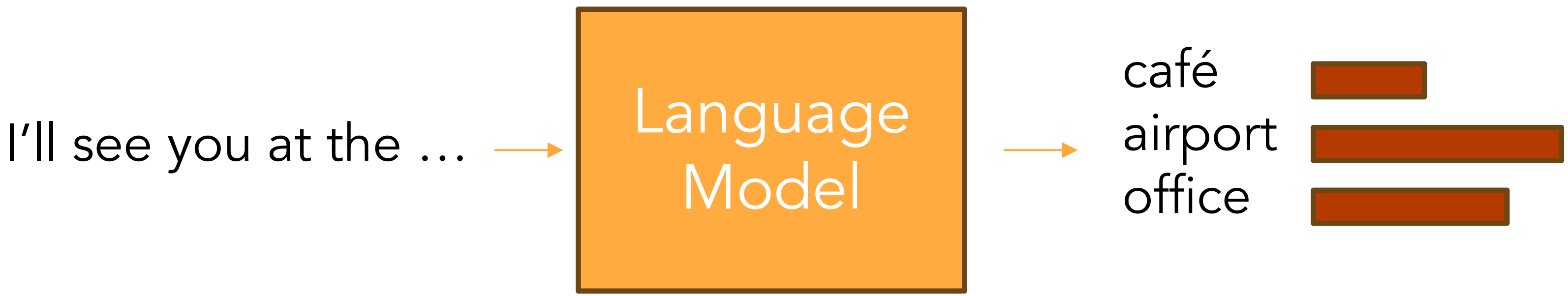
LLM Architectures

LLM Reading Group – 7th July 2023

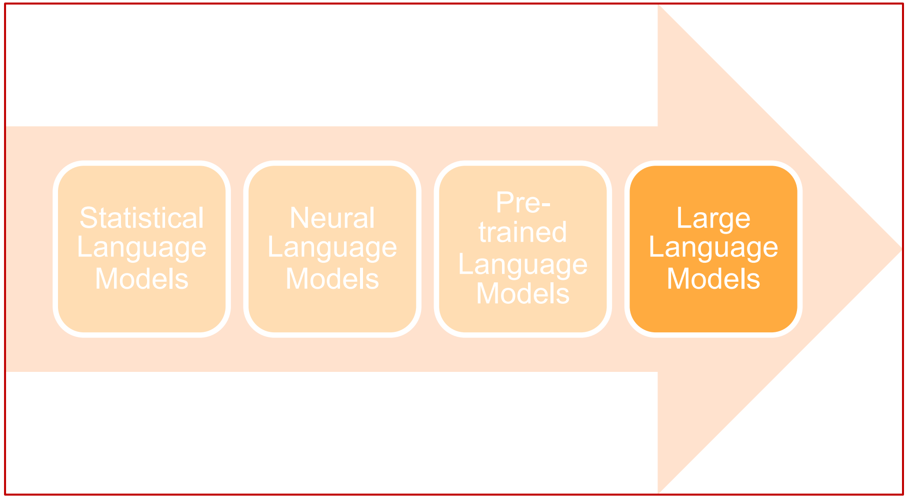
Objectives

- Discuss key components of modern LLM architectures
- Describe encoder only, decoder only, and encoder-decoder architecture using BERT, GPT, and T5 models
- Analyse factors contributing popularity of decoder only architecture

Language Modeling and Language Models



Major development stages in language modelling



The popularity of existing LLMs are attributed to transformers architectures and pre-training strategies.

[1] https://arxiv.org/pdf/2303.18223.pdf

LLM Success Factor – 1
Pretraining / Fine-tuning Paradigm

Pretraining allows model to learn general language representations which then can be fine tuned

<table>
<tr>
<td>

Step-1 Pretrain

Lots of data, learn general aspects of language

- I sit on __
- I [M] on chair.
- I [M] chair.
- I sit on chair. It is a bird.

