Lecture-5: A closer look at GPT

1. Zero shot Vs Few shot learning

Zero-Shot: Ability to generalize to completely unseen tasks without any pair.

- The model predict the answer given only a natural language description of the task. No gradient updates are performed.

specific examples:

1. Translate English to French:
2. cheese ⇒ French

(task description)

One-Shot:

- In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

eg.

1. Translate English to French:
2. sea otter ⇒ loutre de mer
3. cheese ← prompt

(task description; example)

Few Shot: Learning from a minimum number of examples which the user provides as input.

- In addition to the task description, the model see a few examples of the task. No gradient updates are performed.

eg.

1. Translate English to French:
2. sea otter ← example
3. peppermit ← example
4. plush giraffe
5. cheese ← prompt

(task descriptor)

Zero-Shot Vs Few Shot

	Input	output
Complete task without example. ↳ Zero-shot	Translate english to French	petit-dejeuner
Complete task with a few examples. ↳ few-shot	gaot → goat sheo → shoe phone →	phone

2. Utilizing Large Datasets

Let's us look at pretraining datasets of GPT-3 LLM.

Dataset name	Dataset description	Number of tokens	Proportion in training data
CommonCrawl (filtered)	Web crawl data	410 billion	60%
WebText2	Web crawl data	19 billion	22%
Books1	Internet-based book corpus	12 billion	8%
Books2	Internet-based book corpus	55 billion	8%
Wikipedia	High-quality text	3 billion	3%

A token is a unit of text which the model reads. For now, you can think of 1 token = 1 word.

Here take reference from GPT-3 main paper

* The total pre-training cost for GPT-3 ≈ 4.6 Billions dollors.

* These pretrained models are base / foundational models which can be used for further finetuning.

* Many pretrained LLMs are available as open-source models → can be used as general purpose tools to write, extract and edit text which was not part of the training data.

3. GPT Architecture

GPT → Generative pretrained Transformer

Original paper 2018

↳ GPT-3 is a scaled up version of this model, implemented on a larger dataset

* GPT models are simply trained on "next-word" prediction tasks.

The lion roams in the jungle.
next-word

* with this training, they can do a wide range of other tasks like translation, spelling correction etc.

* Next-word prediction: self-supervised learning
↓
Self-labelling

* we don't collect labels for training data, but use the structure of of the data itself.

next word in sentence is used as label

Auto regressive model.
use previous outputs as input for future prediction

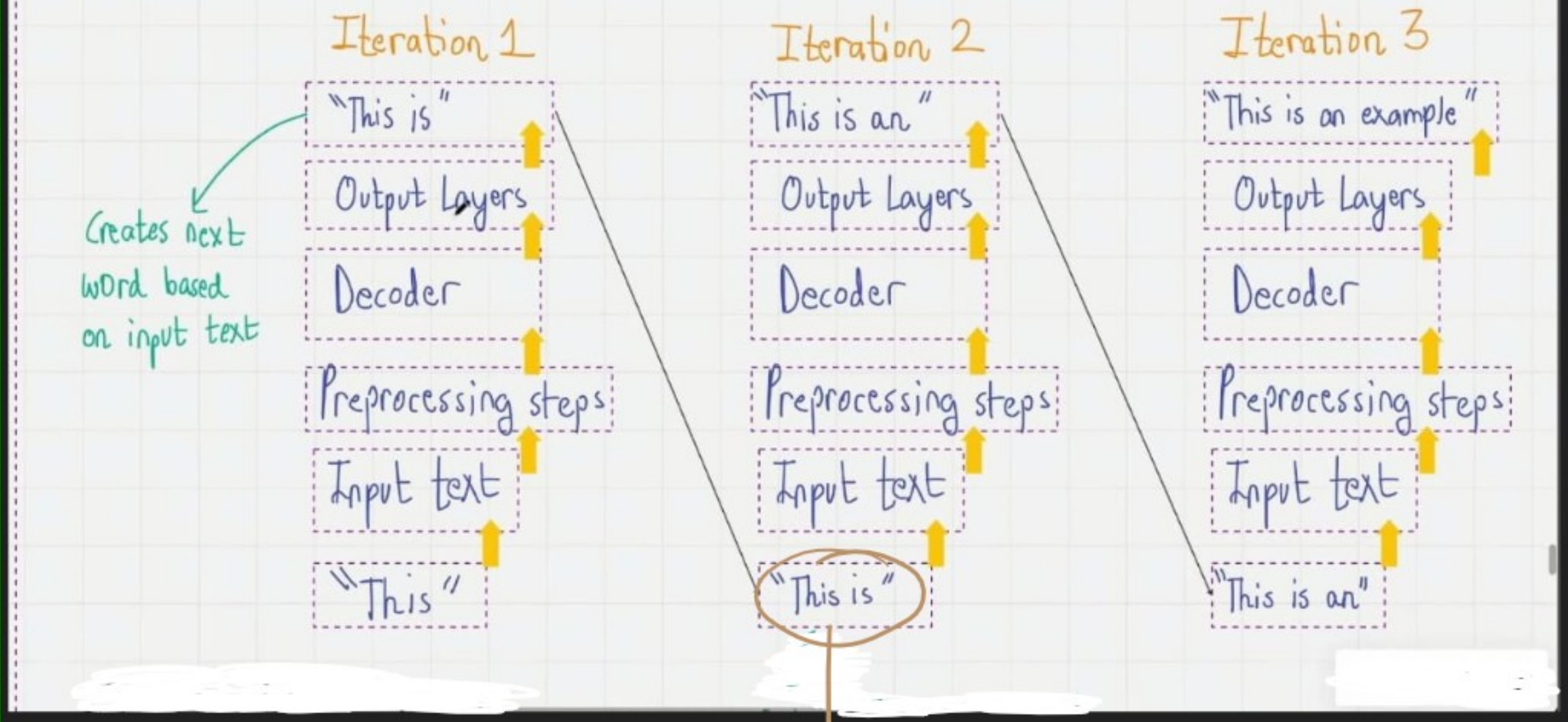
*** pretraining part of GPT model.
↓
Supervised
↓
Auto-regressive model

* compared to original transformer architecture GPT architecture is simpler.

* In GPT architecture: there is no encoder we just have the decoder.

Original transformer = 6 encoder-decoder block

GPT-3: 96 transformer layers,
175 billions parameters



GPT Architecture (decoder only)

output of previous round serve as input to next round.

Although trained only for next word prediction GPT model can perform other tasks like language translation.

This is called "emergent behaviour

Ability of a model to perform tasks that the model wasn't explicity trained to perform.