Multi-modal approach to Deep Hashing for fast Image Retrieval

Abstract

Fast content-based image retrieval in large image databases like ImageNet [1] or BigEarthNet [2] poses a significant challenge. Hashing algorithms can help by indexing the data. Recently convolutional neural networks (CCNs) have been used successfully to hash images [3]. However, images are often captioned or have associated text meta-data which is not taken into account in these methods. We propose researching a multi-modal neural net producing binary code hashes from image as well as textual input.

Problem

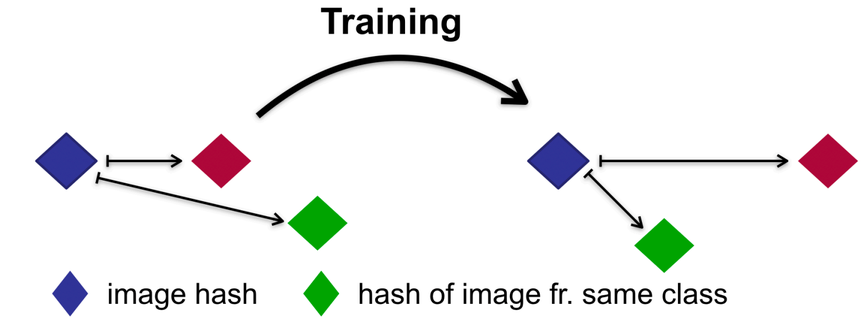
When retrieving images based on a reference image most of the time additional describing text is available, especially in the internet era. This information is often not incorporated in the retrieval process, thus a lot of useful information is potentially unused.

Goal

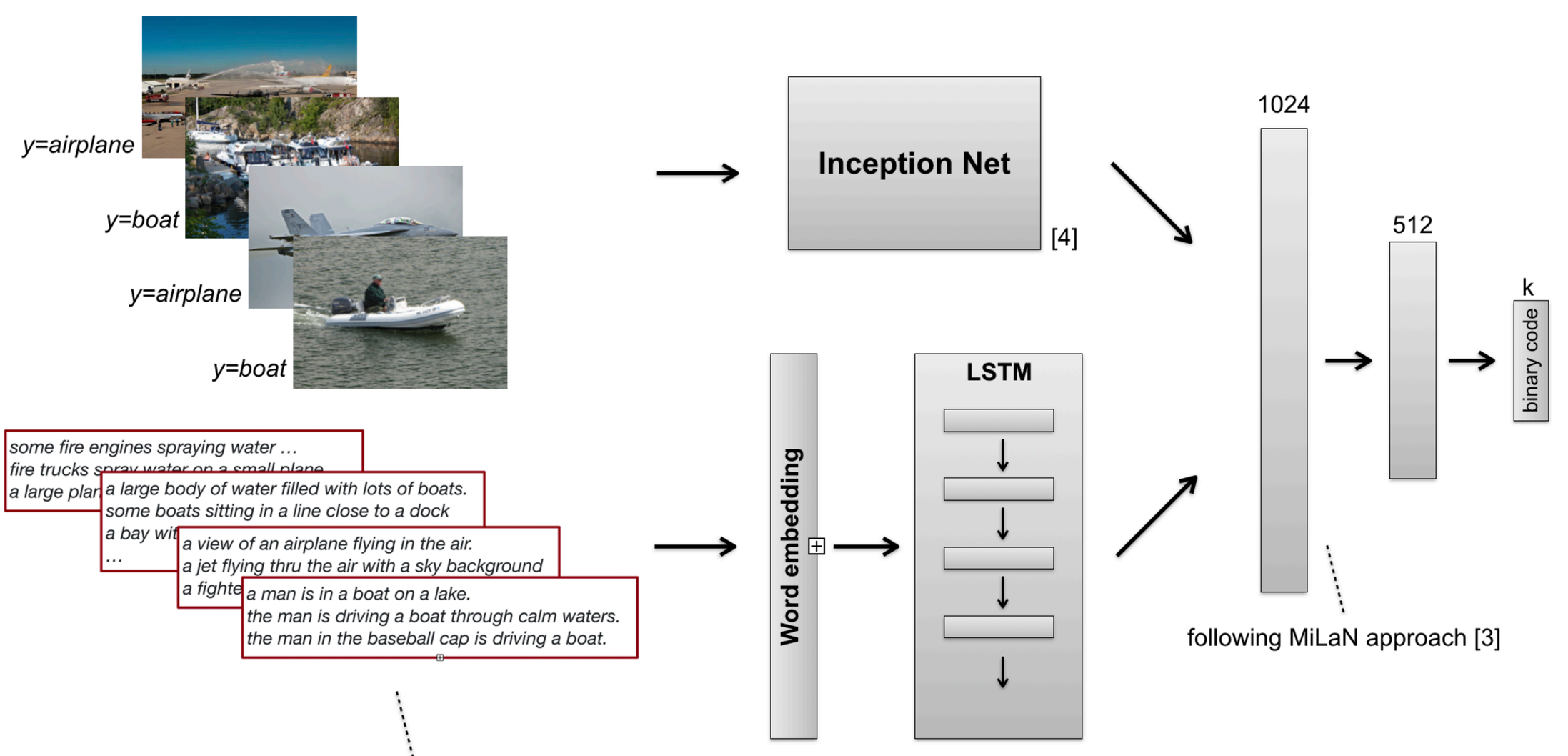
The resulting hash algorithm should map image data which includes text information (e.g. using the CoCo dataset [5] for evaluation) to more representative hash buckets than state-of-the-art hashing algorithms wich only rely on the image itself.

Solution

The idea is to train a multi-modal neural network which is able to take two different types of input data. It consists of a LSTM (or RNN) part reading text data (i.e. image captions and meta-data) and a CNN classifying the image itself. The LSTM can be based on pre-trained word embedding models like GloVe or word2vec. As the network's loss function the similarity measure 'triplet loss' can be used [6]. The output should be a binary code of length k, serving as the hash value.
The networks architecture can be seen below.



Triplet Loss Intuition



texts and images from CoCo dataset [5]

[1] image-net.org
[2] bigearth.net
[3] Metric-Learning based Deep Hashing Network for Content Based Retrieval of Remote Sensing Images; Subhankar Roy, Enver Sangineto, Begüm Demir, Nicu Sebe; CoRR; 2019

[4] Rethinking the inception architecture for computer vision; Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna; CVPR; 2016
[5] cocodataset.org
[6] FaceNet: A Unified Embedding for Face Recognition and Clustering; Florian Schroff, Dmitry Kalenichenko and James Philbin; CoRR; 2015

Emanuel Metzenthin

Master Student, Data Engineering
Hasso Plattner Institute, Potsdam, Germany

E-Mail: emanuel.metzenthin@student.hpi.de

HPI Hasso Plattner Institut