Deep Learning Techniques for Remote Sensing Image Retrieval

Ushasi Chaudhuri

Under the guidance of
Biplab Banerjee, Avik Bhattacharya & Mihai Datcu

Center of Studies in Resources Engineering (CSRE)
IIT Bombay

February 2019

Introduction

- Due to availability of wide range of satellite sensors, accumulation of an unprecedented volume of remote sensing images.
- Numerous applications: Classification, retrieval, segmentation, change detection..
- Availability of large volumes of data, arises the necessity of sophisticated information extraction strategies.
- Image retrieval is one such popular mechanism where the aim is to retrieve a number of visually coherent images from a predefined database given a query.
- Query can be in terms of text or images (text based or content based CBIR).

Introduction

- CBIR performance depends on 2 factors:
  1. **representation capability** of extracted features
  2. efficiency of **similarity measure**.
- If extracted features are of very high dimension, system is affected by the curse of dimensionality.
- Also, feature space should be discriminative to avoid class overlapping problem.
- Problem of exhaustive high dimensional feature matching:
  1. improve the **search criteria**
  2. **dimensionality reduction** in the feature space.

Literature Survey

An Introduction to Convolutional Neural Networks

Keiron O'Shea[1] and Ryan Nash[2]

[1] Department of Computer Science, Aberystwyth University, Ceredigion, SY23 3DB
keo7@aber.ac.uk

[2] School of Computing and Communications, Lancaster University, Lancashire, LA1 4YW
nashrd@live.lancs.ac.uk

Abstract. The field of machine learning has taken a dramatic twist in recent times, with the rise of the Artificial Neural Network (ANN). These biologically inspired computational models are able to far exceed the performance of previous forms of artificial intelligence in common machine learning tasks. One of the most impressive forms of ANN architecture is that of the Convolutional Neural Network (CNN). CNNs are primarily used to solve difficult image-driven pattern recognition tasks and with their precise yet simple architecture, offers a simplified method of getting started with ANNs.

This document provides a brief introduction to CNNs, discussing recently published papers and newly formed techniques in developing these brilliantly fantastic image recognition models. This introduction assumes you are familiar with the fundamentals of ANNs and machine learning.

Literature Survey- 2015

Published as a conference paper at ICLR 2015

VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

Karen Simonyan* **& Andrew Zisserman+**
Visual Geometry Group, Department of Engineering Science, University of Oxford
{karen,az}@robots.ox.ac.uk

ABSTRACT

In this work we investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Our main contribution is a thorough evaluation of networks of increasing depth using an architecture with very small (3 × 3) convolution filters, which shows that a significant improvement on the prior-art configurations can
weight layers. These findings wer
submission, where our team secured
tion and classification tracks respec
generalise well to other datasets, w
have made our two best-performing
tate further research on the use of d

1 INTRODUCTION

Convolutional networks (ConvNets) have
age and video recognition (Krizhevsky et al
Simonyan & Zisserman, 2014) which has be
ries, such as ImageNet (Deng et al., 2009), an
or large-scale distributed clusters (Dean et al
of deep visual recognition architectures has b

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512

Literature Survey

Information Mining in Remote Sensing Image Archives: System Evaluation

Herbert Daschiel and Mihai Datcu

***Abstract*—In this paper, we present an algorithmic protocol for the evaluation of a content-based remote sensing image information mining system. In order to provide users fast access to the content of large image databases, the system is composed of two main modules. The first includes computationally intensive algorithms for off-line data ingestion in the archive, image feature extraction, and indexing. The second module consists of a graphical man–machine interface that manages the information fusion for interactive interpretation and the image information mining functions. According to the system architecture, the proposed evaluation methodology aims to determine the objective technical quality of the system and includes subjective human factors as well. Since the query performance of a content-based image retrieval system mainly depends on the datasets stored in the archive, we first analyze the complexity of image data. Then, we determine the accuracy of the interactive training that can be considered as a supervised Bayesian classification of the entire archive. Based on the stochastic nature of user-defined cover types, the system retrieves images using probabilistic measurements. The information quality of the queried results is measured by target and misclassified images, precision and recall, and the probability to forget and to overretrieve images. Since the queried images are the result of a number of interactions between user and system, we analyze the man–machine communication dialogue and the system operation, too. Finally, we compare the objective component of the evaluation protocol with the users' degree of satisfaction to point out the significance of the computed measurements.**

To overcome these limitations, we implemented an intelligent knowledge-driven information mining (KIM) system, which easily enables users the access to remote sensing image archives via Internet communication, the analysis of information details by an intelligent man–machine interface (MMI), and the search of the whole archive for cover type definitions. A full and detailed description of KIM is presented in [4].

In general the state-of-the-art mining systems have a common architecture. Additionally, the knowledge-driven image information mining system has a clear hierarchical information representation where the image content is modeled in a Bayesian way using different levels of semantic abstraction. From the computational point of view, the hierarchy can be partitioned into two major parts: a computational intensive off-line part which aims at the extraction and description of the image content in a completely, application-free way, and the fast, user-specific definition of semantic cover types. In Fig. 1, the KIM system architecture is presented. At time of data ingestion, multimission images are tiled in subimages, indexed and stored in a repository. From these images, primitive features are extracted, such as spectral, textural and geometrical attributes. To obtain a quasicomplete description of the entire image content, textural and geometrical features are extracted at multiple scales.

Literature Survey- GCN 2017

SEMI-SUPERVISED CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS

Thomas N. Kipf
University of Amsterdam
T.N.Kipf@uva.nl

Max Welling
University of Amsterdam
Canadian Institute for Advanced Research (CIFAR)
M.Welling@uva.nl

ABSTRACT

We present a scalable approach for semi-supervised learning on graph-structured data that is based on an efficient variant of convolutional neural networks which operate directly on graphs. We motivate the choice of our convolutional architecture via a localized first-order approximation of spectral graph convolutions. Our model scales linearly in the number of graph edges and learns hidden layer representations that encode both local graph structure and features of nodes. In a number of experiments on citation networks and on a knowledge graph dataset we demonstrate that our approach outperforms related methods by a significant margin.

