TRANSFORMING

Image Recognition and Classification

(Using CIFAR 100 & Google Landmarks Dataset v2)

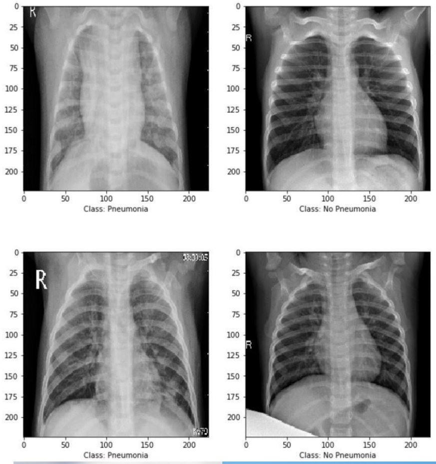
Spring 2023 Term: W281 Computer Vision Final Project

Waqas Ali | Pedro Melendez | Prakash Krishnan

Berkeley
UNIVERSITY OF CALIFORNIA

Image Classification

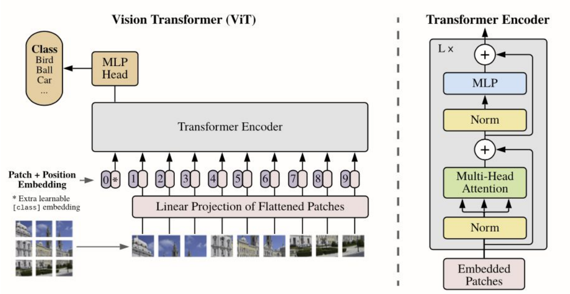
- ❑ Essential Task in Computer Vision
- ❑ Considerable Research on Handcrafted Techniques with Traditional Classifiers (SVM, RF, Gradient Boost) and deep learning models (CNN and ViT)
- ❑ Use Cases – Medical Imaging, Satellite Imaging, Self Driving Cars
- ❑ Emergence of Vision-Language Models



Berkeley
UNIVERSITY OF CALIFORNIA

Our Inspiration

- An Image is Worth 16x16 Words - Transformers for Image Recognition at Scale (Dosovitskiy et al., Google Research, 2021)
- Self-Attention based transformer architecture (Source: Attention Is All You Need, Ashish Vaswani et al., 2017) adapted for images
- Vision Transformer Steps
    - Break image into patches (fixed sizes)
    - Integrate positional embeddings
    - Feed the sequence to the transformer encoder
    - Pre-train the ViT model with image labels
    - Fine-tune the downstream dataset for image classification



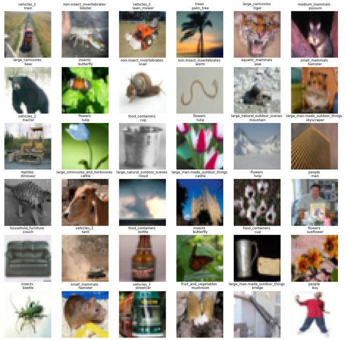
Source: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Berkeley UNIVERSITY OF CALIFORNIA

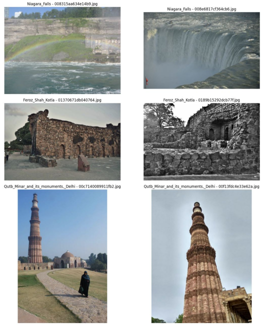
Research Questions

1. How well does a Pre-Trained Vision Transformer Model (ViT) that is Fine Tuned compare to CNN models (ResNet 9 & ResNet 50) and models built on manual feature extraction techniques (HOG+SVM, SIFT+SVM, GLCM+SVM) ?
2. How does the state-of-the-art ViT model generalize and scale ?
3. What is the future of ViTs ?

