Comparative Study Between Vision Transformer and EfficientNet on Marsh Grass Classification

Conrad Testagrose, Mehlam Shabbir, Braden Weaver, Xudong Liu
School of Computing
University of North Florida
Jacksonville, Florida, USA
{n01464016, n01474570, n00914984@, xudong.liu}@unf.edu

Abstract

Due to rapidly changing ecosystems, effective environmental protection often calls for the monitoring of the vegetation for any environmental changes. Vegetation monitoring is essential in assessing the changes and impacts to environmentally valuable ecosystems such as marshlands. While vegetation monitoring of marsh grasses is crucial to the maintenance and protection of marshlands, it is a tedious and time-consuming task that involves careful examination of individual pixels within large resolution images. In this study we compare the use of Vision Transformers (ViT) and two different EfficientNet models on automated marsh grass identification using the GTMNERR Marsh Grass Species data set. Our results show that the use of a ViT allowed for an increase in the accuracy of marsh grass identification. The Vision Transformer was also able to better distinguish between the 6 classes in the data set and provided competitive training time to the smaller of the two EfficientNet models tested in this study.

Introduction

Marshlands and coastal wetlands are among the most important and endangered ecosystems. Coastal wetlands and marshlands provide not only a nursery for a variety of flora and fauna but also coastal protection from storms and major contribution to environmental carbon sequestration (Barbier et al. 2011). These vitally important ecosystems are also easily disturbed by habitat loss as well as any rise in sea-levels (Warren and Niering 1993). In order to effectively preserve marshlands, vegetation monitoring is commonly implemented as a means to assess vegetation coverage and the vegetation composition within the marsh. The monitoring of vegetation coverage and composition can lend insights into the health of the marsh and the various inhabitants that rely on this vegetation. Such monitoring is commonly carried out in research preserves such as the Guana Tolomato Matanzas National Estuarine Research Reserve (GTMNERR).

Researchers at GTMNERR utilize high resolution cameras in their efforts to monitor vegetation density and composition (Bacopoulos, Tritinger, and Dix 2019). The images taken using these cameras are one meter-square and contain a variety of vegetated or non-vegetated regions of the marsh. Randomly selected snippets from each one meter-square image are then manually labeled by trained experts and volunteers. These trained individuals will label each snippet with the percentage of vegetation coverage; manually tallying the species present with one of 6 labels. Due to the high resolution image size as well as the manual tallying of species in each snippet of the high resolution image, this task can be tedious and time consuming. The automation of such time-consuming tasks has the possibility of saving both time and important resources. The automation of this manual identification of marsh species has previously been investigated using convolutional neural networks (CNN) (Welch et al. 2021; Welch and Liu 2021). While CNNs are often the algorithm of choice for object detection and image classification tasks (LeCun, Bengio, and Hinton 2015), a recently proposed network architecture known as a Vision Transformer (ViT) has shown to provide equivalent or greater performance at these tasks (Dosovitskiy et al. 2020). We are not aware of any research that has investigated the usage of the use of the ViT architecture as a solution to marsh species identification while also simultaneously comparing the findings to a CNN with comparable performance to a ViT. One such CNN that touts performance comparable to ViT is EfficientNet(Tan and Le 2019). As of this study, both EfficientNet and ViT architectures rank in the top 10 algorithms for Top 1 Accuracy on ImageNet classification (Krizhevsky, Sutskever, and Hinton 2012).

The goal of this study is to 1. Train both EfficientNet and ViT marsh species classification algorithms and 2. Compare the performance between both algorithms in their ability to classify marsh species within the GTMNERR data set. We compare algorithm performance using the accuracy, Area under the Receiver Operator Characteristic (AUROC), and the time required to train each algorithm in seconds per epoch for each of the models included in this study.

Methodology

GTMNERR Marsh Grass Species dataset

The GTMNERR Marsh Grass Species dataset (Guana Tolomato Matanzas National Estuarine Research Reserve, Ponte Vedra Beach, FL, https://gtmnerr.org/) contains 1 meter by 1 meter photoquadrats of marsh that contains various grass species. One example is given in Figure 1.

