HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 3 (2024)

A Deep Learning Pipeline for Drought Assessment using Spatial Satellite Images and Vision Transformers

Aaryan Doshi[1] and Joe Kim[#]

[1]Monta Vista High School, USA
[#]Advisor

ABSTRACT

700 million people are in danger of being displaced due to inept drought prediction and prevention systems. Current research on drought assessment focuses solely on factors such as soil moisture and rainfall, which require painstaking measurements and lab samples, and can often be misleading. This research eliminates this requirement by proposing an end-to-end pipeline to detect and prevent droughts in at-risk areas using satellite images and vision transformers. The dataset is comprised of over 86,000 satellite images labeled by pastoralists and divided with an 80-20 ratio for training and validation. First, using feature filtering, normalization, and a Gaussian filter, the images in the dataset are modified to yield a better performance. Next, a deep vision transformer model with multi-headed attention is constructed, consisting of four heads, three transformer layers, and a patch size of five. The final MLP head produces logits for drought severity prediction level. Overall, the best transformer model achieves 78.3% accuracy in predicting drought conditions on a validation set of 10,000, unseen satellite images. In addition, this method outperforms state-of-the-art convolutional neural networks on this classification task, as compared to VGG-16, ResNet-50 and DenseNet-121 models. The model harnesses AWS cloud computing, deep vision transformers, and specific image augmentation to achieve state-of-the-art results in drought prediction and prevention. With this research, scientists have the potential to assess droughts quickly and accurately, revolutionizing our ability to provide resources and care to those affected by the increasingly common droughts caused by the climate crisis worldwide.

Background

Droughts worldwide present serious agricultural, ecological and socio-economic risks worldwide. According to the World Health Organization, an estimated 55 million people globally are affected by droughts every year (WHO, 2021). Furthermore, it is the most serious hazard to livestock and crops in nearly every part of the world. Droughts threaten livelihoods, increase the risk of disease and death, and fuel mass migration. In fact, per World Health Organization as many as 700 million people are at-risk of being displaced as a result of drought by 2030 (WHO, 2021). Climate change is gradually accelerating drought occurrences. Droughts do not always offer the same immediate and dramatic visuals associated with events such as wildfires and hurricanes, but they still have a huge price tag. In fact, droughts rank second in types of phenomena associated with billion-dollar weather disasters during the past three decades (APGA 2015). With annual losses nearing $9 billion per year, according to National Centers for Environmental Information, droughts are a serious hazard with substantial socioeconomic risks for the United States (NCEI 2019).

Drought forecasting and drought assessment maintain two sides of the same coin – it is impossible to understand drought without also assessing its severity. According to Wilhite, D.A, the key areas that should be

addressed are: (1) monitoring and early warning, which identifies drought status in a timely fashion; (2) vulnerability and impact assessment, determining the location and what is at risk of drought and why; and (3) mitigation and response, describing actions and measures needed to mitigate drought impacts and respond to drought emergencies (Wilhite, 2014). Most measures of drought stress severity and status assessment rely upon soil moisture sensors (Jones et. al, 2004). These traditional approaches have low efficiency and limited indirect and spatial area, making it a poor way to address the above three areas (Mangus et. al, 2016).

With the development of computer vision, machine learning and image processing techniques, a deep learning-based drought assessment system with real-time surveillance cameras has the advantage of providing constant and accurate monitoring over large spatial areas, compared to traditional techniques. Advancements in satellite-based image-processing technologies enable rapid monitoring of very large land masses for detecting diminishing quantities of forage after the outbreak of a drought. Additionally, vision transformers (Dosovitsky et. al, 2021) provide excellent image classification and object detection by extracting features and patterns from images. These capabilities can be employed to aid in drought assessment.

Introduction

While the debate on causes and severity of climate change rages on, the ecological and environmental havoc that has resulted from changing climate patterns is indisputable. The duration of droughts has become longer (Zhang et. al. 2021) (Behzadi et. al. 2024), which has led to increasingly erratic harvesting seasons. These impacts will be felt directly by hundreds of millions of people displaced due to these extended, multi-year droughts and indirectly by an even larger population due to food shortages.

