UNIVERSITY of HOUSTON

CULLEN COLLEGE of ENGINEERING

DOTA: Detect and Omit Weak Attentions for Scalable Transformer Acceleration

Zheng Qu*
zhengqu@ucsb.edu
UC Santa Barbara
United States

Liu Liu*
liu_liu@ucsb.edu
UC Santa Barbara
United States

Fengbin Tu
fengbintu@ucsb.edu
UC Santa Barbara
United States

Zhaodong Chen
chenzd15thu@ucsb.edu
UC Santa Barbara
United States

Yufei Ding
yufeiding@cs.ucsb.edu
UC Santa Barbara
United States

Yuan Xie
yuanxie@ece.ucsb.edu
UC Santa Barbara
United States

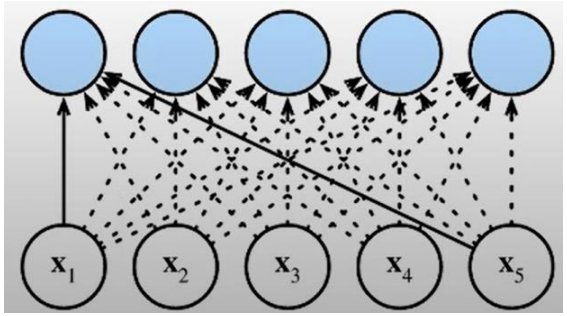
Presented By: Alan Devkota

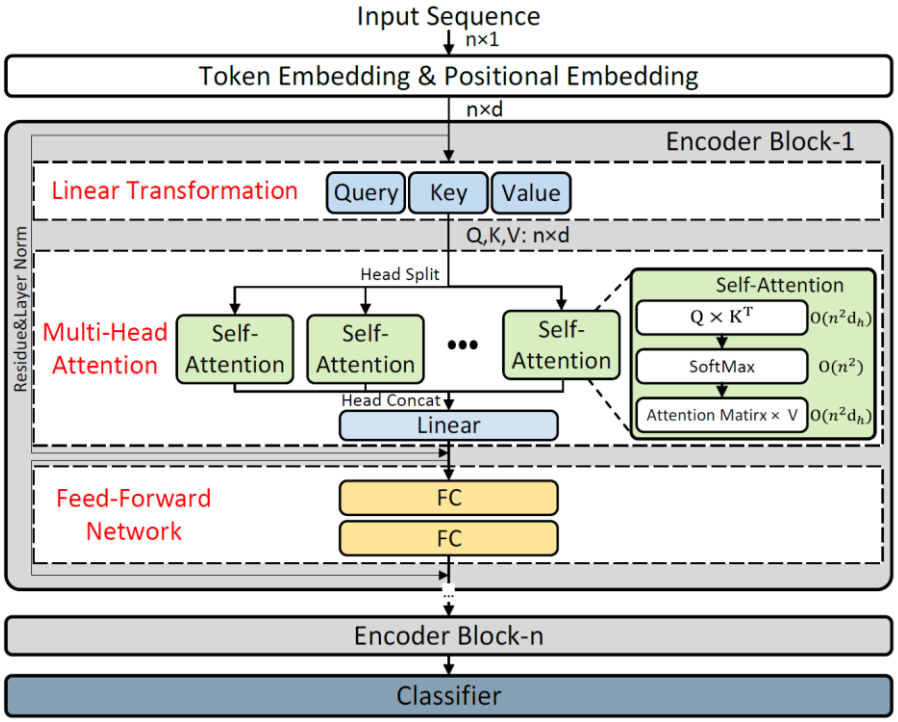
ECE 601

Feb 15, 2023

Department of ECE, University of Houston
Houston, TX, USA

Transformer Neural Network





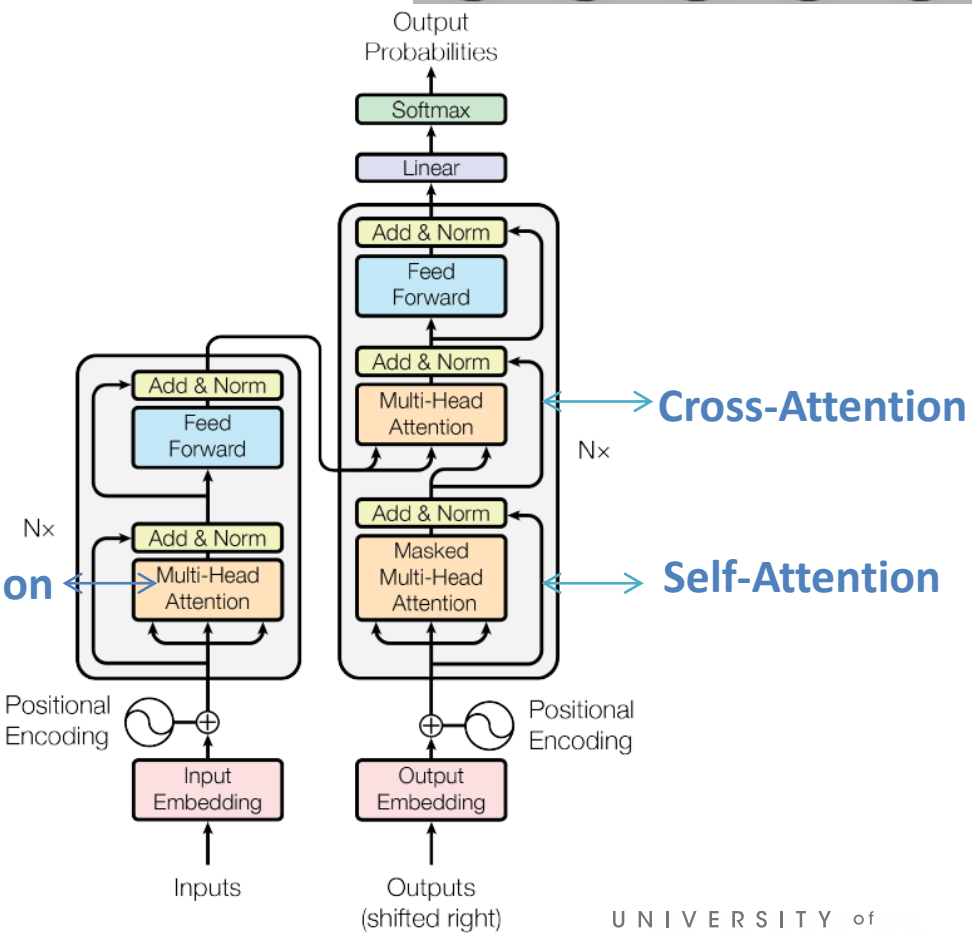


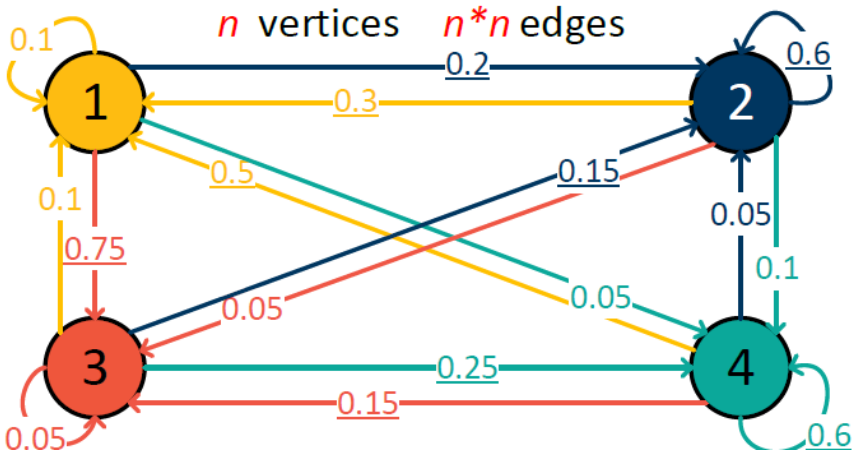
Figure 1: The Transformer - model architecture.

