INTRODUZIONE ALL’AI E AL MACHINE LEARNING PER SPECIALISTI DELL’INGEGNERIA – QUARTO INCONTRO

AI GENERATIVA E LARGE SCALE LANGUAGE MODELS

AGENDA

- **CONVEGNO ON LINE 1: Martedì 10 Ottobre, ore 15.00 – 17.00**
- Introduzione ai sistema informativi, Introduzione alle applicazioni data-driven: dalle basi di dati ai dati di addestramento per l'AI, Elementi di Data Management: dai modelli relazionali alle basi di conoscenza.

- **CONVEGNO ON LINE 2: Martedì 17 Ottobre, ore 15.00 – 17.00**
- Introduzione all'Intelligenza Artificiale: tra rappresentazione della conoscenza, ragionamento e apprendimento automatico

- **CONVEGNO ON LINE 3: Martedì 31 Ottobre, ore 15.00 – 18.00**
- Intelligenza nel trattamento dei dati strutturati e semi-strutturati: il Machine Learning

- **CONVEGNO ON LINE 4: Venerdì 10 Novembre, ore 15.00 – 18.00**
- AI Generativa e Large Scale Language Models

OVERVIEW

CONSIGLIO NAZIONALE DEGLI INGEGNERI

CENTENARIO DELL'ALBO DEGLI INGEGNERI

Con la collaborazione incondizionata della Associazione Italiana di Intelligenza Artificiale

Associazione Italiana per l'Intelligenza Artificiale

- Le Reti Neurali: dai percettroni ai Transfomers
  - I *Multilayer Perceptron*
  - Le reti Convoluzionali e le immagini.
  - Reti Ricorrenti
- Applicazioni avanzate ai dati non strutturati
  - ImageNet: Image Processing, Classification, Automated Captioning
  - Visual Question Answering, Multimodality
- Reti attenzionali, trasformers e autoregressive autoencoders
- Modelli Generativi: la famiglia GPT, e chatGPT

RETI NEURALI (RECAP)

PERCETTRONI E *MULTILAYER PERCEPTRONS*

RETI NEURALI

LE RETI CONVOLUZIONALI E LE IMMAGINI

APPLICAZIONI DELLE RETI NEURALI

IMMAGINI: OBJECT DETECTION, ENCODING, MAP COLOURING

OBJECT DETECTION WITH CNNS

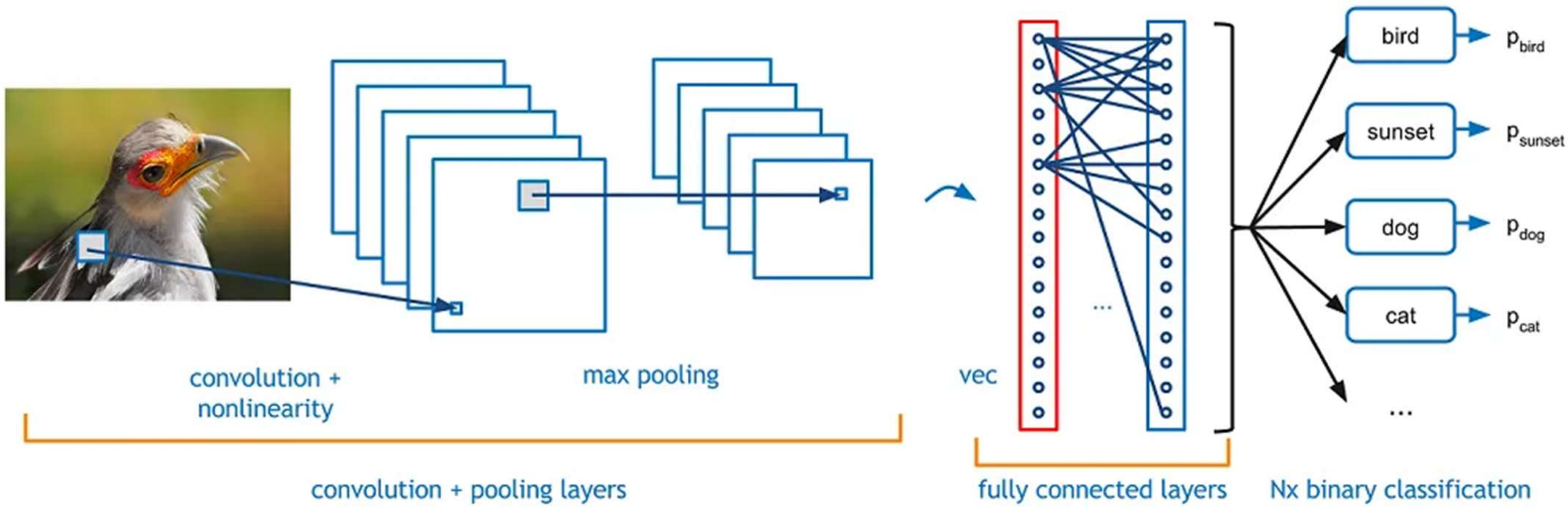
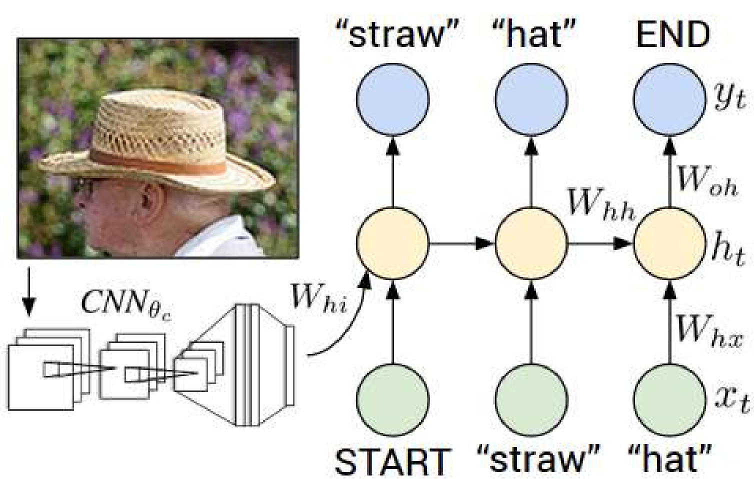


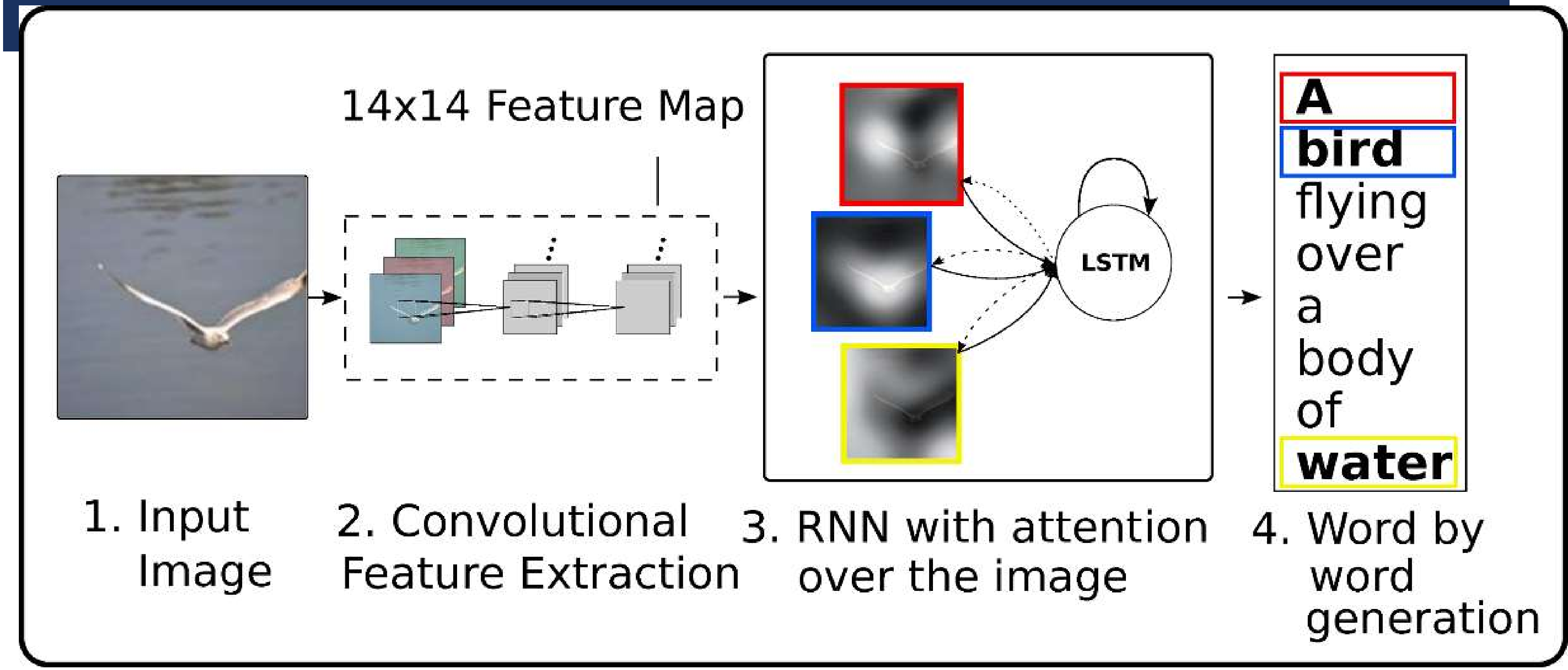
IMAGE CAPTIONING: ADVANCED ARCHITECTURES

- Image to captions
  - Convolutional Neural Network to learn a representation of the image
  - (Bi-directional) Recurrent Neural Network to generate a caption describing the image
    - its input is the representation computed from the CNN
    - its output is a sequence of words, i.e. the caption



"baseball player is throwing ball in game."

THE ARCHITECTURE



ATTENTION: A BRODGE BETWEEN VISION AND LANGUAGE

INTEGRATED VISION AND LANGUAGE PROCESSING: IMAGE CAPTIONING AND ATTENTION

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

ESEMPI


DALL·E
History
Collections
Edit
In
New Chat
Today
Hello and Hi
June
Canzone per Mamma
February
Train Neural Model for NWM
January
New chat
vibrant portrait painting of Salvador Dalí with a robotic half face
a shiba inu wearing a beret and black turtleneck
a close up of a handpalm with leaves growing from it
an espresso machine that makes coffee from human souls, artstation
panda mad scientist mixing sparkling chemicals, artstation
a corgi's head depicted as an explosion of a nebula

NEURAL ENCODING-DECODING FOR DALL-E

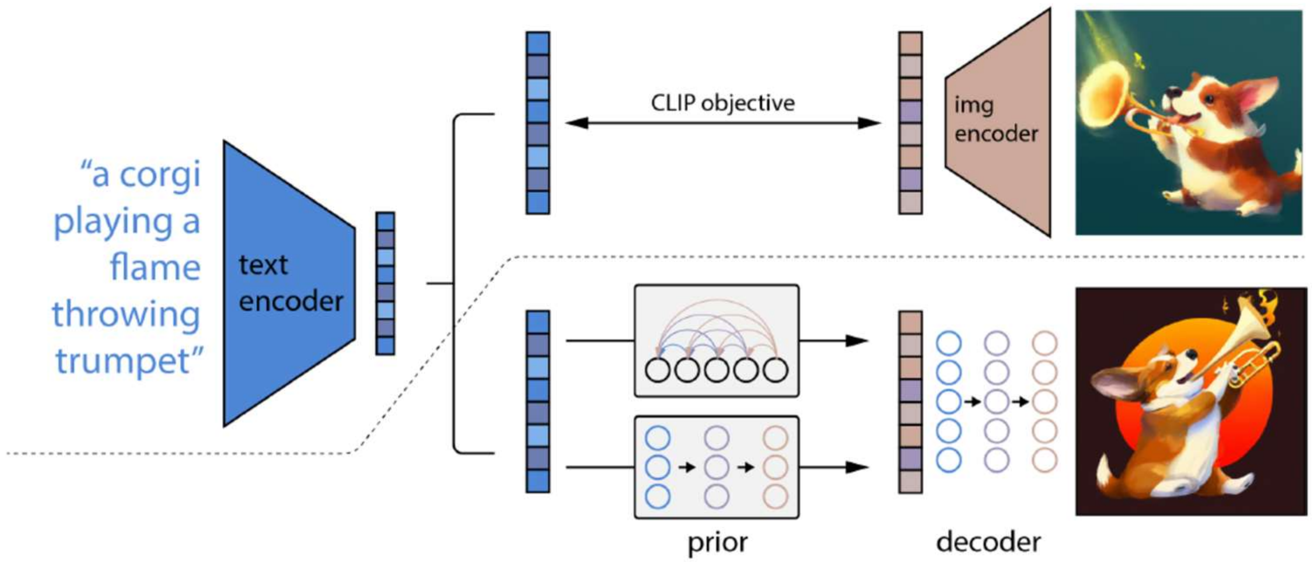
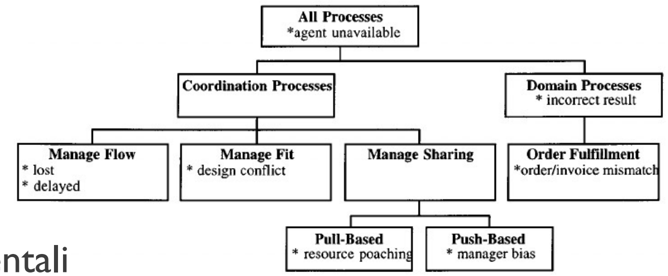


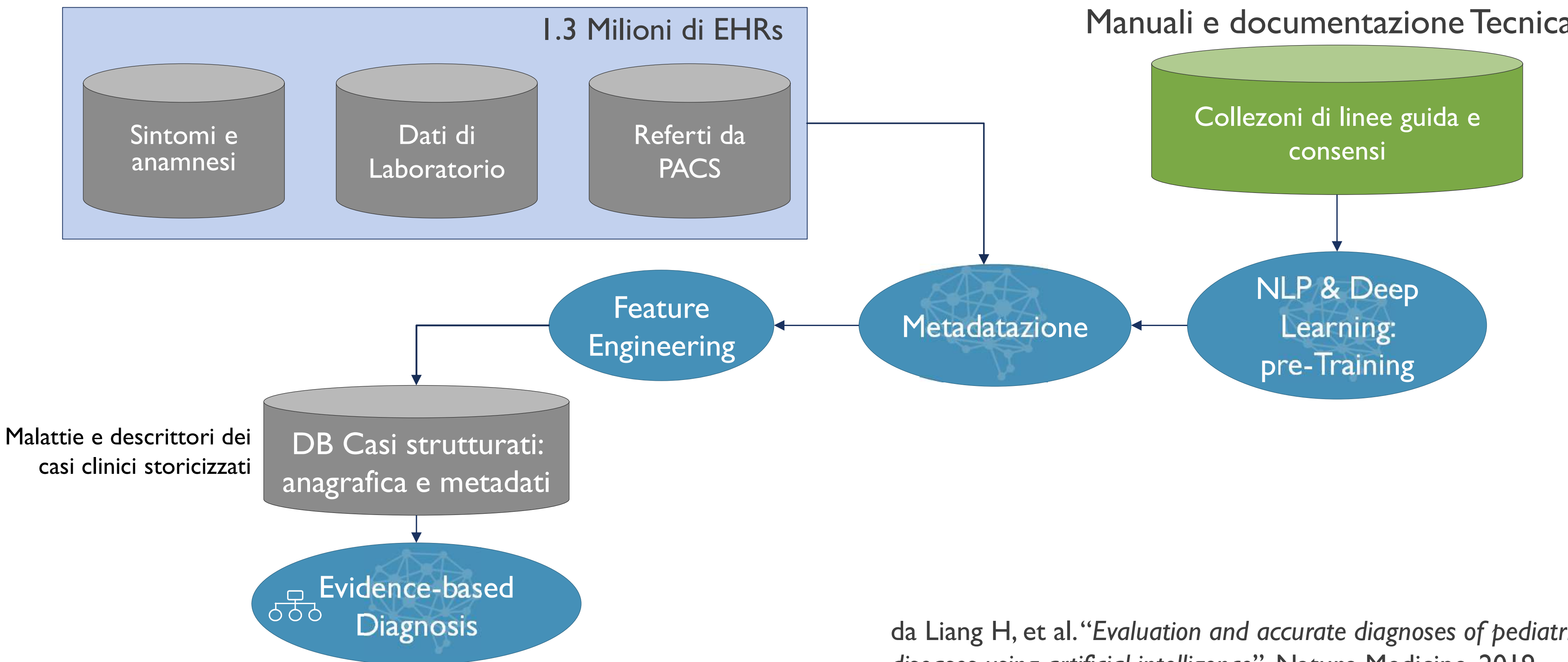
Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

BANKING: ABILABERT IN DECODE



- 5 banche coordinate da ABILAB
- Una Process Taxonomy condivisa e differenti Basi di Dati Documentali
- Automatic Text-driven Process Mapping basato su reti neurali Trasformers

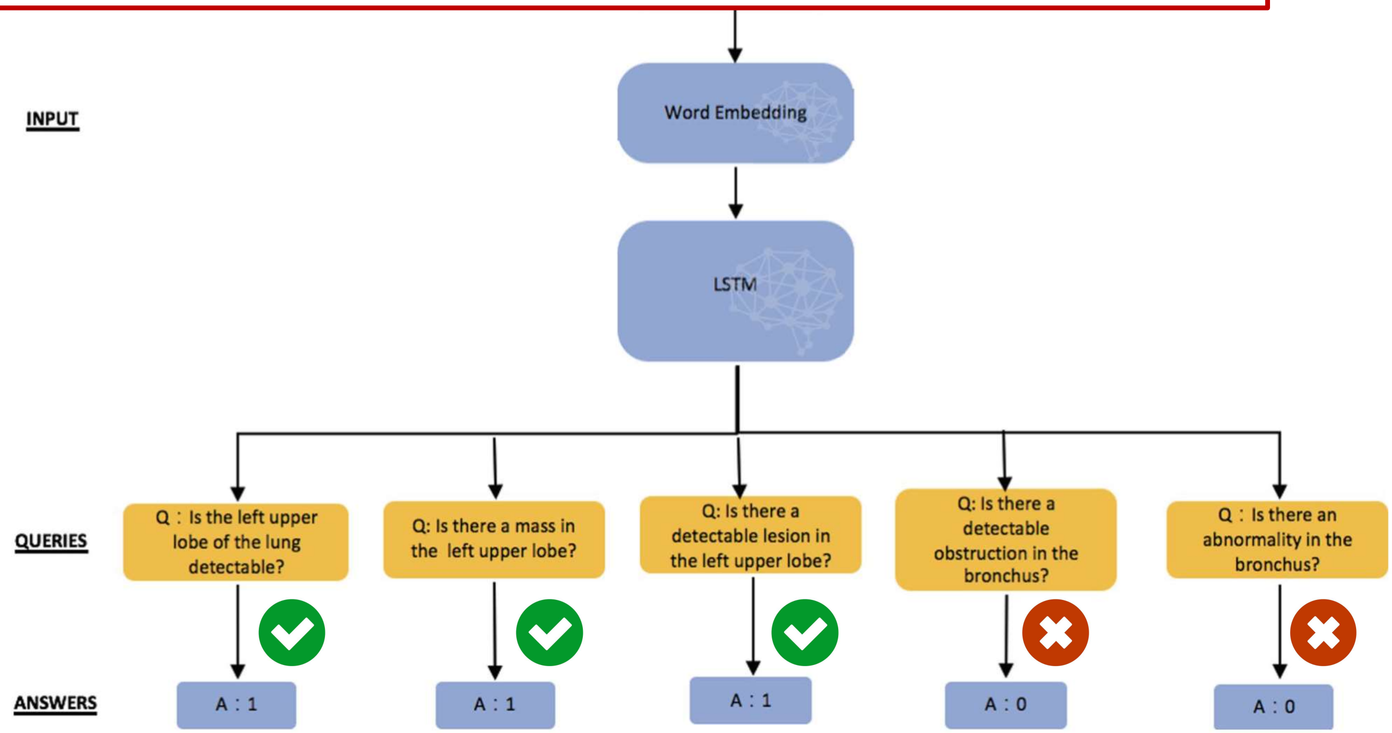
DIAGNOSI MALATTIE PEDIATRICHE: UN WORKFLOW ORIENTATO AL ML



da Liang H, et al. "*Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence*", Nature Medicine, 2019

MEDICAL INFORMATION EXTRACTION

INPUT: "Si **osserva** ***una lesione*** nel ***lobo superiore sinistro*** del ***polmone del paziente*** . ..."

