GNR652 Project: Review of Fusion Attention Network (CVPRw 2020)

Ruturaj D.
IIT Bombay

***Abstract*—Due to the development in sensor technology, different forms of data are available for the same problem. In remote sensing, the data is available in the form of hyperspectral and lidar data. Although both of them can be used separately to build a classification model, fusion of the two leads to more interesting features and better classification accuracy. The fusion of the two modalities is non-trivial. In this project, I have explored one such method as described in the CVPR 2020 workshop paper, FusAtNet: Dual Attention-based Spectro Spatial Multimodal Fusion Network for Hyperspectral and LiDAR Classification.**

I. Introduction

With recent advances in sensing, we can obtain multiple forms of data of the same. Hyperspectral data and Lidar data are easily available in the remote sensing domain. Each form of data is able to capture different qualities of the region of interest, which aids in overall understanding of the region. Lidar data provides information regarding the elevation of different objects; hence it can be used to differentiate between high-rise buildings, roads, etc. Hyperspectral data consists of 144 different spectrums, each having different interaction with the objects. Some objects reflect certain frequencies but don't reflect others. Such data can be used to differentiate between the material of objects, pollution monitoring, etc.

Several strategies have been used to implement this data fusion: support vector machine(SVM), random forests(RF). In the era of deep learning, this paper follows a multi-stream architecture using convolutional neural networks. Although the performance of such an approach is excellent, the main disadvantage is that features are extracted separately. Hence this causes some important shared features to be missed. There also exists the possibility of feature imbalance leading to features not being equally represented. Simple concatenation may lead to redundant data, causing overfitting. Concatenation of features also causes large dimensionality, which may cause the model to suffer. The above-mentioned challenges have led to new methods being adopted for multimodal fusion.

The salient aspect of this project is using attention to selectively emphasize important features and suppress the prominent ones. Along with self-attention, cross-modality attention is also used in this network. Cross attention is implemented by using the attention map derived from lidar data on spectral data, highlighting the important features that would not have been noticed if only self-attention was used. After implementing the method discussed in the paper, I have achieved an accuracy of 79.86% after 50 epochs on Houston 2013 data.

II. Background and Previous Work

In the remote sensing domain, several classical techniques such as Support Vector Machines, Random Forest, Extreme Learning Machines have been applied to achieve multimodal fusion. With the advent of deep learning, it has been actively used in multimodal learning and feature attention. Spectral attention frameworks have been suggested [2] to discover material-dependent features for better classification accuracy. Multi-stream models have also been introduced in [3]. The previous techniques overlook attention-based feature learning. FusAtNet introduces cross-modal attention, which is a new idea in this field.

III. Datasets

The paper tests the architecture on three datasets: Houston 2013 dataset, MUUFL Gulfport dataset, Trento dataset. I have tested my model on the Houston dataset. It was introduced in GRSS Data Fusion Contest 2013. It consists of 144 spectral bands capturing data around the Houston university campus. 15029 ground truth samples are available segregated into 15 classes. There are 2832 training and 12197 testing samples. The main challenge I had to deal with was reading the data since it was in .tif image format. I used the rasterio library for the same. Understanding the data was also challenging due to its size (144,1905,349). The link to the dataset is https://hyperspectral.ee.uh.edu/?page_id=459

IV. Procedure and Experiments

After downloading the dataset in Google Colab, I converted the dataset into an appropriately scaled NumPy array. That includes both the hyperspectral and lidar datasets. Then, from the data text files, I collected the training and testing data in pandas dataframes. Corresponding to each data sample in the training dataframe, I added an 11*11 patch into the training NumPy array. The same procedure was repeated for the testing array. This was followed by making Conv Unit (Pad and No Pad), Residual Unit 1, and Residual Unit 2 functions, which are the basic building blocks for the six significant modules.