</td>
<td>

Step-2 Fine-tune

Adapt pretrained model to task

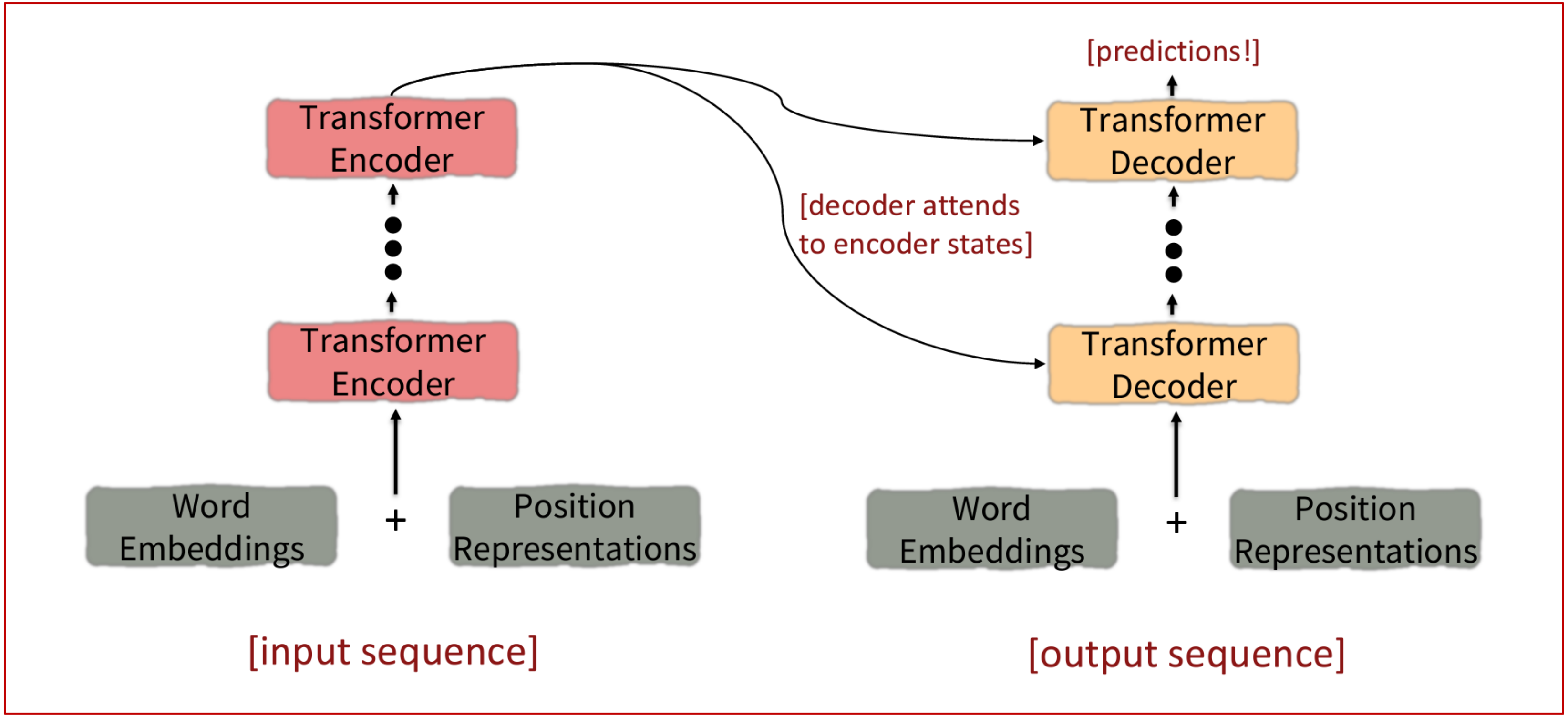
- Full finetuning vs parameter efficient finetuning

</td>
</tr>
</table>

[2] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/slides/cs224n-2021-lecture10-pretraining.pdf

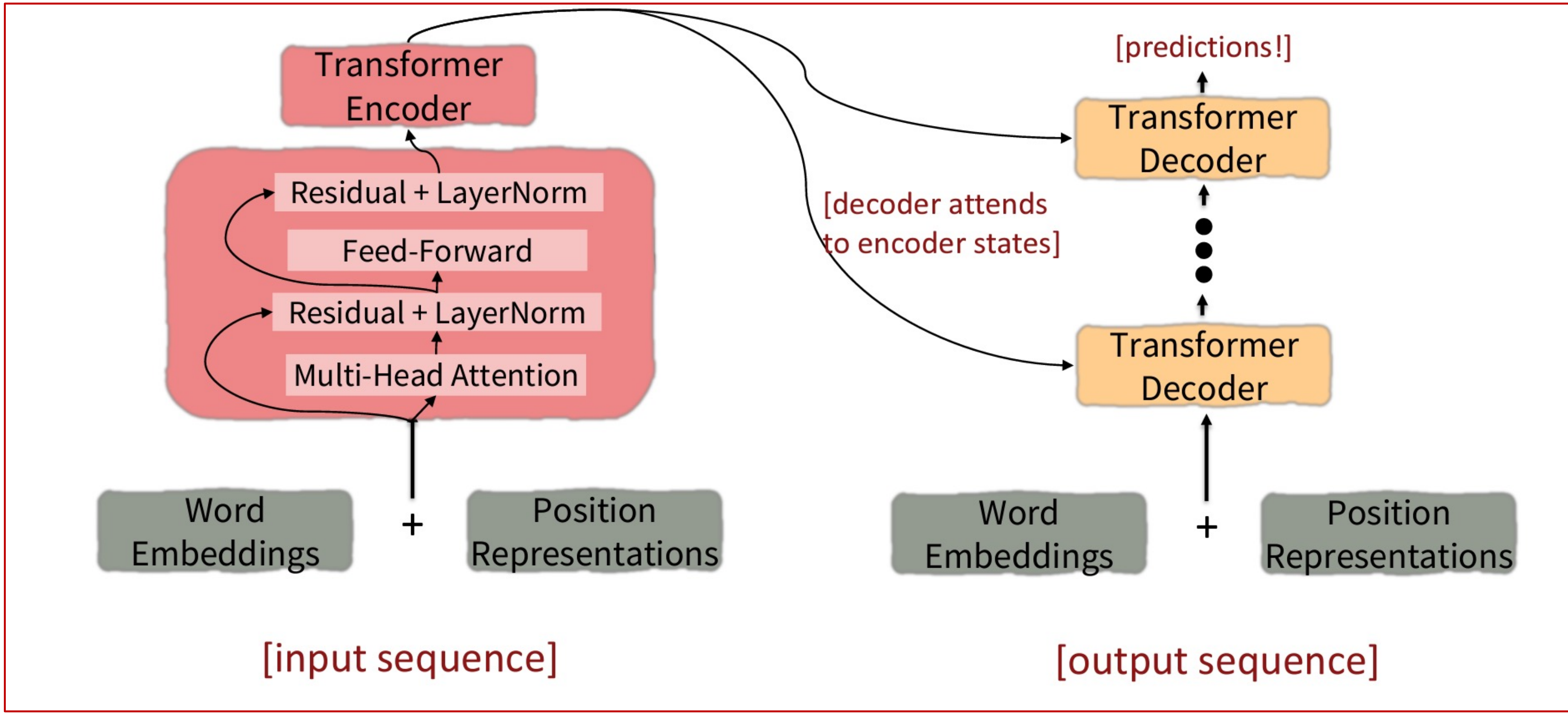
LLM Success Factor – 2
Transformers Architecture

