Got It AI
Transforming Conversational AI

Got It

PhotoStudy

Transformers, Large Scale Pre-trained Language Models & Applications in Conversational AI

Chandra Khatri, Chief Scientist and Head of AI Research

Agenda

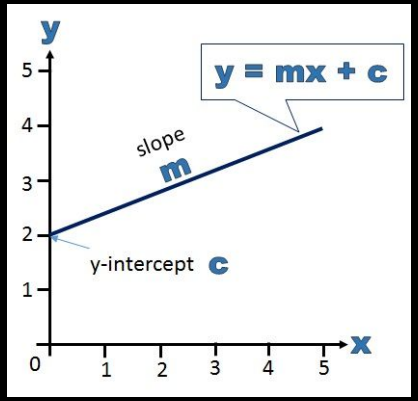
1. Background:
   a. Neural Networks
   b. Recurrent Neural Networks
   c. Language Models (LMs)
   d. Word Embeddings

2. Transformers:
   a. Building blocks of Transformers
   b. Large Scale Pre-trained LMs
   c. From Millions to Trillions of Parameters
   d. Different kinds of Transformers: Encoders, Decoders, Seq2Seq

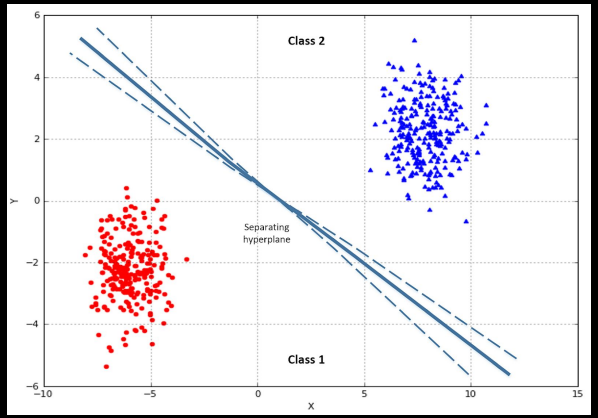
3. Applications:
   a. Transformers for NL Technologies: State of the Art and Recommendations
   b. Evolution of Conversational AI and how Transformers are democratizing the space
   c. Transformers at Got It AI

Neural Networks Background

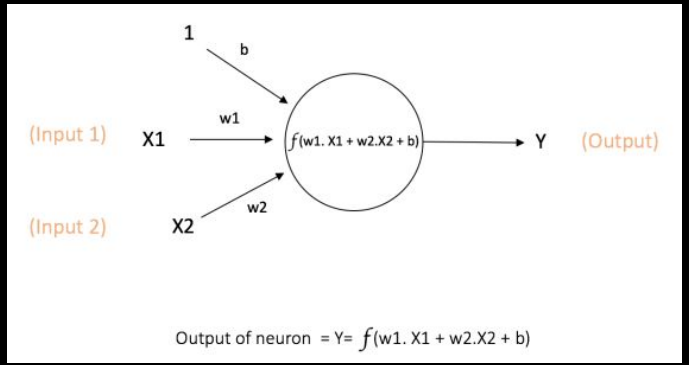
Fundamental Unit of Neural Networks and Deep Learning: Linear Unit



Linear Equation

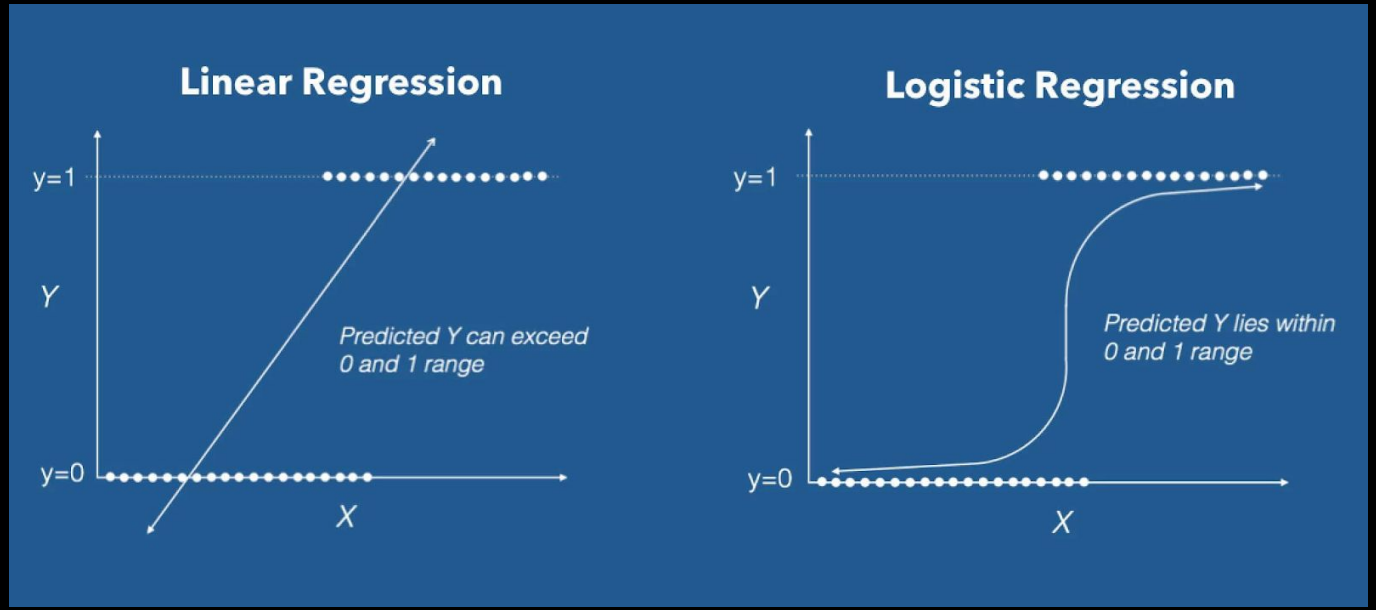


Linear Classification



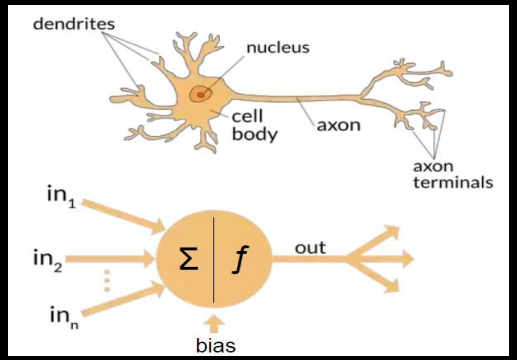
Mathematical representation of a Linear Unit / Perceptron

