Deep Learning CS577

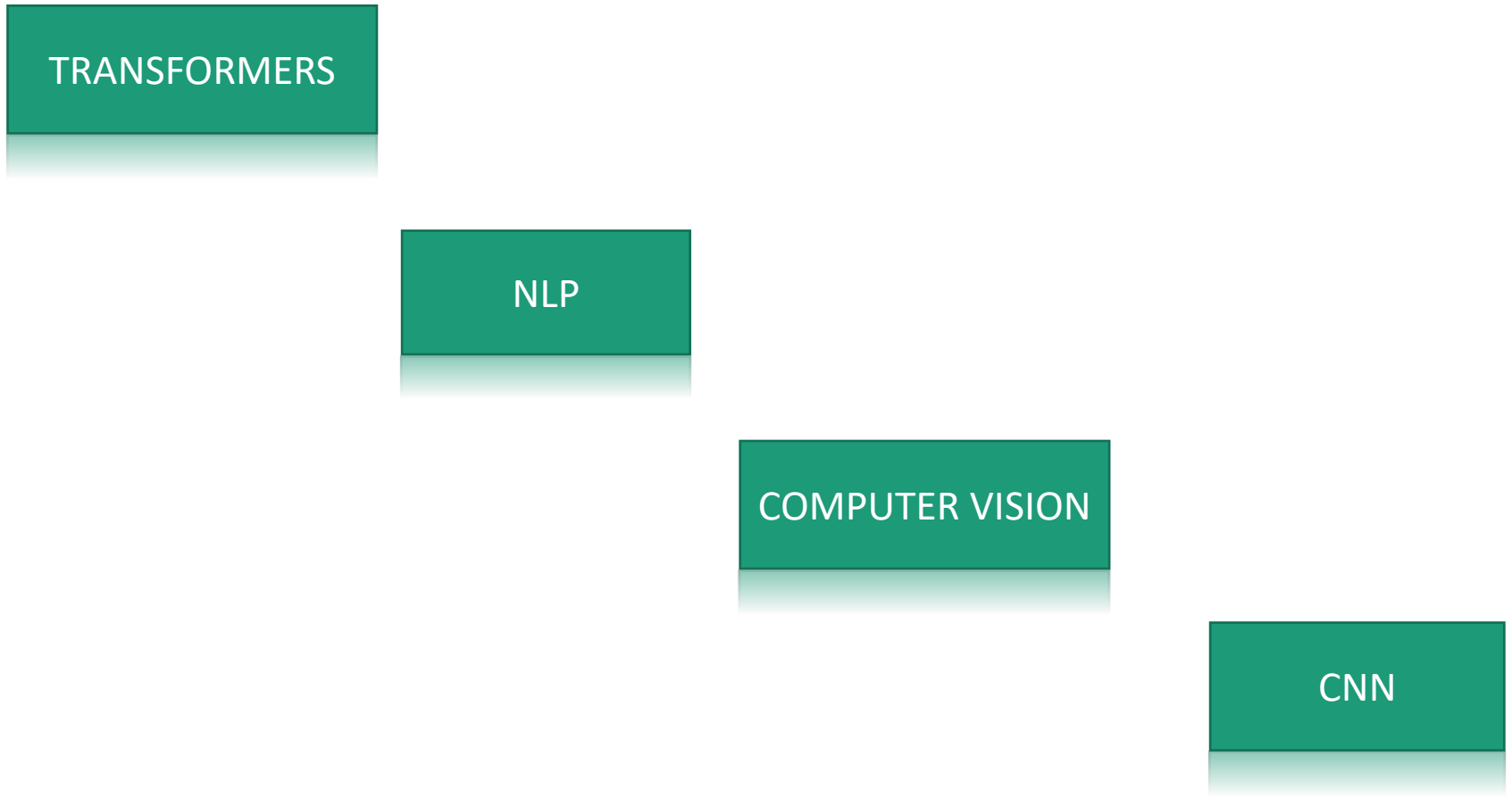
IMAGE CLASSIFICATION WITH VISION TRANSFORMER

Mohammad Ausaf Ali Haqqani
(A20516413)

Bhoomika Panduranga
(A20503493)