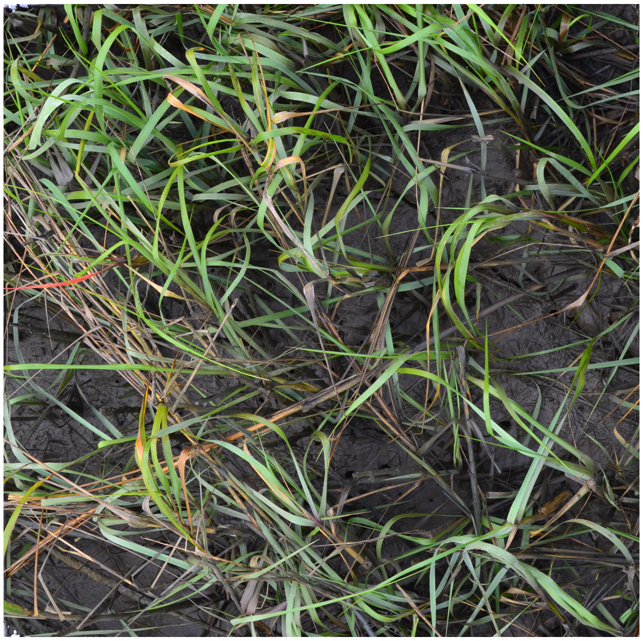

Figure 1: GTMNERR photoquadrat example

Welch et al. (Welch et al. 2021; Welch and Liu 2021) sampled 77,630 image snippets from these photoquadrats taken at the Guana Tolomato Matanzas marine sanctuary [1]. These snippets are each of size 33 x 33 and labeled with either Avicennia, Bare/Non-vegetated, Batis Maritima, Spartina, Juncus, and Sarcoconia. Figure 2 shows an example of each of the labels present in the data set. Images were matched with their labels and were indexed into a JSON file which could be used to load the images and refer to their corresponding label. Creation of 5 folds of the entire dataset of 77,630 image snippets was performed so that determination of any variation in algorithm performance could be observed.

Transfer Learning

Transfer learning is the process of transferring the weights of a model trained on a different more generalized data set to a model that will be trained on a more specific data set. The idea that most images all share basic common characteristics allows us to fine tune the weights of a transferred model to better perform on our more specific data set of images. We have implemented transfer learning for both the Vision Transformer and the EfficientNet convolutional neural networks we have implemented in this project. The models that we have transferred the weights from were both trained on the Imagenet (Krizhevsky, Sutskever, and Hinton 2012) data set which contains over 1000 classes of images making it a very general data set of images that we can then fine tune to classify the images of the marsh species.

EfficientNet

Convolutional Neural Networks(CNN) were developed due to the study of the brain's visual cortex, and have been around since the 1980s (Géron 2019). Convolutional layers are used to identify small patterns, with successive layers working to piece together larger and larger patterns. This approach works very well with images, as they often follow such hierarchical patterns (Géron 2019). Much research has been done in improving the performance of CNNs from the 1980s with LeNet (LeCun et al. 1989) to contemporary research such as EfficientNet.

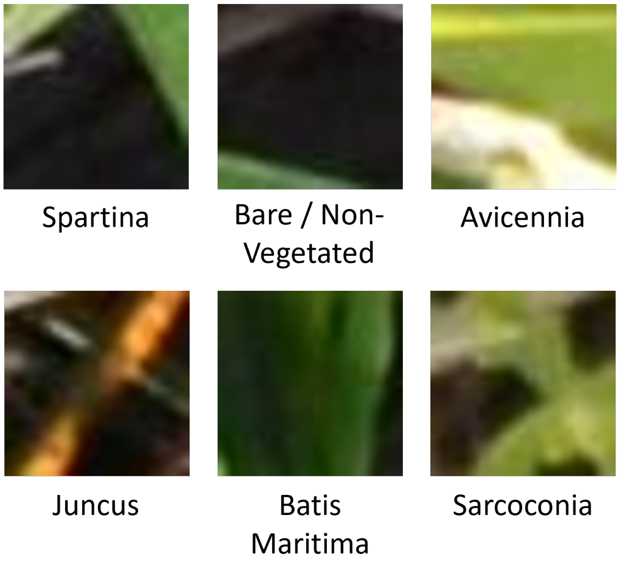

Figure 2: Examples of the different classes contained within the data set. Each snippet is randomly selected from the one meter-square images taken at the GTMNERR and labeled by a trained expert. All snippets are a size of 33 x 33.

EfficientNet which was proposed by Tan and Le (Tan and Le 2019) is a convolutional neural network that uniformly scales the dimensions of depth, width, and resolution using a compound coefficient. In their research they show that their approach is effective in scaling up MobileNets and ResNets to increase the accuracy. They then used their approach to design a baseline network and used their scaling approach to create the family of EfficientNet model. They show that this family of models is capable of achieving better accuracy and are more efficient than previously outlined ConvNets. The current list of available EfficientNet models included the models of efficientnet-b0 to efficientnet-b7 and there has been work done to add more models to the available list (Tan and Le 2021). They show that their EfficientNet-B7 model is over 8 times smaller and over 6 times faster than the best existing ConvNet on the ImageNet data set at the publishing of their article (Tan and Le 2019). We hope to compare the performance of two EfficientNet family models against the performance of a Vision Transformer.

