Comparison of CNN to Vision Transformer Model Trained on HiRISE Mars Satellite Images

Aniruddha Prasad, Andrew Hartnett
Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA, USA

Abstract Since the introduction of the Transformer architecture for image classification tasks in 2020, they have replaced the convolutional neural network (CNN) for the state-of-the-art when pre-trained on significantly large datasets. In this work, we revisit a classification task conducted by Wagstaff et al. using the High Resolution Imaging Experiment (HiRISE) dataset of Mars satellite images. The task involved training a model on images of the Earth and conducting transfer learning to determine accuracy when classifying terrain images of Mars. As the discovery of Transformers in image classification was after this paper, we seek to test the CNN and Transformer model against each other for a similar task on the same dataset. This will allow us to determine whether using the Transformer architecture could provide an improvement in accuracy if the transfer learning task were to be revisited.

***Index Terms*— Transformer, Convolutional Neural Network (CNN)**

I. INTRODUCTION

The HiRISE dataset is a publicly available collection of satellite images taken by the Mars Reconnaissance Orbiter in November of 2017 [1]. A paper by Wagstaff et al., "Deep Mars: CNN Classification of Mars Imagery for the PDS Imaging Atlas.", was created based on this data to handle an image classification task involving transfer learning Earth terrain information onto the Mars terrain [2]. With this paper being published in 2018, "Deep Mars" utilized the state-of-the-art image classification model of the time, the Convolutional Neural Network (CNN).

CNNs are built off of traditional Artificial Neural Networks (ANNs), in that they are made up of neurons that self optimize through learning. Each neuron receives an input and performs an operation just like an ANN. The only real difference is CNNs are primarily used in the field of pattern recognition within images allowing for encoding of image specific features into the architecture, making the network more image focused. Although some simple image datasets, such as the MNIST handwritten digits dataset, can be processed by an ANN, for more complex images with colors and other features, the CNN is most effective [3].

In 2020, the introduction of the Transformer architecture by Dosovitskiy et al. as the new state-of-the-art for image classification tasks opened the door for significant innovation [4][5]. It showed that when pre-trained on a large enough dataset, the Transformer could benefit from accuracy and computing power from fine-tuning onto smaller datasets for more specialized tasks.

This work seeks to implement the newly discovered application of the Transformer architecture onto classification of the HiRISE dataset. In this paper, we train a custom CNN and a pre-trained Transformer model fine-tuned on the Mars satellite images and compare their training time and accuracy. Then, to analyze the benefits of the Transformer, we conduct the same experiment again, this time with pre-training on data sets of three different sizes. We compare the results of the two custom models and determine whether the Transformer could be considered a replacement to the CNN for this task, as well as provide our inductive reasoning as for why one would be better suited than the other.

II. BACKGROUND

The following subsections will explore the original "Deep Mars" paper training a CNN on the HiRISE dataset [2]. After this, the Transformer architecture will be introduced with its preliminary usage in image classification.

A. "Deep Mars" Analysis

Although there are several papers that use the Mars HiRISE Dataset for their own research applications, the one that has done a significant amount of work and lines up with the work that is being done in this paper, as mentioned above, is the "Deep Mars" paper. Within this paper the researchers used the AlexNet CNN trained on images of the Earth. They then applied transfer learning to the algorithm to adapt it to work with the Mars HiRISE images. The original AlexNet CNN was trained on 1.2 million images of the earth with 1000 classes The network was adapted by removing the final fully connected layer and re-defining the output classes and then re-training it with Caffe.

To train their CNN named HiRISENET, they first split the HiRISE images into training, validation and test sets. They however split the sets by the HiRISE source image identifier to ensure that the landmarks from the same source images did not appear in more than one data set. Once the algorithm is trained, the researchers observed a 90.3% accuracy on the

validation set and a 94.5% accuracy on the test set [2]. The figure below shows the accuracy of the neural network as a function of the confidence threshold.

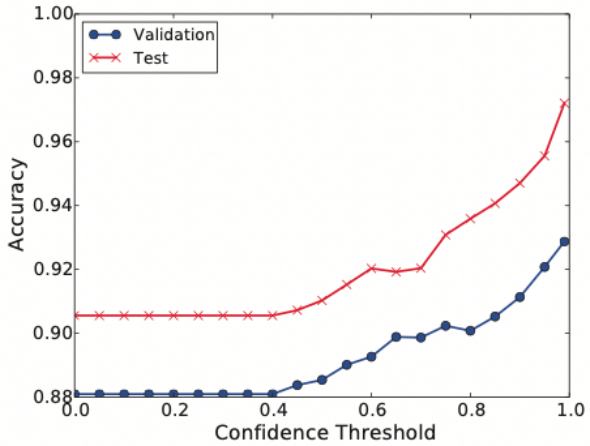

Figure 1: HiRISENET image classification performance as a function of confidence threshold (from [2]).

