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

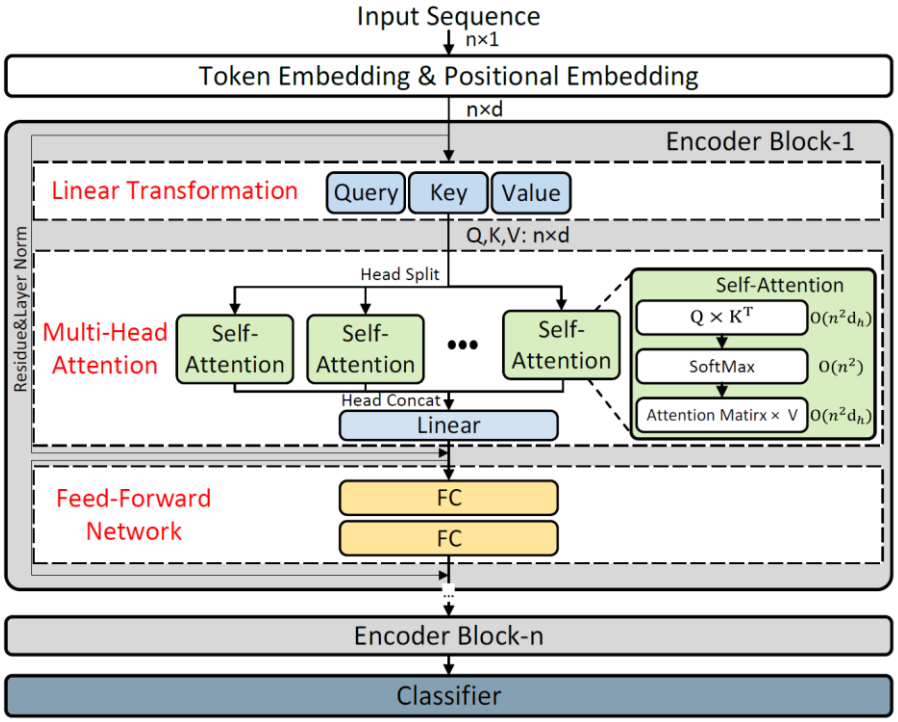
Presented By: Alan Devkota

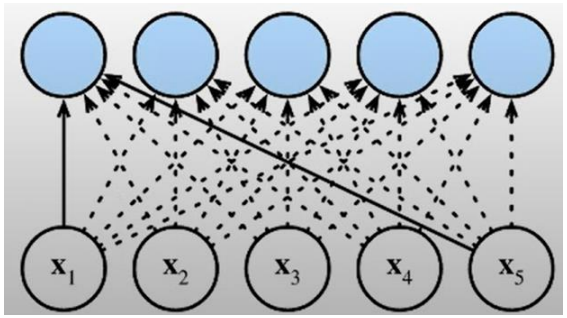
ECE 601

Feb 15, 2023

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

Transformer Neural Network





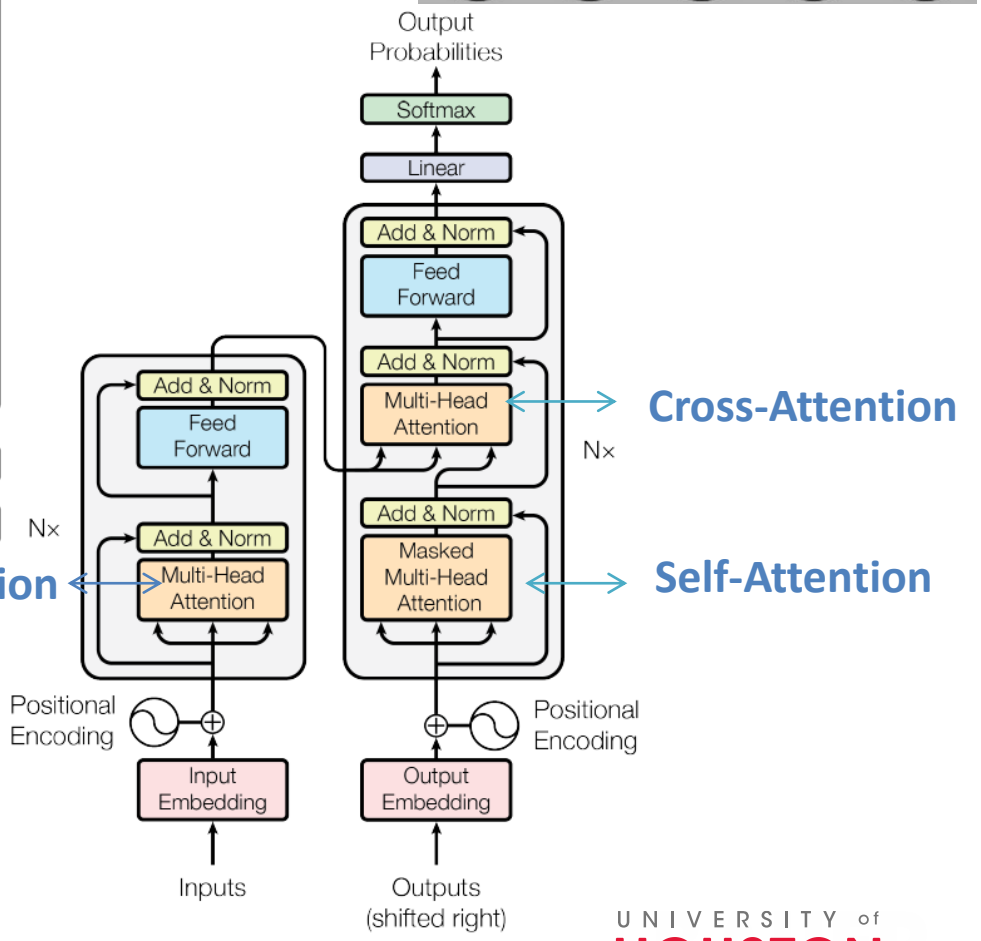


