Large Language Models

Advanced Attention Mechanisms - I

ELL881 · AIL821

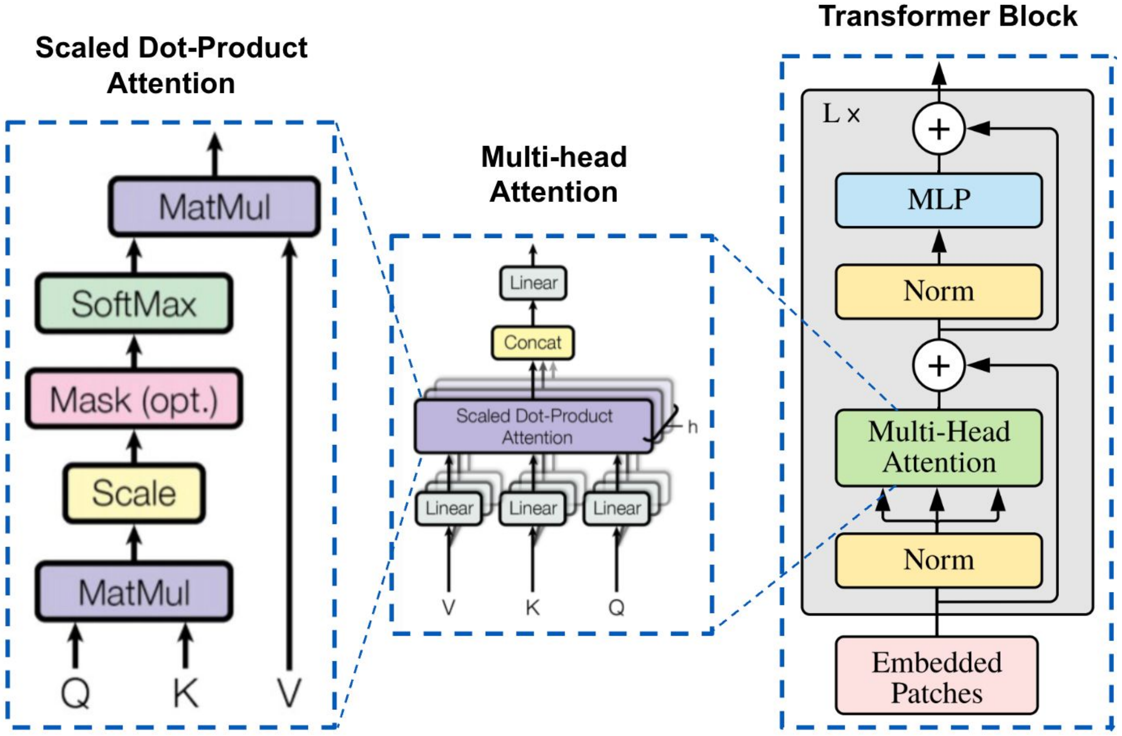
Sourish Dasgupta
Assistant Professor, DA-IICT, Gandhinagar
https://daiict.ac.in/faculty/sourish-dasgupta

Year: 2017, NeurIPS

Google Brain Team

Make machines intelligent. Improve people's lives.

Self Attention



$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\frac{\underset{(6,\ 4096)}{Q} \times \underset{(4096,\ 6)}{K^T}}{\sqrt{4096}} =$$

	THE	CAT	IS	ON	A	CHAIR
THE	0.268	0.119	0.134	0.148	0.179	0.152
CAT	0.124	0.278	0.201	0.128	0.154	0.115
IS	0.147	0.132	0.262	0.097	0.218	0.145
ON	0.210	0.128	0.206	0.212	0.119	0.125
A	0.146	0.158	0.152	0.143	0.227	0.174
CHAIR	0.195	0.114	0.203	0.103	0.157	0.229

(6, 6)

Year: 2017, NeurIPS

Google Brain Team

Make machines intelligent. Improve people's lives.

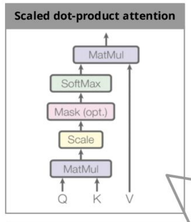
Causal (Forward Masked) Attention

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\frac{\underset{(6,\ 4096)}{Q} \times \underset{(4096,\ 6)}{K^T}}{\sqrt{4096}} =$$

	THE	CAT	IS	ON	A	CHAIR
THE	0.268	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
CAT	0.124	0.278	$-\infty$	$-\infty$	$-\infty$	$-\infty$
IS	0.147	0.132	0.262	$-\infty$	$-\infty$	$-\infty$
ON	0.210	0.128	0.206	0.212	$-\infty$	$-\infty$
A	0.146	0.158	0.152	0.143	0.227	$-\infty$
CHAIR	0.195	0.114	0.203	0.103	0.157	0.229