Transformer architecture consists of an encoder and a decoder



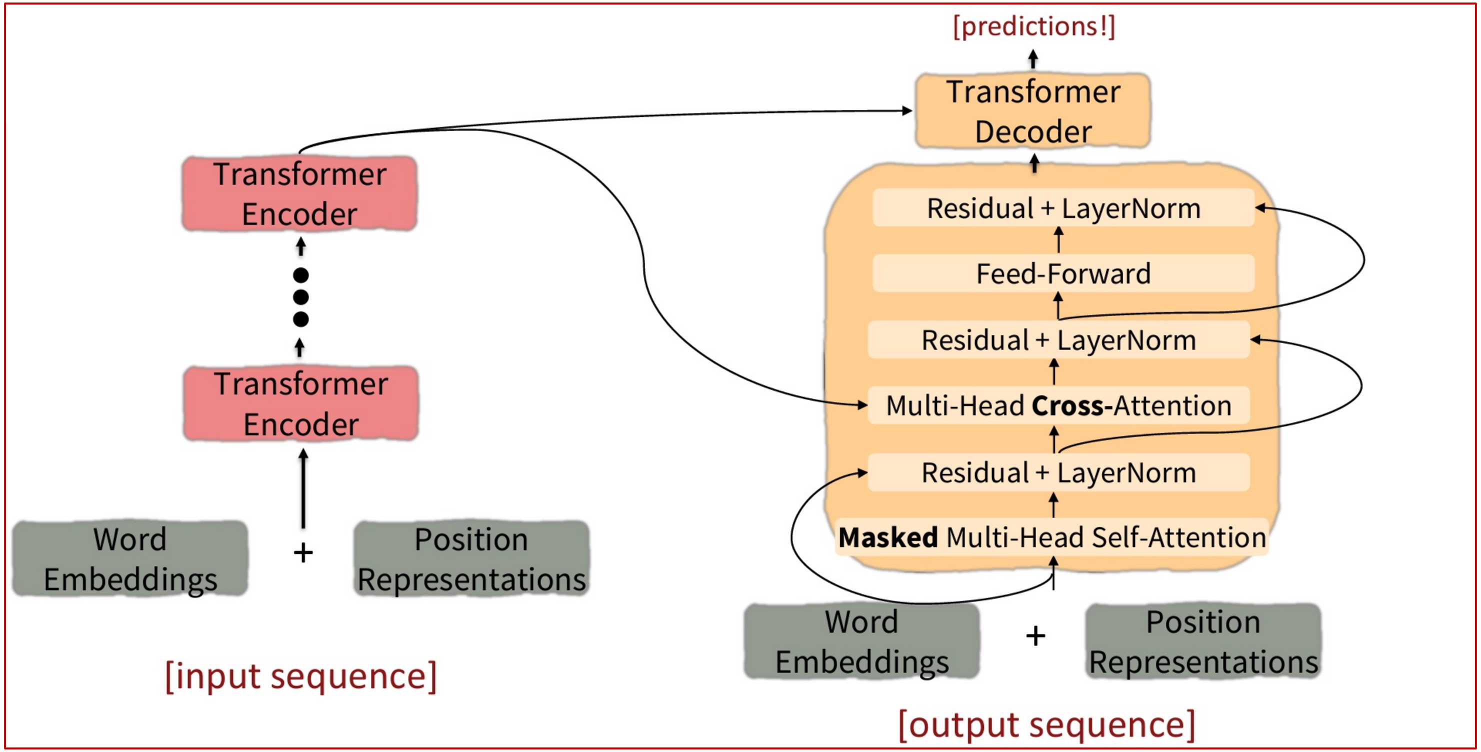
[3] https://arxiv.org/pdf/1706.03762.pdf
[4] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/slides/cs224n-2021-lecture09-transformers.pdf

Transformer encoder block



[3] https://arxiv.org/pdf/1706.03762.pdf
[4] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/slides/cs224n-2021-lecture09-transformers.pdf

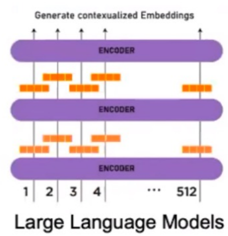
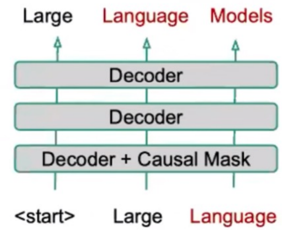
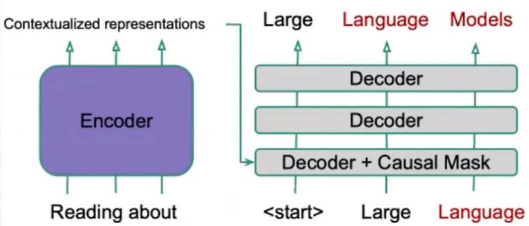
Transformer decoder block



[3] https://arxiv.org/pdf/1706.03762.pdf
[4] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/slides/cs224n-2021-lecture09-transformers.pdf

