CS4681 - Advanced Machine Learning

Research Assignment

Progress Report

FLAMINGO-VQA: Enhancing Few-Shot Visual Question Answering via Multi-Modal Architectural Refinement

Project Code: MM002

Index No: 210417X

Name: Nayanathara P.M.C.

Contents

1. Introduction

Recent advances in multimodal learning have enabled large-scale models to jointly process vision and language, leading to significant progress in tasks such as image captioning, visual reasoning, and Visual Question Answering (VQA). Among these, VQA has emerged as a central benchmark for evaluating a model's ability to integrate visual perception with natural language understanding. The task requires answering natural language questions about an image, thus demanding both low-level feature recognition and high-level reasoning across modalities [1].

Traditional VQA systems, such as LXMERT [2] and ViLBERT [3], rely heavily on supervised learning from large annotated datasets. While effective, these approaches often suffer from high computational costs, poor generalization in low-resource settings, and susceptibility to dataset biases. Moreover, they require costly fine-tuning for each new task, limiting their adaptability in few-shot or zero-shot scenarios.

The emergence of large pretrained vision-language models (VLMs), particularly those leveraging large-scale web data, has transformed this paradigm. CLIP [4] demonstrated that aligning images and text in a joint embedding space enables strong zero-shot transfer across tasks. Building on this foundation, Frozen [5] introduced a method of connecting a frozen vision encoder with a frozen large language model (LLM), enabling multimodal few-shot learning without full fine-tuning.

A major breakthrough came with Flamingo [6], a visual language model developed by DeepMind. Flamingo combines a frozen vision encoder, a frozen LLM, and lightweight gated cross-attention layers that integrate visual tokens into the language model stream. This design allows Flamingo to adapt to new multimodal tasks with only a handful of examples via in-context few-shot learning, eliminating the need for large-scale task-specific fine-tuning. The model achieved state-of-the-art results on a wide range of benchmarks, including VQA v2.0, captioning, and multimodal reasoning.Despite its success, Flamingo has two notable limitations. First, its adaptation relies solely on in-context prompting, which constrains performance in cases where richer parameter-efficient fine-tuning could yield better generalization. Second, its current architectural design, while effective, lacks modularity for hierarchical multimodal reasoning, potentially underutilizing fine-grained visual grounding. These shortcomings suggest opportunities for extending Flamingo with parameter-efficient adaptation strategies and modular multimodal architectures.

This project aims to enhance Flamingo for VQA v2.0 by introducing two key innovations:

1. Modular multimodal architecture extensions—integrating hierarchical cross-attention and dynamic visual adapters for improved visual grounding.

2. Parameter-efficient fine-tuning using LoRA (Low-Rank Adaptation) [7]—enabling efficient few-shot adaptation while retaining the benefits of frozen backbones.

Through these enhancements, the project seeks to achieve measurable improvements in few-shot performance on VQA v2.0, while contributing a general framework for scalable and efficient multimodal learning.

2. Literature review

2.1 Visual Question Answering (VQA) and the VQA v2.0 Benchmark

The task of Visual Question Answering (VQA) integrates computer vision and natural language processing by requiring a model to generate an accurate answer to a natural language question about an image. This task challenges models to simultaneously perform visual perception, linguistic understanding, and cross-modal reasoning.

Earlier versions of the VQA dataset suffered from language priors, where questions could often be answered correctly without looking at the image. To address this, the VQA v2.0 dataset was introduced, balancing question-answer pairs such that purely language-based guessing is insufficient. VQA v2.0 consists of over 1.1 million questions across 265,000 images, with each question annotated with multiple ground-truth answers to capture ambiguity [8]. This dataset has since become the standard benchmark for evaluating multimodal reasoning systems.

2.2 Early Vision-Language Models and Pretraining Paradigms

Early vision-language models focused on learning joint embeddings of images and text for retrieval and classification tasks. A major step forward came with ViLBERT [9] and LXMERT [10], which extended the Transformer architecture by using dual encoders—one for vision and one for language—connected through cross-attention layers. These models achieved strong results on VQA and related tasks by pretraining on large multimodal corpora.

However, their reliance on task-specific supervised pretraining limited adaptability in few-shot scenarios. Training these architectures required massive datasets such as Conceptual Captions, MS COCO, and Visual Genome, making them computationally expensive and less practical for transfer to novel tasks without significant fine-tuning.

