Google DeepMind

Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities

Hexiang Hu†, Yi Luan†, Yang Chen‡, Urvashi Khandelwal†, Mandar Joshi†, Kenton Lee†, Kristina Toutanova†, Ming-Wei Chang†
†: Google DeepMind, ‡: Georgia Tech

(Paper ID: 3031)

Introduction

We introduce a new task called **O**pen-domain **V**isual **E**ntity **R**ecognitio**N**, with the goal of recognizing open-domain visual entities in the wild.

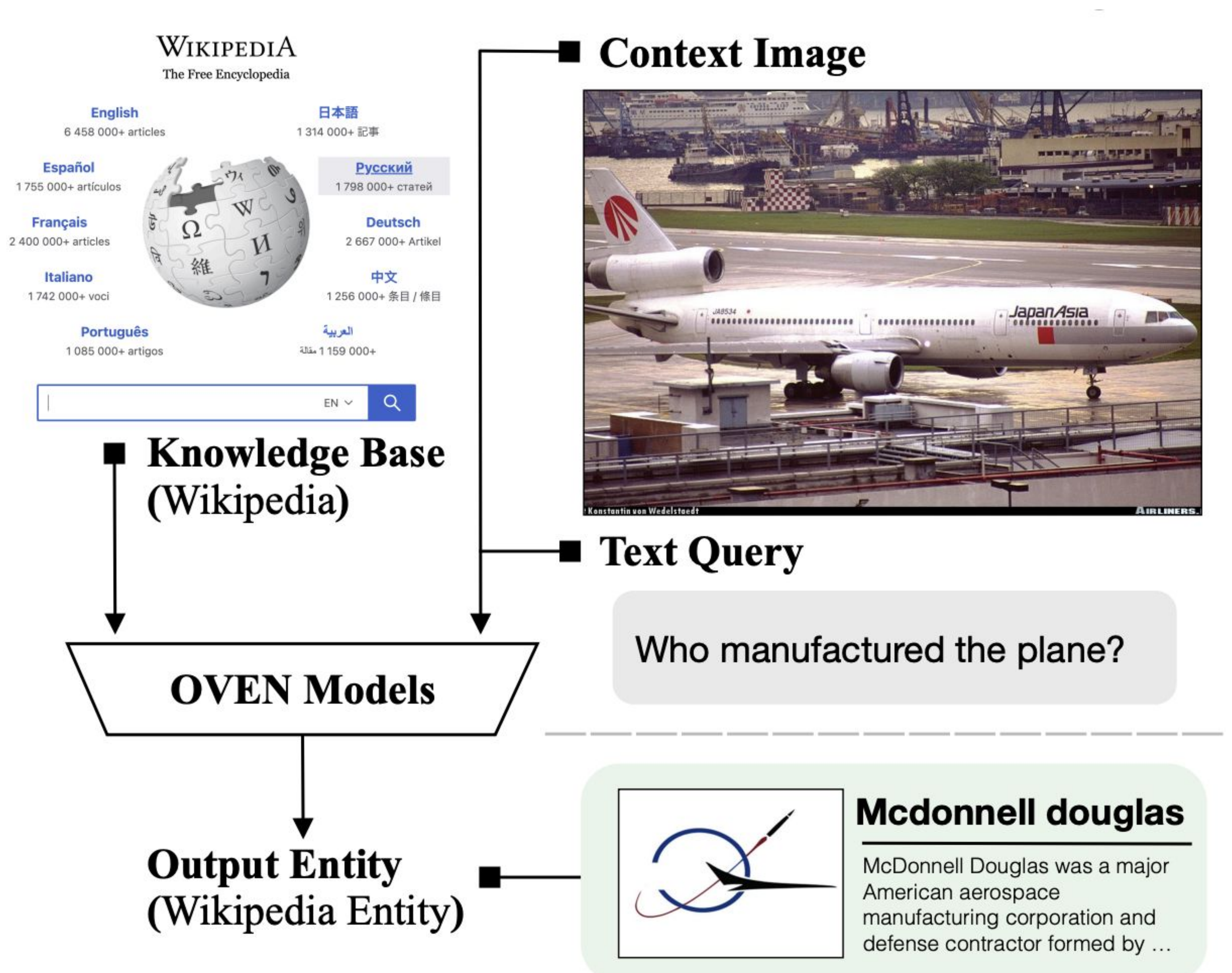
Different from *traditional recognition*, **OVEN** focus on recognizing an queried visual entity from a *very large label space* defined by knowledge base (KB), such as English-Wikipedia, with 6M+ entities.

