SpecAL: Towards Active Learning for Semantic Segmentation of Hyperspectral Imagery *

Aneesh Rangnekar[0000−0002−0079−9495], Emmett Ientilucci, Christopher Kanan, and Matthew Hoffman

Rochester Institute of Technology, Rochester, NY, USA
aneesh.rangnekar@mail.rit.edu, emmett@cis.rit.edu, kanan@rit.edu, mjhsma@rit.edu

Abstract. We investigate active learning towards applied hyperspectral image analysis for semantic segmentation. Active learning stems from initially training on a limited data budget and then gradually querying for additional sets of labeled examples to enrich the overall data distribution and help neural networks increase their task performance. This approach works in favor of the remote sensing tasks, including hyperspectral imagery analysis, where labeling can be intensive and time-consuming as the sensor angle, configured parameters, and atmospheric conditions fluctuate.

In this paper, we tackle active learning for semantic segmentation using the AeroRIT dataset on three fronts - data utilization, neural network design, and formulation of the cost function (also known as acquisition factor, uncertainty estimator). Specifically, we extend the batch ensembles method to semantic segmentation for creating efficient network ensembles to estimate the network's uncertainty as the acquisition factor for querying new sets of images. Our approach reduces the data labeling requirement and achieves competitive performance on the AeroRIT dataset by using only 30% of the entire training data.

Keywords: hyperspectral · active learning · segmentation

1 Introduction

There has been significant development in designing deep neural networks for semantic segmentation across multiple computer vision applications. Most of those networks benefit from large amounts of data [10, 1, 3, 9]. This additional data load comes with an overhead cost of pixel-level annotations that increases exponentially with the number of samples. For example, the AeroRIT semantic segmentation dataset (1973×3975) required around 50 hours of manual labeling and multiple review rounds to ensure a good quality release.

* supported by Dynamic Data Driven Applications Systems Program, Air Force Office of Scientific Research under Grant FA9550-19-1-0021 and Nvidia GPU Grant Program

Multiple branches of machine learning deal with reducing the need for a large number of labeled examples (for example, semi-supervised learning, weakly-supervised learning, and active learning). Active learning is an approach that has gained significant traction to reduce dependency on large amounts of data for semantic segmentation ([12, 14, 8]). This approach focuses on tracking the most informative samples within the unlabeled data pool to add to the labeling pool via a scoring mechanism, most commonly an estimation of the network's uncertainty. This accumulation continues for multiple active learning cycles until one of the two conditions are met: 1) the network under question achieves a preset performance budget (typically, 95% of the entire data performance), or 2) the data labeling budget gets exhausted.

This paper explores the ability of neural networks to capture the information within hyperspectral signatures and function in an active learning data-training framework: we report competitive performance by utilizing significantly less labeled data, at par with fully utilizing the labeled data, which can be achieved under proper training conditions. For our analysis, we use the baseline network provided in the AeroRIT dataset [9] and make our increments to reduce the labeled data requirement by 70%.

2 Related Work

Kendall formulated segmentation as an approximate Bayesian interpretation by applying dropout at selective layers of their architecture to formulate test-time ensembles [4]. Lakshminarayanan *et al.* showed that training multiple iterations of the same network with random initializations acts as ensembles [5], and Wen *et al.* proposed to use multiple rank-1 matrices along with a shared weight matrix to form batch ensembles as an alternative to existing methods [13]. Rangnekar *et al.* studied the effects of using these approaches for uncertainty quantification on hyperspectral imagery by training on the AeroRIT dataset and found that applying dropout during test time gives the most definitive results [7]. In our paper, we build on this approach for uncertainty quantification as the active learning acquisition factor in the learning pipeline.

3 Methodology

The objective of this paper is to get an understanding of how different components of an experiment design can function towards improving the scope of data requirements for hyperspectral semantic segmentation with limited data. We adopt the active learning approach, which works in the following manner: (1) Train on available labeled data, (2) Acquire additional labels by evaluating the network on the unlabeled data pool, (3) Add the freshly labeled data to the existing labeled data pool, and (4) Repeat (1) to (4) till convergence. Given this process, we focus on three essential adjustments: data utilization, neural network design, and acquisition factor.