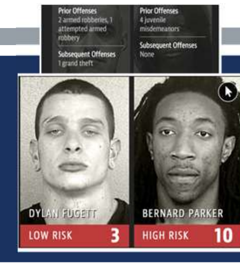


EVIDENCE BASED DIAGNOSIS: RISULTATI (11,926 PAZIENTI)

Table 2 | Illustration of diagnostic performance of our AI model and physicians

Disease conditions	Our model	Physicians				
		Physician group 1	Physician group 2	Physician group 3	Physician group 4	Physician group 5
Asthma	0.920	0.801	0.837	0.904	0.890	0.935
Encephalitis	0.837	0.947	0.961	0.950	0.959	0.965
Gastrointestinal disease	0.865	0.818	0.872	0.854	0.896	0.893
Group: 'Acute laryngitis'	0.786	0.808	0.730	0.879	0.940	0.943
Group: 'Pneumonia'	0.888	0.829	0.767	0.946	0.952	0.972
Group: 'Sinusitis'	0.932	0.839	0.797	0.896	0.873	0.870
Lower respiratory	0.803	0.803	0.815	0.910	0.903	0.935
Mouth-related diseases	0.897	0.818	0.872	0.854	0.896	0.893
Neuropsychiatric disease	0.895	0.925	0.963	0.960	0.962	0.906
Respiratory	0.935	0.808	0.769	0.89	0.907	0.917
Systemic or generalized	0.925	0.879	0.907	0.952	0.907	0.944
Upper respiratory	0.929	0.817	0.754	0.884	0.916	0.916
Root	0.889	0.843	0.863	0.908	0.903	0.912
Average F1 score	**0.885**	**0.841**	**0.839**	**0.907**	**0.915**	**0.923**

COMPAS: PROFILING



- COMPAS dataset (*Correctional Offender Management Profiling for Alternative Sanctions*)
  - raccoglie dati nell'ambito della giustizia penale utilizzati per prevedere il rischio di recidiva di un imputato.
  - pubblicato da ProPublica nel 2016 sulla base dei dati raccolti dalla contea di Broward.

Attributes	Type	Values	#Missing values	Description
sex	Binary	{Male, Female}	0	Sex
age	Numerical	[18 - 96]	0	Age in years
age_cat	Categorical	3	0	Age category
race	Categorical	6	0	Race
juv_fel_count	Numerical	[0 - 20]	0	The juvenile felony count
juv_misd_count	Numerical	[0 - 13]	0	The juvenile misdemeanor count
juv_other_count	Numerical	[0 - 17]	0	The juvenile other offenses count
priors_count	Numerical	[0 - 38]	0	The prior offenses count
c_charge_degree	Binary	{F, M}	0	Charge degree of original crime
score_text	Categorical	3	0	ProPublica-defined category of decile score
v_score_text	Categorical	3	0	ProPublica-defined category of v_decile_score
two_year_recid	Binary	{0, 1}	0	Whether the defendant is rearrested within two years

Caratteristiche Contiene 7.214 istanze. Ogni imputato è descritto da 52 attributi (31 categorici, 6 binari, 14 numerici e un attributo nullo)

Task L'obiettivo è **prevedere se un individuo viene nuovamente arrestato entro due anni dal primo arresto**

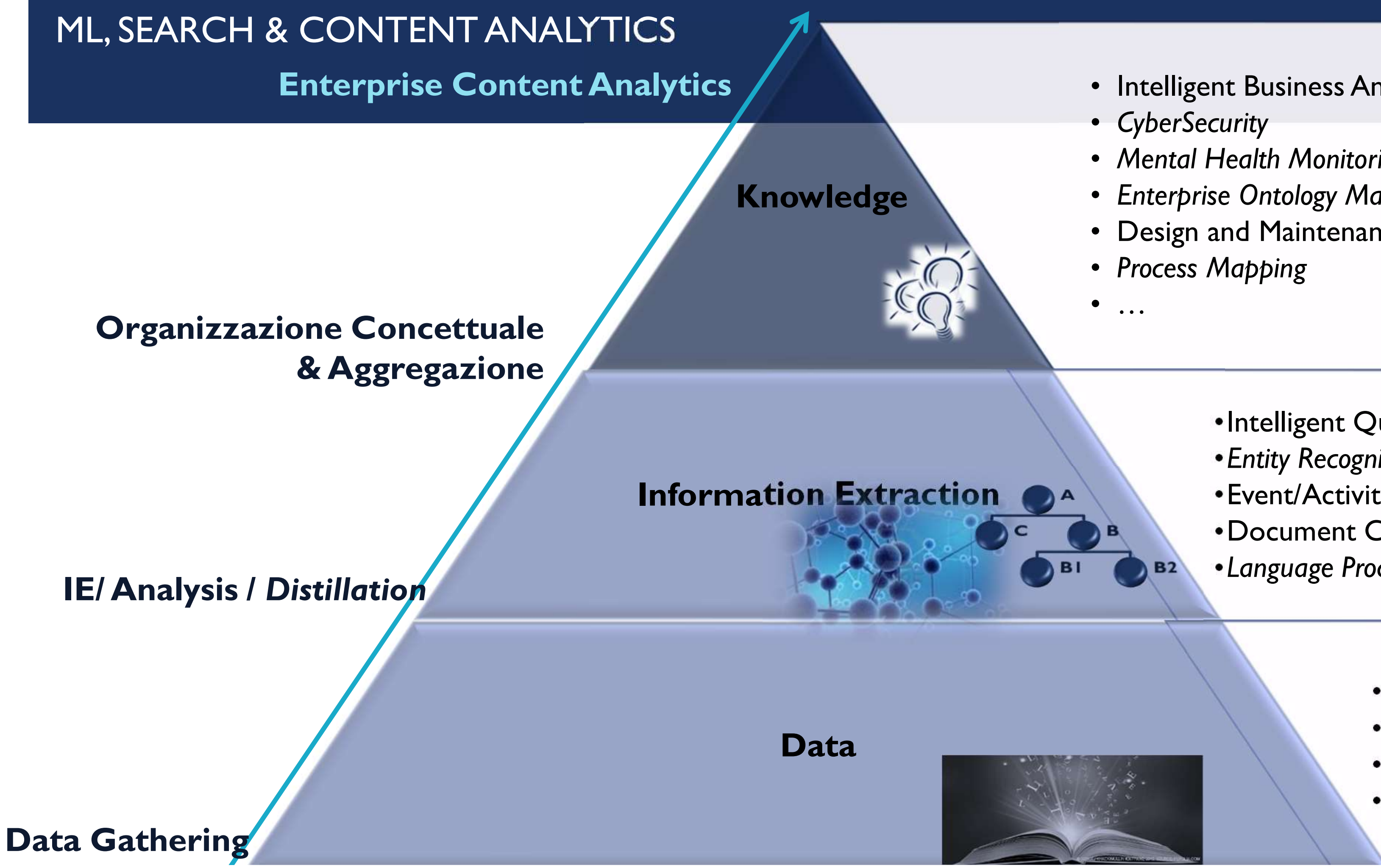
Possibili rischi
Alcuni gruppi sociali (gli afroamericani) hanno maggiori probabilità di essere erroneamente etichettati come a rischio più elevato rispetto agli altri (i caucasici).
Eticamente ingiusto.
Obbiettivo: ottenere un sistema equo tra gruppi sociali diversi.

https://github.com/propublica/compas-analysis

DATA-DRIVEN APPLICATIONS:

ML, SEARCH & CONTENT ANALYTICS

Enterprise Content Analytics



- Intelligent Business Analysis
- *CyberSecurity*
- *Mental Health Monitoring*
- *Enterprise Ontology Management*
- Design and Maintenance of conceptual catalogs
- *Process Mapping*
- …

- Intelligent QueryProcessing
- *Entity Recognition and Linking*
- Event/Activity Recognition
- Document Classification
- *Language Processing*

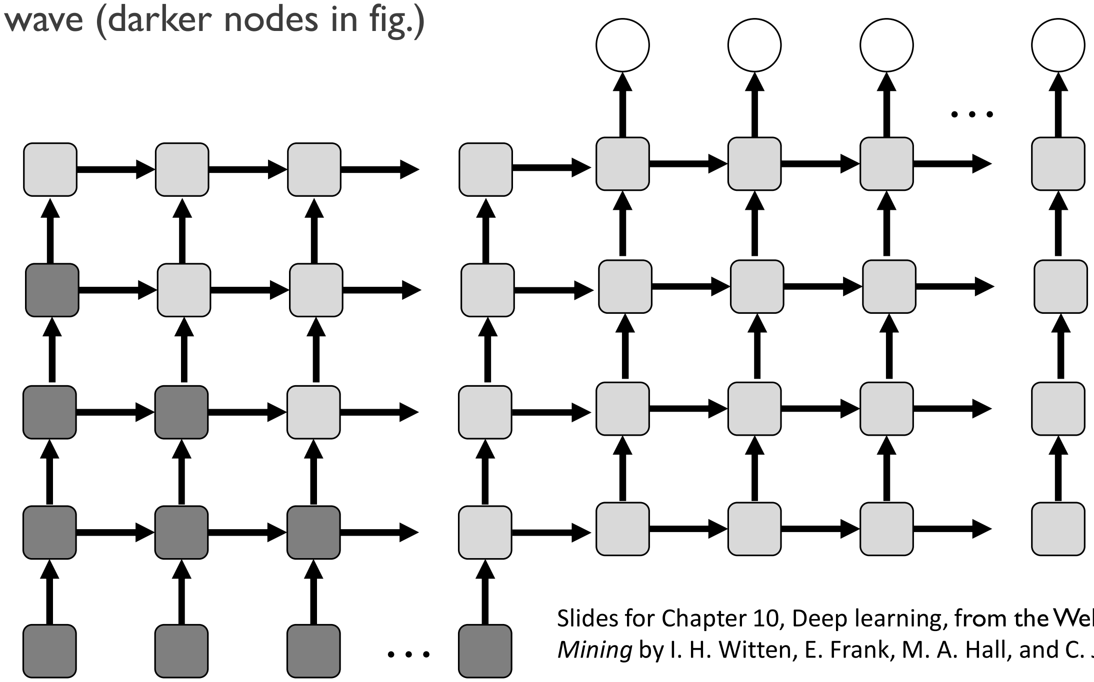
- Narrative and Legal Texts
- Legacy Models
- Social Media
- Open Web sources

RETI NEURALI AVANZATE:
ATTENZIONE E TRANSFORMERS

METODI E ARCHITETTURE

ENCODER-DECODER DEEP ARCHITECTURES

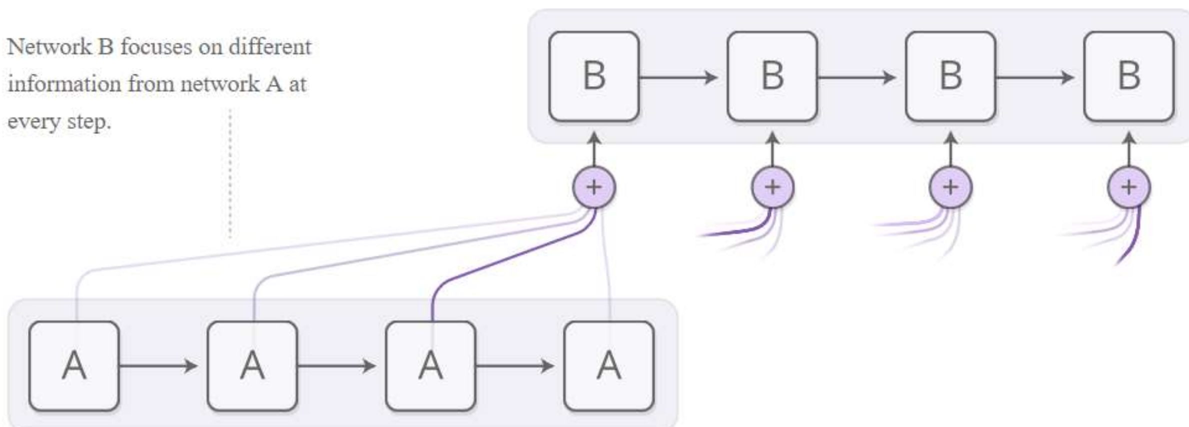
- Given enough data, a deep encoder-decoder architecture (see below) can yield results that compete with hand-engineered translation systems.
- The connectivity structure means that partial computations in the model can flow through the graph in a wave (darker nodes in fig.)



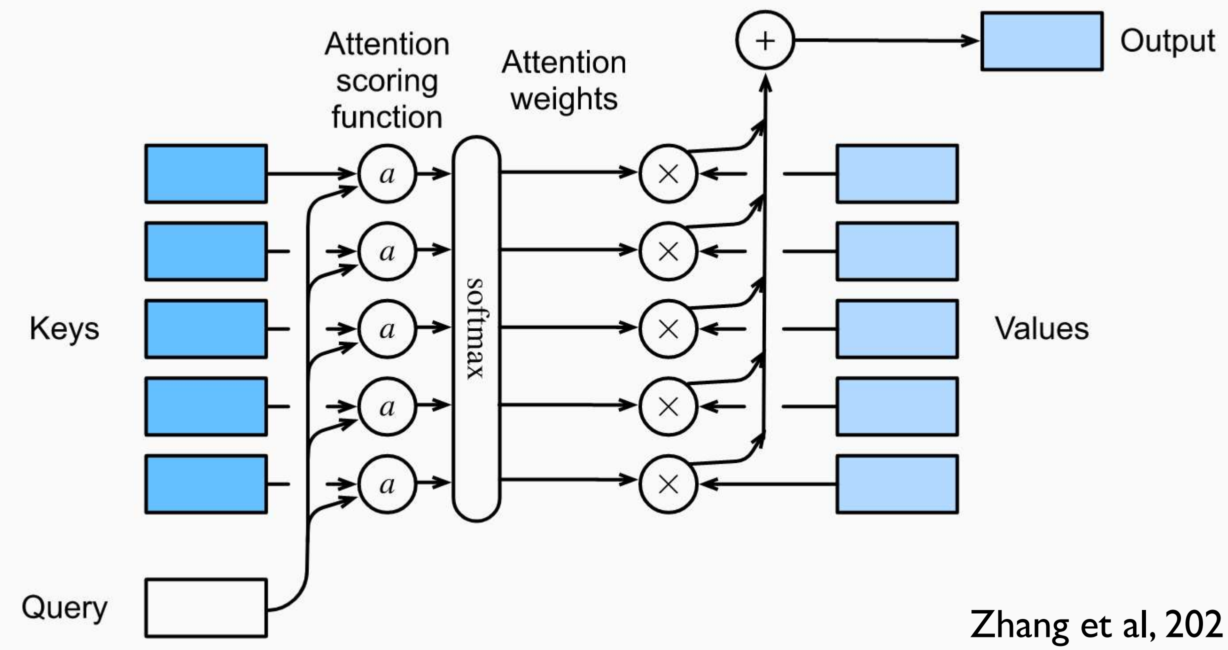
Slides for Chapter 10, Deep learning, from the Weka book, *Data Mining* by I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal

ATTENTION-BASED RNNS

- A NN (e.g. B) is used to attend the outcome of a second network A, e.g. (Vaswani et al., 2017)

