LLAMA  x 3

Tom 

University of Colorado Boulder

Ravi 

LlamaIndex

Eduardo 

aws

AI by Hand ✍️

RAG

MLOps

LLaMA 3
What are the dimensions?

University of Colorado Boulder

8B model parameters

- # of Layers = _____
- # of Attention Heads = _____
- # of Vocabulary Words = _____
- # of Feature Dimensions = _____
- # of Hidden Dimensions = _____
- Context Window Size = _____

meta-llama / **Meta-Llama-3-8B** like 2.08k

main Meta-Llama-3-8B / original / params.json

pcuenq HF STAFF Upload original checkpoint (#1)

raw history blame contribute delete No virus

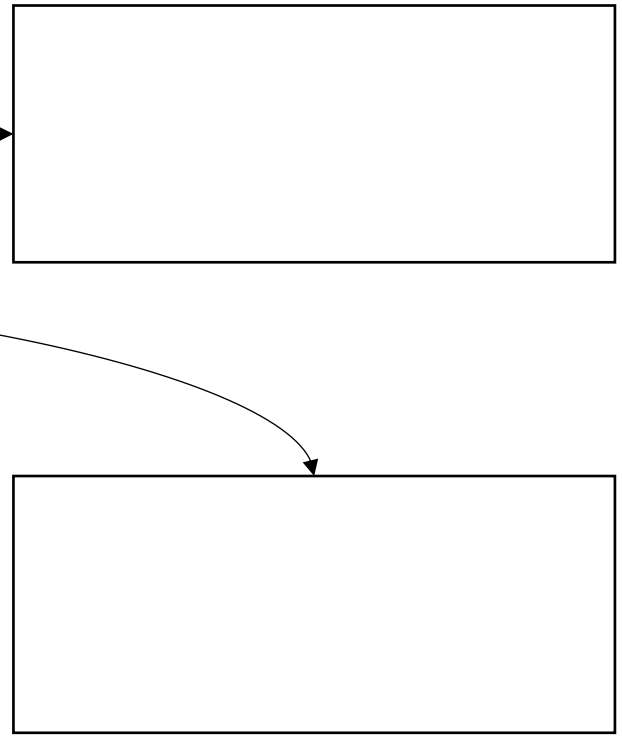
```
{
    "dim": 4096,
    "n_layers": 32,
    "n_heads": 32,
    "n_kv_heads": 8,
    "vocab_size": 128256,
    "multiple_of": 1024,
    "ffn_dim_multiplier": 1.3,
    "norm_eps": 1e-05,
    "rope_theta": 500000.0
}
```

Llama

```
35  class Llama:
36      @staticmethod
37      def build(
38          ckpt_dir: str,
39          tokenizer_path: str,
40          max_seq_len: int,
41          max_batch_size: int,
42          model_parallel_size: Op
43          seed: int = 1,
44      ) -> "Llama":
```

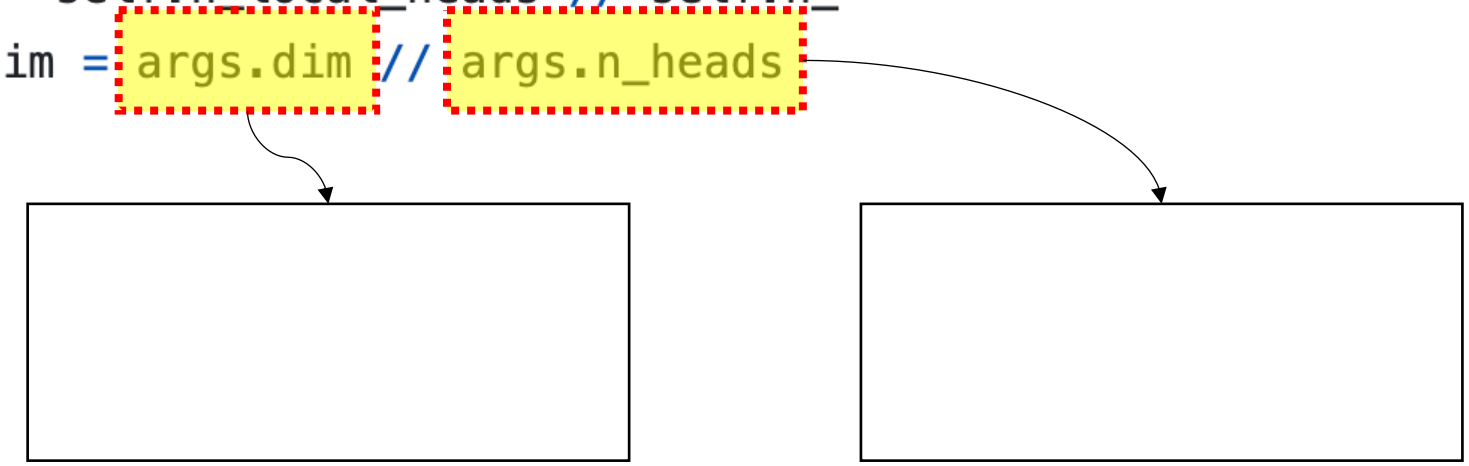
Transformer

```
251  class Transformer(nn.Module):
252      def __init__(self, params:
253          super().__init__()
254          self.params = params
255          self.vocab_size = param
256          self.n_layers = params.
257
```



