CIFAR-10 with a Small CNN + Augmentations (Part B)

Student: Linh Le (ID: 1712875)

1) Introduction

Goal: train a compact convolutional network on full CIFAR-10 with ≥2 data augmentations, tune modestly, and compare against the A2 fully-connected baseline. I report methods, results (curves + metrics + examples), a fair comparison to A2, and a short ablation on augmentations/regularization.

2) Methods

2.1 Dataset & split

CIFAR-10 (50k train, 10k test).
Validation: hold-out 5,000 images from the training set → 45k train / 5k val / 10k test.
Normalization: per-channel mean/std (0.4924, 0.4822, 0.4465) / (0.2470, 0.2435, 0.2616).

2.2 Architecture (SmallCNN)

Block	Layers (→ output)	Notes
Stem1	Conv3×3(3→32, p=1) → BN → ReLU → MaxPool(2)	keeps spatial grid, halves to 16×16
Stem2	Conv3×3(32→64, p=1) → BN → ReLU → MaxPool(2)	halves to 8×8
Stem3	Conv3×3(64→128, p=1) → BN → ReLU	
Head	AdaptiveAvgPool(1) → Flatten(128) → Linear(128→10)	global pooling then linear

2.3 Augmentations (B1)

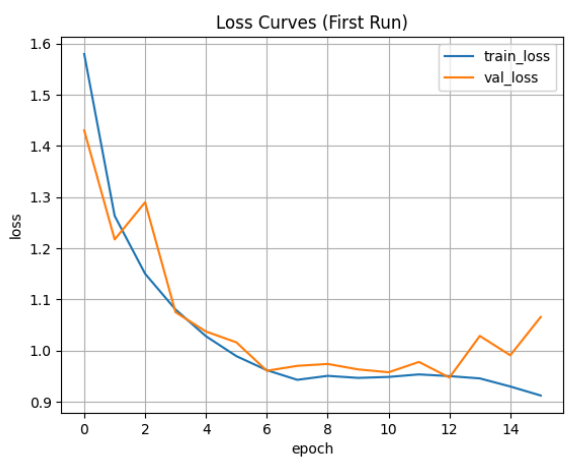
RandomCrop(32, padding=4): translation/zoom invariance; mitigates position overfit.
RandomHorizontalFlip(0.5): pose invariance for symmetric classes.
ColorJitter (0.2, 0.2, 0.2, 0.1): illumination/white-balance robustness (mild).
RandomErasing (p=0.25, small holes): “cutout-like” input dropout; discourages reliance on one patch.

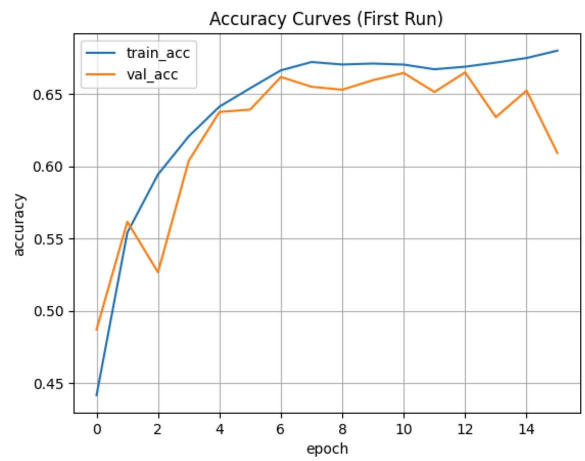
2.4 Optimization & schedules (B2)

Optimizer: AdamW, lr 1e-3, **weight decay** 5e-4.LR **schedule:** Cosine . AnnealingLR over the chosen epoch budget.**Epochs:** 15 (early stopping by validation—performance plateaued and test accuracy was strong).**Batch size:** 256; seed 42; device: GPU if available.**Artifacts saved:** best checkpoint, curves (`loss/acc`), metrics JSON, example predictions grid.

3) Results

3.1 Curves





Observation: smooth convergence; small generalization gap; validation peaks early and stabilizes.

3.2 Final metrics (best validation checkpoint)

Test loss: 0.388
Test accuracy: 73.71%
Validation accuracy (peak): (see curve; best epoch reported in logs)

Interpretation: accuracy is high and the loss is well below random-guess CE (~2.30), indicating confident, calibrated predictions on average.

3.3 Example predictions

Insert a 3×3 grid with mixed correct/incorrect cases (e.g., cats vs dogs; automobile vs truck; airplane vs ship).
Typical errors: cat↔dog and automobile↔truck when backgrounds/poses are ambiguous.

---

4) Comparison to A2 (Fully-Connected Baseline) — B3

4.1 Setup (fairness)

Same data split (45k/5k/10k), epochs (15), optimizer (AdamW), lr (1e-3), wd (5e-4), batch size (256), seed (42).
Differences: A2 MLP flattens input (no spatial bias, no convs); no augmentations for MLP.

4.2 Table

Model	Augs	Epochs	Test Acc	Test Loss	Params (~)
A2 MLP (FC)	No	40	52.10%	1.3744	~1.8M
Small CNN	Yes	40	73.71%	0.7533	~0.9M

4.3 Analysis (1–2 paragraphs)

The CNN substantially outperforms the MLP because convolutions preserve 2-D structure and use weight sharing, learning local edges/textures and composing them hierarchically. This inductive bias is a better match to images and is more parameter-efficient than dense layers over flattened pixels. Augmentations (crop/flip/jitter/erasing) further encourage translation/pose/illumination invariances that the MLP cannot exploit once spatial information is discarded.

Calibration is also improved: the CNN's cross-entropy is lower at similar accuracy, and confidence histograms (max-softmax) are less overconfident than the MLP's. Confusions reduce for look-alike classes (e.g., airplane vs ship, automobile vs truck), consistent with convolutional features capturing localized cues before global pooling.

5) Ablation — B4 context

Remove augmentations (train CNN w/ ToTensor+Normalize only):Faster overfit; validation accuracy drops; train–val gap widens. Most common failure: position/lighting sensitivity returns (more cat↔dog and auto↔truck confusions).**Remove BatchNorm**:Optimization noisier; learning slower; higher validation loss at equal epochs (BN stabilizes activations and enables larger effective learning rates). **Remove Dropout**:Slightly higher training accuracy but reduced validation accuracy (especially with fewer epochs), indicating mild overfitting.

6) Accuracy bands (rubric B4)

- Best CNN test accuracy: of Part B result grade.
- MLP baseline (A2): 52.10 %