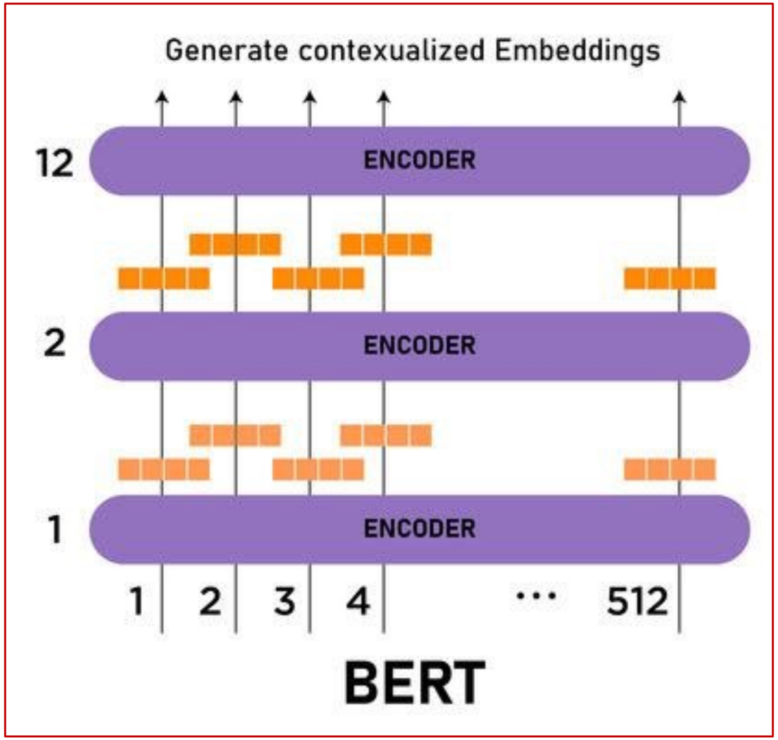
Popular LLM Architectures based on Transformers and Pretraining

Comparing popular LLM architectures

	Encoder	Decoder	Encoder - Decoder
Architecture			
Models	BERT, RoBERTA	GPT-X, Chinchilla, Gopher, PaLM	Flan-T5, UL2, BART
Pre-training	Masked Language Modelling, Next sentence prediction	Next word prediction	Masked Language Modelling, Next word prediction, x-Denosing
	- Generally Discriminative Models - Commonly used for NLU tasks such as text classification, NER, information extraction - More Data efficient	- Generative - Favored for NLG tasks such as Summarization, Machine Translation, Code generation	- Generative - Perform reasonably on both NLU & NLG tasks but not great at either

Encoder-only Architecture

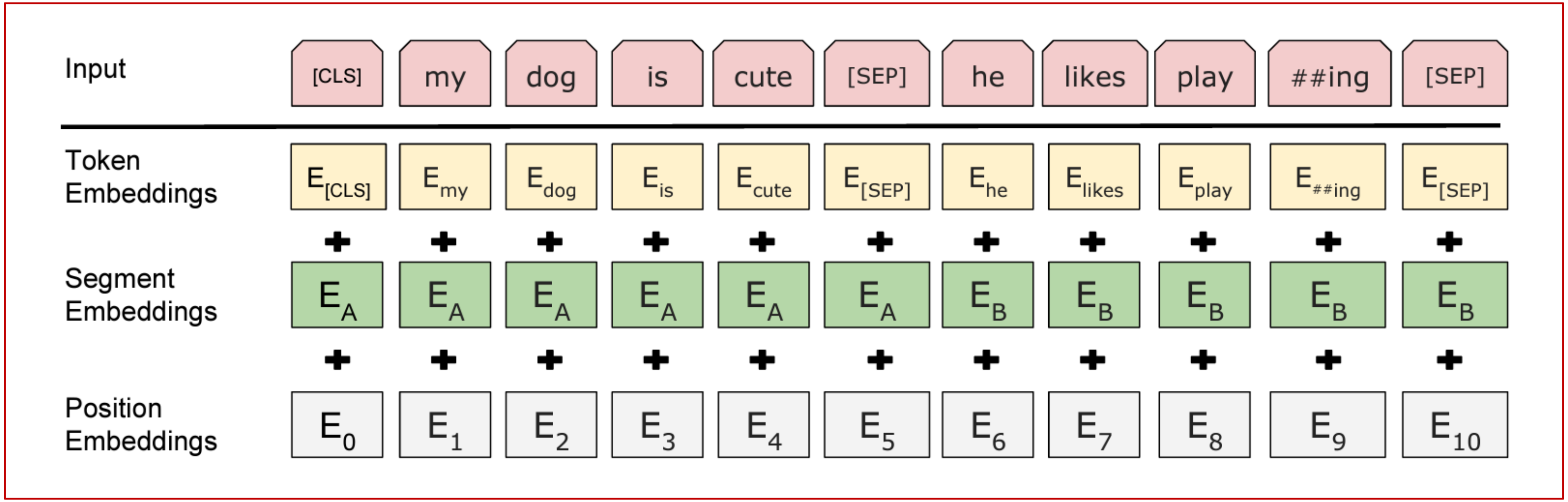
BERT - Bidirectional Encoder Representations from Transformers (2018)



- BERT only uses **Transformers Encoder and not the decoder part**
- It generates contextualized embeddings from **bidirectional context** for input text which can then be used as features for downstream tasks.

