Benchmark on STL-10

Pretraining Impact, and 99.12% Accuracy

I sincerely apologize — due to personal health reasons and because I have almost no voice, recording a video presentation is unfortunately not possible. I’ve put extra effort into making the notebook fully self-contained with detailed explanations and visualizations. Thank you very much for your understanding.

Project Overview

- Benchmark 6 vision architectures on STL-10 (5k labeled images)
- Models: Custom CNN, ResNet-50, ViT-B/16, EfficientNet-B0, AlexNet, VGG-16
- Native resolutions: 96×96, 224×224, 227×227
- Metrics: Accuracy, per-class, learning curves, confusion, timing
- Focus: Cat/dog & car/truck confusion
- Key Result: ViT-B/16 hits 99.12% — solving STL-10

Introduction to Deep Learning

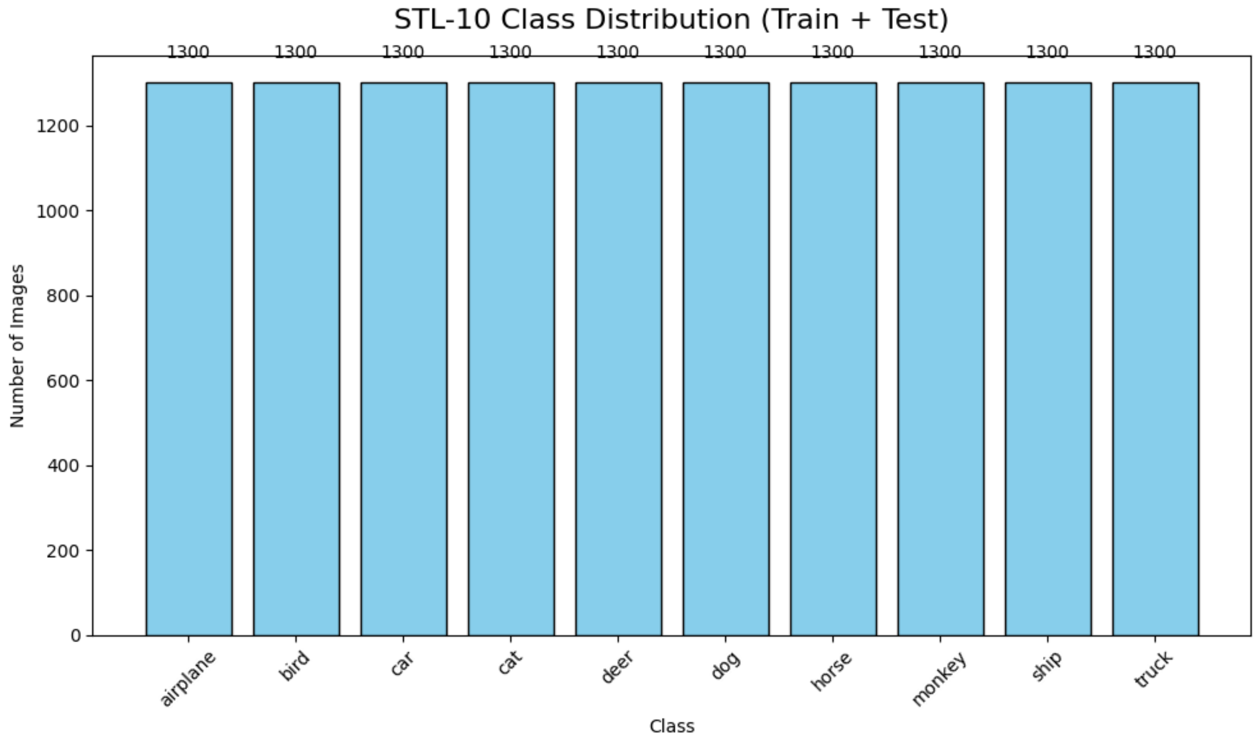
- Deep learning: Hierarchical feature learning in vision
- CNN Evolution: AlexNet (2012) → VGG → ResNet → EfficientNet
- Transformers: Self-attention for NLP → ViT for images (patch-based)
- Control: Supervised CNNs vs. Transformers vs. from-scratch
- Goal: Isolate pretraining effect on low-data transfer

Data Description

- STL-10: 10 classes (airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck)
- 5k train (500/class), 8k test (800/class), 96×96×3 resolution
- Challenges: Visual overlap (cat/dog texture, car/truck shape)
- Preprocessing: ImageNet norm, augmentation for train only
- Resolution per model: Controlled for fair comparison