Different from *visual QA tasks*, **OVEN** focus on generalizable visual recognition, and aims to link queried image with the Web KB.

Contribution.

- Formalize and introduce the task of OVEN.
- Unify 14 image recognition, or VQA datasets, and build a general domain OVEN dataset that recognizes 6M wikipedia entities.
- Perform human annotation on the proposed task, for evaluation and upper-bound performance study.
- Evaluate different type of SoTA multimodal foundation models on our dataset, and characterize the pros and cons of those models.

What is OVEN?



Task Definition. The *input* to an OVEN model is a pair of image x^p and query text x^t, with text x^t expressing the **recognition intent** (e.g. "`what is the model of aircraft?`" vs. "`what is the airline company?`") that corresponding to the image x^p.

Given a knowledge base $\mathcal{K} = \{(e, p(e), t(e)) \mid e \in \mathcal{E}\}$ of triples:

- e: database identity, *i.e., Wikidata id (Q7395937)*
- `t(e)`: textual info of an entity, *i.e., the name of entity.*
- `p(e)`: visual info of an entity, *i.e., Wiki images of the entity.*

The goal of OVEN learner is to predict the entity e of a given input example x = (x^t, x^p) from the KB $\mathcal{K}$

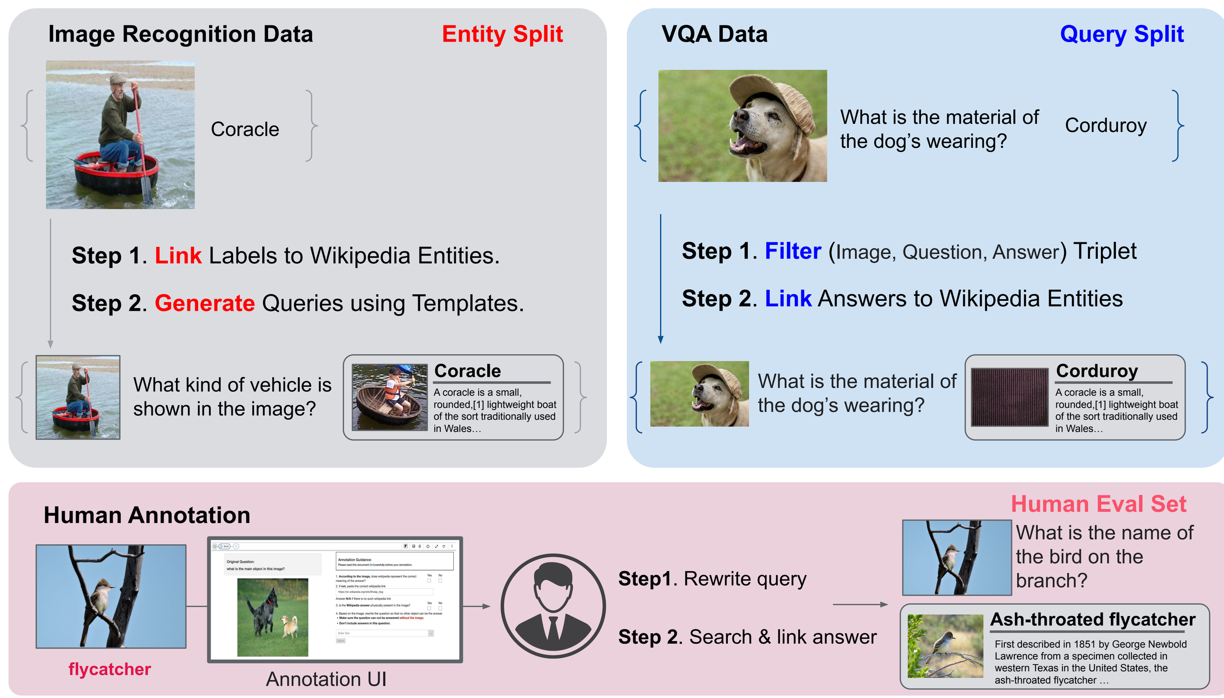
Remark 1. OVEN can be seen as a specialized VQA task, focus on answering "What" questions.