An important observation that was made, that will have to be a consideration within the work done in paper, is the fact that the most common errors were in a small set of images that were labeled "bright dune" and "other". The images with the label "other" were classified as "edge" and the images labeled as "bright dune" were classified as "other". They believe this is because of significant conceptual overlap between the classes.

B. Transformers

Having the idea to propose a new image classification model that could outperform the CNN, the team at Google Brain proposed the application of the Transformer architecture in October 2020 [4]. This architecture was primarily used in natural language processing (NLP) tasks prior to this work, but its focus on self-attention made it a possible candidate for image classification.

A standard Transformer can be broken into its encoder and decoder blocks. The encoding portion is originally built with a multiheaded attention block followed by a multilayer perceptron (MLP). While the MLP behaves as other networks we have seen with trainable weights and parameters, the multi-headed attention block is unique to the Transformer. Attention, in the context of Transformers, refers to the piece of an input sequence that is used more heavily to help predict an output sequence. Many sequential models without attention blocks will innately put greater significance on inputs towards the end of a sequence. This can be detrimental to an NLP model which needs to translate a paragraph with important contextual information at the beginning. This information may end up being less influential on the output without including self-attention into the model [6]. The other half of the Transformer, the decoder, functions similarly to create a sequence of outputs.

The Vision Transformer was developed using the encoder half of the standard Transformer and feeding its output directly into an MLP "head", or softmax layer for classification. The MLP head proves to be significant for fine-tuning towards a smaller model. Pre-training can take significant time and resources but will determine weights within the encoder that will apply to a variety of image classification tasks. The MLP head can be removed and replaced with a fresh layer and retrained on a new smaller dataset. This fine-tunes the final weights towards a specific task [4].

III. HYPOTHESIS

To better investigate an updated Transformer model in place of the CNN used in "Deep Mars", we seek to train each of these models on a subset of the HiRISE dataset and compare their accuracy and training time. In the following subsections, we begin with a compare and contrast of the two architectures, followed by an analysis of the HiRISE dataset. We then state our expectations based on the explored factors of the experiment.

A. CNNs vs Transformers

CNNs have been a standard within the industry for their image classification applications and their ease of use. They have been documented to be models that perform well on the standard benchmarking datasets such as the MNIST Handwritten digits dataset. However the new state-of-the-art transformers now pose a threat to CNNs in performance and accuracy. They only issue being that transformers require a large pre-training dataset size to actually show any improvements over CNNs. CNNs still excel over transformers when a small or medium sized dataset is available. However transformers are more efficient and win over CNNs in training time since most of the time is spent in the pre-training.

B. HiRISE Dataset Review

As mentioned previously, the data set used for this paper is obtained from the High Resolution Imaging Experiment called HiRISE. This satellite camera has the task of taking pictures of vast areas of the Martian terrain while also having a resolution of approximately 1 meter. This camera operates in visible wavelengths, hence producing images that are accurate to how the human eye would see them. The images themselves capture a massive scale of the Martian Terrain with a size of 6km x 60km.

This particular dataset contains a total of 73,031 landmarks. 10,433 landmarks were detected and extracted and 62,598 landmarks were augmented from the original 10,433 landmarks. Each cropped landmark is resided to 227x227 pixels and is then augmented to general 6 additional

landmarks [1]. The 6 general landmarks are listed in the table below.

Class Number	Label Name
0	other
1	crater
2	dark dune
3	slope streak
4	bright dune
5	impact ejecta
6	swiss cheese
7	spider

Table 1: HiRISE dataset classes (from [1]).

C. Expectation for HiRISE

Due to the explosion of Transformers used as state-of-the-art image classification models, we expect to see the ViT model to outperform the CNN provided a large enough pre-training dataset. Besides seeing a significant improvement in the accuracy when a ViT is used rather than a CNN, we also expect the ViT to be more computationally efficient. However, we do also recognize the available pre-training datasets for the Transformer. If not provided a large enough dataset, it is likely that the CNN will outperform the ViT.