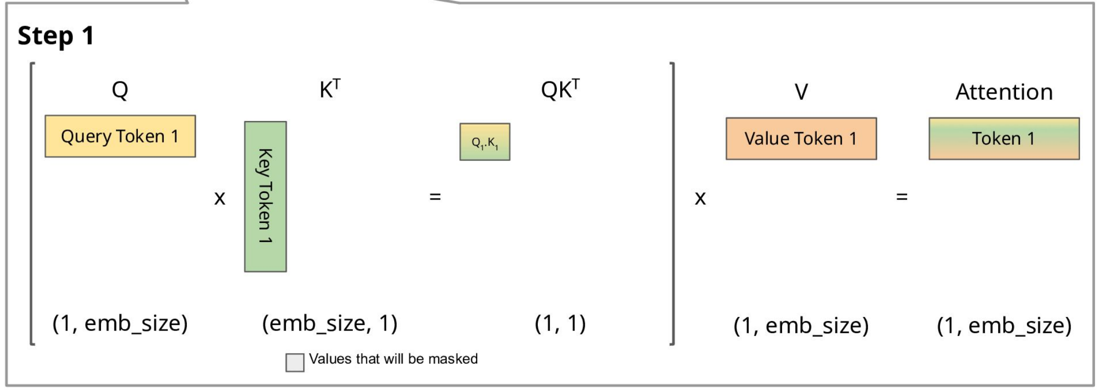
(6, 6)



Time: $O(N^2*d) + O(N^2) + O(N^2*d) = O(N^2*d)$

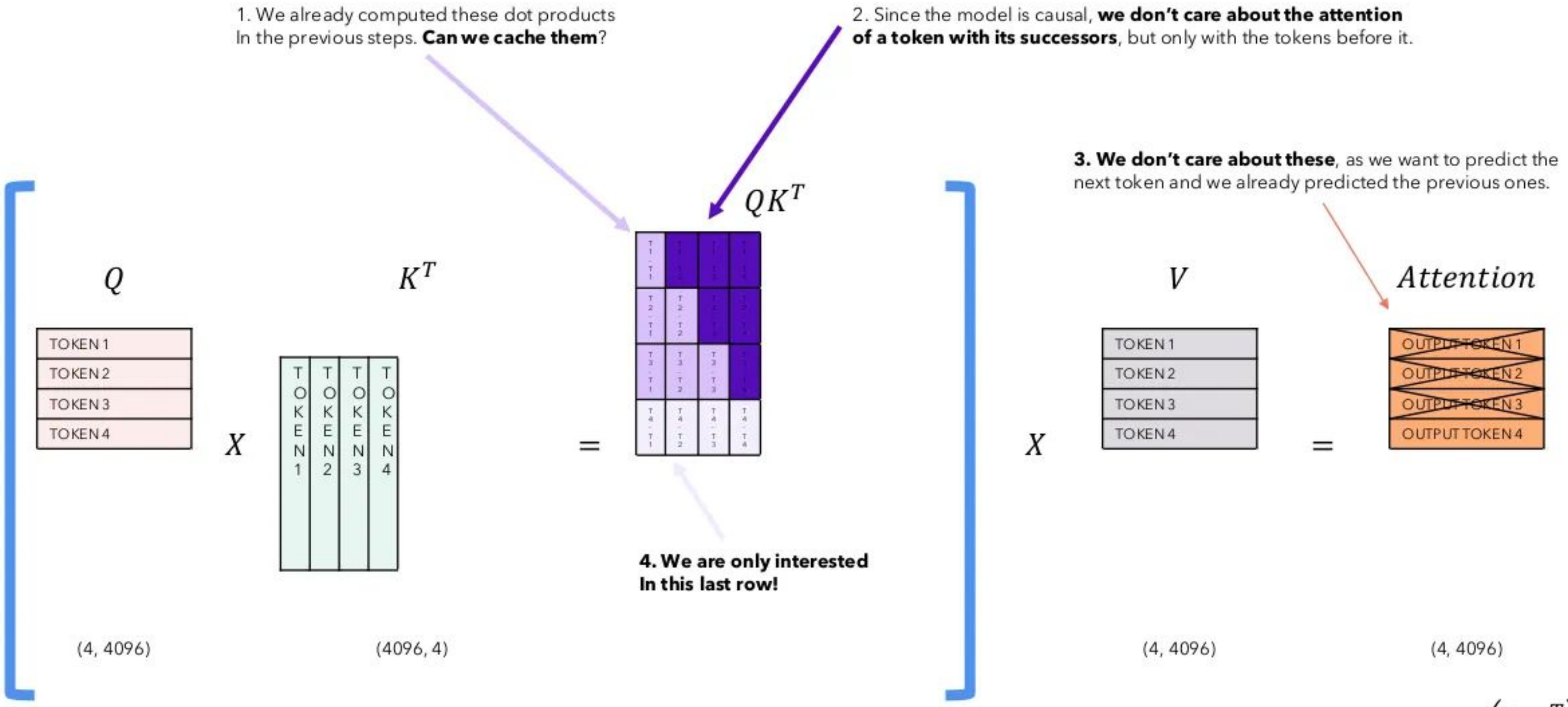
Space: $O(3*N*d) + O(N^2) + O(N*d) = O(N^2 + N*d)$

$= (3\times N\times d+N^2)\times$*Size of a float*



Zoom-in! (simplified without Scale and Softmax)

Why do we need to do better?

