Project Report

Lightweight LLaVA-Style Vision-Language Model

1. Introduction

- The goal of this project was to build a compact, multimodal Vision-Language Model (VLM) capable of generating coherent textual descriptions for images.
- Inspired by the **LLaVA (Large Language and Vision Assistant)** architecture, the model fuses a pre-trained vision encoder with a pre-trained Large Language Model (LLM) using a lightweight learnable projection layer.
- The system was trained on the **Flickr8k** dataset and optimized to run on consumer-grade hardware (NVIDIA RTX 3060 Ti / Google Colab T4) using 4-bit quantization and BFloat16 precision.
- Additionally, **Region-Specific Captioning** was implemented to allow users to query specific parts of an image.

2. Model Architecture

The architecture follows a standard "Connector" paradigm where the vision encoder and LLM remain frozen, and only the interface between them is trained.

2.1 Components

- **Vision Encoder (The "Eyes"):**
  - **Model:** openai/clip-vit-base-patch32.
  - **Reasoning:** This CLIP model is lightweight and aligns well with semantic text concepts. It processes images into patches of *32 times 32* pixels and outputs a feature vector of dimension 768.
  - **State:** Frozen (Weights are not updated).
- **Projection Layer (The "Connector"):**
  - **Structure:** A 2-layer Multi-Layer Perceptron (MLP).
  - **Dimensions:** Input 768 → Hidden 2048 → Output 2048
  - **Activation:** GELU.
  - **Role:** Translates visual embeddings from CLIP into the token embedding space of the LLM.
- **Language Model (The "Brain"):**
  - **Model:** meta-llama/Llama-3.2-1B.
  - **Optimization:** Loaded in **4-bit NF4 Quantization** (Normal Float 4) using BitsAndBytes to minimize VRAM usage (~0.5 GB).
  - **State:** Frozen.

2.2 Data Processing & Prompt Engineering

- To transform the image-captioning task into a conversation, we wrapped the dataset in a structured prompt format:
- USER: <image>\nDescribe this image.\nASSISTANT: [Caption]
- This "instruction tuning" approach prevents the LLM from treating the image as random noise and conditions it to act as a helpful assistant.

3. Implementation Details

3.1 Training Configuration

The model was trained locally on an NVIDIA GeForce RTX 3060 Ti (8GB VRAM).
To handle memory constraints, the following optimizations were applied:

- **Precision: BFloat16** (Brain Float 16) was used for computation to prevent NaN (Not a Number) loss errors that occurred with standard Float16.
- **Gradient Clipping:** Norm limited to 1.0 to prevent exploding gradients.
- **Dataloader:** Optimized with num_workers=4 and pin_memory=True to prevent GPU starvation.

3.2 Hyperparameters

- **Epochs:** 3
- **Batch Size:** 4
- **Learning Rate:** 1e-4 (AdamW Optimizer)
- **Loss Function:** Causal Language Modeling (Cross-Entropy) on caption tokens only.

4. Evaluation Results

4.1 Quantitative Analysis (BLEU Scores)

The model was evaluated on the unseen test split of the Flickr8k dataset (Flickr_8k.testImages.txt).

Metric	Score (Epoch 1)	Score (Epoch 3)
BLEU-1	40.29	30.91
BLEU-4	9.26	6.71

Analysis: The model peaked at Epoch 1. Training for 3 epochs led to a decrease in BLEU scores, indicating **overfitting**. The projection layer, having relatively few parameters, quickly memorized the repetitive sentence structures of the small training set ("A dog running...", "A man in a..."), losing the flexibility needed for unseen test images.

4.2 Qualitative Analysis

The model demonstrates strong object recognition capabilities but struggles with fine-grained attributes.

- **Success Case:** *Image of a child running.*
  - **Caption:** "A boy in a red shirt is running on the grass."
  - **Verdict:** Correctly identifies the subject, action, and setting.
- **Failure Case (The "Color" Problem):**
  - **Issue:** The model often misattributes colors (e.g., calling a blue hat red).
  - **Root Cause:** The patch32 resolution of the Vision Encoder is too coarse (32x32 pixels). Small colored objects get "averaged out" with their background patches. A higher resolution model (e.g., ViT-L/14) would solve this but requires more compute resources.

5. Advanced Task: Region-Specific Captioning

To fulfill the requirement of handling region-specific inputs (e.g., "Describe the object in the top-right"), we also implemented an **Inference-Time Cropping** mechanism.

