HYPER-VIT : A NOVEL LIGHT-WEIGHTED VISUAL TRANSFORMER-BASED SUPERVISED CLASSIFICATION FRAMEWORK FOR HYPERSPECTRAL REMOTE SENSING APPLICATIONS

Bishwas Praveen[†], *and Vineetha Menon*[†]

[†]The University of Alabama in Huntsville, Huntsville, AL, 35899, USA
Email : bp0052@uah.edu ; vineetha.menon@uah.edu

ABSTRACT

Hyperspectral Imagery (HSI) data inherently has the ability to store finer details in the form of reflectance information through its contiguous spectral bands, which is utilized to discriminate between materials included in the data. With the emergence of deep learning (DL) over the last decade and the extent to which it has influenced applications and research in the domain of remote sensing is significant. Convolutional neural networks (CNNs), residual networks (ResNets), recurrent neural networks (RNNs), and other deep learning constructs have been employed to develop remote sensing-based computer vision applications, and have produced excellent results time and time again. However, the aforementioned deep learning constructs lack the intrinsic ability to prioritize data features based on how cardinal they are in mapping inputs to ground-truths, generally called attention, thus not exploiting the underlying architecture's potential to the fullest. Hence, in this work, we explore the breadth of influence of a novel, computationally efficient visual transformer (ViT) based architecture on HSI data based classification tasks. The efficacy of the proposed architecture is evaluated through a series of experimentation and the performance is compared against other state-of-the-art attention based HSI data classification methodologies on two datasets, Salinas and Pavia University. When compared to other approaches discussed, experimental results show that our proposed methodology outperformed them in terms of classification efficacy and computational complexity under limited training samples scenario.

Index Terms**—** Hyperspectral Remote Sensing, Visual Transformer, Deep Learning, HSI Data Classification.

1. INTRODUCTION

With the emergence of hyperspectral imaging, the applications directly related to the domain of target/object detection in warfare [1], urban planning [2], air-borne land-cover surveillance [3], crop monitoring [4], weather forecasting [5], etc. have been greatly influenced in a positive manner. It's indeed worth noting that the introduction of deep learning has had a huge impact on the development of applications in hyperspectral remote sensing for the aforementioned tasks [6]. In particular, computer vision, a major branch of artificial intelligence, has been widely employed in the literature to develop solutions for HSI data-based applications [7, 8].

Hyperspectral imaging is well recognized for encoding reflectance information of materials of interest using contiguous spectral bands which are usually hundreds in number. In general, when using a deep learning based architecture for computer vision tasks using HSI data, such as CNNs [9], ResNets, RNNs, and so on, they have no structural way of prioritizing features based on relevance unless an explicit attention module is manually inserted in the network [10].

Thus, in our work, we propose a novel computationally efficient visual transformer based classification methodology for comprehensive HSI data analysis. Firstly, PCA-based dimensionality reduction (DR) is applied on raw hyperspectral data for effective spectral feature extraction and to enhance the process of noise reduction [11]. In benchmarks for numerous computer vision applications, Vision Transformer, which is directly developed from the traditional transformer architecture, has set the bar high for tasks linked to Natural Language Processing (NLP), and has achieved extremely competitive performance [12]. The multi-head attention module, which recognizes global and local dependencies present in the feature space, selectively emphasizes characteristics that are critical for accurate classification of input data, is the main component of ViT and three such modules have been employed in our research. In addition, a custom data oversampler module is added in our work to alleviate the class imbalance problem, as well as data augmentation as one of the layers in the architecture is introduced. This effectively assists in generalizing the framework while training the network with a limited number of samples. Finally, an FNN-based supervised classification is introduced at the end of our framework to validate the effectiveness of the proposed approach [13].

This research work was supported by an Early Career Research Fellowship from the Gulf Research Program of the National Academies of Sciences, Engineering, and Medicine for supporting this work. DISCLAIMER: "The content is solely the responsibility of the authors and does not necessarily represent the official views of the Gulf Research Program of the National Academies of Sciences, Engineering, and Medicine."