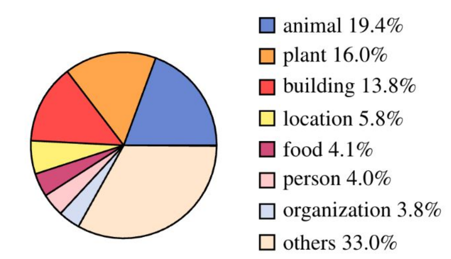
Remark 2. Different from VQA, the answer to OVEN is a visual entity that grounded on the knowledge base (Wikipedia), instead of free-from string, which suppose to have a concrete definition.

Remark 3. OVEN can also be viewed as a recognition task, but without any classification prior (e.g. animal, or vehicle classification). Instead, the text query input x^t specifies the domain and goal of recognition, which reduces ambiguity in open-domain recognition.

Dataset Construction

We re-annotate 14 existing recognition and VQA datasets.





	Train Set	Val Set	Test Set	Human Set
# unique queries	19,129	3,124	18,341	17,669
# SEEN entities	7,943	1,942	10,137	2,487
# SEEN examples	4,958,569	63,366	366,350	14,016
# UNSEEN entities	0	2,004	10,156	2,174
# UNSEEN examples	0	66,124	362,909	10,851
# Total examples	4,958,569	129,490	729,259	24,867

	Wiki_{EN}
# entities	6,084,491
# images	2,032,340
# title	6,084,491
`AvgLen`(title)	2.93

Q: What is this building called?
A: Château de Suscinio (Q2970742)

Q: Which species of fungi is this?
A: Hygrocybe coccinea (Q39528)

Q: What kind of food is this?
A: Sugar candy (Q17119302)

Q: What is the mural of?
A: Golden Gate Bridge (Q44440)

Q: What is attached to the ceiling?
A: Track lighting (Q117208161)

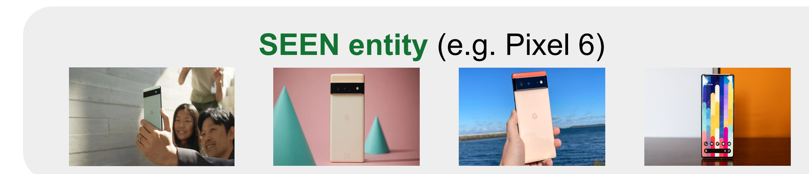
Q: What is the model of this car?
A: Toyota Sequoia (Q1512971)

Q: What is the main object in this image?
A: Coffee percolator

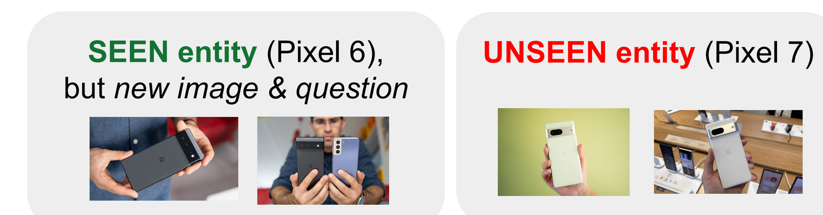
Q: Which category of equipment is shown in the image?
A: Dumbbell

Evaluation focus on **Generalization**

Model Training



Model Evaluation



- Model are evaluated using the **Harmonic mean** over **SEEN** & **UNSEEN** accuracies. (model has to balance fitting and generalization)
- The overall performance is then computed as the **Harmonic mean** of **Entity** and **Query** split performance.