Linear Regression and Logistic Regression

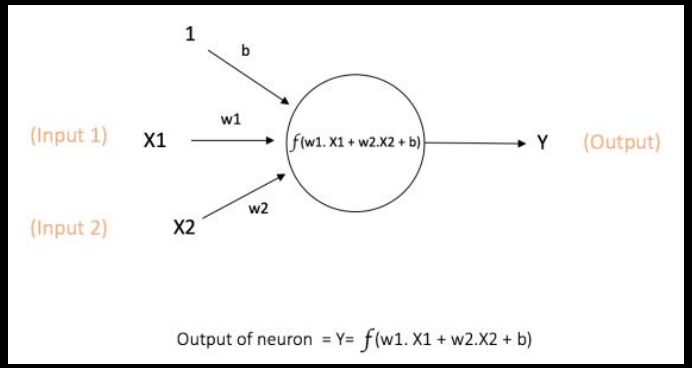


Neural Network

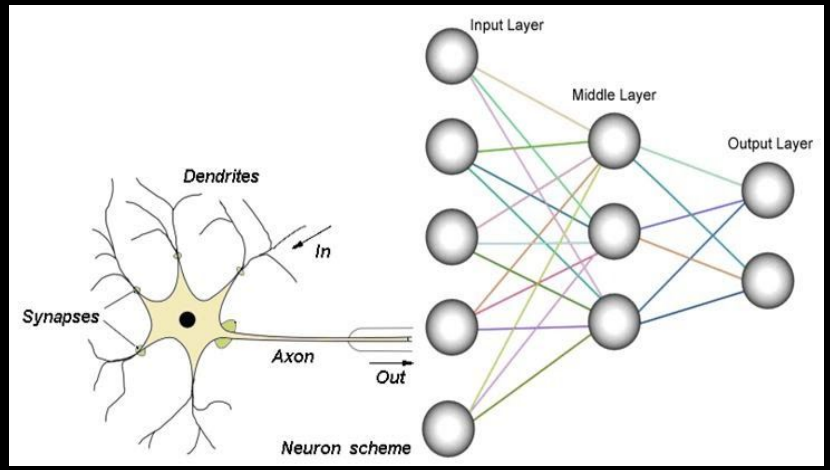
Neuron



Linear Unit/Perceptron

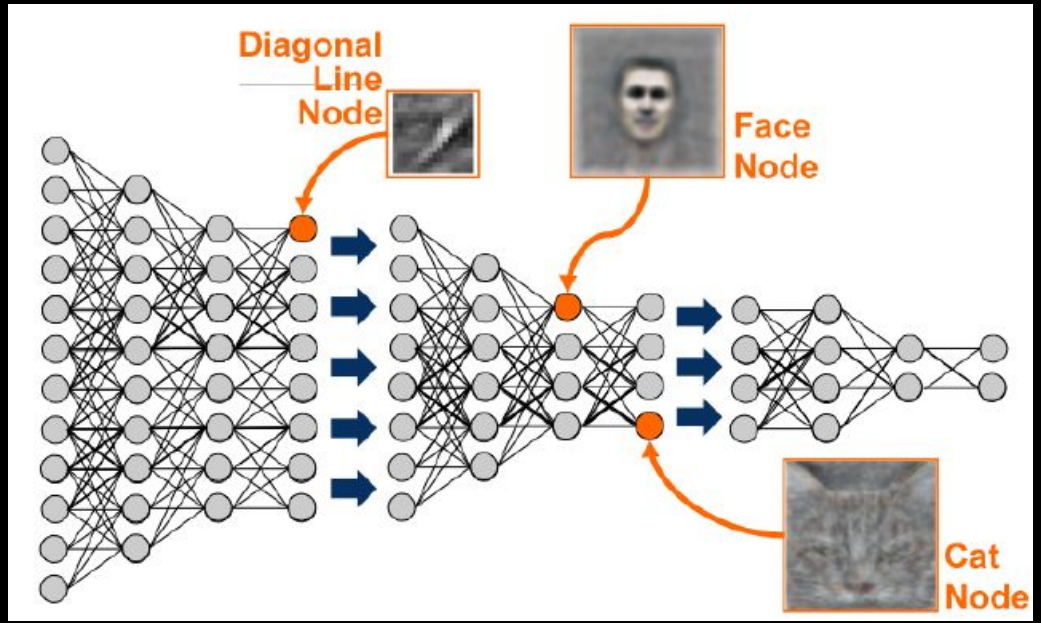


Neural Network: Multi-Layer Perceptrons



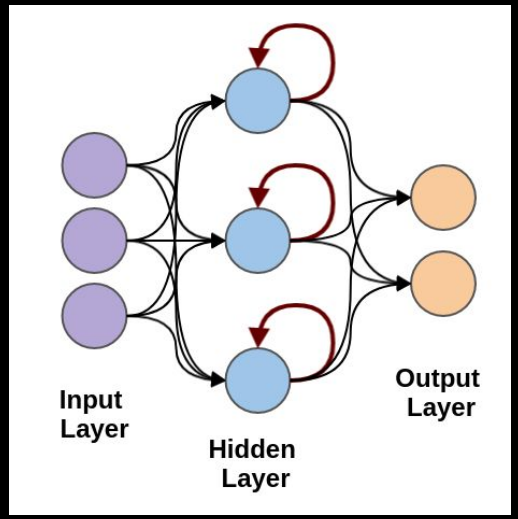
Deep Neural Networks

Fancy new term for Multi-layer Neural Networks with efficient ways of training



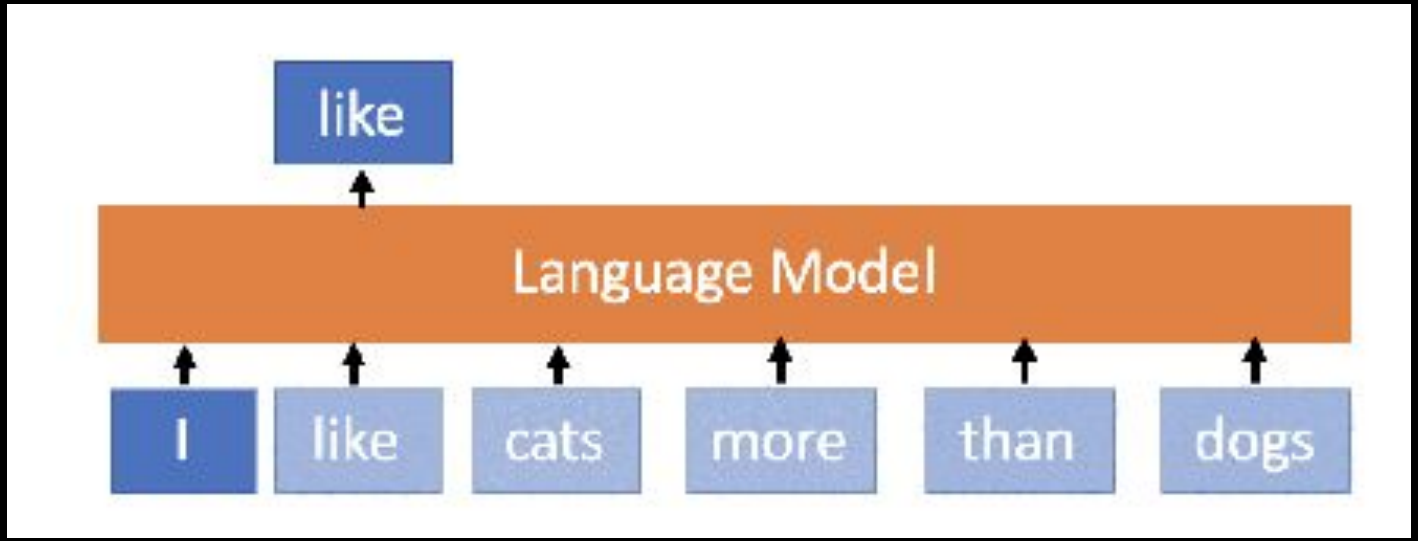
Recurrent Neural Network

Fancy RNN Architecture

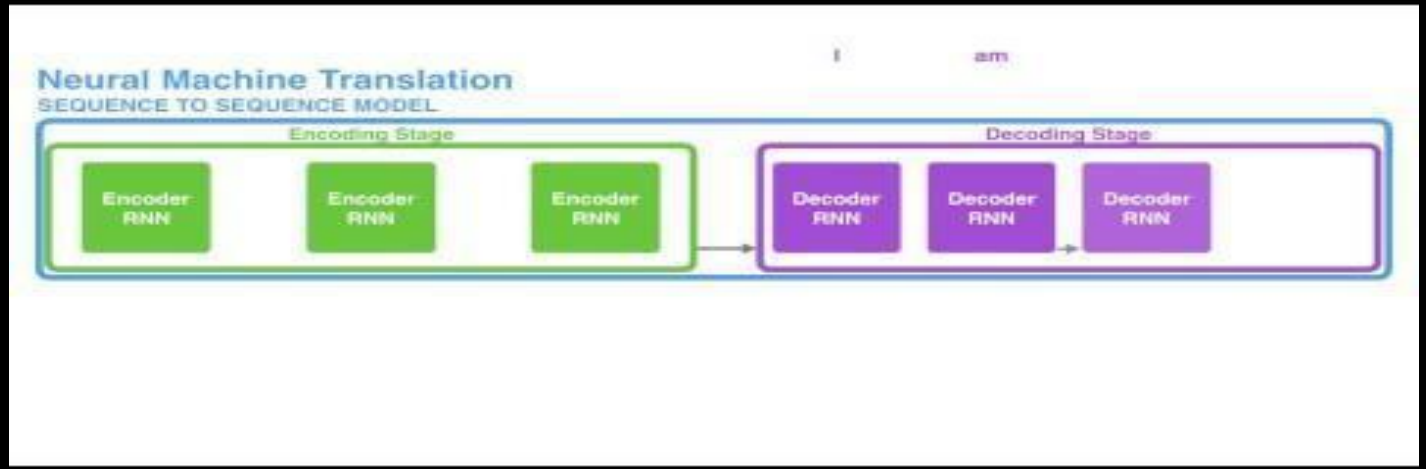


Reality: Memory, i.e. keep track of previous states for future prediction

