burning 100K GPU/hours to train a GPT-7B model for a mid-sized language.

1. Increased compute

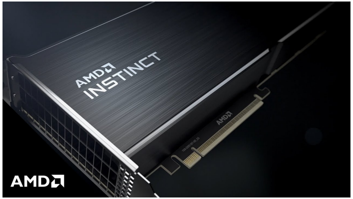
- Hardware capabilities are growing: graphic processing units (GPUs) and Tensor Processing Units (TPUs). They excel in parallelized matrix multiplication.
- But note compute divide: not everyone can afford burning 100K GPU/hours to train a GPT-7B model for a mid-sized language.



- Publicly funded science is important! Norway has access to LUMI:
  - 5th most powerful supercomputer in the world, 1st in Europe
  - 2978 compute nodes with AMD MI250X GPUs (24 000 GPUs in total)
- `https://www.lumi-supercomputer.eu/`

1. Increased compute

- Hardware capabilities are growing: graphic processing units (GPUs) and Tensor Processing Units (TPUs). They excel in parallelized matrix multiplication.
- But note compute divide: not everyone can afford burning 100K GPU/hours to train a GPT-7B model for a mid-sized language.



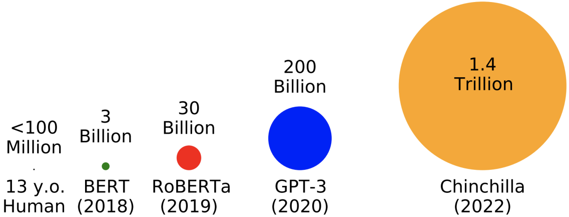
- Publicly funded science is important! Norway has access to LUMI:
  - 5th most powerful supercomputer in the world, 1st in Europe
  - 2978 compute nodes with AMD MI250X GPUs (24 000 GPUs in total)
- https://www.lumi-supercomputer.eu/

IFI Language Technology Group uses LUMI to train open language models for English and Norwegian much faster than before

2. Increased data

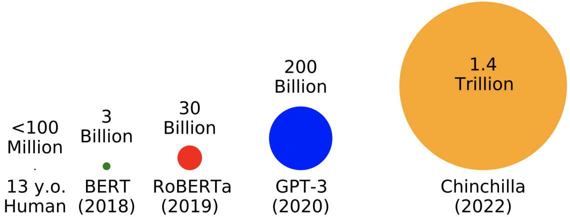
Machine learning models are trained on large datasets: for language models, they are mostly crawled from the Internet (most of it in English).

Training dataset sizes for some famous language models in running words:



2. Increased data

Machine learning models are trained on large datasets: for language models, they are mostly crawled from the Internet (most of it in English).
Training dataset sizes for some famous language models in running words:



- ChatGPT? Size of the training data unknown (but a mix of texts and code).
- Not all languages are equal in the size of available data.
- For Norwegian: not more than 50 billion words publicly available.

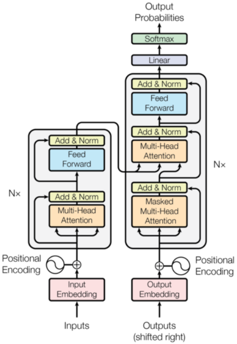
3. Better architectures: transformers

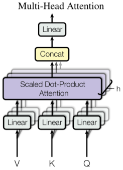
Transformer

- A sequence of feed-forward layers
- multi-headed self-attention
  - model learns what elements in the input sequence to pay attention to during training
  - all input elements are processed simultaneously
  - training easily parallelized across multiple computation units (unlike recurrent neural networks)
  - many heads: solves the under-parameterization problem, different heads excel in different tasks

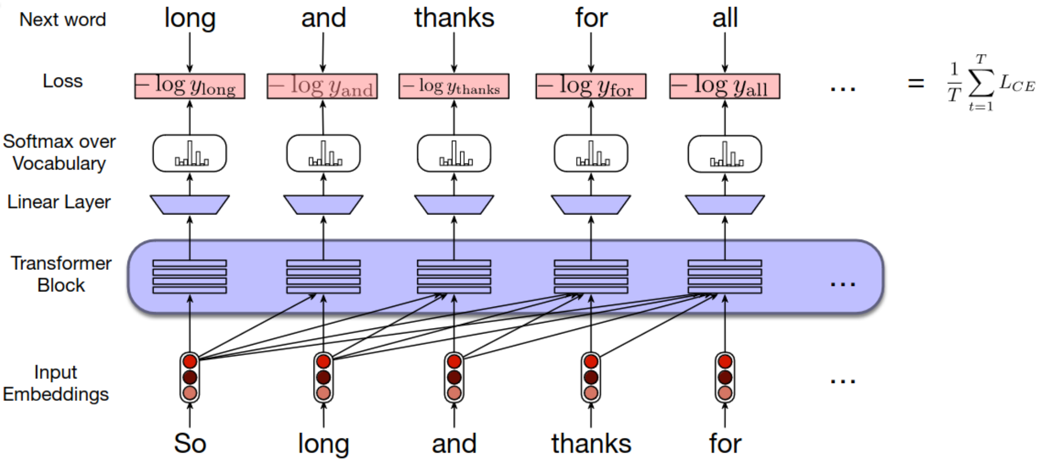
Transformers allowed to use the existing data and compute in the most optimal way.