Vision Transformer

The transformer architecture has more commonly been used in Natural Language Processing (NLP)(Vaswani et al. 2017). Transformers have replaced the recurrent neural network

[1] `http://unfail.ccec.unf.edu/marshdata.html`

(RNN) architectures that were used in NLP prior to the development of the transformer architecture. Transformers utilize a self-attention mechanism to form relationships between words in a sequence of text. Efforts to utilize transformers and their self-attention mechanism with computer vision tasks resulted in the development of the ViT. ViTs, outlined in (Dosovitskiy et al. 2020), are a newer deep learning architecture that utilizes the benefits of the transformer architecture for computing vision tasks. ViTs have been shown to provide comparative performance to CNNs, sometimes even outperforming them (Dosovitskiy et al. 2020). The high-level overview of the ViT architecture is presented in Figure 3. The challenge in using Transformers with image data is that text is one-dimensional while image data is often either 2 or 3 dimensional. In order to get image data in a format that can be use with the transformer architecture, ViTs begin by splitting an input image up into patches. After the patches have been created, they are linearly flattened, resulting in each patch being one-dimensional. Each patch has its original position embedded with the first patch also having an extra learnable class embedding attached to it. These patches are then fed to a transformer encoder before being processed by a Multi-Layer Perceptron for the resulting classification output. Using self-attention, ViTs have the ability to form relationships between the patches in images. Utilizing this self-attention mechanism, ViTs have shown to achieve high levels of accuracy on large general data sets such as ImageNet (Dosovitskiy et al. 2020). We aim to determine if the use of a ViT over EfficientNet for marsh species identification will result in higher levels of performance regarding accuracy. We also want to explore the possible implications to the time to train a ViT over EfficientNet.

Experimentation

All experimentation was carried out using a system with a Intel i7-9700k, Nvidia RTX 2070 Super with 8GB of VRAM, and 16 GB of system RAM. Each model of ViT (source code available at: https://github.com/lukemelas/PyTorch-Pretrained-ViT.git), EfficientNet-b0 and EfficientNet-b7 (source code available at: https://github.com/lukemelas/EfficientNet-PyTorch.git) were trained using 5 fold cross validation. Each algorithm used a learning rate of 0.001 and Stochastic Gradient Descent(SGD) with momentum of 0.9. The code used asks for the image size and number of classes when each model is initialized. By passing these two parameters, the model output layer is modified accordingly. Following training, the trained models were then evaluated on the respective test set for each of the 5 folds to obtain the respective accuracy, weighted F1, and the AUROC for each class. AUROC curves, or Receiving Operator Characteristic Area Under the Curve plots, are a means to compare the ability of each model to differentiate between the classes. In situations of multi-class classification, the AUROC is calculated per class on a one-vs-rest basis. The greater the area under the curve (AUROC) for a class the better the model is at distinguishing that particular class to the rest of the classes. We can compare the accuracy, weighted F1, and the AUROCs to more intimately compare both the ViT and the EfficientNet models against each other on marsh species identification. We also examined the time it took to train each of the implemented models to determine the possible implications that the different model architectures can have on training time.