Figure 1: The Transformer - model architecture.

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

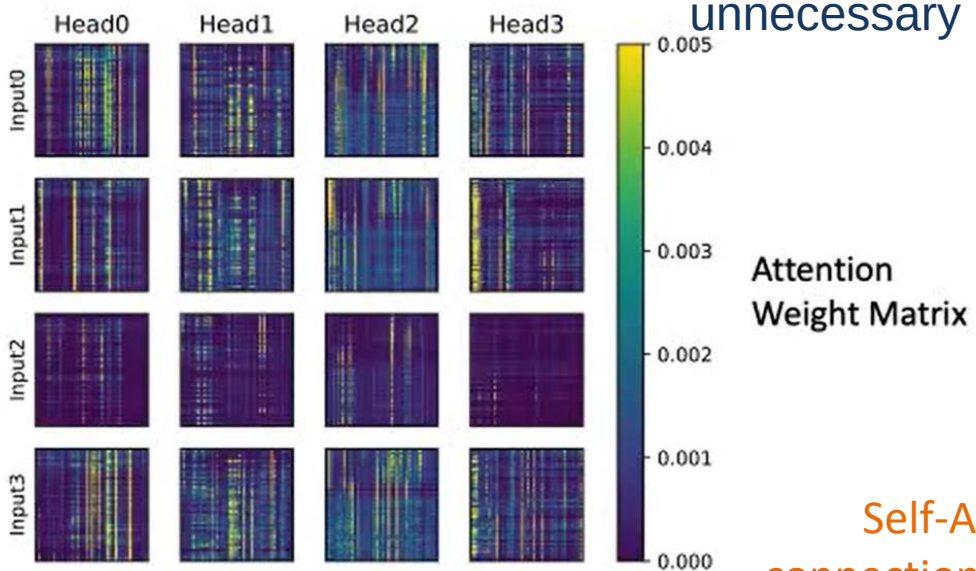
UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Motivation

Dynamic Sparsity

- Repetition of the value “0” in weight or activations allows elimination of unnecessary computations.

