Session 4
Speeding up Attention with long contexts

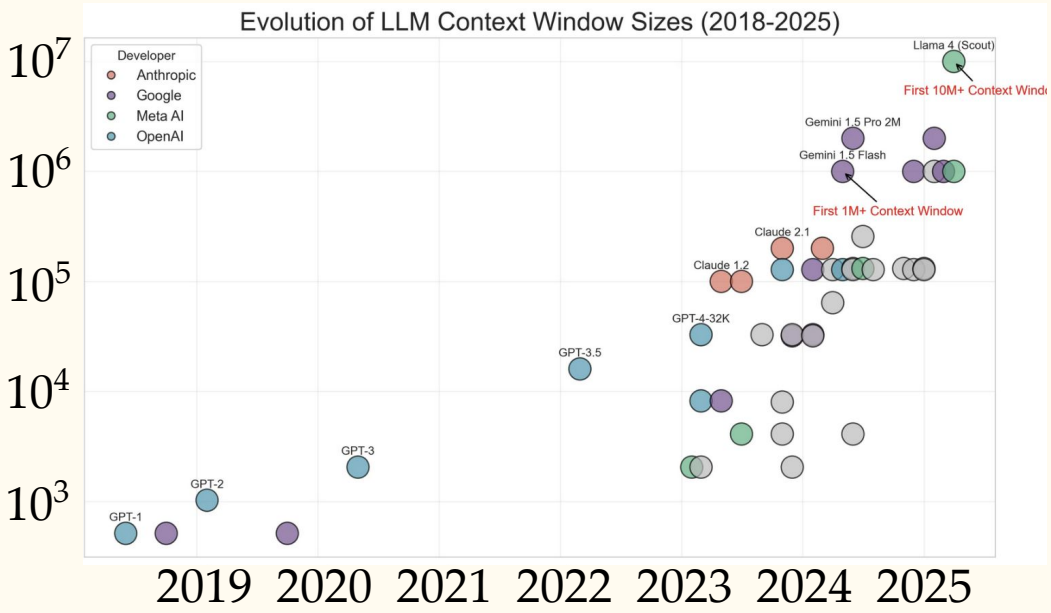
Background: Attention

Attention: $\sigma(QK^\top)V$

$Q\ (n \times d) \quad K\ (n \times d) \quad V\ (n \times d)$

Computation: $O(n^2 d)$ for $S = \sigma(QK)$, $O(n^2 d)$ for SV

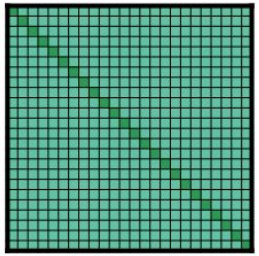
Context Lengths n are Growing



How to improve dependence on n?

Method 1: Sparse Attention

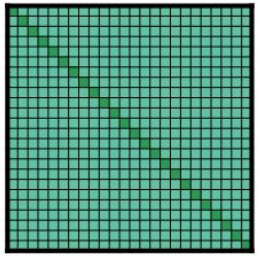
Constant factor improvement via Longformer (Beltagy, Peters, Cohan; 2020)



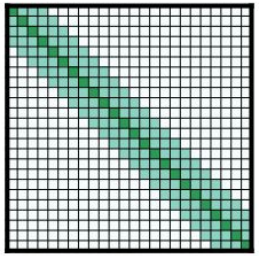
(a) Full n^2 attention

Method 1: Sparse Attention

Constant factor improvement via Longformer (Beltagy, Peters, Cohan; 2020)



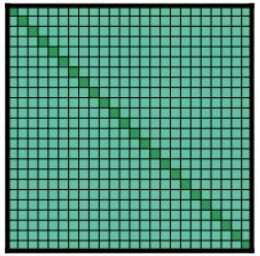
(a) Full n^2 attention



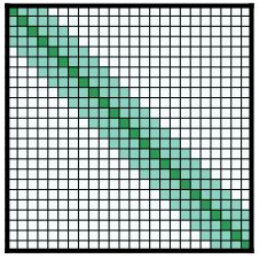
(b) Sliding window attention

Method 1: Sparse Attention

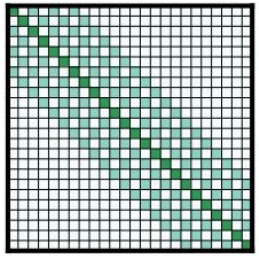
Constant factor improvement via Longformer (Beltagy, Peters, Cohan; 2020)