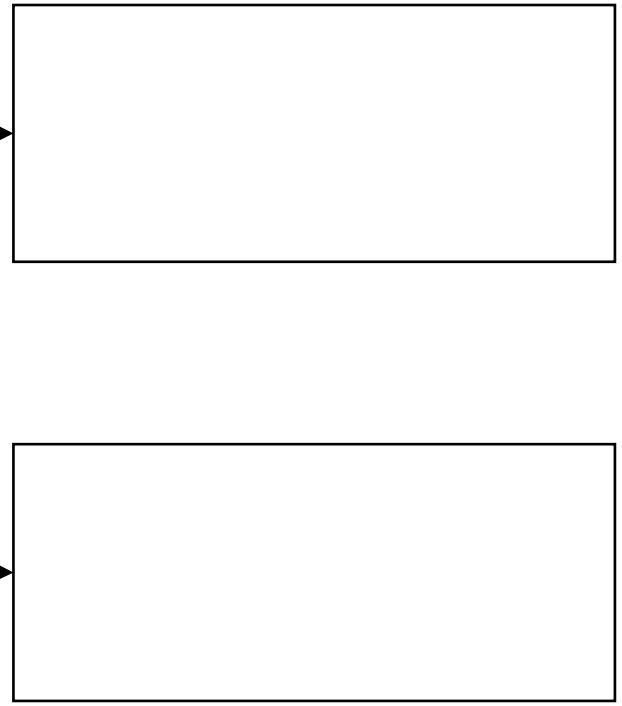
Self Attention

```
90  class Attention(nn.Module):
91      def __init__(self, args: ModelArgs):
92          super().__init__()
93          self.n_kv_heads = args.n_heads if args.n_k
94          model_parallel_size = fs_init.get_model_pa
95          self.n_local_heads = args.n_heads // model
96          self.n_local_kv_heads = self.n_kv_heads //
97          self.n_rep = self.n_local_heads // self.n_
98          self.head_dim = args.dim // args.n_heads
```