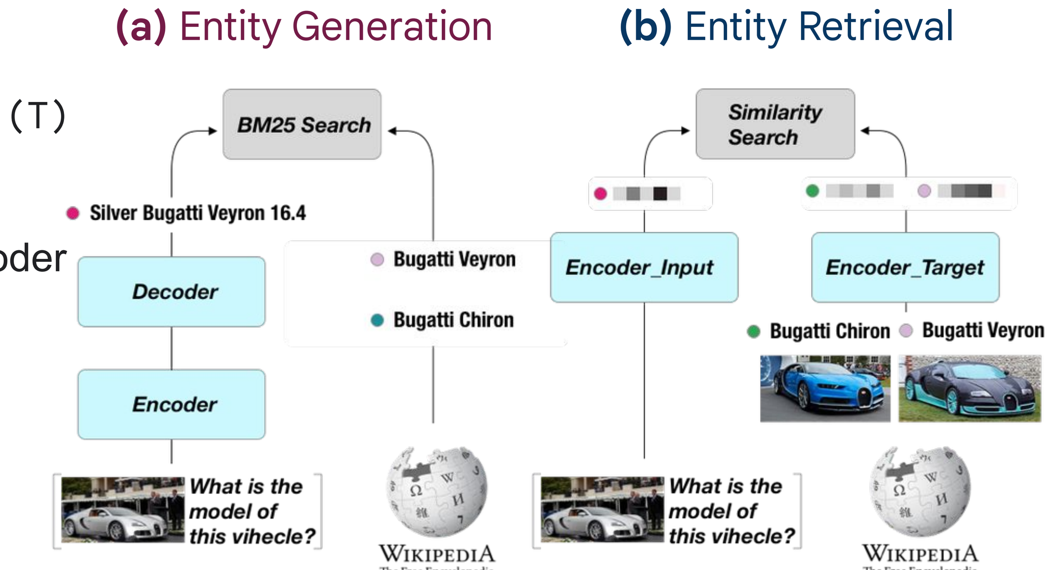
Models for OVEN

(a) Entity Generation Encoder-Decoder:

- Input:
  - `Context Image (I) + Query Text (T)`
- Output:
  - `Entity Name in KB`
- **PaLI-17B**: Multimodal Encoder, Text Decoder

(b) Entity Retrieval Dual Encoders:

- Input:
  - `Context Image + Query Text`
  - `Image & Text of Entity in KB`
- Output:
  - `Entity Name in KB`
- **CLIP2CLIP**: Ensemble of **CLIP** models



Benchmark Results

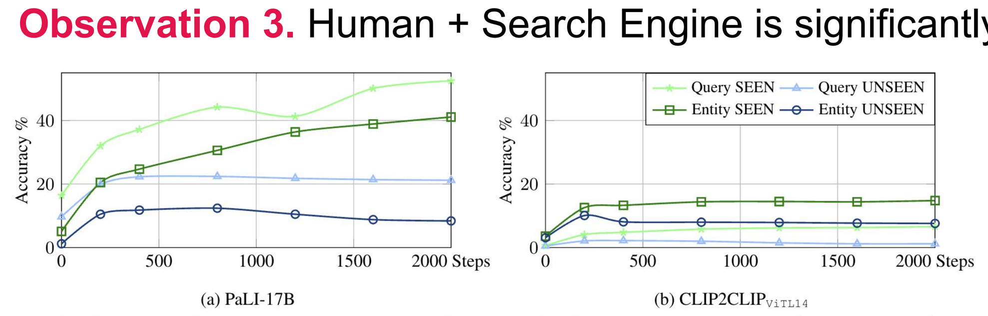
We evaluate prior entity retrieval and generation models (SoTA at the time) on OVEN.

	# Params	Entity Split(Test) SEEN	Entity Split(Test) UNSEEN	Query Split(Test) SEEN	Query Split(Test) UNSEEN	Overall(Test) HM	Human Eval SEEN	Human Eval UNSEEN	Human Eval HM
Dual Encoders:									
● $\text{CLIP}_{\text{ViTL14}}$	0.42B	5.6	4.9	1.3	2.0	2.4	4.6	6.0	5.2
● $\text{CLIP Fusion}_{\text{ViTL14}}$	0.88B	33.6	4.8	25.8	1.4	4.1	18.0	2.9	5.0
● $\text{CLIP2CLIP}_{\text{ViTL14}}$	0.86B	12.6	10.5	3.8	3.2	5.3	14.0	11.1	12.4
Encoder Decoder:									
◆ PaLI-3B	3B	19.1	6.0	27.4	12.0	11.8	30.5	15.8	20.8
◆ PaLI-17B	17B	28.3	11.2	36.2	21.7	20.2	40.3	26.0	31.6
Human+Search [6]	-	-	-	-	-	-	76.1	79.3	77.7