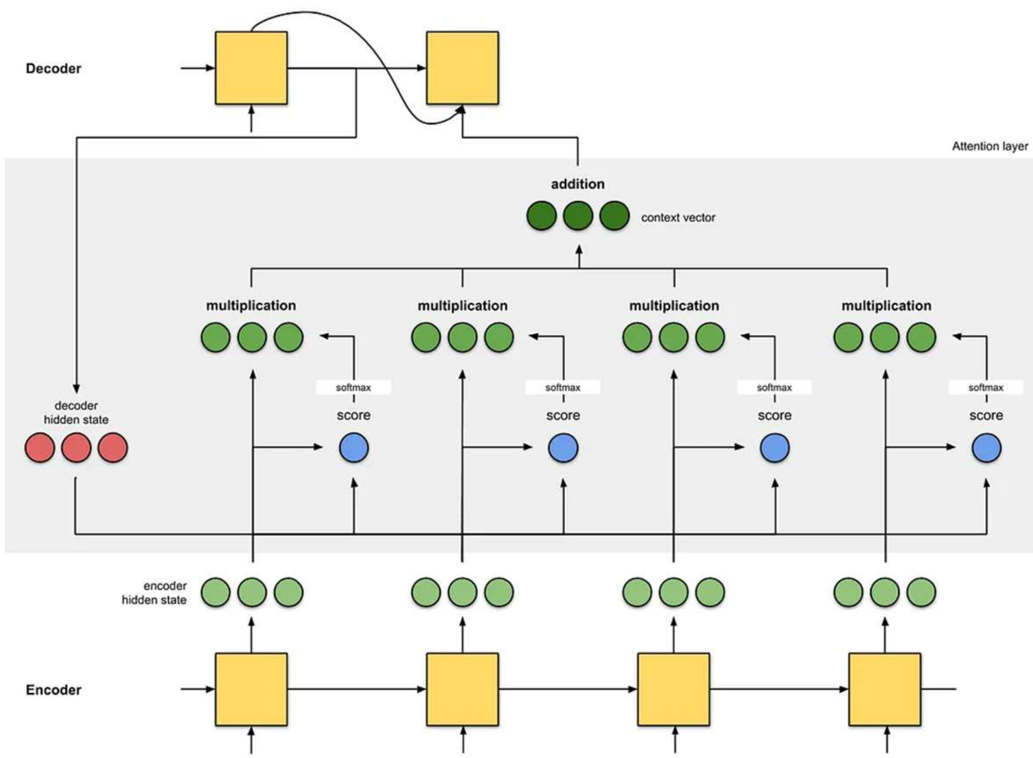


ATTENTION FUNCTIONS

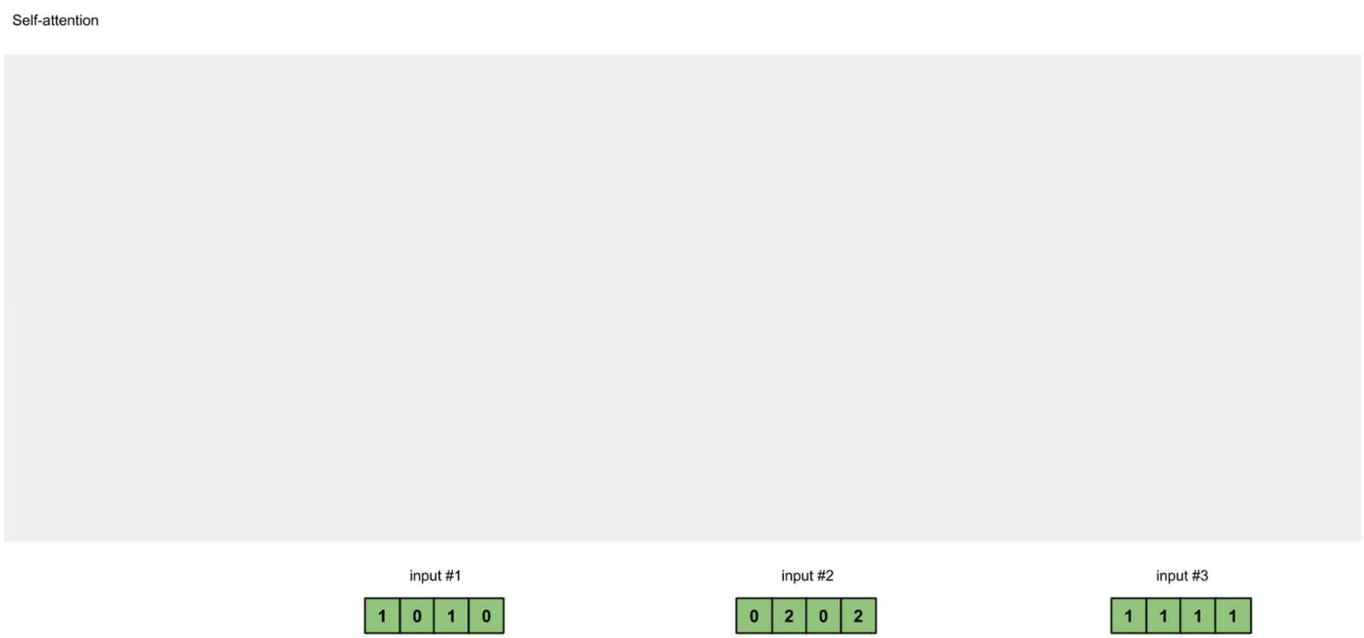


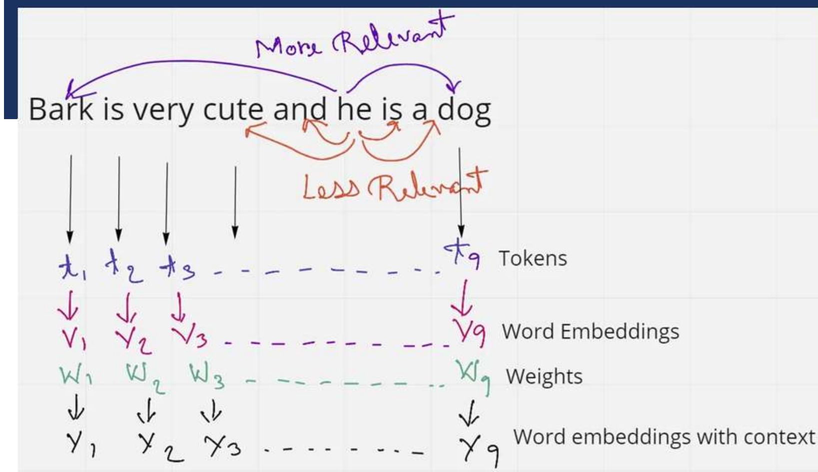
Zhang et al, 2021

ATTENTION IN SEQ2SEQ MODELS

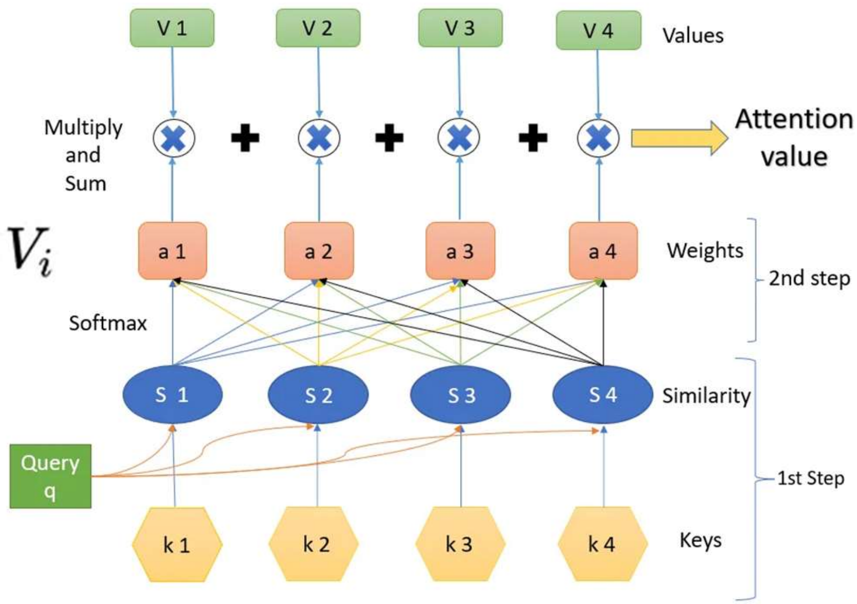


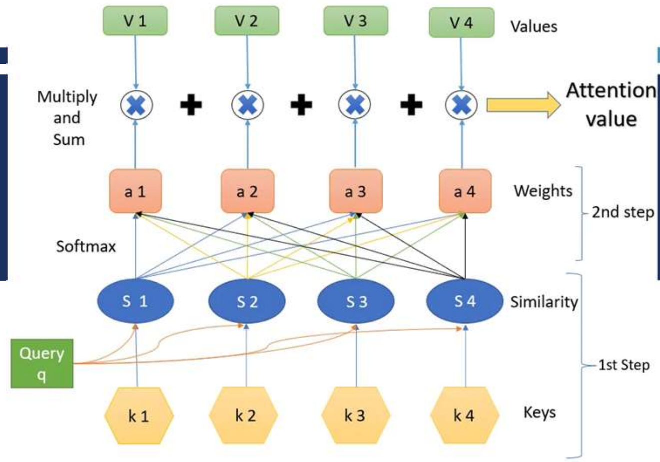
SELF-ATTENTION

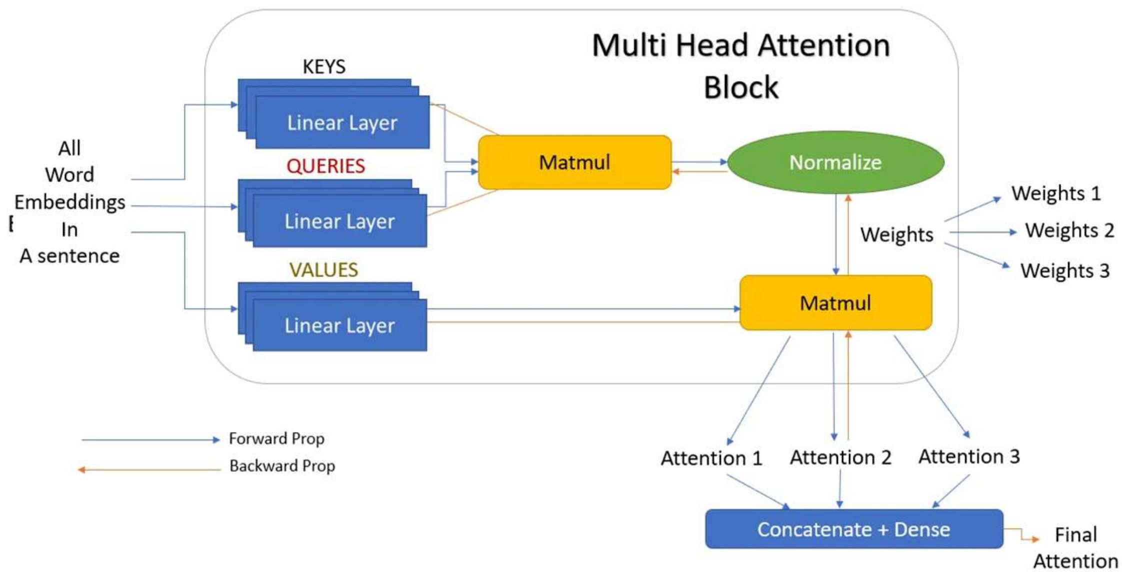




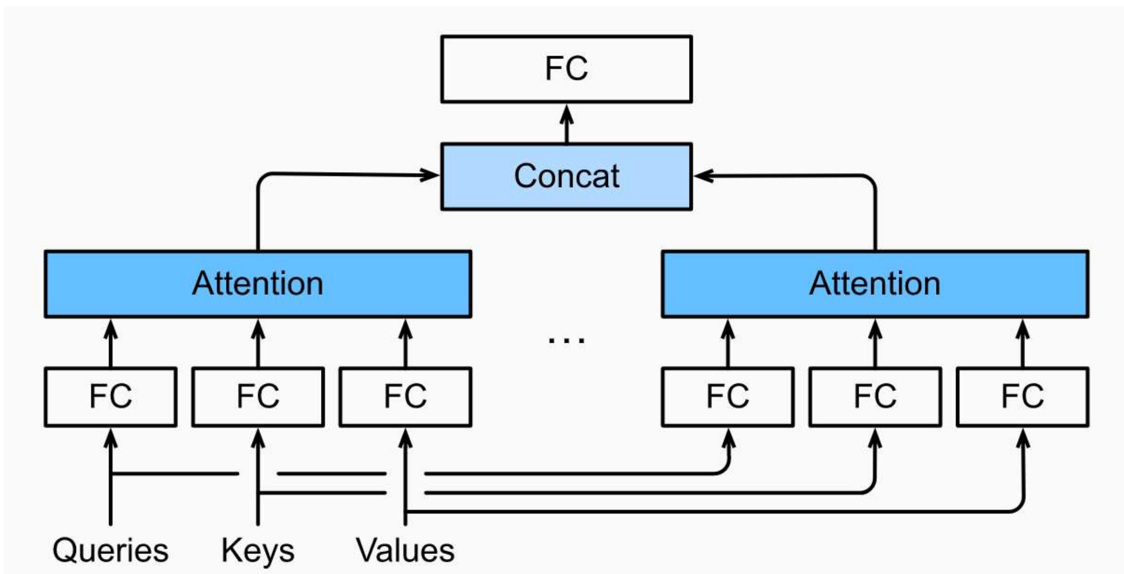
$$attention\ value\ =\ \sum_i a_i V_i$$



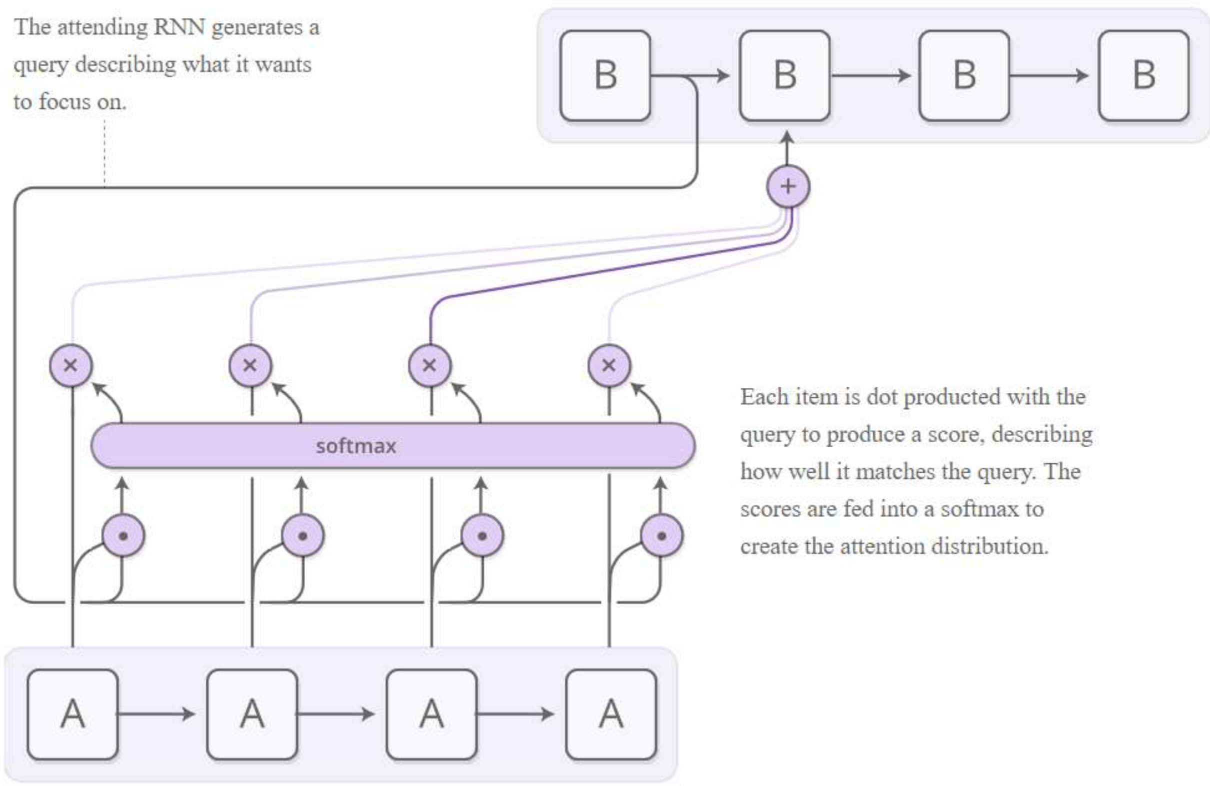

V 1
V 2
V 3
V 4
Values
Multiply and Sum
Attention value
a 1
a 2
a 3
a 4
Weights
2nd step
Softmax
S 1
S 2
S 3
S 4
Similarity
Query q
1st Step
k 1
k 2
k 3
k 4
Keys


Multi Head Attention Block
KEYS
Linear Layer
QUERIES
Linear Layer
VALUES
Linear Layer
All Word Embeddings In A sentence
Matmul
Normalize
Weights
Weights 1
Weights 2
Weights 3
Matmul
Attention 1
Attention 2
Attention 3
Concatenate + Dense
Final Attention
Forward Prop
Backward Prop