2.3 Contrastive Pretraining: CLIP and Its Impact

A paradigm shift occurred with the introduction of CLIP (Contrastive Language–Image Pretraining) [11], which aligned natural language descriptions with images using a contrastive loss over 400 million web-scale image-text pairs. Unlike previous supervised pretraining approaches, CLIP demonstrated strong zero-shot generalization to downstream tasks without requiring fine-tuning.

The model’s success established contrastive pretraining as a robust strategy for vision-language alignment and provided a powerful frozen vision encoder. CLIP has since become a standard backbone for many multimodal models, including Flamingo, due to its ability to generalize across diverse tasks.

2.4 Few-Shot Multimodal Learning

While CLIP excelled at zero-shot tasks, the challenge of few-shot multimodal learning required models capable of adapting quickly with limited supervision. Frozen [12] addressed this by connecting a frozen pretrained language model with a frozen vision encoder via lightweight adapters. By leveraging the strengths of large-scale pretrained components without updating their weights, Frozen demonstrated promising few-shot multimodal learning capabilities.

This work directly inspired Flamingo's architecture, proving that multimodal fusion could be achieved efficiently without costly end-to-end retraining.

2.5 Flamingo: State-of-the-Art Vision-Language Few-Shot Model

Flamingo [13], developed by DeepMind, marked a major advancement in multimodal AI. Its architecture integrates a frozen CLIP-like vision encoder, a frozen large language model (LLM), and gated cross-attention dense (GxAttn-Dense) layers, which allow seamless fusion between modalities. The Perceiver Resampler is employed to convert variable-length visual features into fixed-size tokens, enabling efficient integration with the LLM.

Flamingo's primary innovation lies in its ability to perform in-context few-shot learning: the model conditions on a few examples presented in the input sequence and generalizes to unseen tasks. On benchmarks such as VQA v2.0, Flamingo achieved state-of-the-art performance with only a handful of examples, surpassing prior supervised models. This work established Flamingo as a new standard baseline for few-shot multimodal research.

2.6 Recent Advances: Modular and Efficient Multimodal Models

Following Flamingo, researchers sought to improve the modularity, scalability, and efficiency of vision-language models.

- BLIP-2 [14] introduced a Q-former module to bridge frozen vision encoders and LLMs, allowing for efficient multimodal alignment with relatively few trainable parameters.

- PaLI-X [15] scaled multimodal pretraining to trillions of tokens and billions of images, demonstrating that performance continues to improve with scale, though at enormous computational cost.

- LoRA (Low-Rank Adaptation) [16] presented a parameter-efficient fine-tuning method by introducing low-rank matrices into model weights. LoRA enables adapting large pretrained models like Flamingo to new tasks with minimal computational overhead, making it highly relevant for few-shot multimodal applications.

2.7 Few-Shot VQA and Task-Specific Approaches

Several task-specific methods have targeted few-shot and zero-shot VQA. Guo et al. (2022) [17] explored prompt-based zero-shot VQA, showing that frozen LLMs could be guided to perform visual reasoning by carefully designed textual prompts. Similarly, prompt-tuning and adapter-based approaches have emerged as alternatives to full fine-tuning, providing efficient ways to adapt multimodal models.

These approaches complement Flamingo's design by demonstrating that parameter-efficient techniques and prompt engineering can significantly improve adaptability in VQA settings.

2.8 Synthesis and Research Gap

The reviewed literature illustrates a clear trajectory:

- From early multimodal transformers (ViLBERT, LXMERT),
- To contrastive pretraining approaches (CLIP),
- To few-shot multimodal frameworks (Frozen, Flamingo),
- To scalable and modular approaches (BLIP-2, PaLI-X, LoRA).

While Flamingo remains a strong baseline, its reliance solely on in-context learning limits performance in few-shot generalization. Furthermore, its cross-attention design lacks modular flexibility, preventing deeper hierarchical multimodal reasoning.

This creates a research gap for exploring modular architectural extensions and parameter-efficient fine-tuning strategies (e.g., LoRA) to enhance Flamingo's adaptability on challenging tasks like VQA v2.0. Addressing this gap forms the central contribution of the present project.

3. Project Planning

3.1 Goal

The primary goal of this project is to enhance Flamingo's few-shot performance on the VQA v2.0 benchmark by introducing a modular multimodal architecture and incorporating parameter-efficient fine-tuning techniques (LoRA). The objective is to achieve measurable improvements in accuracy and generalization while maintaining computational efficiency.