1 INTRODUCTION

We consider the problem of classifying nodes (such as documents) in a graph (such as a citation network), where labels are only available for a small subset of nodes. This problem can be framed as graph-based semi-supervised learning, where label information is smoothed over the graph via some form of explicit graph-based regularization (Zhu et al., 2003; Zhou et al., 2004; Belkin et al., 2006; Weston et al., 2012), e.g. by using a graph Laplacian regularization term in the loss function:

$$\mathcal{L} = \mathcal{L}_0 + \lambda \mathcal{L}_{\text{reg}}, \quad \text{with} \quad \mathcal{L}_{\text{reg}} = \sum_{i,j} A_{ij} \| f(X_i) - f(X_j) \|^2 = f(X)^\top \Delta f(X), \qquad (1)$$

Motivation

- **Deep convolutional neural network** (CNN) based data-driven feature extractors, have displaced the conventional features for a number of visual inference tasks in the recent past.
- Standard CNN based feature descriptors summarize the **global** scene level information from the images.
- Such global descriptors, though extensively used, may not yield satisfactory performance for RS images since the local region interaction is also key to recognizing many complex land-cover classes.

Motivation

We plan to focus our attention on two important problems:

1. **global image representation**, highlighting the interactions among the **local scene** constructs
2. discriminative learning of an embedding space from an **irregular spatial distribution** of regions for the purpose of designing a CBIR system for VHR RS data.

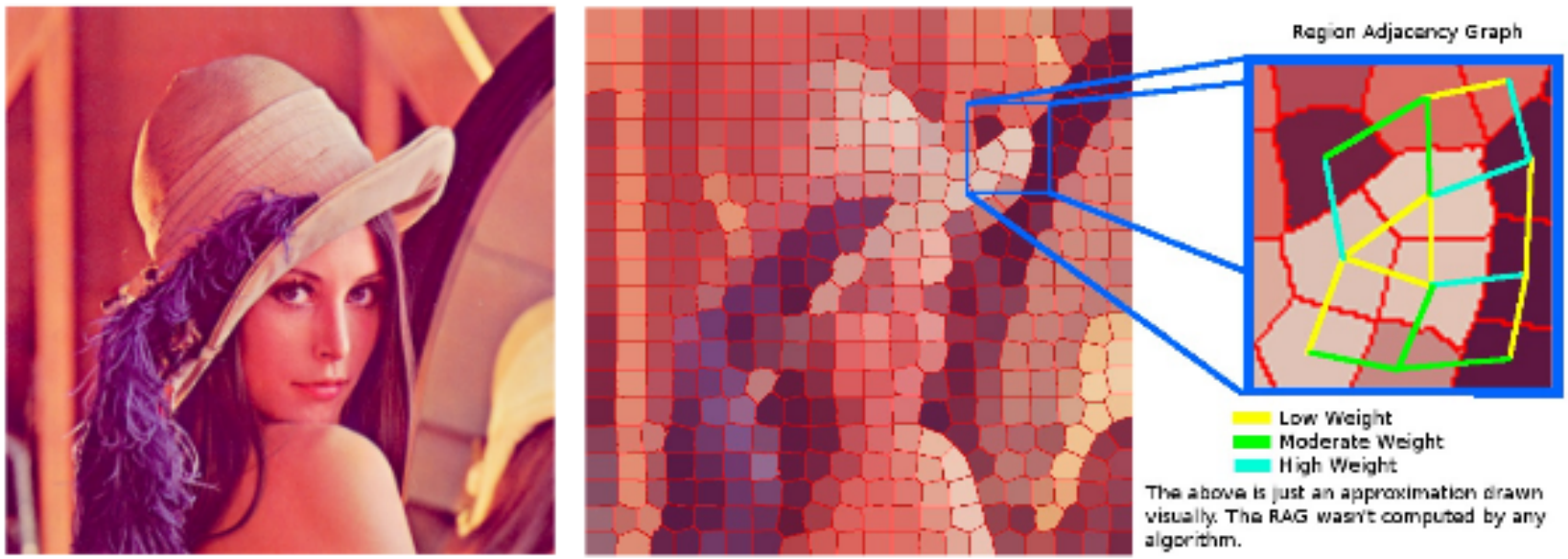


Figure: Region adjacency graph (RAG) construction.

Our Contributions

1. N labeled VHR RS images $\chi = \{x_i, c_i\}_{i=1}^{N}$ from C different classe., we retrieve similar images from χ for the query image x^{query}
2. $x_i = (G_i, A_i)$, where G_i — weighted graph adjacency matrix and A_i — node level attributes.
3. Learn a new embedding space $\widehat{x_i} = f(G_i, A_i)$ following the proposed SGCN model on which the Euclidean distance measure is considered for performing CBIR.
4. For training the SGCN model, positive and negative pairs of graphs from χ are mined.