CIFAR-100, 50k Train, 10k Test, 20 Coarse Class and 100 Fine Class



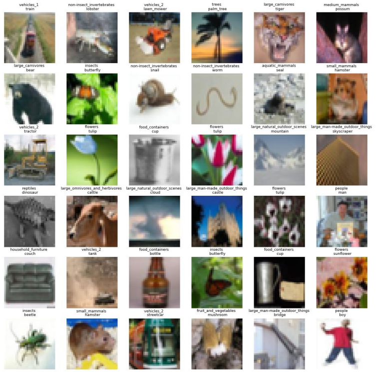
Subset of GLDv2 dataset with only 20 classes and 16,633 images in total



Berkeley
UNIVERSITY OF CALIFORNIA

CIFAR-100 Dataset EDA

- Dataset has 100 classes (fine labels) containing 600 images each.
- There are 500 training images and 100 testing images per class.
- The 100 classes in the CIFAR-100 are grouped into 20 superclasses (course labels).
- Each image (32x32) comes with a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs).

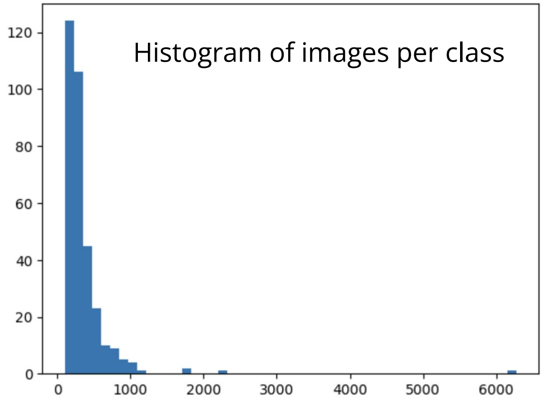


Coarse_Label_Name	Fine_Label_Name
aquatic_mammals	[otter, seal, whale, beaver, dolphin]
fish	[aquarium_fish, shark, flatfish, ray, trout]
flowers	[sunflower, rose, tulip, poppy, orchid]
food_containers	[cup, bottle, plate, bowl, can]
fruit_and_vegetables	[apple, mushroom, sweet_pepper, orange, pear]
household_electrical_devices	[telephone, keyboard, television, clock, lamp]
household_furniture	[table, chair, wardrobe, couch, bed]
insects	[cockroach, butterfly, bee, caterpillar, beetle]
large_carnivores	[wolf, leopard, lion, tiger, bear]
large_man-made_outdoor_things	[castle, skyscraper, road, bridge, house]
large_natural_outdoor_scenes	[cloud, sea, mountain, forest, plain]
large_omnivores_and_herbivores	[cattle, elephant, chimpanzee, camel, kangaroo]
medium_mammals	[possum, skunk, raccoon, fox, porcupine]
non-insect_invertebrates	[lobster, snail, worm, cra, spider]
people	[boy, woman, girl, man, baby]
reptiles	[dinosaur, snake, crocodile, turtle, lizard]
small_mammals	[squirrel, shrew, rabbit, hamster, mouse]
trees	[willow_tree, pine_tree, oak_tree, maple_tree,...
vehicles_1	[train, bicycle, motorcycle, bus, pickup_truck]
vehicles_2	[streetcar, tractor, rocket, tank, lawn_mower]

Berkeley
UNIVERSITY OF CALIFORNIA

GLDv2 Dataset EDA

- Main dataset contains over 4 million images of landmarks across the word, we used Kaggle subset with:
  - 1,580,470 images and 81,313 Classes (105.52 GB)
  - Only 48 Classes had more than 500 images
  - We used the top 20 classes with less than 1,000 images to avoid outliers and unbalanced classes
  - Most images had 800x800 resolution so we blurred and downsized to 200x200
  - Our dataset included 16,633 images (1.8GB)