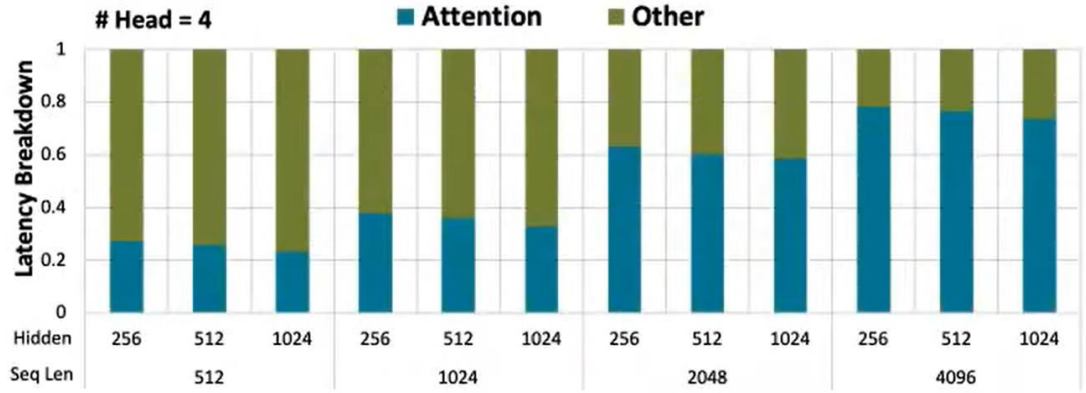
Vanilla Transformer
SQuAD v1.0



Attention
Weight Matrix

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

I like ECE Seminar



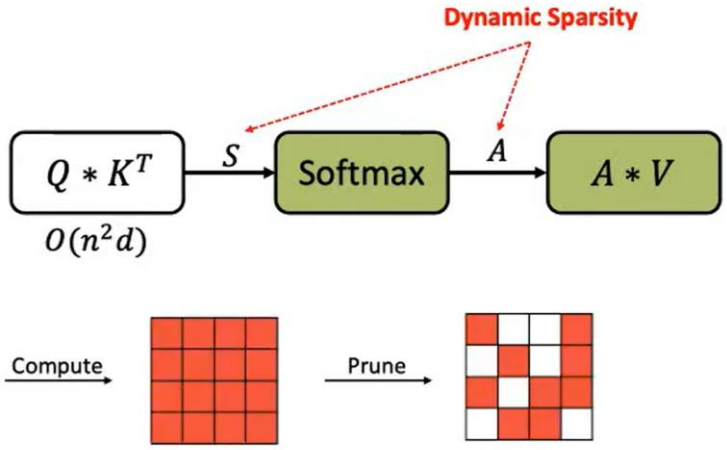
UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Challenges

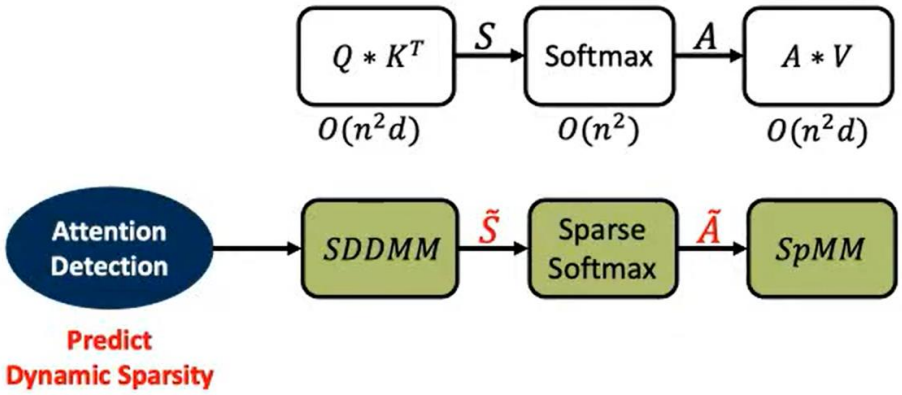
- How to locate weak attention connection?