Language Models

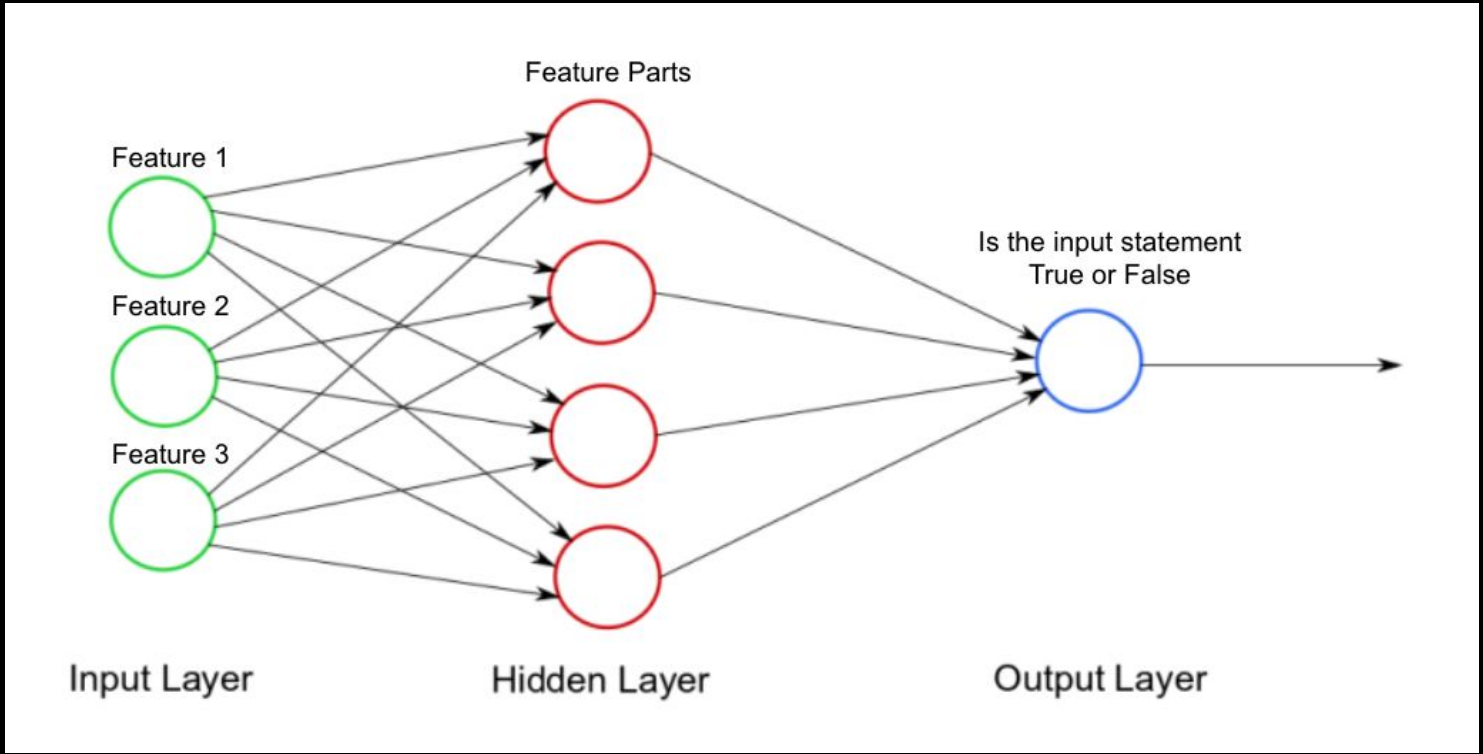


Seq2Seq Illustration



Source: https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Neural Network Features and Labels (Annotations)



Traditional NLP Featurization

Vocabulary Lookup Table

Sentence	the	is	visited	…	president	great	…	obama
Obama is the president of the us	1	1	0	…	1	0	…	1
Obama visited the great wall of china	1	0	1	…	0	1	..	1

One-hot encoding

Input/Sentence:
Obama is the president of the us

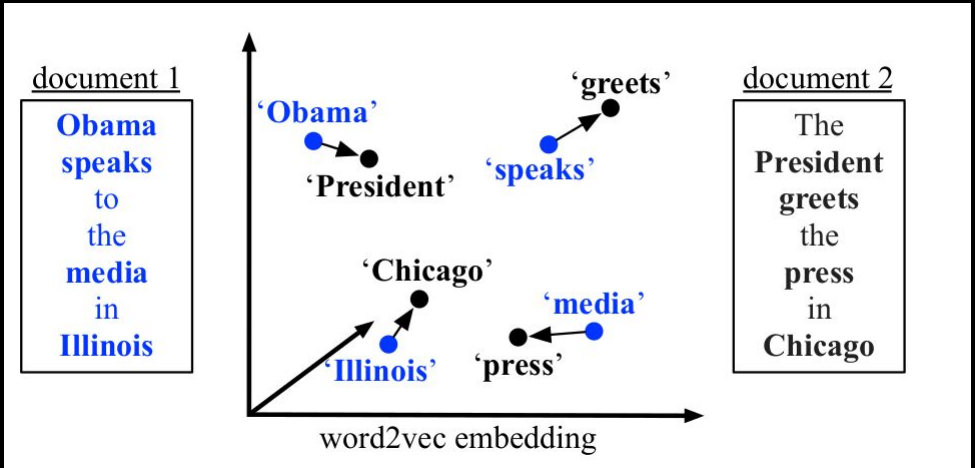
Features: Index of words appearing in the sentence