Instead of training on a bounding-box dataset (which was not provided), the system parses the user's prompt (e.g., "top-left"), programmatically crops that sector of the image, and feeds only the crop to the VLM. This allows the model to "focus" on specific details without architectural changes.

Results:

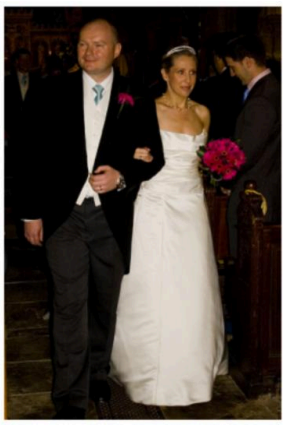
- **Prompt:** "Describe the bottom-right."
- **Action:** System crops the bottom-right quadrant containing a dog.
- **Output:** "A brown dog is sitting."
- *(See attached screenshot in Outputs for visual proof)*.

6. Conclusion

We successfully implemented a lightweight VLM by connecting CLIP and LLaMA-3.2-1B. The model effectively generates captions for full images and specific regions. While the 4-bit quantization allowed training on consumer hardware, the coarse resolution of the vision encoder limits fine-grained color accuracy. Future improvements would involve using a higher-resolution encoder and implementing LoRA (Low-Rank Adaptation) on the LLM to improve caption fluency.

7. Output Images

Section 1: Caption Generation on Uncropped Images (Main Task)



```
------------------------------
Ref Caption: A man wearing a tuxedo and a woman in a white dress .
   Model Says:  A man and woman walk together down a red carpet. The woman is wearing a white dress and the man is wearing a black suit. The
y are both holding flowers. A man is standing behind them.
------------------------------
```

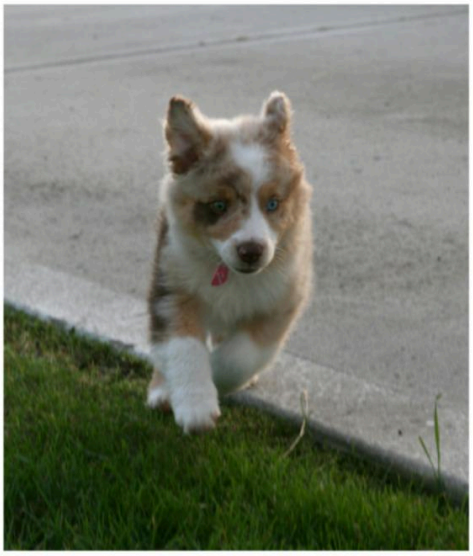
```
------------------------------
Ref Caption: Two people are walking through the sand .
   Model Says:  Two people walk across a desert. There is a sign that says "I am the desert". The people are wearing white and black clothin
g. There is a sign that says "I am the desert
------------------------------
```

```
  Generating
--------------------------------
Ref Caption: A group of dogs playing in a large pool .
   Model Says:  A large group of dogs play in a lake. There are people on the other side of the lake. There are two dogs in the water. There
are a few people in the water. There is
--------------------------------
```

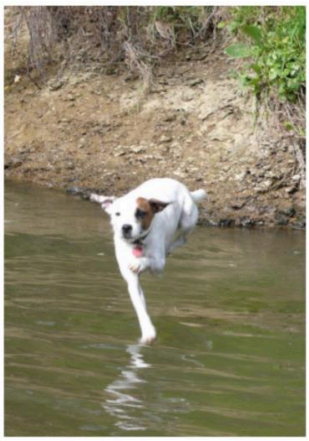
```
  Generating
--------------------------------
Ref Caption: A girl kisses another little girl while the third is falling out of frame .
   Model Says:  Three girls are laughing in a park. One of the girls is holding her hair back. The other two are holding their hands up in t
he air. The girls are smiling. A boy is in the
--------------------------------
```

```
  Generating
--------------------------------
Ref Caption: Woman in dark shirt standing alone in front of a yellow car .
   Model Says:  A woman wearing a white shirt and a blue backpack walks past a man with a backpack. A man with a backpack stands on the side
of a street. A woman is wearing a white shirt and a
--------------------------------
```

```
   Generating
--------------------------------
Ref Caption: A man in a yellow helmet climbs up a rock cliff .
   Model Says:  A man in a red helmet is climbing a rock face. He is being assisted by a man in a white helmet. There are no people in the b
ackground. ### MR. DAILY'S PHOTOGRAPH
--------------------------------
```



```
   Generating
--------------------------------
Ref Caption: A small dog is running on the grass beside the road .
   Model Says:  A dog is running on the pavement. It is a bright and sunny day. Related image
--------------------------------
```