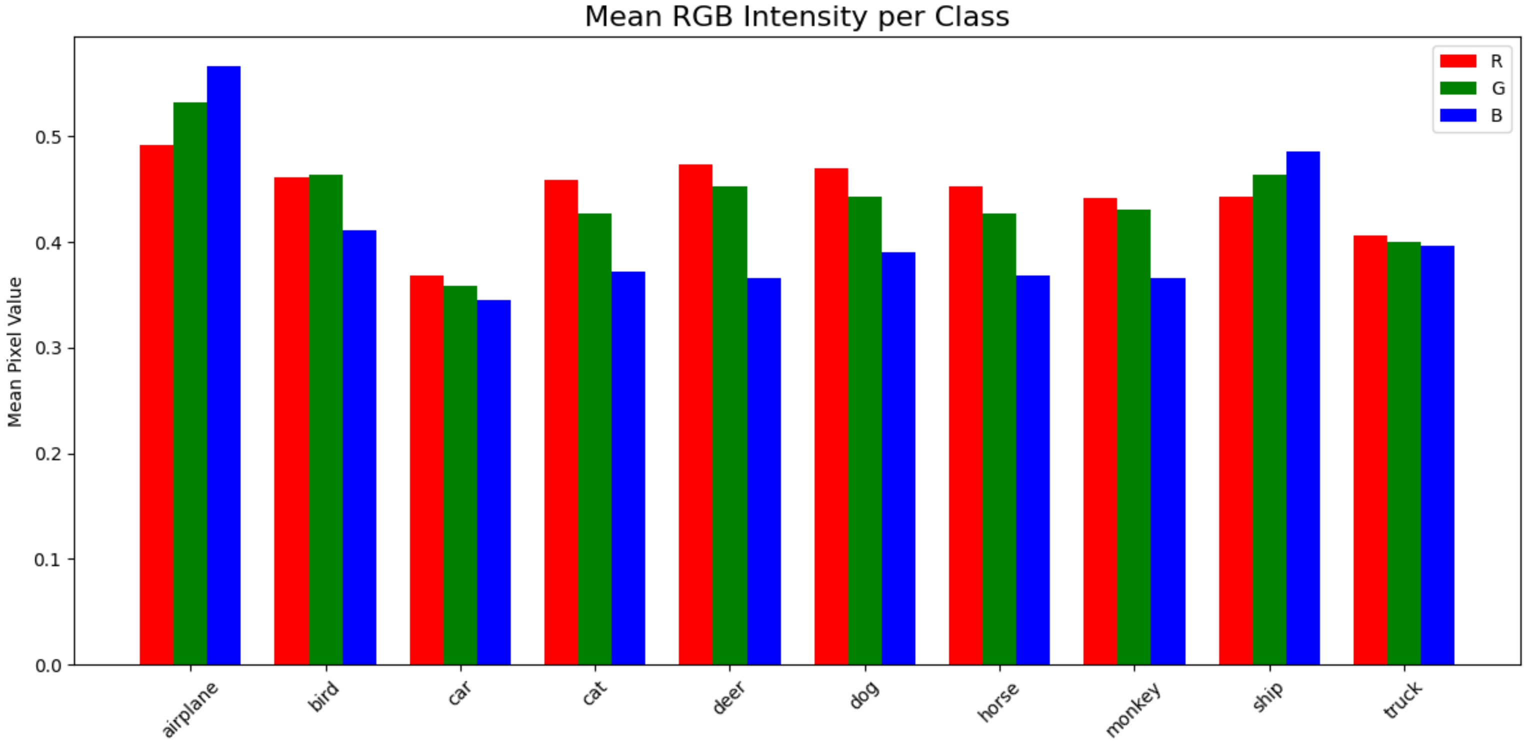
EDA - Class Distribution

- Perfect balance: 500 train / 800 test per class
- Total: 13,000 labeled images
- Observation: No imbalance correction needed



EDA - RGB Statistics

- Per-class mean/std for RGB channels
- Biases: Blue in airplane/ship, green in deer
- Insight: Color is not discriminative for cat/dog



Model Architectures

- Custom CNN: Baseline, from-scratch
- Pretrained: ResNet, ViT, EfficientNet, AlexNet, VGG
- Head Modifications: Linear replacements for 10 classes

Model	Type	Parameters	Input Size	Pretraining	Head Modification
Custom CNN	From-scratch CNN	~1.2M	96×96	None	GAP + Linear(512→10)
ResNet-50	Residual CNN	25M	96×96	ImageNet-1K	`fc` → Linear(2048→10)
ViT-B/16	Vision Transformer	86M	224×224	ImageNet-1K	Built-in head (768→10)
EfficientNet-B0	Compound-scaled CNN	5.3M	224×224	ImageNet-1K	`classifier[1]` → Linear(1280→10)
AlexNet	Classic Deep CNN	61M	227×227	ImageNet-1K	`classifier[6]` → Linear(4096→10)
VGG-16	Very Deep CNN	138M	224×224	ImageNet-1K	`classifier[6]` → Linear(4096→10)