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

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



Compute and Prune



Solution : (Detect and Omit)

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

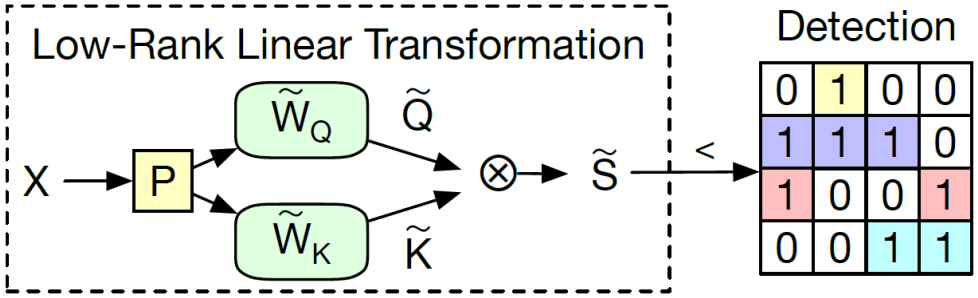
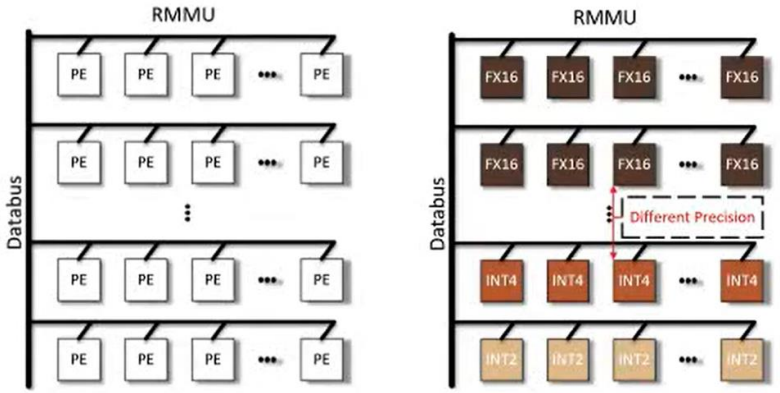


Figure 4: Weak attention detection from estimated attention scores computed by low-rank linear transformations.



Reconfigurable MMU

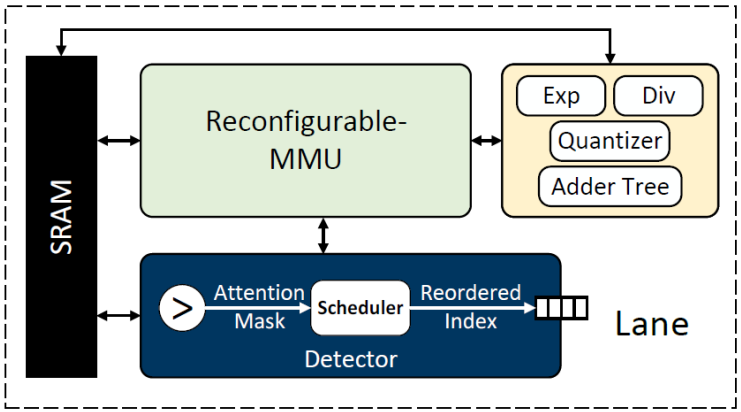


Figure 6: Architecture of each compute Lane.