Overall Block Diagram

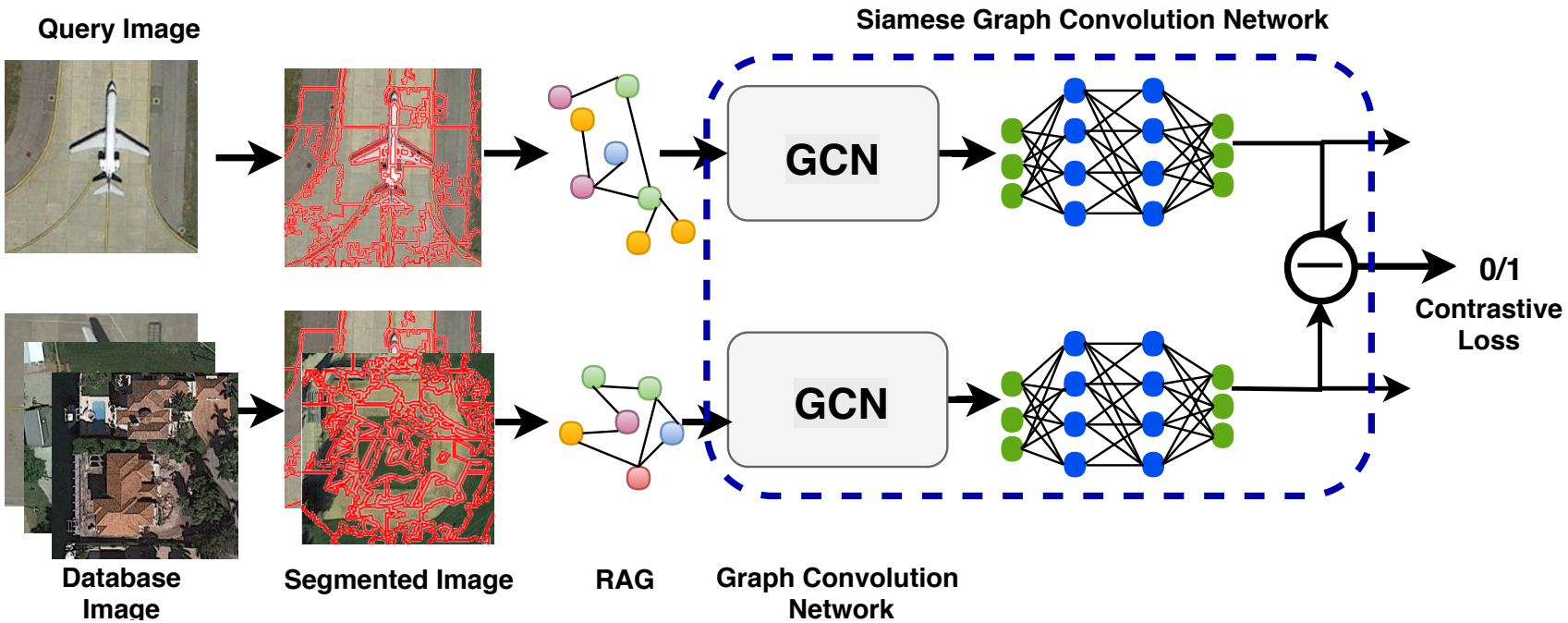


Figure: Pipeline of the proposed SGCN network. The images are first segmented and their segment wise features are extracted. From this segmented image, a RAG is formed, where each node contains their corresponding segment feature vector. These graphs are fed into the GCN layer for a pre-training, which then is fed to the SGCN layer for a complete training. The SGCN layer is trained like a siamese twin pair with two images at a time, and the contrastive loss is minimized.

1. Feature extraction

Given an input image x_i, the aim is to segment it into n_i non-overlapping homogeneous regions $\{r_i^1, r_i^2, \cdots, r_i^{n_i}\}$, using the SLIC super-pixel and Graph-cut algorithm (tried separately).

- **shape features**: Fourier descriptors (mean and standard deviation of the magnitude and phase) and contour-based features (area, convex area, filled area, perimeter, extent, solidity and eccentricity, equivalent diameter, Euler number, major and minor axes lengths, orientations, and bounding box).
- **color features**: color histogram , and color moments (mean, standard deviation and skewness)
- **texture features**: spectral histogram, entropy, local binary pattern (LBP) features and local phase quantization (LPQ) features.

1. Feature extraction...

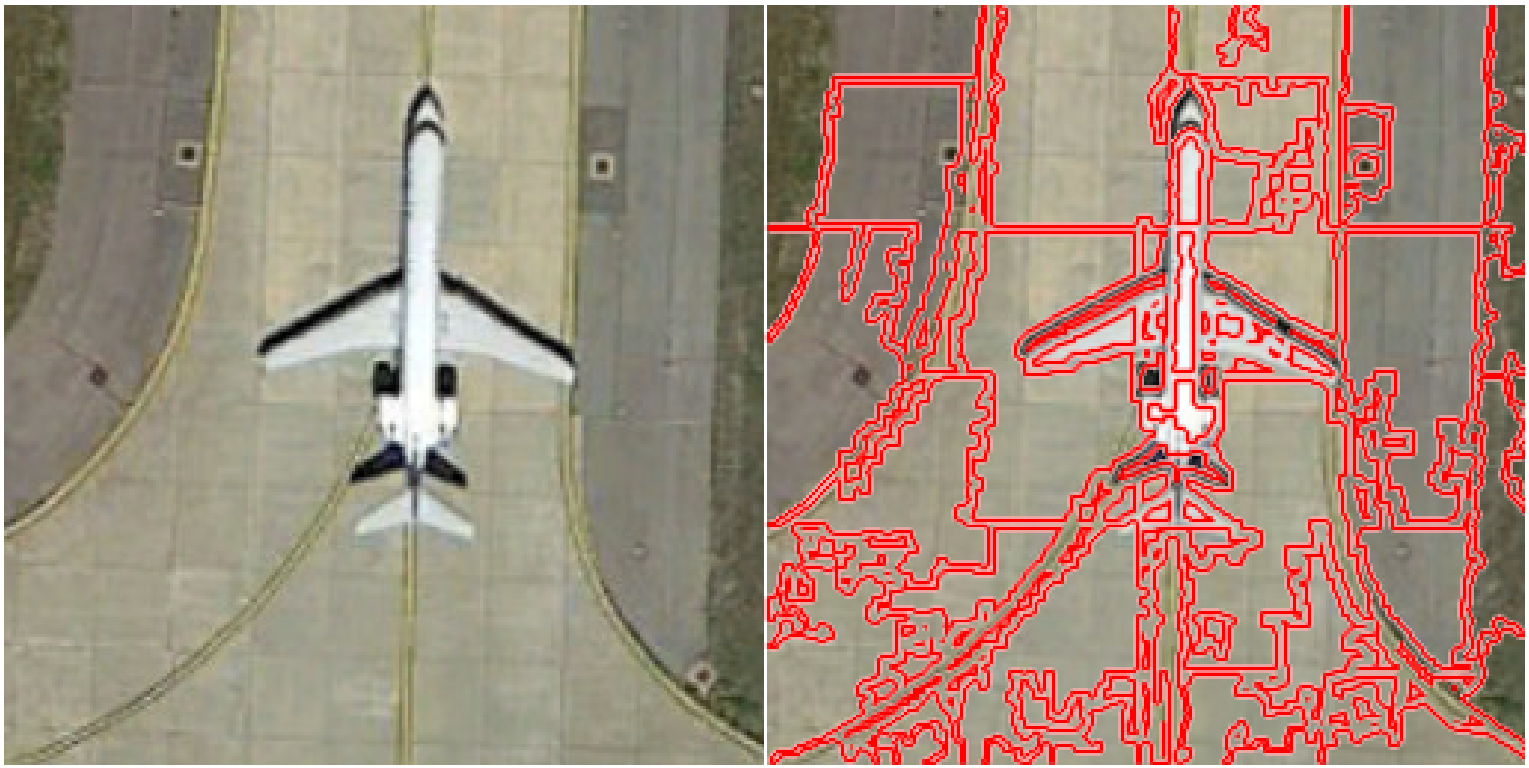


Figure: Segmentation using VIFeat SLIC superpixels. Features from each segments are extracted for the formation of RAG.

2. Region Adjacency Graph (RAG) formation

- A region adjacency graph is created using the regments and their region properties.
- Edges exist between pairs of vertices if the corresponding regions are adjacent to each other and the edge weight $E_i(r_i^j, r_i^k)$ between a pair of regions is calculated following Equation , else E is set to zero.

$$G_i(r_i^j, r_i^k) = \alpha_1 ||c_{r_i^j} - c_{r_i^k}||_2 + \alpha_2 |\theta_{r_i^j} - \theta_{r_i^k}| \quad (1)$$

- $c_{r_i^j}$ represents the centroid pixel of the j^{th} region. $\theta_{r_i^j}$ is the orientation angle (angle between horizontal axis and the major axis of the ellipse having same second moment as the region and $\theta \in [-90, +90]$).
- weights α_1 and α_2 are fixed empirically.