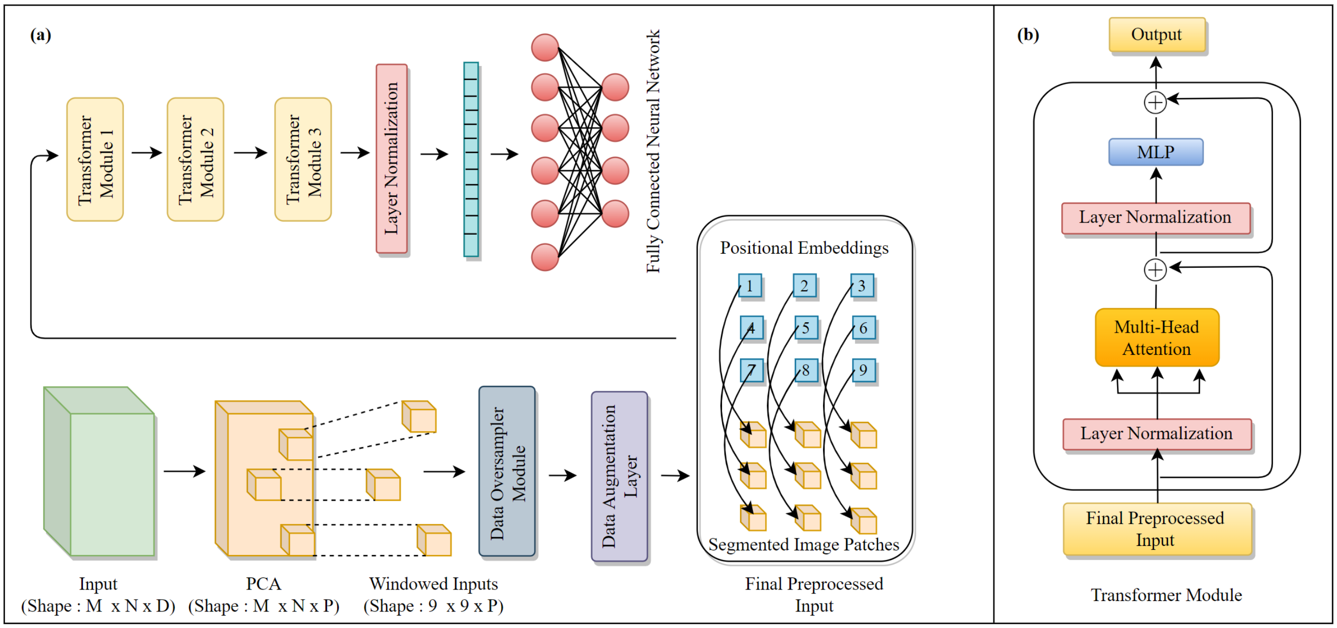


Fig. 1. **(a)** The proposed visual transformer-based supervised classification framework for hyperspectral data classification (HYPER-VIT). **(b)** The internal composition of a transformer module in the proposed architecture.

2. APPROACH OVERVIEW (HYPER-VIT)

In our proposed approach, raw hyperspectral data with the shape $(M \times N \times D)$, where D denotes the spectral dimension, is first projected to a lower dimensional subspace P using a PCA-based linear DR technique, which transforms data to the final shape of $(M \times N \times P)$. This procedure not only aids in spectral feature extraction but also comprehensively reduces noise present in the input hyperspectral data. The output generated from the above PCA-based DR module is now windowed through all the three dimensions in the shape of $(9\times9\times P)$, where the data point of interest lies at the center of the windowed block and is used as an input for supervised learning-based HSI classification approach.