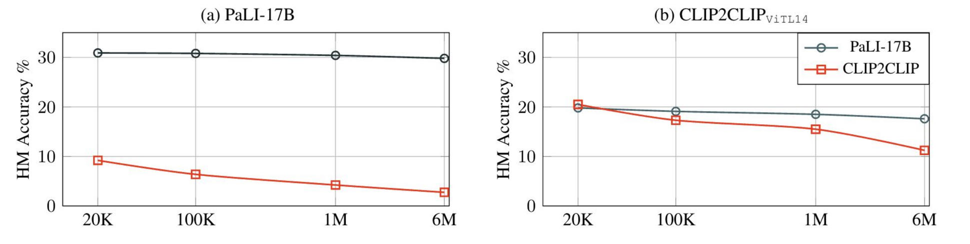
Observation 1. PaLI-based models are significantly better than CLIP
(Performance gap on **Query Split** is bigger)

Observation 2. Scaling PaLI from 3B to 17B creates significant improvement
(this scaling includes both change in language model: 1B to 13B, and change in visual model: ~2B to ~4B)

Observation 3. Human + Search Engine is significantly better than current models



Ablation 1. Over-finetuning models on OVEN leads to strong SEEN acc but weak UNSEEN acc, thus bad overall performance

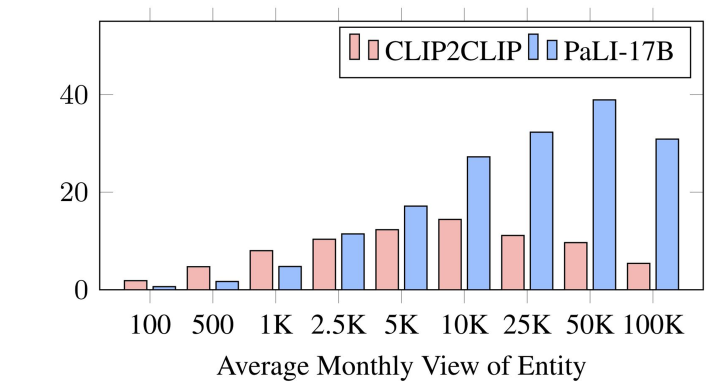


Ablation 2. As the # of Wikipedia candidate space grows, the intrinsic task difficulty grows. Meanwhile, the performance of retrieval model is more affected.