Traditionally, previous research in the field has delved into upstream science of drought forecasting and prevention (Hao et. al., 2018) (Zhang et. al., 2024) (Dikshit 2021). However, empirical, data-driven assessments of the severity of drought conditions continue to lag behind. Current drought severity characterizations leverage indices such as Standardized Precipitation Index (SPI) (McKee et. al., 1993) to determine the magnitude of the drought, using data on rainfall and soil quality. However, these measures rely heavily on deviation from past norms to characterize drought severity – they notably exclude present data, where predictions are full of deviations induced by climate change. For example, extensive amounts of tillable land will face drought for the first time as climate change becomes more severe (Fitton et. al., 2019) (Nuccitelli, 2022). As a result, drought assistance programs employed to provide resources to agricultural communities have become incredibly ad-hoc. A lack of robust drought assessment data drastically inhibits an empirical approach towards assistance distribution and other agronomic areas such as building risk models for drought insurance.

This research aims to address the challenge of drought assessment and characterization via an end-to-end pipeline involving the application of vision transformer models (Dosovitsky et. al, 2021) on Synthetic Aperture Radar (SAR) satellite images of current ground conditions. The pipeline is composed of three major stages. First, we perform feature map generation on satellite images using a variety of image pattern detection techniques that include feature filtering, normalization, and a Gaussian filter (Hayeyer et. al., 1989) to optimize model performance. Next, these images become the input to a deep, multi-layer vision transformer model, consisting of four attention heads, three transformer layers, and dimensionality reduction via patches of size sixteen. Finally, the outputs from the transformer are fed into an MLP layer, which produces probability logits for each of four classes of drought severity.

This research also tests the pipeline using multiple established neural network models to identify which model performs best and visualizes the attention maps of the final method to assess its validity. Since the approach leverages current image sets (instead of deviations from the past norm), it is better suited to assess climate pattern change driven droughts. Furthermore, as this approach leverages LANDSAT-8 satellite images taken from space, it is inherently better able to cover vast swaths of land mass faster, which increases the reach of drought assessment models. Drought assistance programs can then tailor the distribution of their resources

based on the drought assessment predicted by the model resulting in less waste and focused, improved benefits. This research has the potential to transform agronomical industries via a data driven approach to offer drought insurance and aid.

Hypothesis

Vision transformers have proved invaluable in image classification domains. The goal of this work is to explore whether these capabilities similar transfer when training on hyperspectral multi-band images for drought severity prediction. In addition, a secondary goal of the work is to ensure that performance is equal or better than the state-of-the-art convolutional neural network counterparts.

With these factors in mind, the hypothesis of this research is that:

a. We can build a drought severity prediction pipeline by training a vision-transformer to perform image classification on satellite images

b. The vision transformer's architecture lends itself to global and local attention within an image, rather than the local inductive biases imposed by CNNs, allowing it to better capture information relevant for image classification. Ultimately, the transformer's performance should be comparable to or better than the most widely used of its CNN counterparts.

Methods

History & Overview of Vision Transformers

Vision transformers (ViTs) represent a groundbreaking shift in the landscape of computer vision, diverging from the traditional Convolutional Neural Networks (CNNs) that have dominated the field for decades. Originating from the transformer architecture (Vaswani et. al, 2017), which revolutionized natural language processing (NLP), vision transformers adapt the self-attention mechanism to process visual data, allowing for a more flexible and comprehensive analysis of images.

The foundational work by Alexey Dosovitskiy et al., which presented the vision transformer model for image recognition tasks (Dosovitsky et. al, 2021), demonstrated how transformers could be effectively applied to image classification, a pivotal moment in the evolution of machine learning techniques for computer vision. In their approach, an image is divided into a sequence of patches, analogous to words in a sentence, which are then linearly embedded and processed through multiple layers of self-attention mechanisms. This method enables the model to consider the global context of an entire image, a significant departure from the local view provided by CNNs. For image classification task specifically, a fully-connected layer is appended to the end of the network, taking in the attention outputs and converting them to probability logits for each class.

At the core of the Vision Transformer architecture is the transformer encoder, which consists of alternating layers of multi-head self-attention and position-wise fully connected feed-forward networks. Each layer also includes normalization steps and residual connections, enhancing the model's ability to learn deep representations without degradation. The self-attention mechanism allows the model to weigh the importance of different parts of the image, regardless of their spatial proximity, enabling it to focus on relevant features for the task at hand.

