Redes Neurais e Aprendizado Profundo

Artur Jordão
Escola Politécnica – Engenharia de Computação e Sistemas Digitais
Universidade de São Paulo

Multi-Query Attention and Grouped-Query Attention

Introduction

Multi-query attention (MQA) & Grouped-Query Attention (GQA)

- Language models are expensive for inference primarily due to the memory bandwidth overhead from loading keys and values

- Multi-query attention (MQA) and Grouped-Query attention (GQA) reduce this overhead
  - Both methods provide a compromise between model capacity/quality and speed-up

- Ainslie et al.[1] propose to convert multi-head attention (MHA – the original Transformer architecture) models to multi-query and grouped-query models

[1] Ainslie et al. *GQA Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints.* EMNLP, 2023

Preliminaries

Multi-query attention (MQA) & Grouped-Query Attention (GQA)

- MQA and GQA employ the Uptraining recipe
- Uptraining involves initializing a model from a pre-trained checkpoint
- Given the checkpoint, the recipe pre-trains for a further α proportion of original pre-training steps
  - Uptraining by Ainslie et al.[1] employs the original pre-training setup and dataset

[1] Ainslie et al. *GQA Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints.* EMNLP, 2023

Multi-Query Attention

Multi-query attention (MQA) & Grouped-Query Attention (GQA)

- Multi-query attention (MQA) employ a single key-value head

- Going from Multi-head attention (MHA) to MQA reduces H key and value heads to a single key and value head
    - It reduces the size of the key-value cache and therefore amount of data that needs to be loaded by a factor of H
    - It drastically speeds up decoder inference

- Unfortunately, MQA can lead to quality degradation and training instability[1]

[1] Ainslie et al. *GQA Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints.* EMNLP, 2023

Multi-Query Attention

Multi-query attention (MQA) & Grouped-Query Attention (GQA)

- Conversion from a multi-head model to a multi-query model takes place in two steps
  - Converting a checkpoint (i.e., a pre-trained MHA model)
  - Additional pre-training to allow the model to adapt to its new structure

- The projection matrices for key and value heads are mean-pooled into single projection matrices

[1] Ainslie et al. *GQA Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints*. EMNLP, 2023

Grouped-Query Attention

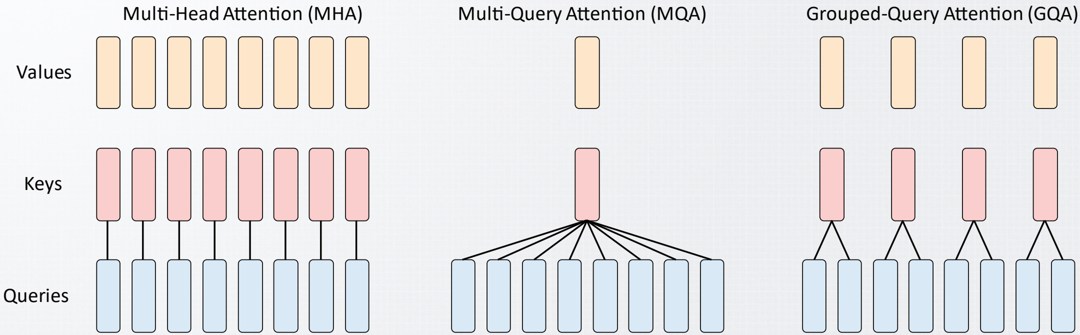
Multi-query attention (MQA) & Grouped-Query Attention (GQA)

- Grouped-query attention divides the query components into G groups
    - Each group $g \in G$ shares a **single** key head and value head

- The GQA constructs each group key and value head by mean-pooling all the original heads within that group

Overall Architectures

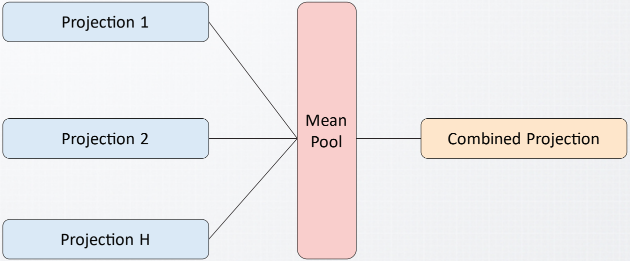
Multi-query attention (MQA) & Grouped-Query Attention (GQA)

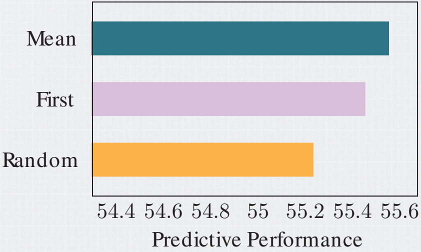

Multi-Head Attention (MHA)
Multi-Query Attention (MQA)
Grouped-Query Attention (GQA)
Values
Keys
Queries

Overall Architectures

Multi-query attention (MQA) & Grouped-Query Attention (GQA)

- Ainslie et al.[1] find this strategy works better than selecting a single key and value head or randomly initializing new key and value heads from scratch

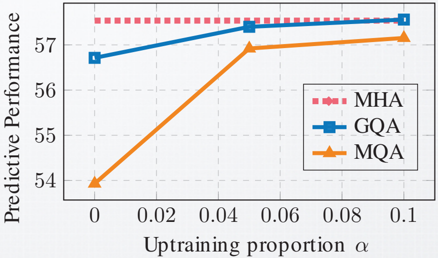


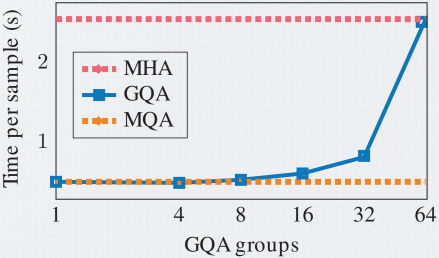


[1] Ainslie et al. *GQA Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints.* EMNLP, 2023

Overall Architectures

Multi-query attention (MQA) & Grouped-Query Attention (GQA)


Predictive Performance
Uptraining proportion α
MHA
GQA
MQA


Time per sample (s)
GQA groups
MHA
GQA
MQA