[5] https://www.geeksforgeeks.org/explanation-of-bert-model-nlp/

BERT – Input representations



[6] https://arxiv.org/pdf/1810.04805.pdf

BERT – Pretraining strategies

1. Masked Language Modeling

Original : "`The cat sat on the mat.`"

Masked : "`The [MASK] sat on the mat.`"

- Forces BERT to learn contextualized usage of word

2. Next Sentence Prediction

Input : `[CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]`

Label : `IsNext`

- Later work has argued this "next sentence prediction" is not necessary

[6] https://arxiv.org/pdf/1810.04805.pdf

BERT - State of the art results on a broad range of tasks

- Finetuning BERT led to new state-of-the-art results on a variety of tasks – Indicating BERT is useful for natural language understanding tasks

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

[6] https://arxiv.org/pdf/1810.04805.pdf

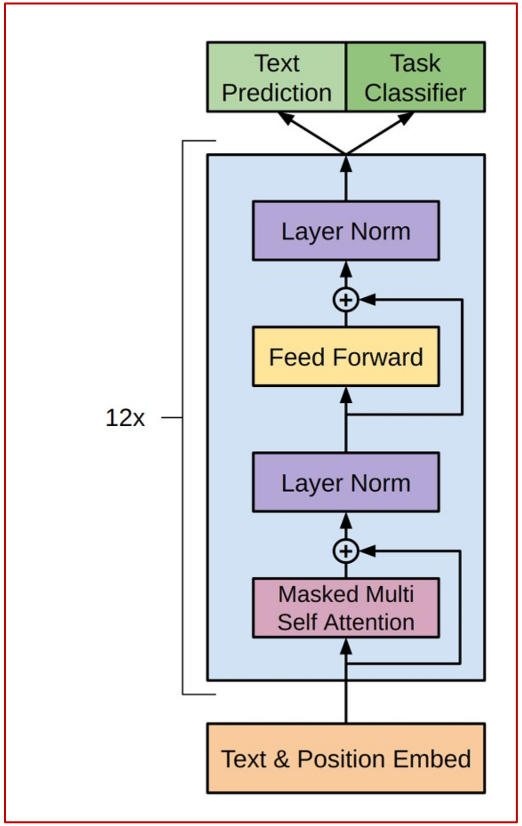
BERT - Limitations

- Lack of generative capabilities
- Lack of Sequence-to-Sequence Modeling
- Lack of Autoregressive Training

[2] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/slides/cs224n-2021-lecture10-pretraining.pdf

Decoder-only Architecture

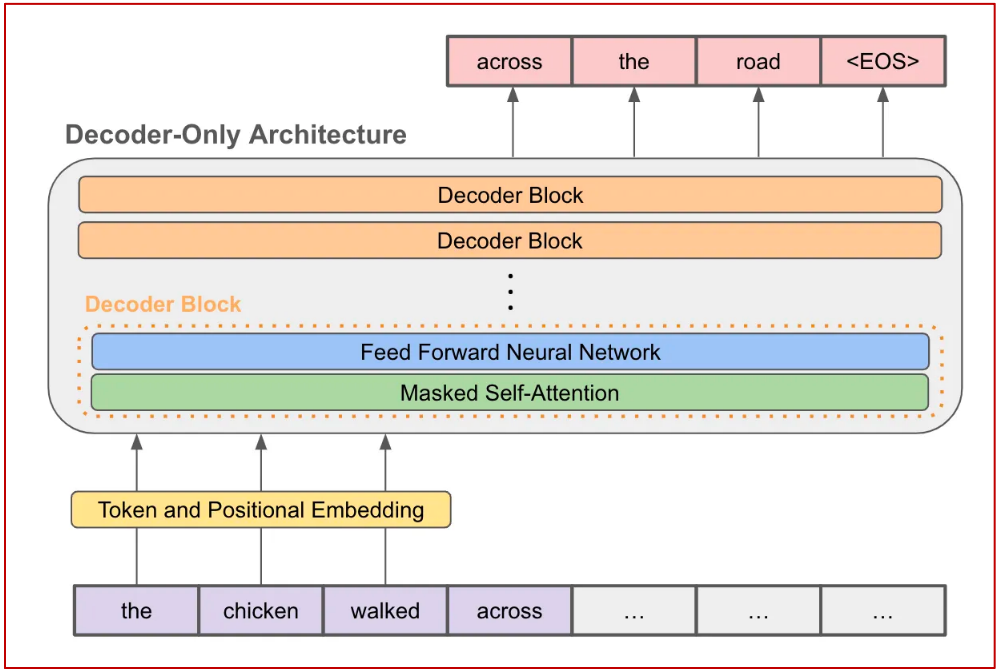
GPT – Generative Pretrained Transformer (2018)



- It only has transformer **decoder part and not the encoder**
- It predicts the next word in a sequence given the previous words, allowing it to generate text that is coherent and contextually appropriate.

[7] https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

GPT – Input / Output representations



[8] https://towardsdatascience.com/language-models-gpt-and-gpt-2-8bdb9867c50a

GPT – Pretraining strategy

- Autoregressive Language Modeling Objective as opposed to Masked Language Modeling in BERT
- This makes them inherent language models!

[2] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/slides/cs224n-2021-lecture10-pretraining.pdf

GPT and GPT 2 – Great results!