ATTENTION: MULTIHEAD



ATTENTION-BASED RNNS



ATTENTION MECHANISMS IN MACHINE TRANSLATION

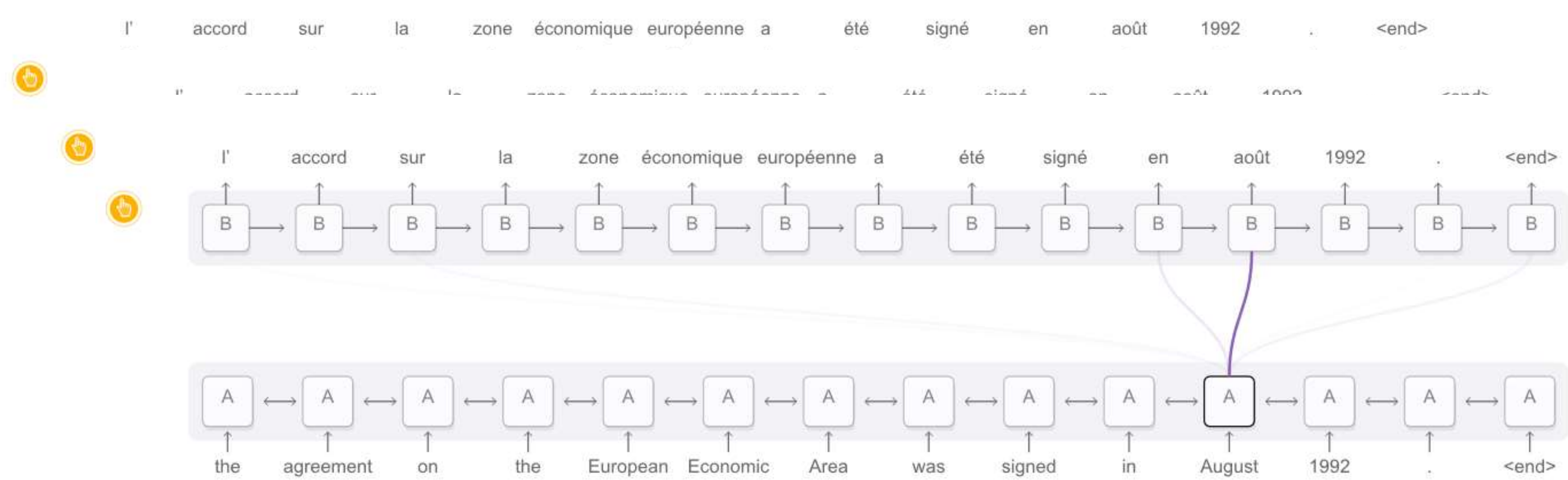


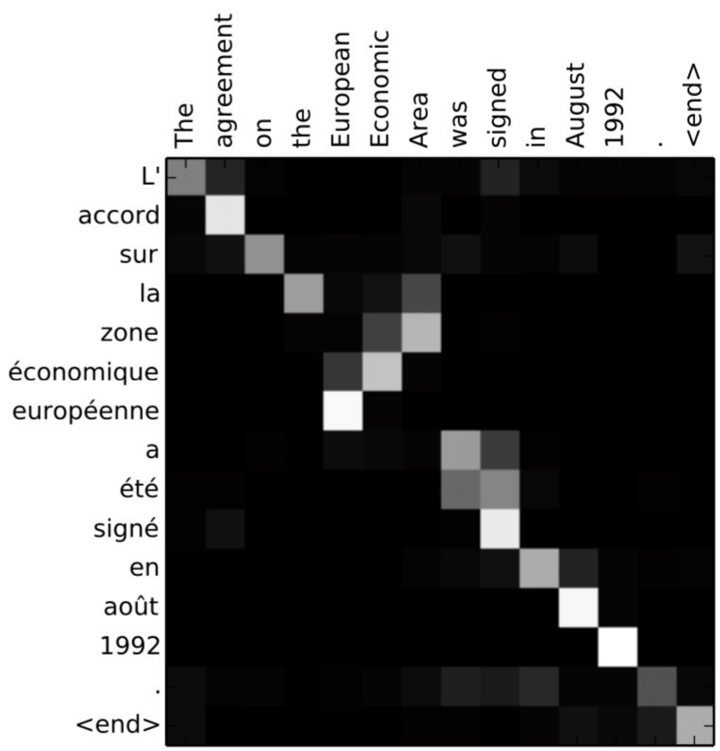
Diagram derived from Fig. 3 of Bahdanau, *et al.* 2014

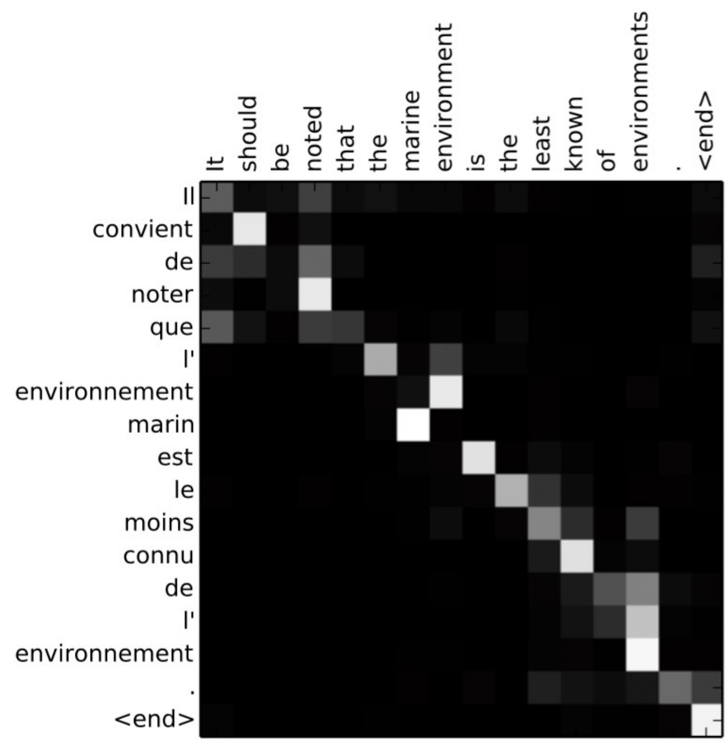
ATTENTION & ENCONDING

- In a decoding process (e.g. machine translation) there are **three** kinds of dependencies for neural architectures
- Dependencies can establish between
  1. the ***input and output*** tokens
  2. the ***input tokens themselves***
  3. the ***output tokens themselves***

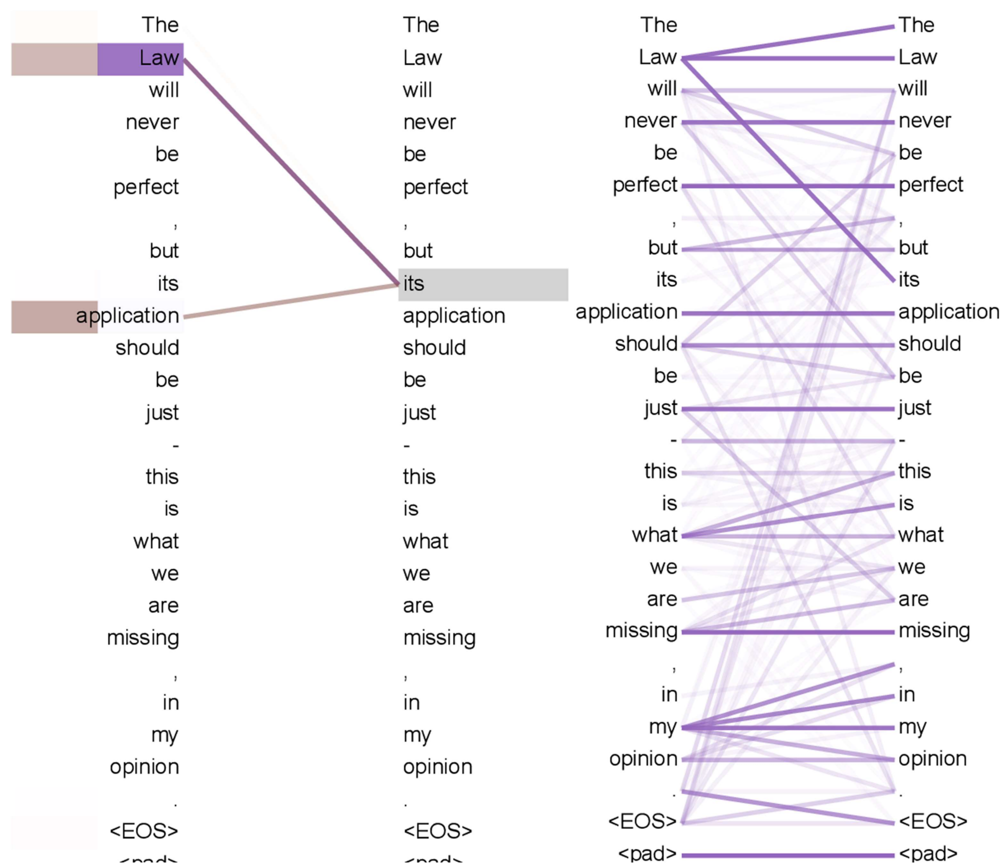
- Examples:
  - Machine Translation
  - QA where the query the answer paragraph is the input and the matched answer is the output

ATTENTION IN MACHINE TRANSLATION

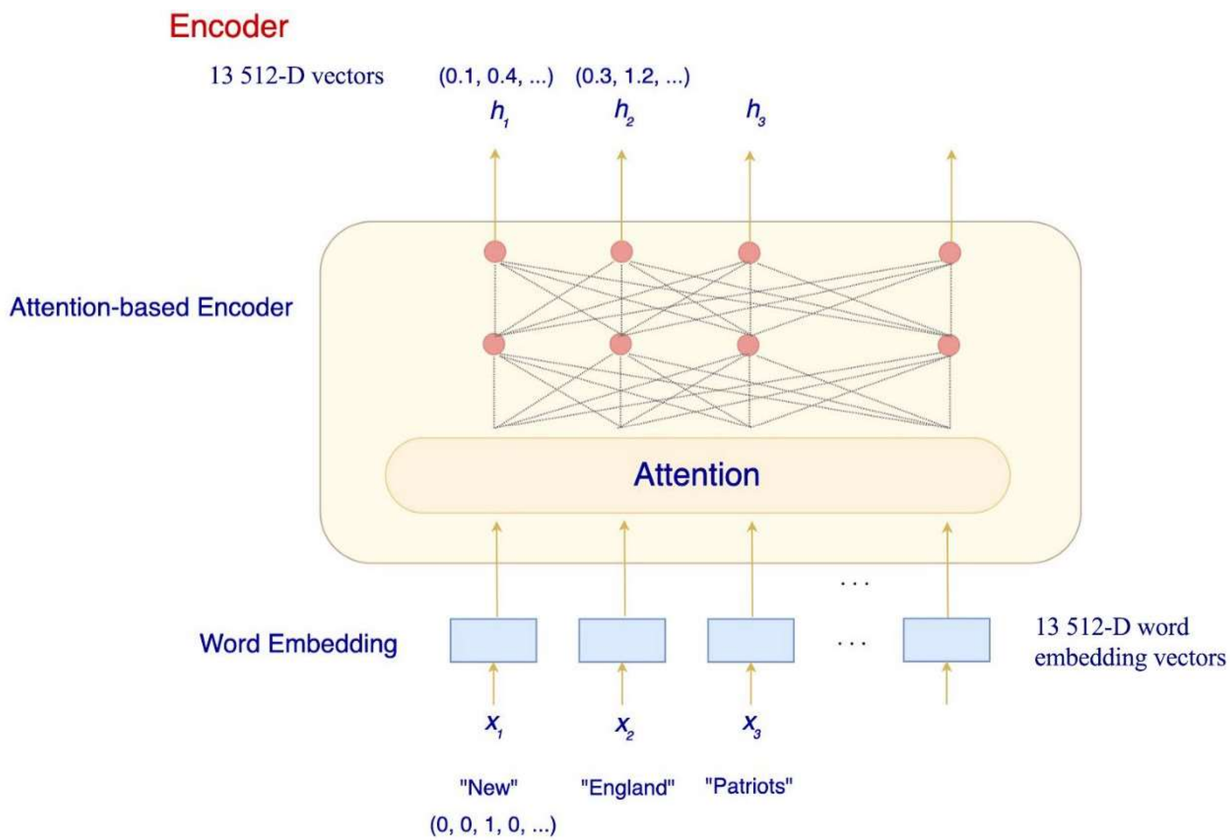


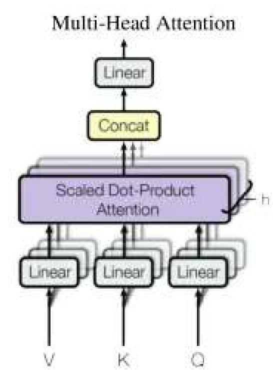


ATTENTION AND ANAPHORA



BERT & NLP


Encoder
13 512-D vectors
(0.1, 0.4, ...)
(0.3, 1.2, ...)
h_1
h_2
h_3
Attention-based Encoder
Attention
Word Embedding
13 512-D word embedding vectors
x_1
x_2
x_3
"New"
"England"
"Patriots"
(0, 0, 1, 0, ...)


Multi-Head Attention
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q

BERT & NLP (2)

- How to *train* (i.e. optimize) the encoding?
- General and complex tasks defined in (Devlin et al., 2018) are
  - Masked Language Modeling (15%)
    - Inpired by Distributional Hypothesis
    - Can be Simulated and does not require any labeling
  - Next Sentence Prediction
    - Inspired by Textual Inference tasks (e.g. Textual Entailment)
    - Can be Simulated and does not require any labeling
- Source Representations
  - Words? And why not subword? (in the BERT jargon) Word Pieces!!
    - Useful to deal with out-of-vocabulary phenomena

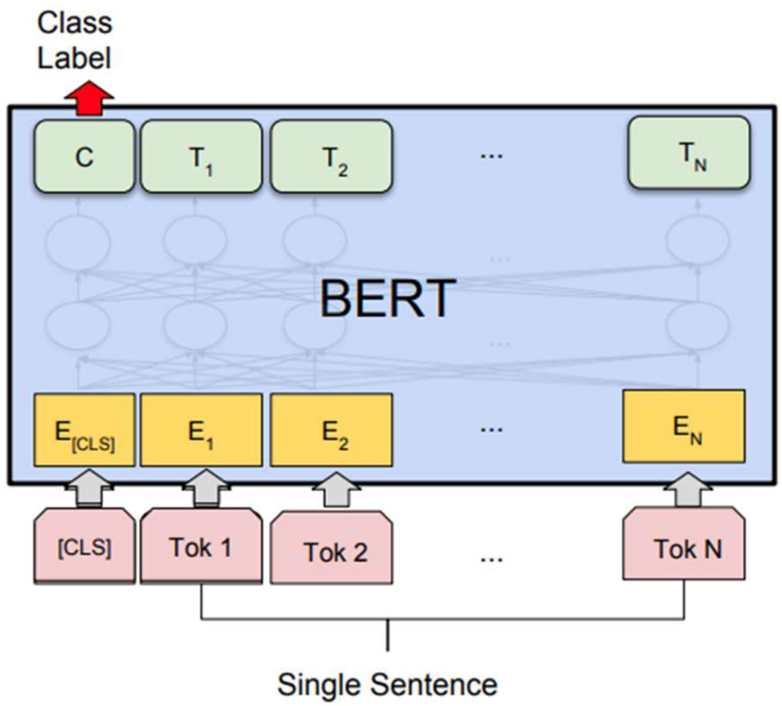
BERT (DEVLIN ET AL. '18)

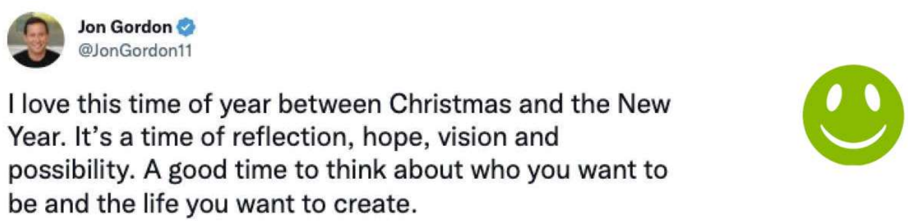
Pretraining on two unsupervised prediction tasks:

- **Masked Language Model**: given a sentence *s* with missing words, reconstruct *s*
  - Example: Amazon <MASK> amazing → Amazon is amazing
  - In BERT the language modeling is deeply Bidirectional, while in ELMo the forward and backward LMs were two independent branches of the NN

- **Next Sentence Prediction**: given two sentences s_1 and s_2, the task is to understand whether s_2 is the actual sentence that follows s_1
  - 50% of the training data are positive examples: s_1 and s_2 are actually consecutive sentences
  - 50% of the training data are negative examples: s_1 and s_2 are randomly chosen from the corpus

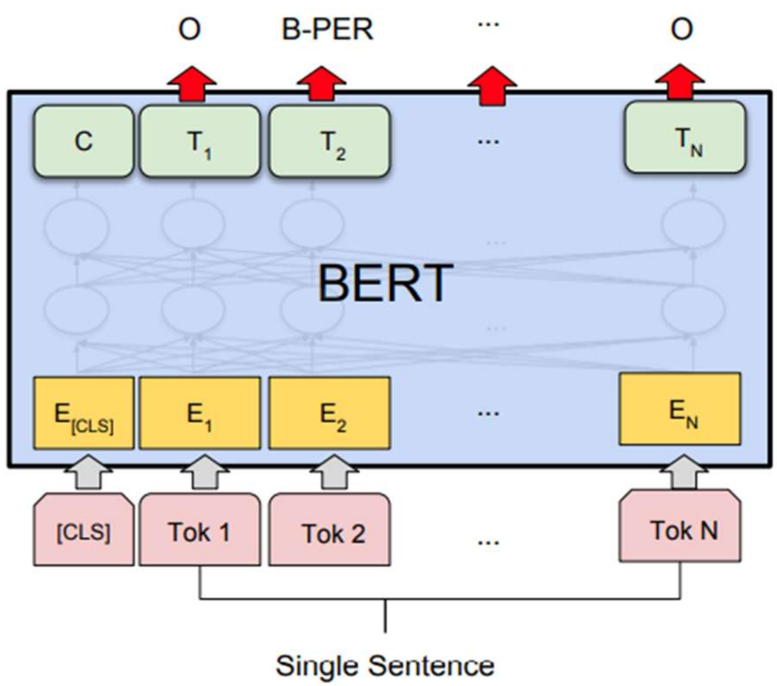
BERT (DEVLIN ET AL. '18): TASKS





BERT for single sentence classification (Sentiment analysis, Intent Classification, etc.)

BERT (DEVLIN ET AL. '18)

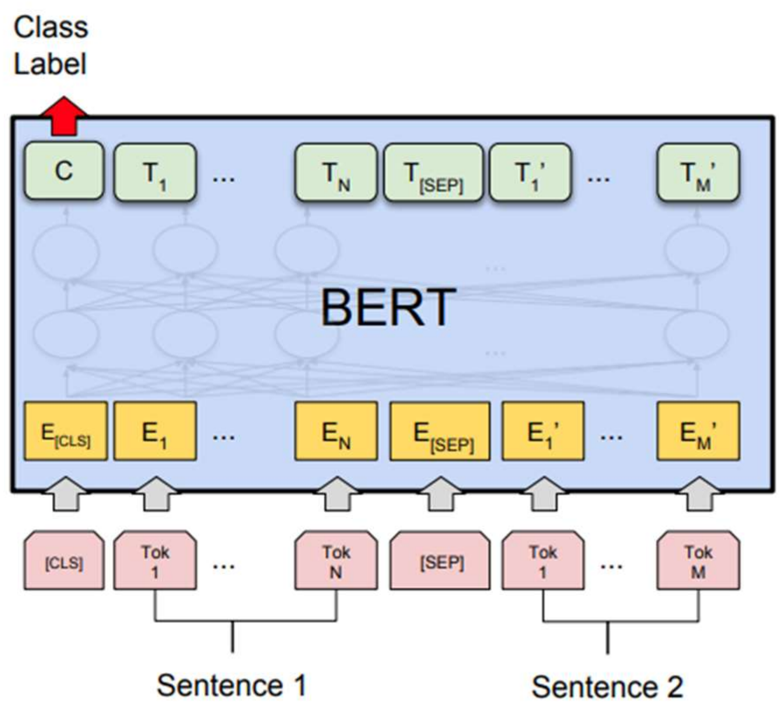


Task: Slot tagging

```
|x 178:1 |# BOS      |y 128:1 |# O
|x 770:1 |# show     |y 128:1 |# O
|x 429:1 |# flights  |y 128:1 |# O
|x 444:1 |# from     |y 128:1 |# O
|x 272:1 |# burbank  |y 48:1  |# B-fromloc.city_name
|x 851:1 |# to       |y 128:1 |# O
|x 789:1 |# st.      |y 78:1  |# B-toloc.city_name
|x 564:1 |# louis    |y 125:1 |# I-toloc.city_name
|x 654:1 |# on       |y 128:1 |# O
|x 601:1 |# monday   |y 26:1  |# B-depart_date.day_name
|x 179:1 |# EOS      |y 128:1 |# O
```

BERT for Sequence Tagging Tasks (e.g., POS tagging, Named Entity Recognition, etc.)