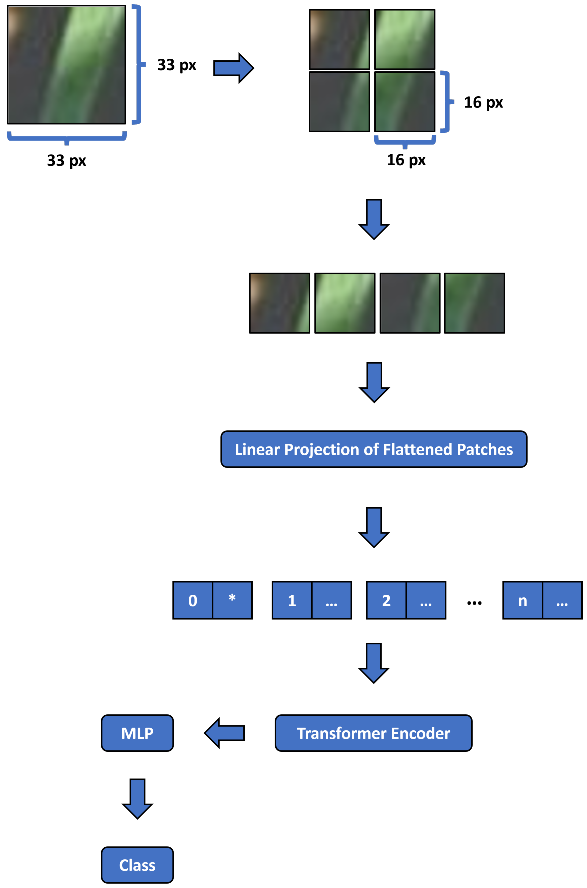

Figure 3: High level overview of the ViT architecture. Images are patched into smaller individual patches. The smaller patches are then linearly flattened and fed to a transformer encoder. The output of the transformer encoder is then fed to an MLP to obtain the classification result. In the case that image resolution is not divisible by the number of patches, the image is resized accordingly.

Results

Table 1 shows the results for the difference between the average class AUROC for each of the models used in this study. The AUROC can help us determine how well each model is at distinguishing between each class by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). By recording the AUROC across 5-folds of the dataset for both EfficientNet-b0 and EfficientNet-b7, we can determine which EfficientNet is performing better in our particular classification task. We can then determine which EfficientNet model is performing more optimally and compare this model to ViT. On average, EfficientNet-b0 provided higher

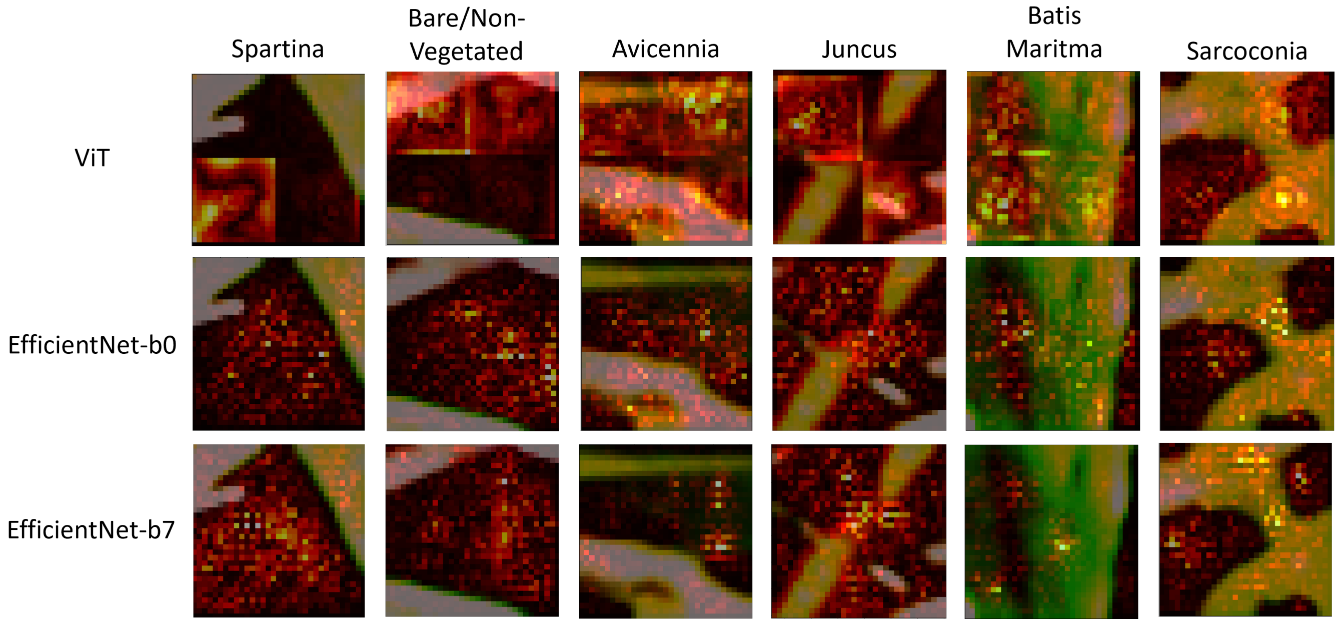

Figure 4: Saliency maps for ViT (top), EfficientNet-b0 (middle), and EfficientNet-b7 (bottom). Saliency maps show which regions of an image are helping to guide the network's decision on the classification of the image.

AUROC results than EfficientNet-b7 on classes a, m, j and p. EfficientNet-b7 did outperform EfficientNet-b0 regarding the AUROC for classes b and s. When comparing the AUROC for the slightly better performing EfficientNet-b0 and a ViT, we see that the ViT outperformed EfficientNet-b0 on all 6 classes contained within the data set.