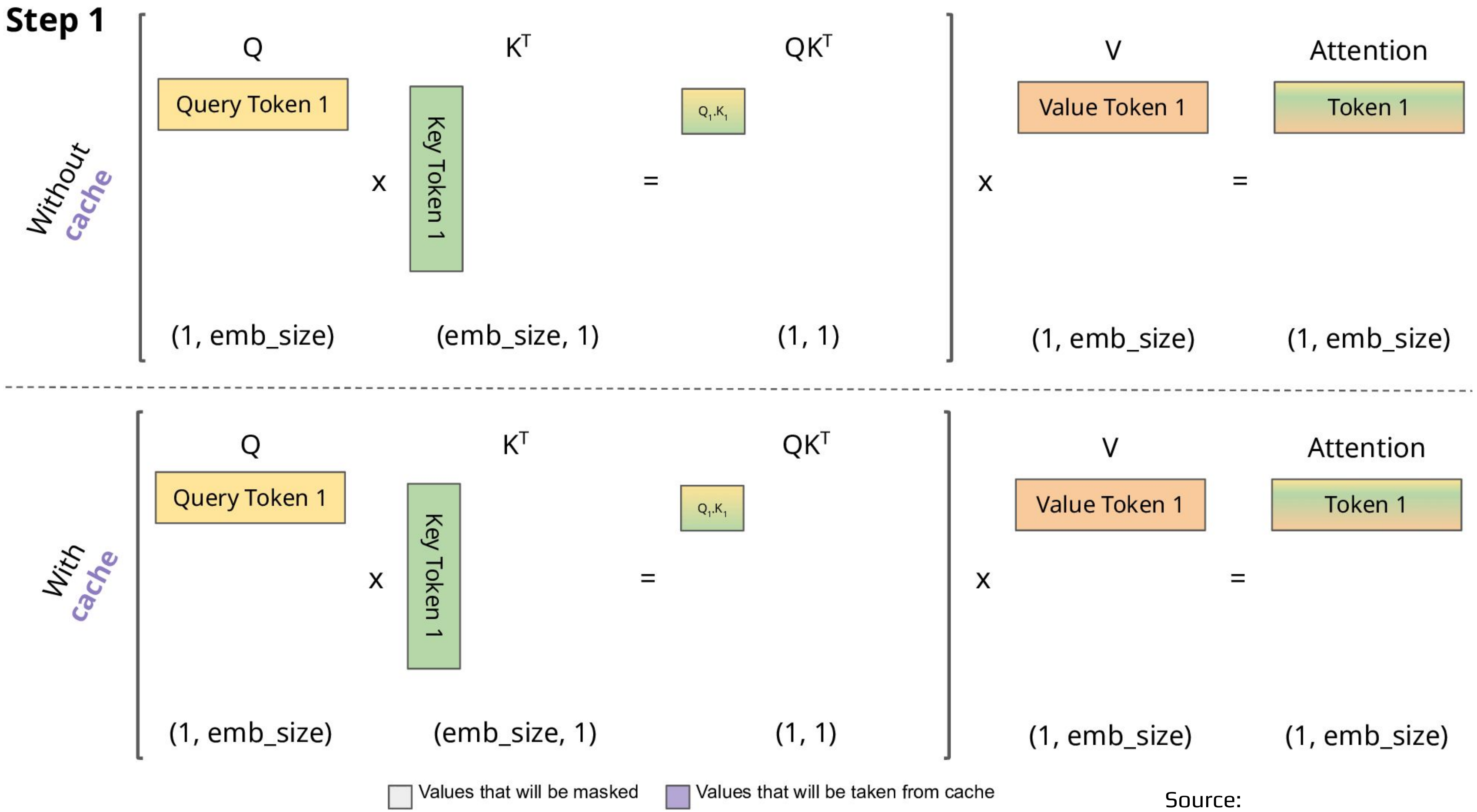


Inference
T = 4

(4, 4)

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

KV Cache based (Forward Masked) Attention



KV Cache Storage = O(N*d)

(2×N×d)×Size of a float

vs.