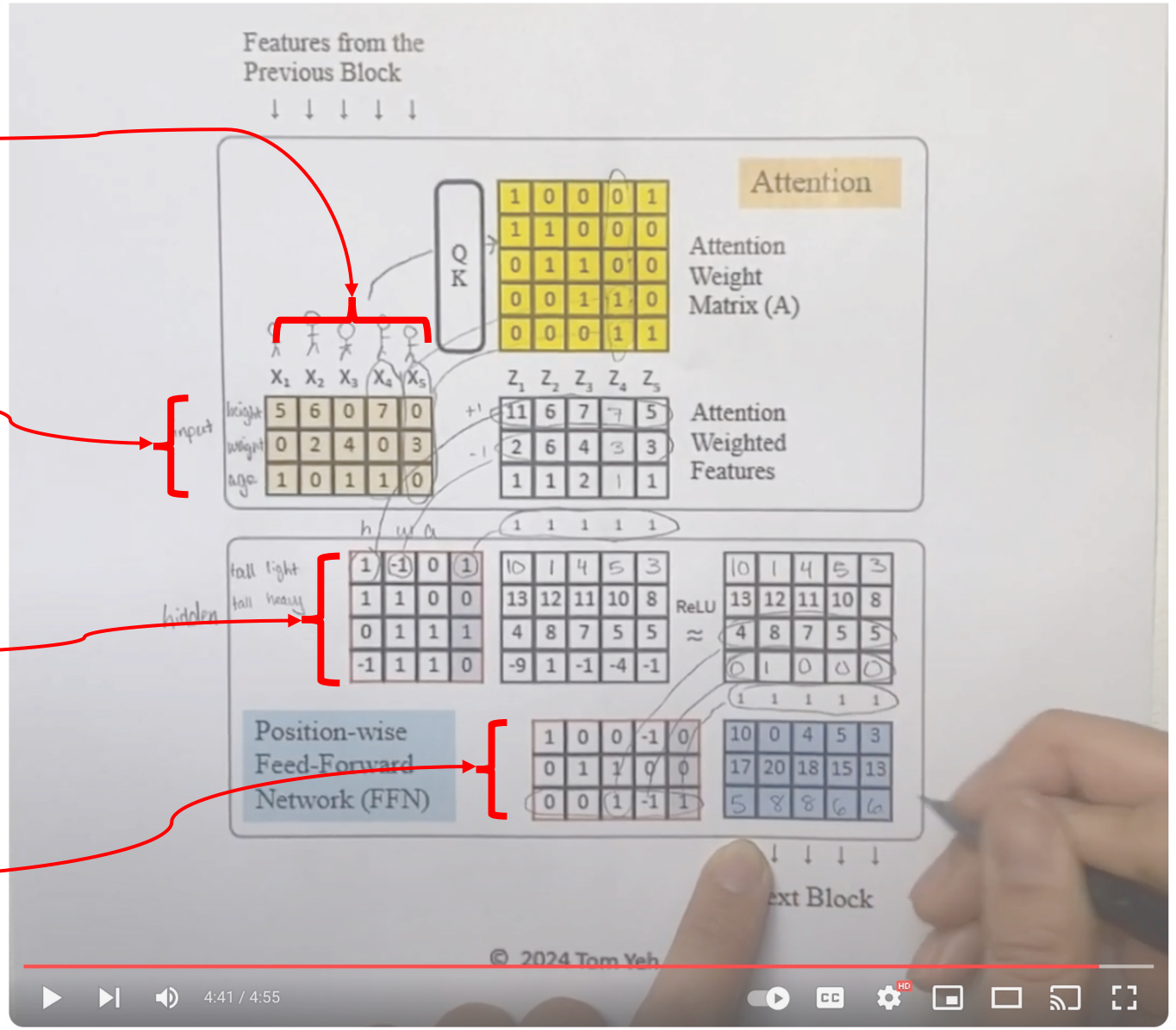


Feed Forward

```
193  class FeedForward(nn.Module):
194      def __init__(
195          self,
196          dim: int,
197          hidden_dim: int,
198          multiple_of: int,
199          ffn_dim_multiplier: Optio
200      ):
```

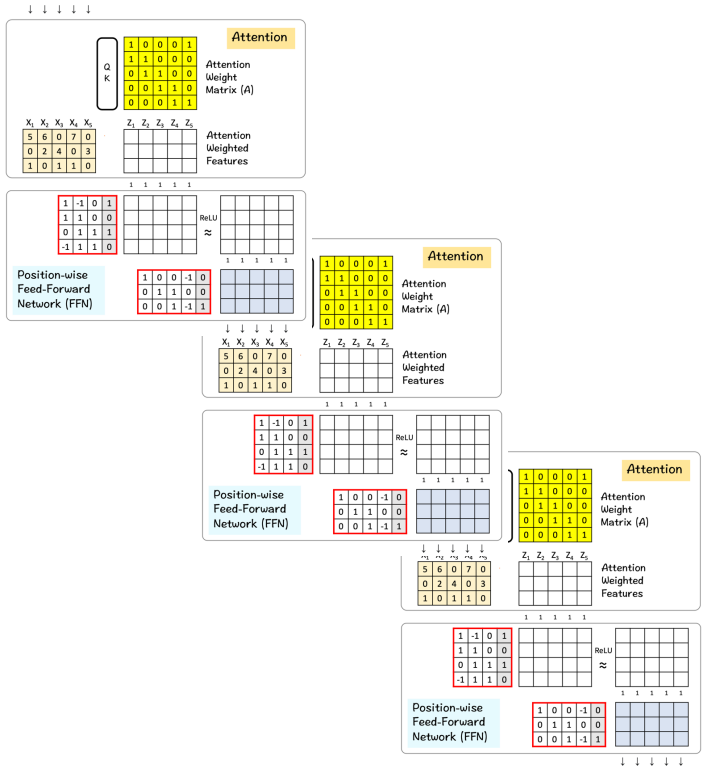


Transformer Block

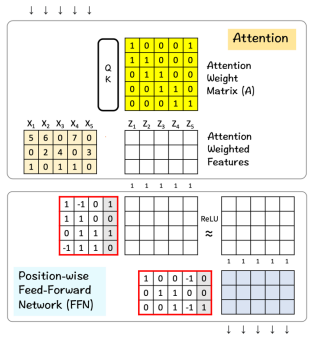


S2.E1 - Transformer - AI by Hand ✍️ with Anna

Layers



. . . . . .



X

Input / Output