BERT (DEVLIN ET AL. '18)



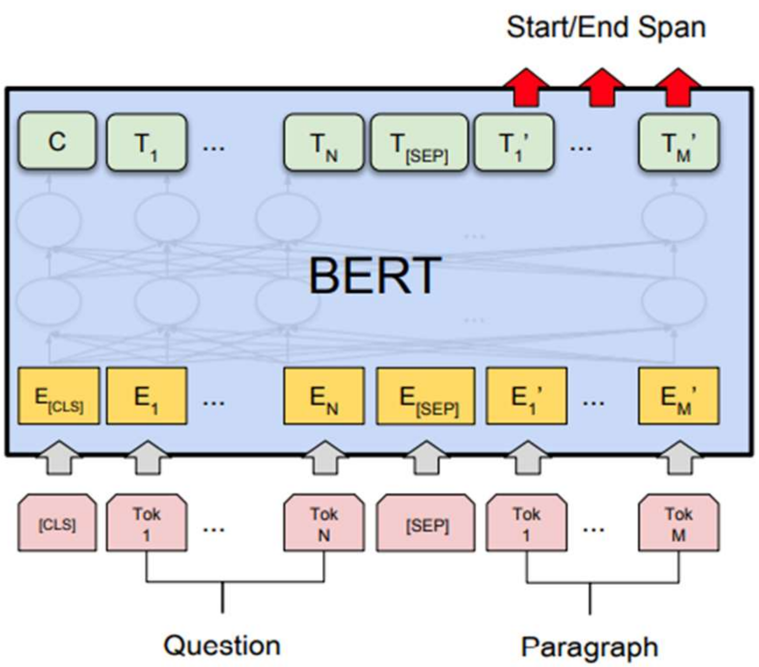
Answer selection in QA: Decide if Q contains an answer to A:
A:"What is the Capital of Italy?"
Q:"Rome, as the capital of Italy, ….."

RTE: Given P decide if H is true (or not)
P:"Roma is the Capital of Italy."
H:"Rome is in Italy."

RTE: Given S1 and S2 decide if they are paraphrases (or not)
S:1 "Roma is the Capital of Italy."
S2:"Italy has Rome as its own Capital town."

BERT for sentence pairs classification (Paraphrase Identification, answer selection in QA, Recognizing Textual Entailment)

BERT (DEVLIN ET AL. '18)



Answer Span Selection in QA:

Decide which part of Q corresponds to the answer to A:

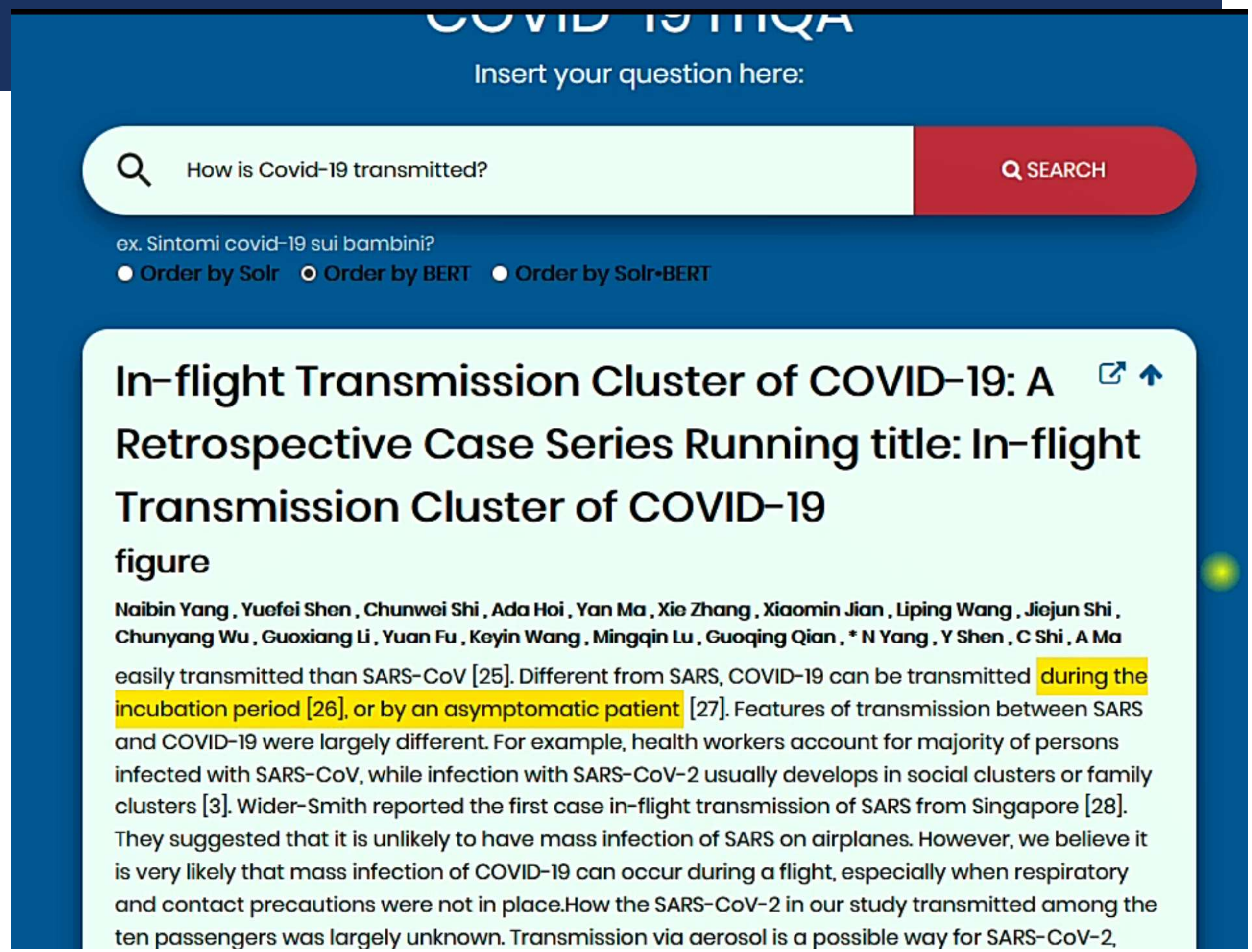
A: "What is the Capital of Italy?"

Q:"<Start>Rome<End>, as the capital of Italy, ….."

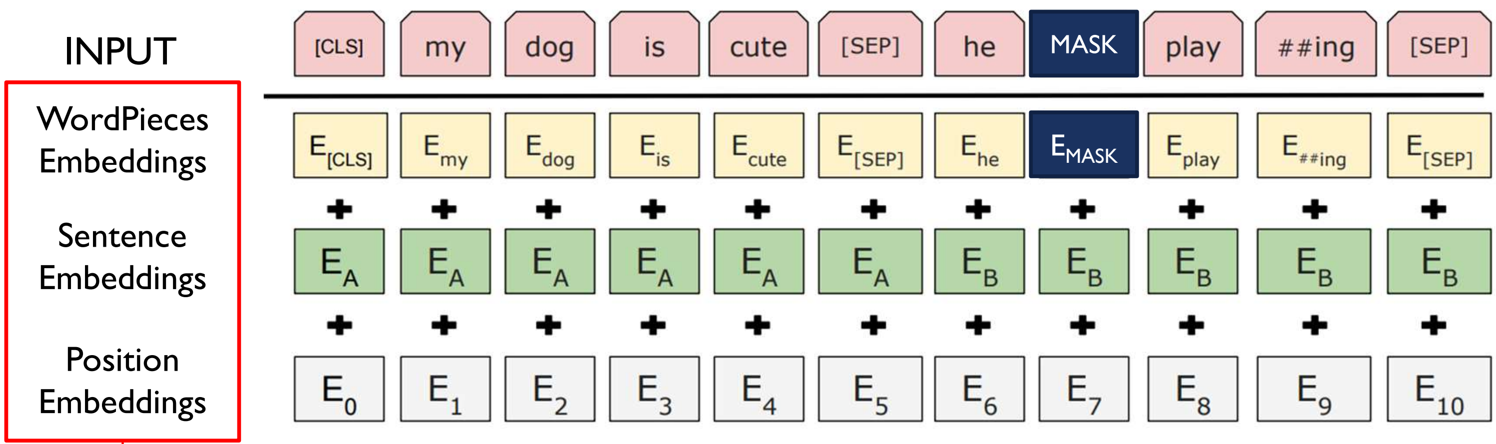
BERT for Answer Span Selection in Question Answering

A QA EXAMPLE ON SQUAD

- Cross-lingual Question Answering



BERT PRETRAINING: INPUT REPRESENTATIONS



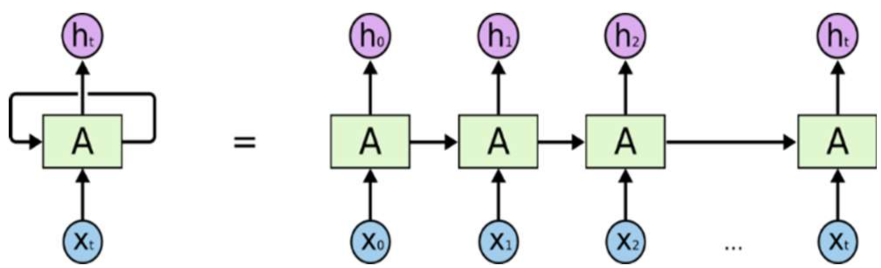
All these embeddings are learned during the (pre)training process

In pre-training 15% of the input tokens are masked for the masked LM task

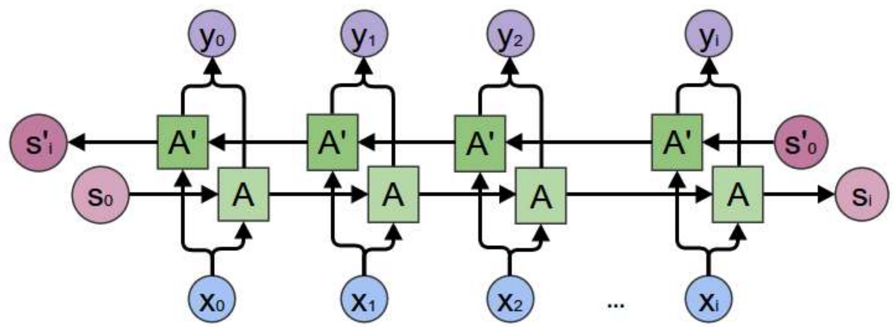
RETI NEURALI AVANZATE:
DALL'AUTOENCODING ALLA IA GENERATIVA

METODI E ARCHITETTURE

RECAP: Machine learning paradigms underlying ChatGPT



Machine learning paradigms underlying ChatGPT

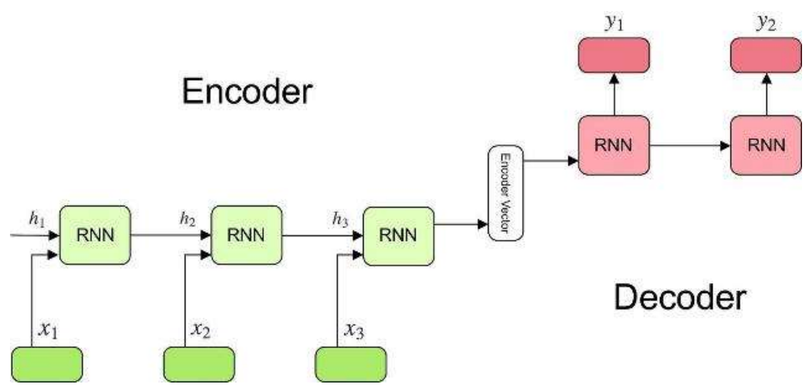


Machine learning paradigms underlying ChatGPT

RNNs
1986

Bidirectional
RNNs
1997

Encoder-Decoder
RNNs
2014



Sutskever, O.Vinyals, & Q.V. Le, 2014

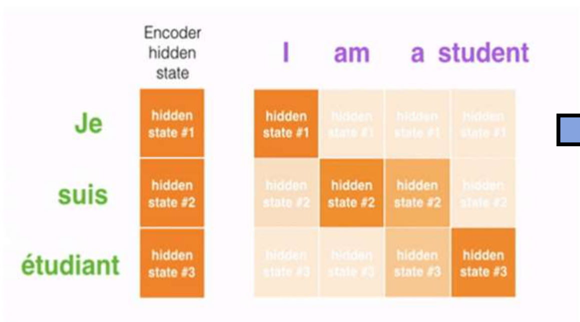
Machine learning paradigms underlying ChatGPT