(a) Full n^2 attention



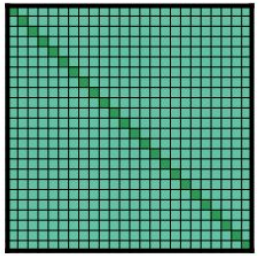
(b) Sliding window attention



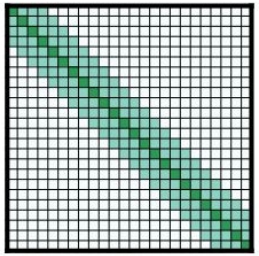
(c) Dilated sliding window

Method 1: Sparse Attention

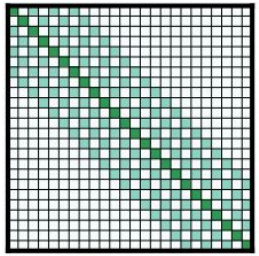
Constant factor improvement via Longformer (Beltagy, Peters, Cohan; 2020)



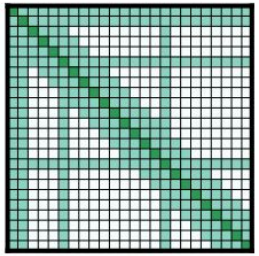
(a) Full n^2 attention



(b) Sliding window attention

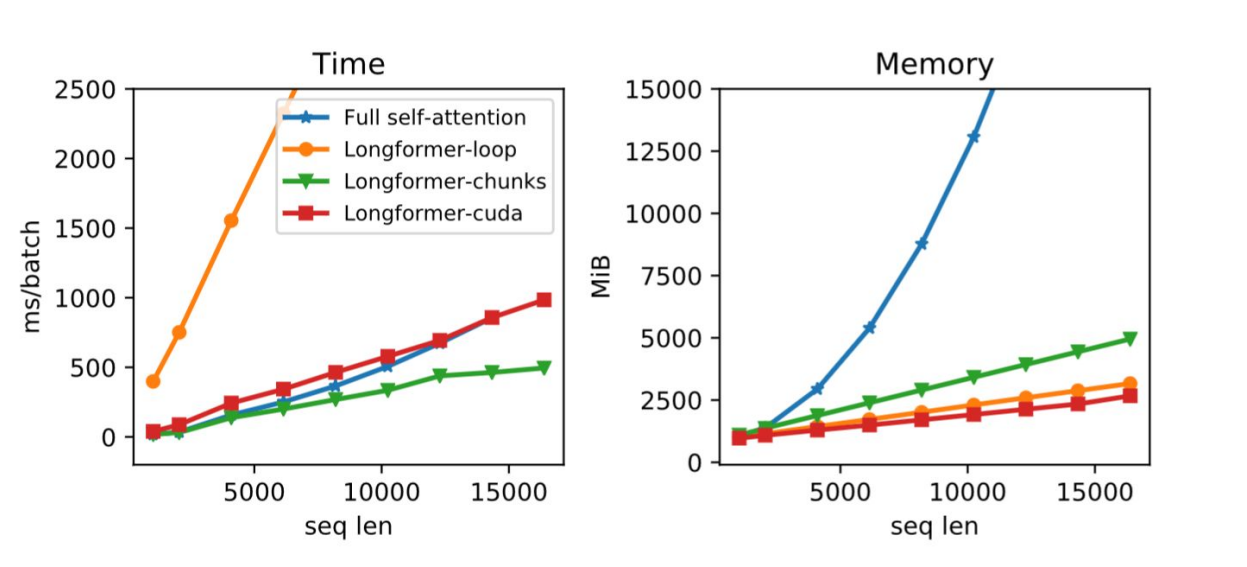


(c) Dilated sliding window



(d) Global+sliding window

Compute Performance of Longformer

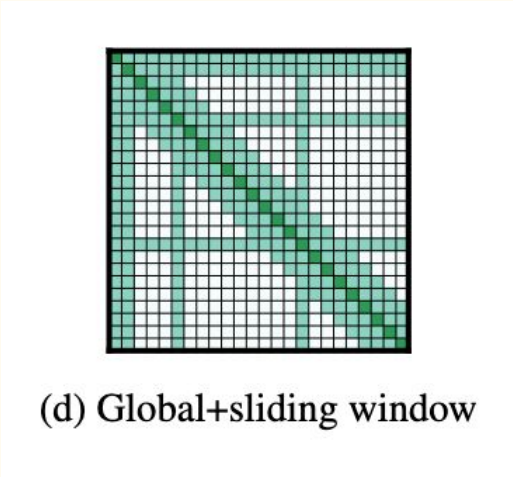


*Practical Performance of Longformer

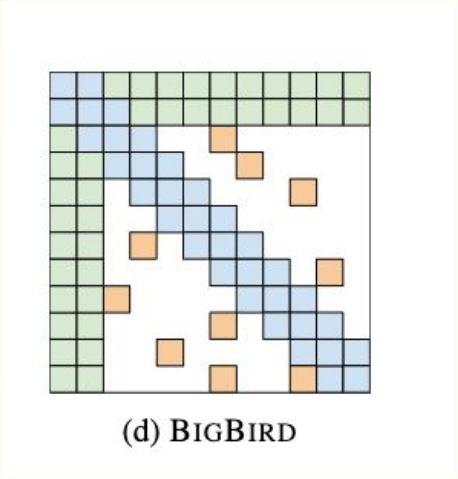
Model	MMLU-pro score	Ratio of linear vs full attention
Mistral-7B	60.1	1 : 0 ratio of sliding-window : full attention
Gemma-2	75.2	1 : 1 ratio of sliding-window : full attention
Gemma-3	78.6	5 : 1 ratio of sliding-window : full attention

Many variations, e.g., Sparse Attention

BigBird = Longformer + random sparseattention (Zaheer et al.; NeurIPS'20)



(d) Global+sliding window



(d) BIGBIRD

Method 2: Linear Attention

		Compute
Full-Attention	$\sigma(QK^\top)V$	$O(n^2d)$
Linear Attention	$(QK^\top)V$	$O(n^2d)$
	$Q(K^\top V)$	$O(nd^2)$

Models using linear attention

