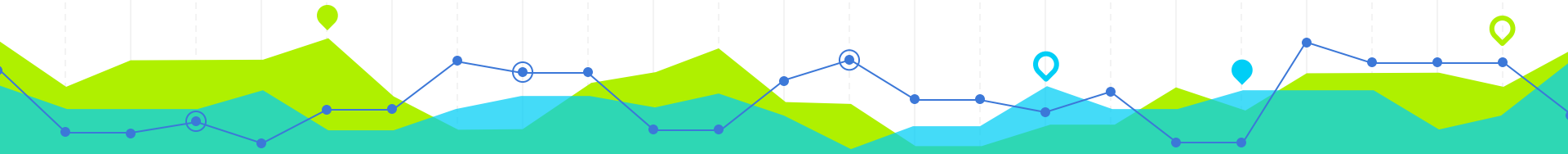


Using Transformers in Image Classification

Zainab Mohammad
Rajit Puzhakkarezhath
Aviral Mehrotra
Nicholas Chiu

Introduction

- Transformers were introduced in 2017 and have become widely used in NLP applications
- In 2021, vision transformers were adapted for computer vision applications

"AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE"
http://arxiv-export-lb.library.cornell.edu/pdf/2010.11929

- Vision transformers (ViT) are scalable and outperform while still being relatively cheap to fine-tune

Problem

- *Motivation:* Image classification now typically uses a CNN, so what is the best moving forward?
- **Goal:** Build and train our own vision transformer model and compare it to CNN models (our own and ResNet50)
- Will our results show similarities with the original paper? What does this say about the performance and approach of using transformers in the domain of computer vision?

Dataset

- EuroSAT
  - 27,000 labelled 64x64 satellite images
  - Classified into 10 different land use classes (e.g. highway, forest, river, etc)
- Using real map data as opposed to simpler images better reflects the viability of using transformers in image classification.

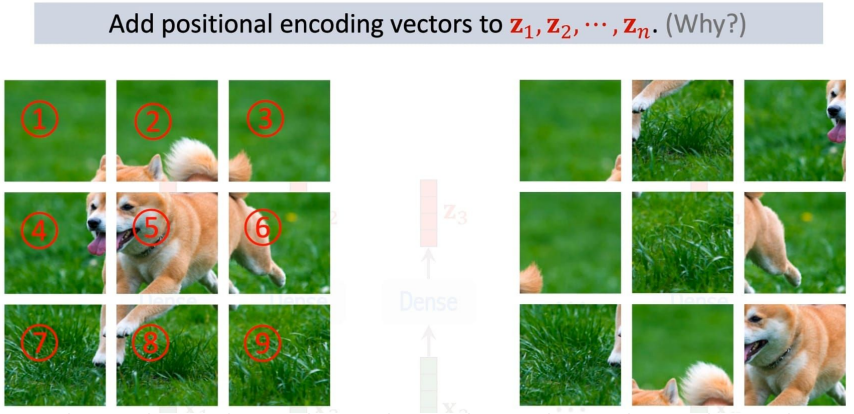
Annual Crop

Forest

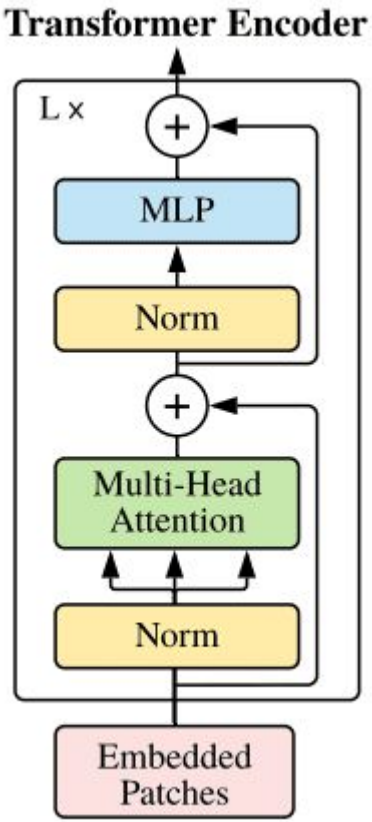
Herbaceous Vegetation

Approach

- First, we convert the images to pixel values.
- We then perform data augmentation and we take image patches and convert them to 1D sequences.
- Positional embeddings are added to these sequences.