3. Graph Convolution and Pooling

A graph convolution network (GCN) has a topology similar to CNN, but follows a very different approach to carry out its convolution and pooling operations.

- **Convolution:**

$$\mathbf{H} = h_0\mathbf{I} + h_1\mathbf{A_1}, \qquad \mathbf{H} \in \mathbb{R}^{N\times N} \tag{2}$$

$$\mathbf{V}_o^{(k)} = \sum_{l=1}^{I} \mathbf{H}^{(l,k)}\mathbf{V}_i^{(l)} + b \tag{3}$$

  Input vertex features - $\mathbf{V}_i$; First-order approximation of their adjacency matrix - $\mathbf{H}$; zeroth-order and first-order adjacency matrices - h_0 and h_1.

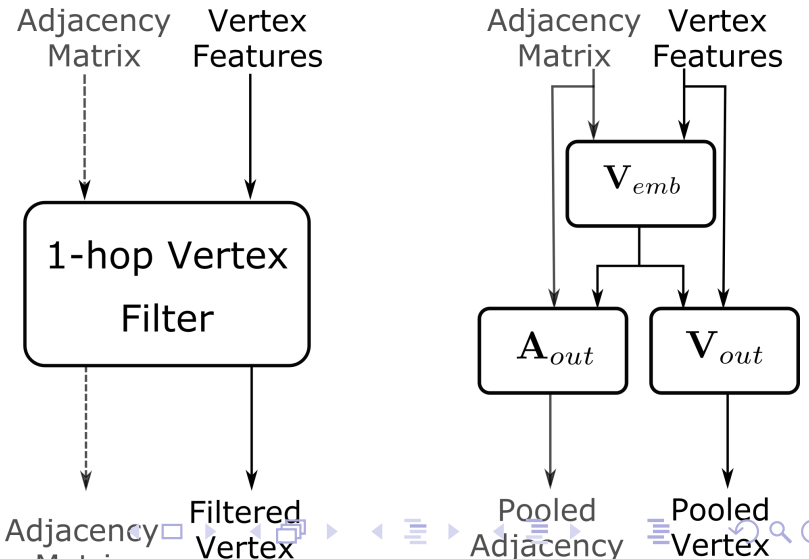
- This filter helps in preserving the structural non-linearities, while achieving a high receptive field.
- Region neighbourhood information preserved.

3. Graph Convolution and Pooling...

- **Pooling:** Graph Embed Pooling layer maps the input graph of any structure and size to a fixed structured output.

$$\mathbf{V}_{emb}^{(n')} = \sum_{l=1}^{k} \mathbf{H}_{emb}^{(l,n')} \mathbf{V}_{i}^{(l)} + b \tag{4}$$

- Helps in attaining a translational in variance for the region of interest.
- pooling constrains the computational load, while increasing the receptive field of the subsequent convolution layer.



4. Siamese Graph Networks (SGCN)

- Siamese CNN network is used to learn an embedding space where images from similar classes are forced to reside closely while making images from different classes to drift apart from each other.
- Obtained embedding space is expected to be more discriminative in comparison to the input feature space.
- Siamese network consists of two identical networks (with sharable weight parameters)
- After performing a series of graph convolution and pooling operations on the input data, a fully-connected layer is used in order to project the graph structured data in the space of one-dimensional vectors ($fc1$)

4. Siamese Graph Networks (SGCN)...

- Let $\tilde{x_i}$ and $\tilde{x_j}$ be 2 feature vectors from *fc1* (input data points x_i and x_j and W_S - shared parameter vector to be learnt).
- To train W_S, the Contrastive loss is optimized to obtain $\widehat{x} = \tilde{x} W_S$ through subsequent fully connected layers (*fc2*)

$$L^k = \sum_{n=1}^{B} \Bigg((1 - y_{nij})(\widehat{x}_{ni} - \widehat{x}_{nj})^2 + y_{nij} \left\{ \max(0, m - (\widehat{x}_{ni} - \widehat{x}_{nj})) \right\}^2 + \alpha \left(|\widehat{x}_{ni}|_2 + |\widehat{x}_{nj}|_2 \right) + \beta |\theta|_2 \Bigg) \quad (5)$$

- We add an Euclidean norm of the features ($|\widehat{x}_i|_2$ and $|\widehat{x}_j|_2$) to our loss function to force our features into a controlled space.
- Positive class: $y_{ij} = 0$; Negative class: $y_{ij} = 1$.

5. Selection of positive and negative pairs

- For training, a pair of images are to be fed to this system.
- When $y_{ij} = 0$, $(l_{ij} = (1 - y_{ij})(\widehat{x}_i - \widehat{x}_j)^2)$. When $y_{ij} = 1$, $(l_{ij} = y_{ij}\{max(0, m - (\widehat{x}_i - \widehat{x}_j)\}^2)$.
- Results in pushing two images far in the feature space, further away.
- Brings similar class images closer in feature space.
- For the negative pair selection, images from a different class which lie closest, as well as the image which lie farthest from the other pair image are chosen randomly.

6. Image retrieval using SGCN based feature embedding

- The final layer fully-connected features (*fc2*), given by $\widehat{x}$ for the entire dataset χ of VHR RS images is stored.
- *k*-nearest neighbours are then found for a given query image x^{query} from all the classes.
- Both the l_1 distance and the l_2 distances have been considered for the nearest neighbour distance calculation.

Results

Datasets:

1. UC Merced land use dataset. 21 classes of land use, each containing 100 images. Total 2100 images. patial resolution of this VHR public dataset is 1 foot.
2. PatternNet. 38 broad categories of high resolution remote sensing images. Each class consists of 800 images, making a total of 30400 images. Created from Google Earth imagery or via the Google Map API for some US cities.

Results

Experiment protocol:

- Retrieval was done using k-NN with l_2 norm (Euclidean distance). The performance evaluation on the **UC-Merced** dataset was done with:
  1. a standard CNN
  2. color histogram of 32-bin size for each RGB spectrum
  3. first principle component for RAG (G-KNN)
  4. entire RAG (RAG-KNN)
  5. VLFeat (bag of visual words + VLAD + Fisher linear vector)
  6. VGG-VD16
  7. VGG-VD19
  8. GoogLeNet
- **PatternNet dataset:** the comparisons have been shown only with CNN and GCN method, as it showed significantly better results than the other methods on the Merced dataset.