Vectorization: Initialize a vector of dimension V with all zeros except the ones appearing in the sentence
[1, 1, 0, 0, 1, 0, 1, 0, 0, 1, ….., 1]

This vector is the feature representation of the given sentence

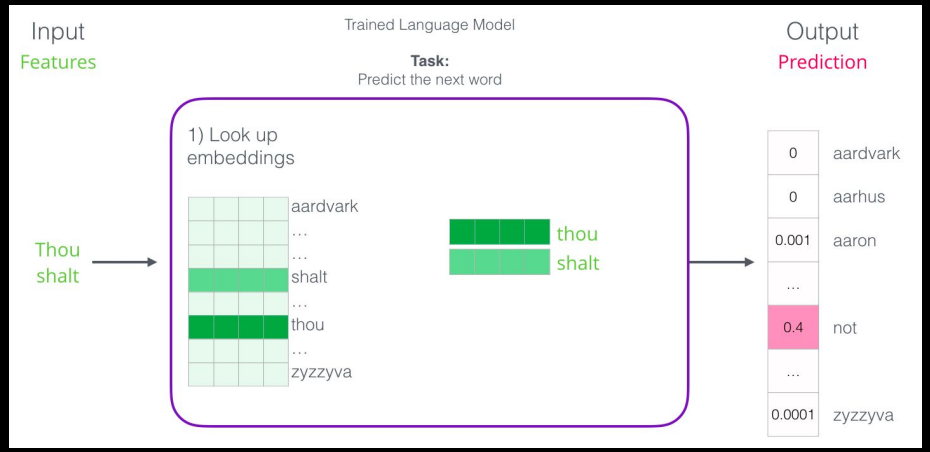
Word Embeddings: Semantic Repres. of Language Units

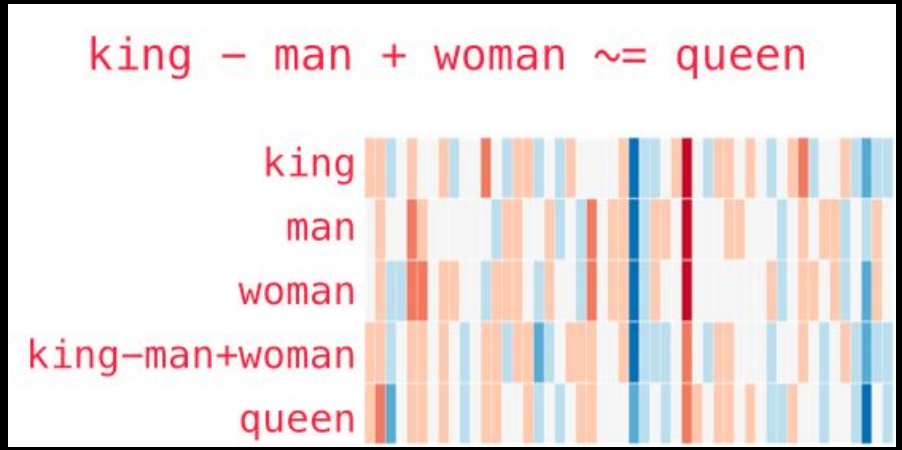
Word Embeddings



word	Dim 1	Dim 2	Dim …	Dim 300
obama	0.1	0.5	…	0.9
president	0.15	0.4	…	0.85
chicago	0.3	0.7	…	0.6
illinois	0.25	0.72	…	0.58
the	0.01	0.02	…	0.07
…				
press	…	…	…	…

Word2Vec Illustration





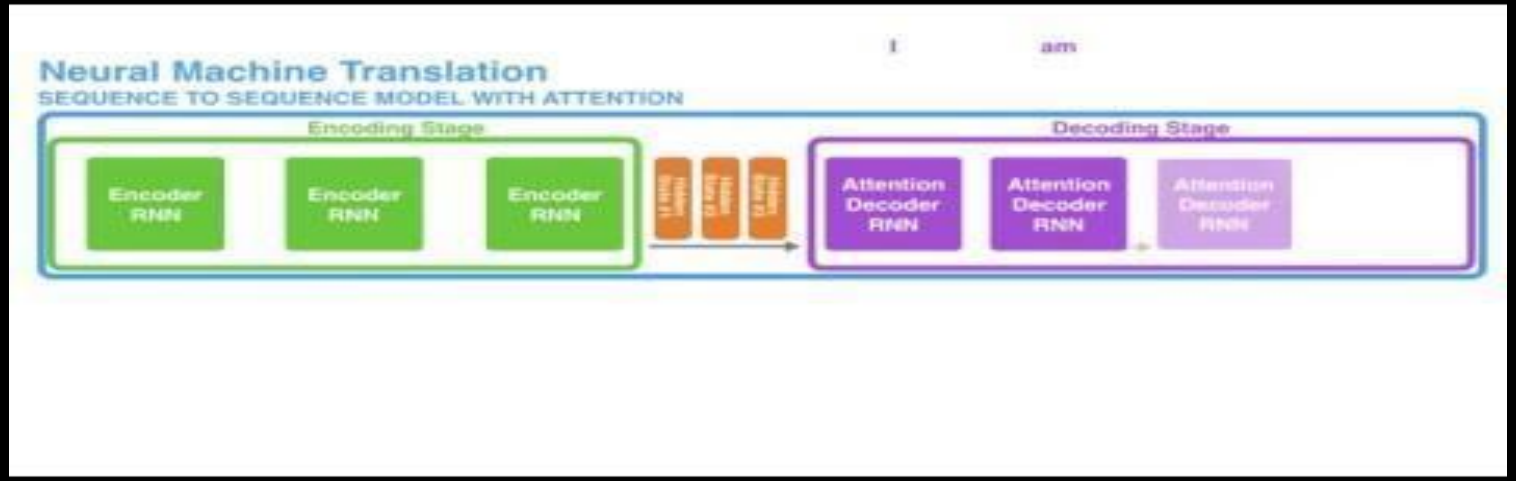
Are Word Embeddings contextual?

Transformers: Attention is all you need!