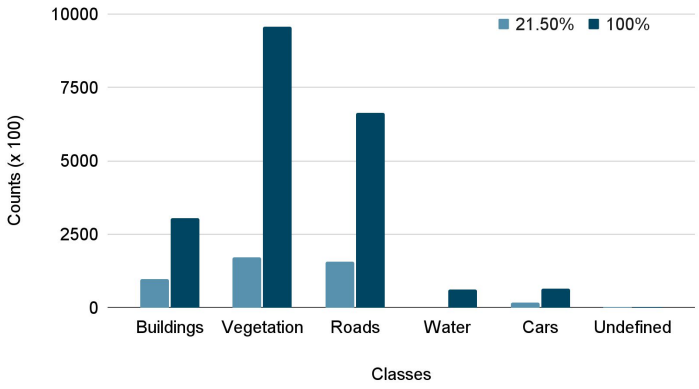


Fig. 1: The distribution of classes within the AeroRIT train set at 100% and our selected starting set with only 21.50% of the data.

3.1 Data utilization

We use the AeroRIT dataset as the reference for our learning pipeline [9]. The train set consists of 502 image patches of 64 $\times$ 64 resolution, post ignoring the overlap and ortho-rectification artifact patches used in the original studies presented in the paper. From this train set, we fixate on a small patch of the area as our starting point, as shown in Fig. 3a. This patch consists of 108 image patches, thus giving us a starting set of 21.5% of the dataset. We treat the rest of the image patches within the train set as the unlabeled pool of available data during the learning pipeline. Fig. 1 shows the statistics of each class within the respective sets, and we observe that the water class is not present in the initial labeled set. We hypothesize a well-trained network will have high uncertainty towards regions consisting of water and hence, do not consider it a concern. This would not be the case when dealing with standard gray-scale or color imagery, but the discriminative nature of hyperspectral imagery allows us to make this hypothesis.

We think it is unrealistic for any neural network to be able to achieve comparable performance to its fully labeled counterpart when using limited data. To this extent, we use data augmentation to increase the number of possible examples within the scene by applying random horizontal and vertical flips, with additive Gaussian noise in randomly selected spectral bands, random resizing, and CutOut [2]. This enables us to increase exposure to underlying data distribution and fully utilize the available data.

We increase the scope of learning further by increasing the learning schedule to account for relatively fewer data samples seen per training epoch. The network sees 502 samples per epoch on the fully labeled dataset to learn representations from scratch. Conversely, in the limited data regime, the network

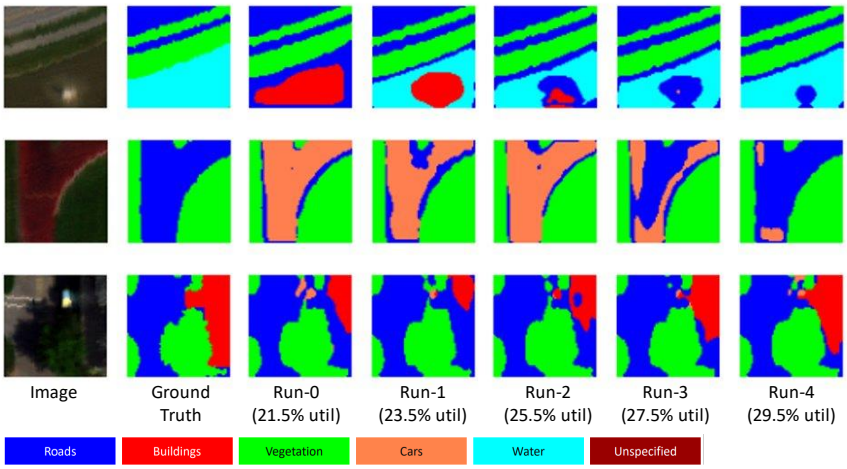


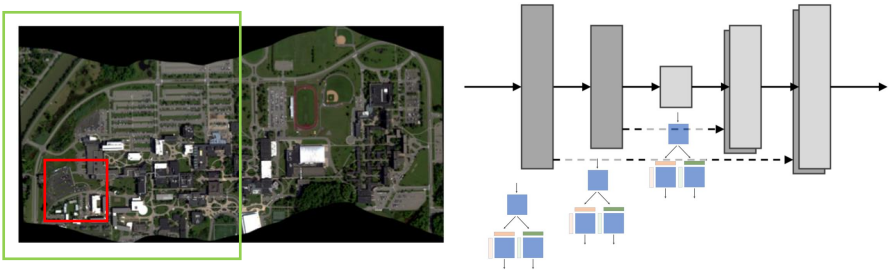
Fig. 2: Example predictions as a progressive acquisition step in the active learning pipeline. We observe a distinct improvement across all classes as the percentage of data used grows through the active learning cycles.

only sees 108 samples per epoch; hence, we do not expect it to learn at total capacity due to the nearly 1/5th reduction of samples per epoch. Hence, our first significant adjustment is to increase the number of samples seen per epoch during training in the limited data regime. We show in our experiments section that this dramatically improves the network's ability to learn representations from the limited set of available data.