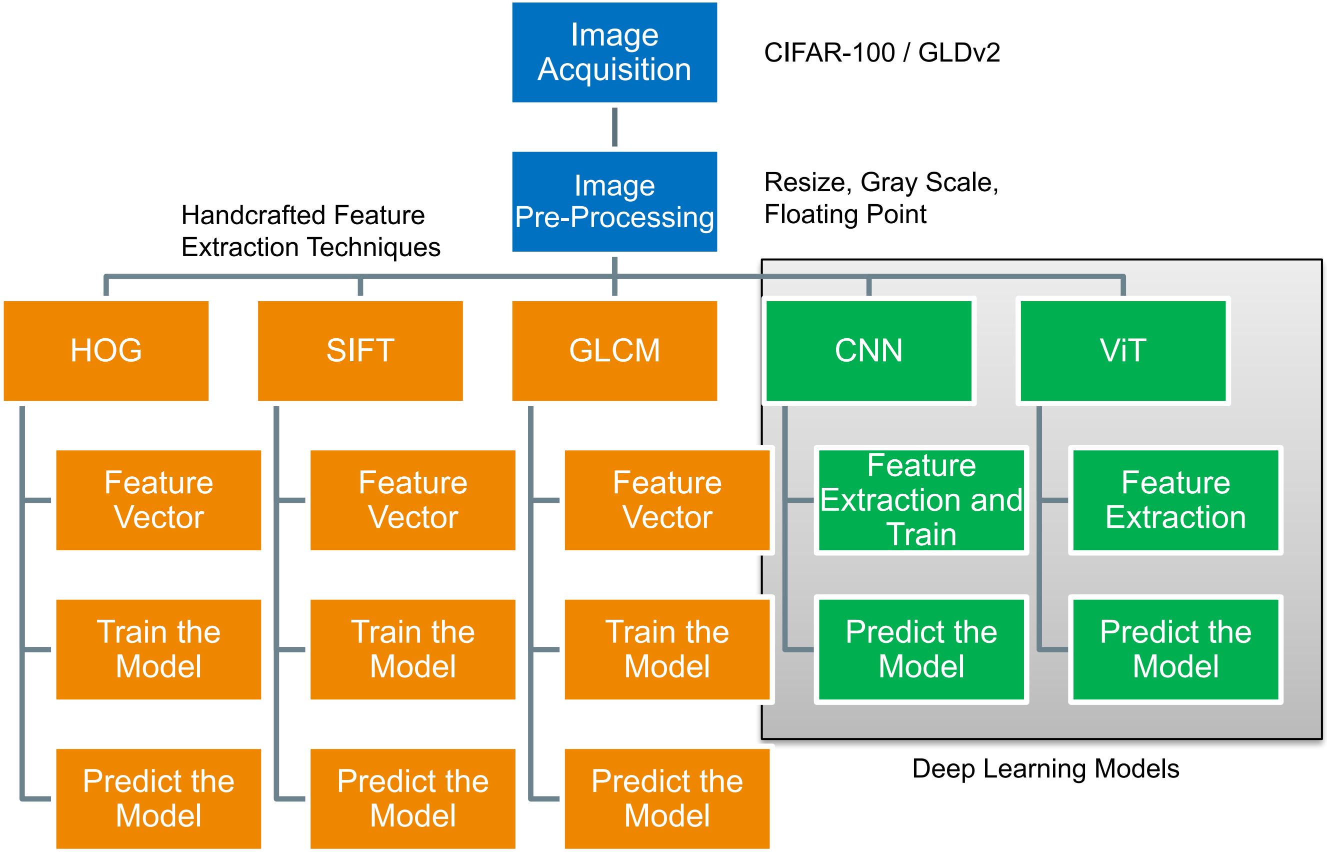


Berkeley
UNIVERSITY OF CALIFORNIA

Methodology

Evaluation Metrics on Test Data

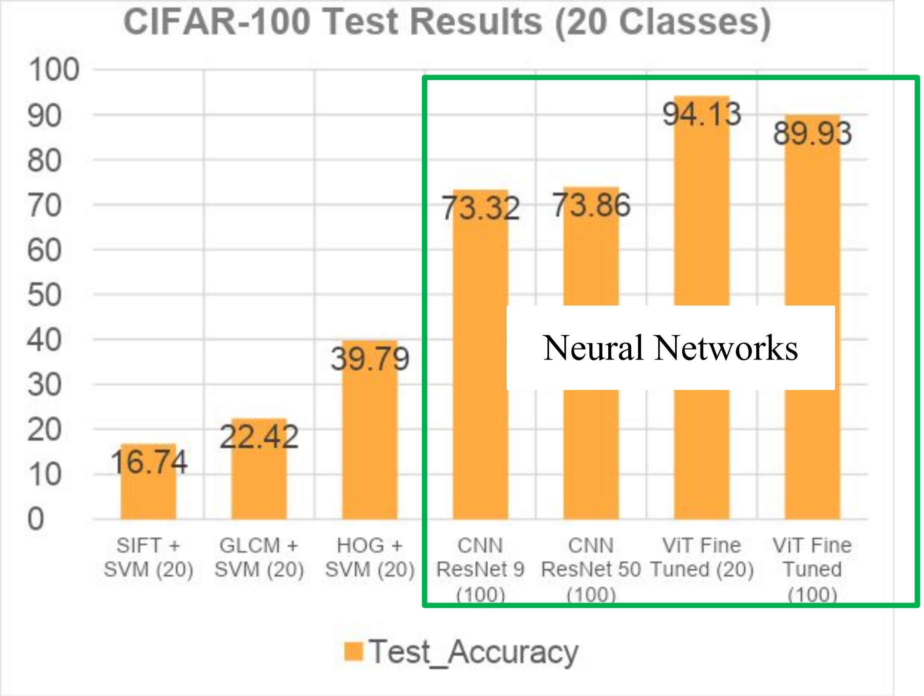
- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix
- Visual Inspection of Predicted Labels