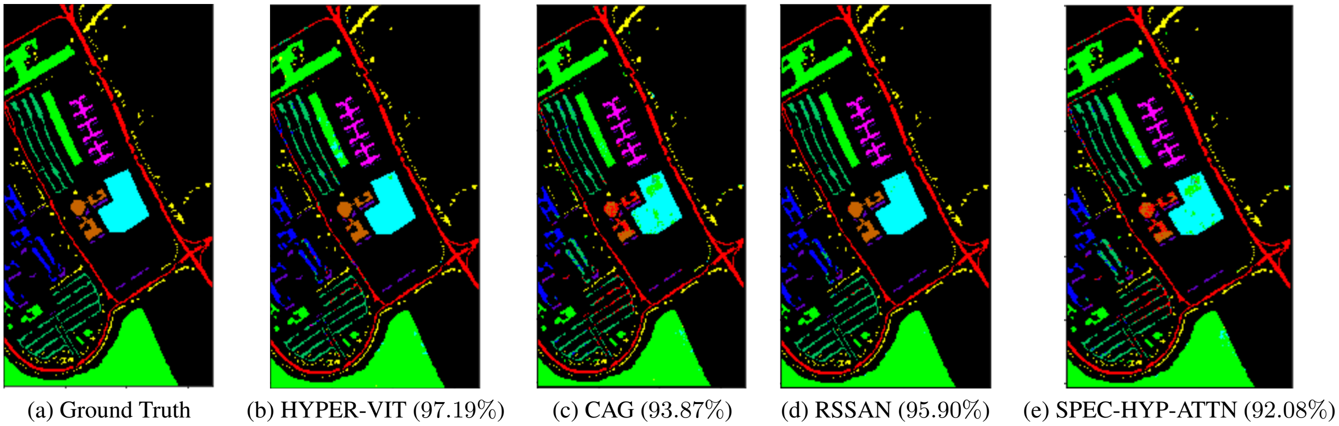
Further, the windowed data points are divided into training and testing sets, with the training set ranging from 1% through 5% of the total dataset to validate our approach and the rest of the samples are used for testing. Later, the data oversampler module uses the training samples as an input to address the class imbalance problem by duplicating data points from a minority class to roughly match the number of samples present in the class with the most samples, allowing the underlying classification framework to generalize better in a scenario with limited training samples.

It is now that the inputs are introduced to the data-augmentation layer of our proposed methodology. When the inputs enter this part of the framework, we adapt a technique that randomly flips the inputs horizontally and vertically by $180°$, allowing the network to train on data samples in different orientations at different stages of the training process, thus, pushing the network to better generalize and forecast the ground-truth of the samples during validation.

Later, the data samples are patched and appended with positional embeddings to be fed as an input to a series of transformer modules as shown in Figure 1a. For an input $x \in \mathcal{R}^{9\times9\times P}$, and a patch size of S, N_p number of patches are created and denoted by x_S, where $x_S \in \mathcal{R}^{N_p\times(S^2\times P)}$, and $N_p = \frac{(9*9)}{S^2}$. In our case, S is empirically set to 3, resulting in x_s to consist 9 patches of size $(3 \times 3 \times P)$ for every sample x. These patches are now concatenated with $1D$ positional embeddings to retain positional information of the patches in the input data point x.

The final pre-processed input is now forwarded to a series of three transformer modules where the cross-section of every transformer module used in our work is as shown in Figure 1b. The most principal component of this module is the multi-head attention layer. It is a module which internally prioritizes features based on their significance. Several of these modules are layered and operated simultaneously, such that these attention modules build attention maps to extract meaningful features from different portions of the image.

Given a query Q, key K and value matrix V, which are all mapped down to x_s in our case, mathematically, the operation of multi-head attention module can be denoted as $MultiHead(Q, K, V) = MultiHead(x_s, x_s, x_s) = [head_1, ..., head_h]W_0$, where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) = Attention(x_sW_i^{x_s}, x_sW_i^{x_s}, x_sW_i^{x_s})$ and all the $W's$ mentioned above are learnable parameters. The output from these three transformer modules are passed on to



(a) Ground Truth (b) HYPER-VIT (97.19%) (c) CAG (93.87%) (d) RSSAN (95.90%) (e) SPEC-HYP-ATTN (92.08%)

Fig. 2. Classification maps of Pavia University dataset for the proposed visual transformer based supervised classification framework and methodologies used for comparison using 2% training data.

a normalisation layer to effectively stabilize the hidden state dynamics in the classification network. The outputs are then flattened and passed on to three dense layers with 1024, 100 and n nodes respectively with a dropout of 0.5 between every two layers for an FNN based supervised classification, where n is the number of ground-truth classes in datasets chosen for experimentation.