Learn more in the IN5550 Master course :-)





Transformer as a language model



(image from Jurafsky and Martin, 2023)

Contents

1. Deep learning: why are we here
    - 1. Increased compute
    - 2. Increased data
    - 3. Better architectures: transformers

2. AI hype

AI hype

- In 2023, the same Geoffrey Hinton left his Google job
- ...and focused on talking about the dangers of AI
- `https://edition.cnn.com/2023/05/01/tech/geoffrey-hinton-leaves-google-ai-fears/index.html`
- ...more on that in later lectures.

AI hype

- In 2023, the same Geoffrey Hinton left his Google job
- ...and focused on talking about the dangers of AI
- `https://edition.cnn.com/2023/05/01/tech/geoffrey-hinton-leaves-google-ai-fears/index.html`
- ...more on that in later lectures.

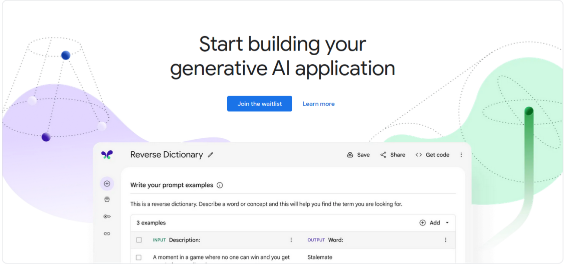
This was mostly because of recent advances in large language models as chatbots.

AI hype

(ChatGPT, a generative language model by OpenAI)
`https://openai.com/blog/chatgpt/`

AI hype

(ChatGPT, a generative language model by OpenAI)
`https://openai.com/blog/chatgpt/`



Start building your generative AI application

Reverse Dictionary

Write your prompt examples

This is a reverse dictionary. Describe a word or concept and this will help you find the term you are looking for.

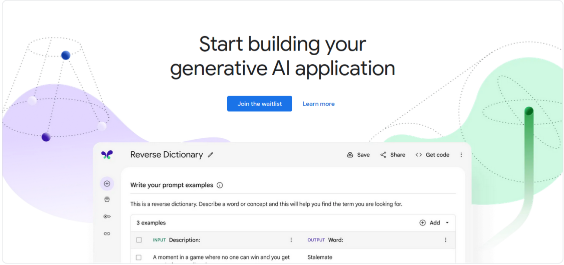
3 examples

	INPUT Description:	OUTPUT Word:
□	A moment in a game where no one can win and you get	Stalemate

(PaLM 2, a generative language model announced by Google in May 2023)
`https://developers.generativeai.google/products/palm`

AI hype

(ChatGPT, a generative language model by OpenAI)
`https://openai.com/blog/chatgpt/`



Start building your generative AI application

Write your prompt examples

This is a reverse dictionary. Describe a word or concept and this will help you find the term you are looking for.

3 examples

□	INPUT Description:	OUTPUT Word:
□	A moment in a game where no one can win and you get	Stalemate

(PaLM 2, a generative language model announced by Google in May 2023)
`https://developers.generativeai.google/products/palm`

Language models are trained to predict the next word. But...

AI hype

Any language model is a text generator by definition

AI hype

Any language model is a text generator by definition

Autoregressive or causal generation:

- feed a word or a sentence (prompt) into the LM
- get a probability distribution over what words are likely to come next
- pick the most probable word from this distribution (or use some form of sampling)
- feed it right back in the LM together with the previous words
- repeat this process and you're generating text!

Slightly rephrasing `https://karpathy.github.io/2015/05/21/rnn-effectiveness/`

AI hype

Any language model is a text generator by definition

Autoregressive or causal generation:

- feed a word or a sentence (prompt) into the LM
- get a probability distribution over what words are likely to come next
- pick the most probable word from this distribution (or use some form of sampling)
- feed it right back in the LM together with the previous words
- repeat this process and you're generating text!

Slightly rephrasing `https://karpathy.github.io/2015/05/21/rnn-effectiveness/`

This is what ChatGPT or GPT-4 do. Thus, generative language models.
Text generation is not the only task LMs can do, but it pushed them into the headlines.

AI hype

Many of the popular LLMs are closed and only available to the public as black-box services.
Some open language models for Norwegian:

- `https://huggingface.co/norallm`
- not specifically aimed to be used as chat-bots, but you still can play with them as such:
- `https://huggingface.co/spaces/ltg/chat-nort5`

References I

Bengio, Y., Ducharme, R., and Vincent, P. (2003).
A neural probabilistic language model.
Journal of Machine Learning Research, 3:1137–1155.