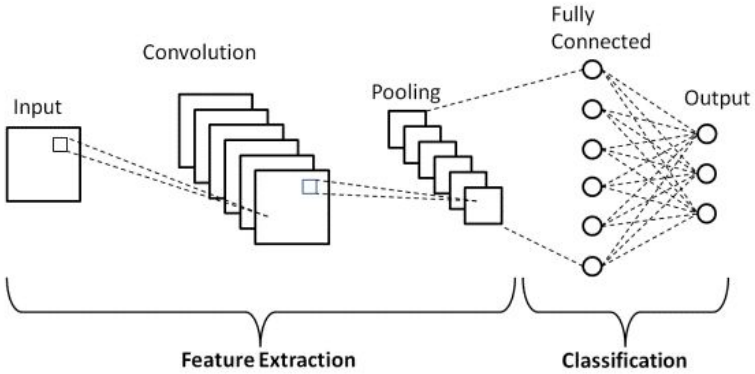


Vision Transformer

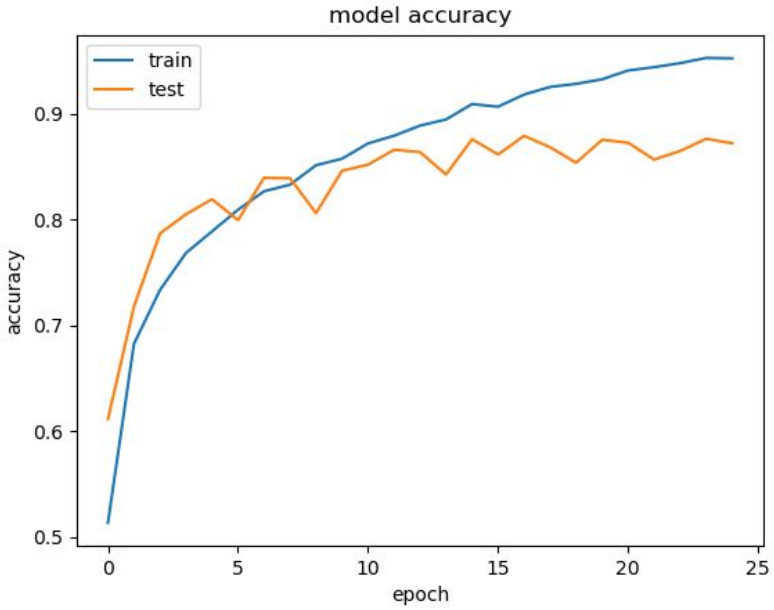


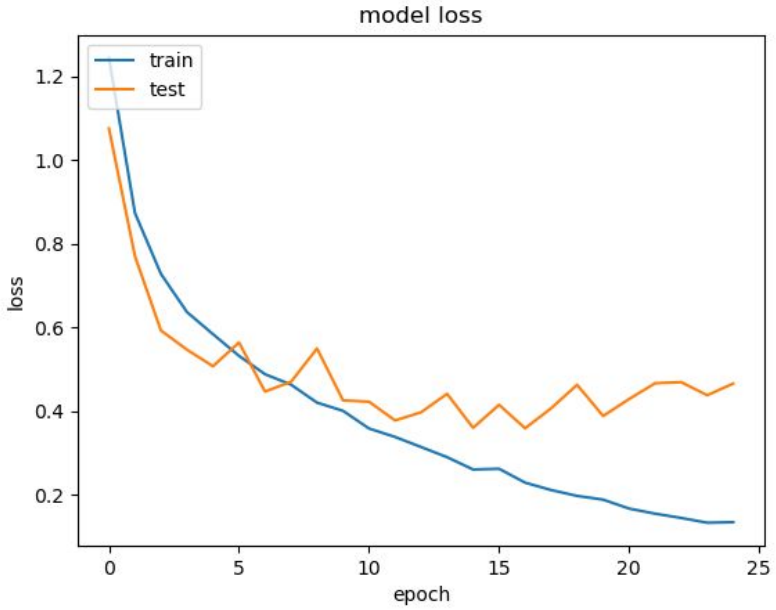
CNN Model

- 5-layer CNN as baseline to compare
  - 3x3 kernel
  - ReLU activation
  - MaxPooling
- Dropout of 10% to prevent overfitting