Model	MMLUpro	Ratio of linear vs full attention
MiniMax-M1	81.1	7 : 1 ratio of linear : full attention
Qwen3-Next-80B	80.6	3 : 1 ratio of linear : full attention

Historical vignette: Relation to RNNs

In linear attention, $\mathrm{y}_t = q_t^\top \sum_{s \le t} k_s v_s^\top$

Equivalently, $y_t = q_t^\top S_t$ for $S_t = \sum_{s \le t} k_s v_s^\top$

Recurrent form: $S_t = S_{t-1} + k_t v_t^\top$

Generalizations of Linear Attention

Linear attention: $S_t = S_{t-1} + k_t v_t^\top$

Generalizations of Linear Attention

Linear attention: $S_t = S_{t-1} + k_t v_t^\top$

Decayed linear attention $S_t = \gamma_t S_{t-1} + k_t v_t^\top$ $\boxed{\gamma_t = f_\theta(x_t)}$

Generalizations of Linear Attention

Linear attention: $S_t = S_{t-1} + k_t v_t^\top$

Decayed linear attention $S_t = \gamma_t S_{t-1} + k_t v_t^\top$

Gated DeltaNet $S_t = \gamma_t (I - \beta_t k_t k_t^\top) S_{t-1} + \beta_t k_t v_t^\top$

Generalizations of Linear Attention

Linear attention: $S_t = S_{t-1} + k_t v_t^\top$

Decayed linear attention $S_t = \gamma_t S_{t-1} + k_t v_t^\top$

Gated DeltaNet $S_t = \gamma_t (I - \beta_t k_t k_t^\top) S_{t-1} + \beta_t k_t v_t^\top$

“Mamba-2” $S_t = \gamma_t S_{t-1} + k_t v_t^\top \qquad y_t = q_t^\top S_t + D v_t$