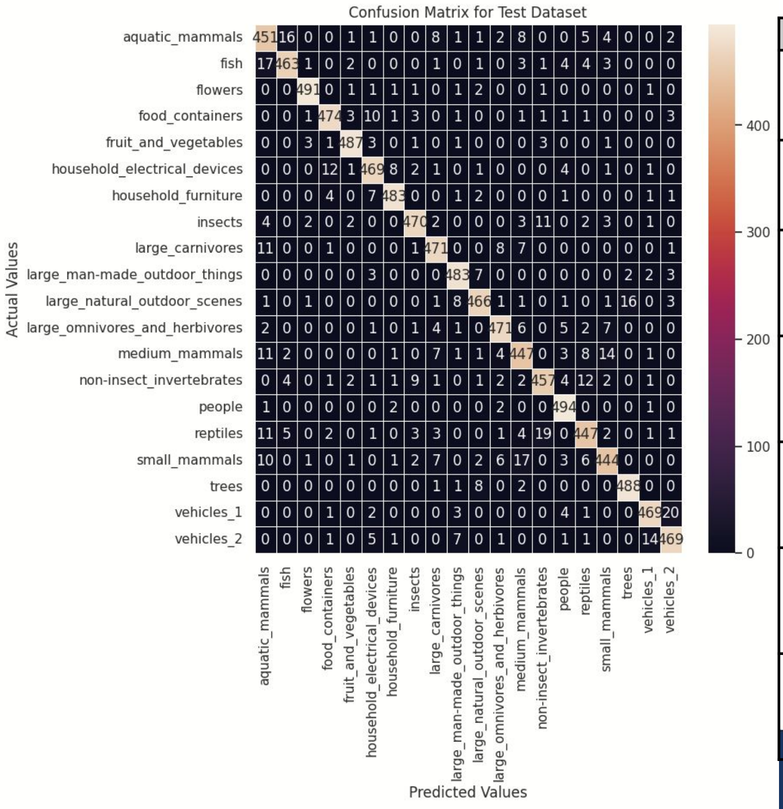
Experimental Results on CIFAR-100

Key Takeaways

1. Vision Transformer outperforms CNN and Handcrafted Models on Test Accuracy
2. Vision Transformer scales well even when number of classes go from 20-100
3. Best CNN performance on ResNet 50 but still inferior to Vision Transformer Model
4. Best performance from handcrafted models is HOG + SVM. But performance is inferior to deep learning models
5. Handcrafted feature extraction techniques do not scale well when dealing with multiple features like edges, blobs, regions, textures, shapes.