IV. EXPERIMENTS

The CNN and Transformer code were run in a Jupyter Notebook environment using Python 3.7. External libraries and packages were all standard to previous use in the ECE 697 ML course, with the exception of the HugsVision Vision Transformer wrapper [7]. All imported libraries are listed in Appendix A.

The following subsections will explain data preprocessing that occurred, as well as notes regarding training the two models individually. Then, results between the most efficient CNN and ViT models will be compared. Finally, we will look at how the ViT model changes based on pre-training dataset size and batch size.

A. Preparing the Models

Before the models were developed, the data had to be preprocessed at first. With the dataset of images, a text file containing the image file names and their respective labels is included. So using this, the labels for each image had to be assigned to the image file itself. The images are then normalized to a 0-1 range, and converted to numpy matrices. They are then split to training and test sets and a fourth dimension is added to allow the CNN and Transformer to be able to fit and train the data without any errors. One-hot encoding is then applied to the labels to optimize them for the Machine Learning algorithms. [8] was used as reference for the preprocessing of the data.

B. Training on HiRISE - CNN

The CNN is comprised of seven layers. The first four are made of two 2D convolutional layers each paired with a max pooling layer to down sample the features. They are followed by a flattening layer, and two dense layers. All layers except the last output layer have relu activation. The last layer has a softmax transformation as it is the final layer for classification. The CNN has an ‘adam’ optimizer and a categorical cross-entropy loss function as this is a multiclass classification problem. The metric that the CNN is evaluated on is accuracy. The model was prepared using the code developed by GitHub user Niehusst [8]

To further tune the performance of the CNN, initially a grid search with 5 fold cross validation was attempted. The grid search aimed to find the optimal number of epochs for the CNN while using precision score and accuracy score as metrics. However due to the structure of the CNN, the grid search attempt was met with an error code stating “*can't pickle _thread._local objects”*. Hence it was abandoned in favor of a more brute force method which involved training and testing the CNN for multiple epochs.

The CNN model has a batch size of 32. This batch size was chosen for low computational load and high speed while not compromising the training. It is then trained on 11 epochs and the training accuracy is observed. As expected the training accuracy for the model significantly increases with every epoch with a 99% training accuracy on 11 epochs. This however is not an indication of the model performing well on a test set because an accuracy this high is almost always an indication of overfitting. This is further confirmed by looking at the accuracy on the test set. When the model is trained on 5 epochs, the test accuracy is approximately 80% however, when trained on 11 epochs, that training accuracy drops down to 78%. Indicating that 5 epochs is good enough and higher number of epochs lead to overfitting.

C. Training on HiRISE - Transformer

Vision Transformer code was provided by Yanis Labrak, a contributor to the HugsVision Vision Transformer wrapper [7][9]. The portion of the code modified by this project was the pre-training datasets and batch size, as these were of the few available parameters to tweak.

The four variations of Transformers were trained to compare to the CNNs. The variations can be separated based on the pre-training model they took their weights from, as well as the batch size used in training. The names used to refer to each Transformer and their parameters can be found in Table 2 below.

Transformer Name	Pre-Training Dataset	Batch Size
ViT/12	ImageNet-21k, FT 2012	12
ViT/12-21k	ImageNet-21k	12
ViT/16	ImageNet-21k, FT 2012	16
ViT/16-21k	ImageNet-21k	16

Table 2 - Pre-training dataset and batch size used for four ViT variants. "FT 2012" means the model was later fine-tuned on ImageNet 2012.

The pre-training datasets used were ImageNet-21k and a variant of the same dataset, except later fine-tuned on the ImageNet 2012 [7]. This pre-training determines the starting weights found within the ViT encoder. ImageNet 2012 and ImageNet-21k contain 1.3M and 14M images, respectively [10]. They were chosen due to availability with the HugsVision wrapper used for training the Transformers [8]. Each model was trained on the combination of map-proj/ and map-proj-v3/ images in the HiRISE dataset with data augmentation. This created a total of 8886 images for the training set and 1569 for the test set.

D. Results

This section should analyze any results we get from evaluating the models trained in the previous sections. We should include plots and graphs that show either accuracy or training time between the two models. Compare this to our expected outcome and try to explain why we are seeing the output we are. If the output is unexpected, say why it goes against the expectation. This ideally will be a lengthy section.

The optimal number of epochs to train the CNN for is 5 epochs, stated previously. The resulting classification report for the test set is displayed below.