What are Transformers? Key Ideas

- Attention based Neural Networks
- Contextul Embeddings using:
  - Word position
  - Dynamic attention
- Auto-Regressive:
  - Use output generated at time “t” to generate at “t+1” time
- More Scalable:
  - Parallelizable: Vertical Architecture [RNN - Horizontal]
  - Faster, Deeper,

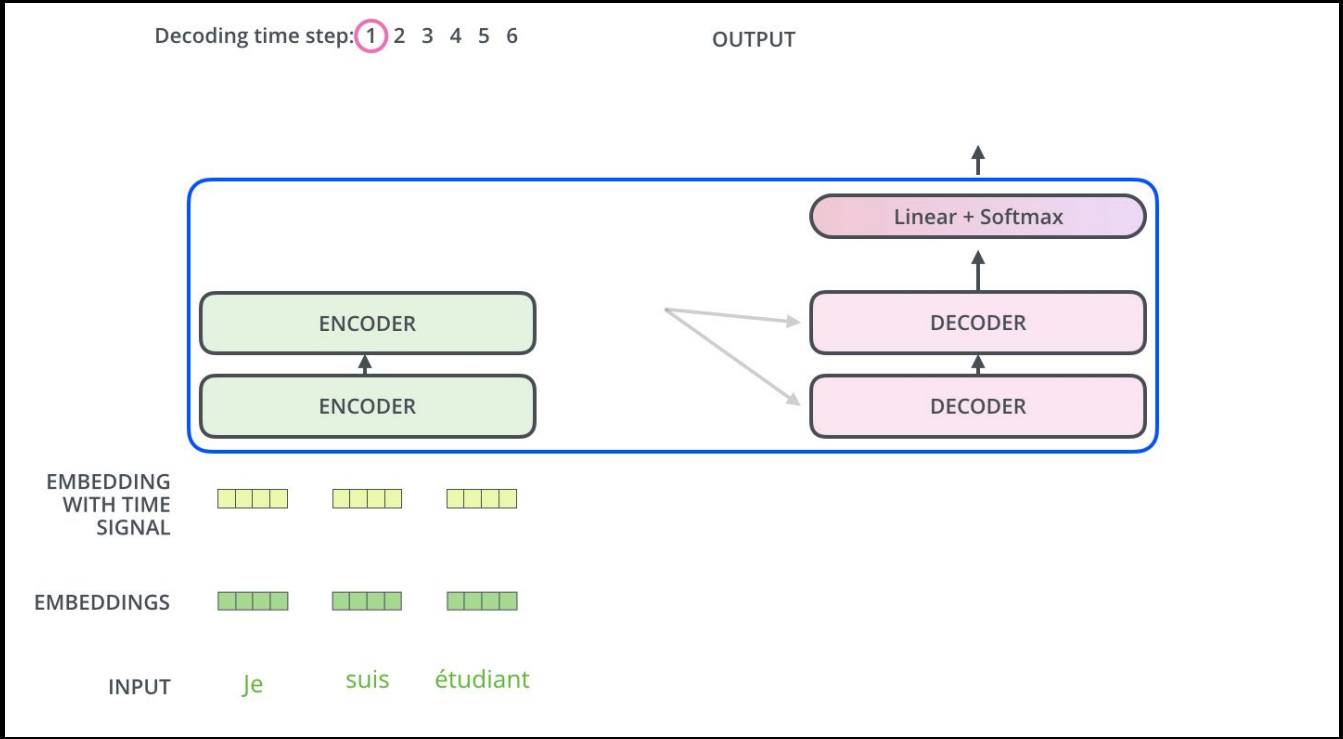
Seq2Seq with Attention



Source: https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

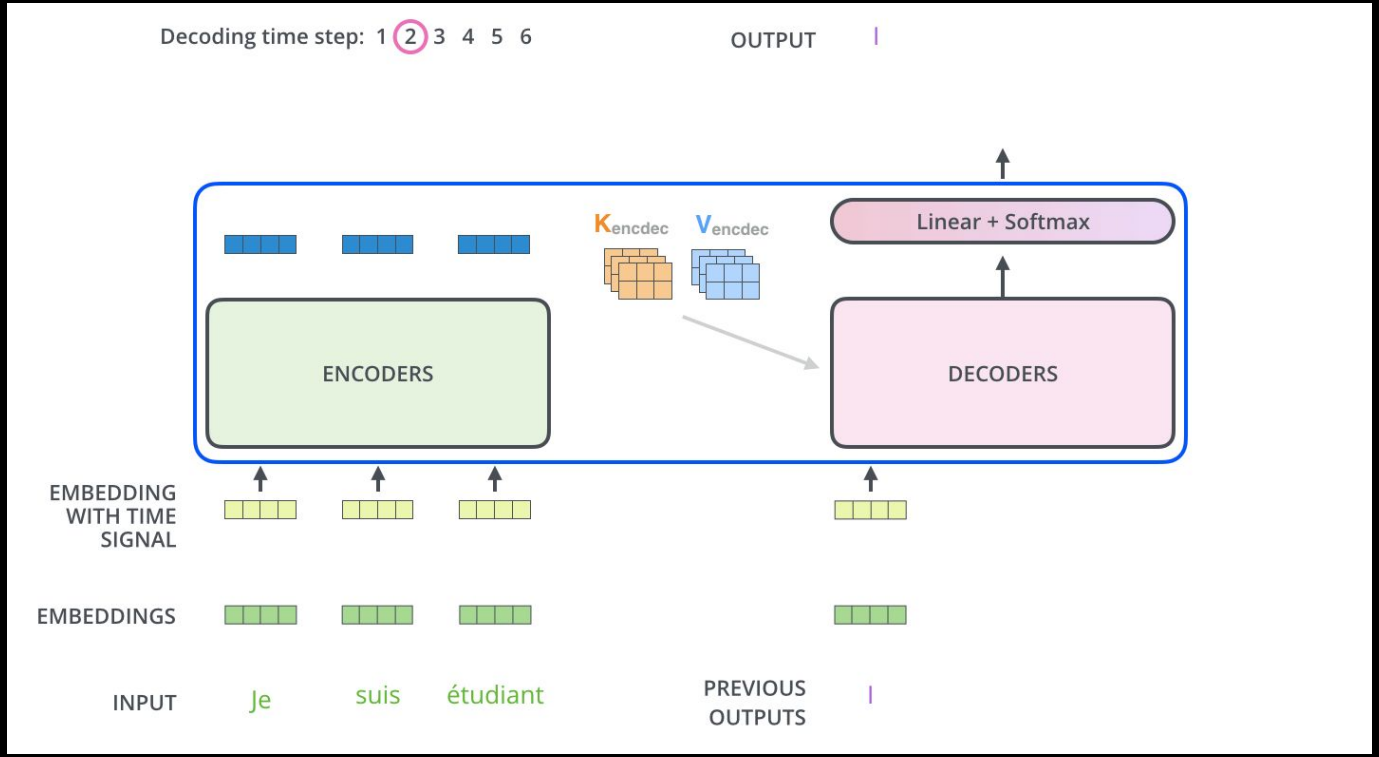
Attention in Neural Networks

Dynamic Attention

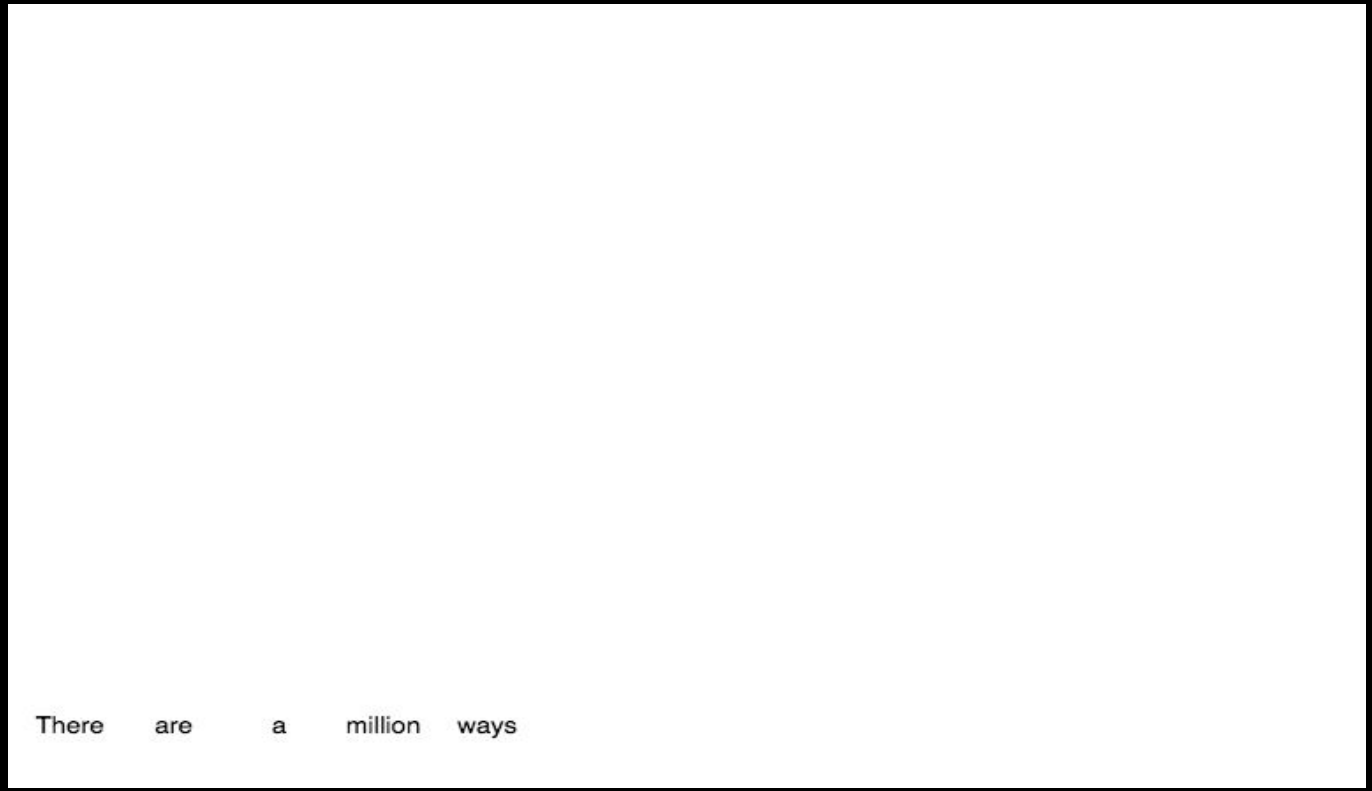


