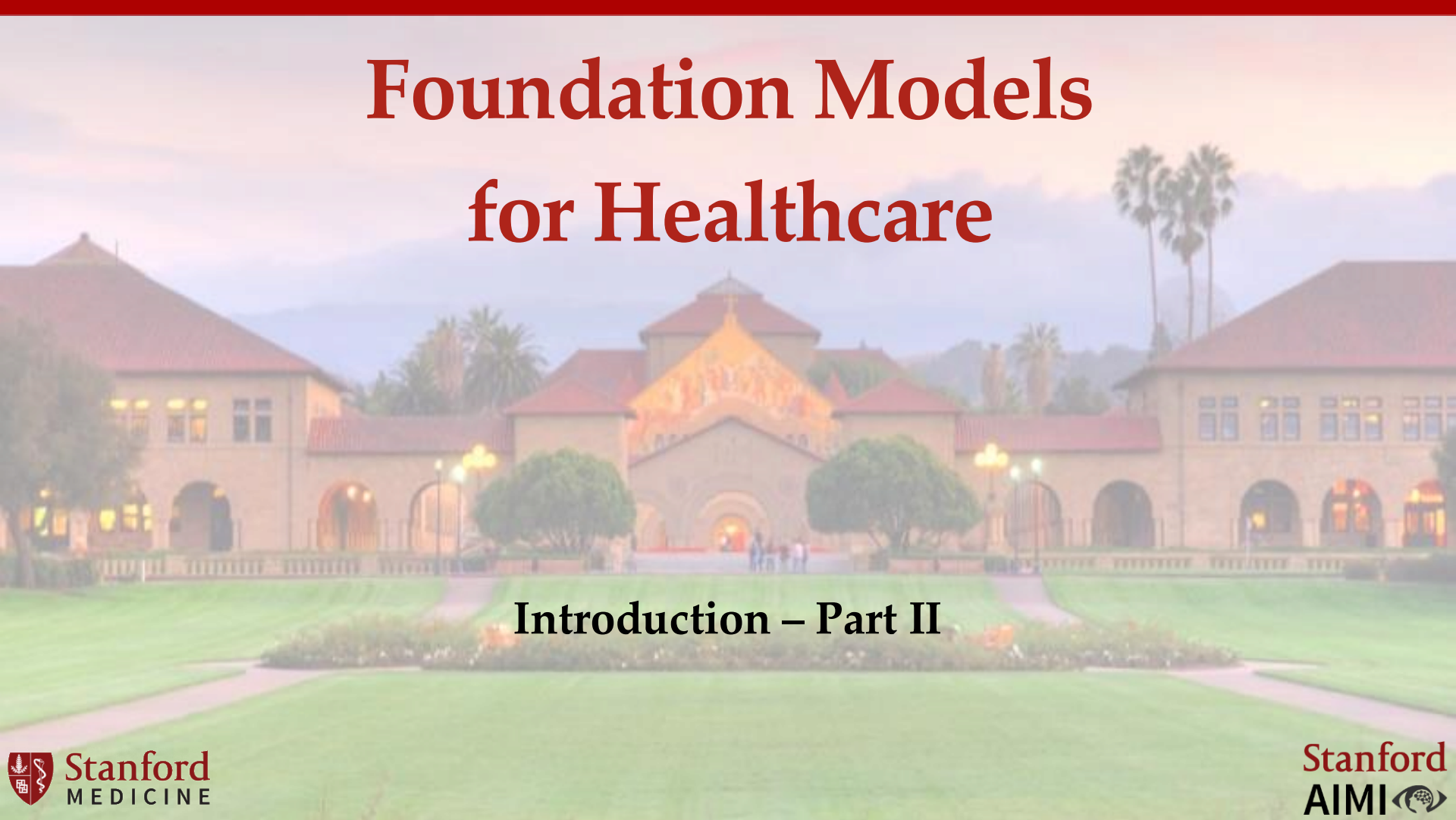
Foundation Models for Healthcare

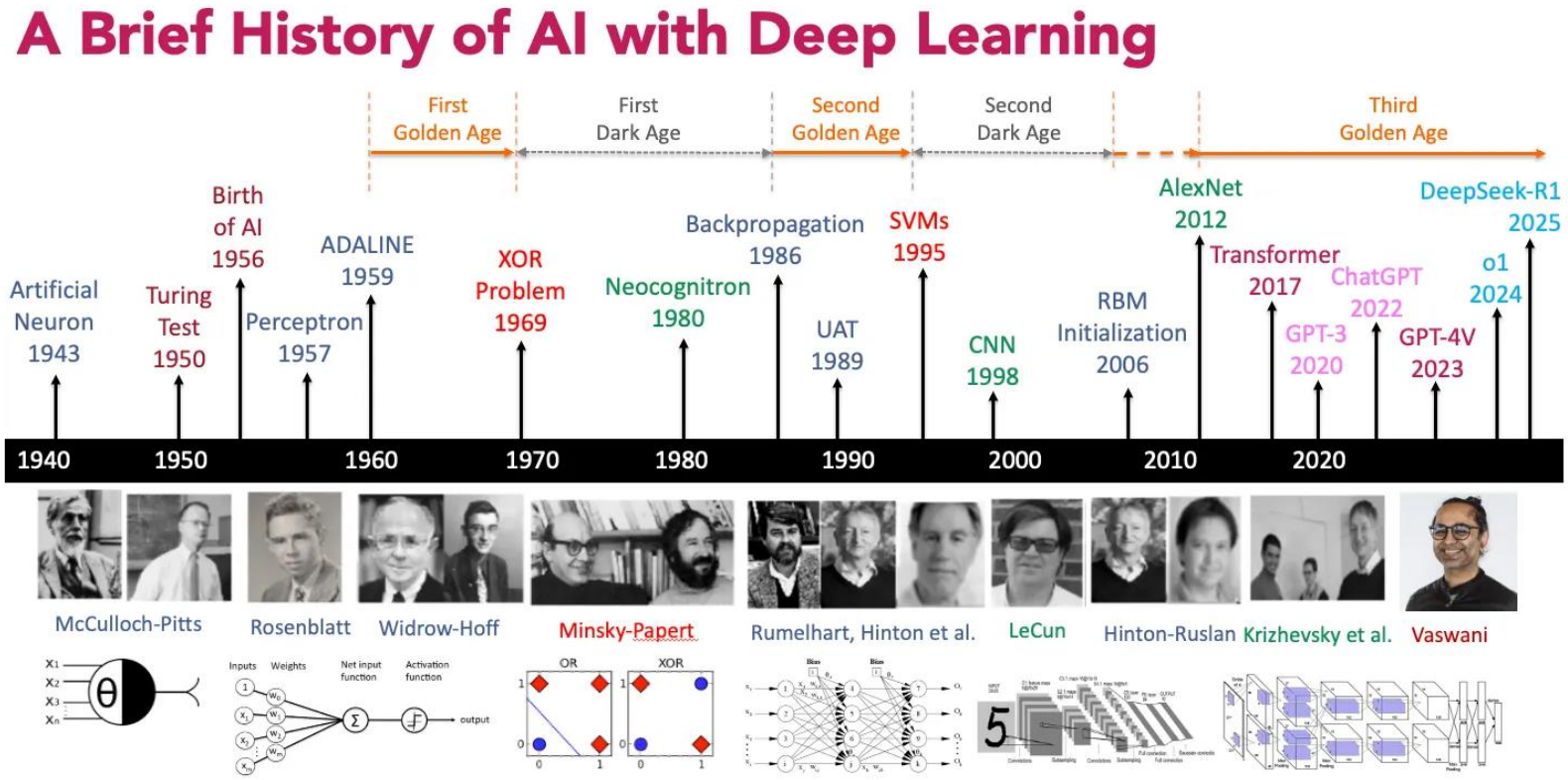


Introduction – Part II

Stanford MEDICINE

Stanford AIMI

Historically

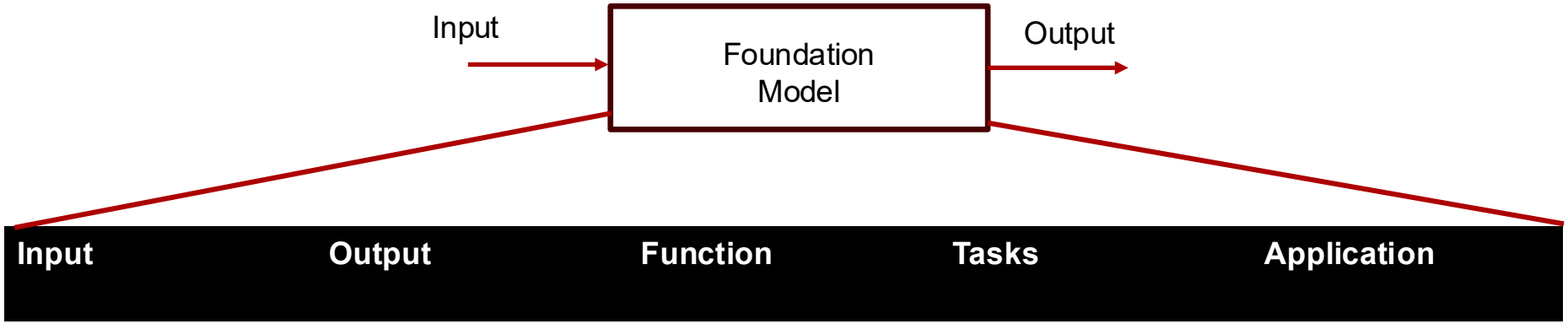


https://medium.com/@lmpo/a-brief-history-of-ai-with-deep-learning-26f7948bc87b

ML Models

Input	Output	Function	Tasks	Application
• Structured • Time series • Text • Image • Audio • Video • Table • Domain-specific • Multi-D • Multimodal	• Label • Regression • Text • Image • Audio • Video • Table • Domain-specific	• Clustering • Classification • Prediction • Regression • Synthesis	• Recognition • Detection • Segmentation • Captioning • Image/text/synthesis • Audio/Video/multimodal synthesis • Question answering • Text generation • Autocompletion • Translation • Summarization • Sentiment analysis • Navigation • Search/IR • Recommendation •	• Customer service • Social Media • Marketing • Code generation • Automated reporting • Domain-specific

