Introduction

Three small pretrained Vision Transformer (ViT) were used for image classification and attention visualization.

- google/vit-base-patch16-224
- facebook/deit-base-patch16-224
- google/vit-large-patch16-224

The models were pretrained on ImageNet. Images were resized to 224×224 before being fed into the model. The main goals were to

01. Record top-1 predictions and check if they make sense,
02. Visualize patch attention to see where the model looks,
03. Experiment with patch masking, and
04. Compare linear probes with different pooling methods.

Results using google/vit-base-patch16-224

Top-1 and Top-5 Predictions

We tested the model on multiple images, some results are:

1. A cat lying on the floor

Top-1: tabby, tabby cat (p=0.722)
Top-5:

1.	tabby, tabby cat	p=0.722
2.	tiger cat	p=0.164
3.	Egyptian cat	p=0.067
4.	Persian cat	p=0.030
5.	lynx, catamount	p=0.007

True Label: tabby cat

Top-5 included other cat breeds and “Egyptian cat.” Shows the model was very confident it is a cat.

2. A chair in a room

Top-1: sliding door (p=0.087)
Top-5:

1. rocking chair, rocker p=0.482
2. sliding door p=0.087
3. studio couch, day bed p=0.073
4. dining table, board p=0.036
5. window shade p=0.028

True Label: chair

Top-5 included rocking chair, studio couch,day bed, sliding door, dining table etc. This shows the model has attention both on door in the background and chair in foreground . Model is confusing background objects like (sliding door or window)

3. A red city bus/ Metro

Top-1: passenger car, coach, carriage (p=0.538)
Top-5:

passenger car, coach, carriage	p=0.538
orange	p=0.007
Granny Smith	p=0.002
banana	p=0.000
grocery store, grocery, food market, market	p=0.000

True Label: Metro Bus

Attention Visualization

Attention maps were created using the CLS token from the last layer (averaged across heads). These were upsampled to image size and overlaid as heatmaps.

Cat lying: Strong attention around the face
Chair: The model focused on seat edges, legs and room.
Bus: Attention concentrated on the front and road.
Cat with hat: The model concentrated attention on the hat and predicted the ferret.
Lemons: Multiple spots lit up, especially shiny areas.

This shows that ViT's have divided attentions across images like the cat with hat was wrongly predicted as ferret but had cat in top5. So ViTs generally pays attention to semantically meaningful parts, even when unusual elements (like the cat's hat) are present. With better training or models the accuracy can be increased.

A List of Results of all three models

== google/vit-base-patch16-224 ==
Sample 0: Top-1 -> **basset**, basset hound (0.26) | CIFAR10 label: **dog**
Sample 1: Top-1 -> tailed **frog**, bell toad, ribbed toad, tailed toad, Ascaphus trui (0.98) | CIFAR10 label: **frog**
Sample 2: Top-1 -> warplane, military **plane** (0.06) | CIFAR10 label: **airplane**
Sample 3: Top-1 -> f**ox squirrel**, eastern fox squirrel, Sciurus niger (0.17) | CIFAR10 label: **bird**
Sample 4: Top-1 -> moving **van** (0.81) | CIFAR10 label: **truck**
Sample 5: Top-1 -> **ostrich**, Struthio camelus (0.98) | CIFAR10 label: **bird**

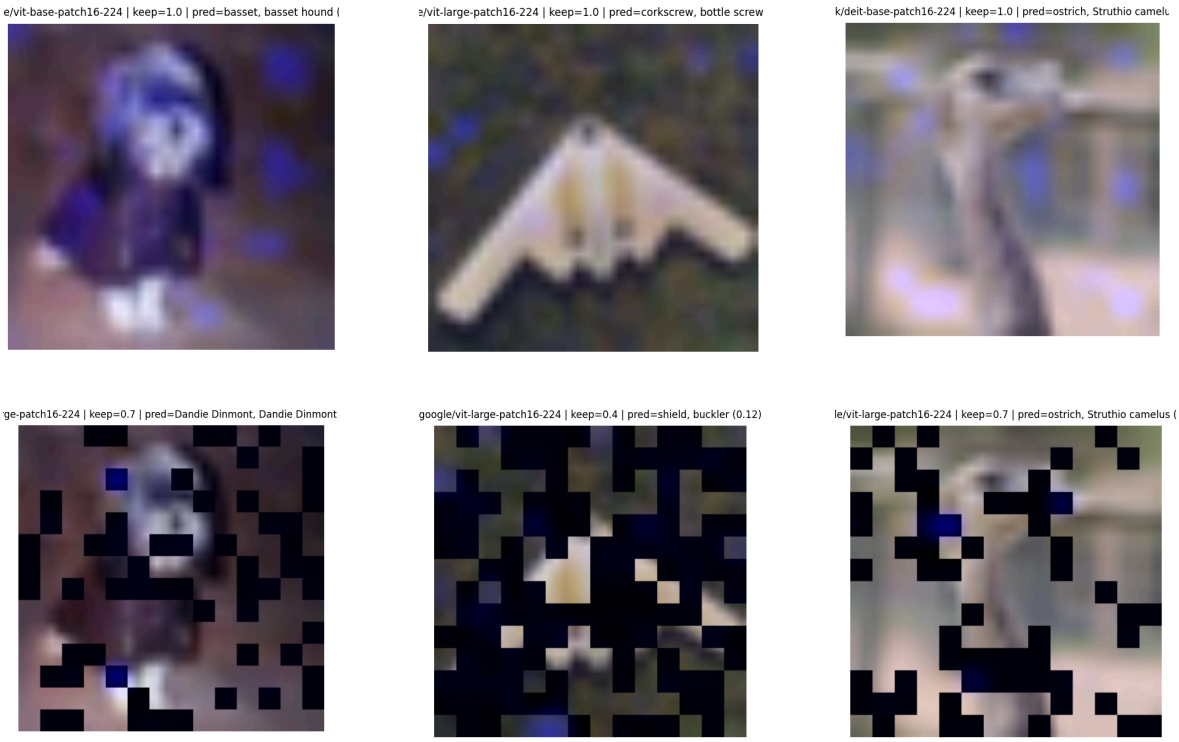
== facebook/deit-base-patch16-224 ==
Sample 0: Top-1 -> **Shih-Tzu** (0.16) | CIFAR10 label: **dog**