One of the key advantages of vision transformers over CNNs is their scalability and adaptability to various image sizes and resolutions. By adjusting the number of transformer blocks and attention heads, vision transformers can be tailored to specific computational budgets and performance requirements. This flexibility, combined with their ability to capture long-range dependencies in the data, has propelled vision transformers to the forefront of research and application in computer vision, challenging the dominance of CNNs.

The application of vision transformers extends beyond image classification to tasks such as object detection, semantic segmentation, and more. Their ability to learn rich, global representations of visual data has shown promise in improving performance across a range of challenging datasets and benchmarks. Moreover, the ongoing research and development in this area are rapidly expanding the capabilities and efficiency of vision transformers, making them a critical component of the modern computer vision toolkit.

As the field continues to evolve, vision transformers are expected to play a central role in advancing our understanding and processing of visual information, opening new avenues for innovation and application in areas such as autonomous vehicles, medical image analysis, and augmented reality.

Figure 1 below illustrates the architecture of a Vision Transformer, showcasing the process from image patch embedding to the final classification output.

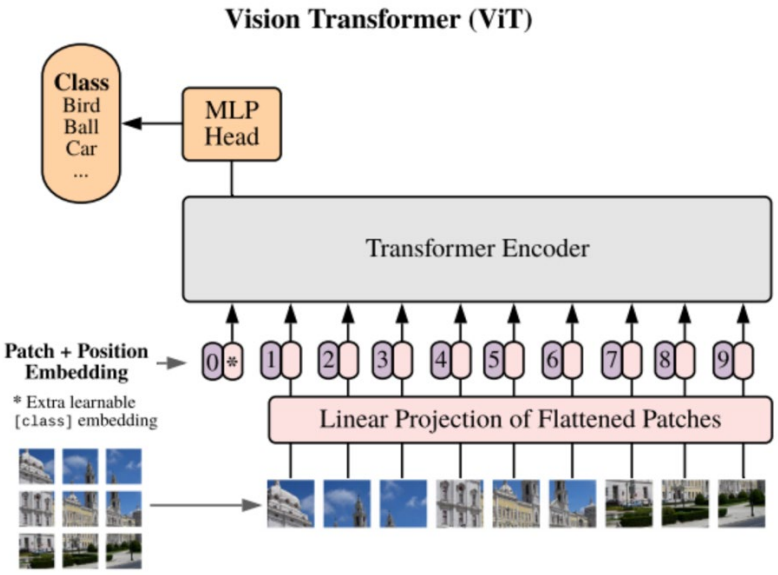


Figure 1. Vision Transformer for Image classification

Dataset

Our dataset is comprised of:

- A training image dataset of 86,217 images consisting of ground-based labels and photos from pastoralists assessing forage conditions on site.
- A test validation image dataset of 10,778 satellite images from NASA's LANDSAT12 mission.

Training Image Data Set Labeling

As mentioned above, in order for a vision transformer model to perform classification of input images, it first needs to be trained on a pre-existing labeled dataset. Labeling any image dataset requires domain expertise, and generating labels on aerial, satellite images can be a daunting task. Our training image dataset was labeled by human pastoralists who harnessed domain expertise to assess the amount of forage the geographic center of the land could sustain.

The image dimensions are 65x65 pixels each representing an area of 1.95 kilometers across. At each reporting point, pastoralists chose from a range of 0, 1, 2 or 3+ cows that could eat for a single day within 20

meters of their location, which corresponds to an area slightly larger than a single pixel. Additionally, Gaussian blur filter (a filter whose weights are sampled from a Gaussian distribution) and image normalization techniques were harnessed to optimize the images for training and testing.

Satellite Test Images

Each labeled location was complemented by a 65x65 pixel LANDSAT8 satellite image. Each pixel in a LANDSAT image represents a 30 * 30 meter area. A distinguishing characteristic of the LANDSAT 8 images are the 8 additional bands in addition to the standard RGB bands that help in more granular discernment of features within these images especially around aridity, cloud, and precipitation.

This approach yields a supervised labeling of forage quality on satellite images. It can be further extrapolated if required to assess other kinds of ground data such as crop quality or aridity for future research. Furthermore, once the ViT is trained on the labeled dataset, the same model can be leveraged on satellite images across other regions enabling rapid applicability of the approach across larger geographic regions.

