Data-Aware Proxy Hashing for Cross-modal Retrieval

基本信息

- 发表刊物：SIGIR 2023
- 作者：Rong-Cheng Tu; Xian-Ling Mao; Wenjin Ji; Wei Wei; Heyan Huang
- 第一完成单位：Beijing Institute of Technology（北京理工大学）
- 关键词：数据感知，跨模态，哈希
- keywords: Data-Aware, Corss-Modal, Hashing

论文内容

背景和动机

- 现存的哈希编码生成仅基于数据集的类别信息或数据标签，没有考虑数据本身
- 生成的编码可能会产生偏差，特别是在编码长度较短的时候，MAP 不如基于数据相似性的方法

主要方法

- 构建一个数据感知网络，将数据点、数据标签向量、数据集类别向量作为输入，生成数据感知基于类别的（class-based data-aware）、图像感知混合标签（label-fused image-aware）的、文本感知混合标签（label-fused text-aware）的哈希编码
- 提出一种新的哈希损失（hash loss），将这三种类型的哈希编码作为监督信息，训练不同模态的哈希网络
- 损失收敛后，这些特定模态的哈希模型就可以用于训练生成哈希编码

实现细节

问题定义

- l 个类别，n 个文本-图像实例对，表示为$o_i = \{x_i^v, x_i^t\}$，上标 $m \in \{v, t\}$ 指明模态是图像还是文本
- 实例标签 $y_i \in \{0,1\}^l$，$y_{ij} = 1$ 即该实例属于第 j 类；构建类别矩阵 $C \in \{0,1\}^{l \times l}$
- 特定模态哈希网络将数据转化为哈希编码 $h_i^m = sgn(F^m(x_i^m)) \in \{-1, 1\}^k$
- 最终目的是保存数据点的语义相似性，i.e. 当 x_i^v 相似于 x_j^t 时，h_i^v 和 h_j^t 之间的汉明距离也应该是小的

数据感知网络

data-aware proxy network (DPN)

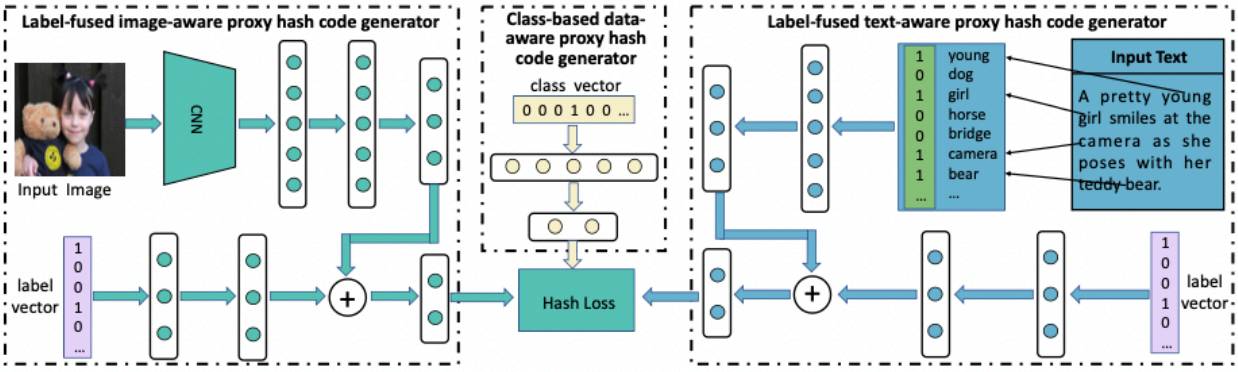

Figure 1: The architecture of our proposed DPN.

- label-fused image-aware proxy hash code generator (LIPG)
    - Alexnet，最后一层换成全连接层以提取图像特征 f_i^v
    - 两层 multi-layer percept (MLP) 将标签向量 y_i 映射为标签特征 f_i^{lv}
    - 把 l_2 正则化后的 f_i^v 和 f_i^{lv} 之和输入 k 维全连接层，加上 $sgn(\cdot)$ 函数，生成 x_i^v 对应的图像感知混合标签哈希码 b_i^v
- label-fused text-aware proxy hash code generator (LTPG)
    - 两层 MLP 提取文本特征 f_i^t
    - 两层 MLP 获得标签特征 f_i^{lv}
    - 步骤同 LIPG，生成文本感知混合标签哈希码 b_i^t

- class-based data-aware proxy hash code generator (CDPG)
    - 两层 MLP 将类别向量 c_i 映射为数据感知哈希码 b_i^c