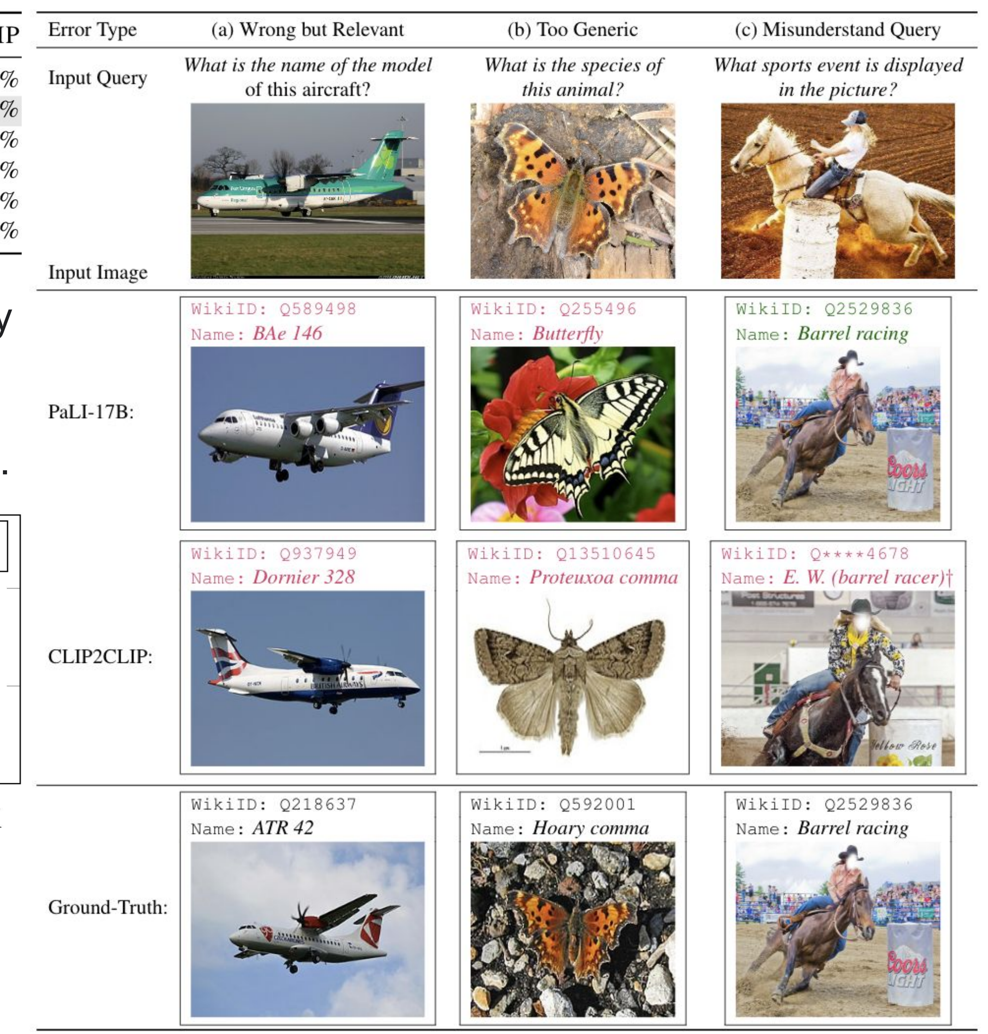
Model Analysis

	PaLI-17B	CLIP2CLIP
Correct	29%	15%
In-correct	71%	85%
→ (A) Wrong But Relevant	23%	27%
→ (B) Too Generic	15%	1%
→ (C) Misunderstand Query	7%	37%
→ (D) Miscellaneous	24%	20%

- PaLI tends to answer generically when it is uncertain
- Most CLIP errors are due to misunderstanding the questions.

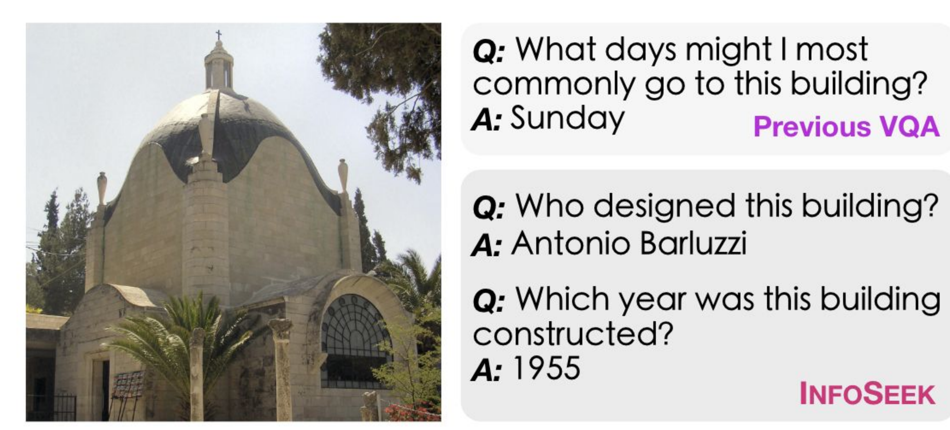


- PaLI show clear win on recognizing head entity
- CLIP outperforms PaLI on tail entities



Towards Understanding Visual Info-Seeking Question

In a follow-up work (dubbed InfoSeek), we propose another task that extend the scope of open-domain visual recognition to open-domain visual info-seeking question answering.



We construct datasets to support Knowledge-intensive VQA, s.t.

- Question are visual info-seeking (asking unknown rather than common sense)
- Answers are fine-grained
- It shows that SoTA multimodal foundation model still can not answer such question well

Resources

Dataset: https://open-vision-language.github.io/oven
Contributed Baseline & Eval: https://github.com/edchengg/oven_eval
Follow-up InfoSeek Project: https://open-vision-language.github.io/infoseek