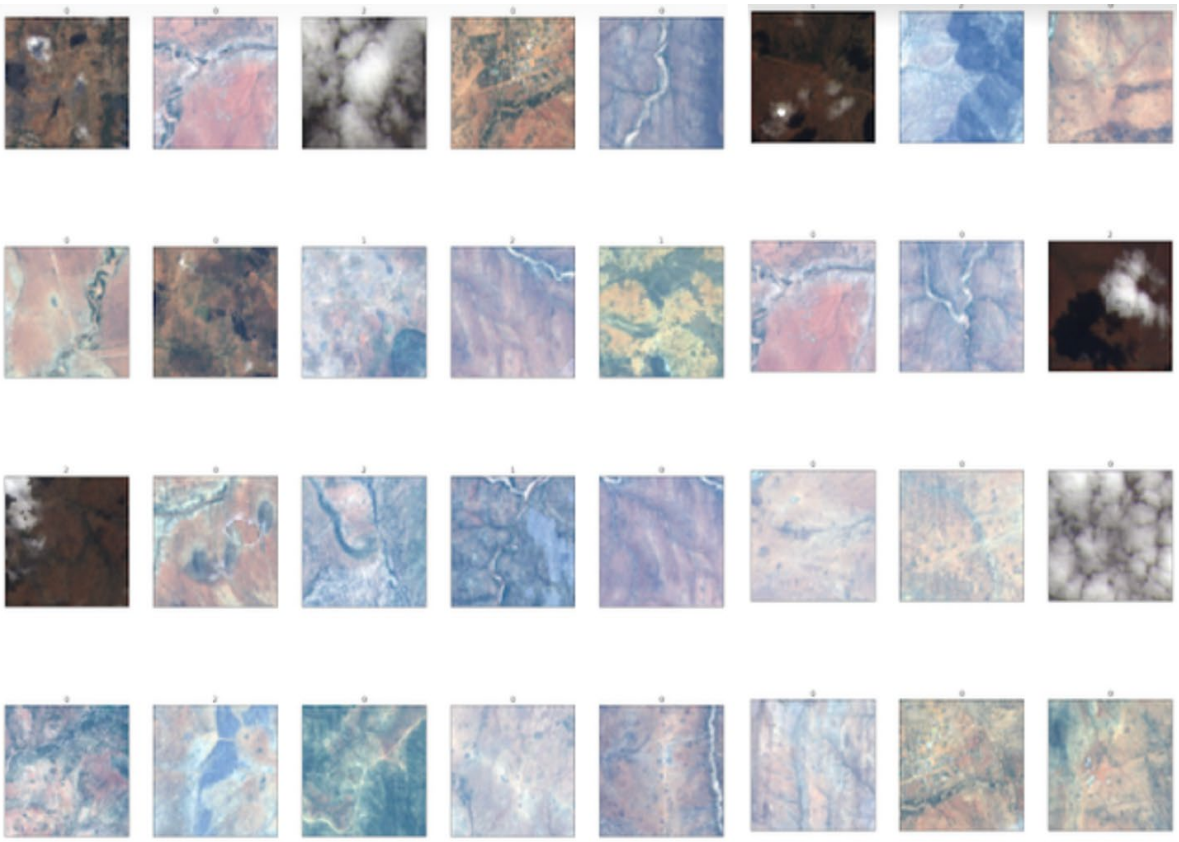


Figure 2. Labeled dataset of forage conditions by human pastoralists

Transformer Configuration Details

The final transformer model was trained using multi-head attention and several layers. The specific architecture configuration is detailed in Table 1.

Table 1. Training configuration of final multi-head, multi-layer transformer architecture.

Patch Size	5
# Transformer Layers	3
# Attention Heads	4
Transformer Units	(256, 256, 128)
Optimizer	Adam
MLP Dimension	128

Environment Setup

To maximize efficiency and to develop an approach for future extensions to this work, this project was powered by the deep learning AMI of Amazon Web Services (AWS) cloud, leveraging TensorFlow and Ubuntu frameworks.

I used an AWS G3 xlarge EC2 instance 14 with 1 GPU, 4vCPU and 30.5 G of RAM to ensure there was enough computing power (processing and memory) to run the different models.

Results

The results below demonstrate the performance of our transformer architecture in comparison to various, traditional convolutional neural network architectures. On the validation set of over 10,000 images, our transformer model outperforms all baselines, ultimately achieving validation accuracy of 78.3%.

