TU Berlin / Faculty IV
Remote Sensing Image Analysis Group

Proposal for a Bachelor / Master Thesis

Type of thesis / Line of study:	Master Thesis Computer Science Summer Semester 2020 6 Months
Title of the thesis:	Content Based Image Retrieval by Deep Multi Modal Hashing for Sentinel Images
Candidate:	Bank, Hasan
Matriculation number:	396582
Advisors(s):	Dr. Kang, Jian
Supervisors(s):	Prof. Dr. Demir, Begüm
Planned period:	June, 2020 until December 2020

1. Introduction / Scientific Background / Related Work

Hashing is getting more critical to retrieve data in terms of time and storage efficiency. Hashing methods convert original high dimensional images into small binary hash codes. Therefore, storage requirements of original data have been reduced significantly, and similarities of the images are measured by the calculated Hamming distance between the binary hash codes. Decreased cost of storage and less searching time are the fundamental objects to use hashing in image retrieval.

Lin et al. (2016) proposed an unsupervised deep hashing approach for image retrieval in a single modality[1]. It did not require labeled training data and had three goals during the creation of binary hash codes. These were minimal quantization loss, evenly distributed codes, and uncorrelated bits.

Thanks to the improvement in earth observation technologies, the volume of remote sensing datasets has been increasing, and the traditional methods to search in these big datasets are not enough anymore. Deep hashing techniques have played a notable role in image retrieval from these remote sensing image archives due to their accuracy and speed.

A supervised deep hashing approach was proposed for fast and accurate image search in remote sensing by Roy et al. (2018)[2]. It utilizes a pre-trained network beside the second stage is trained with different losses such as triplet loss, representation penalty, and a balancing loss.

Instead of having only one modality, some datasets can have multi modalities such as image, audio, text so on so forth. There are several examples of multi modalities. For example, image or audio files can have text descriptors as labels or tags in order to keep some text information which can be related to the class or category of the data.

Cross-modality hashing is an adaptation of the hashing mechanism to use these multi modalities. Using binary hash codes of one modality to retrieve data from another modality is called cross-modality hashing.

Most cross- or multi-modality hashing studies are based on computer vision datasets [3]-[6], rather than remote sensing datasets. However, Sentinel-1 radar images and Sentinel-2 multispectral optical images are used in our approach.

Li et al. (2018) presented a cross-source hashing approach for image retrieval based on remote sensing datasets by using panchromatic images and multispectral images [7]. It is a study in deep cross-modality hashing in remote sensing but distinct from our study in terms of the used data source.

2. Problem Statement / Goals of the Thesis

Remote sensing images from different sources have been continually increasing, yet there are not many studies focused on multi-modalities. Sentinel-1 and Sentinel-2 have different spatial resolutions and image structures, but together they provide better information about geographic areas. Multimodality techniques give a chance to use different modalities together. Different modalities can complement each other to improve the performance of image retrieval.

To create a correlation between Sentinel-1 and Sentinel 2 data sources is very difficult with handcrafted features because of their heterogeneities. That's why deep learning techniques would be used for image retrieval in a cross-modality remote sensing dataset.

Using deep learning networks pre-trained on computer vision datasets for remote sensing is not an effective way to learn features. Remote sensing images are differentiated from natural images with their spatial and spectral resolution. Hence, the creation of a new network is required to have better results. From scratch, separate CNN models per modality would be created and all train-test phases would be conducted by remote sensing images. No pre-trained on computer vision networks would be used.

Captured images from the same coordinates, but different sources are paired. They are categorized in the same land-cover classes. The system will be trained by these paired datasets. After the training phase, the aim is to retrieve similar images by using different modality. If the query is a Sentinel-2 image, the system should return semantically similar images from the Sentinel-1 database or vice-versa.

The images can have single or multi labels. If the dataset is created as a single label per instance, semantically similar images should have the same label. If the dataset is created as multi labels per instance, semantically similar images should have some shared labels, not necessarily to be the same.

3. Thesis Approach / Plan of Implementation

Configurations of the CNNs for image modalities should be decided. AlexNet[8], GoogleNet[9], VGG Net[10], ResNet[11] are some examples of deep convolutional network architectures. They have already provided good results in ImageNet ILSVRC challenges and they are popular models to extract a deep representation of the visual data. One of these architectures can be adapted to our data-sources with some modifications.

BigEarthNet[12] Sentinel-2 archive is used with Sentinel-1 images. The focused area is the country of Serbia. A subset of BigEarthNet has been created. That has only images related to the area of Serbia. Sentinel-1 and Sentinel-2 image pairs have been created. Paired Sentinel-1 and Sentinel-2 images mean that they have the same coordinates and same land-cover classes. These pairs will be used to feed the CNN models.

A joint-loss function will be applied to achieve different aims. There should be a small hamming distance between binary values for images that share some classes. Therefore, binarized query images will retrieve the closest binary values which refer images in the database that have shared classes with the query. There are two scenarios in that case. The semantically similar images can be either in the same data source or in a different data source. Li et al. (2018) provided

IRSC to push binary codes of images closer if they are from the same categories but different sources[7]. IRSC is an intersource loss function. Also, IASC has been provided as a intrasource loss function. It pushes the binary codes of images closer, too. However, there is a difference from IRSC. IASC does not run between the data sources. It means IASC pushes the codes closer when the images are from the same data sources and have the same categories. However, these loss functions were used based on a single label dataset in the aforementioned paper. They should be modified for a multi labels dataset because BigEarthNet is a multi-label archive.

Also, minimal binary quantization loss should be assured between binary hash codes and the original input image. In order to maximize usage of every bit in hashing, the number of 1s and 0s should be closer. Thus, these should be evenly distributed.

4. Time frame

The time frame of the thesis is presented in table 1.

table 1: time frame of the thesis

5. Bibliography

[1] Lin, K., Lu, J., Chen, C.S. and Zhou, J., 2016. Learning compact binary descriptors with unsupervised deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1183-1192).
[2]Roy, S., Sangineto, E., Demir, B. and Sebe, N., 2018, July. Deep metric and hash-code learning for content-based retrieval of remote sensing images. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium (pp. 4539-4542). IEEE.
[3] Jiang, Q.Y. and Li, W.J., 2017. Deep cross-modal hashing. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3232-3240).
[4] Cao, Y., Long, M., Wang, J., Yang, Q. and Yu, P.S., 2016, August. Deep visual-semantic hashing for cross-modal retrieval. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1445-1454).
[5] Wang, W., Yang, X., Ooi, B.C., Zhang, D. and Zhuang, Y., 2016. Effective deep learning-based multi-modal retrieval. The VLDB Journal, 25(1), pp.79-101.
[6] Chen, Z.D., Yu, W.J., Li, C.X., Nie, L. and Xu, X.S., 2018, April. Dual deep neural networks cross-modal hashing. In Thirty-Second AAAI Conference on Artificial Intelligence.
[7] Li, Y., Zhang, Y., Huang, X. and Ma, J., 2018. Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval. IEEE Transactions on Geoscience and Remote Sensing, 56(11), pp.6521-6536.
[8] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
[9] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[10] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[11] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
[12] Sumbul, G., Charfuelan, M., Demir, B. and Markl, V., 2019, July. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium (pp. 5901-5904). IEEE.