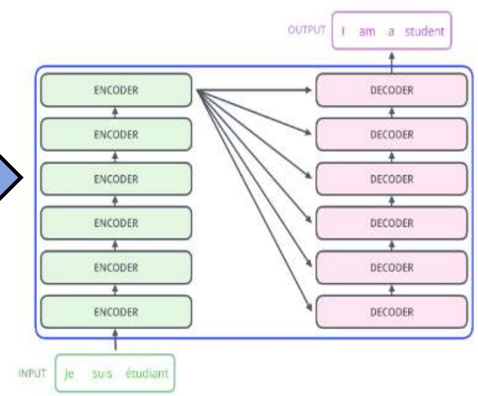
Transformers
2017

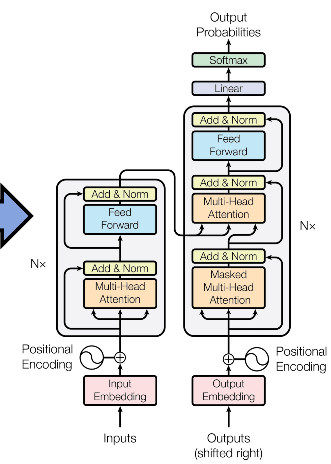
RNNs
1986

Bidirectional RNNs
1997

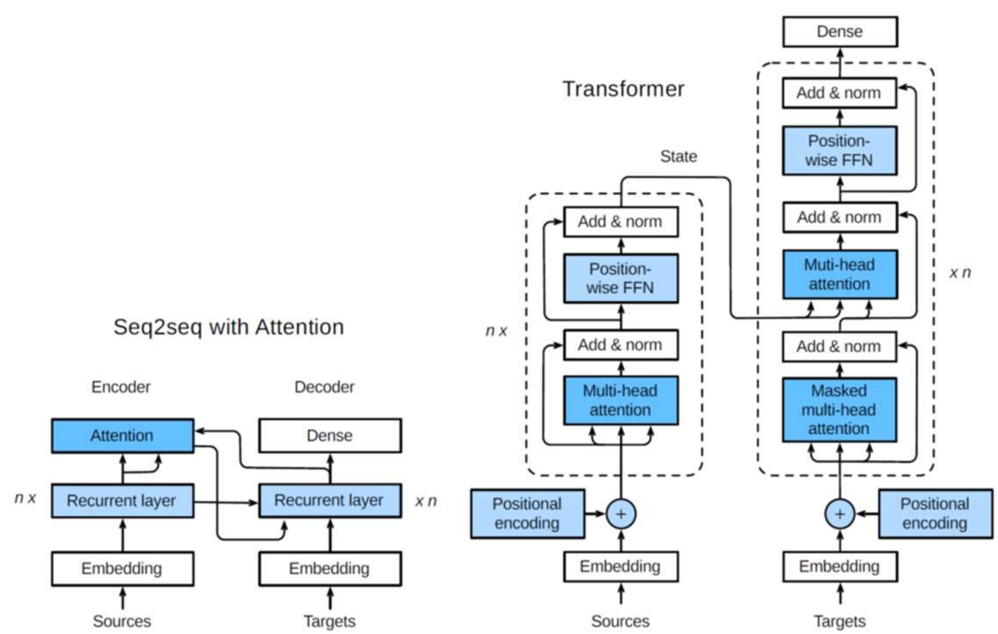
Encoder-Decoder RNNs
2014







FROM ATTENTION TO TRANSFOMERS



Machine learning paradigms underlying ChatGPT

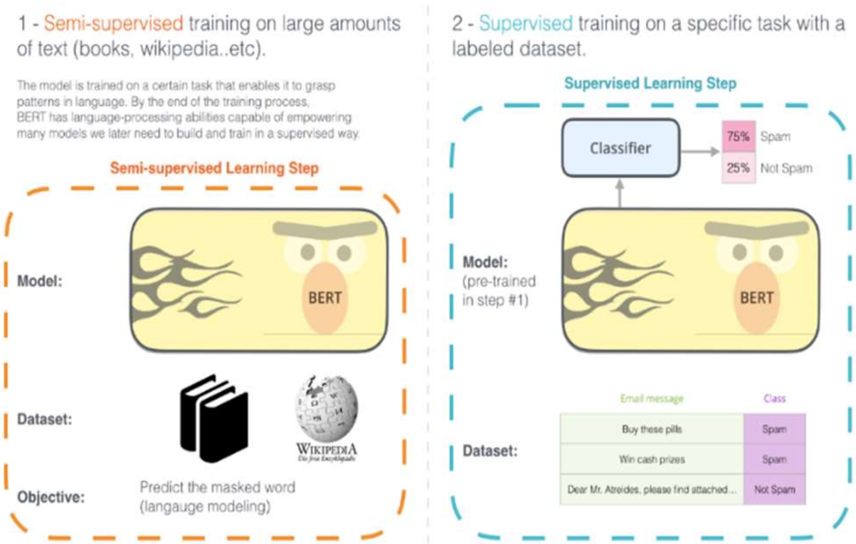
Transformers
2017

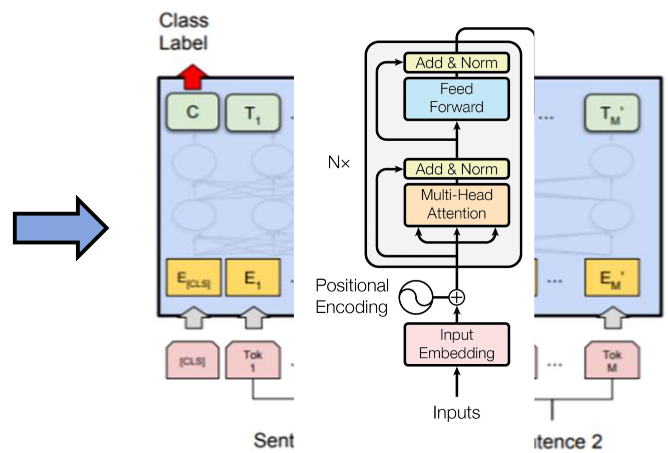
RNNs
1986

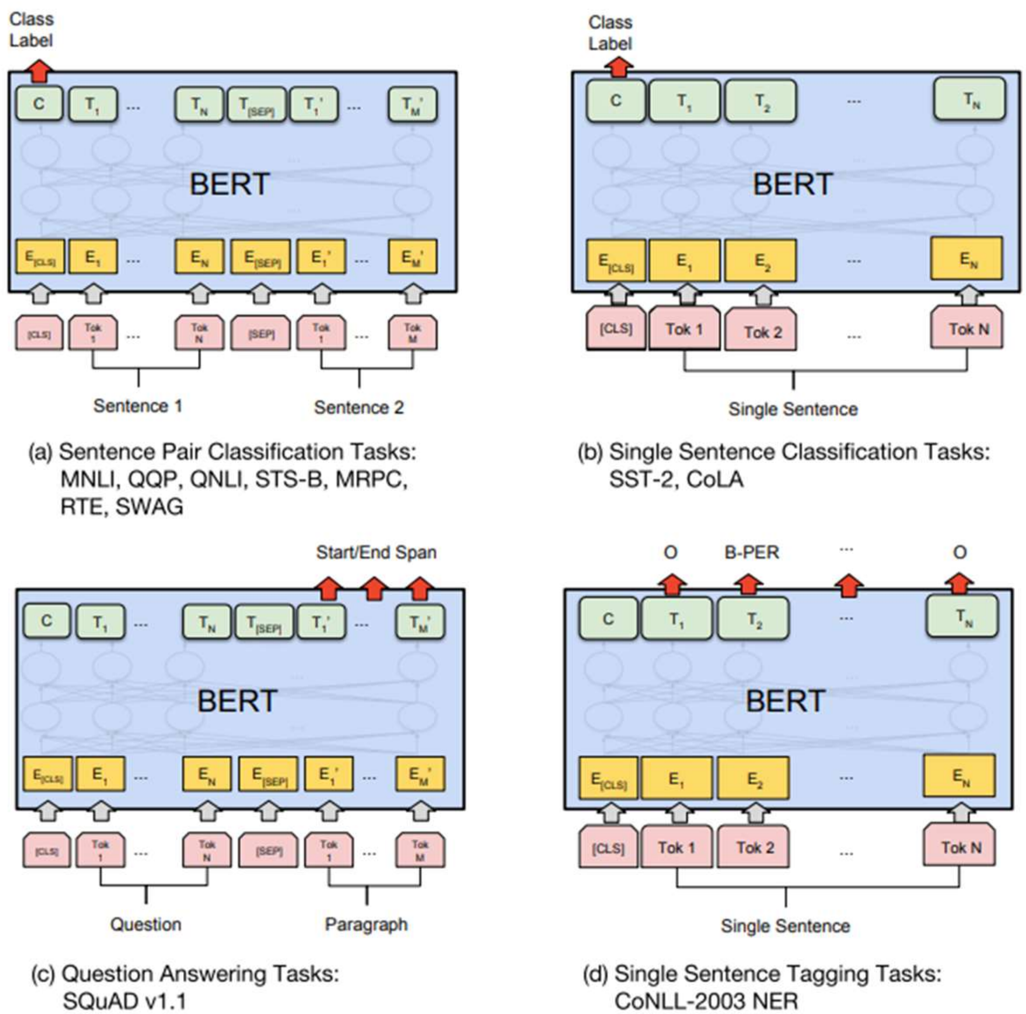
Bidirectional
RNNs
1997

Encoder-Decoder
RNNs
2014

BERT
2018



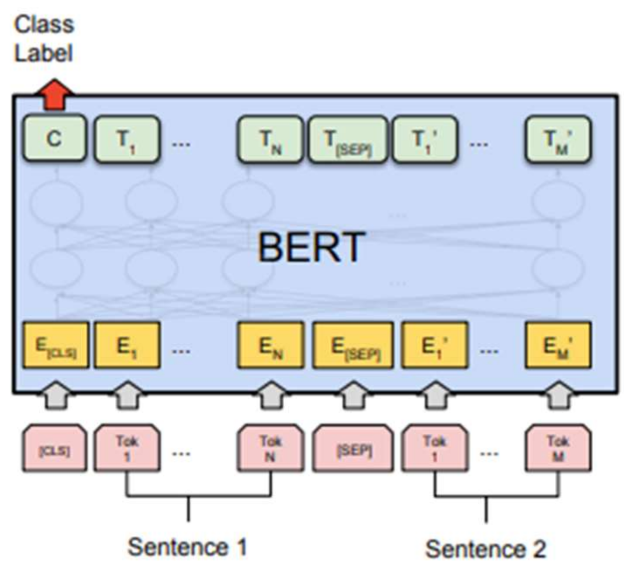



Class Label
BERT
Sentence 1
Sentence 2
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG
Class Label
BERT
Tok 1
Tok 2
Tok N
Single Sentence
(b) Single Sentence Classification Tasks: SST-2, CoLA
Start/End Span
BERT
Question
Paragraph
(c) Question Answering Tasks: SQuAD v1.1
O
B-PER
O
BERT
Tok 1
Tok 2
Tok N
Single Sentence
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

THE ROLE OF TRASFORMERS

- First setting
  - $h(A_i, B_j) = true$ iff $\{\Delta, A_i\} \Vdash B_j$
  - Input given by 2 sentences
  - BERT used as the encoder
  - A stacked classifier is trained on labeled pairs

  - Type of Inference:
    - PARAPHRASING
    - TEXTUAL ENTAILMENT

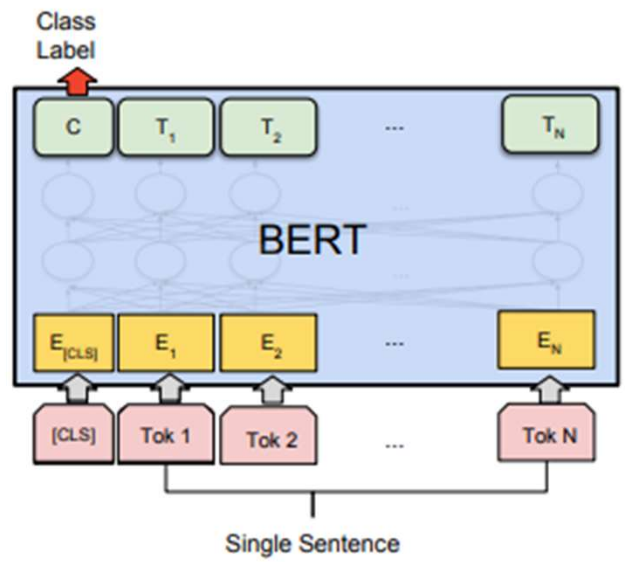


(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

THE ROLE OF TRASFORMERS (2)

- Second setting
  - $h(A_i \rightarrow B_j) = true$ iff $\{\Delta, A_i\} \Vdash B_j$
  - Input given 1 sentence expressing the task over A_i and B_j
  - BERT used as the encoder
  - A stacked classifier is trained on labeled pairs

  - Example (PARAPHRASING):
  - «The sentence B_j has the same meaning of sentence A_i»
  - «Sentence A_i means the same as B_j»

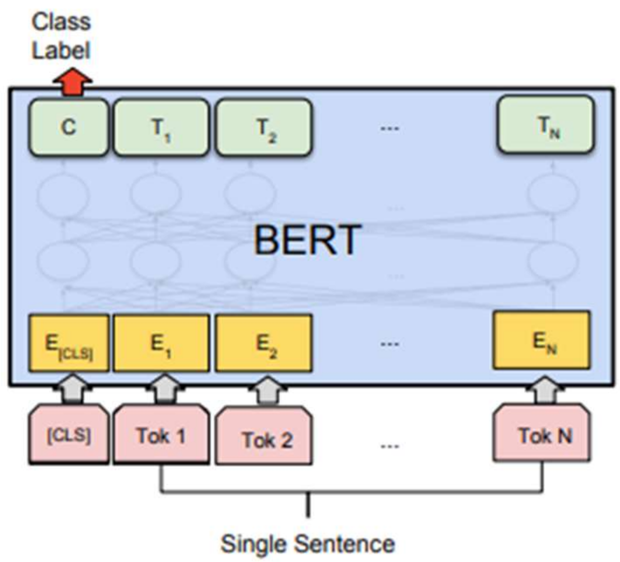


(b) Single Sentence Classification Tasks: SST-2, CoLA

THE ROLE OF TRASFORMERS (3)

- Second setting
  - $h(A_i \rightarrow B_j) = true$ iff $\{\Delta, A_i\} \Vdash B_j$
  - Input given 1 sentence expressing the task over A_i and B_j
  - BERT used as the encoder
  - A stacked classifier is trained on labeled pairs

  - Example (TEXTUAL ENTAILMENT):
  - «The sentence B_j is implied by sentence A_i»
  - «Sentence A_i guarantees the truth of B_j»



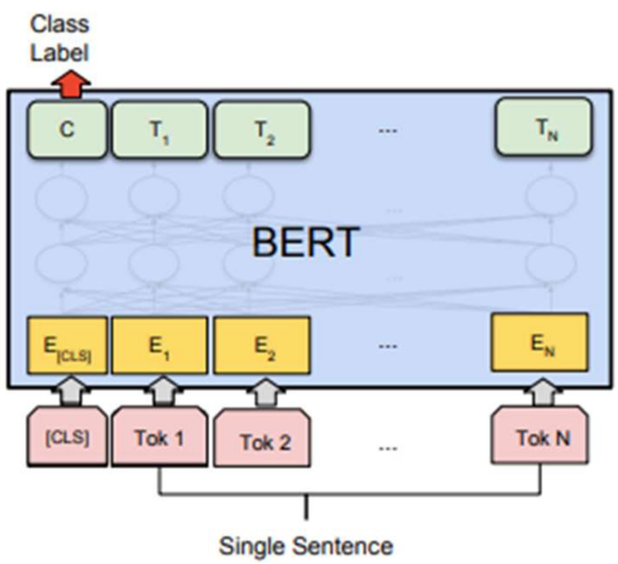
(b) Single Sentence Classification Tasks: SST-2, CoLA

NEURAL ENTAILMENT: APPLICATIONS

- The setting

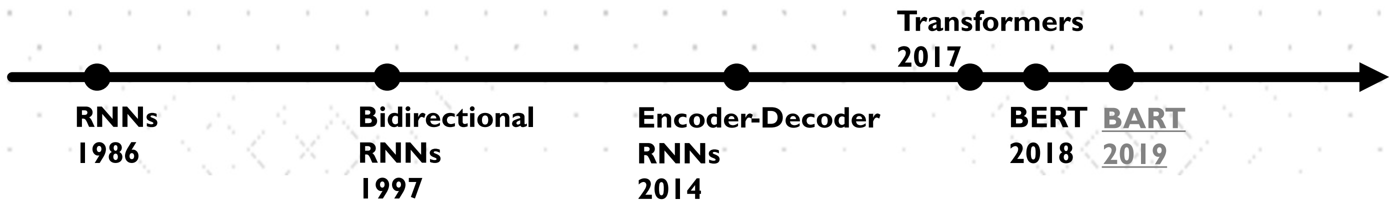
$$h(A_i \rightarrow B_j) = true \text{ iff } \{\Delta, A_i\} \Vdash B_j$$

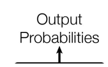
- correspond to sentences that depend on complex interactions between A_i and B_j mapped into an individual sentences
  - BERT is always used as the encoder
  - The stacked classifier is an automatic entailment recognition tool
  - It can be preserved for future TEXTUAL ENTAILMENT tasks, e.g., :
  - Topical Classification
    - «The sentence B_j is classified by label A_i»
    - «Label A_i corresponds to the topic of B_j»
  - Sentiment Analysis:
    - «A_i implies the sentiment label B_j»
    - «A_i expresses sentiment B_j»

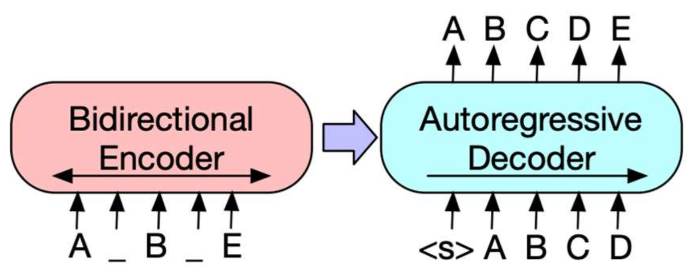


(b) Single Sentence Classification Tasks: SST-2, CoLA

Machine learning paradigms underlying ChatGPT







GPT-2: DECODER ONLY ARCHITECTURES (RADFORD ET AL., 2019)

- "We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText"
- GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages.
- GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text.
- The diversity of the dataset causes this simple goal to contain naturally occurring demonstrations of many tasks across diverse domains.
- GPT-2 is a direct scale-up of GPT, with more than 10X the parameters and trained on more than 10X the amount of data

GPT-2: SOURCES OF INSIPIRATION

- Multitask QA Networks (MQAN) (McCann et al, 2018)

Examples

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

Figure 1: Overview of the decaNLP dataset with one example from each decaNLP task in the order presented in Section 2. They show how the datasets were pre-processed to become question answering problems. Answer words in red are generated by pointing to the context, in green from the question, and in blue if they are generated from a classifier over the output vocabulary.

- Our speculation is that a language model with sufficient capacity will begin to learn to infer and perform the tasks demonstrated in natural language sequences in order to better predict them, regardless of their method of procurement. If a language model is able to do this it will be, in effect, performing unsupervised multitask learning.

GPT-2: ARCHITECTURE (2)

- From (Radford et al., 2017, GPT paper)

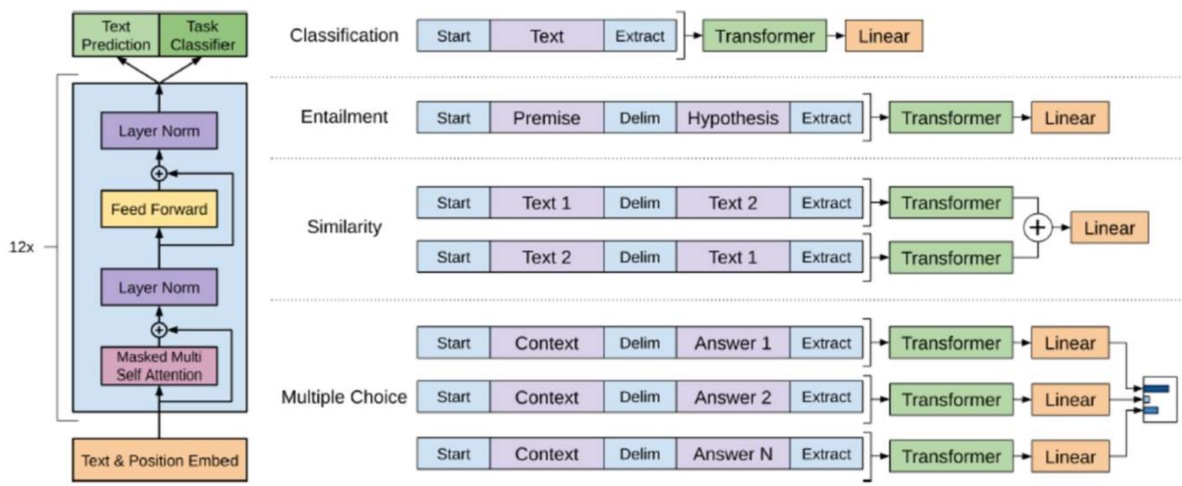


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

GPT-2: RESULTS

Language Models are Unsupervised Multitask Learners

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

- The LAMBADA dataset (Paperno et al., 2016)
  - It tests the ability of systems to model long-range dependencies in text.
  - The task is to predict the final word of sentences which require at least 50 tokens of context for a human to successfully predict.

GPT-2: RESULTS ON LAMBADA

- The LAMBADA dataset (Paperno et al., 2016)
  - It tests the ability of systems to model long-range dependencies in text.
  - The task is to predict the final word of sentences which require at least 50 tokens of context for a human to successfully predict.