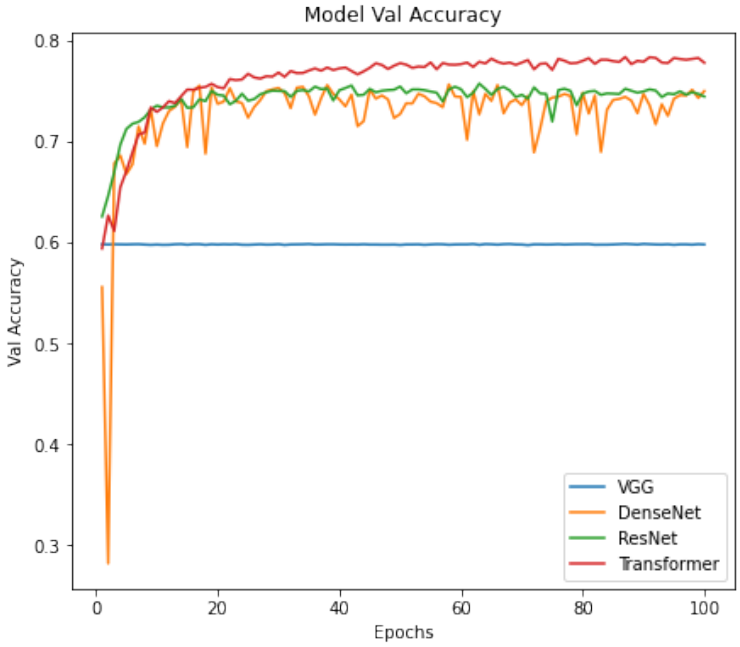


Figure 3. Model Validation Accuracies. The transformer model outperforms convolutional neural network baselines starting at around 25 epochs, converging at 60 epochs. VGG converges very quickly into training,

while the best DenseNet and ResNet architecture converge at 20 epochs, though at a lower value than our transformer architecture.

Table 2. Final Model Validation Accuracies. The validation accuracies averaged across ten runs are displayed. The transformer architecture achieves performance comparable to and slightly higher than the convolutional neural network baselines.

Model	**Validation Accuracy**
VGG-16	59.8%
ResNet-50	75.7%
DenseNet-121	75.6%
Transformer	78.3%

In addition to the validation accuracies, a plot of the per-model validation loss is shown to demonstrate convergence of the models. The transformer model's loss continues to decrease, with the baseline models converging much earlier into training.

