Sequence to Sequence

&

Language Modeling

By: Kanav Bansal
(ThatAIGuy)

Traditional Models vs Language Models

By: Kanav Bansal
(That AI Guy)

What is a model?

Model is a relationship captured b/w inputs & outputs.

Model is used to predict the outcome of an unseen query point.

What is a Language Model?

Model that captures the sequential relationship b/w words & sentences of a language is called as a language model.

Ques: If you can Model a Language, what can you do?

Ans: Build Auto Complete features
Summarize the Text
Build better Chatbots
Build QA systems
etc...

Techniques to Build Language Models ↘

By: Kanav Bansal
(That AI Guy)

Q: How do we build a Language Model?

Huge amounts of data

+

Powerful Architecture

+

Language Modeling Technique

Q: Powerful Architecture?

A: RNN | LSTM, Transformers

Qu: Language Modeling Technique?

A: a) Auto Regressive Task
b) Auto Encoding Task

Auto Regressive vs Auto Encoding Technique ↓

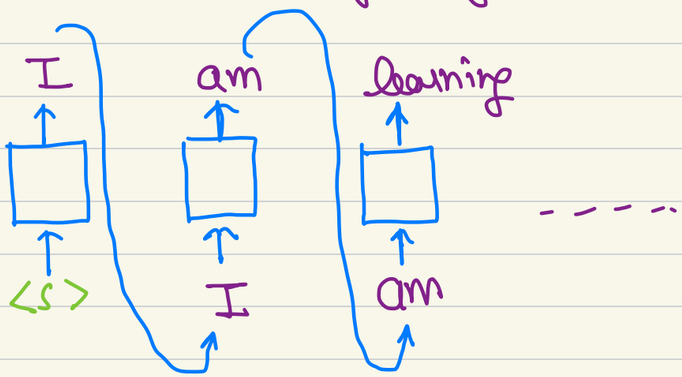
By: Kanav Bansal
(That AI Guy)

Auto Regressive Language Model ↓

They are trained to predict the next token in a sequence, based on the previous tokens.

* A Mask is applied to full sentence.
* Unidirectional

<s> I am learning Language Modeling. </s>

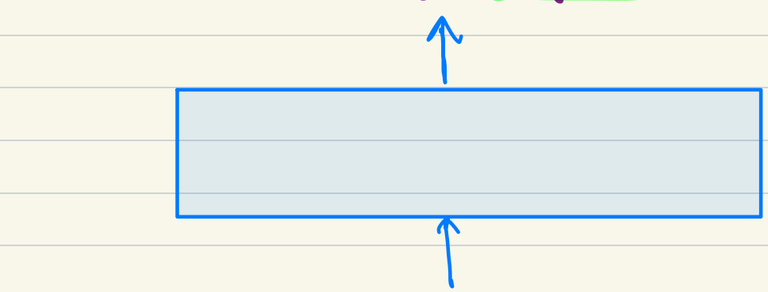


Auto Encoding Language Model ↓

They are trained to reconstruct the original sentence from a corrupted version of the input.

* Certain words in a sentence are replaced with a special token (usually [MASK]).
* Bidirectional

<s>I am learning Language Modeling.</s>



<s>I am learning [MASK] Modeling.</s>

Key Differences:

Predicts ↓

Auto-Regression → Predicts one token at a time.

Training ↓

Auto-Regression → Aims at maximizing the likelihood of the next token given the previous tokens

Predicts ↓

Auto-Encoding → Generates the entire target seq.

Training ↓

Auto-Encoding → Minimize Reconstruction error.

By: Kanav Bansal
(That AI Guy)

Let's Now Discuss the Architecture :

Till 2013 :

- Statistical LM
- Classical ML
- CNN
- ANN
- Word2Vec
- Neural LM RNN/LSTM's

By: Kanav Bansal
(That AI Guy)

text Corpus → Shallow ANN → Model

How to train?
- CBOW
- SkipGram
- SkipGram with Negative Sampling

text Corpus → RNN/ LSTM → Model

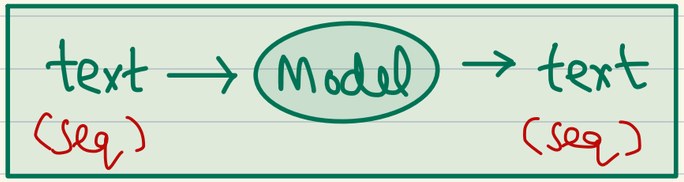
RNN are a special type of a Neural Networks that are designed to cope with sequential data such as time series or text data.
RNN maintains a state variable the evolves over time as the RNN sees more data, giving it the power to model sequential data.
This state variable updates over time by a set of recurrent connections.
The recurrent connections update the current state variable with respect to the past memory the RNN has, enabling the RNN to make a prediction based on the current input as well as the previous input.