(1) *Context:* "Yes, I thought I was going to lose the baby." "I was scared too," he stated, sincerity flooding his eyes. "You were ?" "Yes, of course. Why do you even ask?" "This baby wasn't exactly planned for."
Target sentence: "Do you honestly think that I would want you to have a _____ ?"
Target word: miscarriage

(2) *Context:* "Why?" "I would have thought you'd find him rather dry," she said. "I don't know about that," said Gabriel. "He was a great craftsman," said Heather. "That he was," said Flannery.
Target sentence: "And Polish, to boot," said _____.
Target word: Gabriel

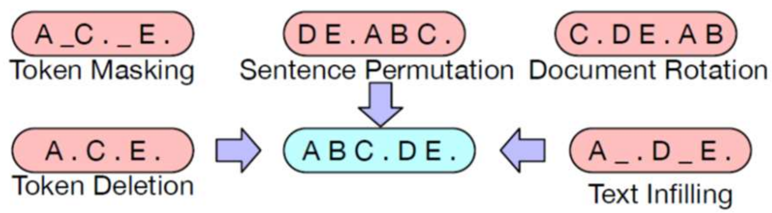
(3) *Context:* Preston had been the last person to wear those chains, and I knew what I'd see and feel if they were slipped onto my skin-the Reaper's unending hatred of me. I'd felt enough of that emotion already in the amphitheater. I didn't want to feel anymore. "Don't put those on me," I whispered. "Please."
Target sentence: Sergei looked at me, surprised by my low, raspy please, but he put down the _____.
Target word: chains

(4) *Context:* They tuned, discussed for a moment, then struck up a lively jig. Everyone joined in, turning the courtyard into an even more chaotic scene, people now dancing in circles, swinging and spinning in circles, everyone making up their own dance steps. I felt my feet tapping, my body wanting to move.
Target sentence: Aside from writing, I 've always loved _____.
Target word: dancing

- GPT-2 improves the state of the art from 99.8 (Grave et al., 2016) to 8.6 perplexity and increases the accuracy of LMs on this test from 19% (Dehghani et al., 2018) to 52.66%. Adding a stop-word filter as an approximation to this further increases accuracy to 63.24%.
- Investigating GPT-2's errors showed most predictions are valid continuations of the sentence, but are not valid final words

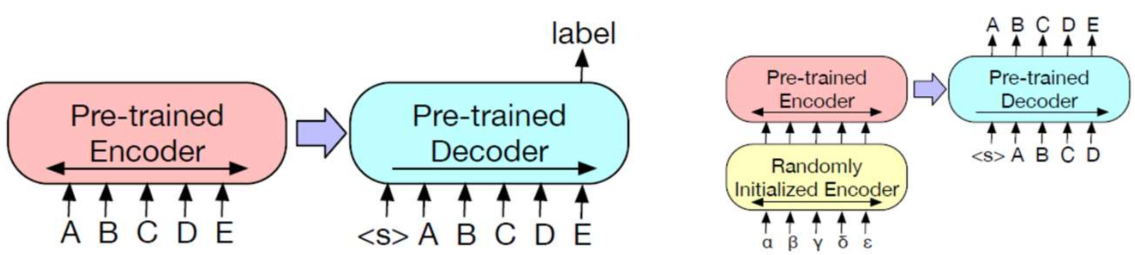
BART (LEWIS ET AL., 2019) - FACEBOOK

- Enconding decoding architecture based on Pretraining and fine tuned towards different tasks such as: RTE, SA, …
- Two stages of PRETRAINING
  - Text is first corrupted with an arbitrary noising function,
  - A sequence-to-sequence model is learned to reconstruct the original text.



- FINE TUNING:
  - **MNLI** (Williams et al., 2017), a bitext classification task to predict whether one sentence **entails** another. The fine-tuned model concatenates the two sentences with appended an EOS token, and passes them to both the BART encoder and decoder. In contrast to BERT, the representation of the EOS token is used to classify the sentences relations.
  - ELI5 (Fan et al., 2019), a long-form abstractive question answering dataset. Models generate answers conditioned on the concatenation of a question and supporting documents.

APPLYING BART



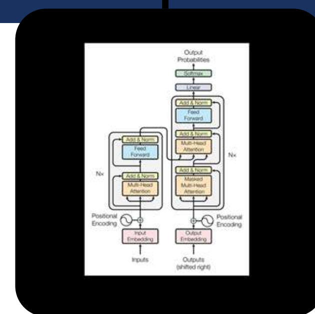
(a) To use BART for classification problems, the same input is fed into the encoder and decoder, and the representation from the final output is used.

(b) For machine translation, we learn a small additional encoder that replaces the word embeddings in BART. The new encoder can use a disjoint vocabulary.

Figure 3: Fine tuning BART for classification and translation.

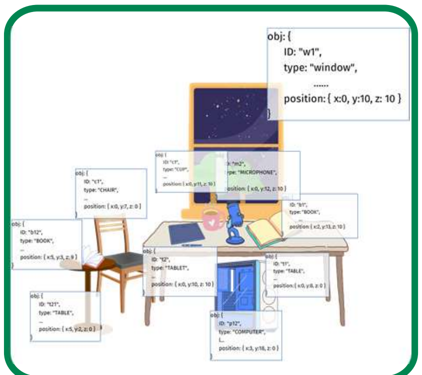
GRUT: THE OVERALL FLOW

Output:
`TAKING(Theme(b1))`



GrUT-IT

Command: *“Prendi il volume sul tavolo vicino la finestra"*



Entities Retrieval

Linguistic Extraction

Input: *Command + MD*

MD: *b1, conosciuto anche come libro o volume, è un’istanza della classe BOOK, t1, conosciuto anche come tavolo o scrivania, è un’istanza della classe TABLE # b1 è vicino t1*

Hromei et al, 2022, "Embedding Contextual Information in Seq2seq Models for Grounded Semantic Role Labeling"

EXPERIMENTAL EVALUATION

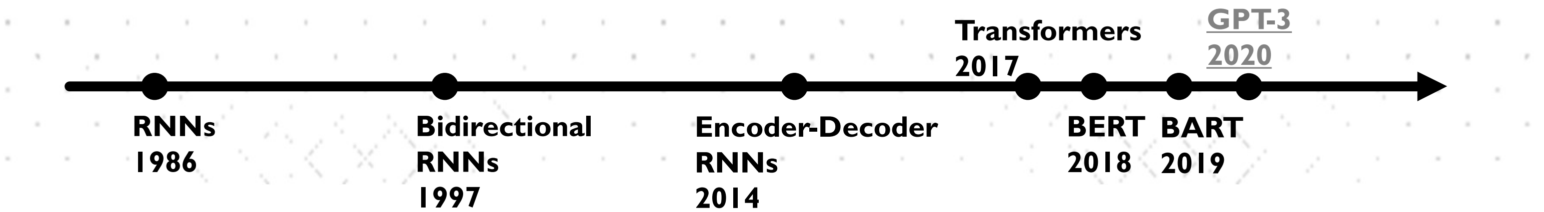
FP = Frame Prediction
AIC = Argument Identification and Classification
EM = Exact Match
HM = Head Match

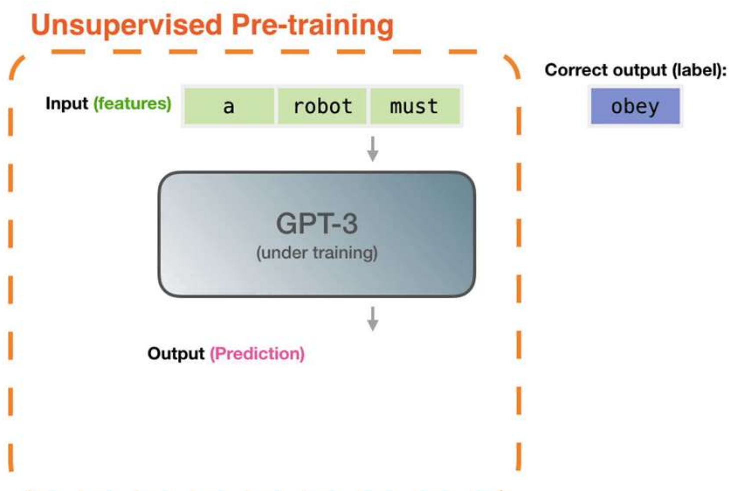
Model	Learning Rate	FP	AIC-Exact Match	AIC-Head Match
LU4R	-	*95.32%*	*77.67%*	*86.35%*
GrUT-IT	$5 \cdot 10^{-5}$	96.86%	82.30%	85.19%

LU4R: `TAKING(Theme("libro"))`
GrUT-IT: `TAKING(Theme(b1))`

Results here are reported as F1 values on 10-fold cross-validation schema with 80/10/10 data split.
Performance for LU4R is reported in *italic* as it is not entirely comparable with.

Machine learning paradigms underlying ChatGPT





GPT3: NOVELTY

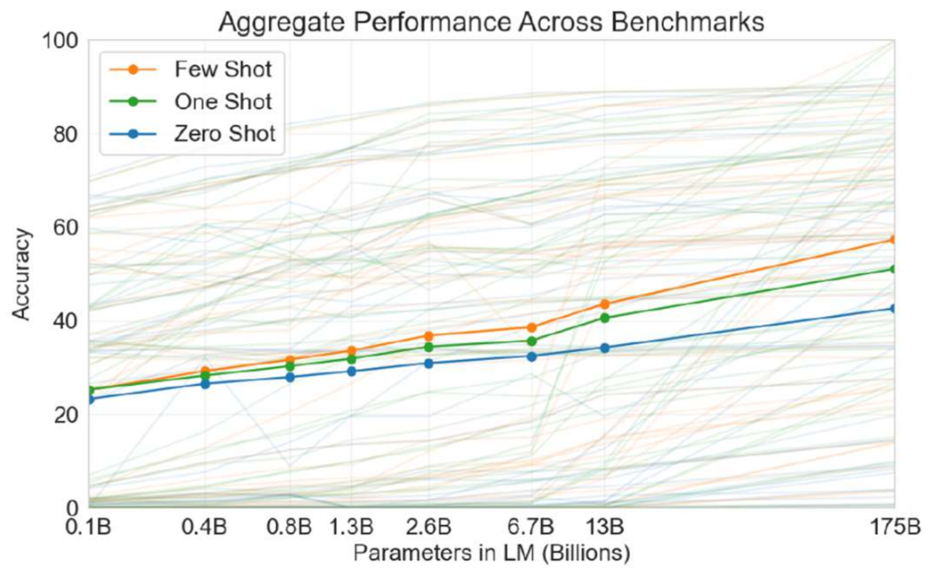
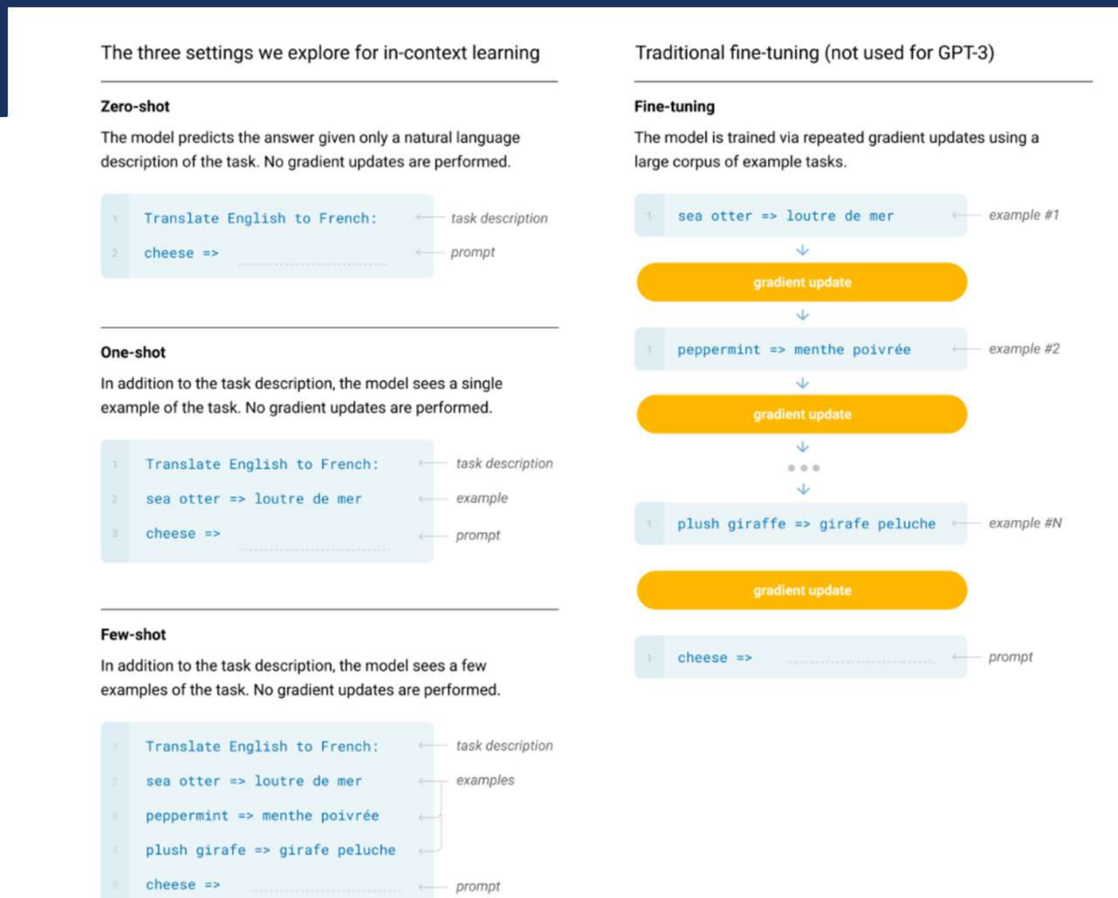


Figure 1.3: Aggregate performance for all 42 accuracy-denominated benchmarks While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning. See Figure 3.8 for a more detailed analysis on SuperGLUE, a standard NLP benchmark suite.



The three settings we explore for in-context learning
Zero-shot
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.
Translate English to French:
cheese =>
task description
prompt
One-shot
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.
Translate English to French:
sea otter => loutre de mer
cheese =>
task description
example
prompt
Few-shot
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
task description
examples
prompt
Traditional fine-tuning (not used for GPT-3)
Fine-tuning
The model is trained via repeated gradient updates using a large corpus of example tasks.
sea otter => loutre de mer
example #1
gradient update
peppermint => menthe poivrée
example #2
gradient update
plush giraffe => girafe peluche
example #N
gradient update
cheese =>
prompt

GPT-3: SIZE

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

- Here n_{params} is the total number of trainable parameters, n_{layers} is the total number of layers, d_{model} is the number of units in each bottleneck layer (we always have the feedforward layer four times the size of the bottleneck layer, d_{ff}=4xd_{model}), and d_{head} is the dimension of each attention head.
- All models use a context window of n_{ctx} = 2048 tokens

Machine learning paradigms underlying ChatGPT

RNNs 1986

Bidirectional RNNs 1997

Encoder-Decoder RNNs 2014

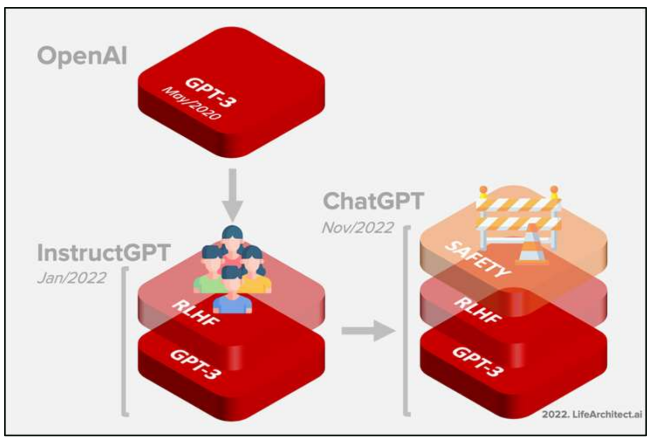
Transformers 2017

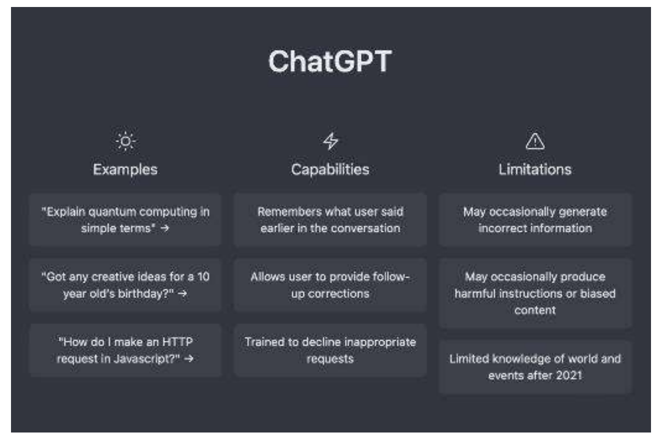
BERT 2018

BART 2019

GPT-3 2020

ChatGPT 2022





LIMITATIONS OF GPT-3

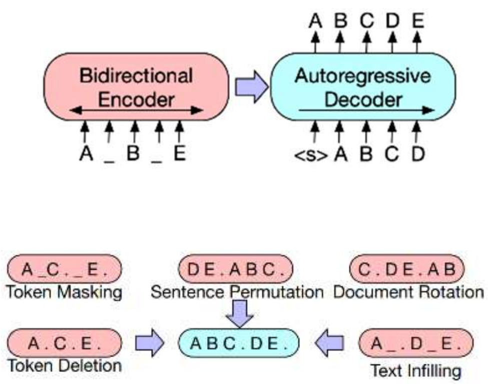
- Large language models often express unintended behaviors such as making up facts, generating biased or toxic text, or simply not following user instructions. This is because the language modeling objective is misaligned.
- The idea: aligning language models by training them to act in accordance with the user's intention (Leike et al., 2018).
  - explicit intentions such as following instructions
  - implicit intentions such as staying truthful, and not being biased, toxic, or otherwise harmful.
- Overall Objective: language models should be helpful (they should help the user solve their task), honest (they shouldn't fabricate information or mislead the user), and harmless (they should not cause physical, psychological, or social harm to people or the environment).

INSTRUCTGPT

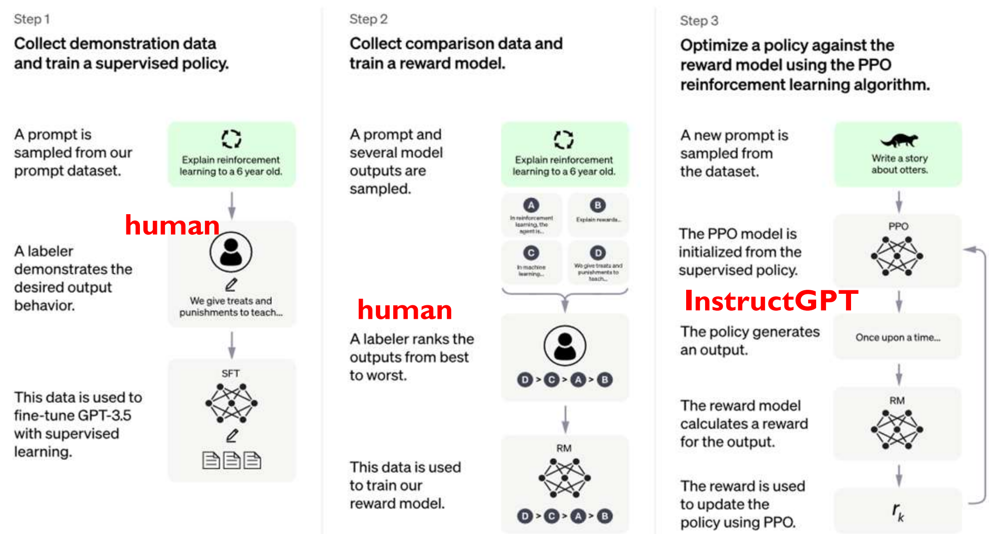
- **Step 1**: Collect demonstration data, and train a supervised policy. Labelers provide demonstrations of the desired behavior on the input prompt distribution. Then, fine-tuning of a pretrained GPT-3 model on this data using supervised learning is carried out.
- **Step 2**: Collect comparison data, and train a reward model. A dataset of comparisons between model outputs is collected: labelers indicate which output they prefer for a given input. A reward model to predict the human-preferred output is then trained.
- **Step 3:** Optimize a policy against the reward model using PPO. We use the output of the RM as a scalar reward. We fine-tune the supervised policy to optimize this reward using the proximal policy optimization (PPO) algorithm (Schulman et al., 2017).