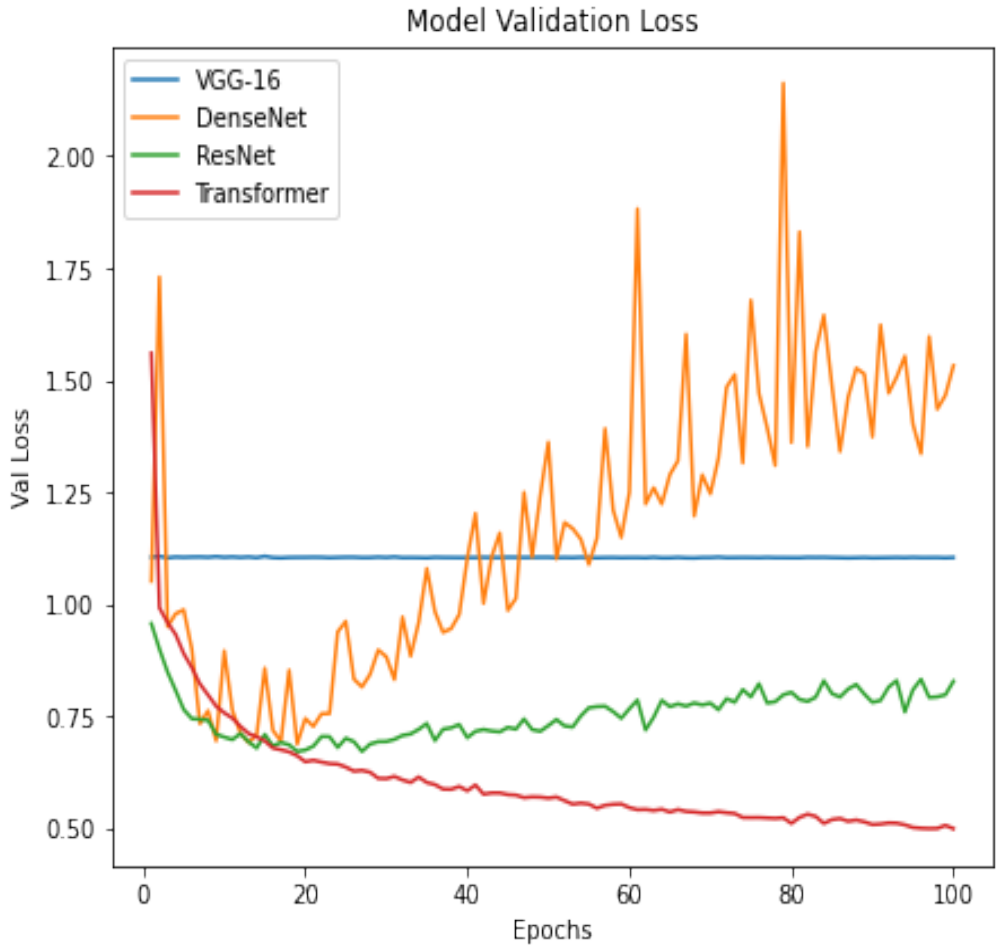


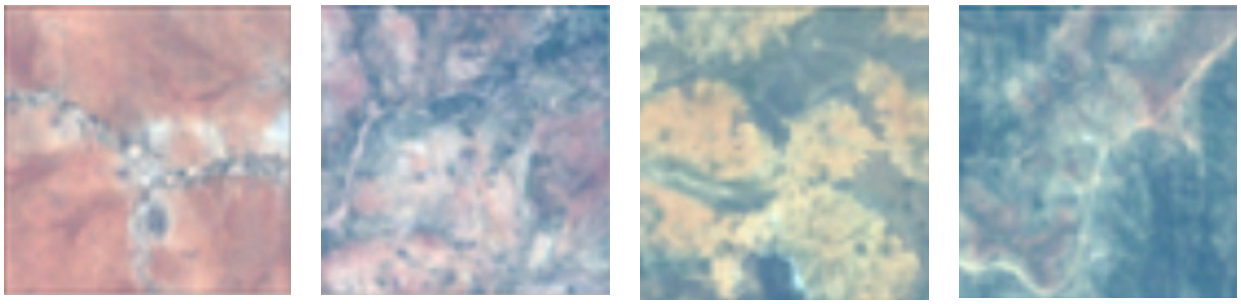
Figure 4. Model Validation Loss. The validation loss for the transformer continues to decrease over the course of training. The VGG validation loss converges within the first three training epochs, while the ResNet and DenseNet losses start to increase at around 30 epochs, indicating overfitting early into training.

In addition, all results are averaged over a set of ten runs, yielding a p-value greater than alpha = 0.05, to ensure that all results are statistically significant.

Table 3. Statistical Significance. The t-values and p-values are computed between each baseline model and the transformer architecture, where we use validation accuracies compiled across ten distinct runs of each model. The p-values all fall below an alpha of 0.05 connoting the statistical significance of these results.

Model Comparison	**t-value**	**p-value**
Transformer vs VGG-16	36.37	0.000003
Transformer vs ResNet-50	10.89	0.000404
Transformer vs DenseNet-121	8.15	0.001235

We also provide examples of our trained transformer model's predictions of forage on images from our validation dataset.



(0) (1) (2) (3)

Figure 5. Model Predictions. The number corresponding to each image is the transformer model's correct prediction of how much forage the land in the image can maximally sustain.

Discussion

The transformer model's performance across the board is comparable to leading convolutional neural network baselines, even minimally outperforming them to a level that is statistically significant. Our transformer achieves state-of-the-art performance, reaching an accuracy of 78.3% on a validation set of 10,000 images. Likely the transformer's ability to integrate both local and global information allow it to express patterns that prove extremely useful in hyperspectral satellite image classification – these patterns are harder to learn with convolutional neural networks, whose inductive biases are more conducive to capturing local information

within an image. Ultimately, the results of this experiment show that it is indeed possible to build a robust drought severity prediction pipeline by utilizing a vision transformer.

Conclusion

The use of vision transformers in the context of this project highlights the potential for machine learning models to contribute effectively to environmental monitoring and disaster response strategies. By harnessing transformer-based prediction pipelines, we can enhance our capabilities to predict and mitigate the effects of climate change, particularly in agriculture and water management sectors.

