Vision Transformer

Multi-Modal Spatial Temporal ViT

University at Buffalo The State University of New York

Introduction

About Team 1:

Brason Dobson	Bhushan Mahajan	Deep Shahane
Computer Science M.S.	Computer Science M.S.	Computer Science M.S.

Index

- Overview
- Objective
- Motivation
- Vision Transformers
  - Architecture
  - Patch Embedding, Positional Embedding
  - Self Attention Mechanism
  - Multi head Attention
- Deep Learning
- Problem Statement of paper
- Challenges Addressed
- Dataset Used
- Proposed Approach(Overview)

Overview

MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer

Fudong Lin[1], Summer Crawford[2], Kaleb Guillot[2], Yihe Zhang[2], Yan Chen[3], Xu Yuan[1*], Li Chen[2], Shelby Williams[2], Robert Minvielle[2], Xiangming Xiao[4], Drew Gholson[5], Nicolas Ashwell[5], Tri Setiyono[6], Brenda Tubana[7], Lu Peng[8], Magdy Bayoumi[2], Nian-Feng Tzeng[2]
[1] University of Delaware, [2] University of Louisiana at Lafayette, [3] University of Connecticut, [4] University of Oklahoma, [5] Mississippi State University, [6] Louisiana State University, [7] LSU AgCenter, [8] Tulane University

Objective

- Explore the theory and application of the MMST-ViT model
- Analyze paper methodology and code implementation on crop yield data
- Clearly and concisely demonstrate an overview of the paper's meaning, importance, intention, applications, and prospects
- We want to understand the vision transformer, potentially reproduce its work, and consider different approaches
- Understand the distinguishing qualities of the multi-modal approach
- Analyze effectiveness of MMST-ViT compared to existing methods (ConvLSTM, CNN-RNN, GNN-RNN, …)
- Demonstrate how the approach refutes previous challenges/setbacks in combining visual and numerical data for agricultural predictions

MMST-ViT

- Key Components: Multi-Modal Transformer, Spatial Transformer, Temporal Transformer
- Distinguished from other ANN architectures/Transformers
- Multi-Modal method considers visual (satellite imagery) and numerical (meteorological data)

  (Covers predictions for corn, cotton, soybean, and winter wheat across multiple U.S. states)

- Better understanding of context between short-term weather variations and long-term climate change effects on crops to capture nuanced, intricate dependencies and improve accuracy of county-level crop yield predictions across the U.S.
- MMST: Considers underlying relationships between visual AND numerical data over space and time
- Inspires multi-modal approaches in other real-world applications

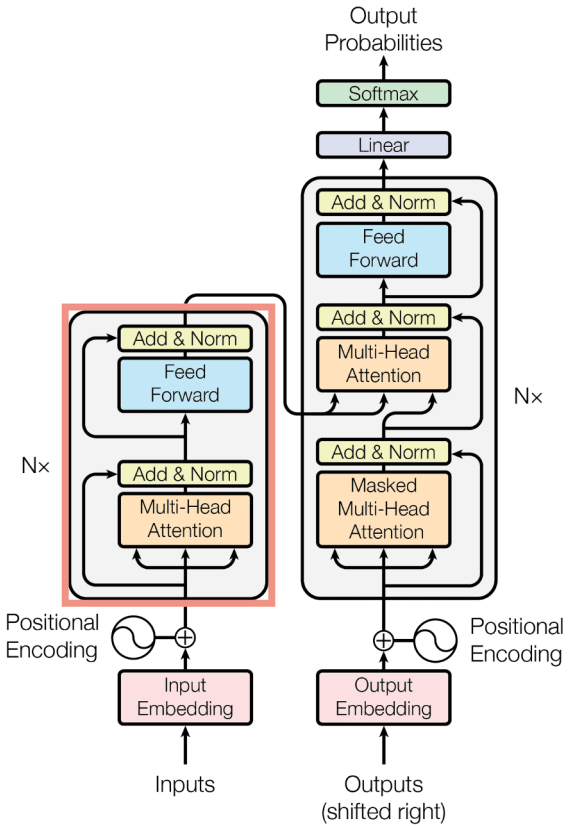
Motivation

- Use technical abilities in our future work to make the world a better place
- Relevance to current agricultural challenges (e.g., food security, climate change adaptation)
- Opportunity to learn about multimodal deep learning approaches and SOTA technologies
- Previous architectures struggle with complex, multimodal data
- Some uses of VISION transformers specifically include medical imaging, facial recognition, image classification (tumors benign or fatal), autonomous vehicles
- Exploring a relatively recent paper (September 2023) to stay up-to-date with a rapidly evolving field