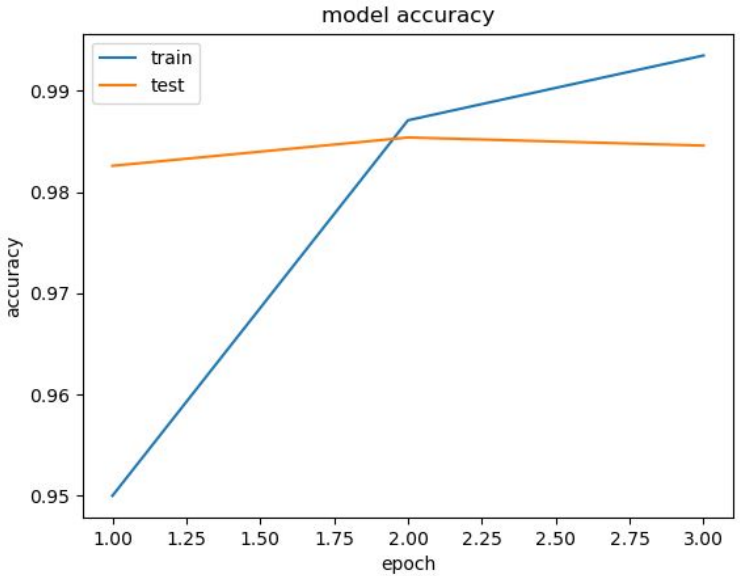


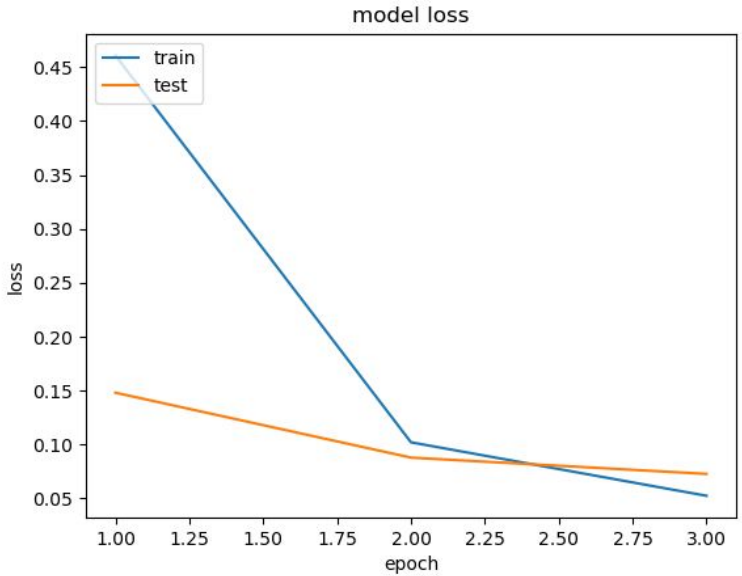
CNN Results


model accuracy
train
test
accuracy
epoch
0.5
0.6
0.7
0.8
0.9
0
5
10
15
20
25


model loss
train
test
loss
epoch
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25

Vision Transformer Results





Evaluation

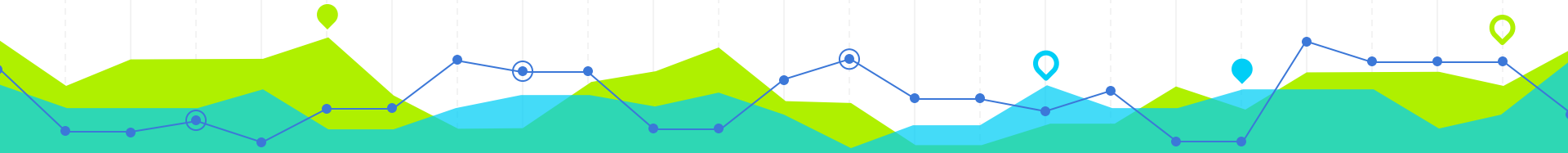
	Our CNN (25 epochs)	ResNet50 (25 epochs)	Vision Transformer (3 epochs)
Trainable Parameters	157,018	17,132,490	85,806,346
Time per Epoch	8-9s	22-23s	>=5000s (83 mins)
Validation Accuracy	87.20%	94.69%	98.46%
Validation Loss	0.4658	0.2988	0.0728

Lessons Learned

- The fundamental approach was to create an image classification problem by using image patches as tokens in a sequence, and then processing it with a transformer.
  - Need to be trained on large amounts of data
- ViT requires huge amounts of computational resources and time when training from scratch
  - Has the potential to beat other state-of-the-art CNN architectures with a more powerful machine and pretrained data

Future Work

- Current trends are suggesting a shift to the transformer architecture.
  - A study in June 2021 added a transformer backend to ResNet which reduced costs and increased accuracy.
  - Tesla is now also using transformers in its autopilot system.
- Future computer vision could move towards transformers instead of CNNs



Thank You!

Questions?