Results

UCMerced precision-recall curves:

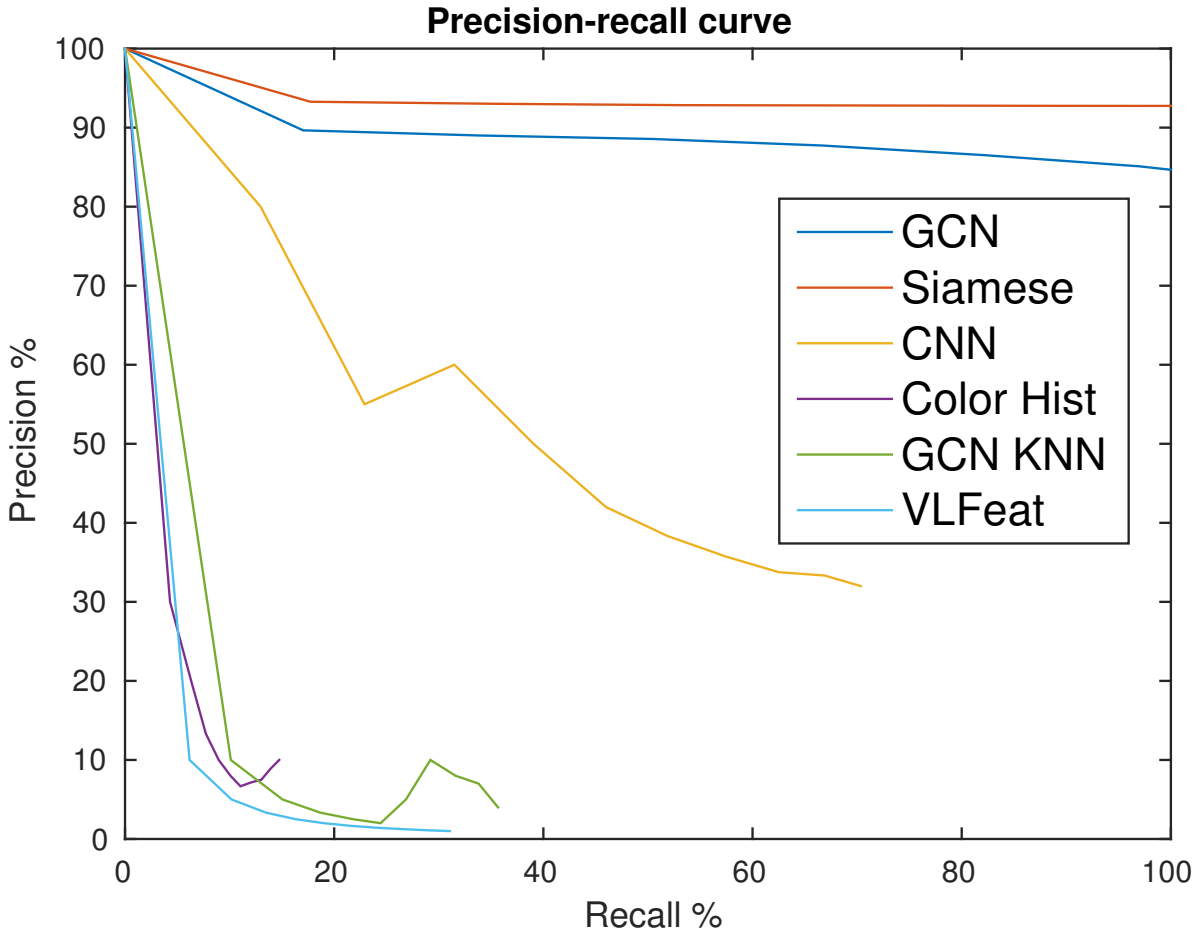
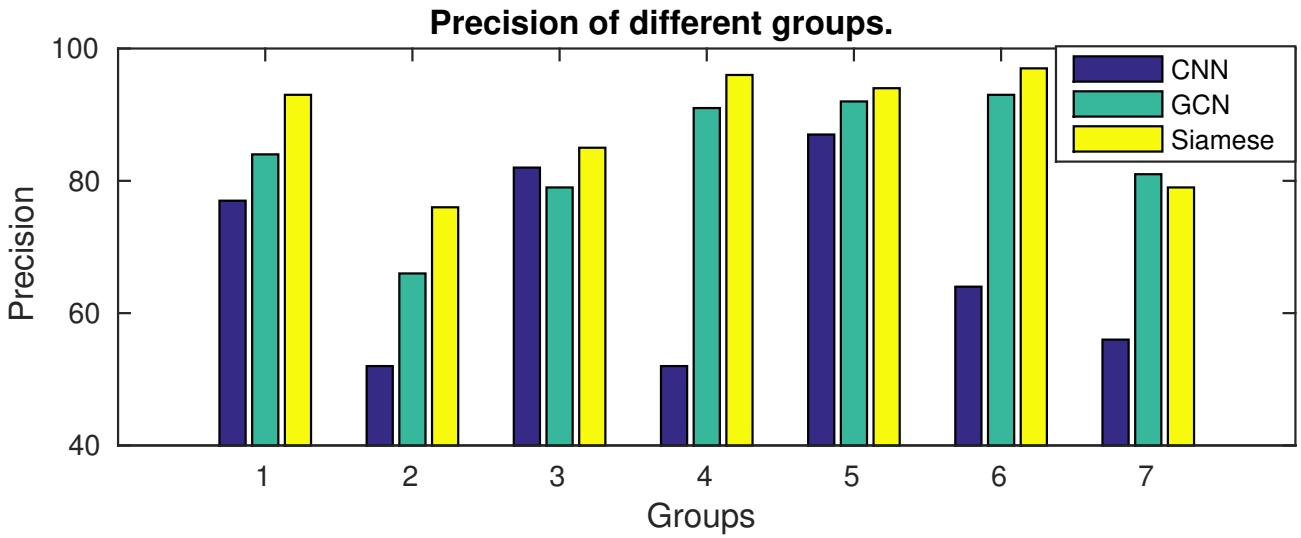


Figure: Precision recall curve of various methods on UCMerced dataset.