Transformers

- The Transformer architecture was introduced in 2017 by Vaswani et al. in the paper “Attention is All You Need”
- It is used in a lot of NLP tasks like translation, summarization, and text generation.
- Unlike RNNs (Recurrent Neural Networks), which process data sequentially, Transformers can process data in parallel, making them more efficient for large datasets
- There are 6 main key components of transformers

1. Positional Encoding
2. Encoder Decoder Structure
3. Self Attention Mechanism
4. Multi Head Attention
5. Feed Forward Layer
6. Parallel Processing



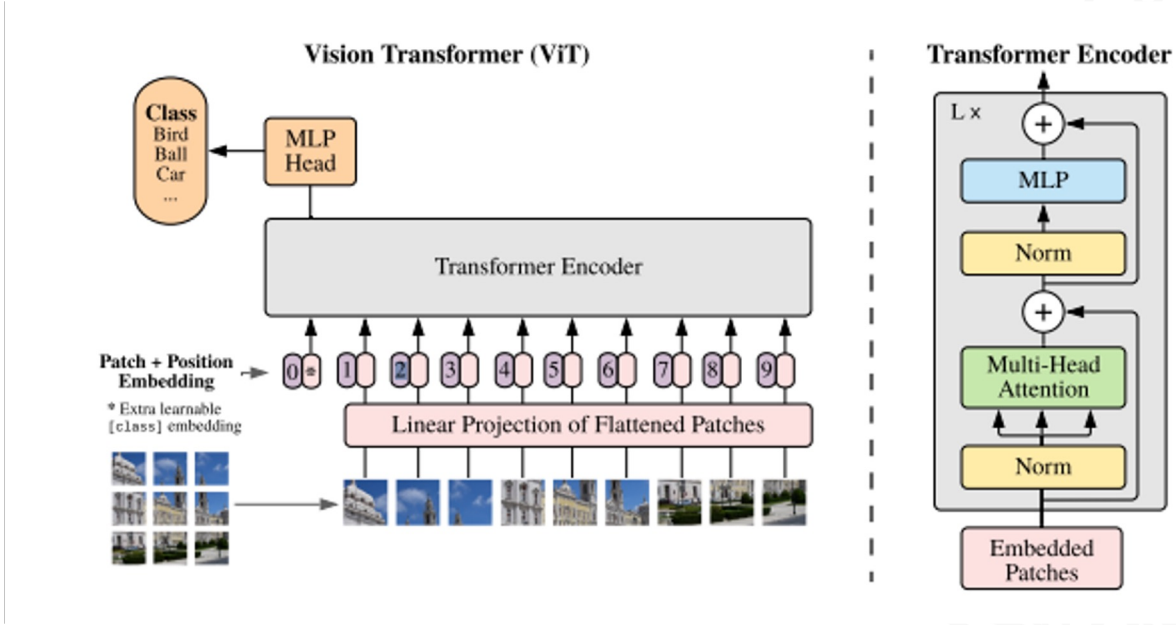
Why Transformer for Vision

- CNN was the most preferred option for image recognition as it is best at detecting patterns in images due to their convolutional layer
- They use a sliding window (kernel) to capture small features, like edges or textures, in an image. These small features are then combined to form higher-level patterns like shapes or objects. Hence capturing long-range dependencies (e.g., relationships between objects far apart in an image) requires deep networks, making training more computationally expensive.
- The research paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" shows that traditional reliance on CNN's is not necessary. The paper proposes the Vision Transformer (ViT), which treats images as sequences of patches

What is Vision Transformer?

- Vision Transformers divide images into fixed-size patches (e.g., 16x16 pixels) and treat each patch as a "token" similar to words in NLP. These patches are flattened into vectors and passed through the Transformer architecture.
- Vision Transformers use the self-attention mechanism, which allows every patch to interact with every other patch in the image. This helps capture long-range dependencies and global relationships across the image.

Vision Transformer Architecture Overview



Patch Embedding

- The input image $x \in R^{H \times W \times C}$ (with height H, width W, and channels C) is divided into non-overlapping patches, each of size P×P. The number of patches N is:

$$N = (H \times W) / P^2$$

- Each patch is flattened and linearly projected into a D-dimensional embedding space using a learnable projection matrix E:
  - $z_p = x_p E$ for $p = 1, 2, \ldots, N$

Position Embedding

Because the Transformer has no inherent understanding of the position of patches, **positional embeddings** E_{pos} are added to the patch embeddings to retain spatial information:

$$z0=[x_{class};z1+ E_{pos} ;z2+ E_{pos} ;\dots ;zN+ E_{pos}]$$

Here, z_0 is the sequence of embeddings that includes the **[class] token** x_{class} and the positional embeddings $E_{pos} \in R^{(N+1)\times D}$

