Effects of retrieval augmented knowledge on visual-instruction tuned language models

Sugi Venugeethan
sugi205@gmail.com

Abstract

We introduce Spoonbill, a retrieval-augmented autoregressive vision-language model fine-tuned on vision-language instruction datasets. Spoonbill is based on Open Flamingo and a custom multi-modal retriever based on pre-trained vision-language encoders.

Instruction following models has become increasingly powerful. We believe fine-tuning the vision-language model on visual instruction following datasets and fine-tuning language components separately on language-only instructions datasets can effectively help our dubbed vision-language model Spoonbill, to effectively follow instructions.

We evaluate the effects of retrieving relevant in-context examples. In task-specific few-shot visual question answering, it is not easy to easily find and use correlated context. By effectively building a multi-modal dense retriever and retrieving relevant documents for in-context learning, we can reduce re-training costs for task-specific knowledge and also improve accuracy. By externalizing the knowledge of the model, we believe accuracy and controllability aspects can improve.

1 Introduction

Recently multi-modal models especially in vision-language space have achieved remarkable progress in image and text generation. These models store all their information in the parameters of the underlying neural network which can lead to a lot of parameters to cover all knowledge.

We are motivated by recent work and the positive effects of instruction tuning and retrieval augmented knowledge on large language models. In addition to language-only instruction training, we like to evaluate visual instruction tuning of the vision language model. The language component in our dubbed Spoonbill is instruction tuned separately.

We focus on learning and experiencing the effects of instruction tuning and retrieval augmentation on top of the Flamingo-based model and create dubbed Spoonbill on limited datasets with a limited GPU budget. We also present dubbed Garuda which is an instruction tuned model based on Llama2 7B-Chat using Alpaca's dataset. Please refer to the model architecture and appendix for additional details on the Spoonbill and Garuda models.

2 Prior Literature

We draw inspiration from MuRAG (Chen et al., 2022) that highlights the importance of augmental retrieval in unearthing more information (especially) in the context of multi-modal data. Having an augmentation layer allows the pre-training of original model to be thin and computationally less efficient. This allows keeping different silos of information in different layer, without needing to retrain the core model with newer information. Recent investigations on prompting techniques and their usage in question answering systems (E.g. GPT) are introducing powerful ways to capture latent information in the LLMs to forefront. In Investigating Prompting Techniques (Awal et al., 2023), the authors investigate various prompting methodologies like CoT, Text and Imagequest and with different templates, a direction that is explored in this paper.

Recently, retrieval-augmented language models have shown promise in natural language processing (NLP) (Karpukhin et al., 2020; Guu et al., 2020; Lewis et al., 2020b; Borgeaud et al., 2022). Given input text, such a model uses a retriever that retrieves relevant documents from an external memory, and uses a generator to generate predictions given the retrieved documents. However, these retrieval-augmented methods are studied originally for text, and extending them to the multimodal setting remains an open problem with challenges. We

draw inspiration from the work done in RA-CM3 (Yasunaga et al., 2022) that exhibits a multimodal incontext learning ability: it can perform controlled image generation by prompting with demonstration examples in context, and it can also perform few-shot image classification. RA-CM3 is the first model that can perform in-context learning for both text and image generation. in our model we use Spoonbill to evaluate the efficacy of Retrieval Augmentation.

In this work, we seek to explore an alternate model to seek further validation of the effects of retriever augmentation. We limit ourselves to question answering on images and text, but hope that we can extend our work to explore video based question answering in future. First, to obtain a multimodal retriever, we use the Dense Retrieval method (Karpukhin et al., 2020) with a mixed-modal encoder that can encode combinations of text and images (e.g., pretrained CLIP; Radford et al. 2021). Given this retriever, we use Faiss (Facebook AI Similarity Search) to index and retrieve diverse and informative documents for the input document. Second, we design the retrieval-augmented generator based on the CM3 architecture (Aghajanyan et al., 2022), which is a Transformer sequence model capable of both text and image generation. For evaluations and understanding appropriate prompting technique, we prepend the retrieved documents as in-context examples to the main input document,

We train our retrieval-augmented, using a subset of MIMICIT dataset (Li et al. 2023) with 1.5M text-image pairs. Based on preliminary human evaluations we notice that the model performs well on image and caption generation, significantly outperforming the baseline Spoonbill with no retrieval. As a future work, we do want to evaluate on a larger dataset and do comparative studies against other models such as DALL-E and Flamingo.

