Deep Neural Network Compression & CV

Ayush Jha, Krish Jain, Sahil Pandey

Prof. Koteswar Rao

26 Nov 2025

Project Overview

Goal

Build a **3-stage compression pipeline** for a VGG/AlexNet-style CNN trained on **CIFAR-10**, targeting efficient deployment on resource-constrained devices.

Pipeline

1. **L1 pruning** to enforce sparsity and **remove 90% of the weights**.
2. **K-Means quantization** to cluster the remaining non-zero weights.
3. **Huffman coding** to losslessly encode the quantized weights.

Target Performance

- Achieve roughly a **9**× **compression ratio**, shrinking the model from about **228 MB** to ∼**25 MB**.
- Maintain an **accuracy drop** of less than **1.5%**.
- Reduce average bits per weight to around **3.57**.

Computer Vision Foundations

Objective

Introduce the **classical vision fundamentals** that form the basis for understanding convolutional neural networks and their feature extraction.

Topics Covered

- Image representations: **RGB vs HSV**, channels, normalization.
- Image processing: **Gaussian blur, sharpening, edge detection**.
- Classical CV operations: **Edge detection, Hough transforms**.
- Frequency domain intuition: **DFT/FFT and filtering**.
- Data augmentation & preprocessing for CNNs.

Outcome

Students will gain the intuition needed to understand how CNNs extract **edges, textures, and patterns** from images.

From CV to CNNs: Feature Extraction Pipeline

Core Goal

Enable students to build the **complete image preprocessing and feature extraction pipeline** used in AlexNet/VGG-style CNN models.

Hands-On Deliverables

- Implement dataset preprocessing:
  - Resize → 256, Crop → 224, RGB conversion.
  - Normalization using ImageNet mean/std.
  - Augmentations: random crop, flip, rotation.
- Visualize and interpret **CNN feature maps** (edges, textures, color blobs).
- Build the feature extraction stage for AlexNet/VGG.

Machine Learning

ML Foundations

Concepts & intuition behind the model we compress.

01: Basic Regression and Classification

- Linear regression and logistic regression.
- Decision trees for non-parametric classification.
- Emphasis on the **mathematics** and loss functions behind these models.

02: Non-linear Models & Generalization

- **Multi-layer perceptrons** (MLPs) for non-linear decision boundaries.
- Bias–variance trade-off, **underfitting** and **overfitting**.
- Practical techniques: regularization, early stopping, etc.

Machine Learning

ML Foundations

Core models used in our compression pipeline.

03: CNNs & Architectures

- Convolutional neural networks for image classification.
- Studying **VGG-16** and **AlexNet** architectures.
- Understanding convolution, pooling, and fully connected layers.

04: K-Means for Compression

- Unsupervised **K-Means clustering** on network weights.
- Replace individual weights with cluster centroids.
- Forms the basis for our **quantization** stage in the compression pipeline.

Compression Algorithm

3-Stage Compression

From dense CNN to compact, deployable model.

01 – Pruning

- Enforce sparsity using **L1 pruning**.
- Remove low-magnitude weights (up to **90%** of parameters).
- Directly reduces model size and computation.

02 – Quantization

- Convert high-precision floating-point weights to **low-bit** representations.
- Apply **K-Means clustering** to non-zero weights; store only cluster centers and indices.
- Significantly reduces memory footprint and enables faster low-precision arithmetic.

Additional Information

Course Project Workflow

- Approximately **15 lectures** will be conducted over a period of 2 months.
- Teams of **5–7 mentees** will be self-formed.
- **Weekly assignments** will be provided, can be individual/ group-based/ presentations.

Assignment Submission Rules

- All assignments must be uploaded to a **single group GitHub repository**.
- Each member must push their own assignment into their **respective folder** (for individual tasks).
- This ensures that by the end of the project, the group has a fully functional, well-organized repository.

Attendance

- Attendance will be taken in each lecture **at any time** using a QR code.

Thank You!

Questions?

Deep Neural Network Compression & CV