- **Model:** Stack of encoder/ decoder blocks
- **Usually 3-Stage processing** procedure: Linear transformation, Multi-head Attention and Feed Forward
- Key Structure: **Self-attention**

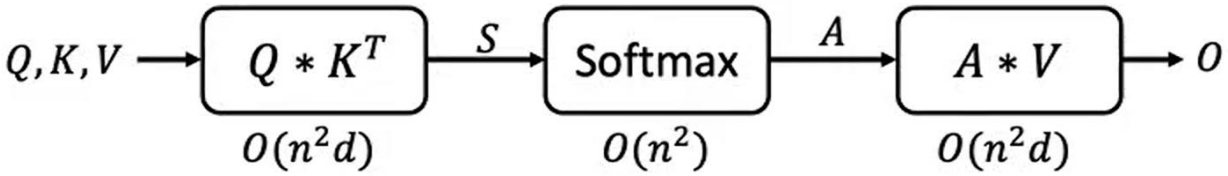
UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Self-Attention Mechanism

- Self-Attention can be understood from a **graph** perspective.
- Each token in the sequence could be regarded as a **vertex**.
- Each directed **edge** represents the attention connection from vertex a to b, and edge **weight** represents attention weight
- We can update each vertex's feature by aggregating the information from its **neighbors**.



I like ECE Seminar

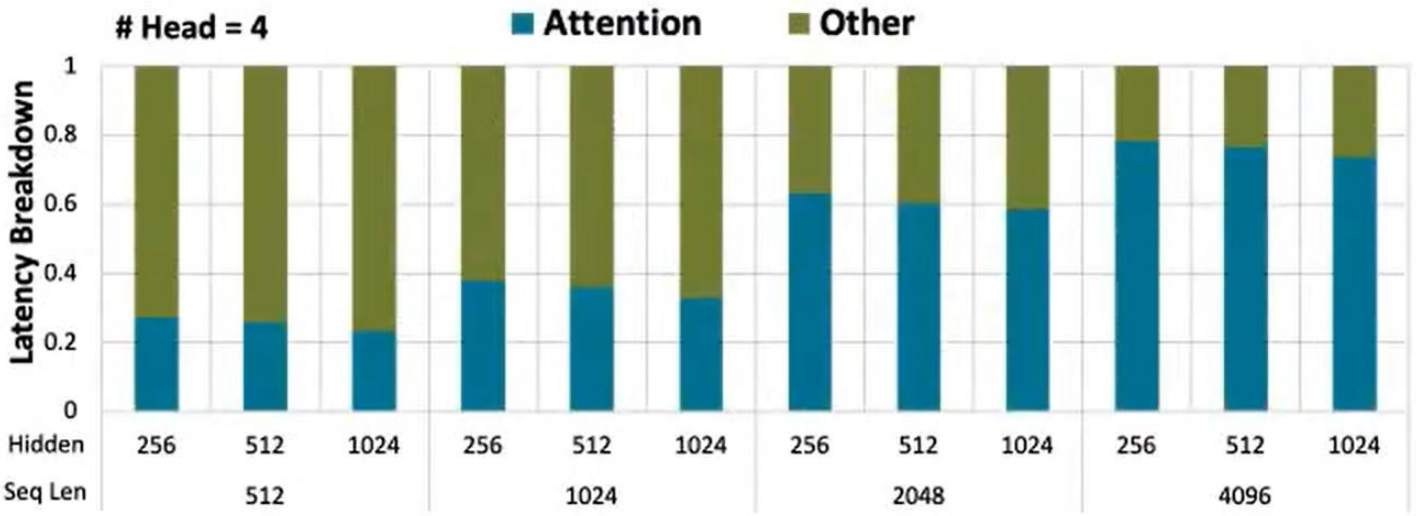


UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Motivation

Self-Attention Cost

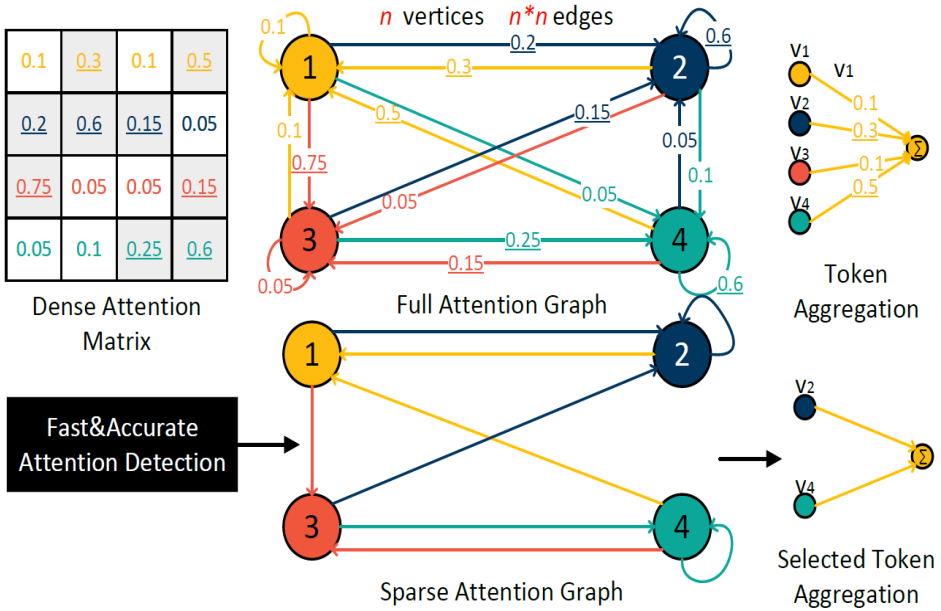
- Quadratically scaled complexity of self-attention → Bottleneck of executing long sequences
- >73% of total execution time with a sequence length of 4096, 8.65GB peak memory usage under 32-bit floating point operation
- Therefore, it is critical to reduce the cost of self-attention blocks.



UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Motivation

Sparsity



- Repetition of the value “0” in weight or activations allows elimination of unnecessary computations.
- Although GPUs are highly optimized in non-sparse dense matrix multiplication, they may not be efficient in reaping the full benefits of sparse matrix multiplication
- Different layers of a neural network exhibit sparsity in unique ways, which necessitates specialized hardware due to variations in the computation nature.