Foundation Model

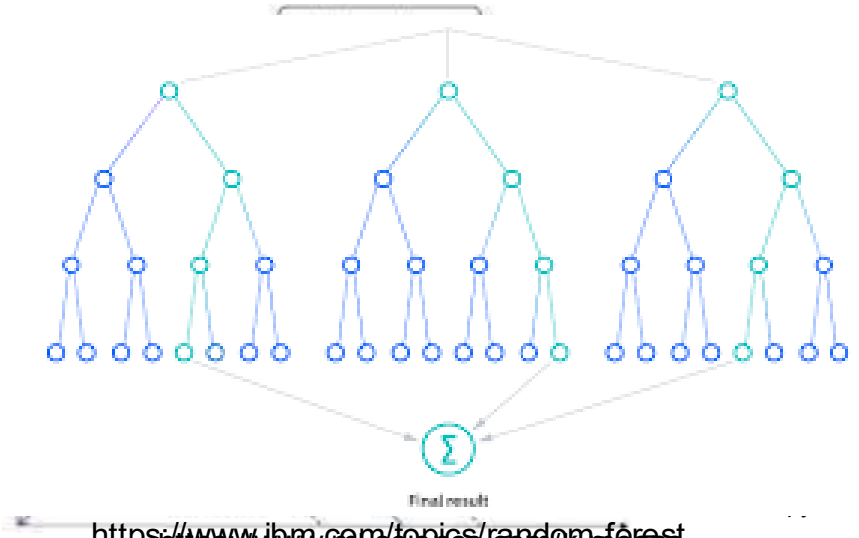


Input	Output	Function	Tasks	Application

• Trained on a variety of data input	• Different types of output based on downstream use	• Defined based on the scope of ability, range of uses, breadth of tasks or types of output and input
		• Encoder
		• Decoder
		• Encoder-Decoder
			• A single model can accomplish multiple tasks:
			  • Sentence completion
			  • Sentiment classification
			  • Summarization
			  • ….
			• Works for arbitrary input
			• Can be fine-tuned
				• Same model can serve multiple applications
				  • Customer service
				  • Code generation
				  • …

Are these Foundational Models?

- Decision Trees
- Clustering
- Support Vector Machines
- Random Forests



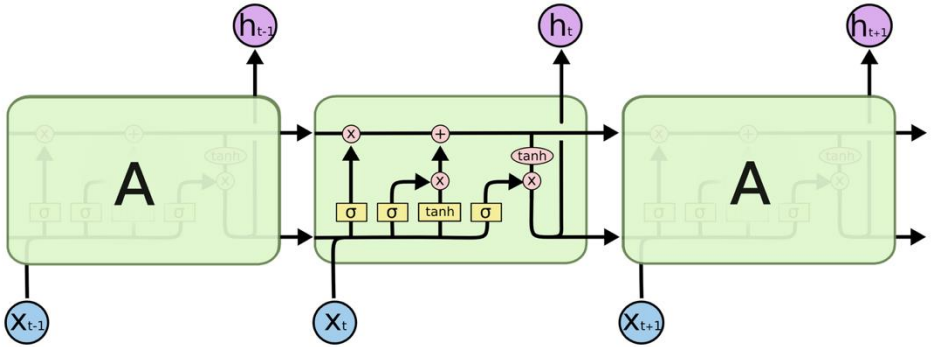
https://www.ibm.com/topics/random-forest

https://www.sciencedirect.com/topics/computer-science/support-vector-machine

clustering

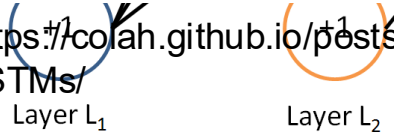
Are these Foundational Models?

- Multi-layer perceptron
- RNN
- LSTM



The repeating module in an LSTM contains four interacting layers.

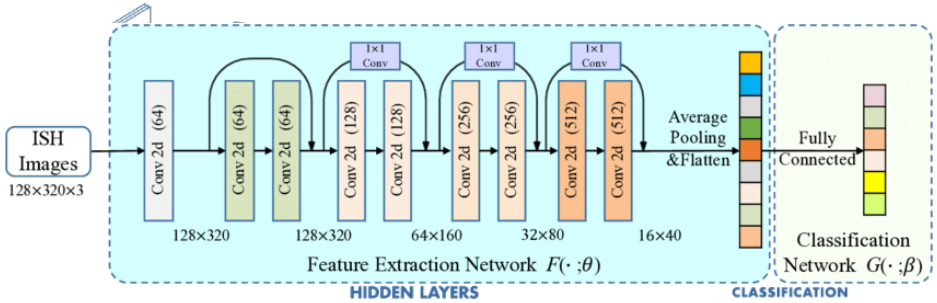
https://colah.github.io/posts/2015-08-Understanding-LSTMs/



http://deeplearning.stanford.edu/tutorial/supervised/MultiLayerNeuralNetworks/

Are these Foundational Models?

- Convolutional neural networks
- ResNet
- DNN

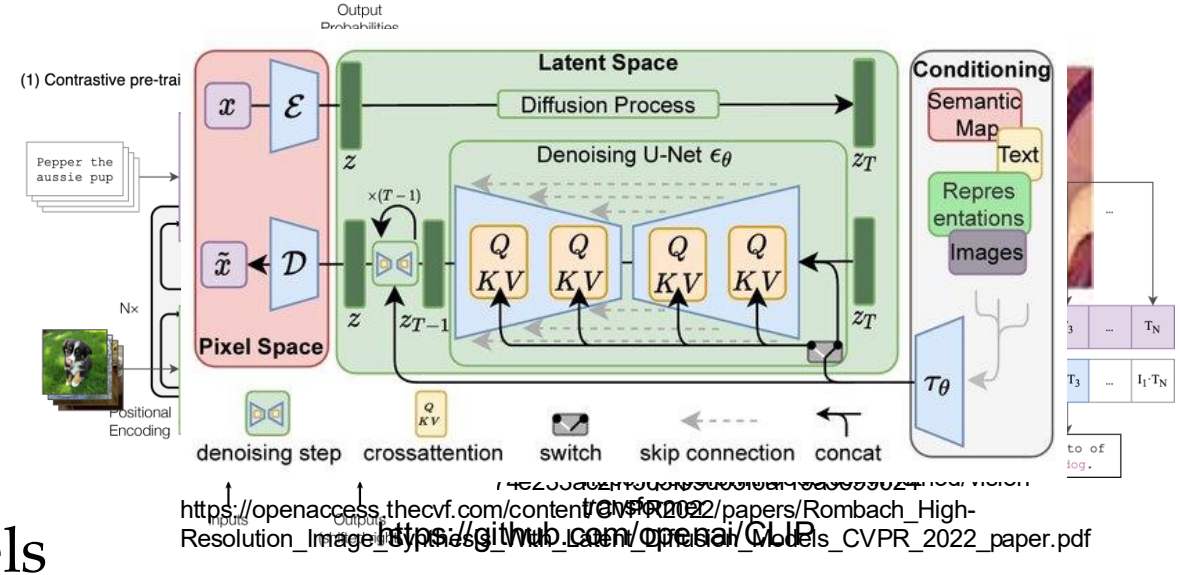


https://www.linkedin.com/pulse/deep-residual-networks-resnet-ayoub-kirouane/

https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939

Are these Foundational Models?