text Corpus → Transformers → Model

One of the primary reasons that the transformer model is so performant is that it has access to the whole sequence of tokens, as opposed to RNN-based models, which look at one token at a time.

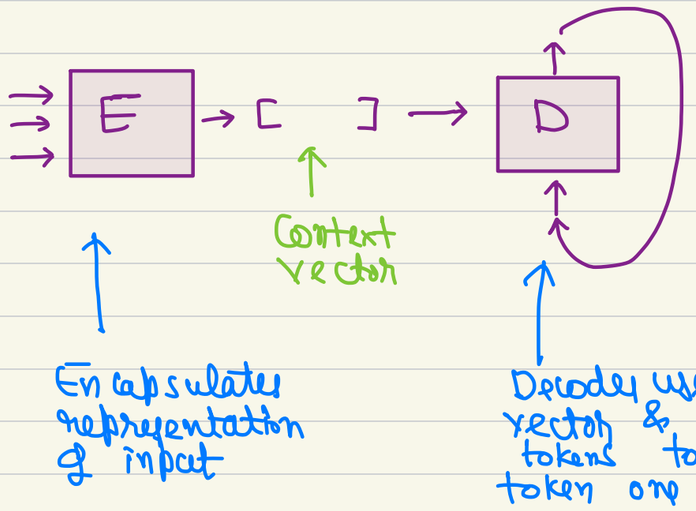
The Inception of Generative Models

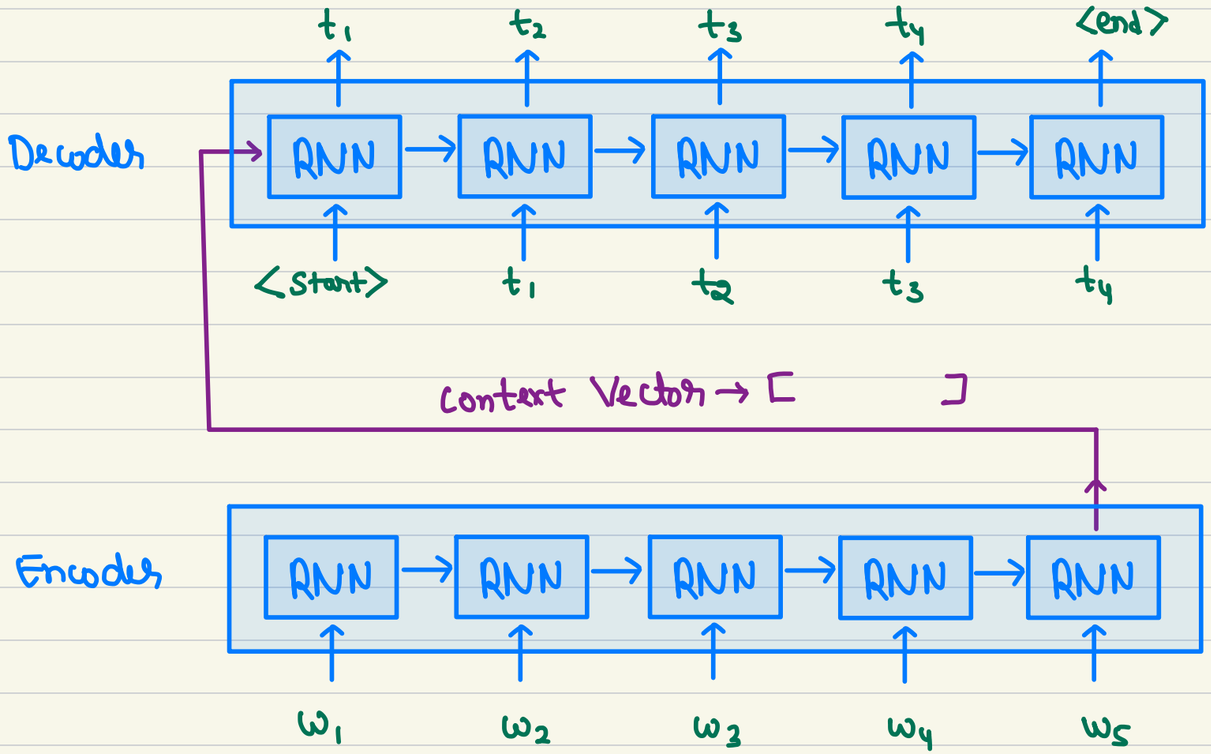
By: Kanav Bansal
(That AI Guy)