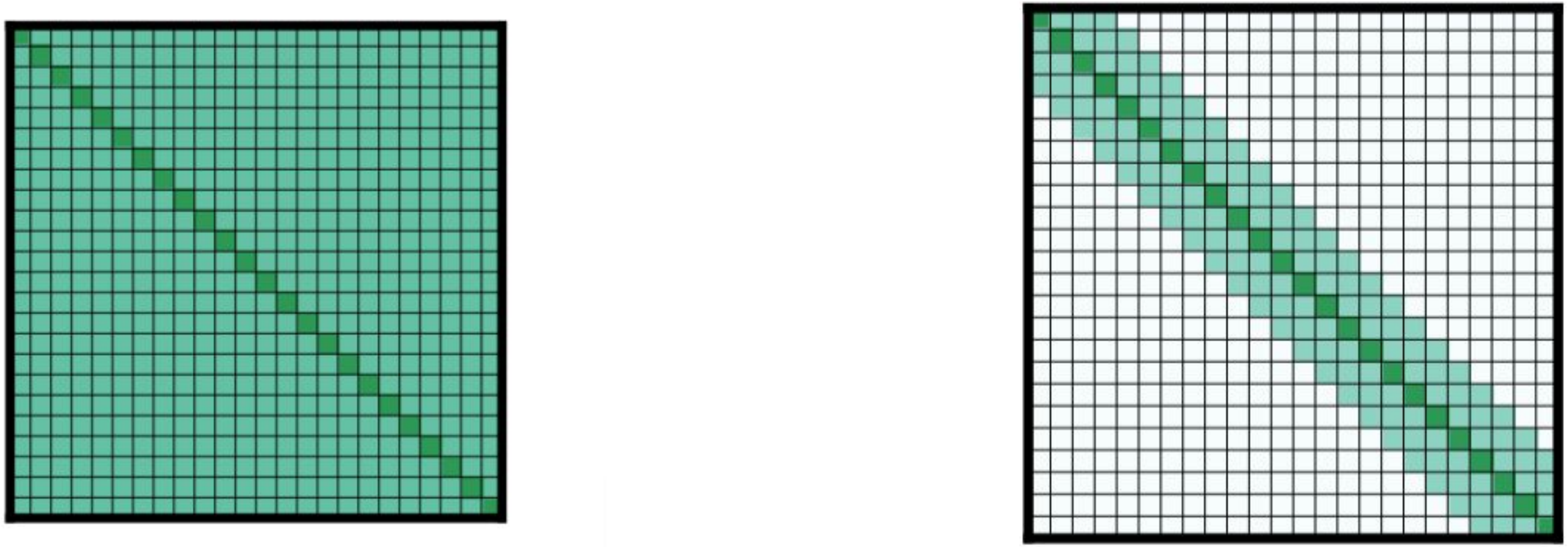
(3×N×d+N^2)×Size of a float

Source: https://medium.com/@joaolages/kv-caching-explained-276520203249

Year: 2020, Arxiv

Allen Institute for AI

Sliding Window Attention



(a) Full n^2 attention

(b) Sliding window attention

Year: 2020, Arxiv

Allen Institute for AI

Sliding Window Attention

$$Attention(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\mathrm{softmax}\left(\frac{\underset{(6,\,4096)}{Q} \times \underset{(4096,\,6)}{K^T}}{\sqrt{4096}}\right) =$$

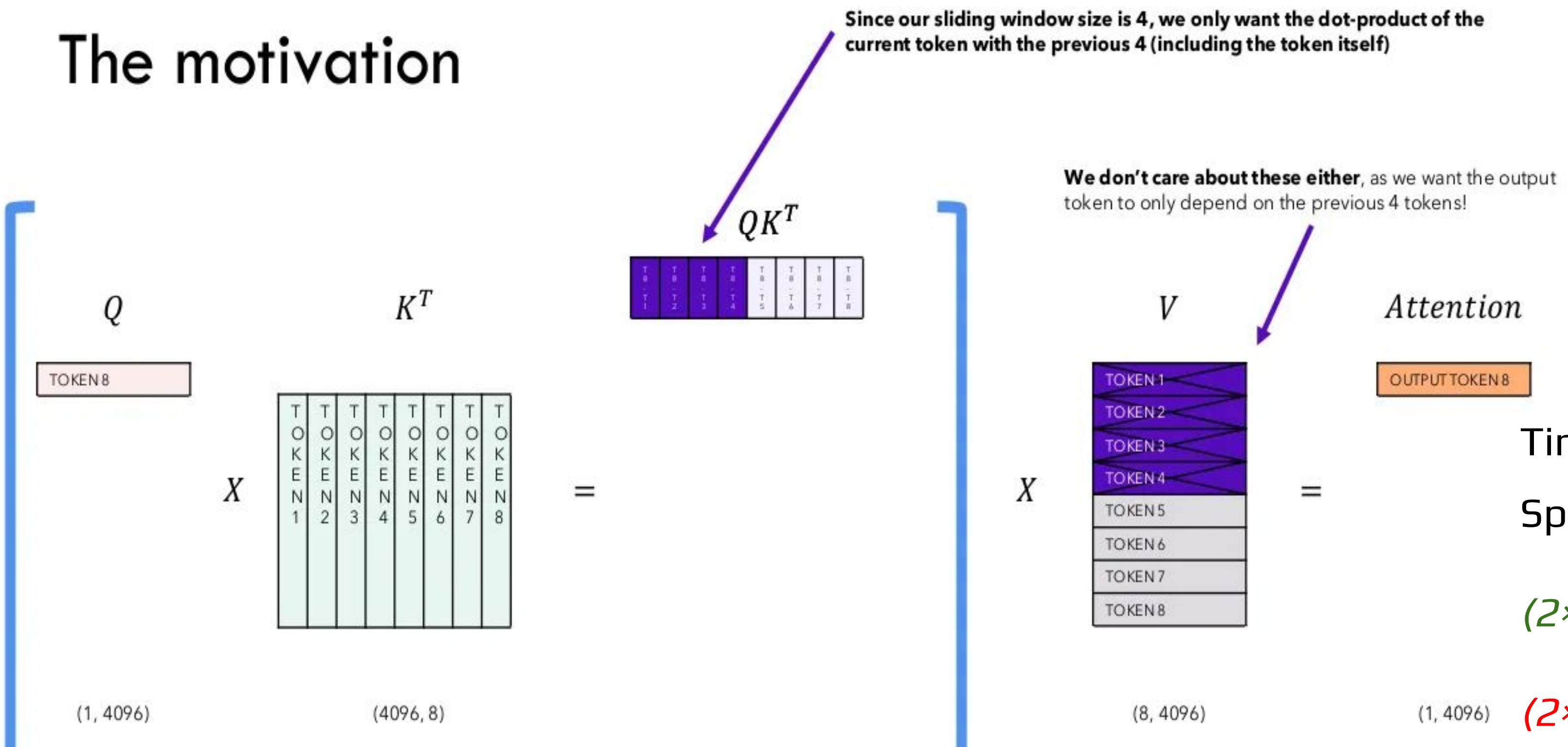
	THE	CAT	IS	ON	A	CHAIR
THE	1.0	0	0	0	0	0
CAT	0.461	0.538	0	0	0	0
IS	0.3219	0.317	0.361	0	0	0
ON	0	0.316	0.341	0.343	0	0
A	0	0	0.326	0.323	0.351	0
CHAIR	0	0	0	0.313	0.331	0.356

Year: 2020, Arxiv

Allen Institute for AI

What happens to the KV Cache?