3. METHODOLOGIES FOR COMPARISON

3.1. SPEC-HYP-ATTN

B. Praveen and V. Menon in [10] propose a computationally efficient bi-directional LSTM-based spectral attention mechanism for HSI classification. In this approach,the input high dimensional hyperspectral data cube is projected onto a lower dimensional feature space using PCA-based DR technique. As a result, the input which was originally $(M \times N \times D)$ is transformed into $(M \times N \times P)$, where P denotes the reduced dimensional space and is set to 50 in their work. Now, this data is windowed in the shape of $(11 \times 11 \times 50)$ and passed onto the 3D-CNN- and bi-directional LSTM-based spectral attention module where spectral attention maps for inputs are generated which effectively prioritizes spectral features based on their importance for effective HSI data classification. The overall architecture of their proposed approach is clearly documented in [10]. The output feature vector derived from the attention module is later passed as an input to a FNN-based supervised classification network to assess the efficacy of the proposed methodology.

3.2. CAG (Cross Attention and Graph Convolution)

W. Cai and Z. Wei in [14] document a novel cross-attention based feature weighing methodology in tandem with a graph convolution based algorithm for effective HSI data classification. Firstly, raw HSI data is dimensionally reduced using PCA to obtain a low-dimensional feature representation and to generate more expressive spectral features. Later, they propose a cross-attention technique which effectively assigns horizontal feature coefficients to every row of features and call it horizontal attention mechanism. Similarly, every column in the feature space is assigned vertical feature coefficients which is called vertical attention mechanism in their work. However, to broaden the differences between horizontal and vertical feature coefficients, two techniques have been proposed : a weight multiplication strategy and maximum weight matching strategy, which is clearly discussed in [14]. The resulting deep features extracted are finally used as an input to a graph convolution-based classification methodology for hyperspectral effective data classification on three datasets namely, Indian Pines, Pavia University and Salinas.

3.3. RSSAN (Residual Spatial Spectral Attention)

M. Zhu et. al. in [15] present an end-to-end residual spatial-spectral attention network for HSI data classification. When studying and constructing applications around high dimensional data, such as HSI data, dimensionality reduction is one of the most prevalent pre-processing stages mentioned in the literature. However, no such DR technique is included in this work. First, the authors put together a spectral attention module which prioritizes and picks spectral bands which are useful by building a spectral attention map around them. This is followed by a spatial attention module designed to adaptively emphasize data points from the same class as the center pixel, in a windowed neighborhood. The novel contribution of this work is how they introduced this spatial-spectral attention module in a 3D-CNN based ResNet block to accelerate training and avoid overfitting. The features thus constructed are input to a (1×1) convolution block and softmax activation for effective HSI data classification as documented in [15].

4. EXPERIMENTAL RESULTS

The efficacy of our approach, HYPER-VIT, is elaborately analysed in this section and compared against three other state-of-the-art attention based HSI data classification frameworks, namely, SPEC-HYP-ATTN, CAG and RSSAN. All the experiments were conducted on two well known remote sensing datasets, Pavia University and Salinas [16]. Pavia university dataset was captures with the ROSIS sensor by flying the sensor over The University of Pavia in Northern Italy. The dataset consists of 103 spectral bands with a spatial dimension of (610×340) consisting of 9 ground-truth classes in total. The second dataset, Salinas, was captured with AVIRIS sensor and has a spatial dimension of (512×217) and 224 spectral bands. However, 20 of those spectral bands have been detected as water absorption bands and are removed from the dataset and the number of ground-truth classes in this dataset is observed to be 16 in number [16].