- GPT results on NLI tasks

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

- GPT - 2 generated convincing sample of text

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

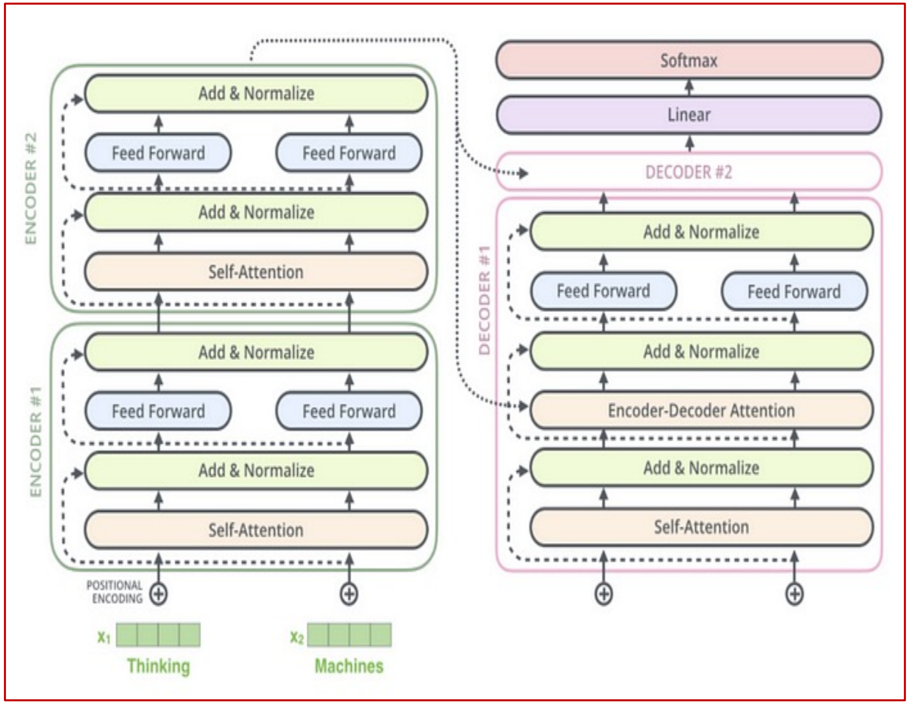
[2] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/slides/cs224n-2021-lecture10-pretraining.pdf

GPT - Limitations

- Lack of Bidirectional Context
- Limited Pre-training Objectives
- Not primarily useful for “Analysis” tasks

Encoder-Decoder Architecture

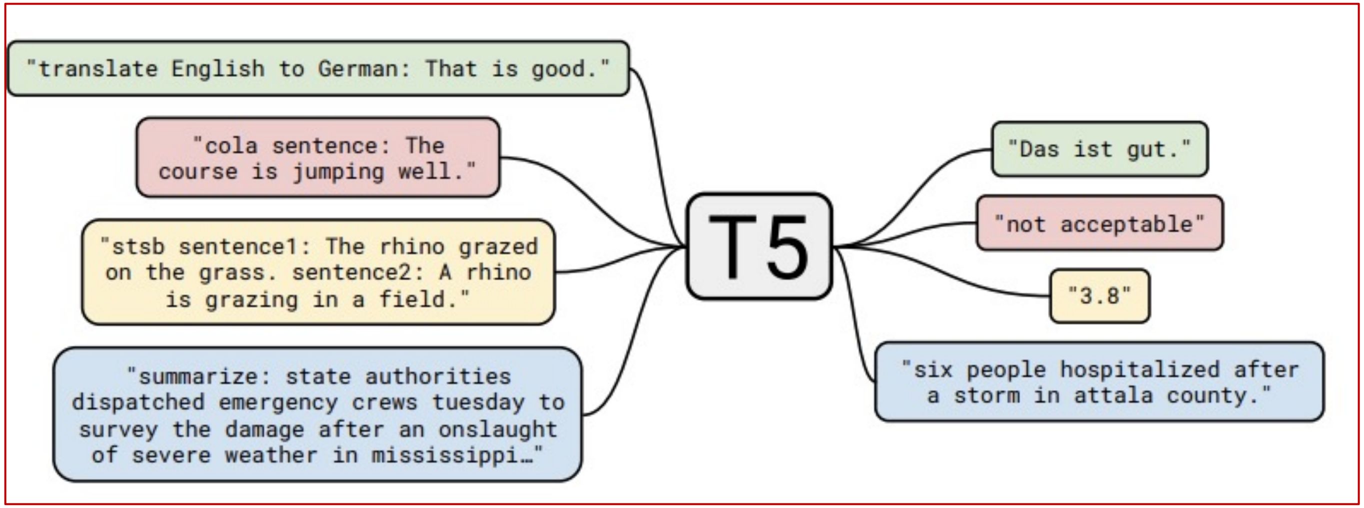
T5 – Text-to-Text Transfer Transformer (2019)



- Utilizes both encoder and decoder parts of transformer
- It has a task-agnostic architecture; meaning same model can be trained on variety of tasks.

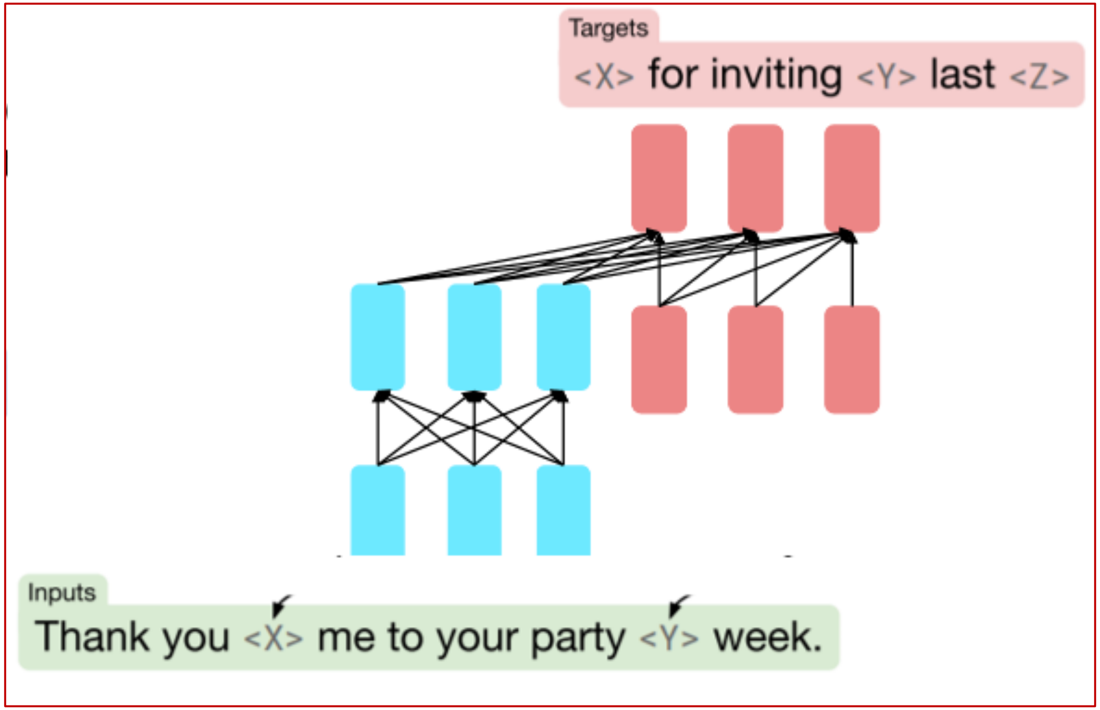
[9] https://medium.com/analytics-vidhya/t5-a-detailed-explanation-a0ac9bc53e51

