Accelerating Position-Aware Top-\(k\) ListNet for Ranking under Custom Precision Regimes

Qiang Li, Erwei Wang, Shane T. Fleming, David B. Thomas, Peter Y. K. Cheung

Imperial College London
Machine-learning based search engine

Ranking plays a key role in information retrieval
Major Approaches for Ranking

**Challenge of listwise approach:** high computational complexity
Position-Aware Sampling

Position-aware approach delivers a better accuracy than traditional stochastic sampling method.
Batch Quantization

4.42x speedup over an Nvidia GTX 1080T GPU implementation with 2% accuracy loss
Thank you!