The motivation



(1, 8)

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Inference
T = 4

Time: O(N*w*d) + O(N*w) + O(N*w*d) = O(N*w*d)

Space: O(2*N*d) + O(N*w) = O(N*w + N*d)

*(2×**w**×d)×Size of a float*

vs.

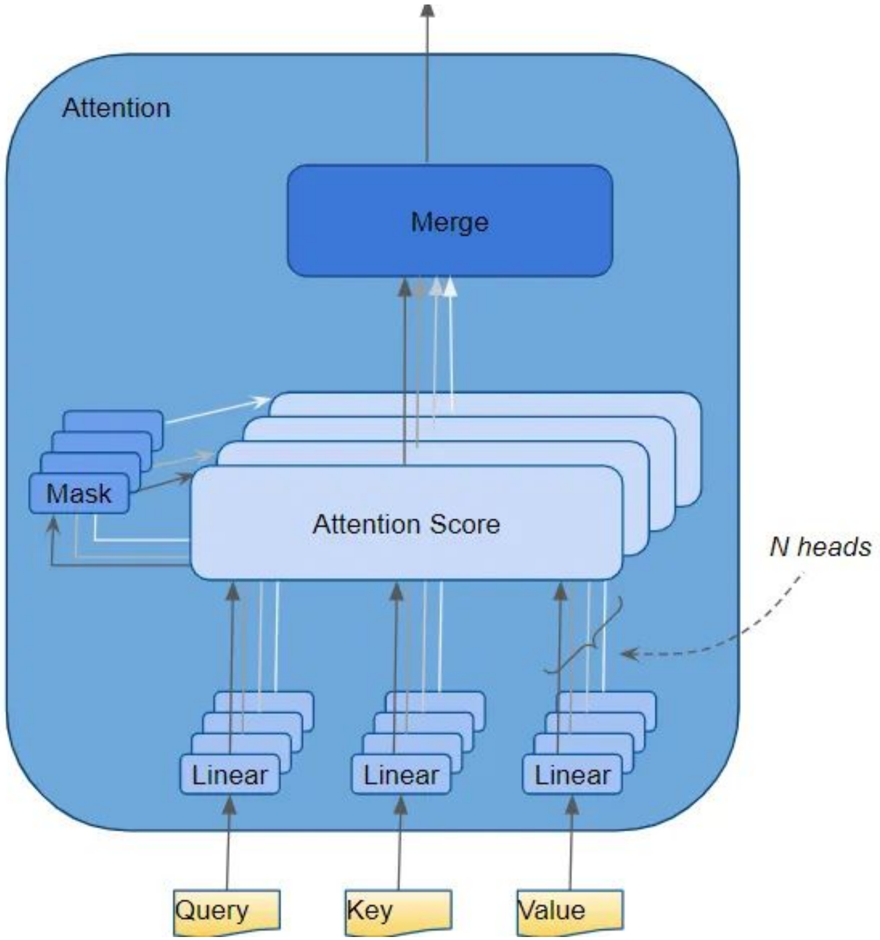
(2×N×d)×Size of a float

