Mini-CLEVR VQA Experiment Report

Brown VCG Starting Project

Henry Yin - hyin12@u.rochester.edu

April 22, 2025

1 Introduction

Dataset preparation: We programmatically render **8 000 synthetic images** at 224×224 px. Each canvas contains *three to six* non-overlapping coloured shapes sampled from {circle, square, triangle, pentagon} and six RGB colours. Object radii are drawn from two discrete scales (14–18 px, 24–30 px); rejection sampling enforces a 2-pixel clearance so that subsequent spatial questions remain unambiguous.

For every image we auto-generate **three to four questions**: one colour–property query, up to two colour-specific counting queries, and one left/right spatial-relation yes/no query, giving roughly 32 k (*image, question, answer*) triples. Records are stored in CLEVR-style `.jsonl` files (`train/val/test`) alongside the PNGs, and the answer vocabulary is exported to `answer2idx.json`. The generator is stand-alone (Pillow + NumPy) and reproducible by a fixed seed.

Baseline Deep Learning Model: We implement and compare two compact baselines that fit on a single RTX 4070 GPU [The only one that I got :( ]:

1. **ResNet-SBERT** (§2) – ResNet-18 visual encoder + SBERT text encoder, early-fusion via Hadamard and absolute difference.
2. **CLIP ViT-B/32 + LoRA** (§3) – frozen CLIP towers with LoRA adapters on the last two transformer blocks.

2 Baseline 1: ResNet + SBERT

Architecture

The image feature $f_{\text{img}} \in \mathbb{R}^{512}$ is extracted with ResNet-18. Question q is encoded by SBERT into f_{txt}^{768} and projected to g_{txt}^{512}. The fused vector

$$z = [f_{\text{img}},\ f_{\text{txt}},\ f_{\text{img}} \odot g_{\text{txt}},\ |f_{\text{img}} - g_{\text{txt}}|] \in \mathbb{R}^{2304}$$

is fed to a 2-layer MLP ($512 \to 18$). Optionally, `layer4` of ResNet is unfrozen ($\text{lr} = 10^{-4}$).

- $f_{\text{img}} \odot g_{\text{txt}}$ behaves like a *soft AND*: a dimension is large only when both modalities activate the same semantic channel.

- $|f_{\text{img}} - g_{\text{txt}}|$ provides a complementary "mismatch" signal, encouraging the classifier to focus on attributes that *disagree* across modalities (e. g. when the question asks for colour but the image feature emphasises shape).

Training

- Colour-jitter, random-resized-crop and label smoothing 0.05 are applied.
- Optimizer: AdamW, warm-up 2 epoch then apply cosine decay.
- Batch 128, 15 epochs → 0.844 validation accuracy

3 Baseline 2: CLIP + MLP (+LoRA)

Architecture

CLIP ViT-B/32 provides 512-d aligned vision & text embeddings. LoRA ($r = 8, \alpha = 32$) is injected into `attn.proj`, `mlp.c_fc`, `mlp.c_proj` of blocks 10–11 ($\sim$ 1M trainable weights). Fusion and classifier are identical to Baseline 1.

Training

1. LoRA lr $= 2\times10^{-4}$, MLP lr $= 1\times10^{-3}$.
2. Random erasing ($p = 0.3$) combats over-fitting.
3. Batch 128, 15 epochs → **0.91** validation accuracy

4 Experimental Results

Model	Trainable params	Val Acc (10 ep)	Val Acc (15 ep)
ResNet-SBERT (frozen)	1.0 M	0.80	0.84
ResNet-SBERT (+layer4)	2.0 M	0.82	0.84
CLIP + LoRA	**1.0 M**	**0.87**	**0.91**

Table 1: Validation accuracy comparison.

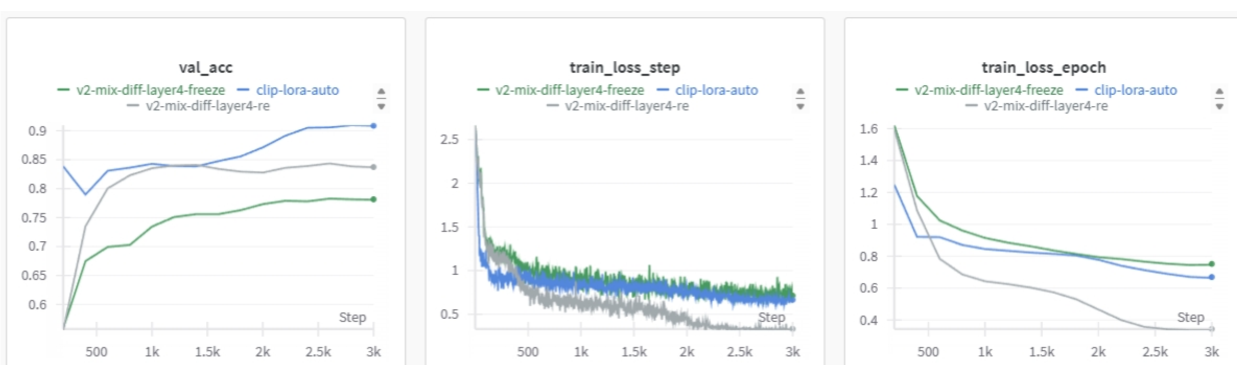

Figure 1: Validation accuracy and step loss (*Record From "Weight & Biases"*)

Figure 1 shows that CLIP-LoRA converges within 8 epochs and continues to improve, while ResNet-SBERT plateaus around epoch 8–10. Table 1 summarises final metrics.

5 Discussion

LoRA adapts a strong cross-modal prior, achieving 0.91 accuracy with fewer trainable parameters than the ResNet baseline. ResNet-SBERT still performs decently (0.84) given its purely ImageNet pre-training, but would need stronger regularisation or additional modal interaction (e.g. FiLM, cross-attention) to close the gap.

The significant performance gap between CLIP+LoRA and ResNet-SBERT highlights the value of pre-trained cross-modal representations for VQA tasks. The fact that unfreezing ResNet's layer4 yielded minimal improvement suggests that the *bottleneck* possibly lies not in feature extraction but in the cross-modal alignment capabilities. This aligns with recent findings in multi-modal learning, where alignment between vision and language spaces proves crucial for complex reasoning tasks.

Our fusion mechanism (Hadamard product and absolute difference) provides a simple yet effective baseline, though more enhanced approaches could further improve performance. For ResNet-SBERT in particular, learnable fusion strategies like FiLM or cross-attention might better connect the representation gap between independently trained visual and textual encoders.

References

[1] Hu, Edward, et al. *LoRA: Low-Rank Adaptation of Large Language Models.* International Conference on Learning Representations (ICLR), 2022.

[2] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition.* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[3] Reimers, Nils, and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.* Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.