- Auto-encoders
- GANs
- Transformers
- GPT
- CLIP
- Diffusion models



https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf

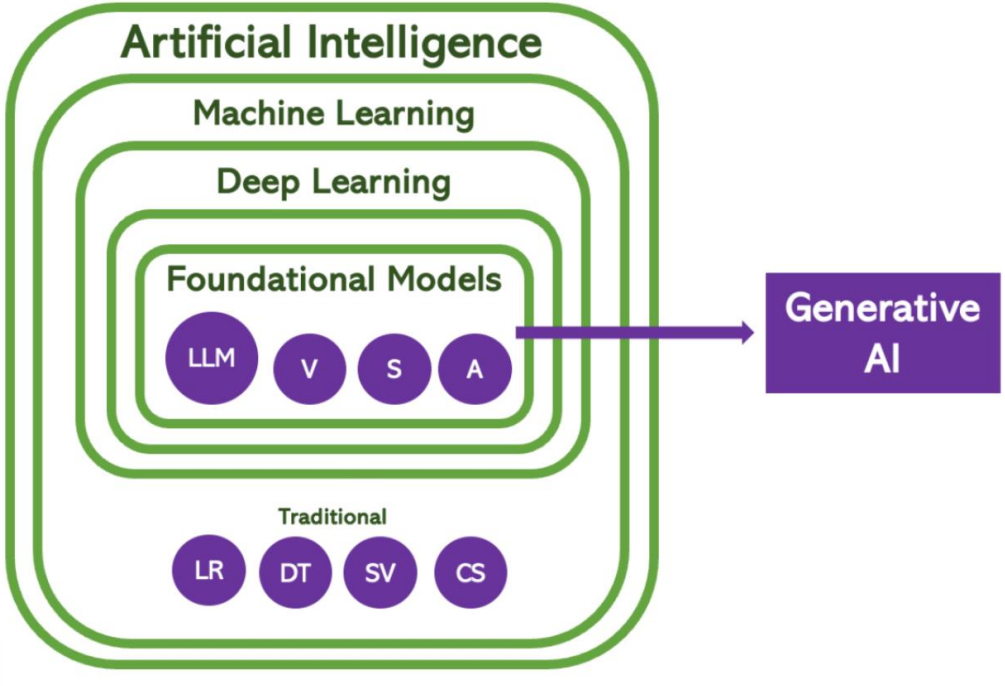
https://github.com/openai/CLIP

https://arxiv.org/abs/1706.03762

How has the landscape changed?

Traditional ML	Foundational models
Engineered features	Auto-derived features
Novel architectures but are they useful now?	Mostly transformer variants.
Multiple independent instances of the product	Single instance adapted for use cases through fine-tuning
Trained with smaller, hand-selected and supervised datasets	Trained with self-supervised on very large datasets
Trained for a specific purpose, often classification.	Trained once and applied to different use cases.
Less compute resources	Power hungry GPU-dependent
Deployment was simple as library calls	Full ML Ops behind training, fine-tuning and deployment.
Humans focused on design and inventions of architectures, less focus on data	More focus on data since architectures are standardized. Most improvements are practical with less radical changes.
Less focus on data collection	A whole industry of annotators, alignment, and LLM training data collection
Custom effort for every instance	Less ongoing maintenance after development
Fewer parameters	Large number of parameters
More domain specificity, less generalizability	Lack of domain specificity
Risk of overfitting	Dataset bias

Evolution of ML models



https://www.linkedin.com/pulse/ai-alphabet-soup-from-buzzwords-brilliance-brian-gruttadauria-fxtie/

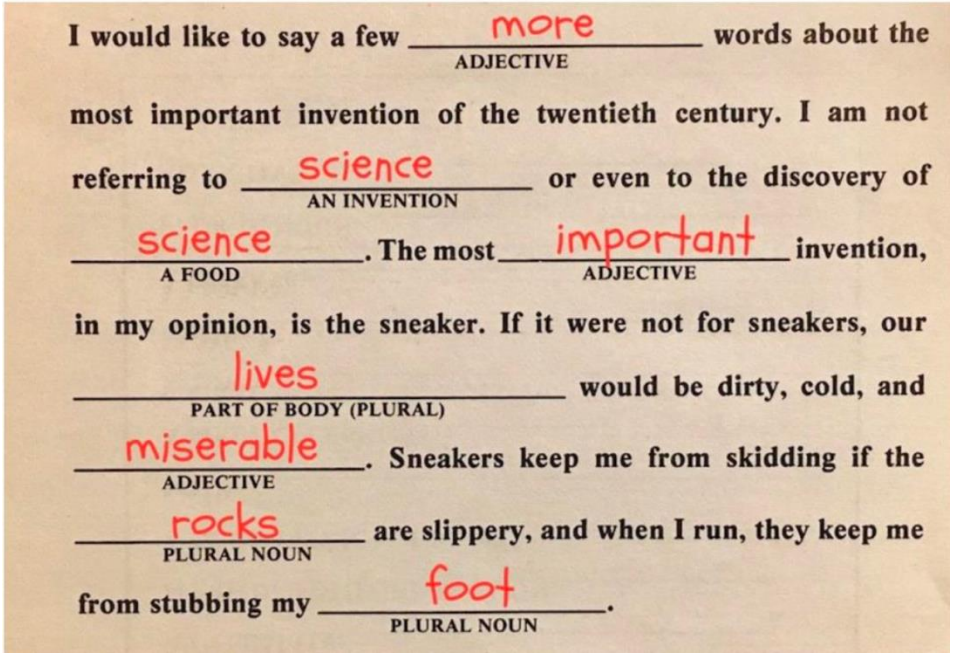
What is the general field headed?

