Wiki Workshop (11th edition) – June 20, 2024

Visual Entity Linking for Structured Data on Wikimedia Commons

Philipp Bielefeld[*1] **Jasmin Geppert**[*1] **Necdet Güven**[*1] **Melna Treesa John**[*1] **Adrian Ziupka**[*1]
Lucie-Aimée Kaffee[2] **Russa Biswas**[1] **Gerard de Melo**[1]
[1]Hasso Plattner Institut, Potsdam, Germany
[2]Hugging Face, lucie.kaffee@huggingface.co

Abstract

Linking Wikidata items to Wikimedia Commons images enables a wide range of automation tasks, such as search and organization, as well as downstream tasks, such as labeling of images or training machine learning models. However, there is currently a lack of structured data-labelled images on Wikimedia Commons. To close this gap, we propose the task of *Visual Entity Linking for Wikimedia Commons*, in which we create new labels for Wikimedia Commons images from Wikidata items. For this, we create a dataset[1] and finetune pre-trained models based on the CLIP architecture. While our best-performing models show promising results, we also acknowledge the drawbacks of the current dataset and the difficulty of the task.

Keywords: Wikimedia Commons, Wikidata, Visual Entity Linking, Multimodality, CLIP

Introduction

Wikimedia Commons serves as Wikimedia's and Wikipedia's image hosting service, providing about 100 million images along with other media files. The image information includes metadata as well as (multilingual) textual descriptions and Wikipedia-like categories. *Wikidata*, Wikimedia's knowledge graph (KG) comprises around 100 million entities, primarily *items*. To facilitate organising and searching Commons images, the *Commons:Structured Data* project was initiated in 2017.[2] For a given image, community members tag relevant Wikidata items that are portrayed in that image, which are added to Commons as structured data via new *depicts* statements. This makes it possible to associate images with universal, language-independent concepts in a machine-friendly way.

Naturally, the greater the coverage of such Commons image annotations, the more useful the structured data becomes. However, as of November 2023, only around 15% of all Commons images have at least one Wikidata item linked. Our work aims to fill this gap by automatically providing suggestions of depicted items in an image. Eventually, this could also be used on newly uploaded Commons images and extended to other use cases drawing on Wikimedia Commons images as training data. To enable this, we employ **Visual Entity Linking**, the multimodal task of linking KG entities (often text) to images displaying these entities to support, e.g., image understanding, visual question answering, and more accurate image searches. Previous work has considered visual entity linking in the domain of people Sun et al. (2022). Hu et al. (2023) link images to over 6 million open domain entities to (the English) Wikipedia, a dataset they create through crowdsourcing. Both approaches use models that are partly based on CLIP (Radford et al., 2021).

Dataset

To create a novel dataset for Visual Entity Linking for Wikimedia Commons, we use the dumps of Wikimedia Commons and Wikidata (Figure 1). We take all Wikidata items of an image's *depicts* statement as its ground truth. The Wikidata KG items in this work are represented by the concatenation of their English name and description. However, around 50% of these items only occur once as ground truth and 90% in total occur less than ten times, meaning the distribution is heavily skewed. For this work, we remove long-tail entities in the following way; we set a minimal threshold of 10 occurrences of Wikidata items. To keep as many labels as possible intact, for the long-tail entities below this threshold, we use the graph structure of Wikidata (*subclass of* and *instance of* statements) to ascend the class hierarchy. If more generic items can be found within a maximum of three hops, the original fine-grained ground truth item is replaced with its more generic parent item(s). We sample 1 million images for our dataset (see Table 3). For example, if the item of *Marie Curie* were annotated fewer than 10 times, it would be replaced with the entity *human*.

Challenges While building this dataset, we identified the following limitations. (1) The guidelines for the *depicted* statement, as many community guidelines, vary across the project, e.g., while sometimes it is advised to

[*] In alphabetical order, these authors contributed equally to this work

[1] Dataset: https://huggingface.co/datasets/aiintelligentsystems/vel_commons_wikidata

[2] https://commons.wikimedia.org/wiki/Commons:Structured_data

not add generic items if more specific ones are already marked[3], in other locations the recommendation is to add *both* generic and specific items.[4] This propagates to the data, i.e., different images with similar content might be annotated differently. (2) Even after filtering with our threshold of 10, there are very specific items; the item Q17447776 *Flintenweg 8, Orvelte* is still present in the dataset despite not even having a description on Wikidata, due to the (relatively) high volume of images annotated with this item. (3) There are many instances of near-identical items, describing similar concepts with different labels.

Experiments & Results

We ran preliminary experiments using the state-of-the-art multimodal model CLIP (Radford et al., 2021) and report the results in Table 1.[5]

Baselines The *random baseline* randomly picks an item from the candidate pool with a probability equal to their frequency in the train split. The *top-k baseline* algorithm picks always the same ten, most frequent items for every image, based on the train data.

Zero-shot model & baseline algorithms The zero-shot CLIP model does not perform well and only achieves a recall score of over 15 at the tenth rank. In a qualitative investigation, we find a trend of the model predicting more specific items. For example, for an image of a person, the model predicts specific names. We believe this results from CLIP's pre-training, where the ground-truth texts were more specific to the image compared to our dataset's labels.

MLP Naive CLIP model. The basic CLIP model with both CLIP encoders frozen and a simple MLP head already performs quite well with a recall score of over 50 at rank ten, i.e., on average the model suggests a correct item on every second image.