Self Attention Mechanism

The Main operation in Transformers is the **self-attention** mechanism. Each patch embedding z_p attends to all other patch embeddings. The attention for each patch is computed as: $A(Q,K,V)=\text{softmax}(QK^T/\text{sqrt}(D))*V$ where:

- $Q=zW_Q$, $K=zW_K$, and $V=zW_V$ are the **query**, **key**, and **value** matrices obtained by projecting the patch embeddings z
- $W_Q,W_K,W_V \in R^{D\times d}$ are the learnable weight matrices for the queries, keys, and values.
- The **softmax** function ensures the attention weights sum to 1

Query (Q):

- Represents the current patch's embedding and asks, "How much attention should I give to other patches?" It’s responsible for determining the focus of each patch in the image.

Key (K):

- Acts like an identifier for each patch. Each patch’s key is compared with the query to measure how relevant that patch is to the current patch being processed.

Value (V):

- Contains the actual information of each patch that is passed forward. The values are combined according to the attention scores derived from the query-key comparisons.

Multi head Attention

To capture different aspects of patch relationships, **multi-head attention** is used, where multiple self-attention operations are applied in parallel. The outputs from all attention heads are concatenated:

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head}_1,\ldots,\text{head}_h)WO$$

where $W_O \in R^{h\cdot d\times D}$ is the output projection matrix, and each attention head_i is computed using the self-attention mechanism

Final Classification

After passing through several layers of multi-head attention and feed-forward networks, the **[class] token** is extracted from the final output and passed to a simple **MLP head** for classification:

$$\hat{y} = \text{MLP}(z_0^L)$$

where z_0^L is the final representation of the **[class] token** after L layers of the Transformer encoder, and $\hat{y}$ is the predicted class label.

How Vision Transformer is used here??

- The research introduces the MMST-ViT model, which utilizes Vision Transformers to predict crop yields by analyzing both visual data (satellite imagery) and meteorological data (temperature, precipitation, etc.)

- The Vision Transformer in this paper applies multi-head attention over both visual remote sensing data (from satellites) and numerical meteorological data.

- **Spatial Transformer**: The Vision Transformer is extended to learn spatial dependencies between different counties, allowing precise crop tracking.

- **Temporal Transformer**: It also captures temporal dependencies, learning how long-term climate change impacts crops, a crucial factor for accurate yield predictions.

- **Multi-Modal Transformer**: Combines both satellite images and weather data to understand how different factors, like land and climate, affect crop growth. This helps improve the accuracy of crop yield predictions.

Deep Learning

- Deep learning models combine satellite imagery and meteorological data to capture complex, non-linear relationships affecting crop growth, enabling more accurate and scalable crop yield predictions.
- By automatically learning features from multi-modal data, deep learning models outperform traditional methods, providing valuable insights for agricultural planning and resource allocation.
- GNN(Graph Neural Network)s are used to capture relationships between different regions (e.g., neighboring counties) by modeling the spatial dependencies in crop yield prediction, enhancing the ability to predict yields based on geographical factors.

Problem Statement of Paper:

Objective: Develop a model to predict crop yield at the county level using multi-modal data

- Satellite imagery, short- and long-term meteorological data, and USDA crop data.

Data Types:

- **Satellite Imagery** ($x \in R^{T \times G \times H \times W \times C}$):
  - Captures real-time field conditions (Sentinel-2 imagery).
- **Short-term Meteorological Data** ($y_s \in R^{T \times G \times N1 \times dy}$):
  - Daily weather data (HRRR dataset) during the growing season.
- **Long-term Meteorological Data** ($y_l \in R^{T \times N2 \times dy}$):
  - Monthly weather data over past years (HRRR dataset).
- **Ground-truth Crop Data** ($z \in R^{d}$):
  - Historical crop data (USDA dataset).

Mathematical Formulation:

- $\hat{y} = f(x, y_s, y_l, z)$

Model f learns relationships between satellite imagery, weather data, and ground-truth crop data.

