Wrap up

- A new attention mechanism to train language models

- Tensoring to obtain non-negative sketches

- Learned Sketches

- Exact attention locally

- Significantly faster to train for long contexts

- No apparent drop in model quality compared to softmax transformers

- Constant space/time inference per token

- Can be further optimized using more efficient implementations of the lower triangular multiplication algorithm (Future work)

40

Wrap up

- A new attention mechanism to train language models
  - Tensoring to obtain non-negative sketches
  - Learned Sketches
  - Exact attention locally
  - Significantly faster to train for long contexts
  - No apparent drop in model quality compared to softmax transformers
  - Constant space/time inference per token
  - Can be further optimized using more efficient implementations of the lower triangular multiplication algorithm (Future work)

Fast and Space Optimal Streaming Algorithms

with Mikkel Thorup, Rasmus Pagh and David Woodruff [FOCS '23]