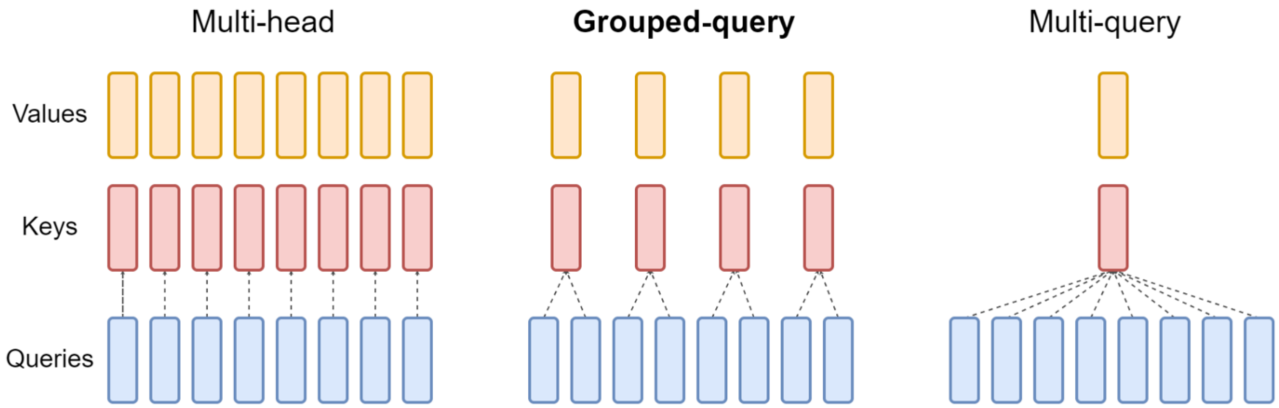
Recent extension – GQA

Don’t go all the way to one dimension of KV – have fewer dims



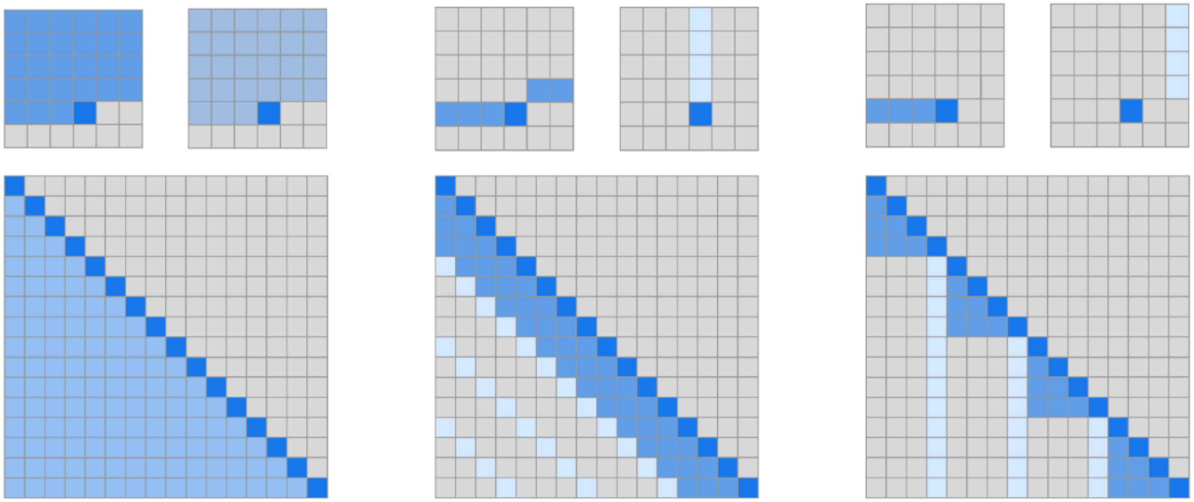
Simple knob to control expressiveness (key-query ratio) and inference efficiency

Borrowed from Stanford CS336

Sparse / sliding window attention

Attending to the entire context can be expensive (quadratic).

Build sparse / structured attention that trades off expressiveness vs runtime (GPT3)



(a) Transformer

(b) Sparse Transformer (strided)

(c) Sparse Transformer (fixed)

[Child et al 2019]

Borrowed from Stanford CS336