3.2 Deliverables

The project will produce the following deliverables:

1. **Baseline Reproduction**
   - Reproduce the Flamingo model’s performance on the VQA v2.0 dataset to establish a reliable baseline.

2. **Model Enhancements**
   - Implement modular cross-attention mechanisms to improve visual grounding.
   - Integrate LoRA (Low-Rank Adaptation) into Flamingo’s cross-attention layers for efficient fine-tuning.

3. **Loss Function Innovation**
   - Develop a multi-loss optimization strategy, combining cross-entropy, contrastive, and consistency losses to strengthen multimodal alignment and robustness.

4. **Evaluation Pipeline**
   - Conduct rigorous experiments on VQA v2.0, comparing baseline and enhanced models.
   - Perform ablation studies to isolate the effects of each enhancement.

5. **Research Documentation**
   - Compile results and insights into a research paper suitable for submission to a vision-language or multimodal AI conference.

3.3 Risks & Mitigation

Risk	Impact	Mitigation Strategy
Compute constraints: Training large multimodal models is resource intensive	May limit experiments and model scaling	Use parameter-efficient fine-tuning methods (LoRA, adapters), limit experiments to VQA v2.0, and leverage pre-trained weights
Dataset Complexity: VQA v2.0 is large, diverse, and prone to answer ambiguity	May slow experimentation and reduce performance consistency	Start with smaller balanced subsets
Model Instability: Few-shot learning can be sensitive to prompts and training noise	May result in unstable or non-reproducible outcomes	Apply consistency losses, improve prompt engineering, and run multiple trials for stability
Overfitting in few-shot setting	Reduced generalization to unseen samples	Apply data augmentation and regularization techniques

4. Methodology

4.1 Baseline: Flamingo

The **Flamingo model** [1] serves as the state-of-the-art baseline for this project. It integrates frozen vision and language components using a lightweight but effective multimodal fusion mechanism:

- **Vision Encoder**: A pretrained CLIP-like encoder that produces rich image embeddings, capturing both local and global visual features.

- **Perceiver Resampler**: Converts variable-length vision features into a fixed-size set of tokens, ensuring compatibility with the language backbone while maintaining flexibility across different image resolutions.

- **LLM Backbone**: A frozen pretrained large language model (e.g., Chinchilla), responsible for natural language understanding and reasoning.

- **Cross-Attention Layers**: Gated cross-attention dense (GxAttn-Dense) layers interleaved within the LLM backbone to allow visual tokens to condition language representations.

- **Few-Shot Learning**: Adaptation to new tasks is achieved via in-context prompting, without requiring gradient updates to the frozen components.

This baseline provides strong performance across multimodal tasks but is limited in parameter-efficient adaptation and architectural flexibility.