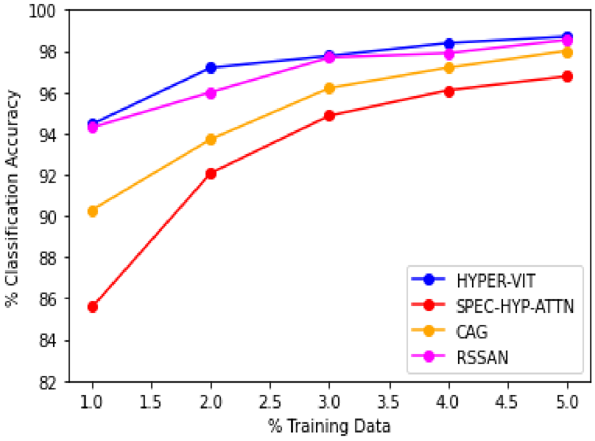


Fig. 3. Overall classification accuracies of Pavia University dataset for varying size of training samples.

In our work, all the parameters that have been used during the phase of experimentation have been empirically set to produce optimal results. Additionally, to avoid bias usually caused due to random sampling of data points, all the results documented are average over three trials. Thus, the value of n is set to 9 and 16 for Pavia University and Salinas datasets respectively. P, which denotes the reduced spectral dimensional space after PCA-based DR, is set to 10 for both the datasets since the first 10 principal components successfully retained more than 99% of variance in the data. All the experiments related to HYPER-VIT use sparse categorical cross-entropy as the objective function with a learning rate of 0.001. A batch size of 64 was chosen empirically and the HYPER-VIT classification network was trained for 60 epochs for both the datasets on Google Colaboratory notebook with a Tesla K80 GPU.

When compared to other attention-based classification methodologies discussed in this work, such as, SPEC-HYP-ATTN, CAG, and, RSSAN, our proposed technique HYPER-VIT produced more coherent classification regions with very few misclassifications as seen in the classification maps documented in Figure 2. It is also clearly visible from Figure 3 and 4, that our proposed vision transformer based HSI classification approach overpowered all the other methodologies in comparison, in terms of classification efficacy when the ratio of training data is spanned from 1% through 5%. The overall execution time of the approaches presented in our study for both the discussed datasets is tabulated in Table 1 for 2% training data. Further, Table 2 presents the overall execution time of proposed approach HYPER-VIT on varying training sizes and effectively proves that our proposed approach superior classification at a reasonable trade-off between computational time and classification performance compared to

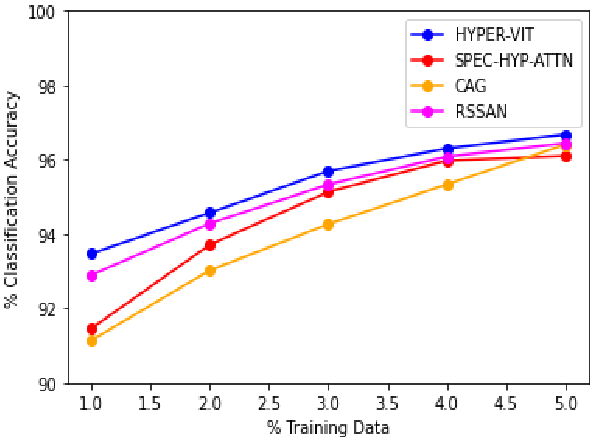


Fig. 4. Overall classification accuracies of Salinas dataset for varying size of training samples.

Table 1. Overall execution time (in minutes) of all the methodologies in comparison for 2% training data.

Dataset	HYPER-VIT	SPEC-HYP-ATTN	CAG	RSSAN
Pavia University	3.66	8.86	9.32	13.26
Salinas	3.91	15.22	17.84	18.33

Table 2. Overall execution time (in minutes) of HYPER-VIT for various training-testing data ratios.

