batch = B
seq_len = L
dim = D
n_heads = Nh
n_kv_heads = Nkv
hidden_dim = Dh
n_layer = Nl
vocab_size = V

Llama model architecture accoding to facebookresearch/llama/llama/model.py and karpathy/llama2.c/model.py

Dh = 4 * 2 * D // 3 then round-up to multiple_of