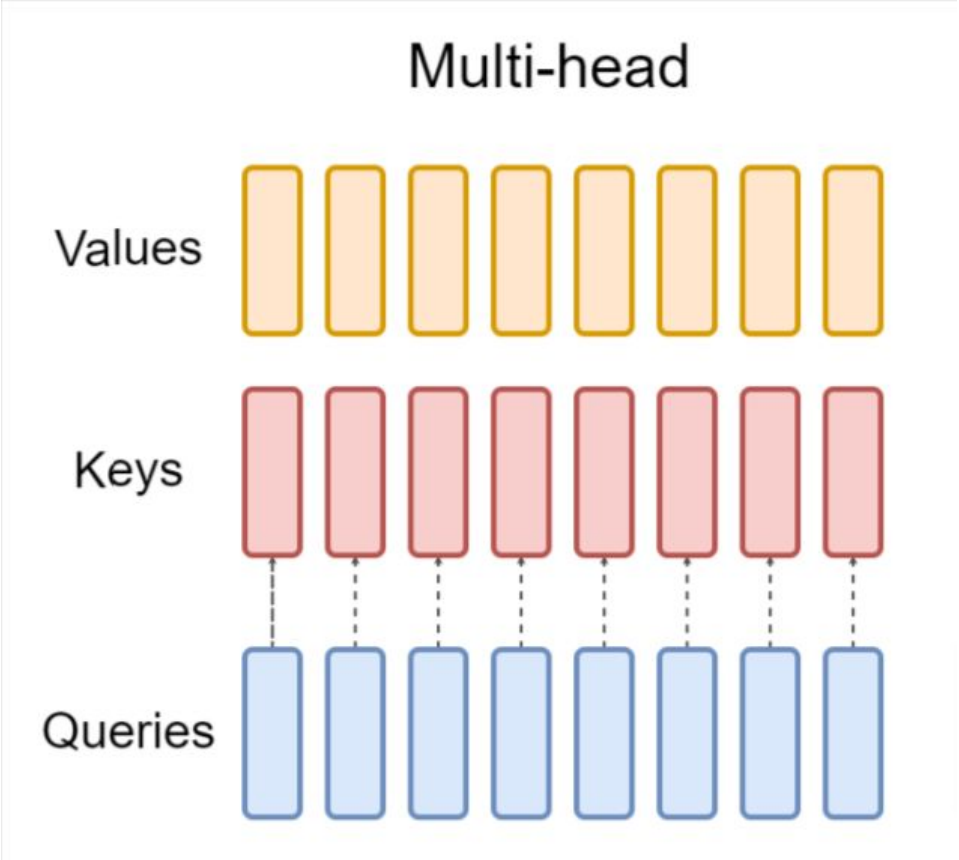
Going back to year: 2017, NeurIPS

Google Brain Team

Make machines intelligent. Improve people's lives.

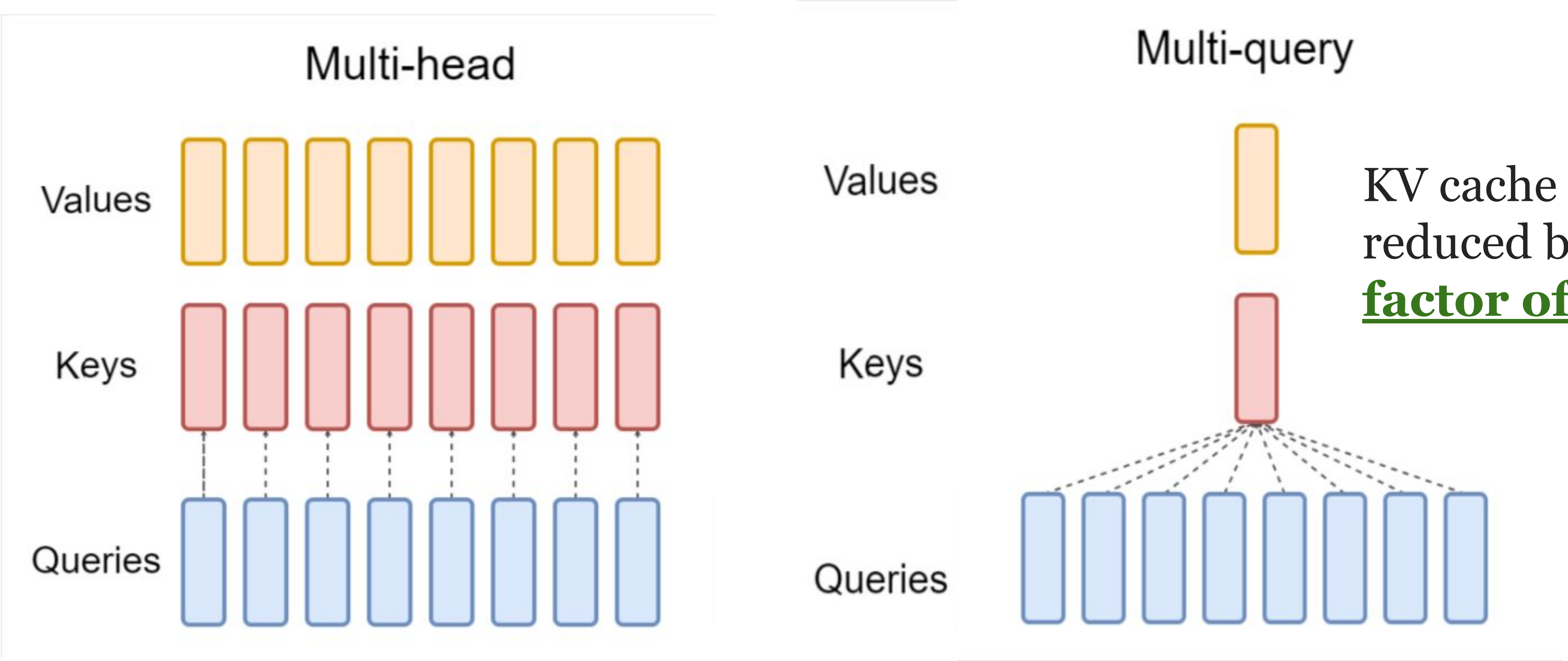
Multi-Head Self Attention





Year: 2019, arxiv Google

Multi-Query Attention (MQA)