At the heart of ChatGPT (from BART to ChatGPT)

BART Training-steps



ChatGPT Training-steps

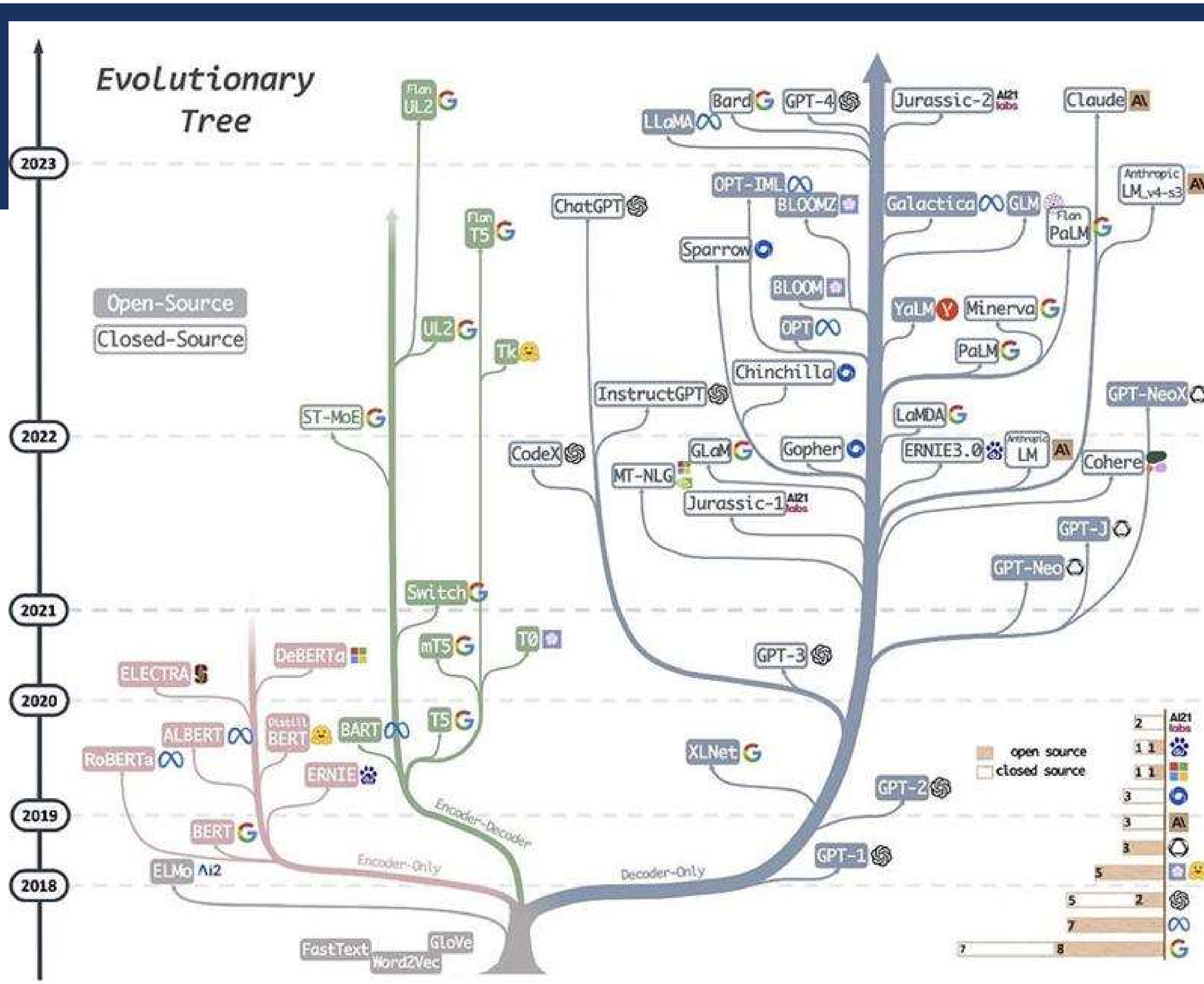


Fine tune text-davinci-003 to get InstructGPT

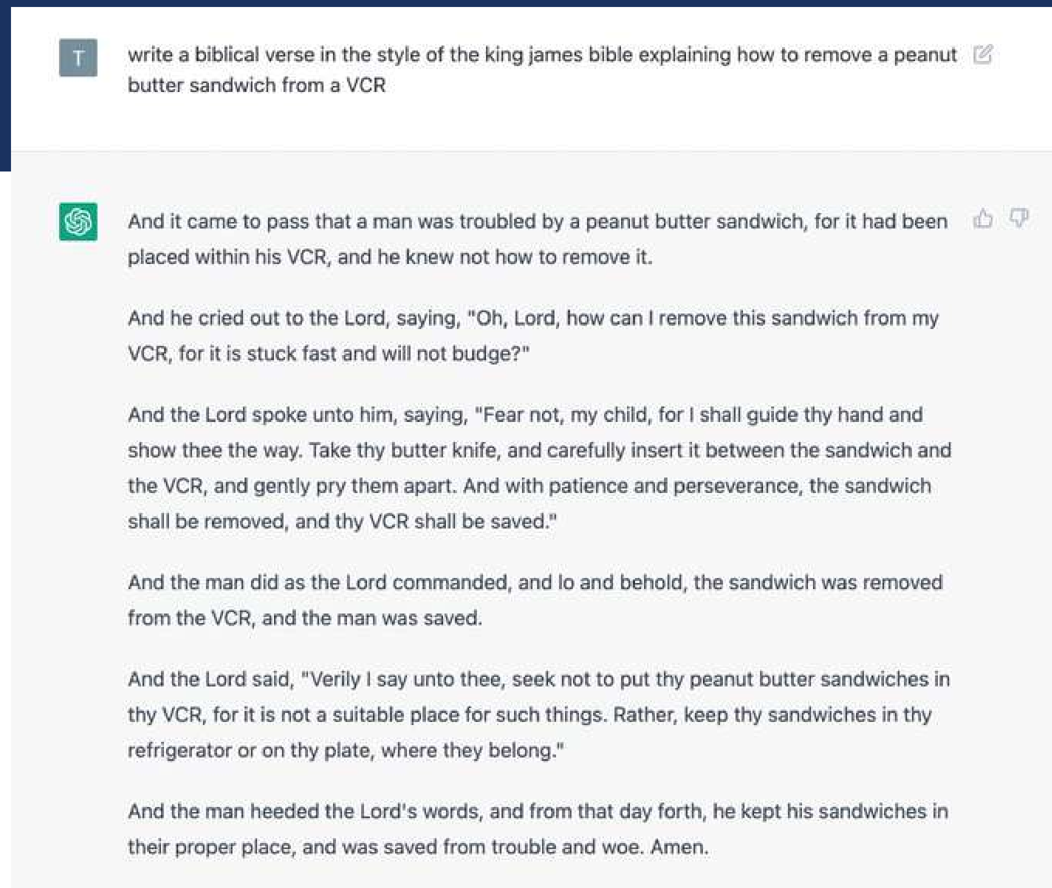
The Environment

from Ouyang, L., Wu, J., Jiang, et al. (2022). *Training language models to follow instructions with human feedback*

FOUNDATIONAL MODELS



NEURAL WRITING



T

write a biblical verse in the style of the king james bible explaining how to remove a peanut butter sandwich from a VCR

And it came to pass that a man was troubled by a peanut butter sandwich, for it had been placed within his VCR, and he knew not how to remove it.

And he cried out to the Lord, saying, "Oh, Lord, how can I remove this sandwich from my VCR, for it is stuck fast and will not budge?"

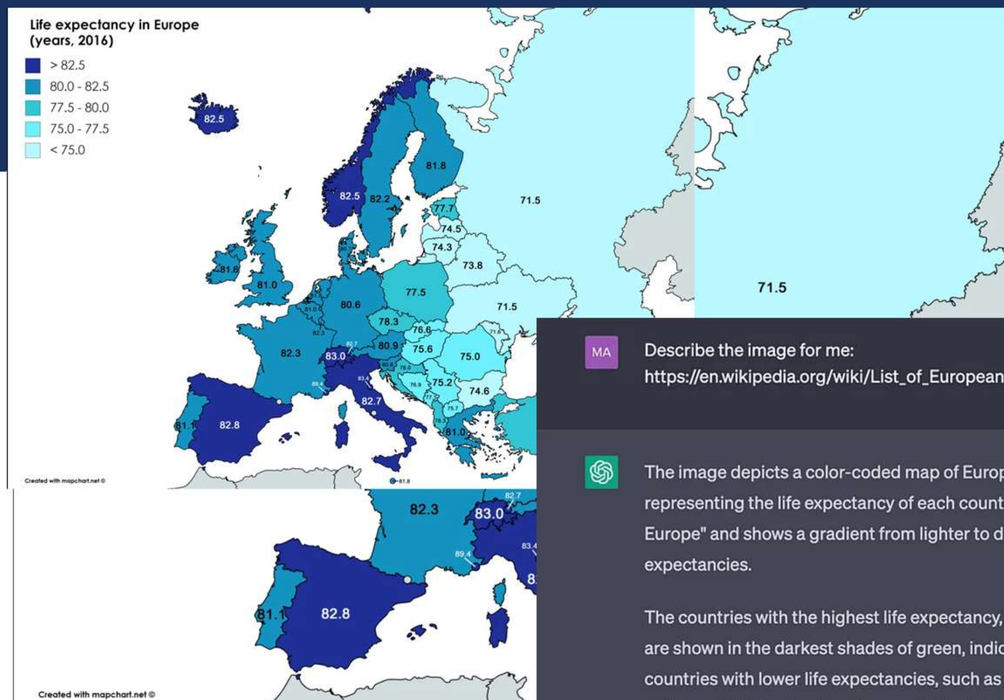
And the Lord spoke unto him, saying, "Fear not, my child, for I shall guide thy hand and show thee the way. Take thy butter knife, and carefully insert it between the sandwich and the VCR, and gently pry them apart. And with patience and perseverance, the sandwich shall be removed, and thy VCR shall be saved."

And the man did as the Lord commanded, and lo and behold, the sandwich was removed from the VCR, and the man was saved.

And the Lord said, "Verily I say unto thee, seek not to put thy peanut butter sandwiches in thy VCR, for it is not a suitable place for such things. Rather, keep thy sandwiches in thy refrigerator or on thy plate, where they belong."

And the man heeded the Lord's words, and from that day forth, he kept his sandwiches in their proper place, and was saved from trouble and woe. Amen.

CHAT GPT4



MA

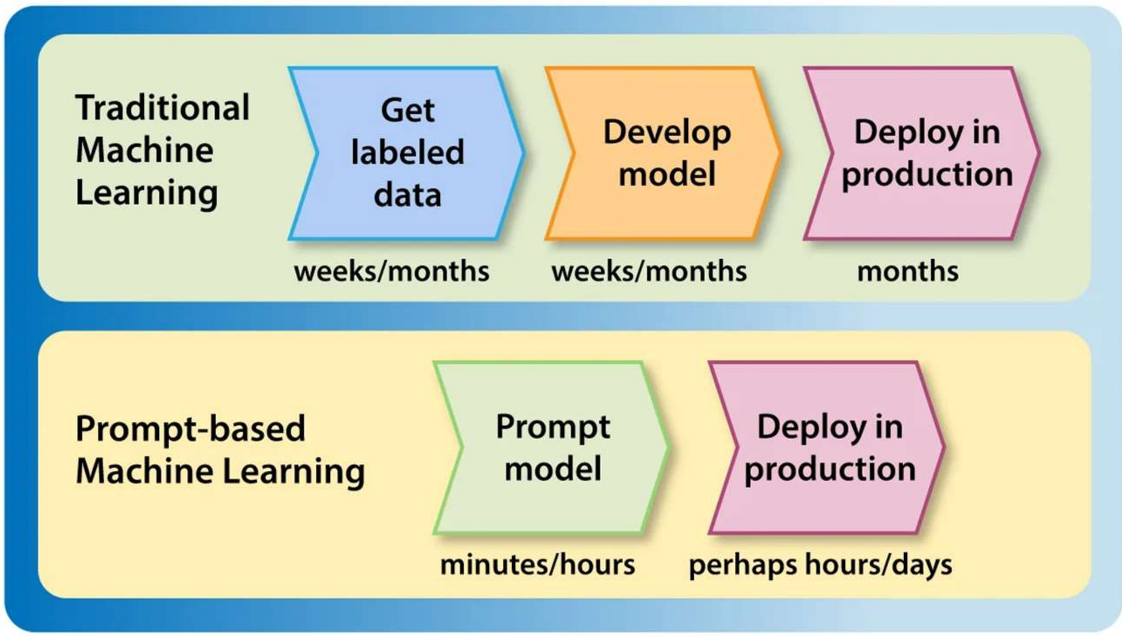
Describe the image for me:
https://en.wikipedia.org/wiki/List_of_European_countries_by_life_expectancy#/media/File:Life_expectancy_in_Europe.png

The image depicts a color-coded map of Europe, with different shades of green and yellow representing the life expectancy of each country. The map is titled "Life expectancy in Europe" and shows a gradient from lighter to darker shades of green, representing higher life expectancies.

The countries with the highest life expectancy, such as Switzerland, Spain, Italy, and France, are shown in the darkest shades of green, indicating life expectancies of over 80 years. The countries with lower life expectancies, such as Russia, Belarus, and Ukraine, are depicted in yellow and lighter shades of green, indicating life expectancies of under 75 years.

The map also includes a legend that provides a color-coded scale and corresponding life expectancy ranges, allowing viewers to easily interpret the data. Overall, the image provides a visual representation of the variation in life expectancy across different European countries, highlighting the disparities between countries and regions.

TRENDS …



Traditional Machine Learning
Get labeled data
Develop model
Deploy in production
weeks/months
weeks/months
months
Prompt-based Machine Learning
Prompt model
Deploy in production
minutes/hours
perhaps hours/days

BEYOND TRANSFORMER BIBLIOGRAPHY

- (Vaswani 2017), Attention is all you need, https://arxiv.org/abs/1706.03762
- (Devlin et al 2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://arxiv.org/abs/1810.04805
- Rocktaschel et al., "Reasoning About Entailment With Neural Attention" (ICLR 2016)
- T5: (Wolf et al, 2019) Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. ArXiv, abs/1910.03771.
- BART Encoding-Decoding: Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.https://arxiv.org/abs/1910.13461
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, "Improving Language Understanding by Generative Pre-Training", 2019
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei: Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. https://arxiv.org/abs/2005.14165, NeurIPS 2020.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, Paul F. Christiano: Learning to summarize with human feedback. NeurIPS 2022

BIBLIOGRAFIA: AN HISTORICAL OVERVIEW ON NNS

- Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4):115{133, 1943.
- Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6):386, 1958.
- Donald Olding Hebb. The organization of behavior: A neuropsychological theory. Psychology Press, 1949.
- John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences, 79(8):2554-2558, 1982.
- David E Rumelhart, Georey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.
- Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge.
- Teuvo Kohonen. The self-organizing map. Proceedings of the IEEE, 78(9):1464{1480, 1990.
- David H Ackley, Georey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. Cognitive science, 9(1):147-169, 1985.
- Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaected by shift in position. Biological cybernetics, 36(4): 193-202, 1980.
- Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in NIPS'89.

BIBLIOGRAFIA: AN HISTORICAL OVERVIEW ON NNS (2)

- Michael I Jordan. Serial order: A parallel distributed processing approach. Advances in psychology, 121:471-495, 1986.
- Jerey L Elman. Finding structure in time. Cognitive science, 14(2):179-211, 1990.
- AJ Robinson and Frank Fallside. The utility driven dynamic error propagation network. University of Cambridge Department of Engineering, 1987.
- Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11):2673-2681, 1997.
- Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735-1780, 1997.
- Hugo Larochelle and Georey E Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In Advances in neural information processing systems, pages 1243-1251, 2010
- Denil, M., Bazzani, L., Larochelle, H., and de Freitas, N. (2012). Learning where to attend with deep architectures for image tracking. Neural Computation, 24 (8), 2151–2184
- Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle, et al. Greedy layer-wise training of deep networks. Advances in neural information processing systems, 19:153, 2007.

BIBLIOGRAFIA: TRANSFORMERS

- (Vaswani 2017), Attention is all you need, https://arxiv.org/abs/1706.03762
- (Devlin et al 2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://arxiv.org/abs/1810.04805
- An interesting introduction to the attention mechanism:
  - **All you need to know about 'Attention' and 'Transformers' — In-depth Understanding — Part 1,** A. Sarkar, URL: https://towardsdatascience.com/all-you-need-to-know-about-attention-and-transformers-in-depth-understanding-part-1-552f0b41d021#4c16
- Other Task specific works:
  - Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473, 2014.
  - Effective Approaches to Attention-based Neural Machine Translation, Minh-Thang Luong Hieu Pham Christopher D. Manning, 2015, https://arxiv.org/abs/1508.04025v5
  - Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks. In International Conference on Learning Representations, 2017.

BIBLIOGRAFIA: *BEYOND TRANSFORMER*

- (Vaswani 2017), Attention is all you need, https://arxiv.org/abs/1706.03762
- (Devlin et al 2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://arxiv.org/abs/1810.04805
- Rocktaschel et al., "Reasoning About Entailment With Neural Attention" (ICLR 2016)
- T5: (Wolf et al, 2019) Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. ArXiv, abs/1910.03771.
- BART Encoding-Decoding: Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.https://arxiv.org/abs/1910.13461
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, "Improving Language Understanding by Generative Pre-Training", 2019
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei: Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. https://arxiv.org/abs/2005.14165, NeurIPS 2020.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, Paul F. Christiano: Learning to summarize with human feedback. NeurIPS 2022

GRAZIE DELL'ATTENZIONE

BASILI@INFO.UNIROMA2.IT