Using Knowledge Graphs to harvest datasets for efficient CLIP model training

Simon Ging, Sebastian Walter, Jelena Bratulić, Johannes Dienert,
Hannah Bast, Thomas Brox. gings@cs.uni-freiburg.de

GCPR 2025
German Conference on Pattern Recognition

Overview

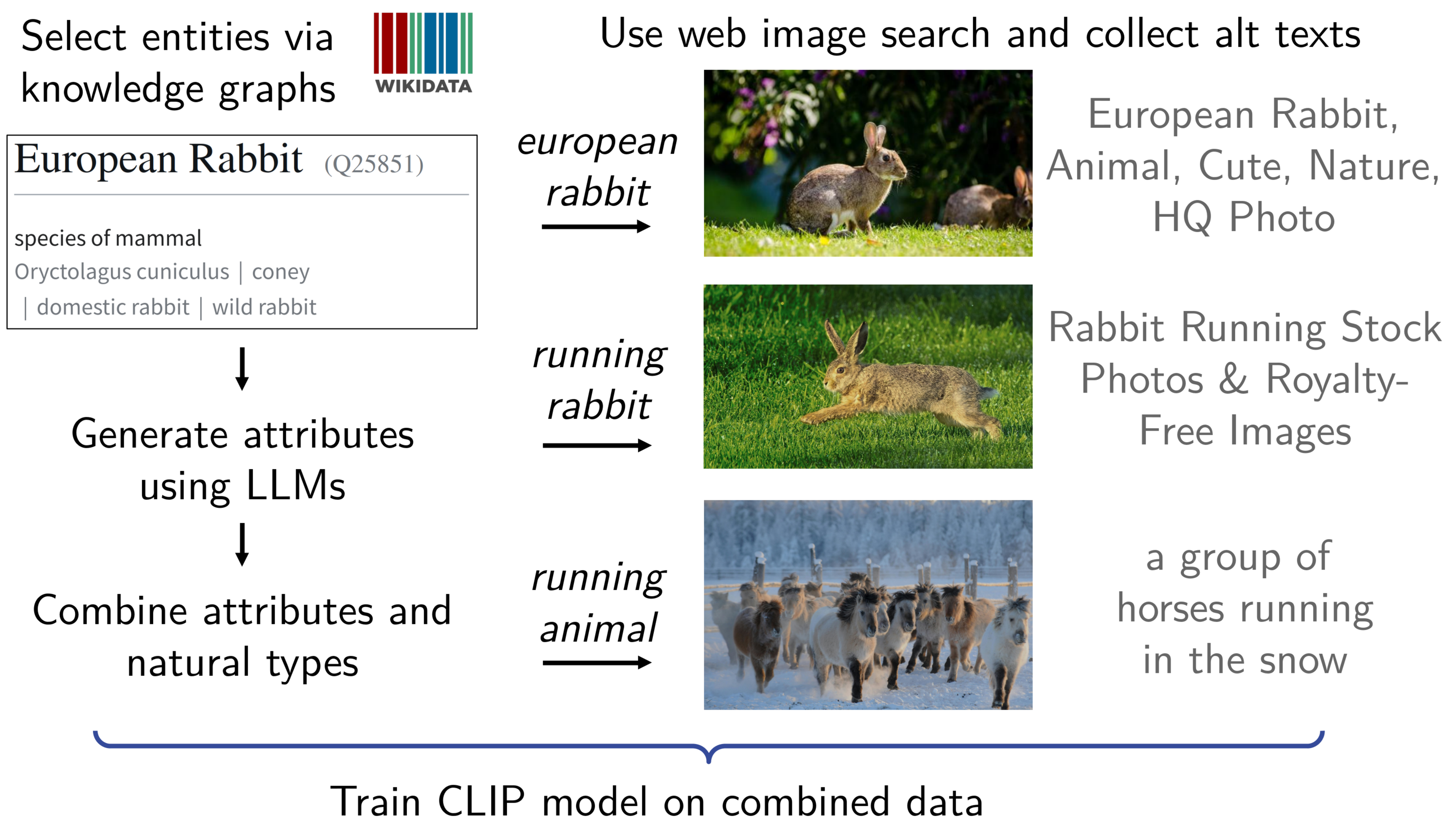
Training high-quality CLIP models requires enormous datasets. This:

- limits the development of domain-specific models, and
- complicates scientific research that investigates training CLIP models.

We show how a robust CLIP model can be trained from scratch with less data:

- We propose a dataset creation method based on a knowledge graph and an image search engine,
- build an expert foundation model for living organisms using 10M images,
- introduce *EntityNet*, comprising 33M images paired with 46M texts, and
- train a generic CLIP model on *EntityNet* in significantly reduced time.