PAPER: Seq2Seq Learning with NN

2014 → Proposed Encoder-Decoder Architecture to solve Machine Translation Task (Seq2Seq)





Pro:
- → Handles variable length input-output sequences.
- → This makes it suitable for Translation/Summarization & QA Problems.

Cons:
- → Both encoder-decoder used RNN/LSTM
- → Generate fixed length context vector which leads to info-loss for long seq.

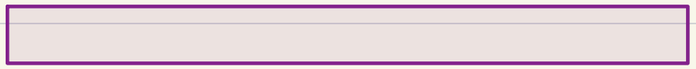
By: Kanav Bansal
(That AI Guy)

Paper: Neutral Machine Translation by Joint learning to Align & Translate

2015 → Proposed Bahdanau Attention to perform better Translation.

Que: Should you rely on the complete sentence for translation?

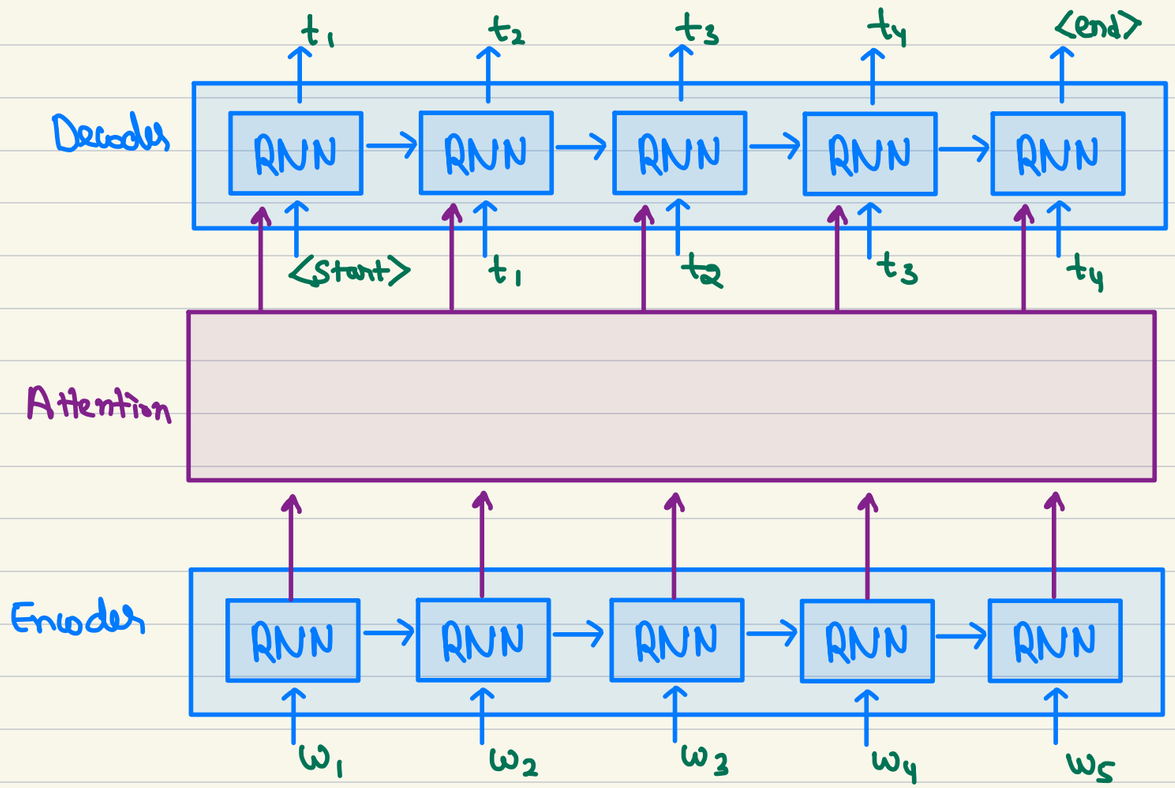
Eg: Hi, how are you?