Contents:

Abstract



Problem Statement

CNN Classifier << Pure Transformer

Proposed Solution

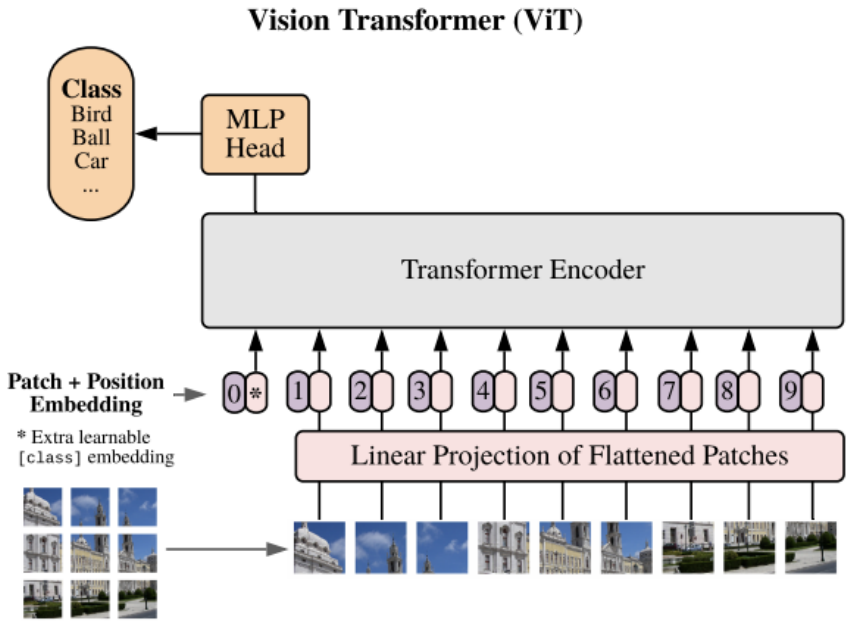
ViT

CIFAR100

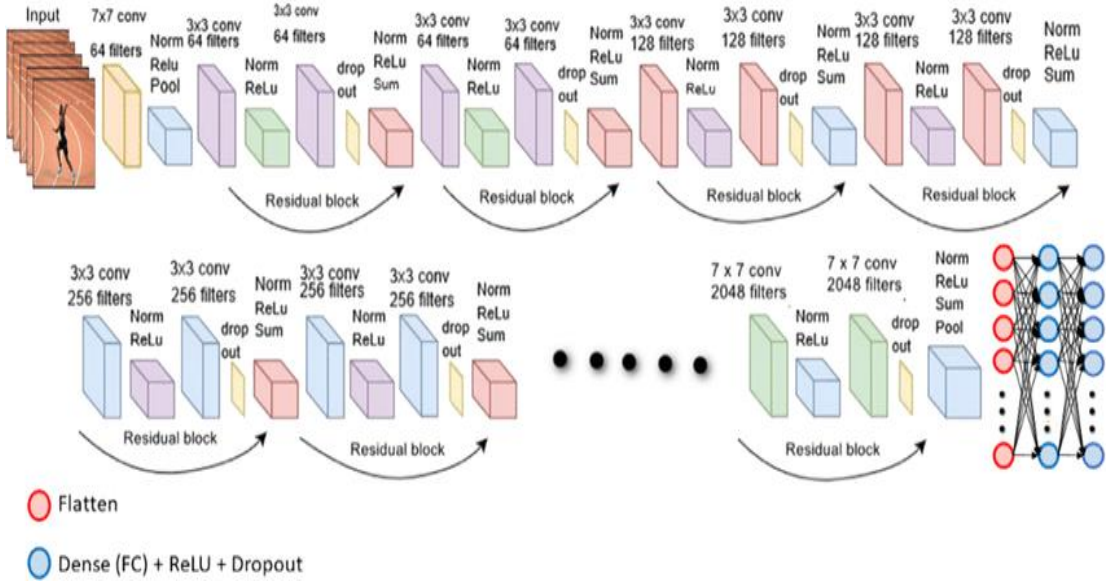
Tokens ->Patches

Supervised Learning

IMPLEMENTATION

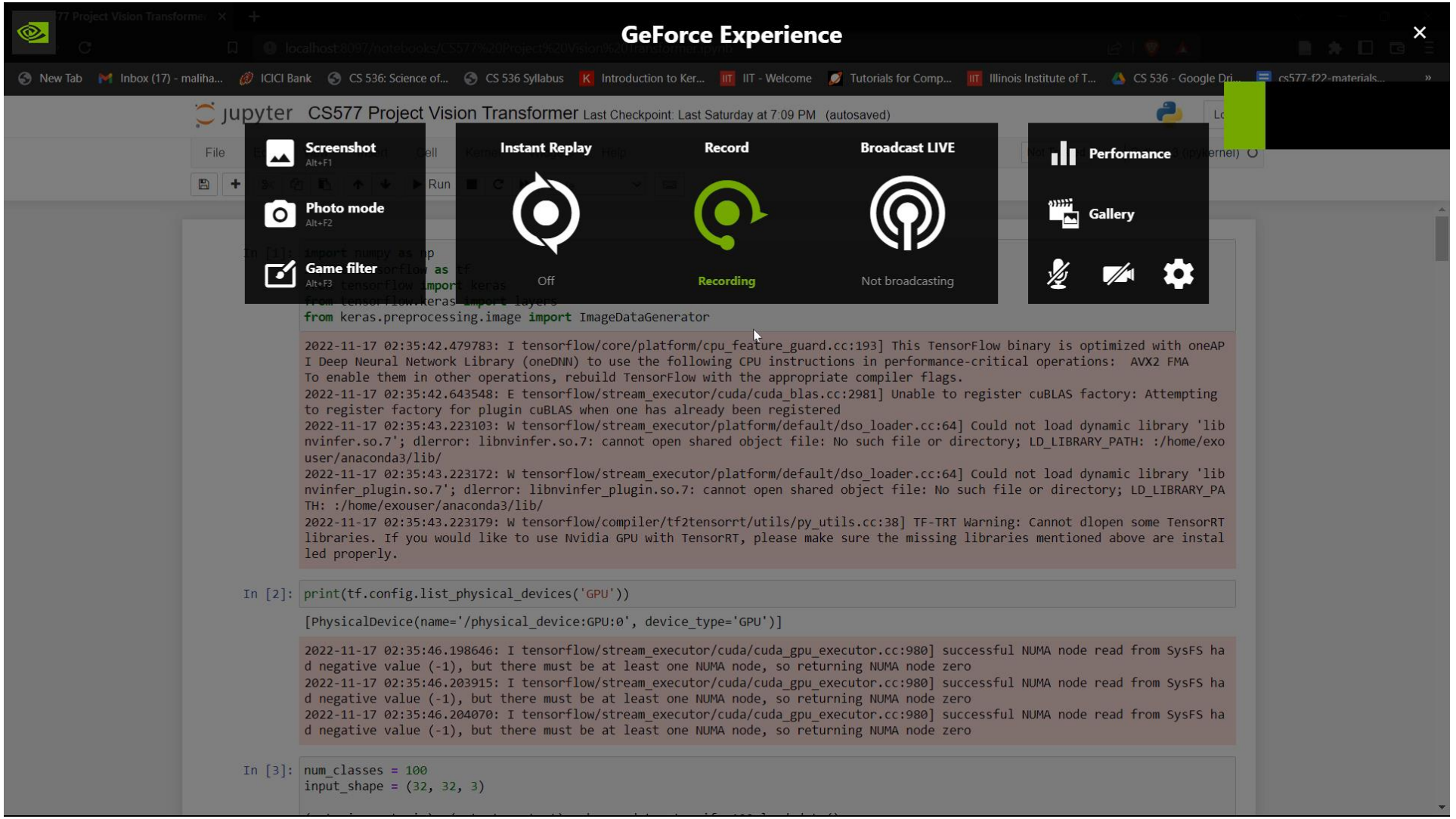


Transformer Architecture

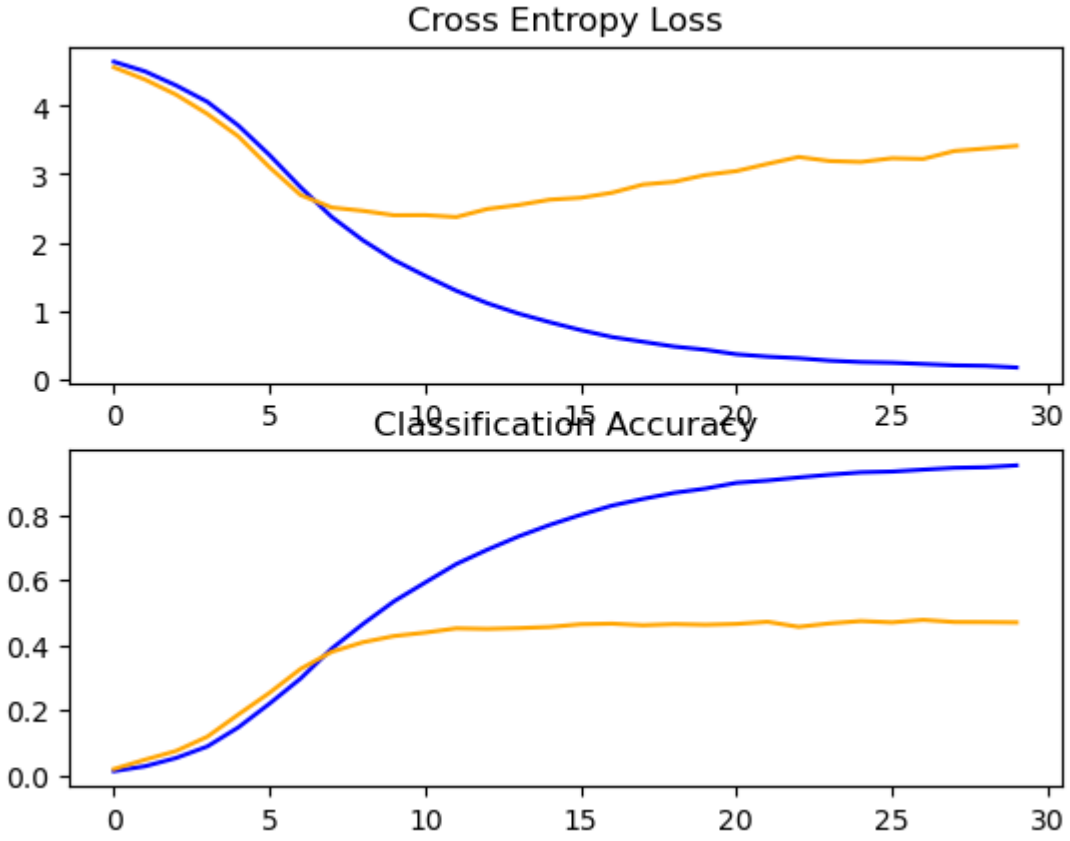