Classes	precision	recall	f1-score	support
0	0.76	0.89	0.82	301
1	0.70	0.39	0.50	59
2	0.73	0.78	0.75	63
3	0.00	0.00	0.00	4
4	0.67	0.27	0.39	22
5	0.00	0.00	0.00	0
6	0.93	0.83	0.88	124
micro avg	0.79	0.79	0.79	573
macro avg	0.54	0.45	0.48	573
weighted avg	0.78	0.79	0.77	573
samples avg	0.79	0.79	0.79	573

Table 3 - F1 score and precision/recall values for each class in the CNN.

An interesting observation within this classification report is the lack of any information for classes 3 and 5. This means that the CNN is predicting the existence of samples with 3 and 5 within the test set of images. However this lack of scores means that the true labels for the test set did not contain any images of classes 3 or 5. This means that the CNN is predicting classes that do not exist and is getting confused. It may be because these classes have some attributes that are very similar to other classes and it may be mistaking those for classes 3 or 5. A way to fix this may be to increase the complexity of the layers within the CNN so the image is broken down into further features and hence has more information. This would allow it to make better distinctions between similar images.

However besides this, the CNN has acceptable performance on the rest of the dataset with favorable metrics. There is room for improvement, especially seen in the poor F1 scores seen from class 1 and 4.

The best performing ViT model, ViT/12-21k, is able to produce an accuracy of 96.88% on the test set. This shows a significant outperformance of the CNN, which produced a ~79% accuracy on its set of test images. The confusion matrix for ViT/16-21k is shown in Figure 2.

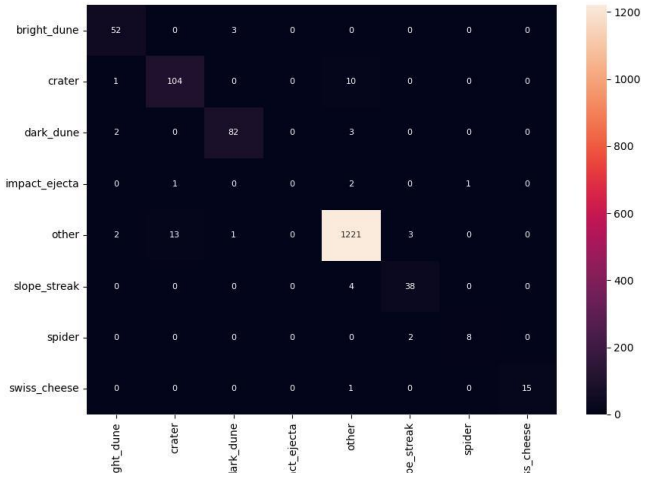

Figure 2 - Confusion matrix for ViT/16-21k.

The confusion matrix shows a heavily biased dataset, containing a significant number of images that fall under the "other" category. When attempting to account for this by balancing the inputs based on class, we discovered that the least represented class, "impact_ejecta", would skew the training set and lead to the ViT models being trained on only 80 images total. Due to this skew in data, we believe that this test produces inconclusive results that require further working with the dataset.

E. Transformer Parameter Analysis

This section should detail the changes we made to the 3 moThe classification reports for each of the four ViT variants can be found in Appendix B. The test set accuracies for each

variant show that the ViT/16-21k, the Vision Transformer trained on size 16 batches and pre-trained on ImageNet-21k, did produce the highest accuracy. However, we do recognize that the range from maximum to minimum accuracies was only 0.45%. This is extremely thin and very likely the cause of error within the training data. Looking at the "support" columns, we see that images from the classes spider and impact_ejecta were seen in the single-digit number of times during testing. This does not represent a balanced and unbiased dataset. To solve this, more images regarding these classes could be added, or permutations could be given to existing images to increase the size of the dataset.

We also know from previous works that the pre-training dataset size should have a significant effect on model accuracy [4]. This leads us to believe that when switching from pre-training dataset ImageNet 2012 to ImageNet-21k, we should see a better performing model. While this holds true for batch size = 16, the batch size = 12 models do not demonstrate this behavior.

V. CONCLUSION

This work sought to investigate the "Deep Mars" paper and its use of the HiRISE Mars satellite image dataset. In particular, we posed the question of whether the recently introduced Vision Transformer model, rivaling state-of-the-art architectures for image classification, would be more suitable for the task in place of the CNN used in the work. While we did see promising results from the Vision Transformer as opposed to the CNN, we ultimately find that work is needed to clean the data used in training for both models as well as balancing the representation of classes.

Future work utilizing these two models could include larger pre-training datasets, or even pre-training datasets that apply closer to the final task, such as classification of landscapes. As the initial inspiratory work involved transfer learning of Earth satellite images onto HiRISE, conducting this task with both CNN and ViT models would be another interesting area of study as well.

APPENDIX A: Required Software Packages

List of imported software packages/libraries:

For CNN Training / Evaluation:

- Tensorflow, Keras
- Math, Numpy, Matplotlib, Scikit-Learn
- PIL
- Tqdm

For Transformer Training / Evaluation:

- HugsVision [7], Transformers
- Pandas
- Seaborn

APPENDIX B: Transformer Variant Evaluation

This section includes the ViT variants displayed earlier in Table 2 with their test set accuracy. Screenshots of these accuracies in Jupyter Notebook are also provided.

Transformer Name	Pre-Training Dataset	Batch Size	Test Set Accuracy
ViT/12	ImageNet-21k, FT 2012	12	96.75%
ViT/12-21k	ImageNet-21k	12	96.43%
ViT/16	ImageNet-21k, FT 2012	16	96.43%
ViT/16-21k	ImageNet-21k	16	96.88%

Table 2 - Pre-training dataset and batch size used for four ViT variants. "FT 2012" means the model was later fine-tuned on ImageNet 2012.

ViT/12

accuracy			0.9675	1569
macro avg	0.9493	0.8711	0.8880	1569
weighted avg	0.9675	0.9675	0.9671	1569

ViT/12-21k

accuracy			0.9643	1569
macro avg	0.9564	0.9343	0.9441	1569
weighted avg	0.9643	0.9643	0.9642	1569

ViT/16

accuracy			0.9643	1569
macro avg	0.9402	0.8618	0.8782	1569
weighted avg	0.9641	0.9643	0.9636	1569

ViT/16-21k

accuracy			0.9688	1569
macro avg	0.8130	0.8024	0.8072	1569
weighted avg	0.9665	0.9688	0.9676	1569

CONTRIBUTIONS

Aniruddha Prasad - Worked on the CNN setup, training and evaluation. Also assisted in acquiring the dataset and brainstorming the idea for the project.

Andrew Hartnett - Implemented code for Vision Transformers, as well as their comparison. Conducted background research for the Transformer model.

ACKNOWLEDGMENT

We would like to express our gratitude to Mario Parente for inspiring in us a passion and enthusiasm for machine learning.

REFERENCES

[1] Mars orbital image (HiRISE) label data set. November 13, 2017. DOI:10.5281/zenodo.1048301.

[2] Kiri L. Wagstaff, You Lu, Alice Stanboli, Kevin Grimes, Thamme Gowda, and Jordan Padams. "Deep Mars: CNN Classification of Mars Imagery for the PDS Imaging Atlas." Proceedings of the Thirtieth Annual Conference on Innovative Applications of Artificial Intelligence, 2018.

[3] O'Shea, K., & Nash, R. (n.d.). An Introduction to Convolutional Neural Networks. *Neural and Evolutionary Computing*. https://doi.org/10.48550/arXiv.1511.08458

[4] Dosovitskiy, Alexey and Beyer, Lucas and Kolesnikov, Alexander and Weissenborn, Dirk and Zhai, Xiaohua and Unterthiner, Thomas and Dehghani, Mostafa and Minderer, Matthias and Heigold, Georg and Gelly, Sylvain and Uszkoreit, Jakob and Houlsby, Neil. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." 2020. DOI:10.48550/ARXIV.2010.11929. Accessed: https://arxiv.org/abs/2010.11929.

[5] Vision Transformer GitHub Repository. Accessed: https://github.com/google-research/vision_transformer

[6] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N. and Kaiser, Lukasz and Polosukhin, Illia. "Attention Is All You Need." 12 Jun 2017. DOI:10.48550/ARXIV.1706.03762. Accessed: https://arxiv.org/abs/1706.03762

[7] Hugsvision GitHub Repository. Accessed: https://github.com/qanastek/HugsVision

[8] Niehus-Staab, L. (2022, May 3). HiRISE-Net. GitHub. https://github.com/niehusst/HiRISE-Net

[9] "How to Train a Custom Vision Transformer (ViT) Image Classifier to Help Endoscopists in Less than 5 min." yanis labrak. 2 Sep 2021. Accessed: https://medium.com/@yanis.labrak/how-to-train-a-custom-vision-transformer-vit-image-classifier-to-help-endoscopists-in-under-5-min-2e7e4110a353

[10] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[11] mars.nasa.gov. (n.d.). HiRISE. Mars.nasa.gov. https://mars.nasa.gov/mro/mission/instruments/hirise/