- Foundational models still have limitations
  - Not truly domain independent despite the large knowledge given
  - Reasoning powers are still limited.
  - Largely unsuccessful in healthcare world
- New thinking is needed for AI to make a larger impact.
  - Domain understanding
    - New datasets interdisciplinary emerging
  - Domain reasoning
  - New architectures
    - From just scaling models to architectural efficiency, better compute/memory, and flexible designs for innovation
      - Best Paper at NeurIPs - gating mechanisms in transformer-based LLMs to improve efficiency and stability (bringing back LSTM ideas into transformers)
- More from larger to smaller models
- Ethical concerns still looming
- Real cognition still far away

Deep dive into some of the basics

- Guage understanding of the fundamentals
- Demos

Masked language prediction – A key element of foundational modeling

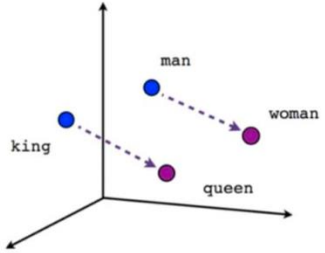


I would like to say a few ___more___ (ADJECTIVE) words about the most important invention of the twentieth century. I am not referring to ___science___ (AN INVENTION) or even to the discovery of ___science___ (A FOOD). The most ___important___ (ADJECTIVE) invention, in my opinion, is the sneaker. If it were not for sneakers, our ___lives___ (PART OF BODY (PLURAL)) would be dirty, cold, and ___miserable___ (ADJECTIVE). Sneakers keep me from skidding if the ___rocks___ (PLURAL NOUN) are slippery, and when I run, they keep me from stubbing my ___foot___ (PLURAL NOUN).

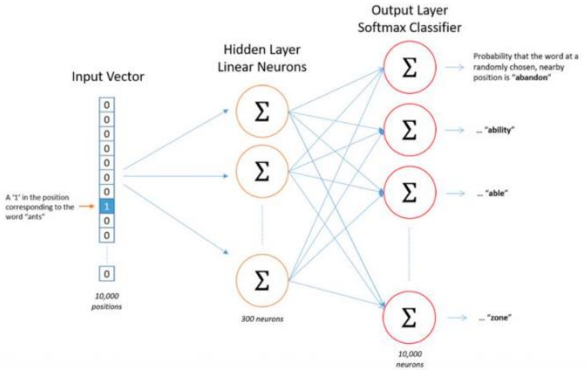
(source: Medium article 2018)

Word2Vec model

- From Words to numeric vectors
- 2-layer NN
- Input : one-hot vector of vocabulary
- Hidden layer : A vector per vocabulary word serving as the embedding
- Output: predicted probabilities of seeing another word within distance k from the input word
- Captures both syntactic and semantic similarity

