Chapter 1 Starts from July 13

Wednesday, July 12, 2023 1:33 PM

Zero-shot classification:-

texts not labelled

Specify labels

X FT

return Probability scores of labels

text-generation:-

provide a prompt

auto-complete it.

can control no. of o/p generated & max. length.

Mask filling:-

Text Generation:-

NER:-

QA:-

Summarization:-

History of Transformers:-

June 2017.

GPT: first PT Xmer.

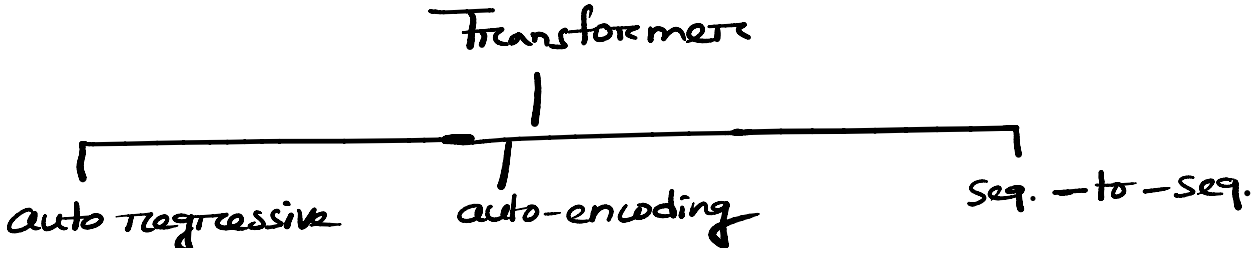
BERT:-

GPT-2:-

DistilBERT:-

BART, T5:-

GPT-3:- Zero-shot learning.



auto regressive | auto-encoding | seq.-to-seq.

Transformers are LMs:-
- trained on large amount of raw texts. } Pre-training
- Self-supervised w/o label data
- Statistical understanding of lang.
- × useful for specific tasks.

Transfer learning
FT in supervised way } Fine-tuning.

Causal LM:-
- i/p:- past, present i/p
- o/p:- predict next word.

Masked LM:-
- i/p এর কিছু token masked করে দেওয়া হয়।

Transformers are big models:-
- কিছু outliers (DistilBERT) ছাড়া ↑ performance পাওয়ার জন্য ↑ model size & ↑ amout of data.
- Costly in time & compute.
- environment impact.
  - footprint of finding best hyperparameters ↑.
  - Scratch থেকে train করার cost ↑
    - Soln:- sharing

Transfer learning:-
- PT training from scratch.
- Large amount of data

Large amount of data
several weeks.

FT after PT
perform + training dataset specific to your task.

Reasons for not training directly for final task:-
Sim. bet^n PT & FT dataset.
take adv. of K acquired during PT
PT data ↑
FT data ↓
amount of time & resource ↓
environmental impact ↓

General Architecture:- of Transformers:-
Two blocks:-
Encoder:- receive i/p
build a rep. (features) of i/p
acquire K from i/p.
Decoder:- receive encoder's rep. + other i/p = target seq.
Optimized for generating o/p.

Encoder only models:- good for task req. understanding of i/p.
Sentence classification
NER
Decoder " " :- generative tasks such as text generation.
En-Decoder / seq-to-seq models:- generative task req. i/p.
translation / Summarization

Attn layers:- pay specific attn to certain words
Translation করার সময় certain u depends on certain other words.
Sense dependent.

Translation করার সময় certain w depends on certain other words.

Subject/noun এর উপর dependent.

meaning of a word affected by context.

Original Architecture:- en. i/p: source lang. + full sentence

de. " :- target lang. + only gen. words so far. + en. rep.

Decoder's Attn layers:—

1st: attn to previous decoder's o/p.

2nd:- " " encoder's o/p.

helpful যেটা context দেবে আগের as full sentence access করতে পারে।

Archi- Skeleton of model.

def. of each layer & op. within the model.

checkpoints- weights loaded in a given archi.

Encoders Archi.:-

Only use encoder. of Xmers.

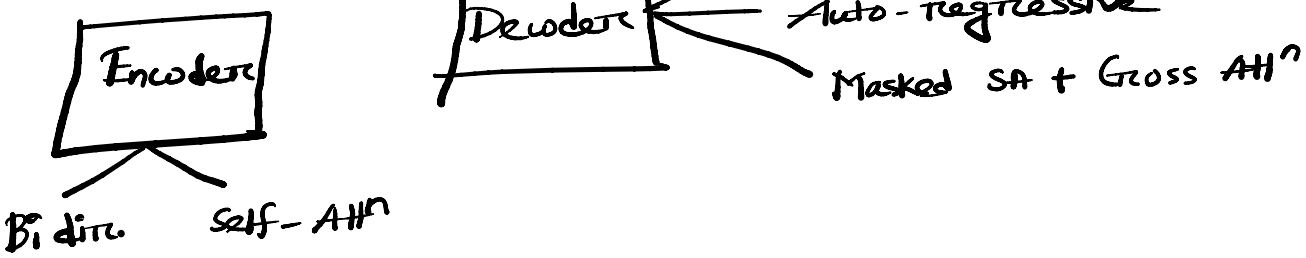
attn layers access all words in initial sentence.

bidirn. attn

auto-encoding models.

PT by MLM (Masked Lang. Modeling)

best suited for tasks req. understanding of full sentence.



Bi dirc. Self-Attn

Decoder models:-

only use decoder of Xmer

Attn access only before/left words.

Auto-regressive

generate seq. of texts by modeling the conditional probabilities. Of previous tokens.

CTRL

GPT-1, 2, 3

Xmer-XL

Seq-to-seq./En.-Decoder Models:-

Autoregressive manner

previous o/p of de. acts as i/p

Bias & limitations:-

PT on lots of data

best + worst data.