EntityNet dataset creation



1. Extract entities from knowledge graphs
2. Generate attributes and natural types
3. Run web image search
4. Vision-language pretraining on alt texts and knowledge graph labels.

Entity	*Description*	*Sitelinks*	*Aliases*
tiger	species of big cat	216	tigress, tigers, Panthera tigris
chest	box-shaped type of furniture	51	coffer, kist
muscle car	type of high-performance car	30	high performance car

Entity	*Category*	*Attribute*	*Search query*
rock	Pattern and texture	porous	porous rock
wolf	Environment	snow	wolf in the snow
residence	Parts	arches	arches in residence architecture
garlic	Shape and size	big	big garlic bulb
farm	Other	tourist	tourists visiting a farm
boot	Color	multicolor	multicolored boots

Examples of **entities** (top) and **attributes** (bottom) extracted from knowledge graphs and LLMs.

Query set	Img.	Queries	Ent.	Aliases	Attrib.	Alt texts	Example query
World entity	23M	158k	74k	101k	-	23M	ship
+ attribute	19M	139k	6k	16k	20k	16M	small handbag
Living entity	9M	72k	63k	51k	-	8M	kohlrabi
+ attribute	9M	53k	1k	3k	5k	6M	tropical plant
All	33M	416k	135k	149k	23k	45M	-

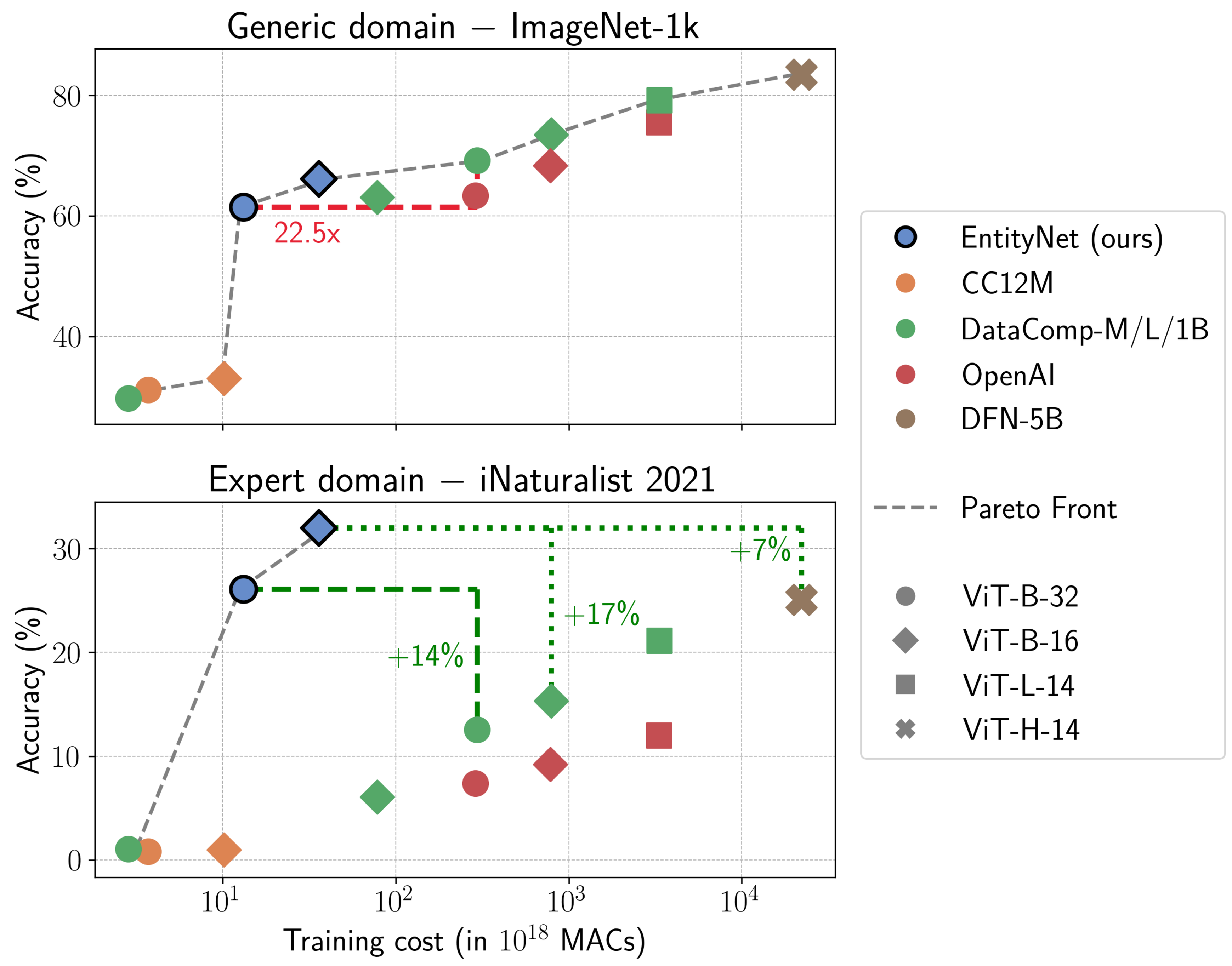
EntityNet dataset statistics, unique elements per column.

