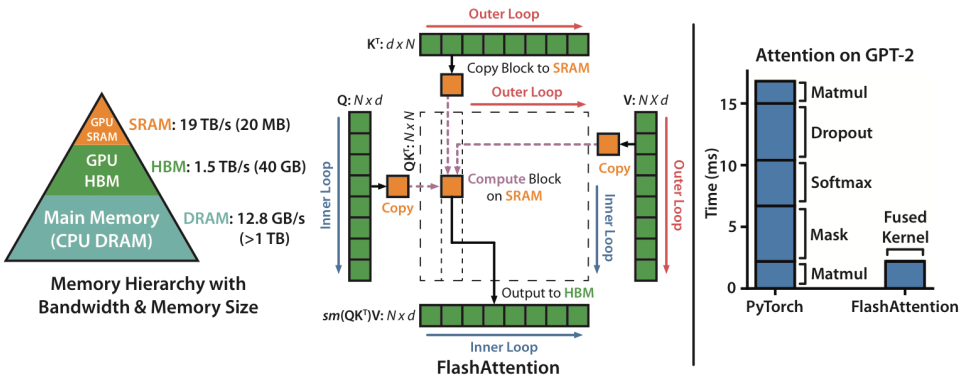

Flash Attention from first principles

Umar Jamil

Topics

- Intro to Multi-Head Attention
- Safe softmax
- Online softmax
- Intro to CUDA & GPUs
- Tensor layouts
- Examples of CUDA Kernels
- Block Matrix Multiplication
- From CUDA to Triton
- Software Pipelining
- Flash Attention (Forward Pass)
- Autograd
- Derivatives and gradients
- Gradient of the MatMul operation
- Gradient of the softmax operation
- Flash Attention (Backward Pass)

Prerequisites:

- High school calculus (derivatives)
- Basics of linear algebra (matrix multiplication, transpose...)
- Basic knowledge of the attention mechanism.
- Lots of patience