I like ECE Seminar

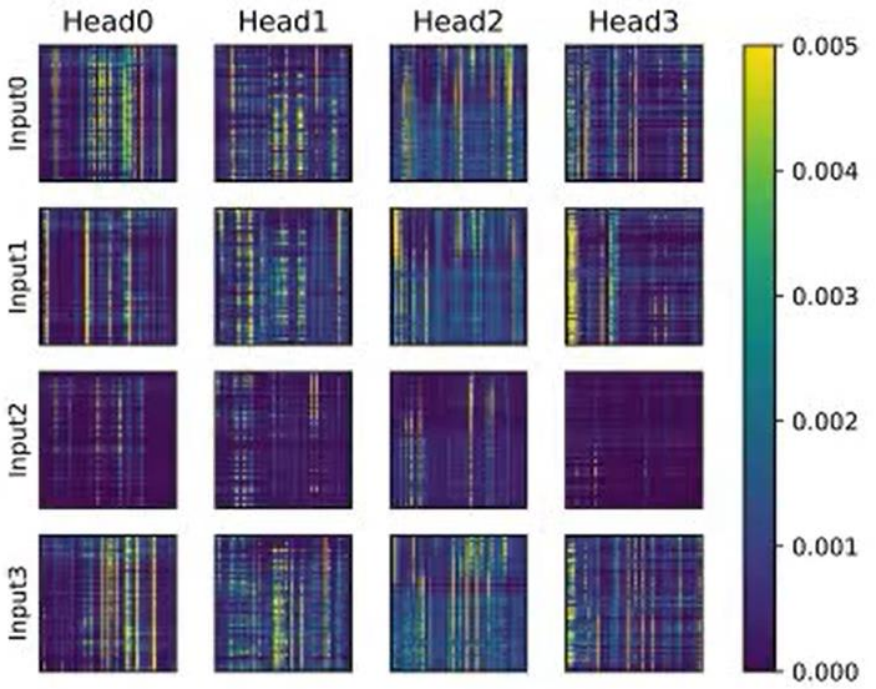
Self-Attention have many **weak** connections that contributes very little to the final output of the feature aggregation!!

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Opportunity: Dynamic Sparsity in Attention Graphs

- Only very few attentions have weight value larger than 0.005
- Model can achieve **on-par accuracy** while running on **pruned** sparse attention graph, which much **reduced memory and computation**

Vanilla Transformer
SQuAD v1.0



Attention
Weight Matrix

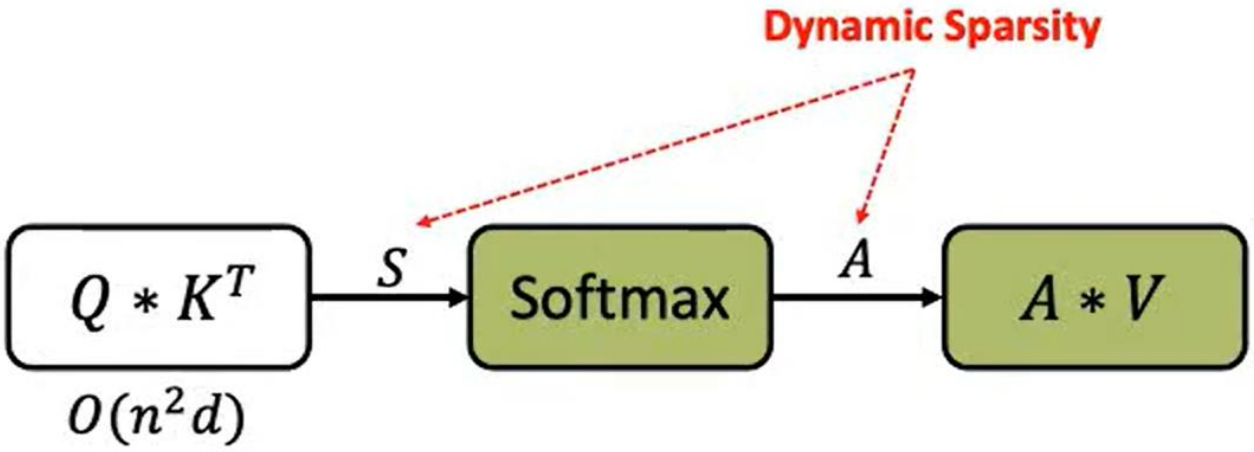
Self-Attention have many **weak** connections that contributes very little to the final output of the feature aggregation!!

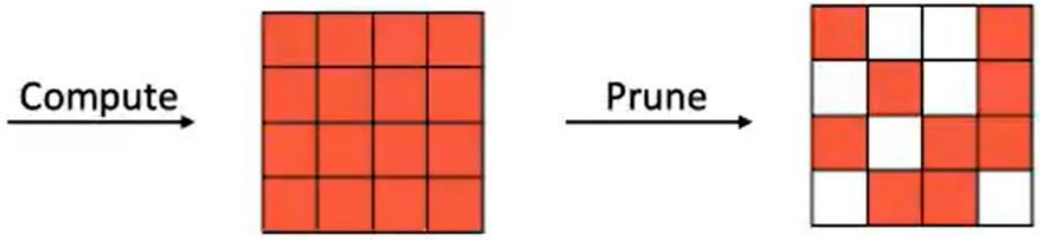
UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Challenges

- How to locate weak attention connection?

(Compute A, Take A as a reference → compare and select only important ones.)



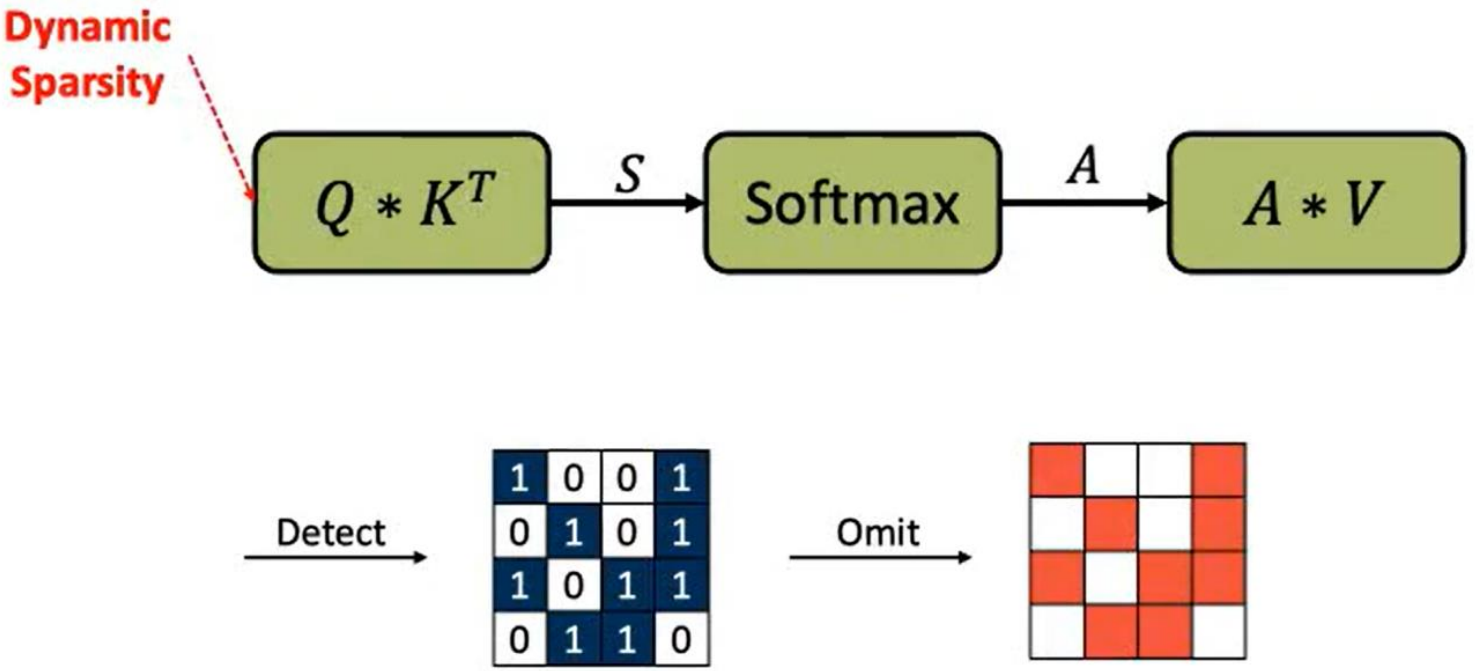


Requires computing the complete attention matrix!!