Table 1: Average Area Under the Curve

Class	EfficientNet-b0	EfficientNet-b7	ViT
Avicennia (a)	0.926	0.916	0.944
Bare/Non-Vegetated (b)	0.912	0.922	0.948
Batis Maritima (m)	0.962	0.954	0.976
Spartina (s)	0.862	0.874	0.920
Juncus (j)	0.840	0.780	0.896
Sarcoconia (p)	0.866	0.856	0.930

Table 2: Average Accuracy and Seconds per Epoch

Metric	EfficientNet-b0	EfficientNet-b7	ViT
Accuracy	0.818	0.810	0.840
Weighted F1	0.808	0.792	0.830
Seconds/Epoch	59.17	110.20	65.46

Table 2 shows the results for the accuracy on the test set for each of the models on all five of the cross validation folds. On average, ViT outperformed both the EfficientNet variants in Accuracy and Weighted F1. Table 3 shows the p-values calculated using the Two-tailed T-Test function defined in Microsoft Excel. When comparing the metric results for the better performing EfficientNet-b0 and ViT, the p-values show that the increases to these metrics were significant. While EfficientNet-b0 did not outperform ViT, it did provide the fastest time to train on average. EfficientNet-b7 took the longest time to train most likely due to the increased size of the network. This increased time to train did not result in a better accuracy or better AUROC.

Table 3: ViT vs. EfficientNet-b0 p-values

Metric	p-value
Accuracy	0.00039
Weighted F1	0.00039
AUROC Avicennia (a)	0.00084
AUROC Bare/Non-Vegetated (b)	0.00012
AUROC Batis Maritima (m)	0.02490
AUROC Spartina (s)	0.00010
AUROC Juncus (j)	0.00015
AUROC Sarcoconia (p)	0.00023

To help understand how each model was making its classification decision, we also visualized saliency maps. Saliency maps help show which regions of an image are aiding in the network's classification decision. Figure 4 shows examples of saliency maps for each model. By examining the saliency maps for each model we can gain insight into how each model is performing. The saliency maps for ViT show interesting behavior with some patches having high levels of activation while the other patches around it have reduced activation. We believe this to show that the ViT is lending more attention to particular patches when making its classification. The saliency maps for both EfficientNet-b0 and EfficientNet-b7 appear to be similar with saliency maps for EfficientNet-b0 appearing to show more widespread activation in some images.

Conclusion and Future Work

After running experimentation and reviewing the results, we can conclude that the Vision Transformer outperformed both of the EfficientNet models we trained on the GTMNERR marsh species data set. Although it provided slightly higher training times than the smaller EfficientNet-b0, it did offer a significant change to the average class AUROCs and to the average accuracy. This comparative study shows that the use of a vision transformer for the purpose of marsh species identification could in fact provide better overall accuracy to researchers looking to automate this process. Due to hardware constraints, more extreme adjustments to the ViT model architecture was not feasible and is a significant limitation of this study. Future research in this area should utilize high performance computing solutions to explore the impact of different image transforms or the increasing of the input image resolution or ViT patch size on the model performance.

References

Bacopoulos, P.; Tritinger, A. S.; and Dix, N. G. 2019. Sea-level rise impact on salt marsh sustainability and migration for a subtropical estuary: Gtmnerr (guana tolomato matanzas national estuarine research reserve). *Environmental Modeling & Assessment* 24(2):163–184.

Barbier, E. B.; Hacker, S. D.; Kennedy, C.; Koch, E. W.; Stier, A. C.; and Silliman, B. R. 2011. The value of estuarine and coastal ecosystem services. *Ecological monographs* 81(2):169–193.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Géron, A. 2019. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.".

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In Pereira, F.; Burges, C.; Bottou, L.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521(7553):436–444.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1(4):541–551.

Tan, M., and Le, Q. V. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks.

Tan, M., and Le, Q. V. 2021. Efficientnetv2: Smaller models and faster training.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Warren, R. S., and Niering, W. A. 1993. Vegetation change on a northeast tidal marsh: Interaction of sea-level rise and marsh accretion. *Ecology* 74(1):96–103.

Welch, L., and Liu, X. 2021. Measuring vegetation density in marsh grass photographs using deep neural networks (student abstract). In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI).* AAAI Press.

Welch, L.; Liu, X.; Reddivari, S.; Umapathy, K.; and Kahanda, I. 2021. Vegetation coverage in marsh grass photography using convolutional neural networks. In *Proceedings of the 34th International Florida Artificial Intelligence Research Society Conference (FLAIRS).* University of Florida Press.