ResNet101V2 Architecture

IMPLEMENTATION

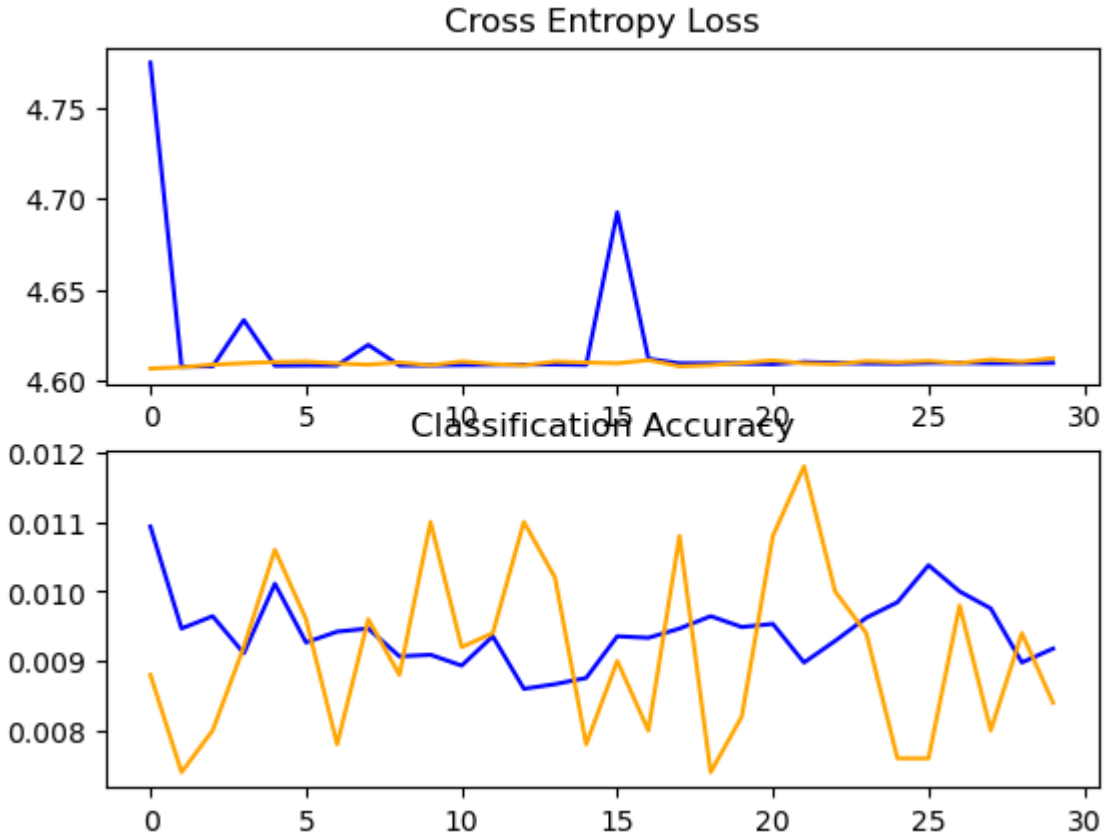


Results



Resnet101v2- 30 epochs

```
313/313 [==============================] - 3s
10ms/step - loss: 3.3324 - accuracy: 0.4793 the
accuracy is: 47.929999232292175
```



Transformer – 30 epochs

```
313/313 [==============================] - 5s