In 2015 → Introduced the concept of attention.

→ By dynamically adjusting the attention weights, the model can focus on important words, leading to more accurate translation.

→ Attention solves the huge fundamental flow of long term dependency in LSTM.



* But _____ problem remains.

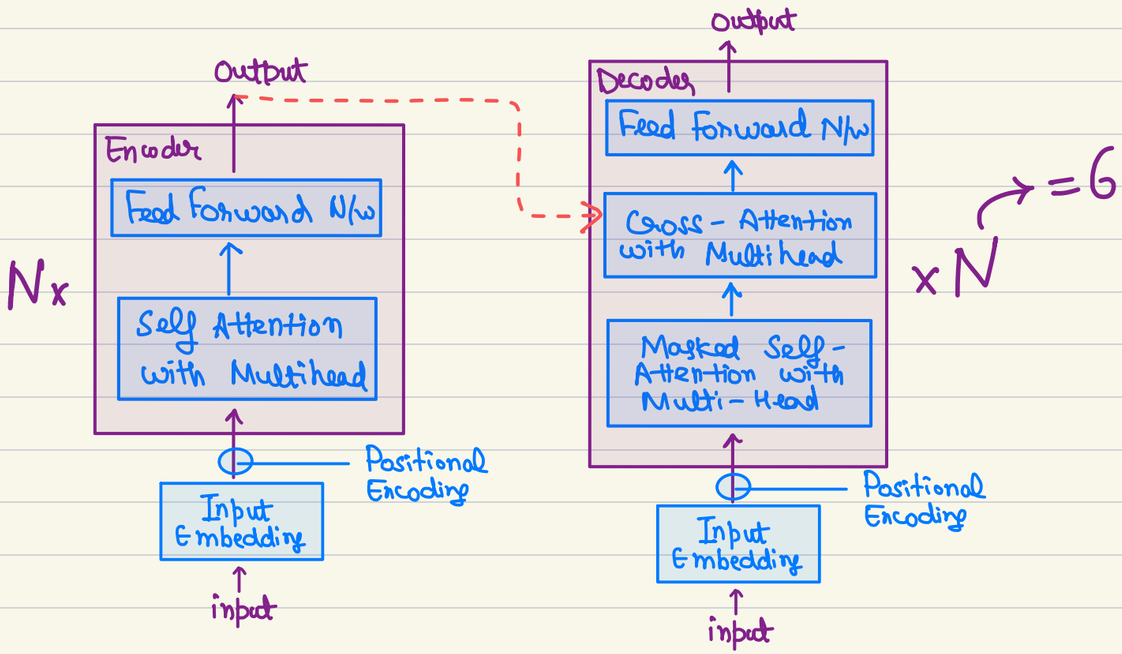
By: Kanav Bansal
(That AI Guy)

Paper: Attention is all you need

2017 → Proposed Transformer Architecture which used:
- → Encoder Decoder Architecture
- → Self-Attention Mechanism

BUILDING BLOCKS

(Note: This is a simplified view of Transformer)



Encoder is good at understanding text

Decoder is good at generating text

ENCODER BUILDING BLOCK

1. Positional Embedding
   - Adds the notion of seq info
2. Self-Attention with Multihead
   - Attention replaced need of recurrence
   - Attention mechanism will find how each word relates to all other words in a sequence.

   Ques: How to find word similarities?
   - Attention will run dot products b/w word vectors & determine the strongest relationships of a word with all other words

   Ques: How to speed up the calculations?
   - For each attention sublayer, the original transformer model runs not only one but eight attention mechanisms in parallel to speed up the calculations. This process is done using 'multi-head attention'.
3. Fully Connected Positionwise Feed Forward N/w
   - Improves the word association by applying non-linear transformations.

By: Kanav Bansal
(That AI Guy)

TRANSFORMER'S

↓

Attention is all you need (2017 Paper)

Why CELEB STATUS ?