DOTA Accelerator Architecture

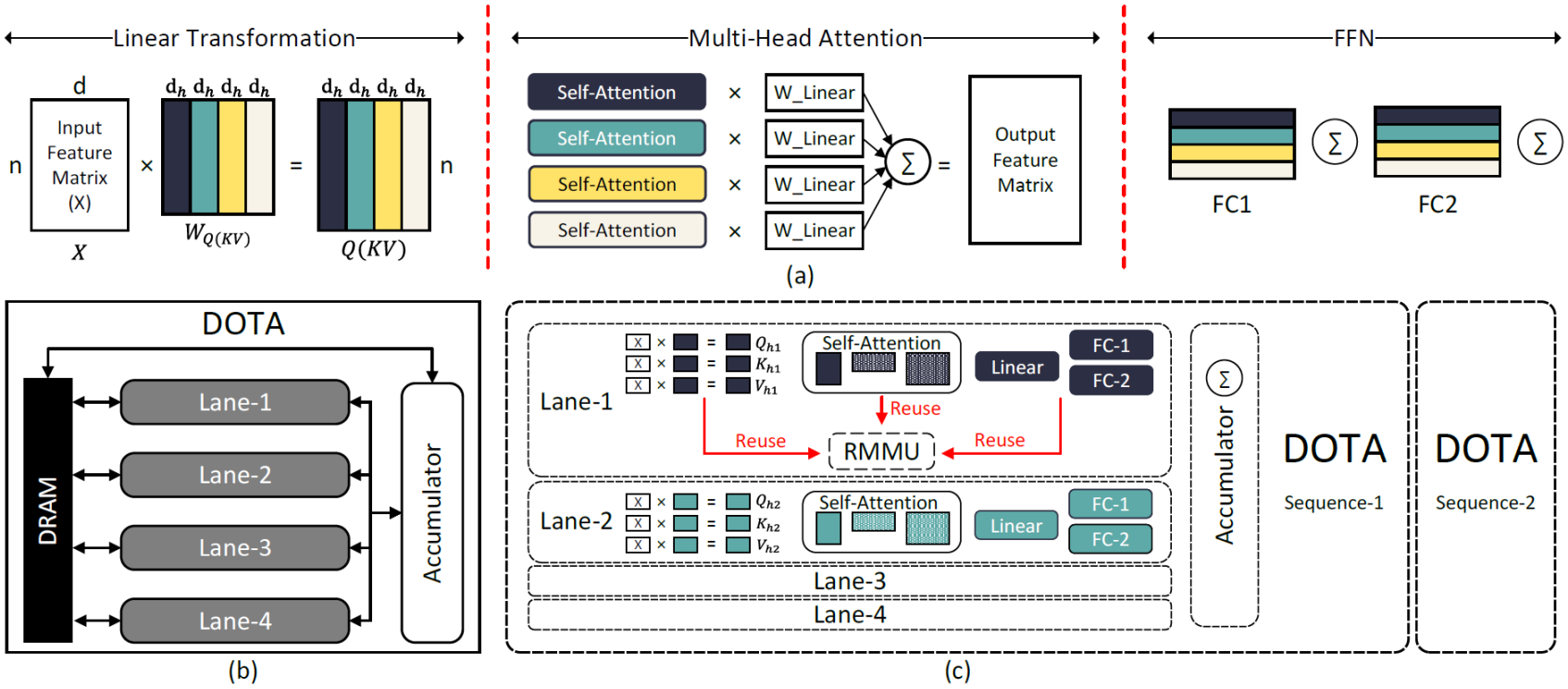


Figure 5: DOTA system design. (a) The abstraction of a single encoder block. We divide each encoder into three sequential stages. Each stage contains multiple GEMM operations that can be further cut into chunks (represented by different colors) and mapped to different compute Lanes. (b) Overall system 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).