T5 – Text-to-Text Framework



[10] https://arxiv.org/pdf/1910.10683.pdf

T5 – Authors found span corruption is better than language modelling objective in pretraining



[2] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/slides/cs224n-2021-lecture10-pretraining.pdf

T5 – SOTA comparisons

Model	GLUE Average	CoLA Matthew's	SST-2 Accuracy	MRPC F1	MRPC Accuracy	STS-B Pearson	STS-B Spearman
Previous best	89.4[a]	69.2[b]	97.1[a]	**93.6**[b]	**91.5**[b]	92.7[b]	92.3[b]
T5-Small	77.4	41.0	91.8	89.7	86.6	85.6	85.0
T5-Base	82.7	51.1	95.2	90.7	87.5	89.4	88.6
T5-Large	86.4	61.2	96.3	92.4	89.9	89.9	89.2
T5-3B	88.5	67.1	97.4	92.5	90.0	90.6	89.8
T5-11B	**90.3**	**71.6**	**97.5**	92.8	90.4	**93.1**	**92.8**

Model	QQP F1	QQP Accuracy	MNLI-m Accuracy	MNLI-mm Accuracy	QNLI Accuracy	RTE Accuracy	WNLI Accuracy
Previous best	74.8[c]	**90.7**[b]	91.3[a]	91.0[a]	**99.2**[a]	89.2[a]	91.8[a]
T5-Small	70.0	88.0	82.4	82.3	90.3	69.9	69.2
T5-Base	72.6	89.4	87.1	86.2	93.7	80.1	78.8
T5-Large	73.9	89.9	89.9	89.6	94.8	87.2	85.6
T5-3B	74.4	89.7	91.4	91.2	96.3	91.1	89.7
T5-11B	**75.1**	90.6	**92.2**	**91.9**	96.9	**92.8**	**94.5**

Model	SQuAD EM	SQuAD F1	SuperGLUE Average	BoolQ Accuracy	CB F1	CB Accuracy	COPA Accuracy
Previous best	90.1[a]	95.5[a]	84.6[d]	87.1[d]	90.5[d]	95.2[d]	90.6[d]
T5-Small	79.10	87.24	63.3	76.4	56.9	81.6	46.0
T5-Base	85.44	92.08	76.2	81.4	86.2	94.0	71.2
T5-Large	86.66	93.79	82.3	85.4	91.6	94.8	83.4
T5-3B	88.53	94.95	86.4	89.9	90.3	94.4	92.0
T5-11B	**91.26**	**96.22**	**88.9**	**91.2**	**93.9**	**96.8**	**94.8**

Model	MultiRC F1a	MultiRC EM	ReCoRD F1	ReCoRD Accuracy	RTE Accuracy	WiC Accuracy	WSC Accuracy
Previous best	84.4[d]	52.5[d]	90.6[d]	90.0[d]	88.2[d]	69.9[d]	89.0[d]
T5-Small	69.3	26.3	56.3	55.4	73.3	66.9	70.5
T5-Base	79.7	43.1	75.0	74.2	81.5	68.3	80.8
T5-Large	83.3	50.7	86.8	85.9	87.8	69.3	86.3
T5-3B	86.8	58.3	91.2	90.4	90.7	72.1	90.4
T5-11B	**88.1**	**63.3**	**94.1**	**93.4**	**92.5**	**76.9**	**93.8**

Model	WMT EnDe BLEU	WMT EnFr BLEU	WMT EnRo BLEU	CNN/DM ROUGE-1	CNN/DM ROUGE-2	CNN/DM ROUGE-L
Previous best	**33.8**[e]	**43.8**[e]	**38.5**[f]	43.47[g]	20.30[g]	40.63[g]
T5-Small	26.7	36.0	26.8	41.12	19.56	38.35
T5-Base	30.9	41.2	28.0	42.05	20.34	39.40
T5-Large	32.0	41.5	28.1	42.50	20.68	39.75
T5-3B	31.8	42.6	28.2	42.72	21.02	39.94
T5-11B	32.1	43.4	28.1	**43.52**	**21.55**	**40.69**

- Overall, state-of-the-art performance is achieved on 18 out of the 24 tasks.
- **This makes T5 useful for variety of downstream tasks including machine translation, summarization, question answering, and more.**