Tensorflow and Keras libraries are used in this step. I then coded each of the six modules: Hyperspectral feature extractor, Spectral attention module, Spatial attention module, Modality

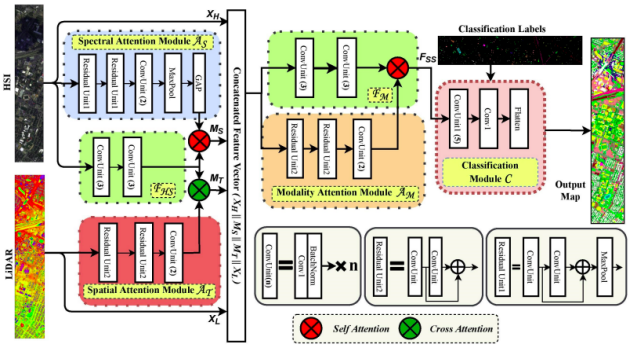

Fig. 1. Schematic of FusAtNet is as shown.
Source: Diagram taken original from FusAtNet CVPR workshop 2020 Paper[1]

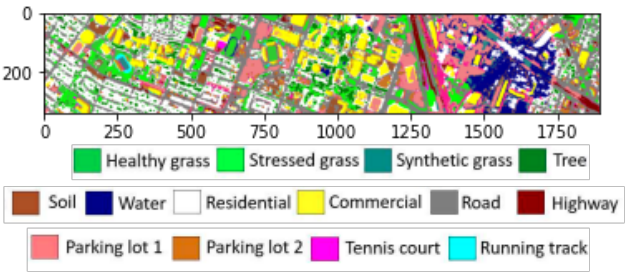

Fig. 2. The generated classification map with individual class labels

feature extractor, Modality attention module, Classification module. After this, I made the model using the Model class in Keras, linking all the modules as described in the paper. I then trained the model on the training set for 50 epochs, Adams optimizer, and sparse categorical cross-entropy loss and batch size=32. The trained model was used to generate a classification map of the entire area. I have used color coding in the map so that it looks as real as possible (e.g., Blue color for water).

V. Results

I have trained my model on Houston Data for 50 epochs, and I observed a testing accuracy of 80%. In the paper, the maximum accuracy reached is 89.98%. The difference in the accuracy can be explained by the fact that in addition to using data augmentation, they have trained for 1000 epochs. The ablation study done in the paper shows that FusAtNet beats current state-of-the-art networks. The classification map generated by this method is also less noisy.

VI. Conclusion

A novel fusion architecture is proposed in this paper, which utilizes hyperspectral and lidar data to improve classification accuracy in the remote sensing domain. FusAtNet uses different attention modules to learn useful features from both modalities. FusAtNet can be used in various applications to deal with several types of modalities.

VII. Statement of Contributions

This is a single student team project. In this project, I, Vikrant Rangnekar, have done the following things: project conception, data collection and assembly, the actual implementation of the paper (everything coded from scratch), report writing, and video making. The coding part took me 30+ hours and the making the report on Overleaf and Video making combined took about 10 hours.

References

[1] Mohla, Satyam & Pande, Shivam & Banerjee, Biplab & Chaudhuri, Subhasis. (2020). FusAtNet: Dual Attention based SpectroSpatial Multimodal Fusion Network for Hyperspectral and LiDAR Classification. 10.21203/rs.3.rs-32802/v1.

[2] Lichao Mou and Xiao Xiang Zhu. Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 2019

[3] Yushi Chen, Chunyang Li, Pedram Ghamisi, Chunyu Shi, and Yanfeng Gu. Deep fusion of hyperspectral and lidar data for thematic classification. In 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pages 3591–3594. IEEE, 2016

[4] https://stats.stackexchange.com/questions/326065/cross-entropy-vs-sparse-cross-entropy-when-to-use-one-over-the-other

[5] https://keras.io/api/models/model/

[6] https://www.pyimagesearch.com/2019/10/28/3-ways-to-create-a-keras-model-with-tensorflow-2-0-sequential-functional-and-model-subclassing/