DPN 的哈希损失

$$L_D = L_1 + \alpha L_2 + L_3 + L_{qtf}$$

- L_1 使得 b_i^c 和 b_j^c 的相似度尽可能小，即不同类别的差异尽可能大
- L_2 是一个 balanced 约束，使得哈希码 b_i^v, b_i^t, b_i^c 均匀分布，表达更多信息
- L_3 有两个目的：1）让 x_i^m 的编码和它所属类别的编码相似，和不所属类别的编码不相似，这样混合标签的编码就可以保存其语义信息；2）基于类别的编码和文本/图像编码有所互动，使前者 b_i^c 在学习时将实际数据点及其标签也纳入计算，以成为监督接下来特定模态哈希网络学习的 hash center
- L_{qtf} 是量化损失，为了抛掉 $sgn(\cdot)$ 函数后补全 error

特定模态哈希网络

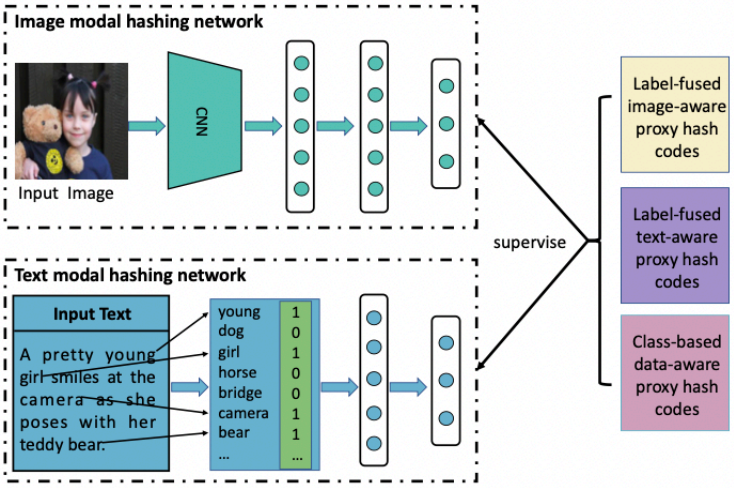

Figure 2: The architecture of modality-specific hashing networks.

- image model hashing network (IMHN)
    - Alexnet，最后一层换成全连接层，加上 $sgn(\cdot)$ 函数，输入图像，输出哈希码
- text model hashing network (TMHN)
    - 两层 MLP，输入文本，输出哈希码

直接使用学习完毕的 B^c, B^v, B^t 监督 IMHN 和 TMHN 的学习过程，构建一个损失函数：

$$L^m = L^m_{class} + L^m_{same} + \eta L^m_{similar}$$

- L^m_{class} 使得 h^m_i 和它所属类别编码之间的平均相似性大于和不所属类别编码的
- L^m_{same} 使得 h^m_i 和与它类别完全相同的 b^u_j 之间的相似性大于和它类别完全不相同的 b^u_q
- $L^m_{similar}$ 和 L^m_{same} 类似，只不过换成了和 h^m_i 类别部分相同的 b^u_p

实验

数据集

- 采用三种 benchmark 数据集：
    - IAPR TC-12：20,000个图像-文本对，255个类别，文本表示为2000维词袋向量
    - NUS-WIDE：选用近200,000个图像-文本对，21个类别，文本表示为1000维词袋向量
    - MS COCO：约120,000个图像-文本对，文本表示为2000维词袋向量

评价指标

- Mean Average Precision (MAP)
- Precision-Recall curves (PR)

实验任务和结果

- 和当前的多种 SOTA 跨模态模型比较（包括基于哈希编码的和数据相似性的），检测 DAPH 的性能
- 进行消融研究：
    - 将三种损失函数排列组合研究其有效性
    - 把另一种 DCPH 方法的哈希编码替换为 class-based data-aware 的编码，研究其优越性
    - 与哈希编码生成和特定模态哈希网络的联合学习进行比较

实验说明的方法优点

- 表现优于所有跨模态检索任务的 SOTA baseline，生成的编码质量更高，保存了更多的语义信息
- 比基于数据相似性的方法表现更好，特别是编码长度较长的时候，表明将类别/标签向量转化为哈希编码能有效提高检索表现
- 相似数据点生成的编码之间的汉明距离也很小
- 三个损失函数叠加的效果最好，说明每个都是有用的
- class-based data-aware 编码的效果明显更好，特别是数据集类别较多的时候，表明将数据本身纳入计算会有效提高检索表现
- DAPH 的效果在所有的编码长度上都好于联合学习

思考

论文优点

- 增加了接收的信息量，并将其作为监督，使得生成编码保存了更多的语义信息
- 通过构建特定模态的模型，更有针对性地优化了检索效果

改进空间

- 仅适用于文本-图像之间的跨模态检索，不适用于其他类型的数据
- 不同数据集上表现最佳的超参数都不一样，要投入应用，泛化性还有待提升
- 进一步优化损失函数的计算公式
- 可以尝试结合其他技术，比如注意力机制等，挖掘潜藏的语义信息