Source: http://jalammar.github.io/illustrated-transformer/

Transformers Language Models



Autoregressive Language Models

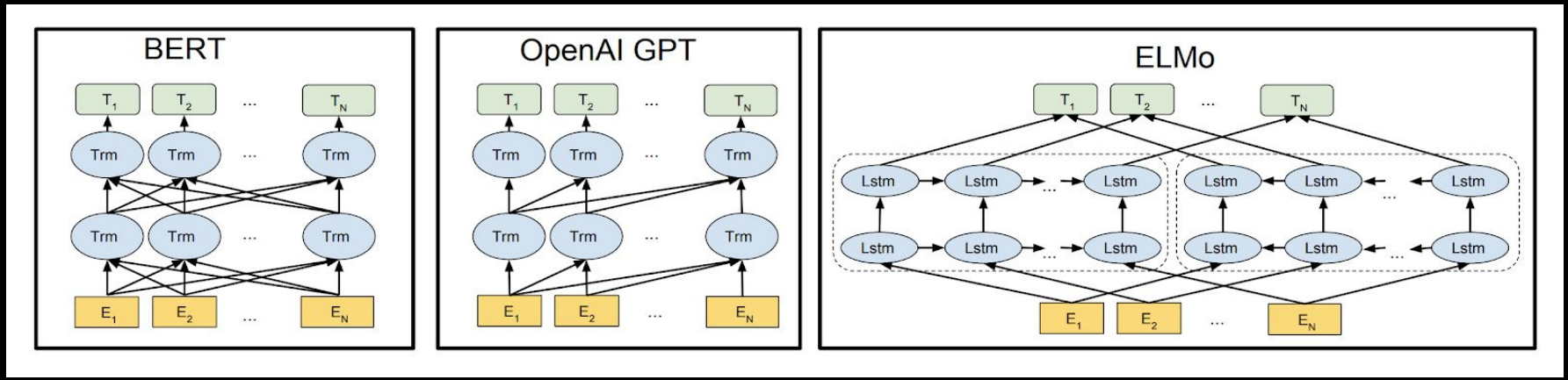


Source: https://eigenfoo.xyz/deep-autoregressive-models/

Contextual Word and Sentence Embeddings: For better NLU/NLP

Transformers

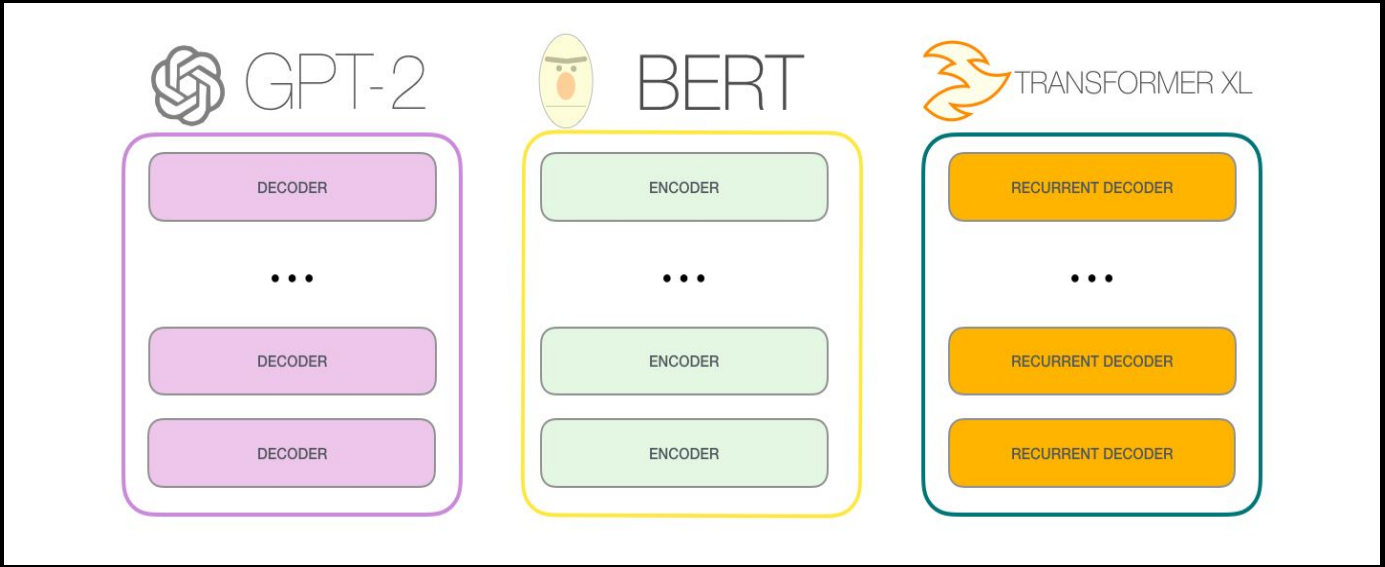
LSTMs/RNNs



Transformer, a Neural Architecture, which learns contextual word and sentence embeddings. Trained with 100s of millions, billions of parameters

Source: https://medium.com/@gauravghati/comparison-between-bert-gpt-2-and-elmo-9ad140cd1cda