3 Data

We have chosen 100K visual instruction samples from combination of dataset described below.

LLaVA-Interleaved (LA-I). Learning with in-context examples is essential for effective instruction tuning. To achieve this, we refine the LLaVA-Instruct-150K [28] dataset by retrieving ten in-context examples for each instruction-response pair in LLaVA-Instruct-150K, building LLaVA-Interleaved (LA-I). We identify each data's in-context examples based on instruction text-to-text similarity or image-image similarity.

Spot The Difference (SD). Learning to discern differences between images is vital for understanding real-world changes. Our study encompasses two interrelated task types in Scene Difference (SD), addressing varying complexity levels in difference identification. The first type, General Scene Difference, involves creating a pair of images by determining the most similar one to the current image, utilizing image-to-image similarity relationships from the COCO2017 [27]. The second type, Subtle Difference, features pairs of similar images with subtle distinctions sourced from the Spot-the-Diff[21], extracted from surveillance footage. For the first type, we prompt ChatGPT using original image captions and object detection annotations, while for the second type, we employ natural language difference descriptions as annotations. The resulting instruction-response pairs focus on identifying differences between the paired images.

Visual Story Telling (VIST). Beyond traditional scene understanding, the ability to generate coherent and engaging narratives based on visual input expands the context comprehension of Visual Language Models (VLMs). To enable this, we propose a task using the Visual Storytelling datase [20], which includes event-based image sequences and corresponding inquiry questions. Given that image annotations often contain narratives and timelines not directly observable, we instruct ChatGPT to act as a viewer answering questions about the images. The prompts also incorporate thought-provoking inquiries to promote creativity. Each task instance comprises multiple images and instruction-response pairs, providing in-context examples.

Dense Captions (DC). Expanding the scope of video understanding, DC features dense captions from [22] corresponding to clips within longer videos. The instructions pose a diverse set of questions, addressing the general visual content of the video, human actions, and behaviors, the chronological sequence of events, and causal relationships. This approach encourages VLMs to delve deeper into the intricacies of video content.

TV Show Captions (TVC). The primary purpose of incorporating TV show clips with high-level captions into the training process of VLMs is to enhance their social reasoning abilities and deepen their understanding of complex character dynamics. By organizing drama clips from [24] to analyze

character relationships and motivations, we aim to challenge VLMs to move beyond mere perception and demonstrate their reasoning capabilities within the context of TV show narratives. This focused approach is crucial for fostering advanced VLMs capable of effectively handling diverse real-world situations and user queries.

Alpaca - 150k language only instruction dataset is used to train Garuda on top of Llam2-7B-Chat

4 Model

4.1 Baseline

For the baseline model, we fine-tuned OpenFlamingo with no retrieval augmentation.

We introduce the dubbed Spoonbill model which is based on the OpenFlamingo model. We also introduce language-only instruction fine-tuned model dubbed Garuda which was fine-tuned on the Alpaca 150K dataset with Llama2-7B as the base model.

Let us introduce some basics of Flamingo model as a first step. Flamingo already demonstrates in-context ability using an interleaved sequence of images with text tokens. Generative vision language models output text conditioned on image text sequence. While other models incorporate only one image in their context, autoregressive vision-language models accept interleaved image-text sequences, enabling in-context learning.

Text tokens attend to their corresponding images via dense cross-attention modules which get attached to a frozen, autoregressive language model. To embed images, they extract patch features from a frozen vision encoder and pass them through a trainable Perciever resampler.

Spoonbill augments OpenFlamingo with visual instruction comprehension capability and preserves its in-context learning ability.

4.2 Spoonbill's architecture

Base model OpenFlamingo 9B was initialized. Garuda 7B chat which is based on Llama2 and fine-tuned on the language-only instruction dataset Alpaca was loaded from its final weights. Now the above-mentioned base OpenFlamingo's language encoder and decoder layers get updated with weights from Garuda 7B chat.

Now the above-said OpenFlamingo model with Garuda injected is finetuned on 100K visual-language instruction following dataset derived from datasources crafted in MIMIC-IT datasets.

The in-context external knowledge learning ability of Spoonbill is achieved through dense retriever which has the same training datasets in external memory indexed.

Each dataset sample consists of a queried image-instruction-answer triplet, with the instruction-answer tailored to the image. We freeze the vision encoder CLIP Vit-L/14 which is part of the base OpenFlamingo 9B model. After injecting the fine-tuned dubbed Garuda model into our base model OpenFlamingo 9B, we also freeze the language encoder. From this point, we only finetune the Perceiver resampler module, cross-attention layers inserted into the language encoder, and input/output embeddings of the language encoder which results in approximately 1.3 billion trainable parameters for the Spoonbill model.

Dataset triplet looks like (Iq, Rq, Xq), where , Iq denotes the q-th instruction in our dataset, Rq represents the response, and Xq refers to the images. Our primary objective is to develop a visual language model p (Rq | (Iq, Xq)) parametrized by trainable parameters , the model generates the response Ri for each query (Iq, Xq).

During training, we use a specific format to prepare training data. The format includes a combination of images, user instruction, and "GPT"-generated answers. [image], [endofchunk] are special tokens preserved from base model OpenFlamingo 9B.

<context> [image] User:<instruction> GPT:[answer] <answer>. [endofchunk] [answer] special token separates the answers from the instruction which helps to mask all the tokens after the special token [answer] during training. The prediction objective thus becomes predicting all masked tokens after the special token <answer>

We adopt "GPT" as a role label as well, because it does not have any specific semantic meaning in vocabulary. We trained our model using cross-entropy loss. Please refer to Figure.1 in the appendix section.

4.3 Inference

During inference, we prompted Spoonbill with Instruction, image pair and asked the question on the image. Spoonbill's dense retriever used both image, instruction pair to figure relevant triplet of instruction, image and answer as in-context demonstrations. Both in-context demonstrations and also the input prompt are concatenated as a final prompt.

Spoonbill was able to effectively learn from instructions and also image examples, to generate the answer for the visual question asked on top of input image.

4.4 Multimodal retrieval

4.4.1 Dense Retriever

A retriever takes a query tuple consisting of a question Q, and an image I. We encode the two parts separately using off-the-shelf frozen CLIP text and image encoders. For the candidate documents stored (using Faiss library), each encoding is stored separately with its own index in the retriever memory. During indexing phase, we go thru the entire instruction dataset and image dataset and index each one separately in the encoding space. Here, we deviate from RA-CM3 Model strategy (Yasunaga et al., 2023), where the approach is to average the two encodings with the L2 norm scaled to 1. The hypothesis here is to not bias the average towards a specific channel of the multi-modal document. We maintain a mapping of instruction to the image - so that we can cross reference each other to produce the correct instruction-image pair needed for in-context learning.

4.4.2 Retrieval strategy

At the retrieval phase, we obtain two lists of candidate documents (top 5) - one by its text relevance and one by its image relevance. Based on the overall top 10 results, we manually pick the top 2 triplet (instruction, image, answer) as candidates for in-context input to the LLM. In the future work, we hope to eliminate the hand-picked strategy by investigating prompting scenarios where the candidates are picked by textual and/or image based relevance.

Caveats: Our training data-set has instructions that can take more than one image for question-answering. For the results where we rely on instruction text relevance, we choose the first image in the sequence of images to form the triplet. Such instructions may have to be filtered out during manual evaluation phase. E.g. instructions like "Find the image among these five images that has a traffic signal", will not be a good example for in-context learning that takes a fixed number of images (in our case it is one) as input.

Shallow Retrieval	0-shot	4-shot	8-shot
Yes	55.75	55.14	42.97
No	55.75	53.97	35.36

Table 1: VQA Accuracy of Spoonbill on VQA-2 Dataset 10000 samples

5 Methods

5.1 Training Methods

Spoonbill was trained using Hugging Face's accelerator framework with deepspeed zero-3 intergration on 4 RTX A 6000 48GB GPU pods. We have used bf16 mixed precision during training. We have also used AdamW optimizer with starting learning rate of 10-5 and a batch size of 4. We started with one epoch and model was hallucinating. We then increased to 6 epochs, with learning rate scheduler.

Garuda (dubbed), a language model which is a result of fine tuning Llama2 with language instruction tuning dataset used for Alpaca. Parameter efficient fine tuning technique QLoRA was used to instruction tune Llama2 using 4-bit precision on a single A 100 80 GB for 3 epochs using supervised fine-tuning trainer.

The above said models are released to Hugging Face and please refer to the Appendix for links.

5.2 Evaluation Of Spoonbill

We use visual question answering accuracy VQA Eval which is like exact match metric used in question answering. We have evaluated the ability of Spoonbill on a subset of VQA-v2 dataset

We have also created vector space embeddings for only images using the Clip model. We have evaluated Spoonbill using zero-shot, 4-shot and 8-shot with relevant in-context retrieval and random in-context retrieval from the training set. We have presented the results in Table.1

6 Results

We are also showing human evaluation results in the Appendix section with Spoonbill. For the input Instruction, Image pair, Spoonbill retrieves relevant in-context demonstration (in below example, two examples) and asks Spoonbill to generate the answer. Please refer to Example-1 and Example-2 in the Appendix.

We notice that model hallucinates which is the drawback inherited from the base language model Garuda which in turn is based on Llama2-7B-Chat.

7 Analysis

Spoonbill has the ability for image captioning, visual question answering on images and text. However we focused only on visual question answering. Spoonbill can still hallucinate for some questions. Fine tuning dense retrieval can improve accuracy of retrieval which when used as in-context demonstrations can further help the model to generate accurate answers.

We believe further tuning on larger visual language instruction datasets for extra epochs with early stopping can help the model to follow instructions.

8 Conclusion

We presented Spoonbill, a retrieval-augmented multimodal model for text, and image which was built on top of OpenFlamingo 9B as a baseline model. Separately, language instruction following a model dubbed Garuda was built from Llama2 7B Chat using the Alpaca dataset. The above-mentioned language model was injected into the base model and then instruction tuned with visual-language instruction following datasets described above to create Spoonbill, a multi-modal retrieval augmented in-context and instruction the following model.

Spoonbill augments OpenFlamingo for task specific or knowledge specific tasks with its ability to depend on external knowledge to retrieve context.

Known Project Limitations

Spoonbill was not trained with any negative training examples. Spoonbill's language model Garuda is based on Llama-2 7B chat which can hallucinate. Hence, Spoonbill also inherits the limitation and can hallucinate.

Authorship Statement

@sugi venugeethan - Lead author and project lead who initiated the project, team, and built Garuda, Spoonbill along with mentoring the fellow authors who helped with this final project.

@aravind Venkatesan and @zihang Liu built the dense retriever and integrated it with Spoonbill. They have also helped in the final evaluation of Spoonbill. They also co-authored this paper on sections related to dense retrieval and prior literature.

References

- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, Pontus Stenetorp 2022, Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity https://arxiv.org/abs/2104.08786
- Laria Reynolds, Kyle McDonell 2021 Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm https://arxiv.org/abs/2102.07350
- Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee 2023 Visual Instruction Tuning https://arxiv.org/abs/2304.08485
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, Ziwei Liu 2023 Otter: A Multi-Modal Model with In-Context Instruction Tuning https://arxiv.org/abs/2305.03726
- Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, Kai Chen 2023 MultiModal-GPT: A Vision and Language Model for Dialogue with Humans https://arxiv.org/abs/2305.04790
- Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, Wen-tau Yih 2023 Retrieval-Augmented Multimodal Language Modeling https://arxiv.org/abs/2211.12561
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, Ludwig Schmidt 2023 OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models https://arxiv.org/abs/2308.01390

Yasunaga, M., Aghajanyan, A., Shi, W., James .R, Leskovec, J., Liang, P., Lewis, Mike., Zettlemoyer, L., Yih, 1 W. Retrieval-Augmented Multimodal Language Modeling *arXiv:2211.12561*

Chen, W., Hu, H., Chen, X., Verga, P., and Cohen, W. W. Murag: Multimodal retrieval-augmented generator for open question answering

over images and text. *In Empirical Methods in Natural Language Processing (EMNLP), 2022a.*

Chen, W., Hu, H., Saharia, C., and Cohen, W. W. Reimagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491, 2022b.*

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198, 2022.*

A Appendix

A.1 HuggingFace Links of Spoonbill and Garuda

1. Garuda from Llama2 7B Chat
2. Spoonbill-Garuda based
3. Spoonbill-Llama2 based

A.2 Dense Retriever

A.2.1 Embeddings

We used the SD and CGD training dataset from MIMIT (https://huggingface.co/datasets/pufanyi/MIMICIT). Each dataset is composed of three parts, including:

1. "xx_instructions.json" file: the instruction-response pairs (also includes image ids and related instructions ids for each instruction-response pair) for each task.
2. "xx_images_preview.json file": Stores the image numbers and their corresponding base64 codes in lossy compressed JPG format.

In the indexing step - the above files were processed to create instruction text embeddings and image embeddings in a vector space using Faiss library. The Faiss system uses a sequential numbering scheme (starting at 0) to index these. So, it is important to map these sequential indexes back to original ids. For this we maintained an in-memory dictionary which was written onto the file-system (13MB) so that it could be referenced at the time of retrieval.

A.2.2 Retriever

The retrieval process involved getting top-5 most relevant images and images from the encoding index. The goal was to return an overall 10 triplets of instruction, answer and image. From these 10 triplets, we wanted to handpick 2 triplets as candidates to in-context learning.

The dataset contains lot of instruction that have instruction based on more than one image. Since the scope of this paper is to do visual question answering with a single image, such instructions are not considered during human evaluation.

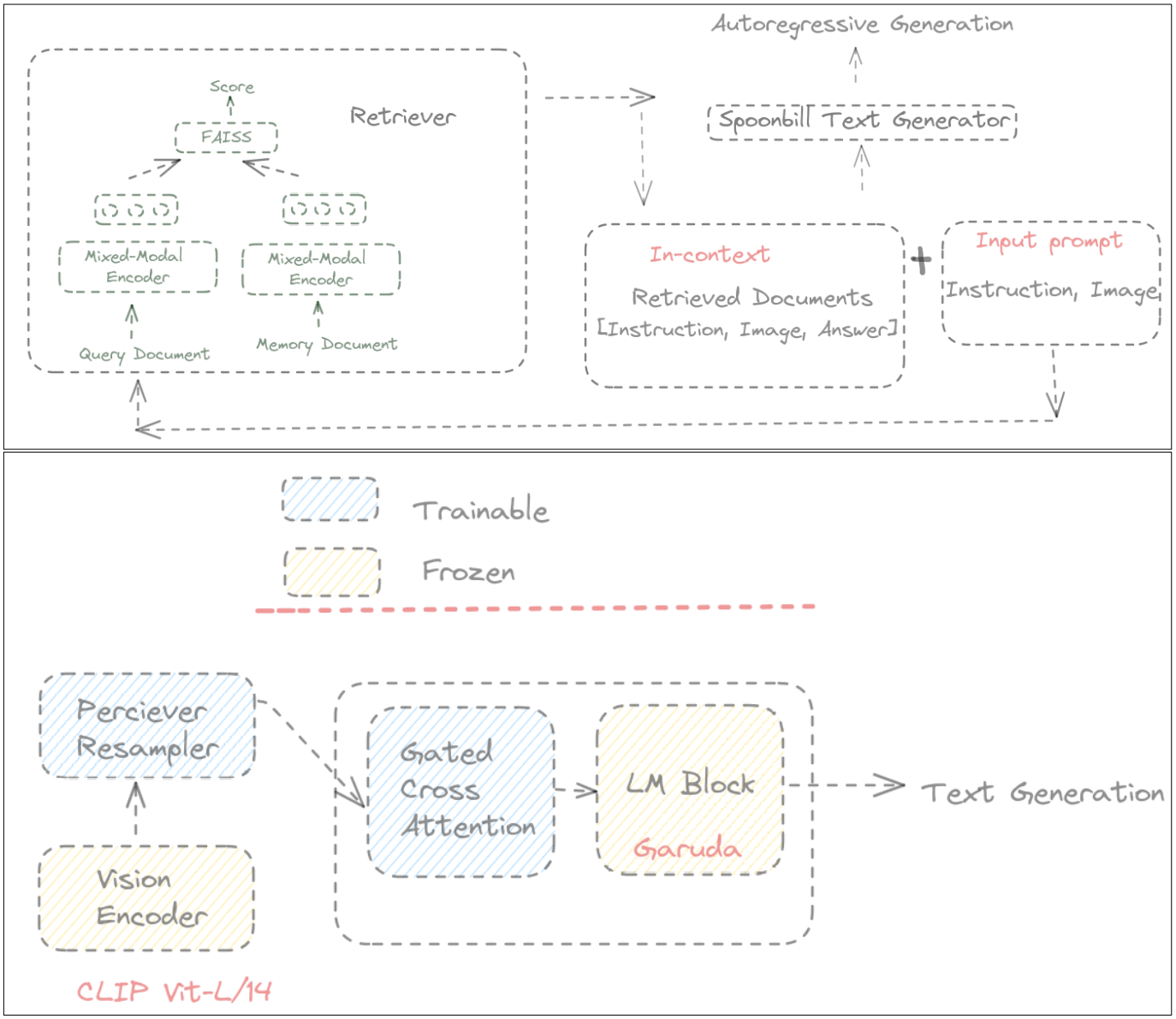


Figure 1: Spoonbill Architecture

What is the topping of the pizza in image?
The pizza is sausage pizza.

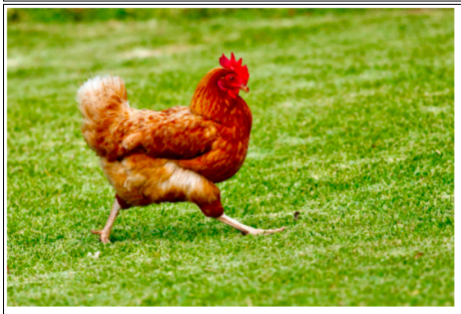
What is the difference in the toppings of the pizza in image?
The pizza in image has ham, cheese, olives and oysters as toppings.

What is the object in the image?
I cannot determine whether it is a cheese and vegetable pizza or a sausage and cheese pizza as it is not mentioned in the description.

Figure 2: Example-1 2-shot VQA with Spoonbill - top two are in-context

What is the animal in the image?
The bird in the image is a chicken.

What is the color of the bird in the image?
The birds in image are brown.



Where is the chicken and What is it doing?
She is standing on a grassy ground and eating grass.

Figure 3: Example-2 2-shot VQA with Spoonbill - top two are in-context