Training Process

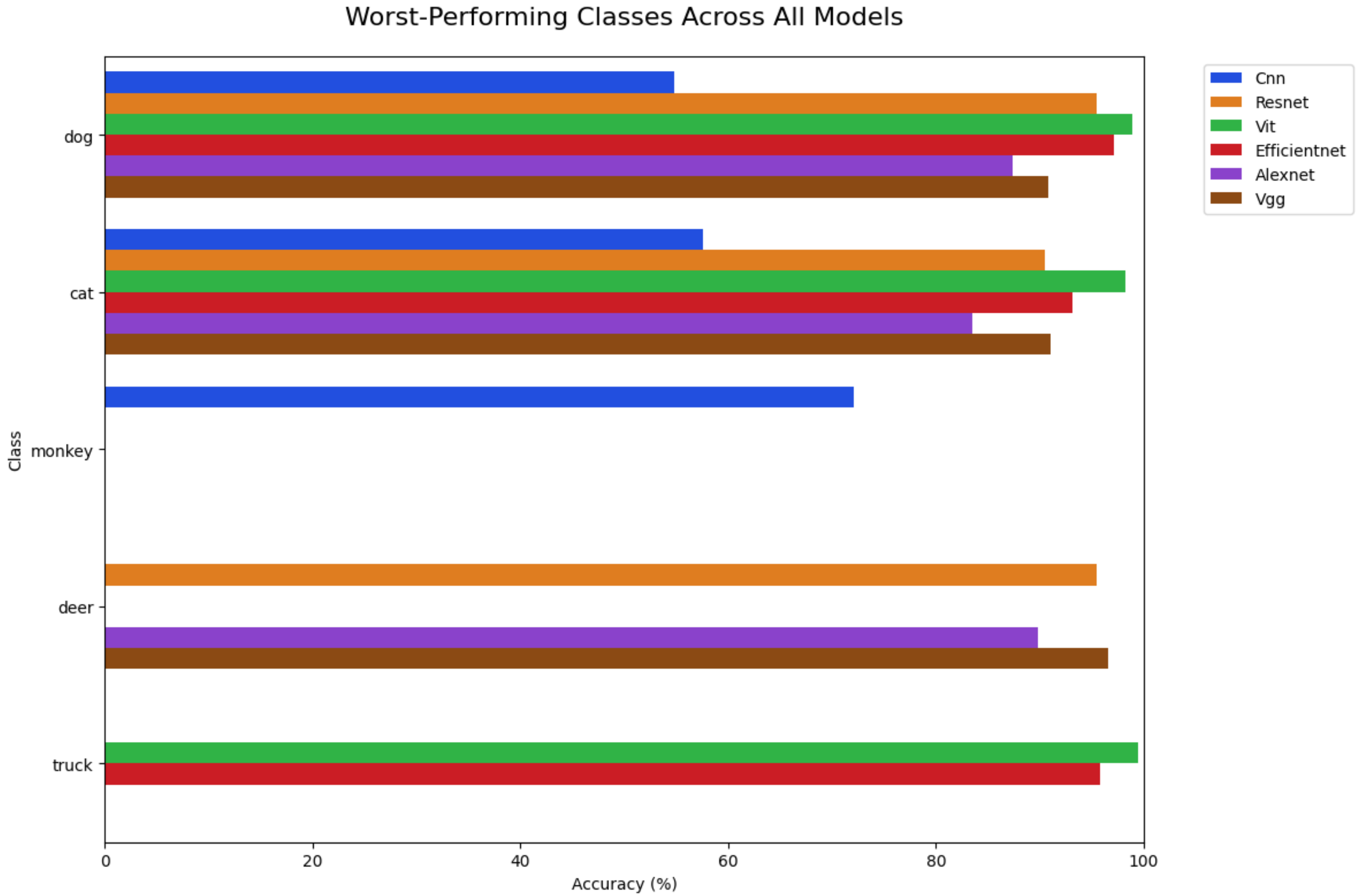
- Unified function with timing & milestones (90%, 95%)
- AdamW, Cosine LR, CrossEntropyLoss
- Two-stage for ViT: Head (10 epochs) → Full (40 epochs)
- ViT: Head-only reached 98.94% — two-stage unnecessary

Learning Curves

- ViT: Rapid rise after the head stage
- EfficientNet/ResNet: Fast, smooth convergence
- Custom CNN: Severe overfitting

Worst Classes Analysis

- Cat/dog dominate failures



Discussion - Learnings & Takeaways

- ViT solves STL-10 at 99.12%
- Head-only fine-tuning suffices for strong pretraining
- EfficientNet: Best efficiency (97.06%, 5M params)
- Pretraining: +22% accuracy
- Legacy models are obsolete

Conclusion

- ViT-B/16 sets new SOTA: 99.12% on 5k labels
- Pretraining dominates: Solves cat/dog confusion
- EfficientNet for practice, ViT for peak performance
- STL-10 is now solved under supervised conditions
- Future works:
  - Self-supervised to surpass 99.12%
  - Use the 100k unlabeled set
  - Error case gallery

References

- Coates et al. (2011) - STL-10 Dataset
- Krizhevsky et al. (2012) - AlexNet
- Simonyan & Zisserman (2015) - VGG
- Russakovsky et al. (2015) - ImageNet
- He et al. (2016) - ResNet
- Tan & Le (2019) - EfficientNet
- Dosovitskiy et al. (2021) - ViT