BERT (RoBERTa) vs GPT (1, 2, 3) vs Transformer XL



Encoders

- Usually bi-directional
- Great for getting good input representations
- Classification, Tagging, Extractive Summarization

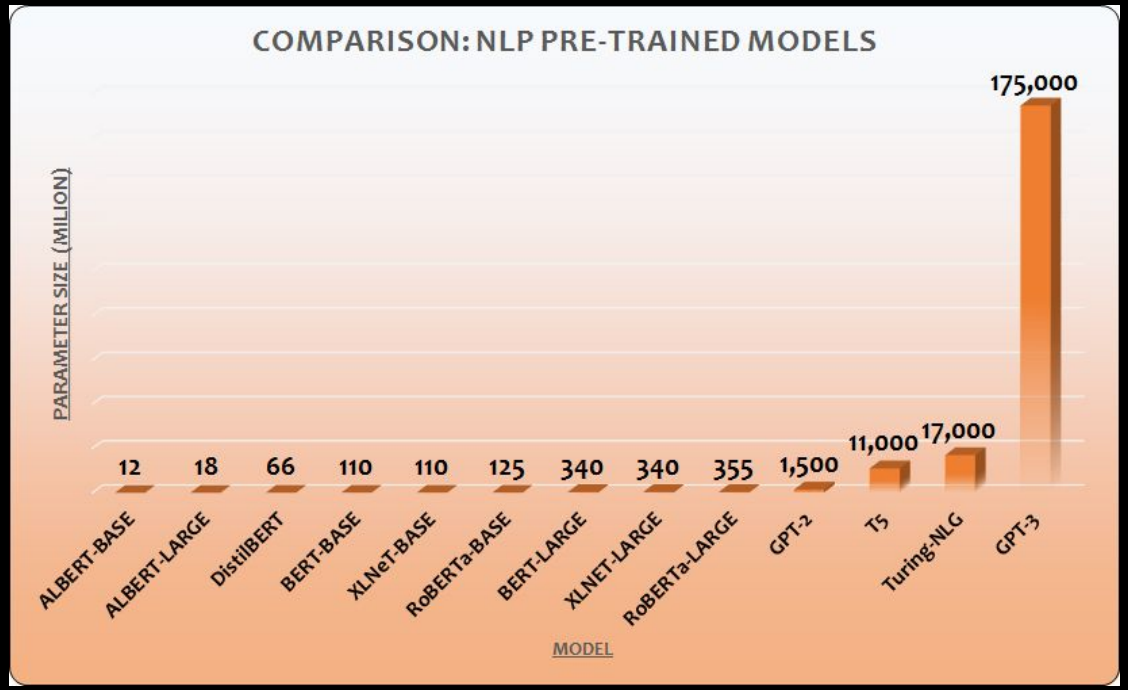
Decoders:

- Uni-directional (forward)
- Great for generative tasks
- NLG, Abstractive Summarization, Translation, QA

Source: http://jalammar.github.io/illustrated-gpt2/

Lots of Transformers!!

Which one to use?

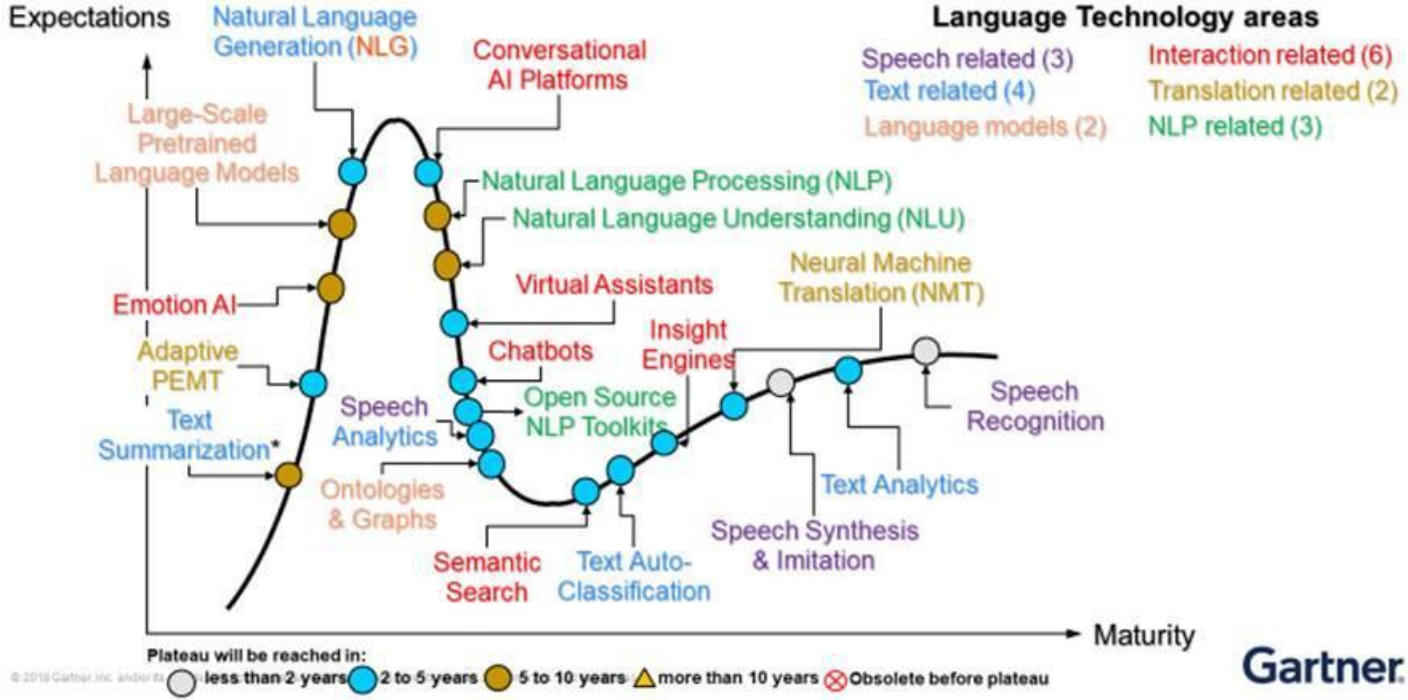


T5XXL: 1,700,000

Source: https://medium.com/analytics-vidhya/openai-gpt-3-language-models-are-few-shot-learners-82531b3d3122