Dataset	1%	2%	3%	4%	5%
	Training	Training	Training	Training	Training
Pavia University	2.84	3.66	5.00	5.89	6.90
Salinas	3.00	3.91	4.92	6.31	7.31

other classification techniques discussed, thus proving the lightweightedness of our HYPER-VIT.

5. CONCLUSION

In this work, a novel vision transformer based HSI data classification approach introduced. Compared to the traditional attention based HSI data analysis and classification methodologies in literature, inclusion of a multi-head attention module, which is the principal component of a visual transformer aids in selectively emphasizing important features present in hyperspectral data, at the same time, suppresses information that is of lesser importance, thus helping the classification framework to produce superior results. When compared to other frameworks discussed, the proposed HSI data classification and analysis framework HYPER-VIT yields outstanding classification performance while being robust under limited training sample scenarios, paving the way for new research avenues in hyperspectral remote sensing.

6. REFERENCES

[1] V. Kumar and J.K. Ghosh, "Camouflage detection using mwir hyperspectral images," *Journal of the Indian Society of Remote Sensing*, vol. 45, no. 1, pp. 139–145, 2017.

[2] Abbate Giulia, L. Fiumi, C. De Lorenzo, and Ruxandra Vintila, "Evaluation of remote sensing data for urban planning. applicative examples by means of multispectral and hyperspectral data," *GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*, pp. 201–205, 2003.

[3] Vakil I. Mohammad, Dalila B. Megherbi, and John A. Malas, "An efficient multi-stage hyper-spectral aerial image registration technique in the presence of differential spatial and temporal sensor uncertainty with application to large critical infrastructures and key resources (cikr) surveillance," *IEEE International Symposium on Technologies for Homeland Security (HST)*, pp. 1–6, 2015.

[4] Zhou Kai, Tao Cheng, Xinqiang Deng, Xia Yao, Yongchao Tian, Yan Zhu, and Weixing Cao, "Assessment of spectral variation between rice canopy components using spectral feature analysis of near-ground hyperspectral imaging data," *8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pp. 1–4, 2016.

[5] Wickert M. Lori, Jeanne B. Percival, William A. Morris, and Jeff R. Harris, "Xrd and infrared spectroscopic validation of weathering surfaces from ultramafic and mafic lithologies examined using hyperspectral imagery, cross lake area, cape smith belt, northern quebec, canada," *IEEE International Geoscience and Remote Sensing Symposium*, vol. 3, pp. III–362, 2008.

[6] Chen Yushi, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected topics in applied earth observations and remote sensing*, vol. 7, no. 6, pp. 2094–2107, 2014.

[7] Makantasis Konstantinos, Konstantinos Karantzalos, Anastasios Doulamis, and Konstantinos Loupos, "Deep learning-based man-made object detection from hyperspectral data," *Springer*, pp. 717–727, 2015.

[8] Bishwas Praveen and Vineetha Menon, "A study of spatial-spectral feature extraction frameworks with 3d convolutional neural network for robust hyperspectral imagery classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020.

[9] Bishwas Praveen and Vineetha Menon, "Novel deep-learning-based spatial-spectral feature extraction for hyperspectral remote sensing applications," *IEEE International Conference on Big Data (Big Data)*, pp. 5444–5452, 2019.

[10] Bishwas Praveen and Vineetha Menon, "A bidirectional deep-learning-based spectral attention mechanism for hyperspectral data classification," *Remote Sensing*, vol. 14, no. 1, pp. 217, 2022.

[11] Jolliffe and T. Ian, "Principal components in regression analysis," *Springer, New York*, pp. 129–155, 1986.

[12] Vaswani Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[13] Fine L Terrence, "Feedforward neural network methodology," *Springer Science and Business Media*, 2006.

[14] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[15] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 449–462, 2020.

[16] Gamba Paolo, "A collection of data for urban area characterization," *IEEE International Geoscience and Remote Sensing Symposium*, vol. 1, 2004.