Challenges

- How about introducing the sparsity before ***Q * K'*** to obtain most computation/ memory reduction?

(Detect and Omit)

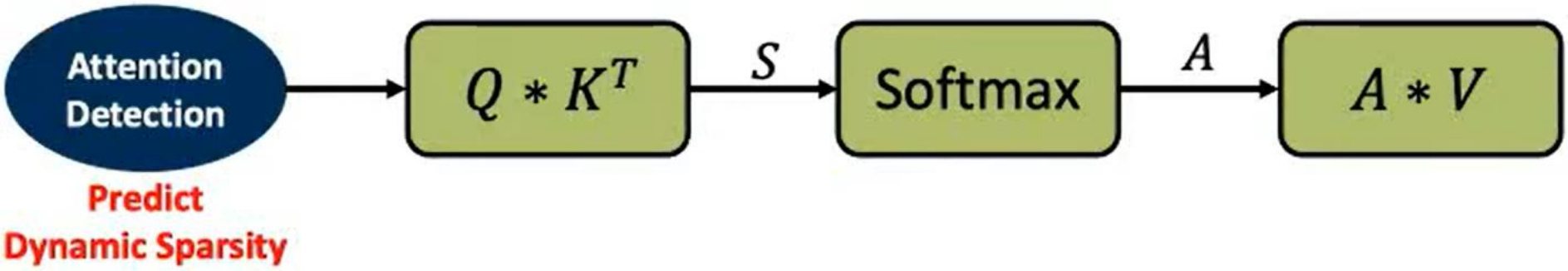


Challenges

- How about introducing the sparsity before ***Q * K'*** to obtain most computation/ memory reduction?

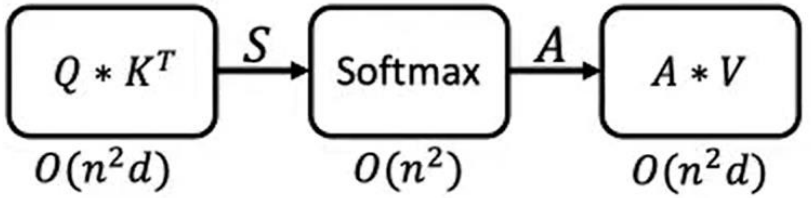
(Detect and Omit)

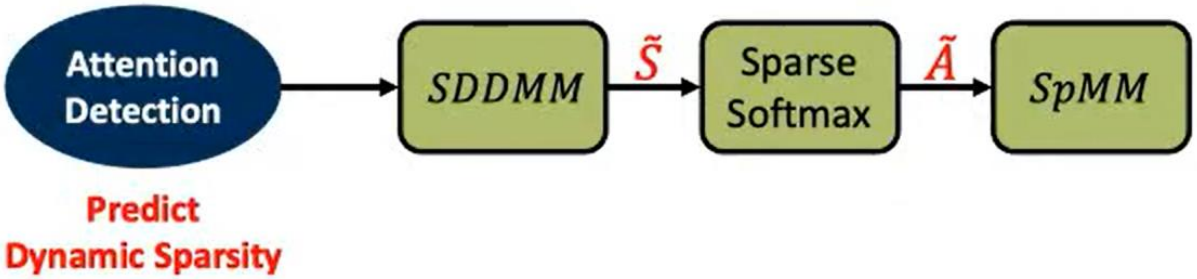
It requires some detection method!!
Should be light-weight, accurate and hardware-efficient (trade-off)



Hardware Challenges

- Should support both detection and execution
  - Random projection
  - Quantization/ dequantization
  - Multi-precision computation
- Should support sparse attention computation
  - SpMM
  - SDDMM
  - Sparse Softmax

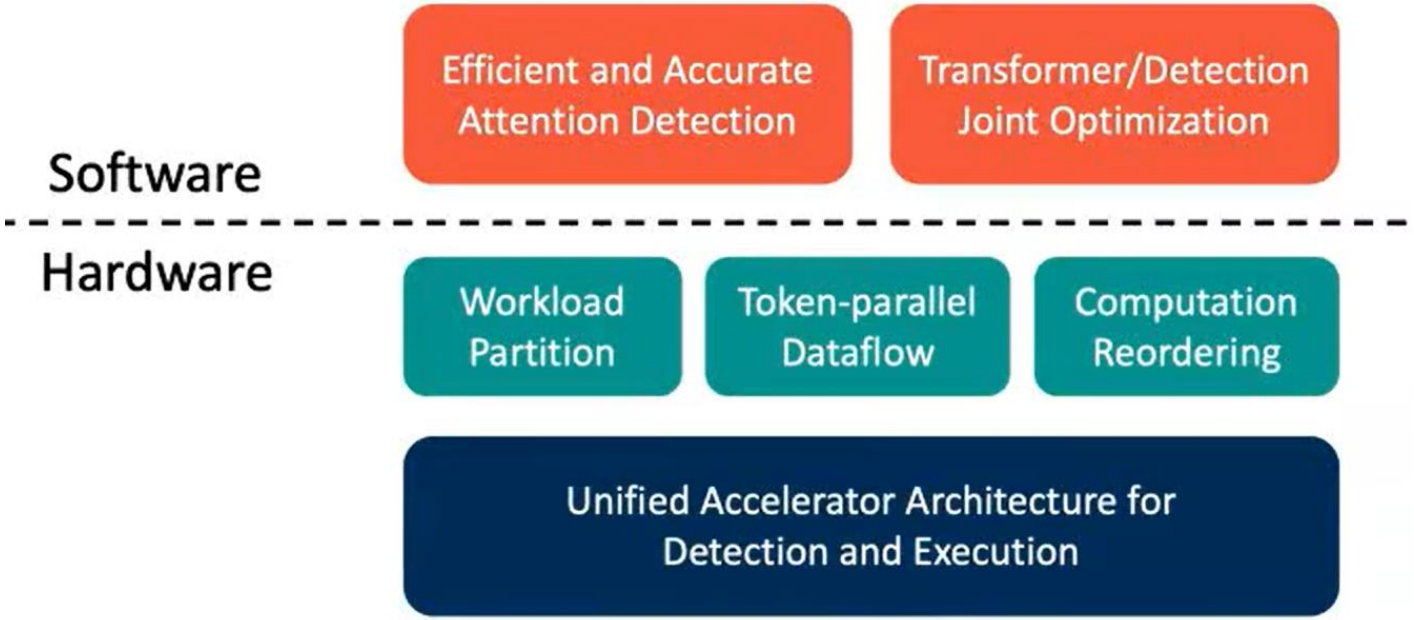




However, some of the sparse operators are difficult to accelerate on GPU, so they have proposed customized architecture.