Challenges Approached in Paper:

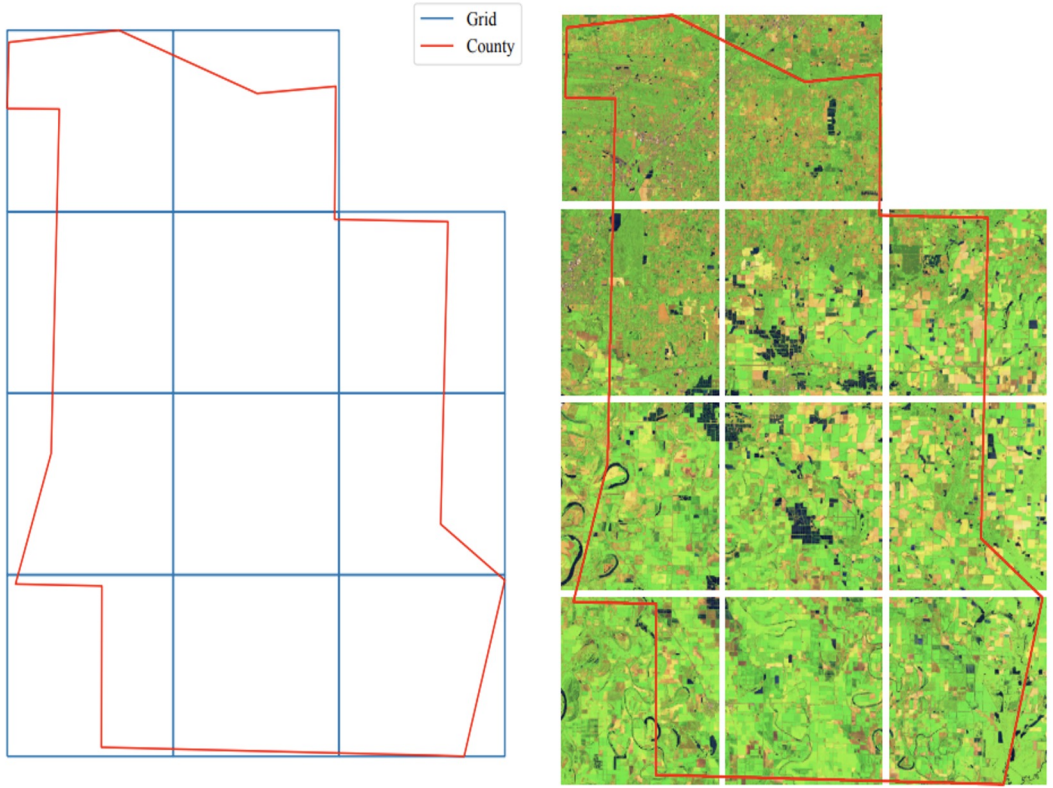
- **Capturing the Effect of Growing Season Weather Variations**:
  - Previous studies focus only on meteorological or remote sensing data, ignoring the direct impact of growing season weather on crop growth.
- **Lack of Pre-training Mechanism for Multi-Modal Models**:
  - Vision Transformer (ViT)-based models are prone to overfitting, needing pre-training for better performance.
  - SimCLR only marginally improve crop yield prediction performance
- **Capturing the Impact of Climate Change on Crops**:
  - Climate change has a gradual negative effect on crop yield, but current models overlook this impact.

Dataset

Sentinel-2 Imagery: A set of images captured by the Sentinel-2 Earth observation satellite.

Features:

- Set of 348x348 RGB Images
- 2- Week interval
- County Partitioned into multiple fine-grained grids(9X9 Km)



(a) Grid Example (b) Sentinel-2 Imagery

https://huggingface.co/datasets/fudong03/Tiny-CropNet

Dataset:

HRRR Computed Dataset: Provides high-resolution meteorological data for the contiguous U.S. continent.

Short-term: Contains **daily** meteorological data, crucial for assessing immediate weather impacts on crop growth

Long-term: Includes **monthly** data, used for understanding and predicting the long-term effects of climatic changes on crop yields.

Features used:

- Averaged Temperature(K)
- Maximal Temperature(K)
- Minimal Temperature(K)
- Precipitation(Kg m**-2)
- Relative Humidity(%)
- Wind Gust(m/s)
- Wind Speed(m/s)
- Downward Shortwave Radiation Flux(w/m^2)
- Vapor Pressure Deficit(kPa)

