Background Context

Semantic Hashing: [Salakhutdinov & Weston 07; Torralba et al 08]
- unsupervised learning, auto-encoder, nonlinear NCA
- results on semantic labelled data not much better than CNR
- loss function?

Minimal Loss Hashing [Norouzi & Fleet 11]
- quantized linear mapping

Learning Objective

Bound on Loss

Stochastic Gradient Descent

Loss

Hamming Distance Metric Learning
MNST

Classification error rates on MNIST test set:

Multi-Index Hashing [CPV12]

Exact NN search in Hamming space.
- search space: Given a corpus of k-bit codes, a search query
- (1) find k codes with k smallest Hamming distances from u.
- (2) find all codes that differ from u in k or less bits.
- Imagine a dataset of 1-bit codes, a search radius of r = 2.
- Black marks depict bits that differ from a query u.
- (The first 3 codes have Hamming distance ≤ r = 2.)

Key Idea: Partition the codes into 3 substrings. Then, instead of searching r = 2 in the full codes, search r = 0 in the substrings.

Result: A single threaded implementation finds 1000 Hamming nearest neighbors of queries from one billion 64-bit codes in under 100ms.
(source code avalible at github.com/norouzi/mih/)