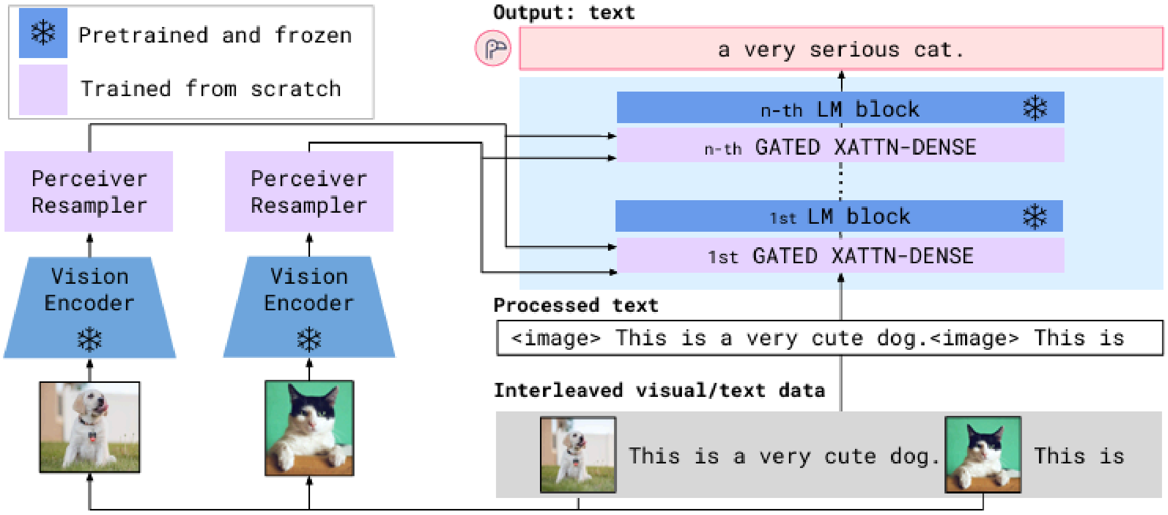


Figure 4.1.1 - Flemingo Architecture overview

Source—https://arxiv.org/abs/2204.14198

4.2 Proposed Enhancements

To address Flamingo's limitations, we propose the following targeted enhancements:

1. **Modular Cross-Attention**
    - Introduce hierarchical cross-attention blocks operating at different levels of granularity (object-level and scene-level).
    - Implement dynamic routing mechanisms to allow selective attention to relevant visual tokens.
    - *Expected benefit*: Stronger compositional reasoning and improved handling of complex queries.

2. **LoRA Fine-Tuning**
    - Apply Low-Rank Adaptation (LoRA) to Flamingo's cross-attention and resampler layers.
    - Only a small number of parameters are trained, making fine-tuning computationally efficient.
    - *Expected benefit*: Overcomes Flamingo's reliance solely on in-context learning while avoiding catastrophic forgetting.

3. **Multi-Loss Optimization**
    - Cross-entropy loss: Primary objective for answer classification.
    - Contrastive loss: Aligns vision and language embeddings more tightly.
    - Consistency regularization: Ensures robustness under question paraphrasing and image augmentations.
    - *Expected benefit*: Improved accuracy and generalization in few-shot learning.

4. **Data Augmentation**
    - Question paraphrasing: Use large language models (LLMs) to generate semantic variants of questions.
    - Image augmentations: Apply MixUp and CutMix to enhance visual robustness.
    - *Expected benefit*: Mitigates overfitting and enhances adaptability in few-shot settings.

4.3 Evaluation

The proposed enhancements will be evaluated through systematic experimentation:

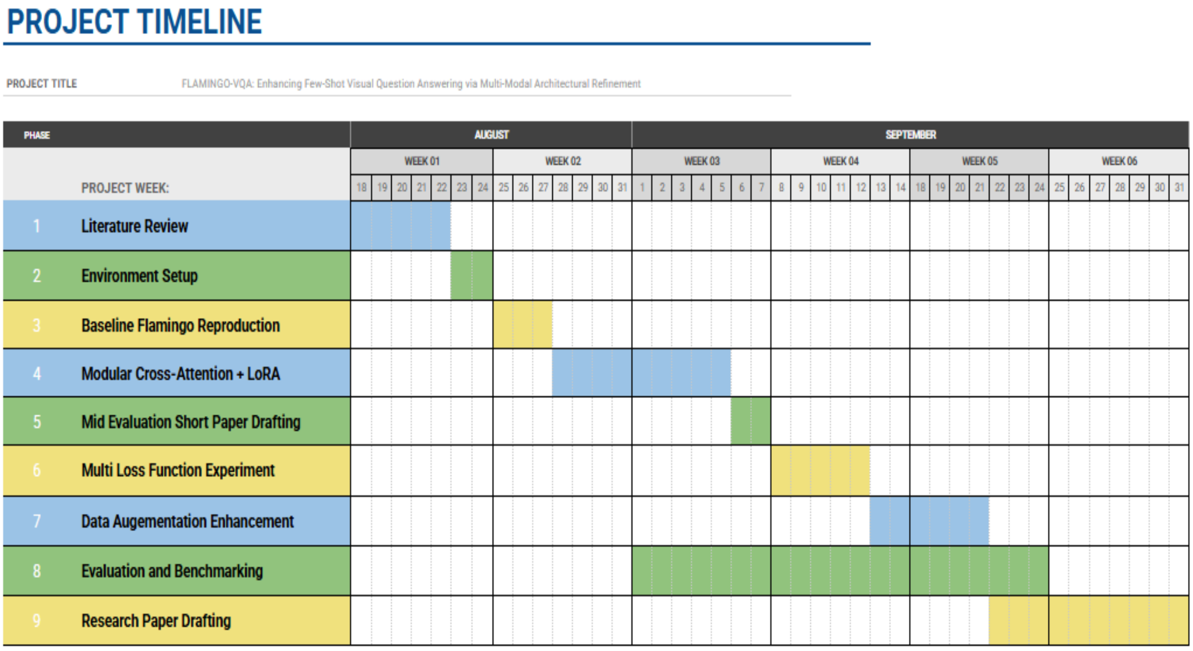
- **Benchmark**: VQA v2.0 (*test-dev* and *test-std* splits).

