**GPT Types:**

There are several versions of the GPT (Generative Pre-trained Transformer) architecture, each with a different number of parameters and capabilities. Here is a list of the currently known versions of GPT:

GPT: The original version of GPT was released by OpenAI in 2018 and had 117 million parameters.

GPT-2: Released in 2019, GPT-2 is a larger version of the original model with 1.5 billion parameters.

GPT-3: Released in 2020, GPT-3 is the largest version of the model to date, with 175 billion parameters. It has been shown to perform well on a wide range of natural language processing tasks, including language translation, question answering, and text generation.

GPT-Neo: GPT-Neo is a series of models developed by EleutherAI, an open-source research organization. The models range in size from 1.3 billion to 2.7 billion parameters.

GPT-4: As of now, there is no official release or announcement for GPT-4 from OpenAI or any other organization.

It's worth noting that there may be other versions of GPT that are not publicly known or that have not yet been released.

**Structure of GPT and GPT-2:**

The structure of GPT-2 (Generative Pre-trained Transformer 2) is very similar to that of the original GPT (Generative Pre-trained Transformer) model. Both models use the Transformer architecture, which includes multi-head self-attention layers and feedforward layers. However, there are some key differences between the two models.

GPT-2 is significantly larger than the original GPT model, with 1.5 billion parameters compared to the original model's 117 million parameters. This allows GPT-2 to generate more complex and nuanced language.

In addition, GPT-2 includes a few architectural tweaks that were not present in the original model, such as layer normalization and a modified position embedding scheme.

Despite these differences, both models are based on the same fundamental architecture and share many of the same design principles.

More depth to GPT ( Decoder only architecture) :

Diagram

Description automatically generated with medium confidence

Diagram

Description automatically generated

**The modifications made to the decoder in the GPT (Generative Pre-trained Transformer) architecture:**

1. Layer normalization: In the original Transformer architecture, layer normalization is applied after each sub-layer, which includes both the self-attention and feedforward layers. However, in the GPT architecture, layer normalization is applied after each sub-layer within the decoder only. This modification helps to stabilize the training process and improves the model's performance.
2. Modified position embedding scheme: In the original Transformer architecture, the position embeddings are added to the input embeddings to provide positional information to the model. However, in the GPT architecture, the position embeddings are added to the output of each sub-layer within the decoder only. This modification is intended to prevent the model from "cheating" by attending to the position embeddings rather than the input text.

Hardware:

1. IMCus
2. GPUs
3. FPU

Memory of the dataset : Less than 100 or 200 GB ( 2 or 3 datasets)

And can be trained in 2 days

Training the model:

1. Training from scratch
2. Pre trained model (fine tuning)

Dataset :

1. WikiText-103 - <https://pytorch.org/text/stable/datasets.html#torchtext.datasets.WikiText103> Models-( RoBERTa, XLNet, GPT-2 , BERT, T5)
2. GloVe. - <https://nlp.stanford.edu/projects/glove/> Models-( RoBERTa, XLNet, GPT-2)
3. MSCOCO - <https://cocodataset.org/#download> Models- ( ViT, DeiT)
4. CIFAR-10 - <https://www.cs.toronto.edu/~kriz/cifar.html> Models- ( ViT, DeiT)