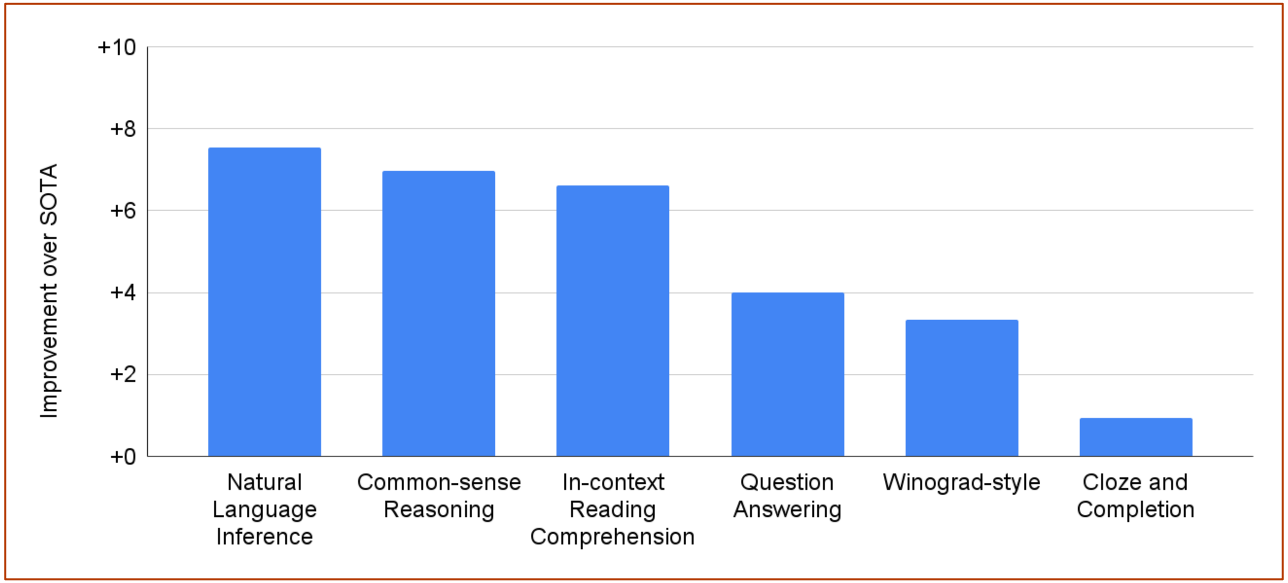
[10] https://arxiv.org/pdf/1910.10683.pdf

T5 - Limitations

- Not suitable if your focus is on specific task instead of broad range of tasks. Fine-tuning a BERT or GPT model on a specific task might still provide superior results due to the task-specific architectures and pretraining objectives.
- For smaller-scale tasks, or smaller labeled datasets, using BERT or GPT models can be more feasible and practical over T5 due to computations resources requirement

Why are decoder only models dominating?

- Breakthrough Capabilities on Language, Reasoning, and Code Tasks (Google PaLM)



[11] https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

Why are decoder only models dominating?

- They are are excelling at the language modelling and text generation
  - Try ChatGPT yourself!

Creative Writing

Write a short story about a mysterious island in 100 words.

Deep within the tempestuous sea, an isle concealed its secrets. Legend whispered of an ancient curse, casting an ethereal aura upon the land. Storms raged perpetually, keeping intruders at bay. Yet, a courageous sailor, driven by insatiable curiosity, braved the treacherous tides. As he set foot on the shore, an ominous mist surrounded him. The island's lush greenery beckoned, hiding enigmatic ruins and untold treasures. Whispers of forgotten civilizations echoed through the wind, guiding him through darkened paths. With every step, the sailor felt the weight of the island's mystery. Would he unravel its enigma or succumb to its eternal allure?

Technical Knowledge

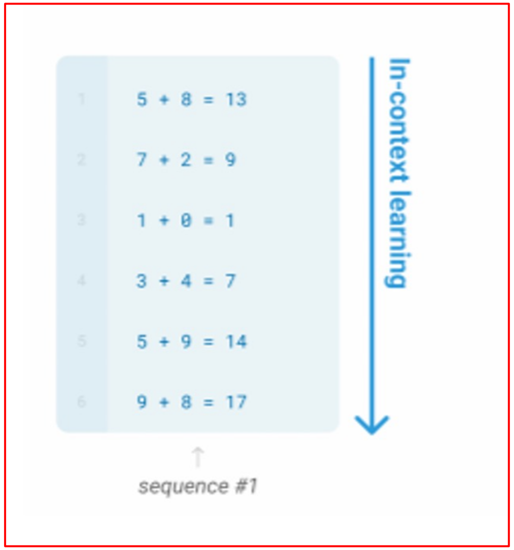
Explain the concept of deep learning in simple terms concisely.

Deep learning is a type of artificial intelligence that imitates the human brain's way of processing information. It involves training neural networks with multiple layers to learn patterns and make predictions from vast amounts of data. By using these deep neural networks, machines can understand complex relationships and perform tasks like recognizing images or understanding speech. It's like teaching a computer to think and learn on its own, allowing it to solve problems and make decisions with remarkable accuracy. Deep learning has revolutionized fields such as computer vision, natural language processing, and robotics, bringing us closer to creating intelligent machines.

[2] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/slides/cs224n-2021-lecture10-pretraining.pdf
[12] https://chat.openai.com/

Why are decoder only models dominating?

- Emergent abilities were observed in GPT-3
    - In-context learning without gradient steps



[2] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/slides/cs224n-2021-lecture10-pretraining.pdf
[13] https://bdtechtalks.com/2022/08/22/llm-emergent-abilities/

Thank you for your attention!