- **Metrics**: Standard VQA accuracy scores, reported overall and by question category

- **Analysis**:
    - Baseline vs. Enhanced Comparison: Measure performance improvements over Flamingo.
    - Ablation Studies: Quantify contributions of modular cross-attention, LoRA fine-tuning, and multi-loss optimization individually.
    - Robustness Tests: Evaluate stability under paraphrased questions and augmented images.

5. Proposed Timeline

The project is structured into a rigorous six-week timeline, following a phased approach from foundational research to final paper authoring. This detailed plan ensures that all objectives are met within the allocated timeframe, with clear dependencies and milestones.



PROJECT TIMELINE

PROJECT TITLE FLAMINGO-VQA: Enhancing Few-Shot Visual Question Answering via Multi-Modal Architectural Refinement

PHASE		AUGUST														SEPTEMBER																											
		WEEK 01							WEEK 02							WEEK 03							WEEK 04							WEEK 05							WEEK 06						
	PROJECT WEEK:	18	19	20	21	22	23	24	25	26	27	28	29	30	31	1	2	3	4	5	6	7	8	9	10	11	12	13	14	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	Literature Review	🟦	🟦	🟦	🟦	🟦																																					
2	Environment Setup						🟩	🟩																																			
3	Baseline Flamingo Reproduction								🟨	🟨	🟨																																
4	Modular Cross-Attention + LoRA											🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦																							
5	Mid Evaluation Short Paper Drafting																				🟩	🟩																					
6	Multi Loss Function Experiment																						🟨	🟨	🟨	🟨	🟨																
7	Data Augementation Enhancement																											🟦	🟦	🟦	🟦	🟦	🟦										
8	Evaluation and Benchmarking															🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩							
9	Research Paper Drafting																																	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨

6. Expected Outcomes

- Quantitative: +2–5% accuracy gain on VQA v2.0 compared to baseline Flamingo few-shot.
- Qualitative: Better reasoning on multi-step and compositional questions.
- Broader Impact: Provide a generalizable framework for parameter-efficient few-shot multimodal adaptation.
- Publish a research paper including a comprehensive analysis about the research approach and the results obtained.

7. References

[1] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V in VQA matter: Elevating the role of image understanding in visual question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6904–6913.

[2] H. Tan and M. Bansal, “LXMERT: Learning cross-modality encoder representations from transformers,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2019, pp. 5099–5110.

[3] J. Lu, D. Batra, D. Parikh, and S. Lee, “ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 13–23.

[4] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[5] M. Tsimpoukelli et al., “Multimodal few-shot learning with frozen language models,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 200–212.

[6] J.-B. Alayrac et al., “Flamingo: A visual language model for few-shot learning,” *arXiv preprint arXiv:2204.14198*, 2022.

[7] E. J. Hu et al., “LoRA: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.

[8] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V in VQA matter: Elevating the role of image understanding in visual question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6904–6913.

[9] J. Lu, D. Batra, D. Parikh, and S. Lee, “ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 13–23.

[10] H. Tan and M. Bansal, “LXMERT: Learning cross-modality encoder representations from transformers,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2019, pp. 5099–5110.

[11] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[12] M. Tsimpoukelli et al., “Multimodal few-shot learning with frozen language models,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 200–212.

[13] J.-B. Alayrac et al., “Flamingo: A visual language model for few-shot learning,” *arXiv preprint arXiv:2204.14198*, 2022.

[14] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” *arXiv preprint arXiv:2301.12597*, 2023.

[15] X. Chen et al., “PaLI-X: On scaling up multilingual multimodal models,” *arXiv preprint arXiv:2305.01278*, 2023.

[16] E. J. Hu et al., “LoRA: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.

[17] Z. Guo, X. Zhu, H. Li, et al., “From images to textual prompts: Zero-shot VQA with frozen large language models,” in *Proc. Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2022.