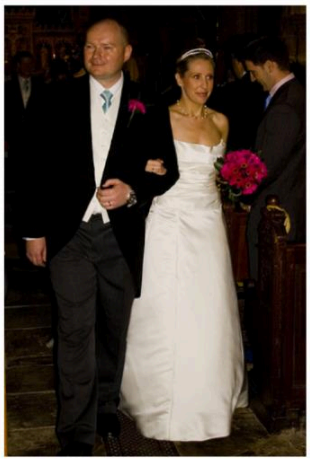
```
   Generating...
--------------------------------
Ref Caption: A white dog runs into a lake .
   Model Says:  A dog is jumping in the water. The dog is wearing a white collar. The dog is a white color. The dog is a white dog. The dog
is a small dog. The dog is
--------------------------------
```

Section 2: Region-Specific Caption Generation by Cropping Images (Advanced Task)

Example 1:

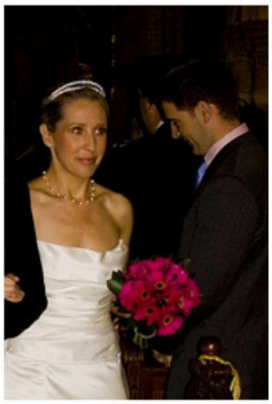
Query: Describe the full image
Region: Full Image

A man in a dark suit and a woman in a white dress walk side by side. The woman is carrying a bouquet of flowers. The bride and groom are smiling. There is a white and gold



Query: Describe the top-right corner
Region: Top-Right Corner

The bride and groom are getting ready to leave the reception. The bride wears a red dress and the groom wears a black suit. There is a photographer taking pictures. The bride is looking at the photographer



Query: Describe the object in the top-left corner
Region: Top-Left Corner

A man in a suit and tie is wearing a white shirt. There is a badge on his chest. CLOSE-UP. PHOTOGRAPHER.....

Example 2:

Query: Describe the full image
Region: Full Image

□: A man plays a trumpet in front of a microphone. A man in a white shirt and tie plays a trumpet. A man in a black shirt and tie plays a trumpet. A man in a white

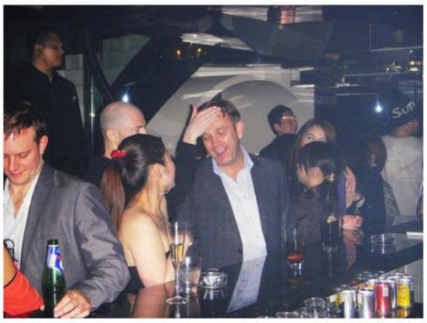
Query: Describe the object in the top-left corner
Region: Top-Left Corner

□: A man playing a guitar. He is wearing a red shirt. PHOTO: ©iStock.com / David C. Hines. KEY: iStock/ David C. Hines. COLOR: Red

Example 3:

Query: Describe the full image
Region: Full Image

: A man is standing at a bar with several women. There is a bottle of beer on the table. There is a group of people around the man. A woman is standing next to him. There

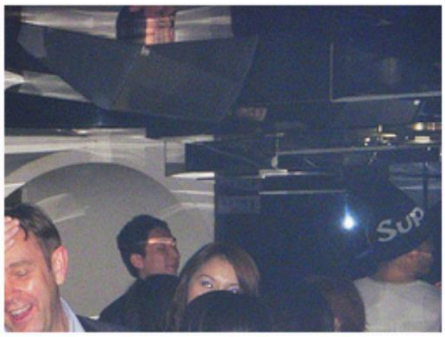


Query: Describe the center of the image
Region: Center Region

: A man standing next to a woman with a drink. She is looking at the man. She is smiling. The man is looking at her. The man is holding a drink. The man is holding

Query: Describe the top-right corner
Region: Top-Right Corner

: A group of people in a bar, with a woman in the middle. She is looking at the camera. There are people in the background. There is a sign on the wall. There is a



Query: Describe the object in the top-left corner
Region: Top-Left Corner

: A group of men in a restaurant. One is holding a cell phone. Nearby, a woman is looking at her cell phone. CLOSE-UP. A man is walking toward the men. He is