Learning a Cross-Modal Hashing Network for Multimedia Search

Venice Erin Liong¹, ³, Jiwen Lu², Yap-Peng Tan³

¹Interdisciplinary Graduate School (IGS) ²School of Electrical and Electronic Engineering (EEE) Nanyang Technological University, Singapore
³School of Automation Tsinghua University, Beijing, China

Overview

✓ Learn compact binary codes for cross-modality multimedia search
✓ We design a deep neural network implementation:
  ✓ Learns unified binary code discretely and discriminatively through a classification-based hinge-loss criterion
✓ Cross-modal hashing network (CMHN), one deep network for each modality, through minimizing the quantization loss between real-valued neural code and binary code, and maximizing the variance of the learned neural codes

Formulation

- Problem:
  \[ f_u : \mathbb{R}^{d_u} \rightarrow \{-1, 1\}^K , f_v : \mathbb{R}^{d_v} \rightarrow \{-1, 1\}^K \]

- The objective function of CMHN:
  \[ \min_{B,M,\theta_u,\theta_v} J = J_1 + \lambda_1 J_2 \]

  ✓ Binary codes (B) by following the assumption that the codes should be able to perform well on a multi-classification problem (using hinge-loss criterion)

  \[ \min_{B,M} J_1 = \|M\|_F^2 + \sum_{n=1}^{N} \xi_n \]

  \[ \forall n, j, y_{n,j}(m_n^T b_n) \geq 1 - \xi_n \]

✓ Network Parameters \( \theta_u, \theta_v \) by minimizing the quantization loss between neural codes and binary code and maximizes the variances

\[ \min_{\theta_u,\theta_v} J_2 = \left( \|B - H_u^T\|_2^2 + \|B - H_v^T\|_2^2 \right) \]

\[ -\alpha (tr(H_u^TH_u^T) + tr(H_v^TH_v^T)) \]

Optimization

We perform optimization by fixing the other variables and solving one variable alternatively and iteratively.

- Classification Step: learn the classification matrix (M) by having a support vector machine (SVM) formulation which can solved through a standard solver (libsvm)

- Binary Code Step: learn B by having a binary quadratic problem which can be solved through a linear gradient technique as follows:

  \[ b_n = \text{sgn}(y_n^TM^T + \lambda_1(h_u^T + h_v^T)) \]


- Implementation details:
  ✓ image network – pretrained CNN-F up to FC7 + new FC layer + hashlayer (tanh)
  ✓ Text network – FC layers [D – 500 - K]

- Out-of-sample extension:
  \[ b_{un} = \text{sgn}(h_{un}^T), b_{vn} = \text{sgn}(h_{vn}^T) \]

Experiments

- IAPRTC12: 19627 image-sentence pairs. Top 22 frequent labels from the 275 concept. Text feature: BoW 1386 dim

- NUSWIDE: 186577 images-tag pairs. Top 10 frequent concepts from 81 concepts. Text feature: BoW 1000 dim

Acknowledgements

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the Infocomm Media Development Authority, Singapore.