Results

Model	UC Merced			PatternNet		
	ANMRR	MAP(%)	P@10 (%)	ANMRR	MAP(%)	P@10
G-KNN	0.92	7.50	10.12	0.88	12.35	13.24
RAG-KNN	0.75	26.74	24.90	0.69	22.56	37.70
VGG-VD16	0.38	53.71	78.34	0.33	59.86	92.04
VGG-VD19	0.39	53.19	77.60	0.34	57.89	91.13
GoogLeNet	0.39	53.13	80.96	0.29	63.11	93.31
GCN	0.33	64.81	87.12	0.28	73.11	95.53
SGCN	**0.30**	**69.89**	**93.63**	**0.21**	**81.79**	**97.14**



Results

PatternNet data:

- GCN network achieved a **classification** accuracy of 99.00% on the PatternNet during pre-training over about 3000 iterations.
- Proposed SGCN algorithm could give a high precision of 97.14%.
- Query images for a few specific classes even resulted with a retrieval of 100% precision for the top 10 query images.
- In spite of PatternNet being a much larger dataset in size than the UC-Merced dataset, it gave far better results than the UC-Merced dataset. This might be due to the contribution of more number of samples for the training, which boosts the performance significantly.

Results

PatternNet precision-recall curves:

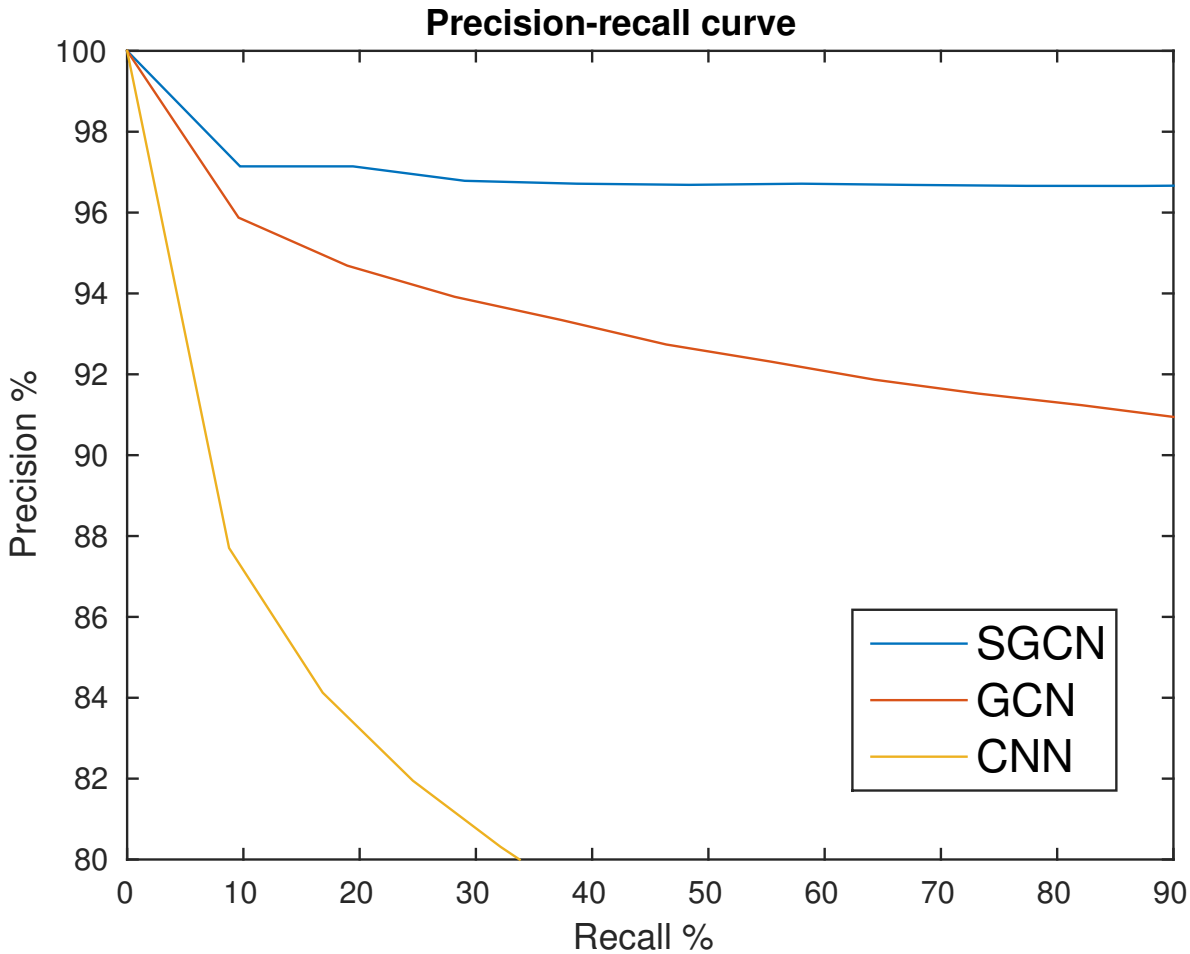


Figure: Precision recall curve of various methods on PatternNet dataset.

Results

Ablation Study:

- **Segmentation Algorithm:**

Model	UC Merced ANMRR	UC Merced MAP(%)	UC Merced P@10(%)
SGCN(1-413)	0.46	59.21	82.23
SGCN(49)	0.38	62.87	88.27
SGCN(169)	**0.30**	**69.89**	**93.63**

Table: Ablation study with different number of segments.

Results

Confusion classes:

Figure: Confusion classes: Sparse Residential Areas and Dense Residential Areas. The first three samples are from the previous class, while the subsequent three images are from the later class. Note the similarity in the classes. The last row shows samples from the class coastal mansion.

Results

Confusion classes and possible future work:

- Confusion classes: few classes like `sparse residential areas` and `dense residential areas` which are very near-by classes, show a fall in the retrieval results.
- Distinguishing between them with a full confidence score for even humans are a challenging task.
- These were the classes which were mainly responsible for the drop in the model's performance.
- The confusion matrix for these classes mainly had each others class mixing.
- Discrimination between such similar classes requires a much more fine-grained training and can be an interesting future problem.

Results

Dataset: Query	1	2	3	4	5

Query image versus the retried images from UC Merced and PatternNet. The first and the second row images are from UC Merced dataset `Airplane` and `Buildings` classes respectively, while the third and the fourth row images are from the PatternNet dataset of `ClosedRoad` and `CoastalMansion` classes.