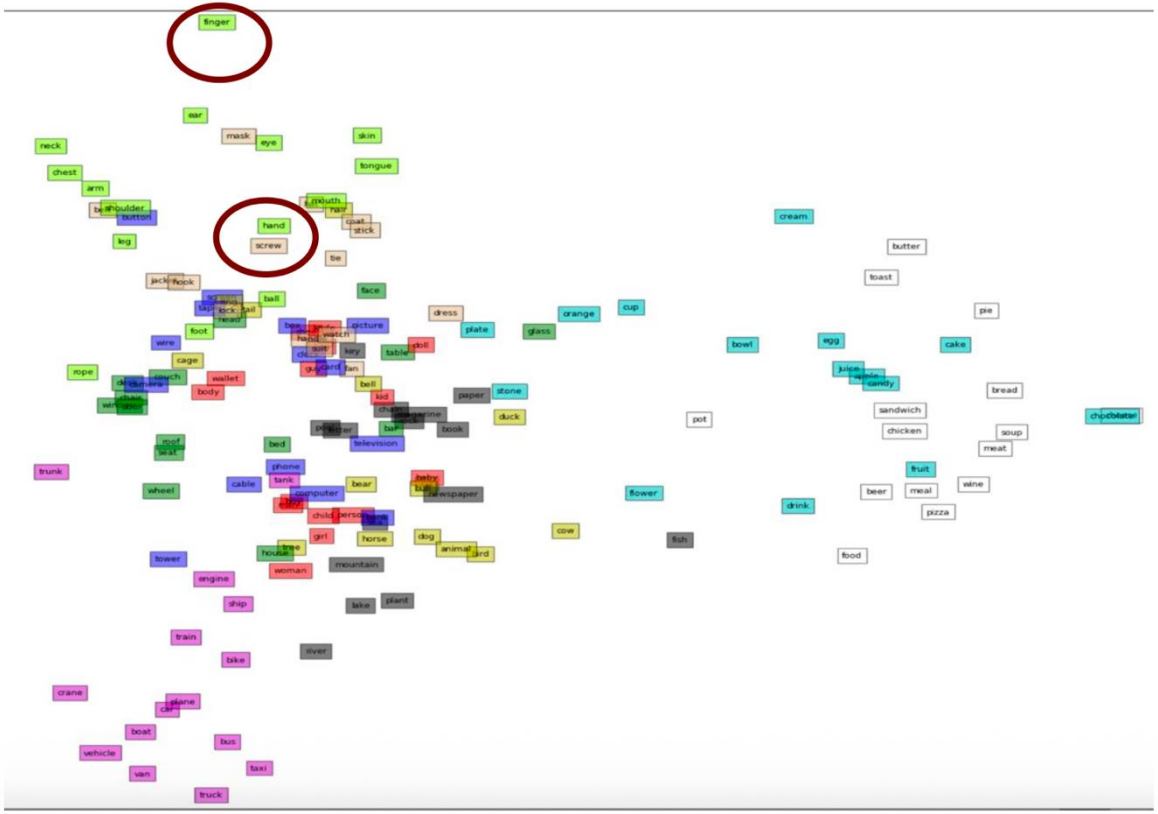


Demo



Understanding the neural network training of Word2Vec model

How good are these representations?



Transformers

- Your kitchen sink of good ideas all rolled into one architecture
  - Tokenizing the input text
  - Positional encoding of tokens to preserve order
  - Attention – the Q,K,V paradigm
  - Multi-head attention – for more combinations to explore
  - Multi-layer encoders for increased abstractions
  - Linear layers for aggregating it all together
  - Explores this in parallel for all tokens unlike RNN
  - Representations created more powerful than CNNs!

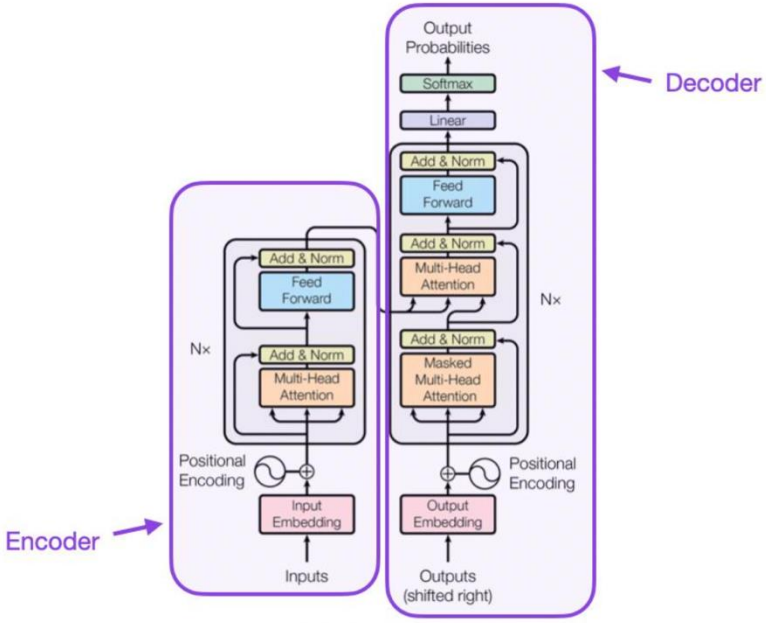


Figure 1: The Transformer - model architecture.

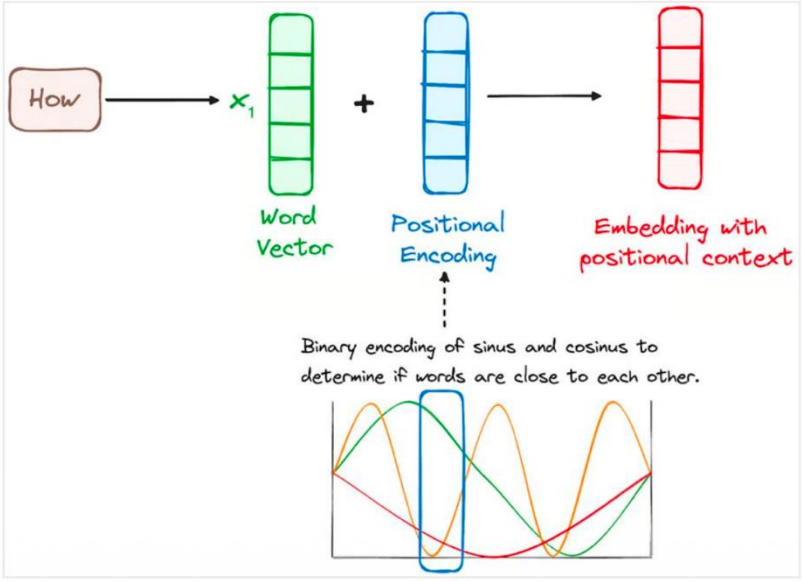
https://www.linkedin.com/pulse/llm-transformer-architecture-shivasish-mahapatra-kj9qf/

Tokenizing the input

- Word2Vec encoding of base vocabulary
  - 30,000+ words
- Word-piece tokenization for handling out-of-vocabulary
- E.g. A sequence of numbers or letters that opens a combination lock
  - ['<s>', 'a', 'Ġsequence', 'Ġof', 'Ġnumbers', 'Ġor', 'Ġletters', 'Ġthat', 'Ġopens', 'Ġa', 'Ġcombination', 'Ġlock', '</s>']
- This combination of terms makes sense to me to describe this beautifying phenomenon .
  - ['<s>', 'This', 'Ġcombination', 'Ġof', 'Ġterms', 'Ġmakes', 'Ġsense', 'Ġto', 'Ġme', 'Ġto', 'Ġdescribe', 'Ġthis', 'Ġbeaut', 'ifying', 'Ġphenomenon', 'Ġ.', '</s>']
- Tokens may be non-sensical from language perspective