Experimental Results on CIFAR-100 - Continued



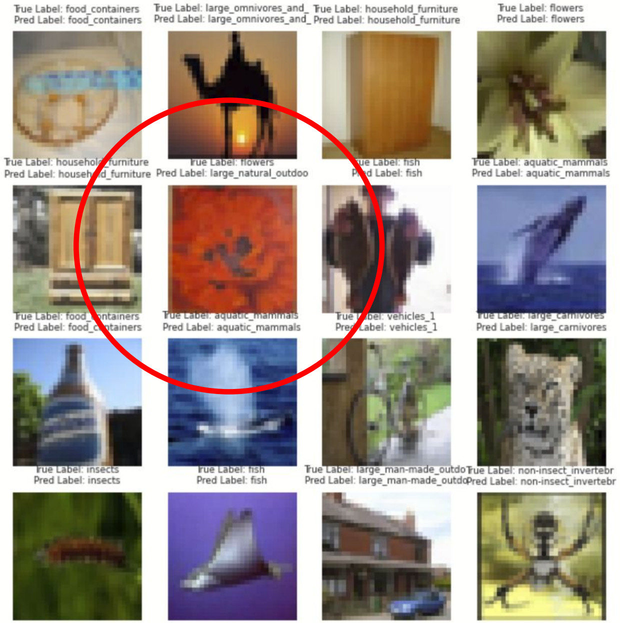
Coarse Class	Accuracy	Coarse Class	Accuracy	Coarse Class	Accuracy
aquatic_mammals	89.6%	insects	93.2%	people	98.2%
fish	93.8%	large_carnivores	92.0%	reptiles	89.6%
flowers	98.4%	large_man-made_outdoor_things	94.8%	small_mammals	91.2%
food_containers	94.0%	large_natural_outdoor_scenes	95.4%	trees	97.0%
fruit_and_vegetables	97.2%	large_omnivores_and_herbivores	94.0%	vehicles_1	93.2%
household_electrical_devices	93.6%	medium_mammals	93.0%	vehicles_2	94.8%
household_furniture	96.0%	non-insect_invertebrates	93.6%		

Key Takeaways

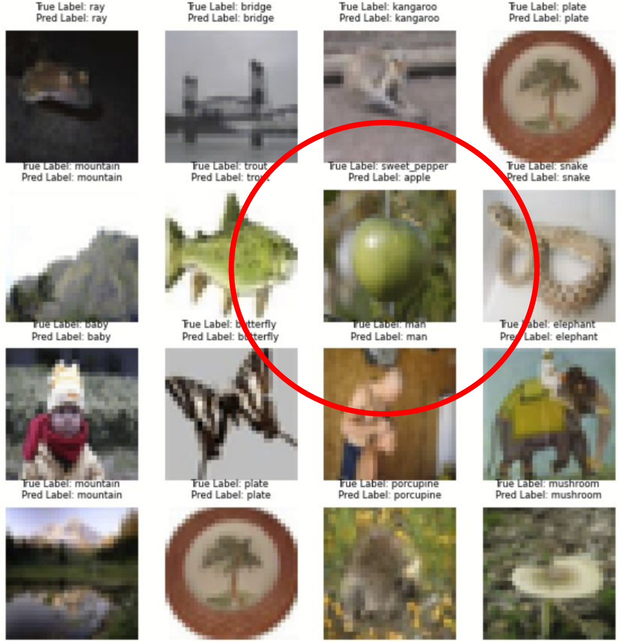
- Classes that are distinct and have good features have higher accuracy rates.
- As an example flowers, people, trees, fruits and vegetables have the highest accuracy rates.
- Aquatic mammals and reptiles have the lowest accuracy rates.
- These classes are not distinctive and have other classes such as small mammals and medium mammals that are close.

