Deployment

CSC2541H1 Topics in Machine Learning, Winter 2025, UToronto

Chris J. Maddison

Announcements

- If you are assigned to **present on April 4, come talk** about presentations after class.
- Presentation marking update.

Questions?

Today

- Deploying LLMs imposes additional latency and efficiency concerns.
- Today's topics:
  - Speculative decoding - speeding up decoding algorithmically
  - Paged attention (and KV cache) - caching to speed up attention
  - Flash attention and decoding - more memory-access efficient attention computation for faster decoding (and training)
  - Smooth quant - post-training quantization for faster inference with less memory use