Capturing sequence information

- Allows for sequence and nearness of words.
- But is additive operation the only possibility?



https://www.datacamp.com/tutorial/how-transformers-work

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}})$$

$$\begin{aligned} f(a+b) &= \begin{bmatrix} \cos(a+b) \\ \sin(a+b) \end{bmatrix} \\ &= \begin{bmatrix} \cos a \cos b - \sin a \sin b \\ \cos a \sin b + \sin a \cos b \end{bmatrix} \\ &= \begin{bmatrix} \cos b & -\sin b \\ \sin b & \cos b \end{bmatrix} \begin{bmatrix} \cos a \\ \sin a \end{bmatrix} \\ &= R(b) f(a) \end{aligned}$$

Why sinusoidal?

where $R(b)$ is the rotation matrix for angle b. A similar identity holds if $f(a)$ is replaced with $f(\omega a)$ for any $\omega \in \mathbb{R}$.

In contrast, suppose f is the binary encoding:

$$f(a) = \begin{bmatrix} a \bmod 2 \\ \lfloor a \div 2 \rfloor \bmod 2 \\ \lfloor a \div 4 \rfloor \bmod 2 \\ \lfloor a \div 8 \rfloor \bmod 2 \\ \cdots \end{bmatrix}$$

Then $f(a+1)$ is not a linear function of $f(a)$.

Attention

- What is (Q,K,V)
  - From Information retrieval literature
    - Key is the index or pointer to the content, q=query, V=actual value stored
    - Also a model for human memory retrieval
  - In the context of sequence modeling:
    - Capturing relations between words.
      - Given a word Q, and its neighbors x-steps away K, can you predict the likelihood of a word V?
      - In Word2Vec we did a fixed K (e.g. immediate neighbor)
      - More general masked language modeling

Attentional computing

- Predominantly based on modeling context by exploring positional relationships between words in sentences

e.g. Hello, how are you ?

Vocabulary

V = {"Hello", ",","how", "are", "you", "?"}

Hello , how are you ?

Hello	,	how	are	you	?
0.23	0.13	0.20	0.23	0.23	0.11
0.45	0.15	0.35	0.55	0.45	0.25
0	1.0	0.4	0.7	0.30	0
1.0	0.34	0	1.0	0.10	0.0
0.45	0.14	0.1	0.45	0.15	0.41
0.67	0.27	0.17	0.27	0.37	0.27

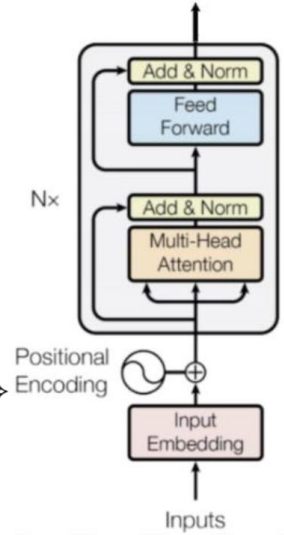
More complex relationships and context modeling

Hello , how are you ?

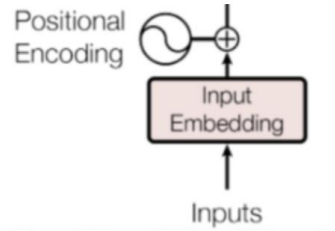
$$Attention(q, k, v) = \underbrace{softmax\left(\frac{qk^T}{\sqrt{d_k}}\right)}_{\text{Attention weights}} v$$

from (q) to (k); vector dimensionality of K, V (d_k)

	Hello	,	how	are	you	?
Hello	78.49	43.29	1.2	41.74	91.43	74.47
,	95.84	28.78	57.13	68.20	−60.94	26.85
how	−95.69	−52.16	17.00	45.71	48.49	64.35
are	−69.92	85.16	94.94	91.04	−92.83	77.49
you	65.85	55.85	62.54	−97.46	76.38	13.20
?	−30.05	−4.52	76.02	42.35	15.29	63.61



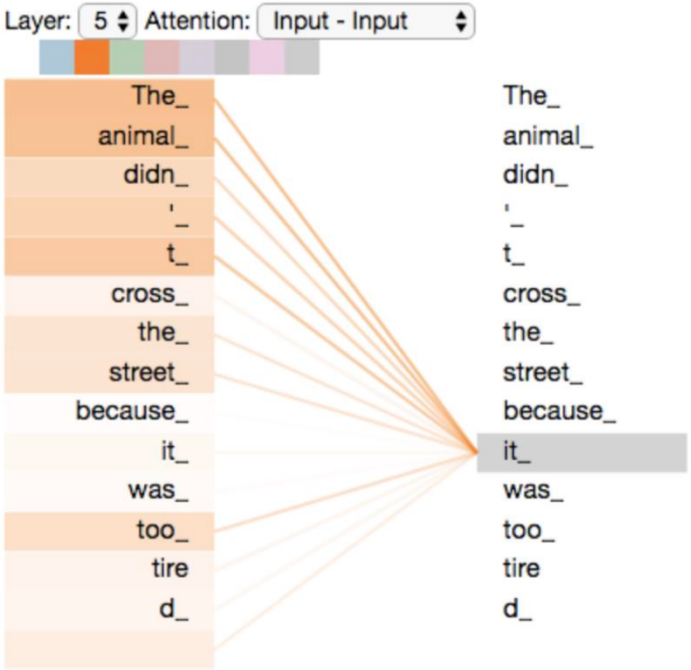
	<	−	d_{emb_dim}	−	>
Hello	123.4	0.32	⋯	94	32
,	83	34	⋯	77	19
how	0.2	50	⋯	33	30
are	289	432.98	⋯	150	92
you	80	46	⋯	23	32
?	41	21	⋯	74	33



https://medium.com/dissecting-bert/dissecting-bert-part-1-d3c3d495cdb3

Attention

e.g. I am going to my home and play with toy house.



