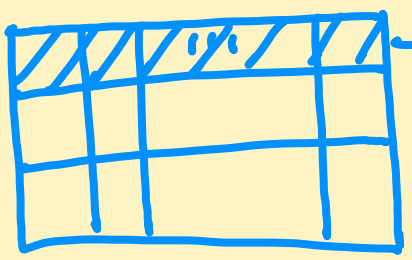
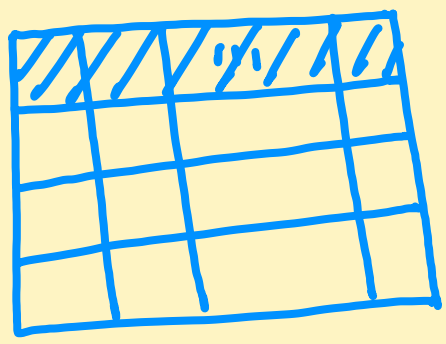
Transformers w. External Memory

(Vaswani et al.)



Q : len_Q x d
(768)



K.V : len_K x d

(len_K = len_V)

Outputs = Softmax(QK^T).V

(len_Q x d x d x len_K) x len_V x d

→ len_Q x d

(Sukhbaatar et al.)

U (ffn) : 4d x d

↑

V (ffn) : d x 4d

↑

Outputs (Q.K.V) : len_Q x d

Outputs_ffn

= Softmax(Outputs.V).U

ReLU(x)

(len_Q x d x d x 4d)

x 4d x d

→ len_Q x d

(External Memory) ∴ V ≈ K̃^T (Key)

U ≈ Ṽ (Value)