1. Scalable & Parallel Training
2. Revolutionized NLP with LLMs
3. Unified DL Approaches for text, image, audio & video data
4. Multi-Modal
5. Accelerated Gen AI

Encoder	Decoder
Has self-Attention with multi-head & Feed forward ANN	Mask self-Attention & Cross-Attention with Multi-head & feed forward ANN
↑ Great at understanding Text	↑ Great at Generating Text
Eg: BERT	Eg: GPT-1

By: Kanav Bansal
(That AI Guy)

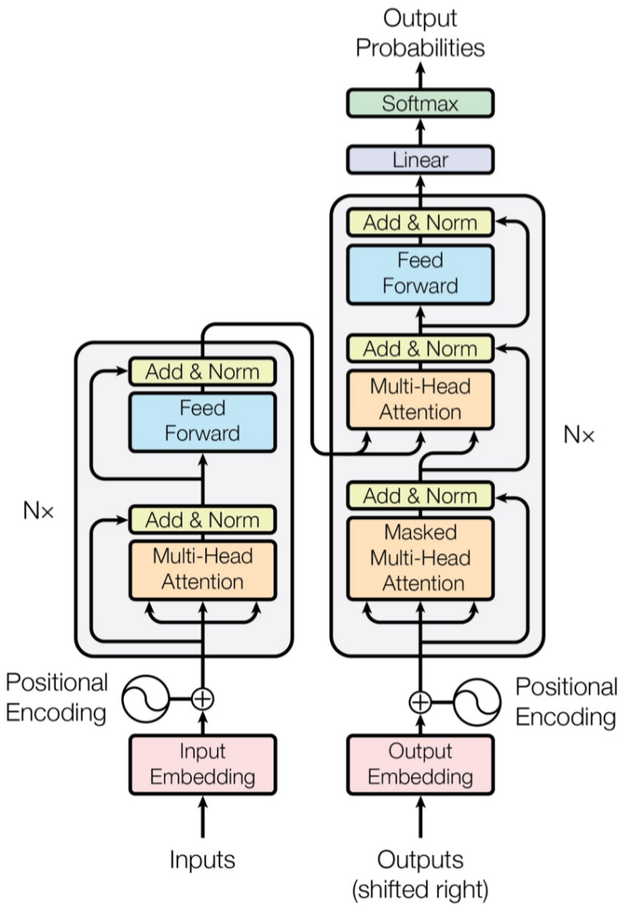


Figure 1: The Transformer - model architecture.

Advantages :

* Parallel & Scalable
* Multi-Modal input & output

Disadvantages :

* Huge compute resources
* Huge amount of data
* Overfit
* Energy

The Concept of Transfer Learning

By: Kanav Bansal
(That AI Guy)

Re-use the knowledge gained from one task inorder to boost the performance on a related task.

For eg: Pre Training Task → Riding a cycle
Down Stream Task → Riding a MotorBike

Ques: Can we use the model trained during machine translation to boost the performance of text summarization task?

Ans: Note that, we can't use machine translation as a pre-trained model for text summarization task as both the tasks are completely different.

Ques: What is the problem with MT as a pre-training task?

Ans:
* Machine translation, Text Summarization, Question Answering Systems, etc... are way too much task specific.
* Lack of data

For eg: Pre Training Task → Riding a cycle
Down Stream Task → Flying a Plane

By: Kanav Bansal
(That AI Guy)

Paper: Universal Language Model Fine-Tuning for Text Classification (ULMFiT)

2018 → Proposed Language Modeling as pre-training task.

* Using Language Modeling as a Pre-Training task, they outperformed the SOTA on six text classification tasks.
* Furthermore, with only 100 labeled examples, they matched the performance of training from scratch on 100x more data.

Ques: What is the benefit of using LM as a pretraining task?

Ans:
* Huge amounts of data is available
* High quality language representation

Summary ↓

By: Kanav Bansal
(That AI Guy)

2014 → Sequence to Sequence Learning with Neural Network
Proposed - Encoder Decoder architecture

2015 → Neural Machine Translation with Joint Learning to align & Translate
Proposed - Bahdanau Attention Mechanism

2017 → Attention is all you need
Proposed - Transformer Architecture with Self-Attention

2018 → Universal Language Model Fine-Tuning for Text Classification
Proposed - Language Modeling as pre-training Task