Component analysis

Arch.	Dataset	MACs (1e18)	Image-Net	Retrie-val	Distr. shift	iNat. 2021	CUB	Rare Species
B-32	Everything	13.1	**61.5**	**37.2**	**41.0**	26.1	79.5	**42.7**
B-32	No living organisms	9.0	39.2	32.1	28.0	0.8	6.2	6.9
B-32	Only living organisms	4.1	36.0	16.5	21.0	**28.6**	**83.2**	42.0
B-32	No attribute queries	8.7	54.8	28.6	33.8	25.6	79.7	39.2
B-32	50% alt text	13.1	**61.5**	37.2	**41.0**	**26.1**	**79.5**	**42.7**
B-32	100% alt text	13.1	59.1	**38.3**	38.1	22.9	78.8	39.7
B-32	0% alt text	13.1	55.7	13.5	35.5	24.2	78.1	29.7

Model effectiveness when varying dataset composition and text sampling. Notably, training on 50% alt text and 50% knowledge graph text performs best.

Results



Harvesting datasets for training CLIP models with an improved quality-cost trade-off, either for a generic domain (top) or an expert domain (bottom).

Arch.	Dataset	MACs (1e18)	Image-Net	Retrie-val	Distr. shift	iNat. 2021	CUB	Rare Species
B-32	CC12M	3.7	28.6	25.6	18.3	0.7	9.2	–
B-32	CommonPool-M	2.9	27.2	20.2	19.8	0.8	10.1	–
B-32	DataComp-M	2.9	29.7	19.5	20.5	1.0	16.8	–
B-32	OpenAI	288.6	63.4	49.6	48.7	7.4	51.8	–
B-32	DataComp-1B	295.4	**69.2**	**54.0**	**56.3**	12.6	73.8	–
B-32	EntityNet (ours)	13.1	61.5	37.2	41.0	**26.1**	**79.5**	**42.7**
B-16	BioCLIP	61.3	18.6	0.8	15.4	–	78.1	38.1
B-16	CommonPool-L	78.2	57.8	45.6	47.0	4.1	35.1	–
B-16	DataComp-L	78.2	63.1	49.4	51.1	6.1	48.1	–
B-16	DataComp-1B	791.4	**73.5**	**57.4**	**64.4**	15.3	79.0	–
B-16	OpenAI	784.6	68.3	52.1	58.6	9.2	56.1	–
B-16	EntityNet (ours)	36.0	66.2	39.8	47.4	**32.0**	**85.3**	**47.1**
L-14	OpenAI	3328.4	75.5	54.3	71.4	12.0	62.9	–
L-14	DataComp-1B	3338.6	79.2	61.8	74.9	21.1	85.5	–
H-14	DFN-5B	22164.0	**83.4**	**68.7**	**76.3**	25.1	**88.1**	–

Training **CLIP** on *EntityNet* from scratch. **MACs** (multiply–accumulate operations) measure training cost. "–": Model does not fulfill zero-shot requirements.

Arch.	Dataset	MACs (1e18)	Image-Net	Retrie-val	Distr. shift	iNat. 2021	CUB	
B-32	DataComp-1B	295.4	69.2	**54.0**	**56.3**	12.6	73.8	–
B-32	EntityNet	13.3	**69.5**	50.8	53.3	29.5	83.3	–
B-32	Only living organisms	4.2	48.2	31.2	33.3	**37.0**	**87.0**	–
B-16	DataComp-1B	791.4	73.5	**57.4**	**64.4**	15.3	79.0	–
B-16	EntityNet	36.1	**73.5**	52.2	61.0	34.9	86.5	–
B-16	Only living organisms	11.3	51.4	34.8	39.2	**42.7**	**90.3**	–

Finetuning **DataComp-1B CLIP** on *EntityNet*. The B-16 CLIP model fine-tuned on living organisms outperforms all other CLIP models on recognizing fine-grained species.

Conclusion

Our dataset is efficient for pretraining CLIP models, both in an expert domain, and a generic domain. It enables controlled training experiments, previously limited to low-quality or computationally expensive models. Finally, our dataset harvesting method can be used to build new datasets given a knowledge graph and an image search engine.

small data MORE FROM LESS CRC 1597
Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation