Lightweight Vision-Language Modeling for Fine-Grained Bird Species Reasoning

Shufan He (she68), Yitong Liu (yliu336), Yunjia Zhao (yzhao291)

Github Link

Introduction

This project aims to develop a domain-specific, lightweight Vision-Language Model (VLM) for Visual Question Answering (VQA) focused on bird species identification. This project explores training a compact ViT-based image encoder on a bird-only dataset and connecting it with a language model through a simple and interpretable modality integration strategy.

Our motivation is to study whether lightweight VLMs, using minimal connection mechanisms such as a linear projection-based integration, can still achieve strong performance in fine-grained visual reasoning tasks. We focus on the bird domain to take advantage of its rich attribute annotations (e.g., color, shape, part-level details) and to explore VLM interpretability in specialized settings.

Related Work

Recent VLMs adopt different strategies for modality integration:

- LLaVA [1] connects image features from a pre-trained SigLIP (ViT) encoder to a language model via a linear projection.
- Flamingo [2] uses cross-attention layers to inject visual features into the language model for conditioned generation.
- Fuyu [3] merges image patches and text tokens into a unified input sequence for joint multimodal modeling.

These designs reflect a trade-off between simplicity (linear layers), controllability (cross-attention), and dense fusion (token merging). Our project focuses on the simplest strategy,linear projection for lightweight and interpretable modeling.

Data

We use two datasets in this project. The main dataset is the Caltech-UCSD Birds-200-2011 (CUB-200-2011) [4] dataset, which contains 11,788 images across 200 bird species along with species labels. This dataset will be used to train the vision encoder for bird classification. In addition, we use a vision-language dataset from Hugging Face [5] for multimodal training. Each instance in this dataset includes an RGB image of a bird from CUB, a text field containing 10 natural language descriptions and a numeric label indicating the bird species. This dataset is suitable for both captioning and VQA-style tasks and enables us to align visual features with rich

textual descriptions. We will preprocess the images (e.g., resizing, normalization) as needed for model input. As a fallback, we also consider the NABirds dataset [6], which contains 126.7k images across 555 categories.

Methodology

Step 1: Vision Encoder Training

- Architecture: ViT-Tiny
- Dataset: CUB-200-2011
- Task: Bird species classification
- Training strategy:
    - Train from scratch for domain-specific representation learning
    - Apply strong regularization (RandAugment, CutMix, Dropout, Stochastic Depth)

Step 2: Vision-Language Integration & Fine-tuning

- Architecture:
    - Freeze the trained ViT-Tiny encoder
    - Use a small pretrained language model (e.g., Qwen 0.5B)
    - Insert a special <|image_pad|> token into the text input
    - Project image features into the language model's embedding space using two linear layers
    - Replace <|image_pad|> with projected image features during the forward pass
- Tasks:
    - Captioning: Generate descriptions from images
    - VQA-style generation: Generate answers from image + question pairs
- Dataset:
    - HuggingFace Bird Dataset (multi-caption, fine-grained descriptions)
    - Preprocess data into VQA-style QA pairs using attribute extraction
- Training:
    - Freeze the vision encoder
    - Only the projection layers and the language model are fine-tuned to reduce compute overhead and isolate the effectiveness of the integration method

Metrics

We'll evaluate the project in two parts. First, we plan to evaluate the vision encoder performs on bird classification using accuracy on the CUB-200-2011 test set — aiming for at least 65%.

Then, for the multimodal part, we plan to use BLEU scores to measure how well the model generates bird descriptions or answers visual questions. For fine-grained attributes (like beak color or wing shape), we may also manually inspect or run rule-based checks.

- Base Goal: Reasonable caption generation and correct answering for simple bird attributes (e.g., color, shape, type).
- Target Goal: Accurate answering for part-based or compositional questions (e.g., "What color is the bird's beak?")
- Stretch Goal: Robust generation for long-tail attributes and complex reasoning (e.g., habitat-based descriptions).

Ethics

Why deep learning? Deep learning is well-suited for this task because bird species identification is a fine-grained visual problem that relies on subtle differences in appearance, such as color patterns, shapes, and textures. Vision Transformers (ViTs) are especially effective at capturing these patterns by learning global attention across image patches. Additionally, using a Vision-Language Model allows us to connect visual understanding with natural language generation, making it possible to answer detailed questions about bird attributes or generate human-like descriptions — something traditional methods would struggle to do.

As for data representation and bias, the images we use in this project are scraped predominantly from North-American birding websites. As a result, species outside of the North American area as well as under-photoed species could be under-represented in our models.

Division of Labor

All team members will collectively contribute to model build up. Shufan and Yunjia will be responsible for model performance testing and finalizing codes for github submission. Yitong will be responsible for poster, presentation slides preparation and final report write-up. If we feel the work distribution becomes imbalanced in any way throughout the process we will adjust it accordingly.

References

[1] Haotian Liu, Chunyuan Li, Qingyang Wu, et al. "Visual Instruction Tuning." *arXiv preprint arXiv:2304.08485*, 2023. https://arxiv.org/abs/2304.08485

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. "Flamingo: Visual language models with frozen vision and language models." *arXiv preprint arXiv:2204.14198*, 2022. https://arxiv.org/abs/2204.14198

[3] Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. "Fuyu-8B: Open-access multimodal models for understanding and generating text and images." *arXiv preprint arXiv:2310.08491*, 2023. https://arxiv.org/abs/2310.08491

[4] Wah, Catherine, et al. "The Caltech-UCSD Birds-200-2011 Dataset." *California Institute of Technology*, 2011. https://www.vision.caltech.edu/datasets/cub_200_2011/

[5] Alkzar90. "CC6204-Hackaton-Cub-Dataset." *HuggingFace Datasets*, 2023. https://huggingface.co/datasets/alkzar90/CC6204-Hackaton-Cub-Dataset

[6]126.7k images
https://images.cv/dataset/bird-image-classification-dataset