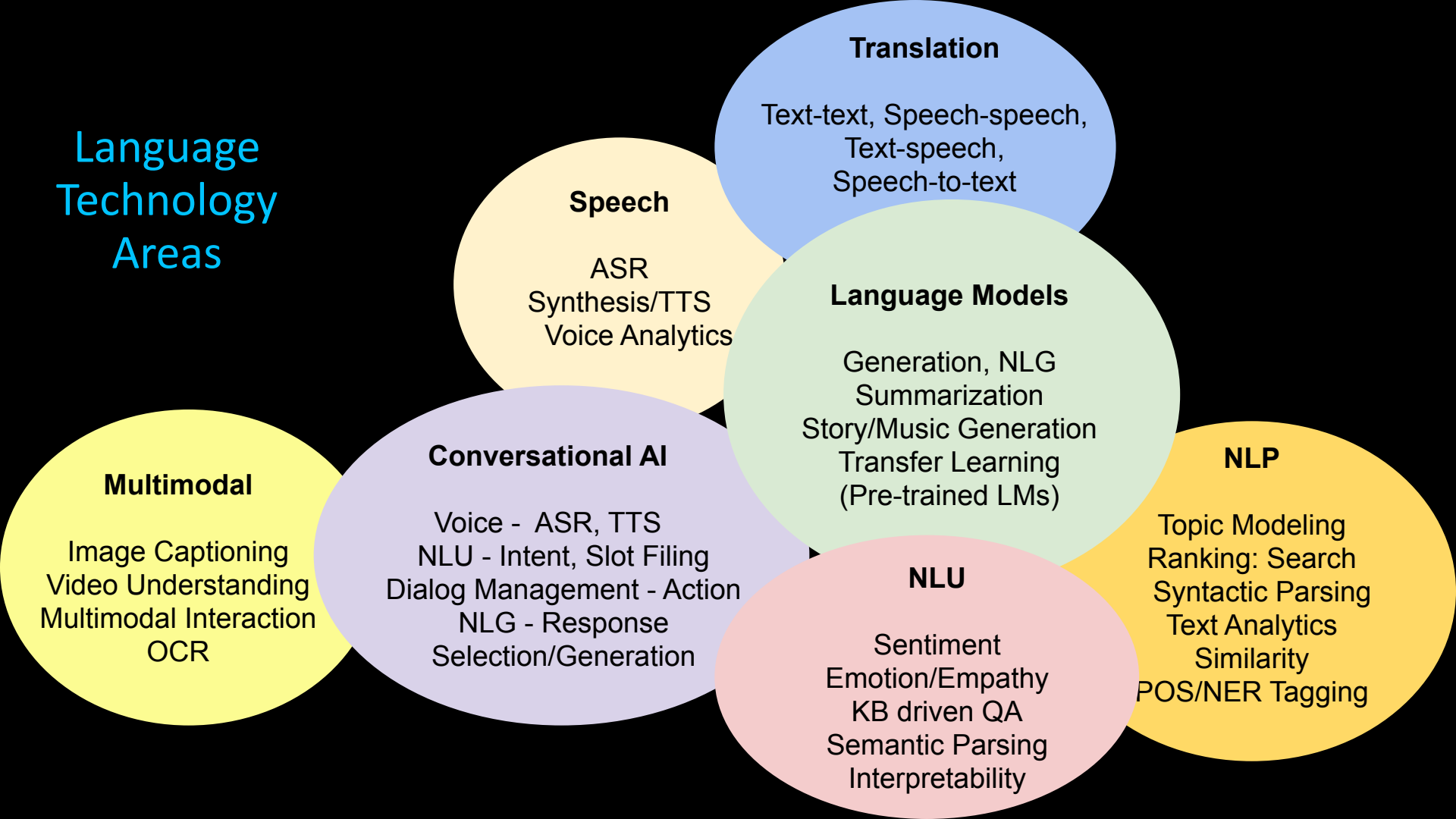
Applications

Hype Cycle for Natural Language Technologies, 2020



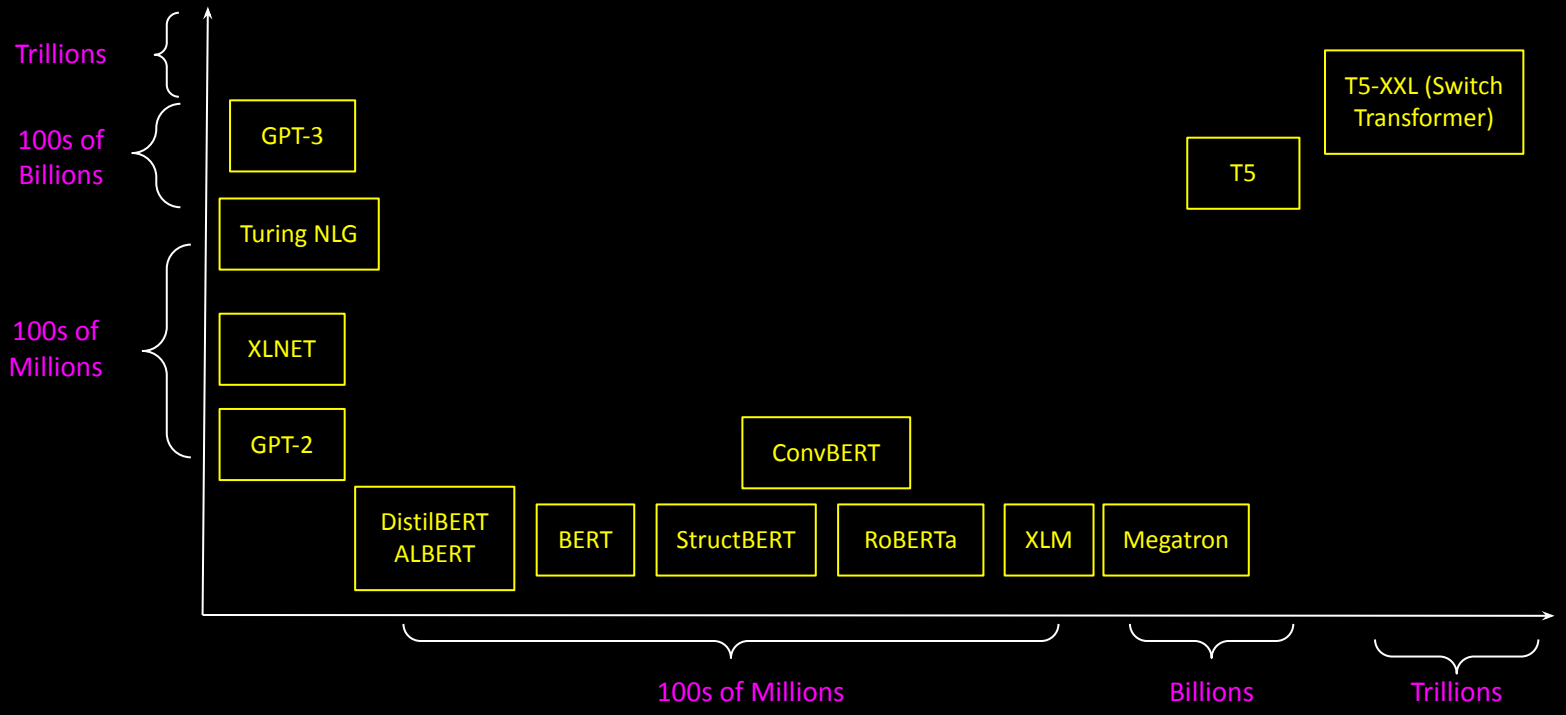
© 2018 Gartner, Inc. and/or its affiliates.

Gartner


Language Technology Areas
Translation
Text-text, Speech-speech, Text-speech, Speech-to-text
Speech
ASR
Synthesis/TTS
Voice Analytics
Language Models
Generation, NLG
Summarization
Story/Music Generation
Transfer Learning
(Pre-trained LMs)
Multimodal
Image Captioning
Video Understanding
Multimodal Interaction
OCR
Conversational AI
Voice - ASR, TTS
NLU - Intent, Slot Filing
Dialog Management - Action
NLG - Response Selection/Generation
NLP
Topic Modeling
Ranking: Search
Syntactic Parsing
Text Analytics
Similarity
POS/NER Tagging
NLU
Sentiment
Emotion/Empathy
KB driven QA
Semantic Parsing
Interpretability

Popular Transformer Architectures & their Applications

Decoders: **Better for Generation Applications**



Encoders: **Better for Classification, Ranking Applications**