CNNs vs Vision Transformers: Image Classification

Kelvin Kuang

An Image is Worth 16x16 Words

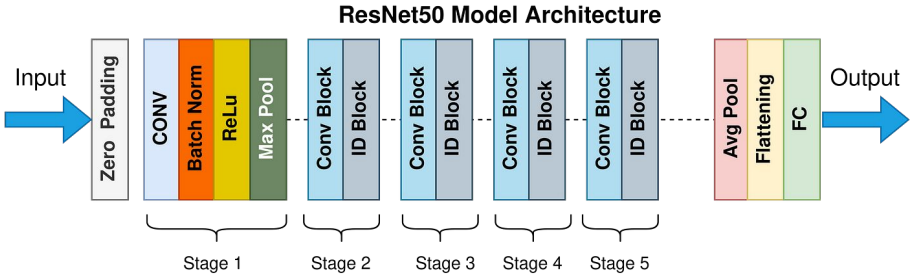
- “Reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks”

name	Epochs	ImageNet	ImageNet ReaL	CIFAR-10	CIFAR-100	Pets	Flowers	exaFLOPs
ViT-B/32	7	80.73	86.27	98.61	90.49	93.40	99.27	55
ViT-B/16	7	84.15	88.85	99.00	91.87	95.80	99.56	224
ViT-L/32	7	84.37	88.28	99.19	92.52	95.83	99.45	196
ViT-L/16	7	86.30	89.43	99.38	93.46	96.81	99.66	783
ViT-L/16	14	87.12	89.99	99.38	94.04	97.11	99.56	1567
ViT-H/14	14	88.08	90.36	99.50	94.71	97.11	99.71	4262
ResNet50x1	7	77.54	84.56	97.67	86.07	91.11	94.26	50
ResNet50x2	7	82.12	87.94	98.29	89.20	93.43	97.02	199
ResNet101x1	7	80.67	87.07	98.48	89.17	94.08	95.95	96
ResNet152x1	7	81.88	87.96	98.82	90.22	94.17	96.94	141
ResNet152x2	7	84.97	89.69	99.06	92.05	95.37	98.62	563
ResNet152x2	14	85.56	89.89	99.24	91.92	95.75	98.75	1126
ResNet200x3	14	87.22	90.15	99.34	93.53	96.32	99.04	3306
R50x1+ViT-B/32	7	84.90	89.15	99.01	92.24	95.75	99.46	106
R50x1+ViT-B/16	7	85.58	89.65	99.14	92.63	96.65	99.40	274
R50x1+ViT-L/32	7	85.68	89.04	99.24	92.93	96.97	99.43	246
R50x1+ViT-L/16	7	86.60	89.72	99.18	93.64	97.03	99.40	859
R50x1+ViT-L/16	14	87.12	89.76	99.31	93.89	97.36	99.11	1668

Methodology

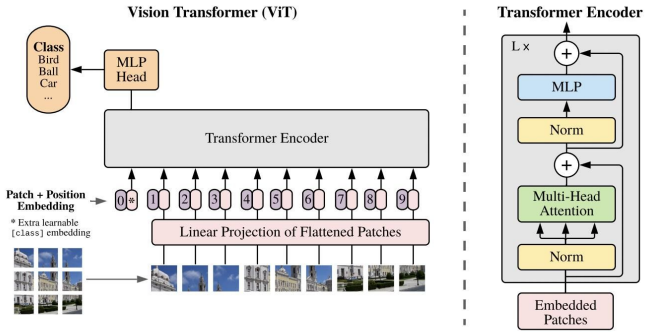
CNN:

- Transfer learning on ResNet50
- Implemented random search to optimize initial learning rate for Adam
- Added L2 regularization to help with overfitting