The output embedding for “house” is mainly attending to “toy” and “play”, while the output embedding for “home” is attending to “my” and “going”.
Even though home and house are synonyms and more related.

$$Attention(q, k, v) = \overbrace{softmax\left(\frac{qk^T}{\sqrt{d_k}}\right)}^{\text{Attention weights}} v$$

from (q) to (k)

vector dimensionality of K, V

Normalizes the score

https://jalammar.github.io/illustrated-transformer/

Does it resolve reference to context?

How is this different from cosine similarity?

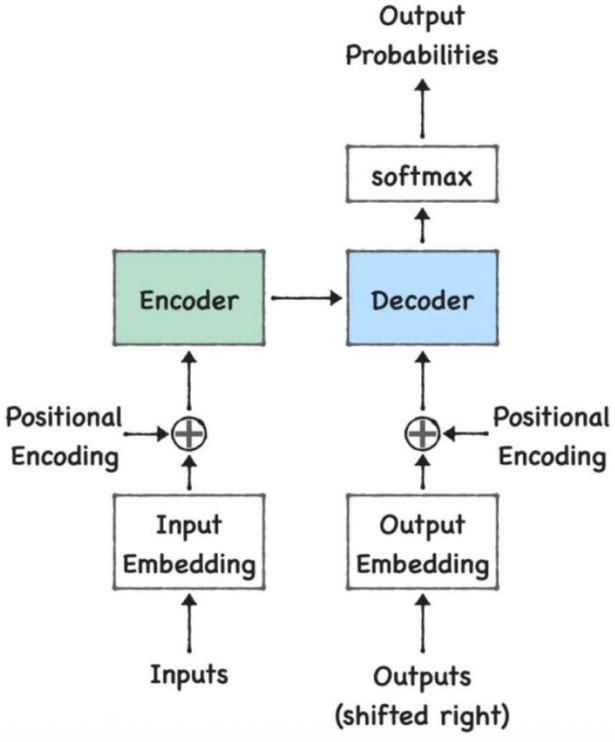
Multi-layer encoders

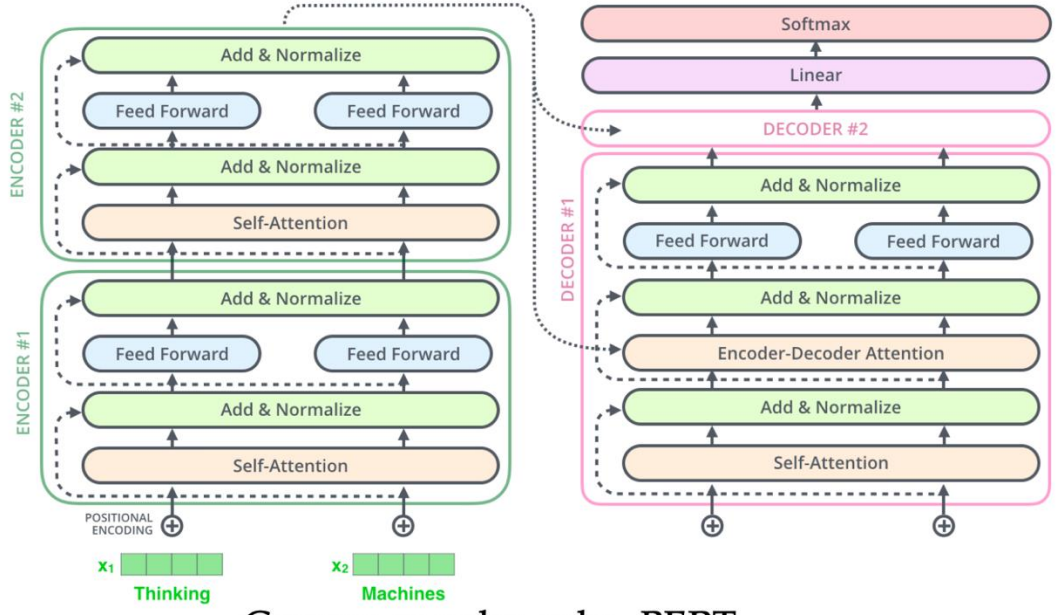
- Why multi-layer when we are already doing multi-head attention?
  - With multiple attention heads, each head learns a separate aspect of the data
  - Multiple layers build progressively more complex and abstract representations of the input sequence. Each layer in the encoder builds upon the previous one, refining the representations and capturing richer contextual information.
    - Similar to CNNs going over larger and larger areas of the image
  - By stacking multiple layers, the model can capture context not just at the local level, but also at a more global level, understanding how different parts of the sequence relate to each oth
    - The first layer gives context for the actual adjacent.
    - And the feedforward layer after that attention layer recognizes phrases.
    - The next attention layer gives context to phrases about other phrases, so that “threw the ball” and “over the wall” get paired together.
    - And then higher and higher features get recognized each layer and the next attention contextualizes those

Why linear layers or feed-forward layers?

- Map input representations to output predictions
- Transforming data within the model's architecture.
- Project vectors into different dimensions
- Enables the model to learn complex relationships, transform data, and generate meaningful output

Summarizing the transformer model





Can use encoder only - BERT
Decoder only – GPT-3
Encoder-decoder models – GPT-4, Llama, ec.

Figure 1: The original encoder-decoder architecture of a transformer model (adapted from Vaswani et al, 2017)