17ms/step - loss: 4.6079 - accuracy: 0.0100 -
top-5-accuracy: 0.0500 Test accuracy: 1.0% Test
top 5 accuracy: 5.0%
```

Conclusion

- Large Dataset
- Scalable
- Economical
- TPU

Future work

- Pretrain on large datasets
- ViT – Segmentation and detection
- Explore self supervised pre- training methods
- Further scaling

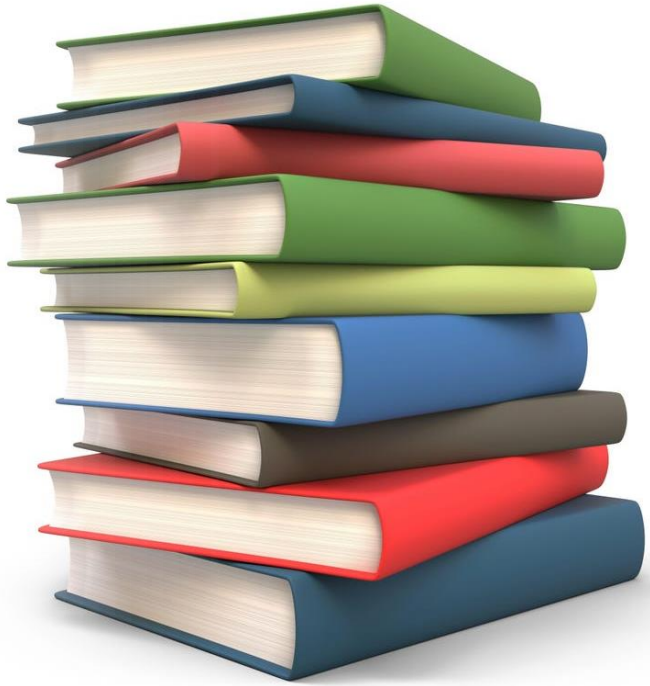
References

[1].https://keras.io/examples/vision/image_classification_with_vision_transformer/

[2]. https://arxiv.org/abs/2010.11929

[3]. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens. Attention augmented convolutional networks. In ICCV, 2019.

[4]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2019.



Thank you