Future Work Ideas

Confusion classes and possible future work:

- While training the SGCN part, assigning just the near or far label (i.e., 0 or 1), resulted in this much of leap in the retrieval performance. However, if we could somehow encompass the information of the class label while pushing to different images far apart, without pushing it closer to another nearby class, probably the performance might boost up even more.
- **Multi-modal data:** The problem that we have tackled was based on a single dataset. We can look into the retrieval problem on a cross-platform/domain of input data.
  1. Cross-modal retrieval: We train a system for SAR images and do the retrieval from a co-registered optical image.
  2. Fusion: SAR data of different resolutions, for ex., using Sentinel 1 (coarse resolution) and TerraSarX (fine resolution) data fusion. An optimized method for fusion, or retrieval from fused data can be an interesting problem.

2. Cross-modal retrieval of Remote sensing and multimedia data.

- Within the same dataset, retrieval is a relatively easy work. If we have different modalities of data, ex Pan images, multi-spectral images, SAR images, optical images, or texts, retrieval amongst cross domain becomes more challenging.
- A network learned for a particular modality may not give good performance on different data.

Definition (No free lunch theorem)

[a] If an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.

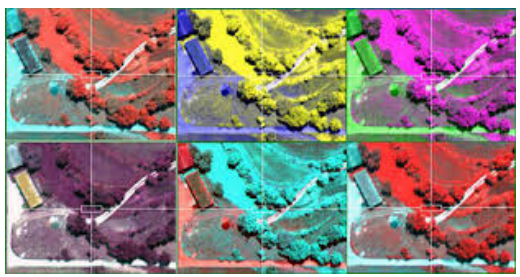
---

[a] Wolpert, D.H., Macready, W.G. (1997), "No Free Lunch Theorems for Optimization", IEEE Transactions on Evolutionary Computation 1, 67.

2. Cross-modal retrieval of pan-chromatic, multi-spectral and SAR images using hashing techniques.

Advantages of each data:

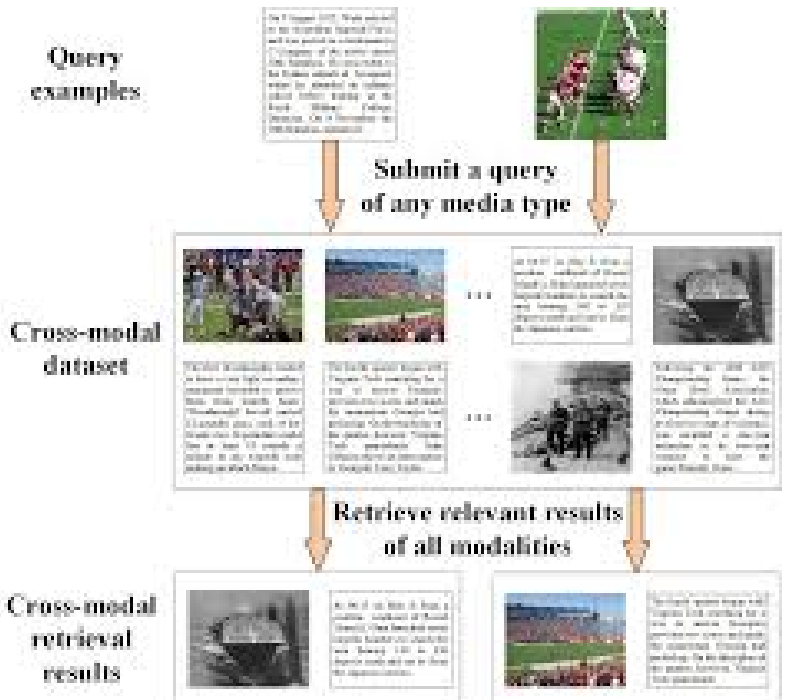
- **PAN images:** High spatial resolution.
- **Multi-spectral images:** high spectral resolution.
- **SAR images:** Polarization information. No cloud clutter.
- **VHR optical image:** High spatial resolution, 3 spectral channel.

1. 1m pan IKONOS image of Valparaiso, Chile
2. source: https://www.sensorsmag.com/components/hyperspectral-and-multispectral-imaging-sensors-find-hyper-flexible-applications
3. Sentinel-1 SAR dataset with C band
4. `sparseresidentialarea` from PatternNet dataset.

Motivation

- Important problems:
  1. PAN $\Leftrightarrow$ Multispectral.
  2. Optical $\Leftrightarrow$ SAR.
  3. RGB $\Leftrightarrow$ DEM.
  4. Image $\Leftrightarrow$ text.



Cross-modal data

Datasets used:

1. DSRSID: Panchromatic - Multi-spectral images (GF-1 satellite). 80,000 images, 8 classes. **(Single-label, paired)**.
2. UCMerced: VHR RS images (RGB) - speech. Merced: 2100 images, 17 multi-labels. speech: 15 samples of each multi-label. **(Multi-label, un-paired)**.

Problem Overview:

Challenge: To design a model which can be tuned for any cross-modal or multi-modal retrieval.

Approach to the problem:

1. Pre-training to obtain the initial embedding space of each domain separately.
2. Unified representation.
3. Learning hash codes for the unique representations (if possible?).
4. Find their classes using the assigned projection vectors for the hash codes.
5. Cross-modal or Uni-modal image retrieval.