Sample 1: Top-1 -> tailed **frog**, bell toad, ribbed toad, tailed toad, Ascaphus trui (0.24) | CIFAR10 label: **frog**
Sample 2: Top-1 -> **panpipe**, pandean pipe, syrinx (0.04) | CIFAR10 label: **airplane**
Sample 3: Top-1 -> **brambling**, Fringilla montifringilla (0.13) | CIFAR10 label: **bird**
Sample 4: Top-1 -> moving **van** (0.08) | CIFAR10 label: **truck**
Sample 5: Top-1 -> **ostrich**, Struthio camelus (0.14) | CIFAR10 label: **bird**

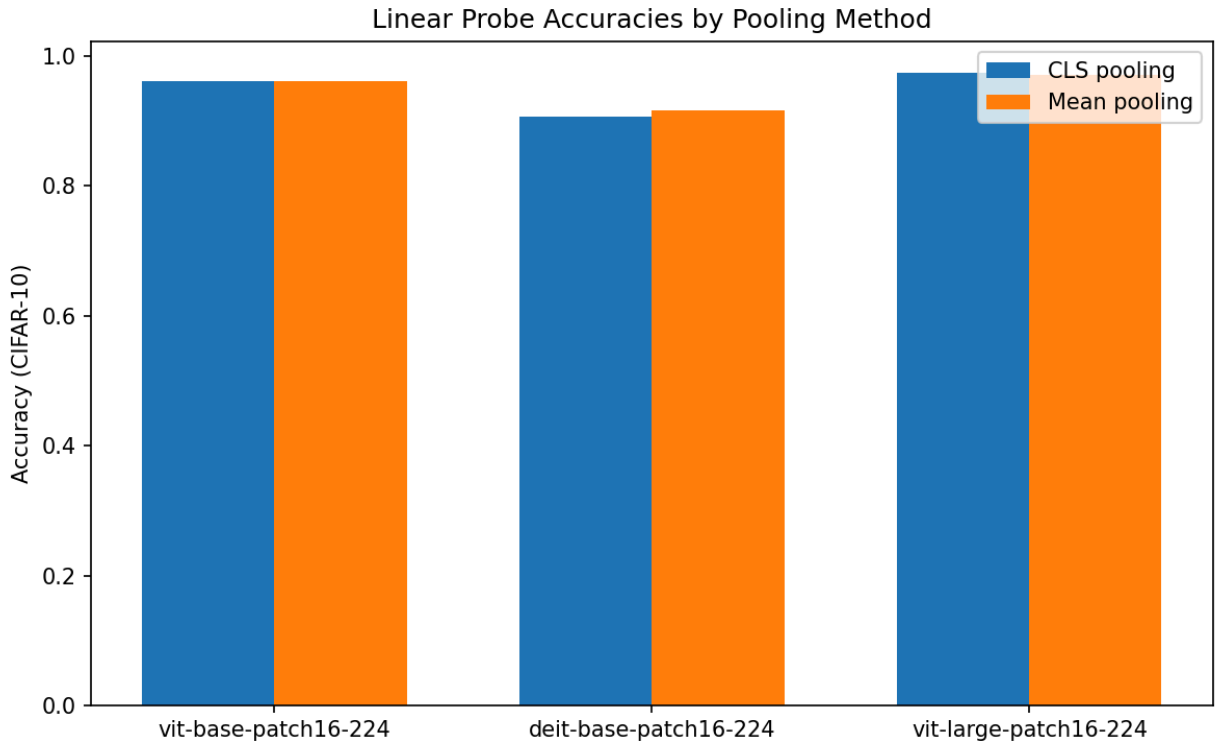
== google/vit-large-patch16-224 ==
Sample 0: Top-1 -> **Dandie Dinmont**, Dandie Dinmont terrier (0.64) | CIFAR10 label: **dog**
Sample 1: Top-1 -> tailed **frog**, bell toad, ribbed toad, tailed toad, Ascaphus trui (0.99) | CIFAR10 label: **frog**
Sample 2: Top-1 -> **corkscrew**, bottle screw (0.34) | CIFAR10 label: **airplane**
Sample 3: Top-1 -> **jacamar** (0.21) | CIFAR10 label: **bird**
Sample 4: Top-1 -> moving **van** (0.71) | CIFAR10 label: **truck**
Sample 5: Top-1 -> **ostrich**, Struthio camelus (0.97) | CIFAR10 label: **bird**

Experimenting with Masking Patches

We tested how robust ViT is to missing patches. We experimented with 70% and 40% of image revealed and the rest masked as shown.



Accuracies were evaluated for all three models on CIFAR-10 dataset using a linear probe (features frozen) .CLS pooling and Mean pooling results were compared



The bar chart shows the probe accuracies. Model size is most important as ViT-large has 97.4 cls and 97.2 mean pooling accuracy, while ViT-Base has same accuracy for cls and mean pooling i.e. 96.2 further DeiT base has performed the least with cls 91.75 and mean pooling accuracy to be 90.5 so bigger models give features that are easier for a linear classifier to separate. From the results CLS performs better for google ViTs.