Evaluation: Model Accuracy

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

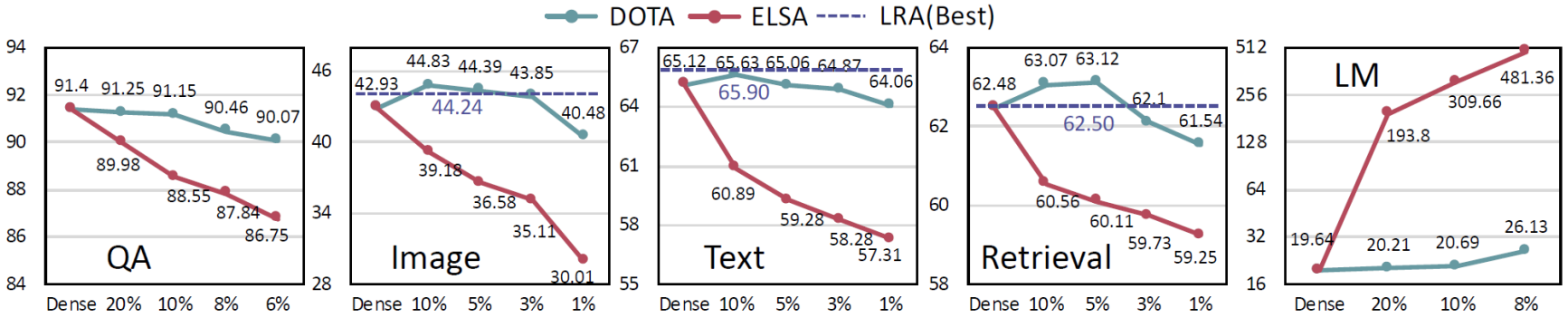


Figure 11: Model accuracy of DOTA comparing with dense baseline and ELSA under different retention ratios across the benchmarks. The performance metric of GPT-2 is perplexity score, the lower the better. The other dataset uses accuracy, the higher the better. The purple line indicates the best results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

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Evaluation

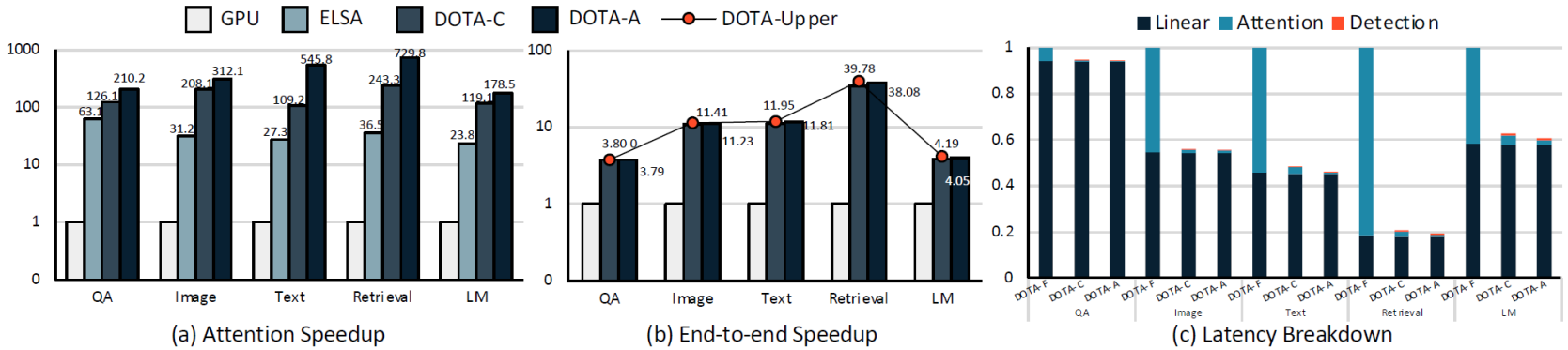


Figure 12: (a) Speedup of DOTA over GPU and ELSA on attention block. (b) End-to-end speedup over GPU. Red dots indicate the theoretical performance upper-bound of an accelerator. (c) Normalized latency breakdown of DOTA. DOTA-F means to compute the *Full* attention graph with DOTA without detection and 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%.

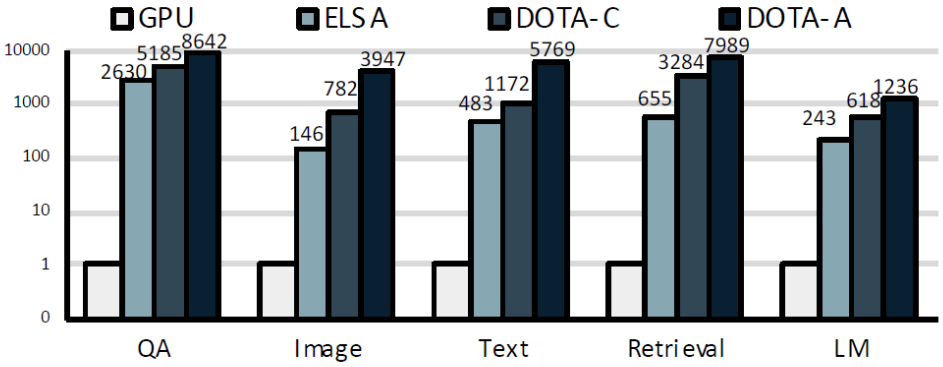


Figure 13: Energy-efficiency comparisons.

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