https://huggingface.co/datasets/fudong03/Tiny-CropNet

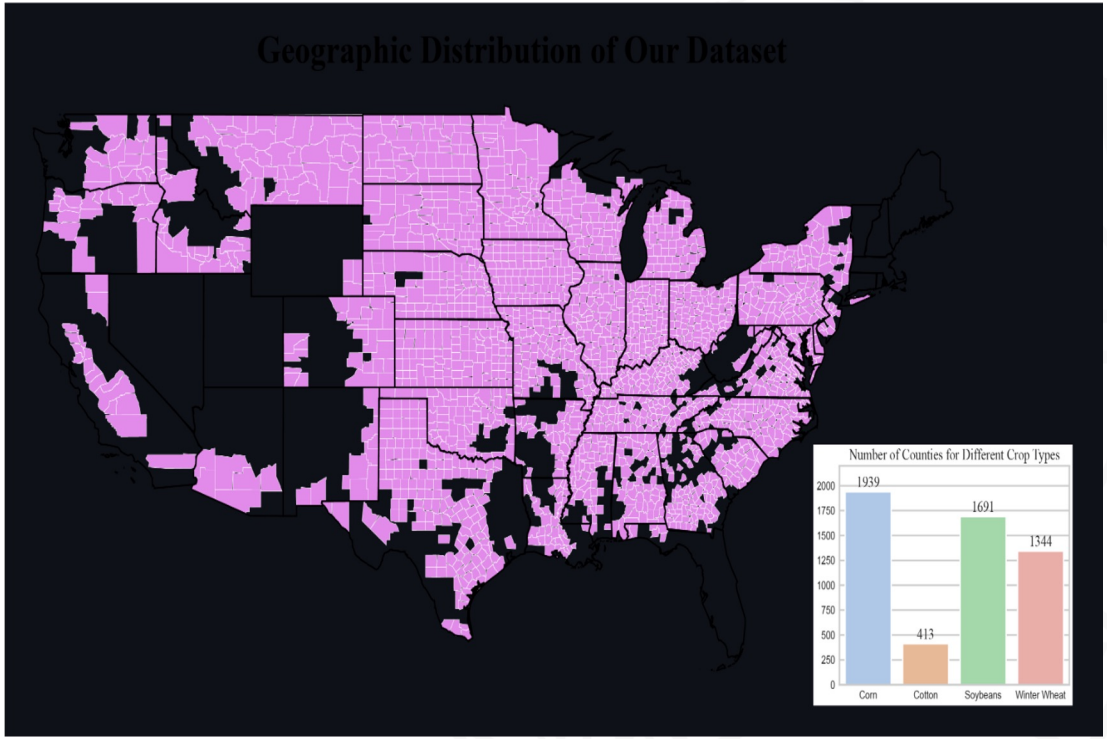
Dataset:

USDA Crop Dataset: Provides annual crop data for major crops grown on a county-level basis.

Crops included, corn, cotton, soybean and winter wheat

Features:

- Production,
- Yield, both measured in BU.



https://huggingface.co/datasets/fudong03/Tiny-CropNet

Proposed Approach

1) **Multi Modal Transformer** - To capture impact of short term meteorological variations on crop growth.
2) **Spatial Transformer** - Learns global spatial information of a county.
3) **Temporal Transformer** - To capture temporal information as well as impact of long term climate change on crop yields.
4) **Linear classifier** for predicting annual crop yields.

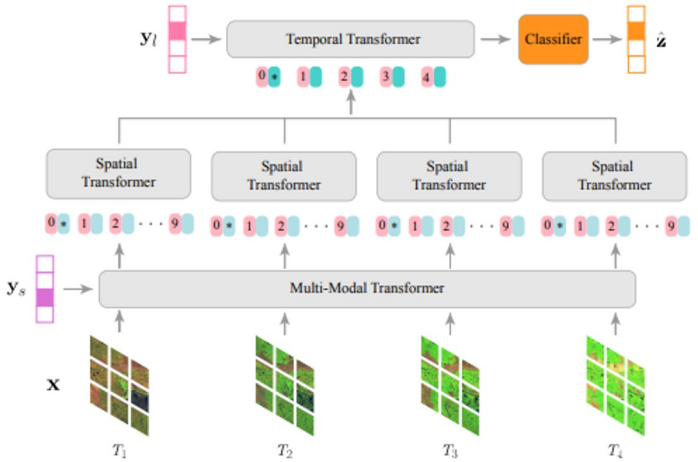


Figure 2: The architecture of our proposed MMST-ViT.

Continuation(Lecture 2)

- Overview of first lecture
- Multi Modal Transformer (Pyramid Vision Transformer)
- SimCLR Technique
- Spatial Transformer
- Temporal Transformer
- Experiments
- Comparative Performance Evaluation
- Conclusion

References:

- MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer
  https://arxiv.org/pdf/2309.09067
- An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
  https://arxiv.org/abs/2010.11929
- A GNN-RNN Approach for Harnessing Geospatial and Temporal Information: Application to Crop Yield Prediction
  https://arxiv.org/pdf/2111.08900
- An Open and Large-Scale Dataset for Multi-Modal Climate Change-aware Crop Yield Predictions
  https://arxiv.org/html/2406.06081v1

Q&A