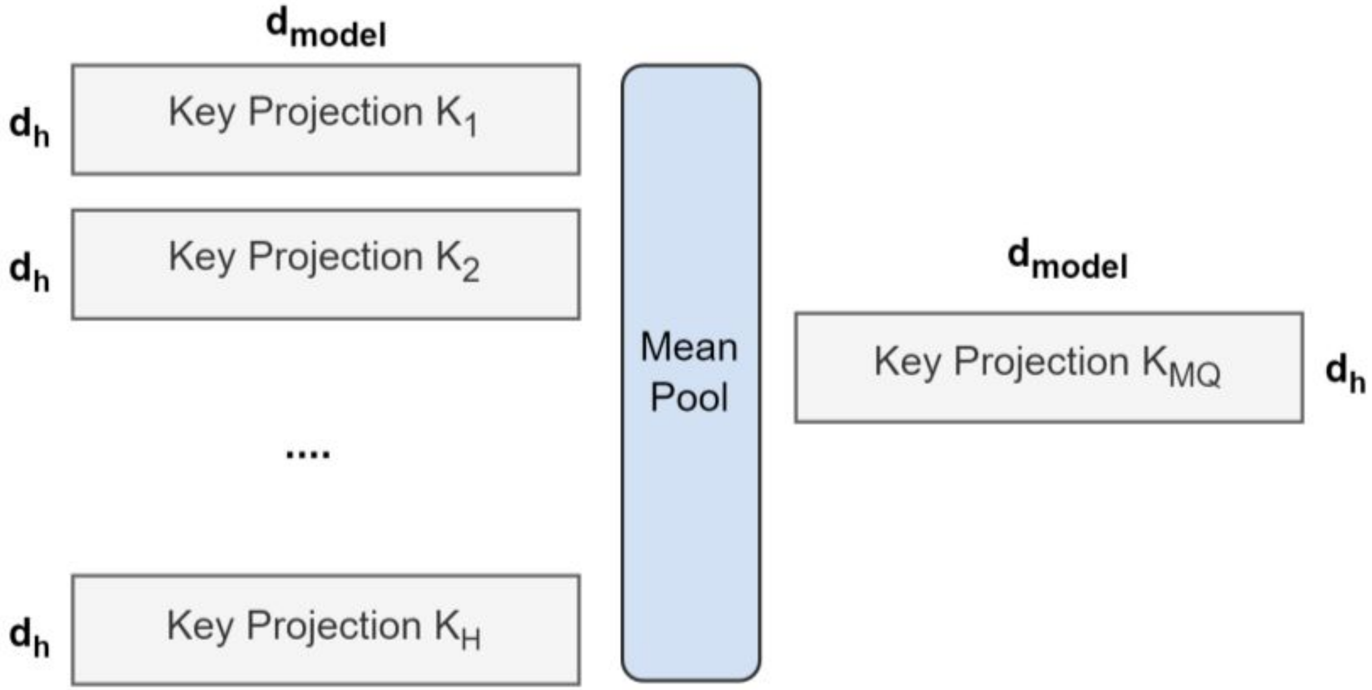
KV cache size is reduced by a **factor of *h***

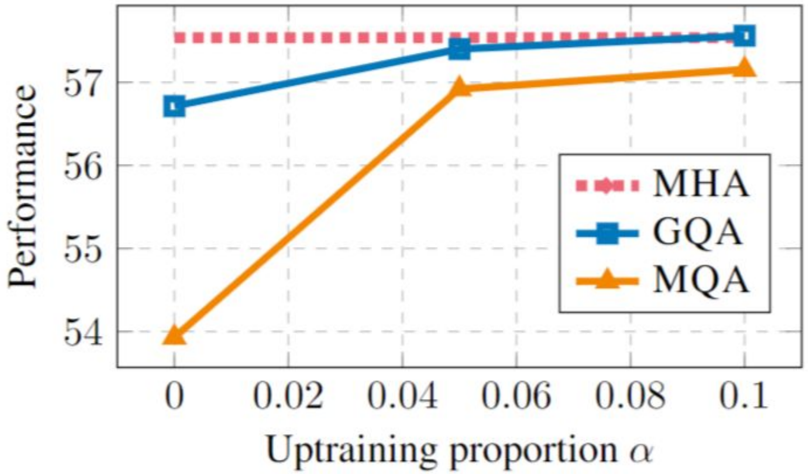
Do we lose out on something?

- Decline in performance quality
- Training instability

Year: 2023; ICLR

Uptraining: Converting MHA to MQA



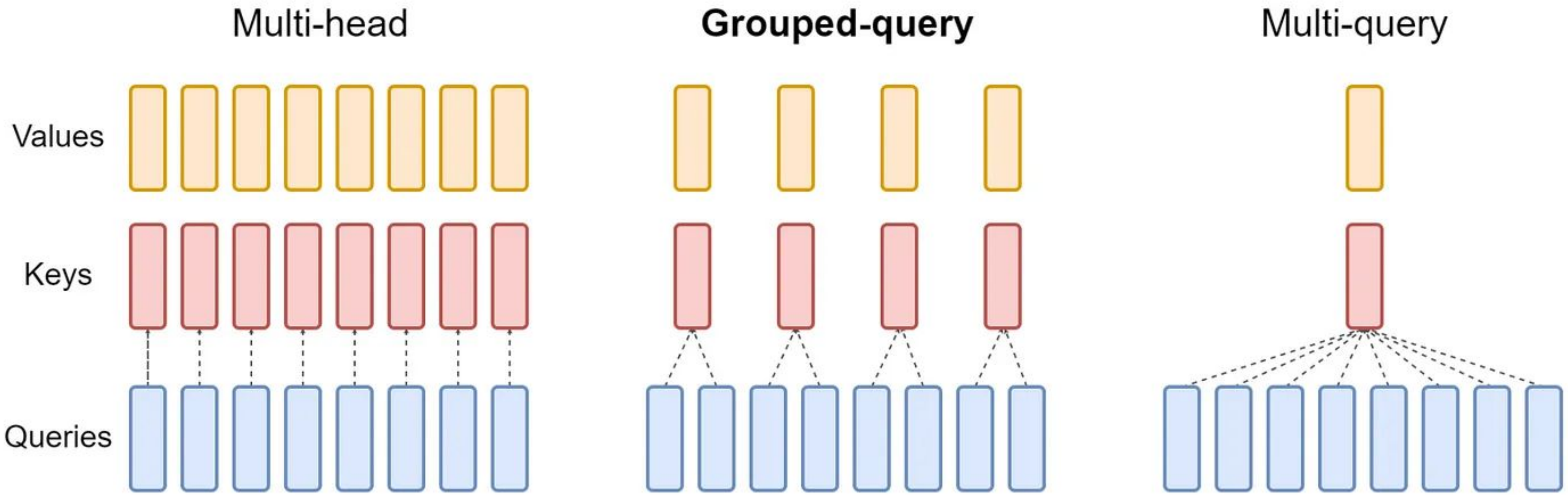


What can still go wrong?