Sample Test Images and Predictions from CIFAR- 100

Coarse Label Prediction



Fine Label Prediction



Experimental Results on GLDv2

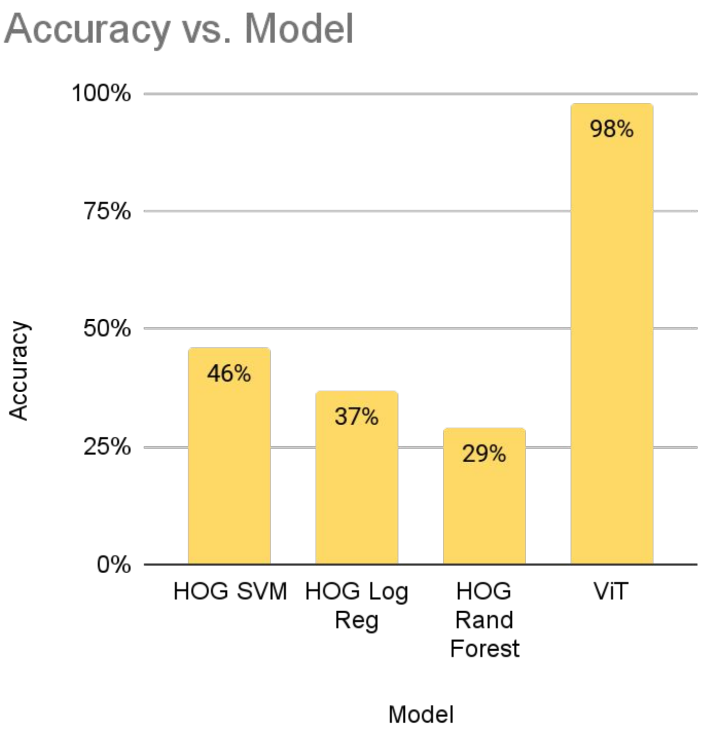
Key Takeaways

1. Extremely high accuracy (98%) with Vision Transformer

2. Vision Transformer accuracy 2x of Handcrafted Models

3. Vision Transformer had better accuracy with GLDV2 compared to CIFAR due to better resolution on images

4. HOG+SVM had higher accuracy compared to HOG+LR and HOG+RF



Generalizability CIFAR-100

	Test Evaluation			
Models	Accuracy	Precision	Recall	F1-Score
ViT - CIFAR-100 (20 Classes)	94.13	94.15	94.13	94.13
ViT - CIFAR-100 (100 Classes)	89.93	90.11	89.30	89.95
CNN-ResNet 9 (100 Classes)	73.32	73.56	73.32	73.32
CNN-ResNet 50 (100 Classes)	73.86	74.1	73.86	73.85

Key Takeaways

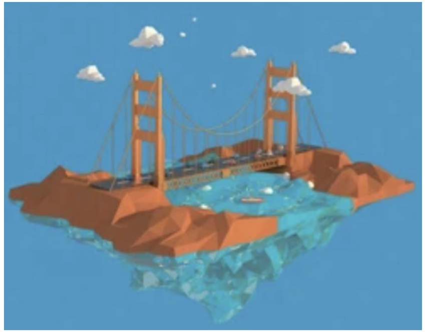
ViT on CIFAR-100 does remarkably well on Test Accuracy with 94.13% for 20 course labels and 89.93% on 100 fine labels.

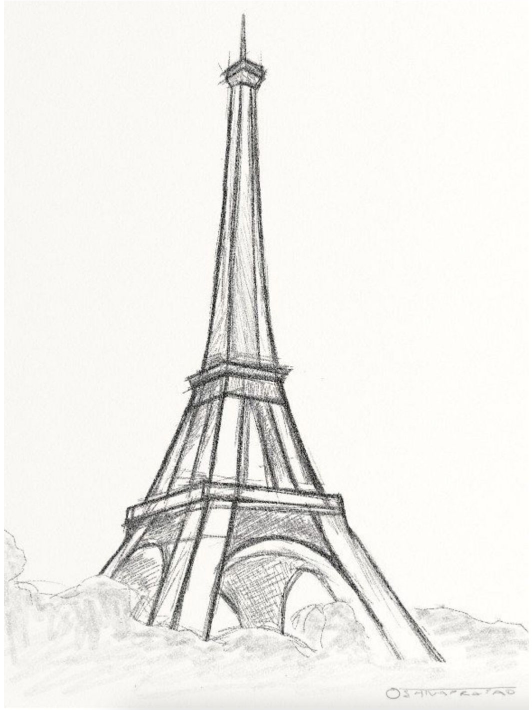
CNN-ResNet 9 and 50 perform well but inferior to ViT (CNN is not pre-trained)

Berkeley
UNIVERSITY OF CALIFORNIA

Generalizability GLDv2

- ViT model classified correctly even draw images of some of the landmarks
- Tested with low detailed sketches and with full colored drawings and still got the correct label





Conclusion

- Vision Transformers achieve outstanding classification accuracy across large, complex datasets (including GLDv2) even when the numbers of classes are large (100).
- Traditional techniques work well on datasets that have homogeneous features but underperform when images have non-homogeneous features.
- Future studies can include experiments to determine under what conditions a ViT may underperform CNN.
- Combining features extracted by a ViT with handcrafted techniques can address some of the intra-class variations and inter-class similarities especially with medical-imaging applications.
- Using multi-modal transformer models with NLP and CV can lead to interesting use cases.