3.2 Neural network design

We use the network described in [9, 7] for fair comparison. The network is based on the U-Net architecture and consists of two downsampling encoder blocks, followed by a bottleneck block, and two upsampling blocks before making the final prediction [10]. Our goal is to modify the network to express uncertainty, and previous studies have shown that ensembles work well for this purpose [4, 7]. The Monte-Carlo Dropout based approach is a clear choice for our task, where dropout is applied during test time for multiple network ensembles. However, dropout-based approaches have a shortcoming regarding reproducibility as the application is a function of the random probability distribution. Hence, we consider an alternative, simpler approach of Batch Ensembles.

Batch Ensembles (BE) work by sharing a tuple of trainable rank-1 matrices for every convolutional filter that is present in the neural network (Fig. 3b). The tuples act as the ensembles and, when combined with the filter weights, act as an



(a) The entire training set (Green) versus the starting set for active learning (Red).

(b) A graphic description of our U-Net with Batch Ensembles blocks.

Fig. 3

individual ensemble members. The network in a standard manner - each mini-batch is passed through the network, wherein it is split according to the number of ensembles (for example, a network with four tuples would split a mini-batch of 32 samples into eight samples per tuple). The core idea is that the convolutional filter weight acts as the shared weight amongst all randomly initialized rank-1 tuples, which learn their own set of representations. During the evaluation (inference) phase, every data point is replicated (for the above example, four times) and passed through the network to get an ensemble prediction, which is further averaged to obtain a final network prediction. In this approach, the tuple weights being fixed post learning ensure consistent predictions every time, unlike the dropout approach, while maintaining a low-cost solution to training individual ensemble instances like the Deep Ensembles approach.

Kendall *et al.* found in their study that applying dropout only to the bottleneck blocks of the encoder and decoder gave them optimum results [4]. With this motivation, we experiment with key areas to apply Batch Ensembles tuples to the convolutional filters within our network and experimentally found out the best and most consistent results were obtained by converting the convolutional filters in the encoder and bottleneck blocks to their ensemble counterparts. Our second significant adjustment is to convert a deterministic neural network into its light-weight ensemble version that can easily express uncertainty without further adjustments.

3.3 Acquisition Factor

Our goal now is to interpret the model's outputs as a function for querying a fixed budget of patches for labeling from the unlabeled pool of image patches. We simulate the process of querying for additional labels in reality by using the ground truth annotations already present for our dataset. These queried image patches are added to the labeled set for another cycle of learning. We experiment with four different approaches as acquisition factors: random (lower-bound), softmax confidence, softmax entropy, and softmax margin. We refer the

Table 1: Quantitative improvements over the baseline performance using a small U-Net on the AeroRIT dataset.

Training Scheme	Data Util.	Building	Vegetation	Roads	Water	Cars	mIoU
Baseline (1x data)	21.50%	61.92	93.39	74.06	0.00	31.39	51.70
5x data	21.50%	72.13	94.12	75.10	0.00	31.66	54.60
5x data (BE)	21.50%	77.00	94.50	76.99	0.00	28.85	55.47
5x data (BE)	29.50%	81.06	93.51	78.83	72.50	35.52	72.28
Oracle	100.0%	82.94	94.82	80.00	63.35	35.82	71.40

readers to [11] for an in-depth explanation of these factors and experimentally find softmax entropy as the best candidate for our approach.

We observe that the networks quickly become confident in their predictions during the training process. Typically, the network sees enough variance in the data to understand the minor differences between classes that may have similar signatures (for example, a black car and a black roof on a building). We observe this in Fig. 2 as the network progressively makes an understanding of the similarities and differences in the signatures with more labeled data. We account for this spurious leap in the network's prediction by penalizing the confident predictions [6]. This results in an elegant win for us as a byproduct of the penalty is higher entropy, which helps express uncertainty better. Our third significant adjustment is to combine confidence penalty regularization with softmax entropy as our acquisition factor.

4 Experiments and Results

4.1 Hyperparameters

We use 50 bands in the AeroRIT dataset, ranging from 400 to 890 nm, in this paper - 31 visible and 20 infrared bands. All chips are clipped to a maximum of 2^{14}, and normalized between 0 and 1 before forward passing through the networks. All networks are initialized with Kaiming initialization, and the rank-1 matrices for batch ensembles are initialized to have a mean of 1 and a standard deviation of 0.5 in accordance with the original paper. We use an initial learning rate of $1e^{-2}$, with drops of 0.1 at 50 and 90 epochs. We train all our networks for 120 epochs with standard cross-entropy loss and use confidence penalty for all limited data training instances. We will release the code post-publication.

