Efficient and Practical Neural Question Answering for Heterogeneous Platforms

Qingqing Cao, joint work with Noah Weber, Harsh Trivedi, Yash Lal, Niranjan Balasubramanian and Aruna Balasubramanian

DeQA: On-Device Question Answering (MobiSys 2019)

QA models don’t fit into a smartphone memory and run extremely slow!

- Document is much longer than the question
- Document processing is the bottleneck in all layers

DeFormer: Decomposing Transformers for Faster Question Answering (ACL 2020)

Transformer for QA

Layer n

Layer 2

Layer 1

Run 4x faster; Uses 70% less memory

Minimal accuracy loss (~1% F1 score)

Decomposed Transformer

Accurate and Interpretable Energy Modeling for Transformers (under review)

Existing software methods measure the total energy by combining energy of hardware components based on utilization

$$e_{total} = \sum_p (u_{dram} e_{dram} + u_{cpu} e_{cpu} + u_{gpu} e_{gpu})$$

(Strubell et al., ACL 2019; Henderson et al., JMLR 2020)

Non-utilization behaviors such as data movement is the major energy bottleneck (>50%)

(Eyeriss, ISSCC 2016)

Idea: combine runtime resources (include util, freq etc) with model features and feed them into a predictive model to estimate energy

Reduce the QA latency on a smartphone from over a minute to under 5s. Reduce energy by >10x.