We further find that the recall score at the ranks 20, 50 and 100 goes up 62.4, 74.8 and 82.4, respectively, with rank 100 still being among the first 0.5% of all candidate items. The actual prediction scores (the cosine similarities) are close to each other: on average 0.29 at rank one and still 0.25 at rank 100, and similarly for the other models.

However, manually inspecting example predictions reveals that the model achieves a good balance between more specific and generic items. It is able to safely pick up the image content, for example when predicting *presenter*, *microphone* or *award* besides the correct *human*[6] instead of outputting the names of specific persons.

Encoder finetuning Following Zhai et al. (2022), we finetune CLIP's text encoder while freezing the image encoder. This model has by far the highest precision of the tested ones, most notably at rank one.

Since we are limited to a batch size of 256 in this scenario, we also finetune another CLIP model on this batch size for a fair comparison, but keep both encoders frozen. We can see that a finetuned encoder achieves much better precision, especially for the top rank, while the other model slightly outperforms it in the recall. Overall, finetuning the text encoder does not yield generally better results than only training the MLP head, but might outperform the other models when used with a larger batch size.

Conclusion

Our work shows the potential of automated visual entity linking to provide suggestions of Wikidata items for Commons images. While the task remains challenging, with proper validation by the community, this project has the potential to greatly increase the amount of structured data on Commons.

References

[Hu et al.2023] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of Wikipedia entities.

[Radford et al.2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

[Sun et al.2022] Wen Sun, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. 2022. Visual named entity linking: A new dataset and a baseline, December.

[Zhai et al.2022] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. LiT: Zero-shot transfer with locked-image text tuning.

[3] `https://commons.wikimedia.org/wiki/Commons:Depicts#What_items_not_to_add`

[4] `https://commons.wikimedia.org/wiki/Commons:Depiction_guidelines#Depicts_level_of_detail` (marked as disputed)

[5] Due to the multilabel challenge of our task, we adjust the loss targets: They can also be the ground truth items' parents in the class hierarchy. Here, we set the number of loss target hops to one. We used a learning rate of 0.001, batch size of 1,024, and AdamW optimization. We rescale the item gradients by the inverse batch frequency (to account for the fact that items like *human* often appear multiple times).

[6] `https://commons.wikimedia.org/?curid=28127864`

Model	Recall			Diversity Recall			mAP		
	@1	@5	@10	@1	@5	@10	@1	@5	@10
Random baseline	2.1	9.6	17.2	2.1	6.5	11.5	2.1	3.1	3.7
Top-k baseline	12.4	29.8	40.8	12.4	20.5	29.8	12.4	14.3	15.9
Zero-shot	4.7	11.5	15.9	4.7	7.5	10.3	4.7	4.7	5.1
MLP Naive CLIP	16.2	**40.5**	**51.8**	16.2	**27.5**	**37.2**	16.2	17.1	18.7
TE Naive CLIP BS 256	**20.6**	37.5	45.0	**20.6**	26.0	31.8	**20.6**	**19.0**	**20.0**
MLP Naive CLIP BS 256	14.2	38.8	49.8	14.2	26.3	35.6	14.2	15.7	17.3

Table 1: Comparison of the performance of various model setups on our test split (zero hops in the metrics). Default batch size is 1,024. ”MLP” = CLIP encoders frozen, ”TE” = finetuned text encoder, ”BS” = batch size.

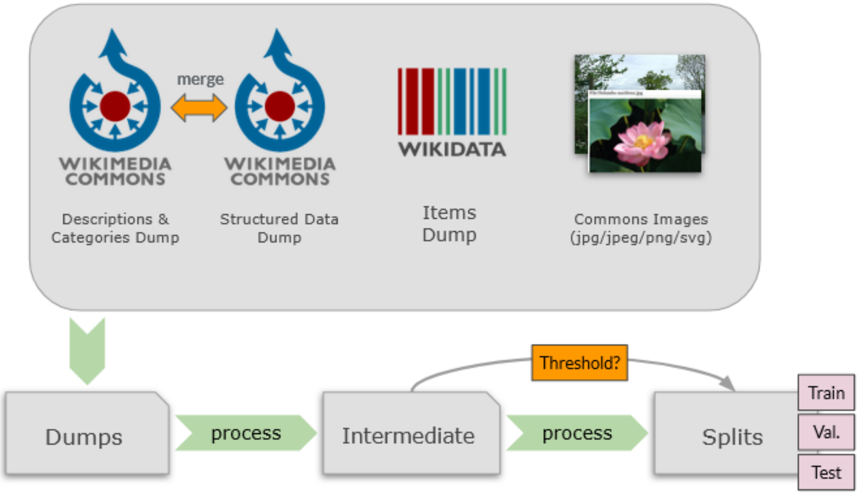

Figure 1: Our pipeline going from raw dumps to experiment-ready datasets.

f=0		f=10	
Label	Freq.	Label	Freq.
road	34,615	human	119,233
village	16,186	painting	55,213
agriculture	16,117	taxon	44,461
path	15,601	village	37,040
house	14,943	road	36,159

Table 2: Most frequent items in the train split, before and after filtering, where f is the entity threshold, for our experiments we use the threshold of 10.

train	800,000
test	100,000
val	100,000
items	18,522

Table 3: Total number of images in the train, test, validation splits, and total number of Wikidata items across the dataset.