Overall Block Diagram

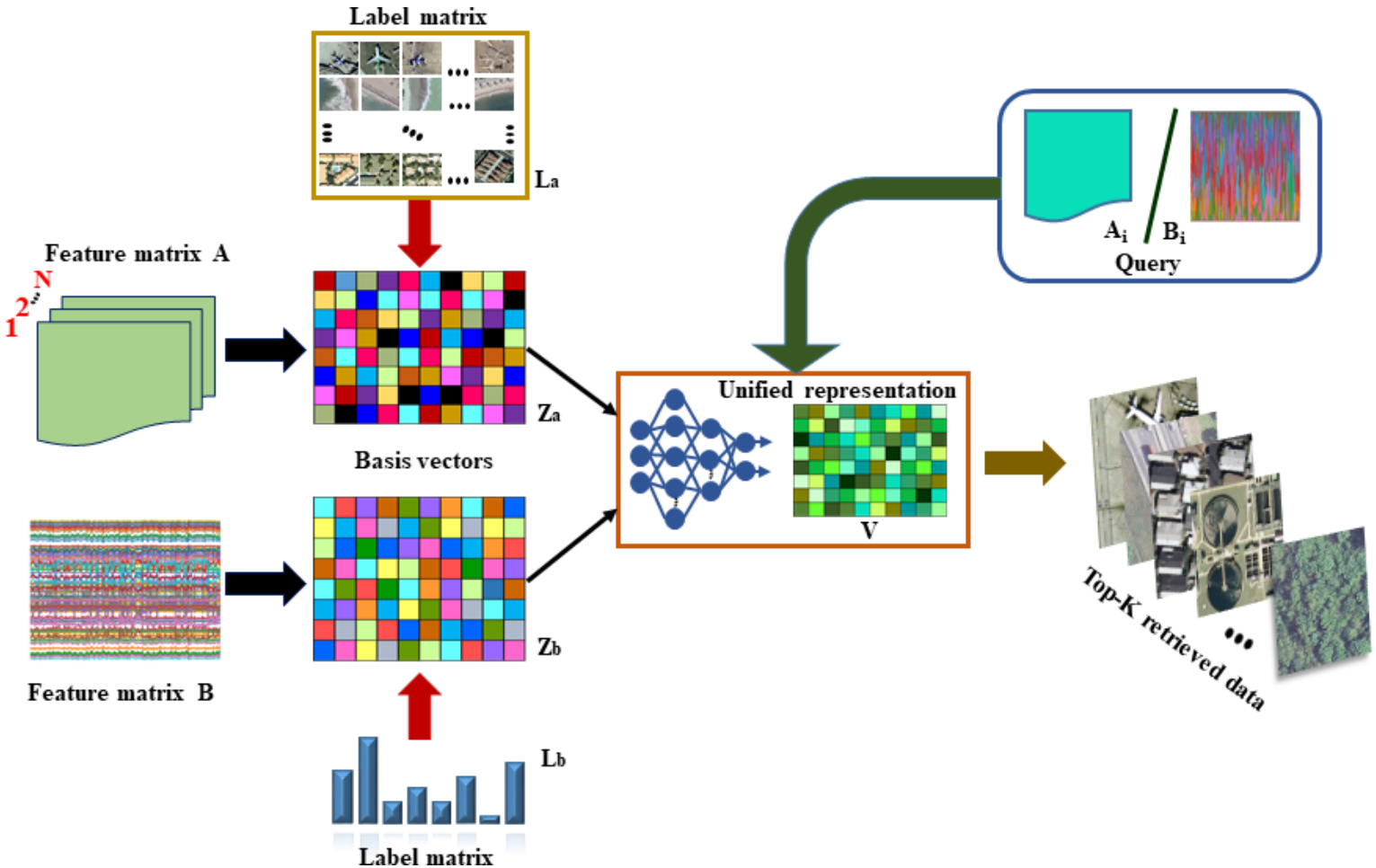


Figure: Rough pipeline of the network.

Problem Formulation

- Let (**A**, **B**) be a set of bi-modal data (**A** and **B**).
- $a_i \in \mathbb{R}^{d_a}$ and $b_i \in \mathbb{R}^{d_b}$ feature embeddings.
- $\mathbf{A} = [a_1, a_2, ..., a_n] \in \mathbb{R}^{d_a \times n}$ and $\mathbf{B} = [b_1, b_2, ..., b_m] \in \mathbb{R}^{d_b \times m}$, n and m - number of samples of data present for each modality.
- Labels $\mathbf{L}_a \in \mathbb{R}^{n \times C}$ and $\mathbf{L}_b \in \mathbb{R}^{m \times C}$, where C is the number of classes.
- Unified representation matrix as $\mathbf{V} = \{v\}_{k=1}^{p}$, where $v_k \in \mathbb{R}^{d_v}$

Loss Functions

Latent loss:

$$L_{latent} = \| w_1 \mathbf{Z}_a - w_2 \mathbf{Z}_b \|_F \tag{6}$$

Classification loss:

$$L_{classification} = \mathrm{CE}(\mathbf{Z}_a) + \mathrm{CE}(\mathbf{Z}_b) \tag{7}$$

Decoder loss:

$$L_{decoder} = \|w_3 \mathbf{V} - \mathbf{Z}_a\|_F + \|w_4 \mathbf{V} - \mathbf{Z}_b\|_F \tag{8}$$

Weight-normalization:

$$L_{norm} = \|w_1 - \mathbf{1}\|_F + \|w_2 - \mathbf{1}\|_F \tag{9}$$

Overall Objective Function and Optimization

Overall loss:

$$L = L_{latent} + L_{classification} + L_{decoder} + L_{norm} \tag{10}$$

Therefore, the total loss can be formulated as:

$$\begin{aligned} L = \lambda \| w_1 \mathbf{Z}_a - w_2 \mathbf{Z}_b \|_F + \mathrm{CE}(\mathbf{Z}_a) + \mathrm{CE}(\mathbf{Z}_b) + \\ \mu \| w_3 \mathbf{V} - \mathbf{Z}_a \|_F + \gamma \| w_4 \mathbf{V} - \mathbf{Z}_b \|_F + \\ \| w_1 - \mathbf{1} \|_F + \| w_2 - \mathbf{1} \|_F \end{aligned} \tag{11}$$

where, λ, μ, and γ are heuristically-chosen non-negative trade-off parameters.

Multiple-modality Extension

The ULSME can be extented from bi-modal data to multi-modal data directly by the following equation.

$$L^p = \sum_{i=1,j=1}^{p} \lambda_i \| w_i \mathbf{Z}_i - w_j \mathbf{Z}_j \|_F + \sum_{i=1}^{p} \mathrm{CE}(\mathbf{Z}_i) + \sum_{i=1,j=1}^{p} \mu_{i,j} \| w_{j'} \mathbf{V} - \mathbf{Z}_i \|_F + \sum_{i=1}^{p} \| w_i - \mathbf{1} \|_F \quad (12)$$

Publications:

1. Ushasi Chaudhuri, Biplab Banerjee, Avik Bhattacharya, Siamese graph convolutional network for content based remote sensing image retrieval, In *Computer Vision and Image Understanding* (undergoing major revision).

2. Ushasi Chaudhuri, Biplab Banerjee, Avik Bhattacharya, An End-to-End Framework For Cross-Modal Multimedia Retrieval in Remote Sensing, *Earth Vision CVPRw* (in submission).

References

- Such, Felipe Petroski, et al. "Robust spatial filtering with graph convolutional neural networks." *IEEE Journal of Selected Topics in Signal Processing* 11.6 (2017): 884-896.
- Zhou, Weixun, et al. "Patternnet: a benchmark dataset for performance evaluation of remote sensing image retrieval." *ISPRS Journal of Photogrammetry and Remote Sensing (2018)*.
- Yang, Yi, and Shawn Newsam. "Bag-of-visual-words and spatial extensions for land-use classification." *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. ACM, 2010.
- Chaudhuri, Bindita, et al. "Region-based retrieval of remote sensing images using an unsupervised graph-theoretic approach." *IEEE Geoscience and Remote Sensing Letters* 13.7 (2016): 987-991.

Thank You