The scalability of this model suggests that it could be implemented on a global scale, utilizing satellite image data in real-time from sources such as NASA to provide much-needed insights into drought severity that will help hasten relief strategies. Future work will focus on refining the model's predictive accuracy and expanding its application to other forms of environmental monitoring.

As climate change continues to affect global weather patterns, the need for advanced tools like our vision transformer for drought severity assessment becomes that much more critical. Continued advancements in AI and machine learning will lead to more robust, accurate, and timely environmental monitoring systems that can significantly reduce the impact of droughts and other natural disasters on vulnerable populations.

In conclusion, this work demonstrates the effectiveness of vision transformers in drought severity assessment via forage prediction, ultimately achieving state-of-the-art performance and enabling more timely and smarter drought relief programs across the world.

Limitations

This project currently tackles drought severity assessment via solely forage prediction. Future work could apply our vision transformer architecture in conjunction with predictions of other drought indicators, such as soil quality or vegetation health.

Acknowledgments

I would like to thank my teacher Mr. Kim, for helping me hone the computer science skills needed to tackle this project; NASA, for open-sourcing their LANDSAT-8 dataset; and my family, for their support.

References

Zhang, F., Biederman, J. A., Dannenberg, M. P., Yan, D., Reed, S. C., & Smith, W. K. (2021). Five Decades of Observed Daily Precipitation Reveal Longer and More Variable Drought Events Across Much of the Western United States. Geophysical Research Letters, 48(7). https://doi.org/10.1029/2020gl092293

Behzadi, F., Javadi, S., Yousefi, H. et al. Projections of meteorological drought severity-duration variations based on CMIP6. Sci Rep 14, 5027 (2024). https://doi.org/10.1038/s41598-024-55340-x

Hao, Z., Singh, V. P., & Xia, Y. (2018). Seasonal Drought Prediction: Advances, Challenges, and Future Prospects. Reviews of Geophysics, 56(1), 108–141. https://doi.org/10.1002/2016rg000549

Drought. (2019, November 8). https://www.who.int/health-topics/drought

Zhang, H., Loaiciga, H. A., & Sauter, T. (2024). A Novel Fusion-Based Methodology for Drought Forecasting. Remote Sensing, 16(5), 828. https://doi.org/10.3390/rs16050828

Dikshit, A., & Pradhan, B. (2021). Explainable AI in drought forecasting. Machine Learning With Applications, 6, 100192. https://doi.org/10.1016/j.mlwa.2021.100192

Fitton, N., Alexander, P., Arnell, N., Bajzelj, B., Calvin, K., Doelman, J., Gerber, J., Havlik, P., Hasegawa, T., Herrero, M., Krisztin, T., Van Meijl, H., Powell, T., Sands, R., Stehfest, E., West, P., & Smith, P. (2019). The vulnerabilities of agricultural land and food production to future water scarcity. Global Environmental Change, 58, 101944. https://doi.org/10.1016/j.gloenvcha.2019.101944

Nuccitelli, D., & Nuccitelli, D. (2022, October 20). UN report: The world's farms stretched to 'a breaking point.' Yale Climate Connections. https://yaleclimateconnections.org/2022/01/un-report-the-worlds-farms-stretched-to-a-breaking-point/

Wilhite, D. A. (n.d.). National Drought Management Policy Guidelines: A Template for Action. DigitalCommons@University of Nebraska - Lincoln. https://digitalcommons.unl.edu/droughtfacpub/83/

Mangus, D. L., Sharda, A., & Zhang, N. (2016). Development and evaluation of thermal infrared imaging system for high spatial and temporal resolution crop water stress monitoring of corn within a greenhouse. Computers and Electronics in Agriculture, 121, 149–159. https://doi.org/10.1016/j.compag.2015.12.007

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2010.11929

D'Haeyer, J. P. (1989). Gaussian filtering of images: A regularization approach. Signal Processing, 18(2), 169–181. https://doi.org/10.1016/0165-1684(89)90048-0

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017a). Attention Is All You Need. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1706.03762

USGS Landsat 8 Collection 1 Tier 1 and Real-Time data Raw Scenes [deprecated]. (n.d.). Google for Developers. https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C01_T1_RT

Drought Tools and Resources. (2015, October 28). APGA. https://www.publicgardens.org/resource/drought-tools-and-resources/