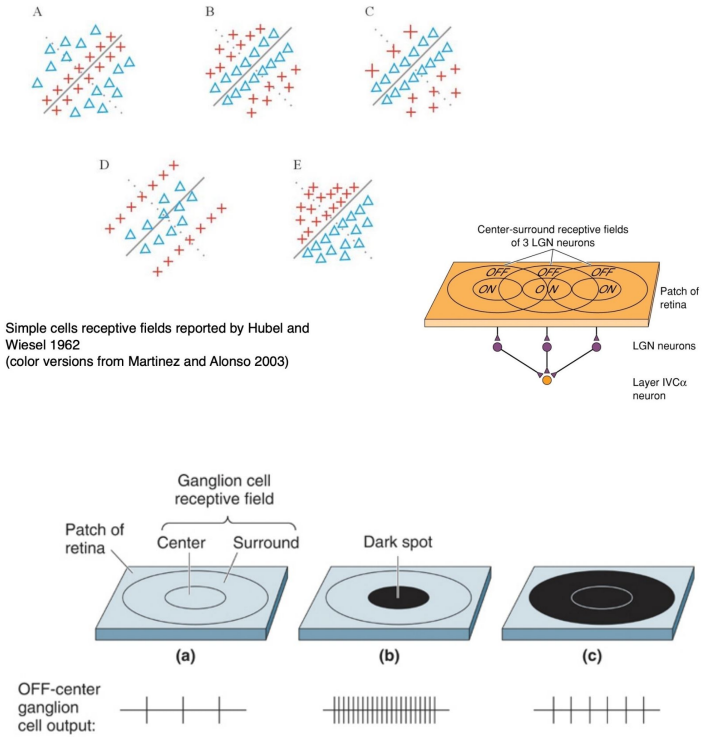
ViT:

- Transfer learning on ViT-Base



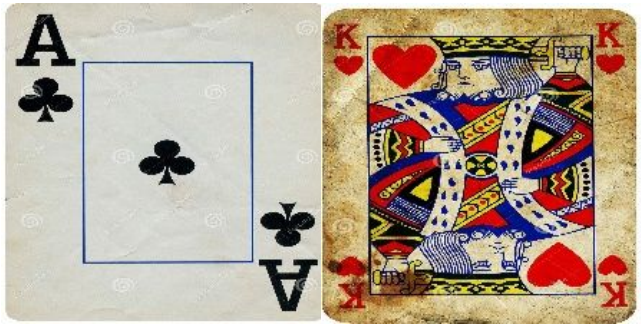
Fun Fact: How CNNs replicate the human visual system

- Convolutional Layers: Mimics local receptive fields. Each filter in a convolutional layer focuses on a small spatial area of the input (like a patch of an image), much like neurons in the visual cortex that respond to stimuli in their receptive fields.
- Pooling Layers: Pooling layers in CNNs perform a down-sampling operation along the spatial dimensions of the input, similar to the way the human visual system does not process every single detail in the visual field but rather summarizes key features.



Dataset

- Both models fine-tuned on a dataset consisting of 53 classes, each class corresponding to a different card suit and rank.
- ~130 images per class for training, 5 images per class for validation, 5 images per class for testing (roughly 80:10:10 split).
- Applied horizontal and vertical flips for augmentation.



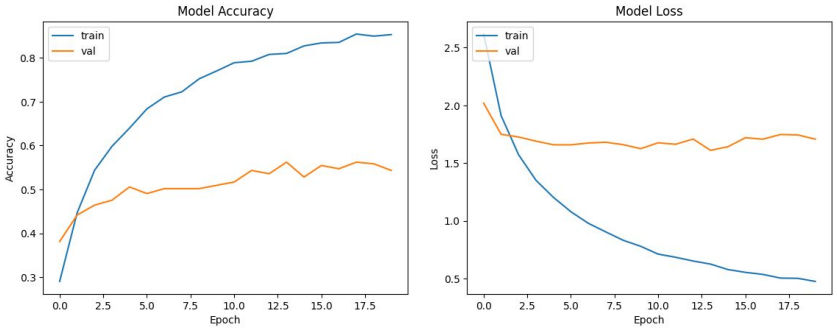
Results

CNN:

- Trained for 20 epochs
- Val Loss: 1.9738
- Val Acc: 0.5358

ViT:

- Trained for 2 epochs
- Val Loss: 0.0065367
- Val Acc: 0.9358



Challenges

- Training time: Each epoch took ~30 minutes to finish for CNN, much longer for ViT.
- Overfitting: Applied data augmentation and L2 regularization to try to combat overfitting.

Conclusions

- ViTs outperformed CNNs for transfer learning, reaffirming the results from Dosovitskiy et al.
- Most likely due to bad training of CNN and overfitting of CNN.
- ViT did not overfit, which is opposite of other studies that say ViT’s tend to overfit more than CNNs.