For the active learning scenario, we start with an initial labeled set of 108 images (21.5% of the data) and iteratively query for 10 images (2%) every active learning cycle. We do not keep a preset data budget but instead strive to obtain performance comparable with the network trained on full data (502 images).

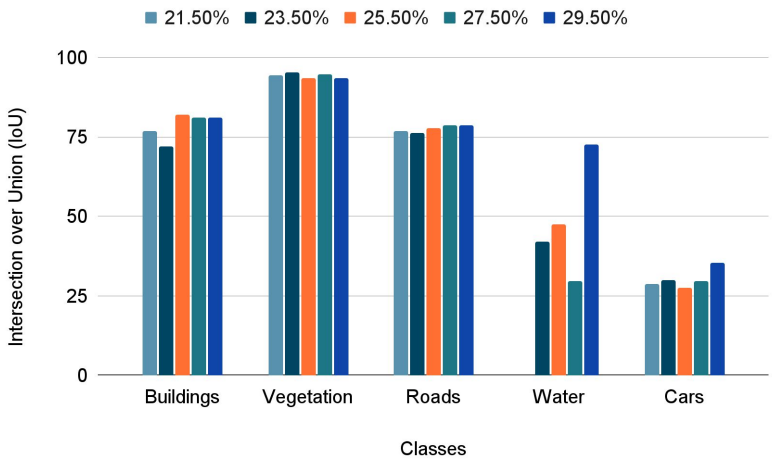


Fig. 4: Quantitative improvements of individual classes in the AeroRIT dataset as a function of active learning (data querying) cycles.

4.2 Results

Tab. 1 shows that the training performance dramatically benefits from increasing the learning schedule throughout the process. An increase in the number of samples per epoch (5x data) results in an improvement from a mIoU of 51.70 to 54.60, most significantly affecting the Building category during the data augmentation schemes. We also observe that shifting the model to its ensemble version (BE) further increases the mIoU by another point, yet again, mainly influencing the Building category that has two distinct white and black signatures throughout our dataset. BE also drops the IoU for the Car category by a few points, which is unexpected but is gradually over-comed through the active learning cycles (Fig. 4). Fig. 4 also shows an interesting trend in the Water category, we immediately observe a leap from 0 points in the IoU to roughly 45 points, before finally improving at the final cycle to 72.5 points and beating the performance of the fully supervised network. This could indicate (and warrants analyzing) wrongly labeled instances within the training set, which the network has successfully chosen to ignore during its learning process.

We observe that using a confidence penalty helps stabilize the performance and ensure reproducibility among various random initializations of the network. We run the entire framework through a rigorous evaluation scheme by further sampling 108 grid patches across random areas in the training set, ensuring that all classes follow the data distribution in Fig. 1. Surprisingly, the networks could reach similar performances in eight of the ten trials. We repeated the same set of experiments and enhanced our analysis with test-time data augmentation via random flips but did not obtain a reasonable difference in performance.

5 Conclusion

We present an approach for learning with limited data in hyperspectral imagery by leveraging the active learning framework. We can obtain performance at par with a fully supervised network using only 30% of the data budget. In closing, our next steps are to explore the domains of self-supervised learning to have a better-initialized network, which can also incorporate pseudo information from the unlabeled data to learn better representations.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
2. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
3. Kemker, R., Salvaggio, C., Kanan, C.: Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS journal of photogrammetry and remote sensing **145**, 60–77 (2018)
4. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680 (2015)
5. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
6. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:1701.06548 (2017)
7. Rangnekar, A., Ientilucci, E., Kanan, C., Hoffman, M.J.: Uncertainty estimation for semantic segmentation of hyperspectral imagery. In: International Conference on Dynamic Data Driven Application Systems. pp. 163–170. Springer (2020)
8. Rangnekar, A., Kanan, C., Hoffman, M.: Semantic segmentation with active semi-supervised learning. arXiv preprint arXiv:2203.10730 (2022)
9. Rangnekar, A., Mokashi, N., Ientilucci, E.J., Kanan, C., Hoffman, M.J.: Aerorit: A new scene for hyperspectral image analysis. IEEE Transactions on Geoscience and Remote Sensing **58**(11), 8116–8124 (2020)
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
11. Siddiqui, Y., Valentin, J., Nießner, M.: Viewal: Active learning with viewpoint entropy for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9433–9443 (2020)
12. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5972–5981 (2019)
13. Wen, Y., Tran, D., Ba, J.: Batchensemble: An alternative approach to efficient ensemble and lifelong learning (2020)
14. Xie, S., Feng, Z., Chen, Y., Sun, S., Ma, C., Song, M.: Deal: Difficulty-aware active learning for semantic segmentation. In: Proceedings of the Asian Conference on Computer Vision (2020)