08

TUESDAY 189-176

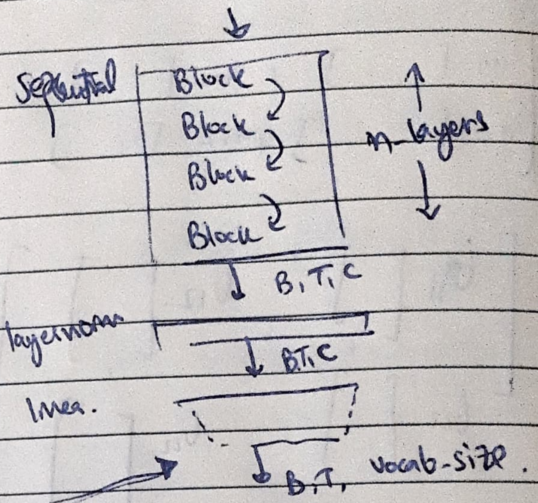
WK 28

C = n_embd

x_b, y_b : [ ] B, T — batch size × block size (16, 32)

x_b, y_b.

x = tok-emb + pos-emb = B,T,C
B,T,C + T,C



① model = Bigram LM()

② logits, loss = model(x_b, y_b).

self.token embds = [ ] vocab size, n embd (65, 64)

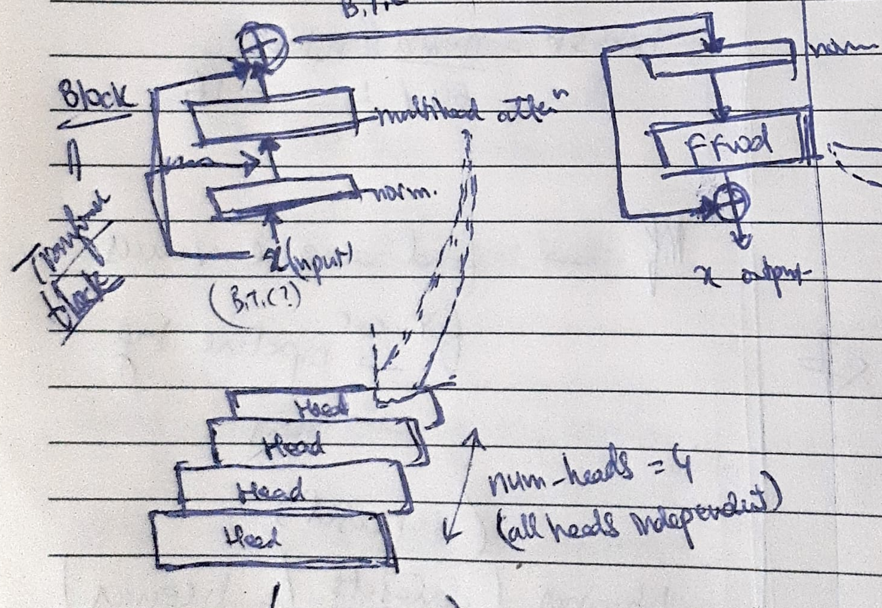
self.pos embds = [ ] block size, n embd (32, 64)

self.blocks = sequential(Block( ),
Block ( )
⋮
Block( ))

n_layer = 4

self.layernorm = nn.LayerNorm (n_embd)

self.lm head = nn.Linear (n embd, vocab size)



B,T,C
↓
seq. (Lin (C, 4C)
↓ B,T,4C
ReLU
↓
Lin (4C, C)
↓ B,T,C
Dropout
↓
B,T,C.

↓
concat (dim=-1)

Notes

[ ]
↓ B,T,C
linear. n_emb, n_emb
[ ]
↓ B,T,C.
dropout
↓
B,T,C

JULY

AUGUST

18	19	20	21	22	23	24	25	26	27	28	29	30	31	11	12	13	14	15	16	17
M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S

B, T, C

Head

head_size = n_embd / n_head = 64 / 4 = 16

K = B, T, C — linear — B, T, head_size

Q = B, T, C — linear — B, T, head_size

V = (same)

wei = B, T, head_size × B, head_size, T

wei = (B, T, T) / √C (C = n_embd)

softmax

(T) . (T)

block size x block size

dropout

B, T, T

out = wei @ V

(B, T, T) @ (B, T, C)

(B, T, C)

wei