- Decline in performance quality
- ~~Training instability~~

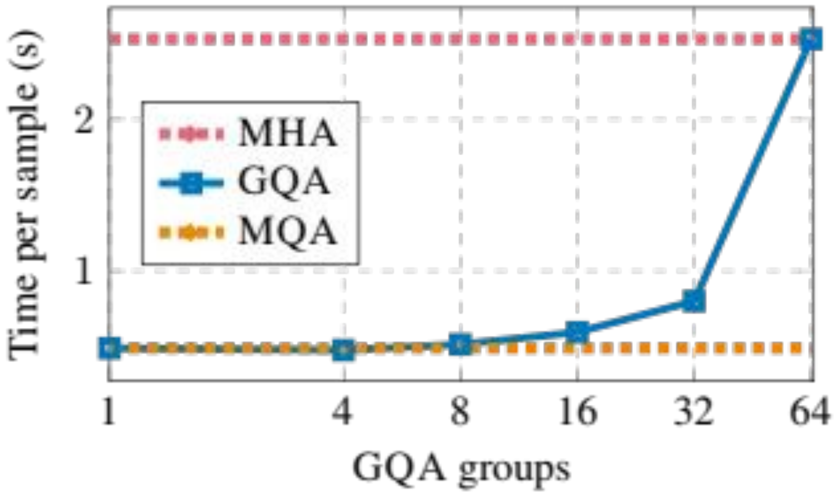
Year: 2023; EMNLP

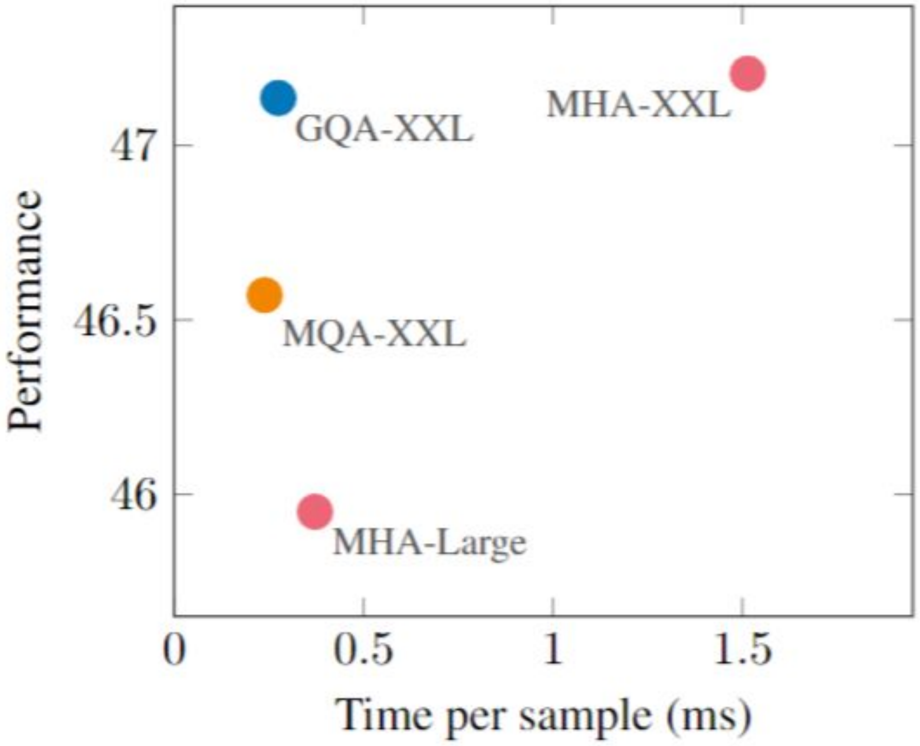
Grouped Query Attention



What did we gain?

Model	T_{infer}	Average	CNN	arXiv	PubMed	MediaSum	MultiNews	WMT	TriviaQA
	s		R_1	R_1	R_1	R_1	R_1	BLEU	F1
MHA-Large	0.37	46.0	42.9	44.6	46.2	35.5	46.6	27.7	78.2
MHA-XXL	1.51	47.2	43.8	45.6	47.5	36.4	46.9	28.4	81.9
MQA-XXL	0.24	46.6	43.0	45.0	46.9	36.1	46.5	28.5	81.3
GQA-8-XXL	0.28	47.1	43.5	45.4	47.7	36.3	47.2	28.4	81.6





So are we all set? Key Takeaways takeaways

- GQA/MQA Aim: *To reduce the need for storing a large amount of KV cache*
- LLM server can handle more requests, larger batch sizes and increased throughput
- *Cannot significantly reduce the computational load*
- *Quality degradation remains*