Comprehensive Report: LoRA Fine-Tuned BLIP2 for Visual Question Answering

Executive Summary

This report documents the development and evaluation of a Visual Question Answering (VQA) system that leverages the BLIP-2 model with Low-Rank Adaptation (LoRA) fine-tuning. The project demonstrates significant improvements in performance, achieving a BERT score increase from 0.66 (baseline) to 0.70 (fine-tuned model).

1. Data Curation

Dataset Source and Structure

The Visual Question Answering dataset was created using the Amazon Berkeley Objects (ABO) dataset as the image source, combined with the Gemini 1.5 Pro model for question-answer generation. The comprehensive data curation process involved:

1. **ABO Dataset Processing**:
   - Structured access to the Amazon Berkeley Objects dataset containing product images
   - Loading and parsing image metadata from CSV files
   - Processing product listings from JSON files with detailed product information
   - Creating a filtered subset of valid products with available images
2. **Question Generation with Gemini 1.5 Pro**:
   - Each product image was processed by the Gemini 1.5 Pro multimodal model
   - Product metadata (name, color, material, description) provided as context
   - Strict guidelines enforced for generating answerable, diverse questions:
     - Variety across 11 question types: color, shape, material, size, pattern, count, texture, position, function, brand, and comparison
     - Distribution across three difficulty levels: easy, medium, and hard
     - Focus on visual attributes observable in the images
3. **Answer Standardization**:
   - All answers constrained to exactly one word for consistent evaluation
   - Numerical answers represented as digits rather than words
   - Post-processing to extract first word if Gemini provided multi-word answers
   - Consistent formatting of question-answer pairs
4. **Dataset Composition and Quality Control**:
   - Final dataset contains thousands of question-answer pairs
   - Each image associated with approximately 10 diverse questions
   - CSV format with columns: image_name, question, answer

- Automatic filtering to ensure answer quality and question diversity
- Optional JSON backup format with full product metadata preserved

Data Processing Pipeline

The preprocessing workflow for model training includes: - Conversion to RGB format and resizing to meet BLIP-2 input requirements - Text prompting using a consistent format: "Question: {question} Answer in one word: {answer}" - Batched processing to optimize memory usage during training

This carefully curated dataset provides a robust foundation for training and evaluating the VQA model, with questions that challenge the model's visual understanding across various attributes and difficulty levels.

2. Model Choices

Primary Model Selection

We selected the **BLIP-2 (Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation)** model as our foundation for several reasons:

1. **Architecture Advantages**:
   - BLIP-2 uses a Querying Transformer (Q-Former) that efficiently connects a frozen image encoder with a frozen LLM
   - The architecture allows for efficient fine-tuning while maintaining strong zero-shot capabilities
   - Superior performance on vision-language tasks compared to previous models
2. **Specific Variant Selection**:
   - We chose `Salesforce/blip2-opt-2.7b` as our primary model due to:
     - Optimal balance between performance and computational requirements (~5GB VRAM)
     - Strong baseline capabilities for visual reasoning tasks
     - Compatibility with LoRA fine-tuning approaches
3. **Alternative Models Considered**:
   - `Salesforce/blip2-opt-6.7b`: Higher performance but requires ~12GB VRAM
   - `Salesforce/blip2-flan-t5-xl`: Alternative architecture with similar VRAM requirements
   - `Salesforce/blip2-flan-t5-xxl`: Highest performance but requires ~20GB VRAM

The chosen model balances accuracy and computational efficiency, making it accessible for researchers with limited GPU resources while maintaining strong performance capabilities.

3. Fine-Tuning Approaches

LoRA Fine-Tuning Strategy

We implemented Low-Rank Adaptation (LoRA) for fine-tuning BLIP-2, which offers several advantages:

1. **LoRA Implementation Details**:
   - Target modules include:
     - Attention layers: query, key, value, and output projections
     - MLP layers: up and down projections
   - Hyperparameters:
     - Rank (r): 16 (balance between adaptation capacity and parameter efficiency)
     - Alpha: 32 (scaling factor for LoRA activations)
     - Dropout: 0.05 (prevents overfitting during fine-tuning)
2. **Two-Stage Processing Approach**:
   - Vision encoder features extracted once and cached to avoid redundant computation
   - Q-Former processes these features to generate query embeddings
   - Language model adapters fine-tuned using LoRA while keeping vision components frozen
3. **Training Configuration**:
   - Learning rate: 1e-4 with a linear warmup schedule
   - Mixed-precision training (fp16) for efficiency
   - Batch size of 4 for optimal GPU memory usage
   - Training-validation split of 80-20

Parameter Efficiency

Our LoRA implementation significantly reduces trainable parameters compared to full fine-tuning: - Total BLIP-2 parameters: ~2.7 billion - Trainable parameters with LoRA: ~3.8 million (~0.14% of total parameters) - This efficiency allowed fine-tuning on consumer-grade hardware while achieving meaningful improvements

4. Evaluation Metrics

BERT Score Analysis

We chose BERT Score as our primary evaluation metric for the following reasons:

1. **BERT Score Rationale**:
   - Semantic understanding: Captures meaning beyond exact word matches
   - Contextual embeddings: Evaluates answers based on semantic similarity
   - Better correlation with human judgment than traditional metrics like exact match

2. **Performance Improvement**:
    - Baseline BLIP-2 (zero-shot): 0.66 BERT Score
    - LoRA fine-tuned BLIP-2: 0.70 BERT Score
    - This 0.04 improvement represents a significant enhancement in answer quality
3. **Score Distribution Analysis**:
    - The fine-tuned model showed more consistent performance across different question types
    - Particularly improved on attribute-based questions (color, size, material)
    - Maintained strong performance on object identification questions

Additional Evaluation Considerations

While BERT Score provided our primary metric, we also considered: - Generation speed: The LoRA fine-tuned model maintained inference efficiency - Parameter efficiency: Achieved improvement with minimal parameter growth - Qualitative assessment: Manual review of answers showed improved precision and relevance

5. Additional Contributions and Novelty

Technical Innovations

1. **Efficient Preprocessing Pipeline**:
    - The custom `VQADatasetPreprocessor` class optimizes BLIP-2 preprocessing
    - Batched processing of vision features reduces memory requirements
    - Fixed tensor dimensionality handling ensures consistent training data
2. **Targeted LoRA Configuration**:
    - Strategic application of LoRA to specific model components
    - Customized adaptation for vision-language tasks beyond standard NLP applications
    - Balance of rank and scaling parameters optimized for VQA tasks
3. **Inference Optimization**:
    - Half-precision inference for faster processing
    - Streamlined prompt structure for one-word answer generation
    - Confidence thresholding for answer selection