Solution: DOTA

- Proposed top-down **Software-hardware** co design
  - **On software side** → Lightweight detection n/w along with optimization of transformer model
  - **Hardware side** → Different accelerator (on top of it workload partition, token-parallel Dataflow, and computation reordering)



UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

DOTA: Dynamic Sparse Attention Algorithm

- Train a lightweight detection n/w to help detect the weak/important connection
  - Low precision and low dimension

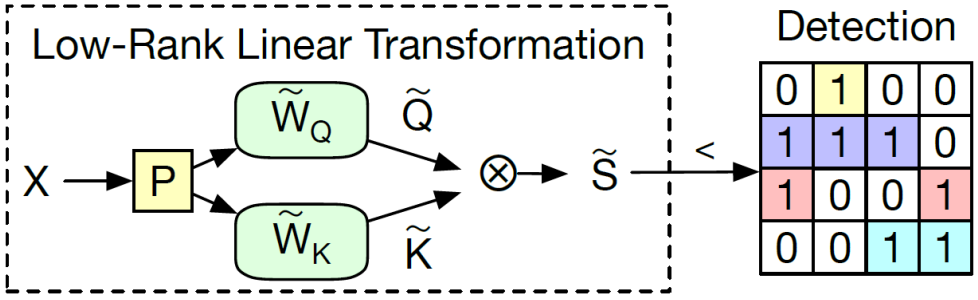


Figure 4: Weak attention detection from estimated attention scores computed by low-rank linear transformations.

Calculate approx. attention weight using expression $\widetilde{S} = XP\widetilde{W}_Q(XP\widetilde{W}_K)^T$

P is sparse random projection and $\widetilde{W}_K$ and $\widetilde{W}_Q$ denote approx. weights for keys and queries

Use $\widetilde{S}$ to preserve or discard the attention scores

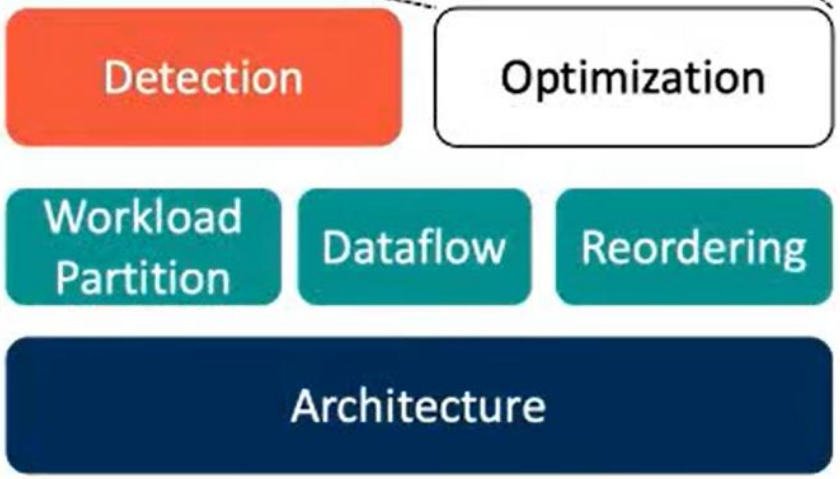
UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

DOTA: Dynamic Sparse Attention Algorithm

- Joint optimization of detector and transformer model to ensure detection accuracy and final model accuracy

$$L_{MSE} = \frac{1}{B}||S - \tilde{S}||_2^2 = \frac{1}{B}||QK^T - \tilde{Q}\tilde{K}^T||_2^2$$

$$L = L_{Model} + \lambda L_{MSE}$$



DOTA Accelerator Architecture

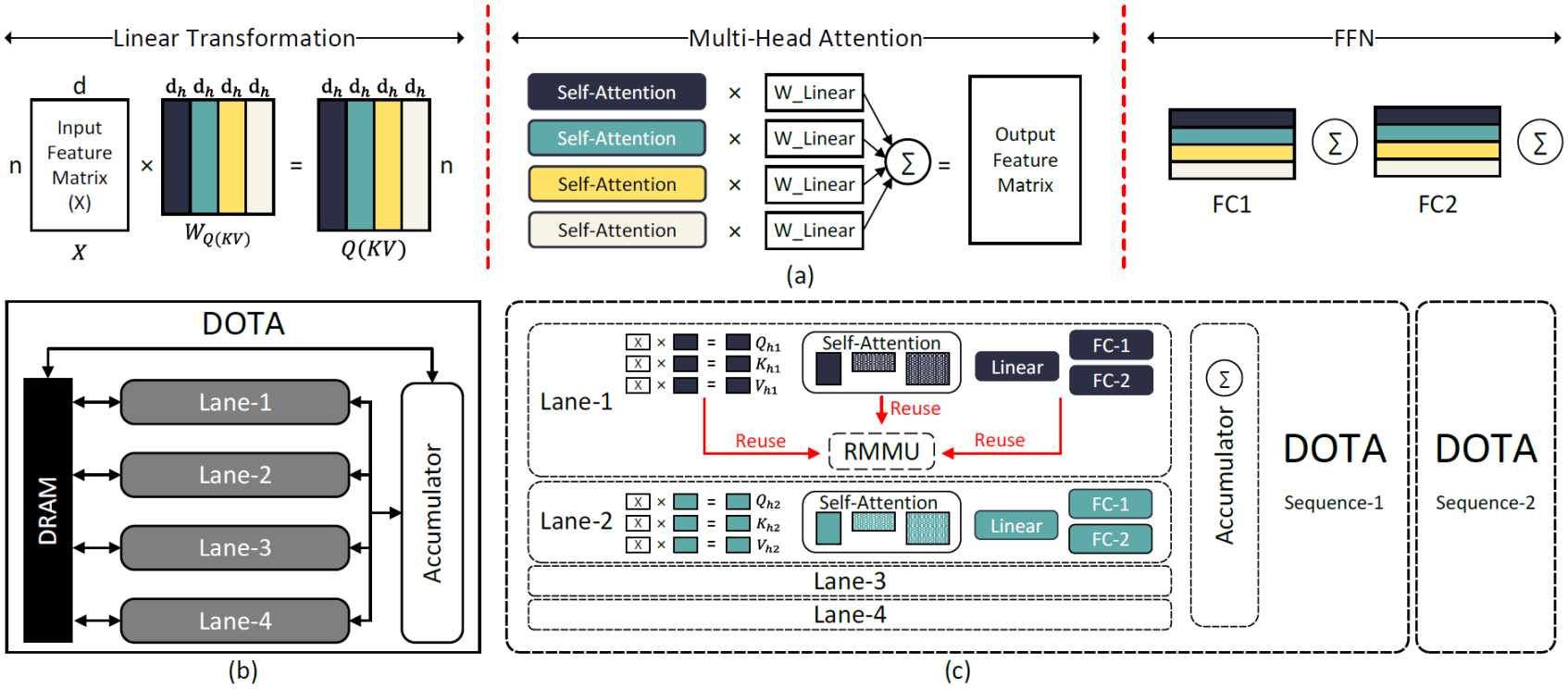


Figure 5: DOTA system design. (a) The abstraction of a single encoder block. We divide each encoder into three sequential stages. Each stage contains multiple GEMM operations that can be further cut into chunks (represented by different colors) and mapped to different compute Lanes. (b) Overall system design of DOTA. Each compute Lane communicates with off-chip DRAM for input feature. The intermediate results are summed up in the Accumulator. (c) Computation mapping between the algorithm and hardware. Each DOTA accelerator processes one input sequence, and each Lane computes for one chunk (color).

Compute Lane

- Reconfigurable Matrix Multiplication Unit – Multi-precision computation
- SRAM Buffer
- Multi-Function Unit – Adder tree for random projection, (De)Quantization Unit, Exp/Div for Softmax function
- Detector – Attention Detection, Hardware-level sparse computation scheduling

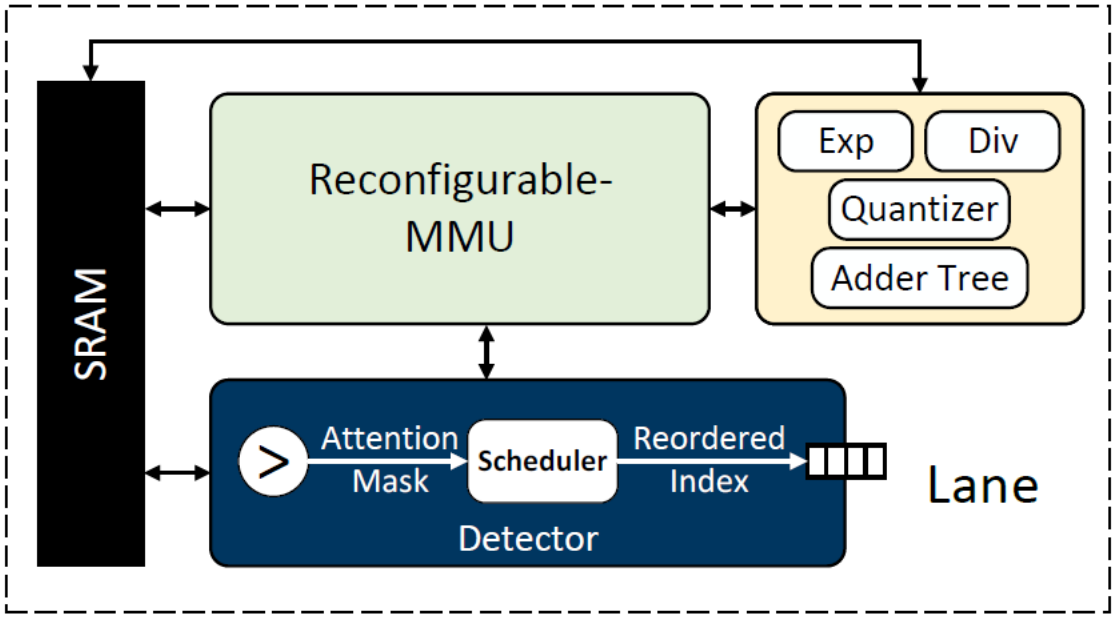
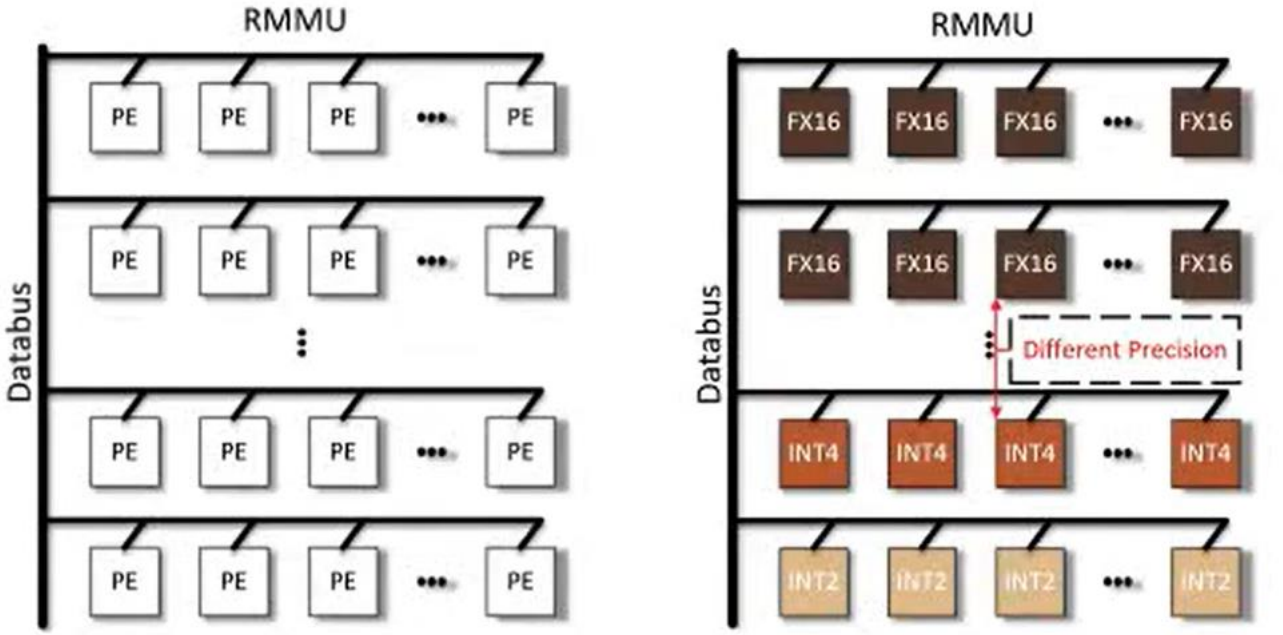


Figure 6: Architecture of each compute Lane.

Reconfigurable MMU

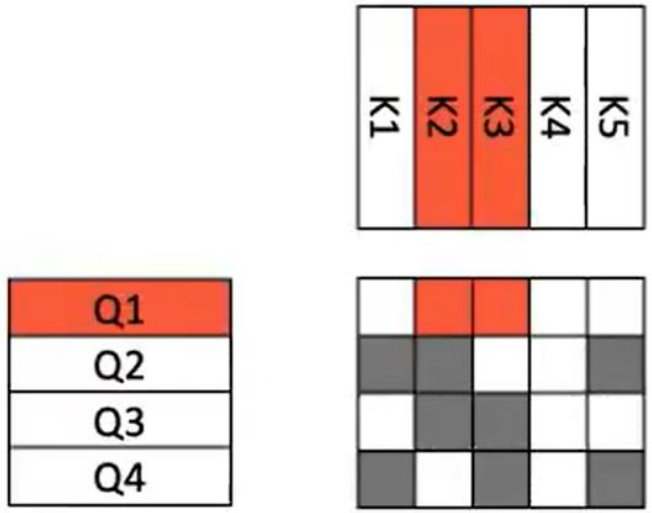
- 2D PE array
- One PE architecture, row-wise reconfigurable precision



Token-Parallel Sparse Attention

- Naïve way – compute the output matrix row by row, from left to right

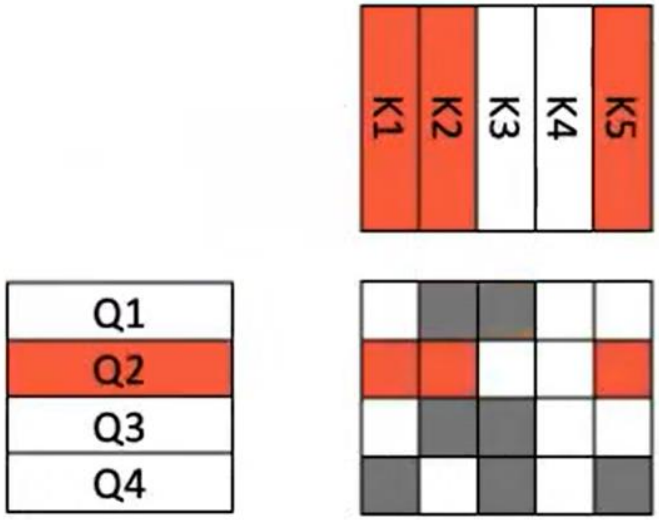
Row-by-Row		
(Q1, K2)	(Q1, K3)	



Token-Parallel Sparse Attention

- Naïve way – compute the output matrix row by row, from left to right

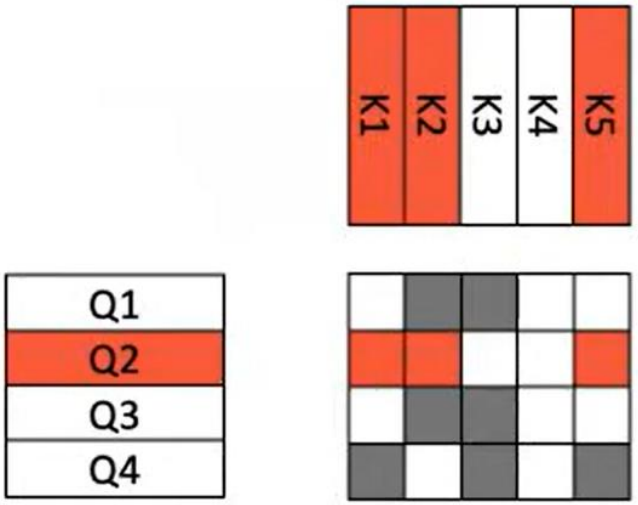
Row-by-Row		
(Q1, K2)	(Q1, K3)	
(Q2, K1)	(Q2, K2)	(Q2, K5)



Token-Parallel Sparse Attention

- Naïve way – compute the output matrix row by row, from left to right

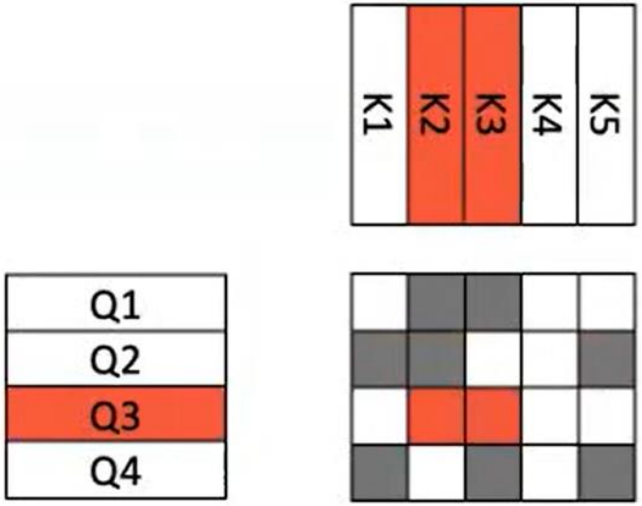
Row-by-Row		
(Q1, K2)	(Q1, K3)	
(Q2, K1)	(Q2, K2)	(Q2, K5)



Token-Parallel Sparse Attention

- Naïve way – compute the output matrix row by row, from left to right

Row-by-Row		
(Q1, K2)	(Q1, K3)	
(Q2, K1)	(Q2, K2)	(Q2, K5)
(Q3, K2)	(Q3, K3)	

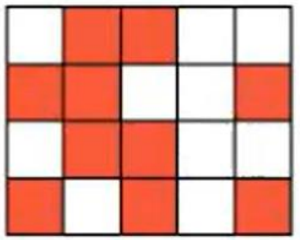


Token-Parallel Sparse Attention

- Naïve way – compute the output matrix row by row, from left to right

Row-by-Row		
(Q1, K2)	(Q1, K3)	
(Q2, K1)	(Q2, K2)	(Q2, K5)
(Q3, K2)	(Q3, K3)	
(Q4, K1)	(Q4, K3)	(Q4, K5)

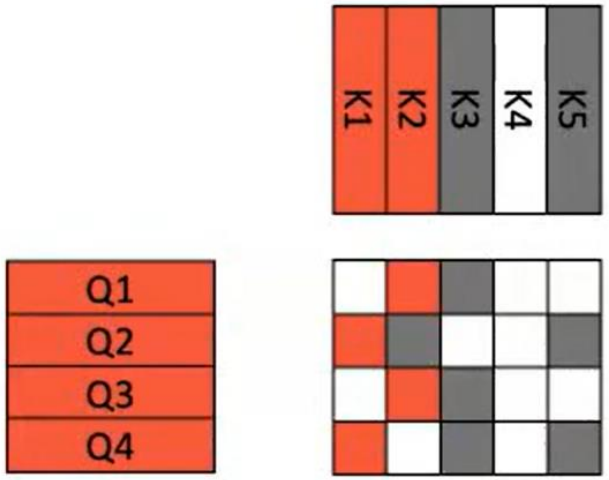
Total Memory Access:
4(Q) + 10(K) = 14 Vecs



Token-Parallel Sparse Attention

- Utilize sparsity, compute output matrix in group of rows, each row from left to right

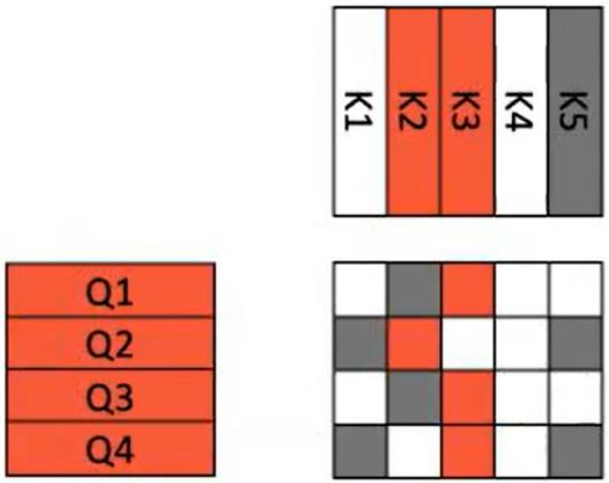
Token Parallel			
(Q1, k2)	(Q2, k1)	(Q3, k2)	(Q4, k1)



Token-Parallel Sparse Attention

- Utilize sparsity, compute output matrix in group of rows, each row from left to right

Token Parallel			
(Q1, k2)	(Q2, k1)	(Q3, k2)	(Q4, k1)
(Q1, k3)	(Q2, k2)	(Q3, k3)	(Q4, k3)

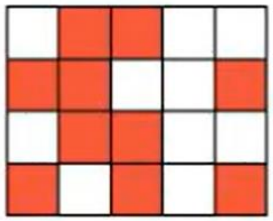


Token-Parallel Sparse Attention

- Utilize sparsity, compute output matrix in group of rows, each row from left to right

Token Parallel			
(Q1, k2)	(Q2, k1)	(Q3, k2)	(Q4, k1)
(Q1, k3)	(Q2, k2)	(Q3, k3)	(Q4, k3)
	(Q2, k5)		(Q4, k5)

Total Memory Access:
4(Q) + 5(K) = 9 Vecs < 14Vecs



Hardware Configuration

Table 2: Configurations, Power, and Area of DOTA under 22nm Technology and 1GHz Frequency.

Hardware Module		Configuration	Power(mW)	Area(mm^2)
Lane		4 Lanes per accelerator	2878.33	2.701
Lane	RMMU	32*16 FX-16	645.98	0.609
	Filter	Token Paral. = 4	9.13	0.003
	MFU	16 Exp, 16 Div 16*16 Adder Tree	60.73	0.060
Accumulator		512 accu/cycle	139.21	0.045
DOTA (w/o SRAM)		2TOPS	3017.54	2.746
SRAM		2.5MB	0.51(Leakage)	1.690

Evaluation: Model Accuracy

- Comparable accuracy with dense models under 90-95% sparsity
- Much better accuracy-sparsity trade-off than prior art (ELSA)

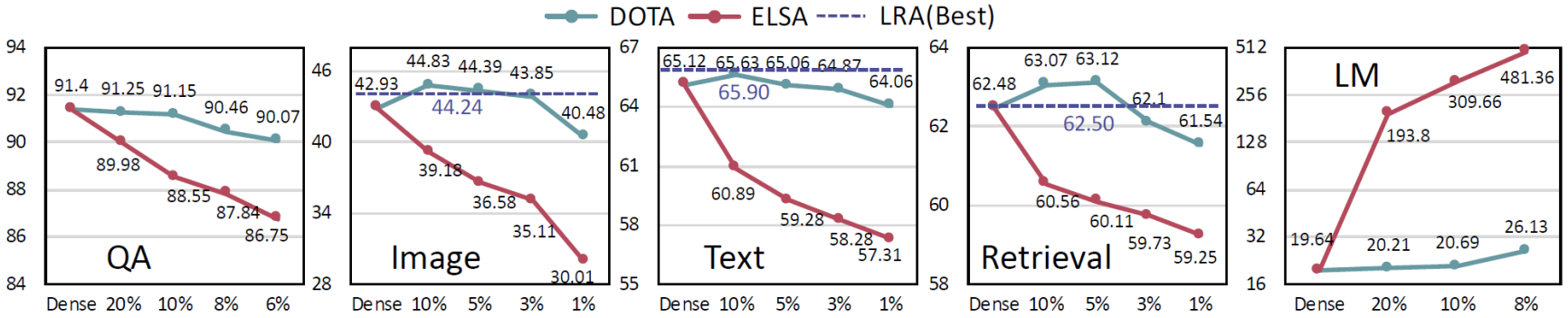


Figure 11: Model accuracy of DOTA comparing with dense baseline and ELSA under different retention ratios across the benchmarks. The performance metric of GPT-2 is perplexity score, the lower the better. The other dataset uses accuracy, the higher the better. The purple line indicates the best results provided by the LRA benchmark.

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Evaluation: Speedup and Latency Breakdown

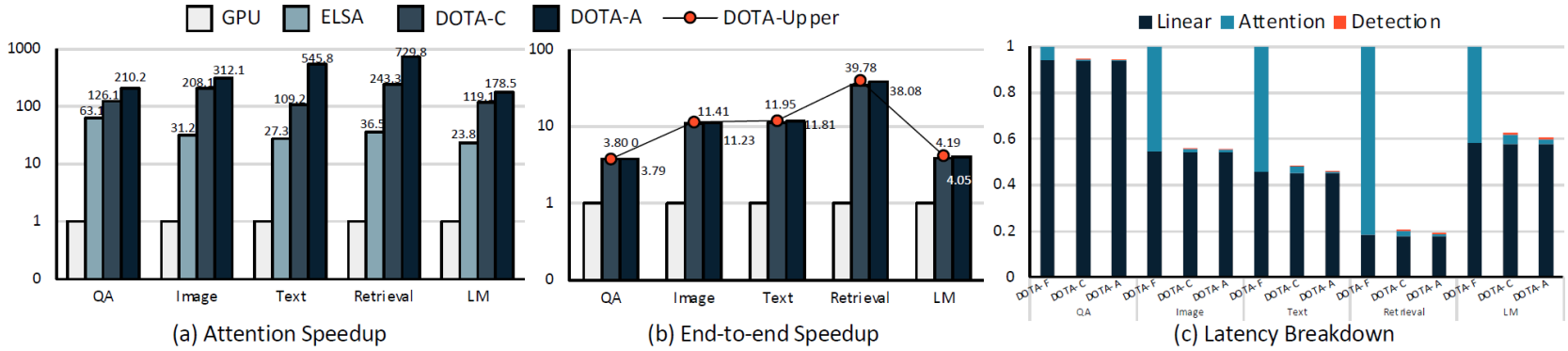


Figure 12: (a) Speedup of DOTA over GPU and ELSA on attention block. (b) End-to-end speedup over GPU. Red dots indicate the theoretical performance upper-bound of an accelerator. (c) Normalized latency breakdown of DOTA. DOTA-F means to compute the *Full* attention graph with DOTA without detection and omission. DOTA-C (Conservative) and DOTA-A (Aggressive) both adopt attention detection, while DOTA-C allows for an accuracy degradation less than 0.5% and DOTA-A allows for 1.5%.

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Evaluation: Energy Efficiency

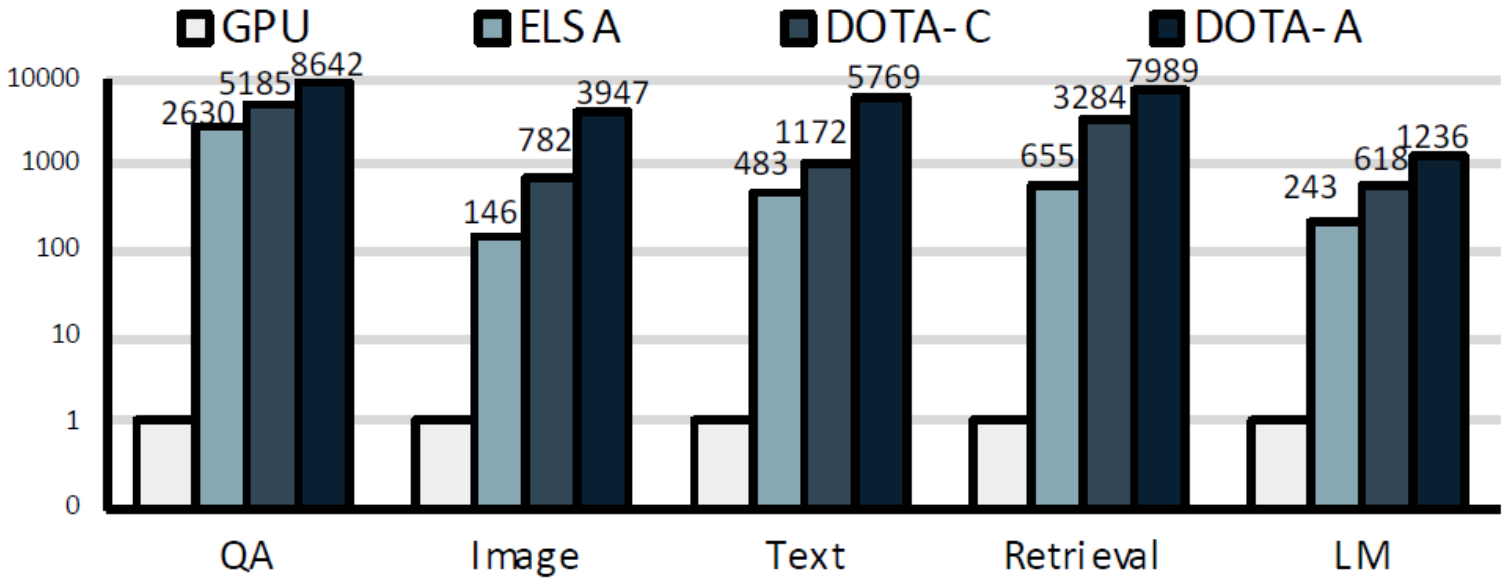


Figure 13: Energy-efficiency comparisons.

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Conclusion

- Proposed way to leverage **weak attention connection** to reduce cost of self-attention mechanism
- Light weight detection network and joint optimization
- Unified hardware-software co-design
- Speedup, energy-efficient with negligible accuracy degradation

Thank you for your attention!

Questions ??

adevkota2@uh.edu

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING