VisualSem: A High-quality Knowledge Graph for Vision & Language

Houda Alberts[1,*] **Ningyuan (Teresa) Huang**[5] **Yash R. Deshpande**[2] **Yibo Liu**[2]
Kyunghyun Cho[2] **Clara Vania**[4] **Iacer Calixto**[2,3]

[1]Rond Consulting, NL [2]New York University [3]ILLC, University of Amsterdam
[4]Amazon, UK [5]Johns Hopkins University, USA.
houda1996@hotmail.com,iacer.calixto@nyu.edu

Abstract

An exciting frontier in natural language understanding (NLU) and generation (NLG) calls for (vision-and-) language models that can efficiently access external structured knowledge repositories. However, many existing knowledge bases only cover limited domains, or suffer from noisy data, and most of all are typically hard to integrate into neural language pipelines. To fill this gap, we release VisualSem: a high-quality knowledge graph (KG) which includes nodes with multilingual glosses, multiple illustrative images, and visually relevant relations. We also release a neural multi-modal retrieval model that can use images or sentences as inputs and retrieves entities in the KG. This multi-modal retrieval model can be integrated into any (neural network) model pipeline. We encourage the research community to use VisualSem for data augmentation and/or as a source of grounding, among other possible uses. VisualSem as well as the multi-modal retrieval models are publicly available and can be downloaded in this URL: https://github.com/iacercalixto/visualsem.

1 Introduction

Publicly-available multilingual resources such as Wikipedia are crucial when used as knowledge bases in recent neural language models (LMs) that retrieve documents in order to solve complex tasks such as open-ended question answering (Guu et al., 2020; Lewis et al., 2020b,a). However, Wikipedia's wealth of visual information is typically hard to use,[1] which prevents models from including rich *multi-modal data*. Although a small number of structured knowledge graphs (KGs) with images exist and are publicly available, they either cover limited domains (Xie et al., 2017; Mousselly Sergieh et al., 2018) or span multiple domains but with noisy images (Navigli and Ponzetto, 2012). To facilitate further progress in this direction, we introduce a new resource to enable research on LMs that efficiently retrieve *textual* and *visual contextual information* from KGs, where this information can be used for grounding, data augmentation, among other uses.

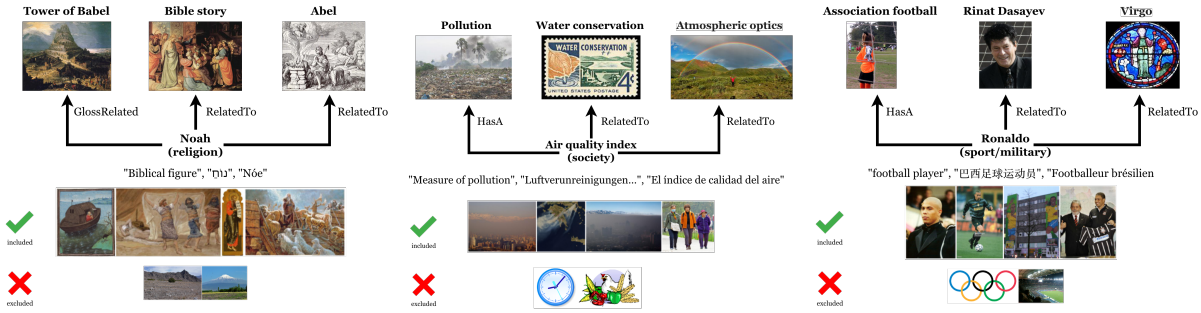
In this paper, we introduce VisualSem, a multilingual and multimodal KG with $\sim$ 90k nodes, $\sim$ 1.3M glosses and $\sim$ 938k images. VisualSem's nodes denote concepts and named entities, include multiple curated illustrative images, as well as glosses in up to 14 diverse languages. Typed semantic relations connect nodes in the graph, and nodes in VisualSem are linked to Wikipedia articles, WordNet synsets, and (when available) high-quality images from ImageNet (Deng et al., 2009). VisualSem integrates seamlessly with existing resources. Compared to existing multimodal KGs, VisualSem includes data from different sources and thus it is also more diverse in terms of the domains. We source the images in VisualSem using BabelNet (Navigli and Ponzetto, 2012), a large multilingual and multimodal resource that semi-automatically aggregates information from many different sources. We address the known issue of noisy images in BabelNet (Colla et al., 2018; Calabrese et al., 2020) by applying multiple filtering steps to ensure that we remove noise while maintaining breadth of coverage and image diversity.

We also release pre-trained models to retrieve entities from VisualSem using either images or sentences as queries in a k-nearest neighbor search. This effectively allows researchers to integrate entities and facts in VisualSem into their (neural) model pipelines. Code to generate and download VisualSem, as well as retrieval models, is publicly available.[2]

[*] Work initiated while in the University of Amsterdam during her MSc. research.

[1]See for instance https://en.wikipedia.org/wiki/Wikipedia:Images and https://commons.wikimedia.org/wiki/Main_Page.

[2]https://github.com/iacercalixto/



(a) Node "Noah". (b) Node "Air quality index". (c) Node "Ronaldo".

Figure 1: Example nodes in VisualSem, some of their glosses and images, and how they relate to other nodes. We also show examples of images we collected for the nodes that were filtered out and that were kept in following our data collection pipeline (Section 2.1).

Our main contributions are:

- We introduce VisualSem, a multi-modal knowledge graph designed to be used in vision and language research that integrates textual descriptions in up to 14 languages and images from curated sources.
- We build an image filtering pipeline to obtain a clean set of images associated to each concept in VisualSem.
- We provide an open source code base one can use to download and generate the KG, as well as multi-modal retrieval models that retrieve entities from the KG given images and sentences.

This paper is organised as follows. In Section 2 we explain how we build VisualSem and provide details on its data collection pipeline. In Section 3 we include dataset statistics and also analyse VisualSem's content and structure qualitatively. In Section 4, we describe the multi-modal retrieval models that use sentences and images to retrieve entities from the KG. In Section 5 we discuss relevant related work, including existing multi-modal knowledge bases and how they compare to our work. Finally, in Section 6 we discuss our main findings as well as provide avenues for future work.

2 Approach

VisualSem is a multilingual and multimodal knowledge graph consisting of $89,896$ unique nodes and $1,481,007$ facts, where each fact is a tuple <n_i, r_j, n_k> that denotes that node n_i is connected to node n_k via relation r_j. Each node n denotes a *concept*, such as a WordNet synset or Wikipedia article (e.g., see example nodes in Figures 1a, 1b, and 1c). Relations r_j can take one of 13 different semantic types. We select VisualSem's nodes and relations carefully: we include nodes that denote concepts and named entities with a strong visual component, and relations that encode relevant visual knowledge and/or knowledge relevant to solving tasks that require visual understanding (discussed next). We extract nodes, relations, and images in VisualSem using BabelNet v4.0 (Navigli and Ponzetto, 2012), a large multilingual KB that consolidates data from multiple knowledge graphs into one single graph.[3]

Relation types We follow previous work to choose relation types in VisualSem. Cui et al. (2018) propose to use 15 relation types with a strong visual component and that therefore are likely to help in vision and language tasks: *is-a*, *has-part*, *related-to*, *used-for*, *used-by*, *subject-of*, *depicts*, *receives-action*, *made-of*, *has-property*, *also-see*, *gloss-related*, *synonym*, *part-of*, and *located-at*. We use 13 out of the 15 proposed relation types, since we do not have examples with *depicts* and *also-see* in the nodes we select. We describe the data collection procedure next.

2.1 Data collection

We start by choosing a set of *seed* nodes we can guarantee is high-quality, well-curated, and visually relevant. We therefore use the BabelNet API[4] to get synsets corresponding to the $1,000$ ImageNet classes used in the ILSVRC image classification

visualsem.

[3] https://babelnet.org
[4] https://babelnet.org/guide

competition (Russakovsky et al., 2015) as our initial seed nodes.[5] We call these nodes our initial *node pool*. We choose this initial node pool since ImageNet already provides high-quality images associated to these nodes. We then follow an iterative procedure where in each step we add nodes to the node pool, until we reach the end of the algorithm. Specifically, in the first step, the node pool includes the $1,000$ seed nodes; in the second step, it will include the $1,000$ seed nodes plus the additional nodes gathered in the first step; and so on, until we reach N nodes ($N = 90,000$). This procedure works by iterating over the following steps:

1. **Retrieve neighbors**: retrieve neighbor nodes for each node in the node pool using the BabelNet API;
2. **Validate images**: validate linked images and optionally remove those that do not meet certain quality criteria;
3. **Filter nodes**: filter out nodes that do not meet inclusion criteria;
4. **Update pool**: accept top-k nodes among remaining nodes after sorting nodes according to their features.

Retrieve neighbors In this step we collect all first-degree neighbors for each node in the node pool, and remove any duplicate node retrieved if any exists. The neighbors are retrieved using the BabelNet v4.0 API (Navigli and Ponzetto, 2012) and can include nodes from multiple sources, such as multilingual Wikipedias, multilingual WordNets, among others. First-degree neighbors n_k of node n_i are those nodes directly connected via a relation which type r_j is one of the 13 relation types, i.e., $<n_i, r_j, n_k = ?>$. In Table 1 we show a list of relation types in BabelNet, and how these were merged into the 13 types used in VisualSem.

Validate images + Node filtering + Update pool We collect images available for all first-degree neighbors retrieved in the previous step, and apply *four filters* to the available images: **i)** we check if images are valid files, **ii)** we remove duplicate or near-duplicate images, **iii)** we train a binary classifier to remove non-photographic images, and **iv)** we use OpenAI's CLIP (Radford et al., 2021) to remove images that do not minimally match *any* of the node's glosses. At the end of each iteration, we require that each node has at least one image associated to it, and that it contains relations with other nodes with a minimum of two different relation types. If nodes do not satisfy these criteria, they are filtered out.

BabelNet	VisualSem
is-a, is_a	is-a
has-part, has_part	has-part
related	related-to
use	used-for
used-by, used_by	used-by
subject-of, subject_of	subject-of
interaction	receives-action
oath-made-by	made-of
has_*	has-property
gloss-related	gloss-related
taxon-synonym	synonym
part-of, part_of	part-of
location, located_*	located-at

Table 1: Relation types in BabelNet and their corresponding types in VisualSem. Asterisks (*) can match any number of characters.

In more detail: **i)** From the total amount of image files we downloaded, an average of $\sim 6.3\%$ were invalid image files and thus removed, e.g., the downloaded file was in fact an audio file. **ii)** We find many images that are (near-)duplicates across different synsets, in which case we remove those duplicates using SHA1 hashing (Eastlake and Jones, 2001). **iii)** We use a binary classifier to validate images associated to each node in the pool of retrieved neighbors following the simple procedure described in Alberts and Calixto (2020). We use their ResNet-based coarse-grained binary classification model to further filter out undesirable images. In short, the image classification model is trained on a total of $6,000$ randomly sampled images (mostly sourced from Wikipedia and ImageNet) where 50% are good quality/photographic images, and the remaining are non-photographic/undesirable ones. We denote undesirable images as unclear/blurred/dark or low-quality images, and non-photographic images such as hand drawings, sketches, maps, icons, flags, graphs and other rendered images. Please refer to Alberts and Calixto (2020) for more details on the image classification model training and evaluation. We apply this step and remove images flagged with this binary classifier, since many images we obtain in practice are very noisy. **iv)** Finally, in our last filter in the image filtering process we use the

[5] https://image-net.org/challenges/LSVRC/2014/browse-synsets

	# langs.	# nodes	# rel. types	# glosses	# images	# train	# valid	# test	sources
WN9-IMG[†]	1	6,555	9	N/A	65,550	11,741	1,337	1,319	WordNet
FB15-IMG[‡]	1	11,757	1,231	N/A	107,570	285,850	29,580	34,863	Freebase
VisualSem	14	89,896	13	1,342,764	938,100	1,441,007	20,000	20,000	Multiple[⋆]

Table 2: VisualSem KG statistics. [⋆]Multiple sources include Wikipedia, WordNet, ImageNet, among others. [†]Xie et al. (2017). [‡]Mousselly Sergieh et al. (2018).

pre-trained CLIP model. CLIP has one text and one image encoder and is trained to maximize the dot-product of correct image-sentence pairs, and at the same time minimize that of random image-sentence pairs. We encode the k English glosses $g_{i,k}$ and the l images $a_{i,l}$ available for node n_i with CLIP's text and image encoder, respectively. We then compute the dot-product between each image $a_{i,l}$ and each gloss $g_{i,k}$, keeping only the images that had at least one dot-product greater than 0.5 with one of the English glosses. We choose the value of 0.5 empirically by manually checking the quality of the image-gloss matches. We note that the motivation for keeping only images that match well with at least one gloss is the fact that the available glosses are by design descriptive of the node, and by making sure images align well with at least one of the available glosses we can make sure we filter out noisy and unrelated images (e.g., please refer to examples in Appendix A for more details).

Step (i) reduces the number of images from ~ 5.6 to ~ 5.3 million, step (ii) brings the number of images from ~ 5.3 to ~ 2.1 million, step (iii) further reduces images from ~ 2.1 to ~ 1.5 million, and finally step (iv) brings us to the final number of images in VisualSem, $938,100$.

After filtering out undesirable images in the retrieved neighbors, we discard any nodes in the neighborhood that do not have at least one associated image. Finally, for each node in the initial node pool we add the top-k neighbour nodes ($k = 10$) to the pool and repeat. We prioritize visually relevant nodes with a larger number of images, nodes that include as many diverse relation types as possible, and especially nodes with less frequent relation types.

3 Data Statistics and Analysis

In this Section, we explore the structure and content of VisualSem. In Section 3.1, we report relevant statistics, as well as show some examples of nodes and relations in it. In Section 3.2, we provide a more qualitative analysis using unsupervised topic models to induce topic distributions for VisualSem.

3.1 Data Statistics

In Table 2, we show statistics comparing VisualSem and multimodal knowledge bases WN9-IMG and FB15-IMG.

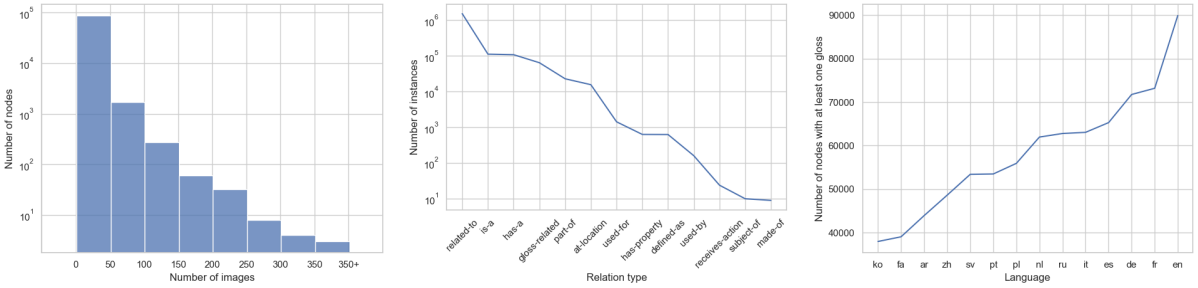
Glosses. VisualSem has a total of $1,342,764$ glosses in 14 different languages: Arabic, Chinese, Dutch, English, Farsi, French, German, Italian, Korean, Polish, Portuguese, Russian, Spanish, and Swedish. The languages were chosen to be representative of diverse linguistic families and to include many different scripts, while at the same time covering a high number of nodes. Nodes have on average 14.9 glosses across all 14 languages, and in Figure 2c we show the number of nodes that have at least one gloss in each language. Nodes with English (Korean) glosses have the highest (lowest) coverage: $89,896$ nodes have at least one English and $37,970$ at least one Korean gloss. **Example:** Node *Ronaldo (Brazilian footballer)* has four glosses in English, and one of these glosses is *Ronaldo Luís Nazário de Lima, commonly known as Ronaldo, is a retired Brazilian professional footballer who played as a striker*[6] (see Figure 1c).

Images. As shown in Table 2, VisualSem has a total of $938,100$ unique images. On average, there are 10.4 images per node—similarly to WN9-IMG and FB15-IMG datasets—and in Figure 2a we show the distribution of images per node. The node with the greatest number of images is the concept *Russian culture* with 848 images.[7]

Relations. VisualSem has 13 different relation types, and all its relations are typed. Relation types include: *is-a*, *has-part*, *related-to*, *used-for*, *used-by*, *subject-of*, *receives-action*, *made-of*,

[6]The same node also has glosses in French (3), German (2), Spanish (4), Italian (4), Russian (1), Dutch (3), Polish (1), Portuguese (3), Swedish (3), Mandarin (1), Arabic (1), Farsi (2), and Korean (1).

[7]It has BabelNet id `bn:01286889n` and is linked to Wikipedia article `https://en.wikipedia.org/wiki/Russian_culture`.



(a) Number of images per node. (b) Number of instances or facts (i.e., tuples $<n_i, r_j, n_k>$) per relation type. (c) Number of nodes with at least one gloss in each language.

Figure 2: VisualSem data statistics.

has-property, *gloss-related*, *synonym*, *part-of*, and *located-at*. In Table 1 we show the mapping from relations in BabelNet to the corresponding types in VisualSem. **Example:** In Figure 1 we show three example nodes and how they relate to other nodes in the KG. We note that most connections have the *related-to* relation type. In Figure 2b, we show the number of instances (i.e., tuples $<n_i, r_j, n_k>$) per relation type. There is a very large imbalance, with the type *related-to* accounting for about 82% of all existing relations in the KB. We note that had we not adopted measures to alleviate this issue in our data collection pipeline—e.g., prioritizing nodes that have less frequent relation types when adding to the node pool—this imbalance would have been much greater (see Section 2.1).

3.2 Topic Models

To further analyze the quality of our KG, we try to understand what topics are salient in VisualSem. We train a neural topic model with 20 latent topics using Embedded Topic Model (ETM; Dieng et al., 2020), which is an extension of Latent Dirichlet Allocation (LDA; Blei et al., 2003) that use pre-trained word embeddings. We let each node be a document, and each document's content be the combination of all its English glosses. In Table 3, we show the six most representative words per topic after stop-word removal. The topics covered are varied and can be clustered in broad domains such as 'sciences' (e.g., physics, chemistry, biology, space), 'society' (e.g., geography, politics, occupations), 'culture' (e.g., religion, history, food, fashion), among others (e.g., material, city). There is a trend towards representing factual knowledge, which is expected since glosses by definition describe facts about nodes. Concepts well covered in Wikipedia are also well covered in VisualSem.

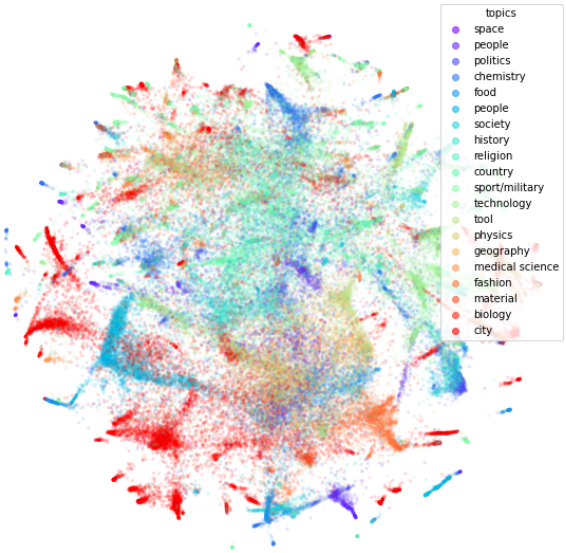


Figure 3: T-SNE plot for node embeddings where each node is represented as the average of its gloss embeddings. Topic assignments are used to colorise node embeddings, and topics are computed with the topic model described in Section 3.2 and Table 3.

3.3 Node visualisation via its glosses

We now visualise nodes with the t-SNE algorithm (van der Maaten and Hinton, 2008). We embed glosses as the average word embedding for each word in the gloss computed with multilingual fastText (Bojanowski et al., 2016). Similarly, we set each node n_i's representation $\boldsymbol{n}_i$ as the average of all its gloss embeddings, across all languages. We use the ETM topic model trained in Section 3.2 to assign a topic for each node in VisualSem (i.e., the highest probability topic induced in the topic model), and colorise the node according to its predicted topic. We apply t-SNE to the node

1. space	2. occupation	3. politics	4. chemistry	5. food	6. occupation	7. society	8. history	9. religion	10. country
planet	dance	party	gas	meat	actor	school	rome	religion	commune
constellation	physicist	rank	formula	bread	composer	institution	emperor	directed	autonomous
boat	painting	officer	atomic	cheese	band	economic	ireland	jesus	saxony
spacecraft	mathematician	politician	acid	sauce	painter	education	pope	jewish	philippine
moon	philosopher	minister	iron	vegetable	singer	agency	dynasty	bible	indonesia
mar	scientist	currency	solid	rice	musician	society	egypt	goddess	finland
11. sports/military	**12. technology**	**13. mixed**	**14. physics**	**15. geography**	**16. medicine**	**17. material**	**18. fashion**	**19. biology**	**20. city**
football	electrical	horse	image	bridge	blood	garment	hair	cat	museum
tank	data	ice	measure	switzerland	bone	clothing	fabric	shark	street
rifle	storage	wall	motion	wine	tissue	dress	soil	temperate	metro
team	electronic	blade	energy	valley	muscle	skirt	colour	subfamily	stockholm
carrier	signal	tool	wave	canton	organism	flag	cloth	beetle	tokyo
stadium	card	stone	radiation	archipelago	organ	garden	hat	grass	korea

Table 3: Topics induced using the Embedded Topic Model on VisualSem English glosses (labels in bold are assigned manually).

embeddings $\boldsymbol{n}_i$ and show the plot in Figure 3.[8] We highlight that nodes cluster well according to their predicted topic, and note that even though there clearly is some structure in the space induced by the KG glosses, finding hard boundaries across topics does not seem straightforward (at least for the majority of nodes closer to the center of the plot).

4 Entity Retrieval

In this section, we describe *retrieval models* that retrieve entities in the KB given a sentence or an image. Our use-case includes tasks with possibly multi-modal inputs—i.e., either a sentence, an image, or both—where we use these inputs to retrieve relevant entities from VisualSem. The idea is to ground the task inputs or to augment the available data to train the task model. This task could be text-only, such as named entity recognition or machine translation, or multi-modal, such as visual question answering or image captioning. We do not include experiments on specific tasks but instead leave that investigation for future work.

We use a standard *retrieval as ranking* framework where the model is trained to rank nodes in VisualSem given an input, and we use the top-k ranked nodes as the results. These retrieval modules allow for an easy integration of VisualSem's entities, glosses, and images into neural pipelines.

We frame entity retrieval using sentences (henceforth *sentence retrieval*, Section 4.1) as a sentence-to-gloss ranking problem, and use glosses available in all languages in the model. Given an input sentence, we transform the resulting ranking over glosses into a ranking over nodes (i.e., by retrieving the nodes these glosses are linked to from their 1-to-1 mappings). We similarly frame retrieval using images (henceforth *image retrieval*, Section 4.2) as an image-to-gloss ranking problem, and use all English glosses available for retrieval.

[8] We use the following hyperparameters with t-SNE: perplexity is set to 70, and number of iterations to 5000.

	# train	# valid	# test
Glosses	1,286,764	28,000	28,000
Images	898,100	20,000	20,000

Table 4: VisualSem gloss and image splits. Gloss validation and test splits include 2,000 entries for each of the 14 languages in VisualSem.

Experiments in this section use the gloss and image splits in Table 4. Training, validation and test splits for glosses and images are chosen randomly in order to facilitate their use by the research community in future experiments. Gloss validation and test splits are balanced regarding different languages, and each split includes 2,000 examples for each of the 14 languages in VisualSem.

4.1 Sentence retrieval

VisualSem has $N = 89,896$ nodes with glosses in up to 14 languages. We encode glosses using Sentence BERT (SBERT; Reimers and Gurevych, 2019), which is a family of models based on bi-encoder architectures trained on the task of sentence similarity that have strong performance when used for retrieval/indexing. SBERT models are by default trained on English-only sentences to map semantically similar sentences to points close in the embedding space. We use a multilingual SBERT model trained using knowledge distillation to transfer knowledge from English to other languages, more specifically the `paraphrase-multilingual-mpnet-base-v2` model (Reimers and Gurevych, 2020) trained on over 50 languages. We select the model from all publicly available models according to validation

	Hits@k			Rank
	1 ↑	3 ↑	10 ↑	mean (std) ↓
ar	37.7	48.9	58.6	2,572 (18,369)
de	48.9	58.3	66.7	1,590 (13,801)
en	56.1	64.0	73.0	15,156 (133,161)
es	60.5	69.3	76.4	693 (6,234)
fr	53.4	62.3	70.1	1,967 (20,850)
it	57.7	66.1	73.2	1,248 (16,216)
ko	44.7	56.8	66.8	1,488 (19,586)
nl	46.2	54.8	62.6	3,110 (31,413)
pt	**73.1**	**79.5**	**83.8**	1,646 (34,586)
ru	28.9	36.4	42.8	16,043 (55,115)
zh	62.9	73.3	81.1	1,691 (26,218)
fa	38.6	49.8	60.1	1,829 (9,089)
pl	49.2	58.0	66.8	3,803 (25,605)
sv	61.7	71.4	78.7	**656 (6,865)**
avg.	51.4	60.6	68.6	3,821 (43,430)

Table 5: Sentence retrieval results on VisualSem test glosses. We report Hits@k, which is the percentage of time the correct node is retrieved in the top-k results (higher is better) and the mean rank of the correct node (lower is better). According to Hits@k, Portuguese and Chinese have the best retrieval results, whereas worst results are obtained for Russian, Arabic, and Farsi. Best mean ranks are obtained with Spanish and Slovenian queries, and worst with Russian and English.

performance.[9]

Let $g_{i,j}$ be the j-th gloss in the set of glosses associated to node i, and $\boldsymbol{g}_{i,j}$ be gloss $g_{i,j}$'s 512-dimensional vector representation computed with SBERT. We implement the sentence retrieval using k-nearest neighbour (k-NN). We directly rank all glosses in the split given the query according to their cosine similarity, and use the node associated to the gloss with the highest cosine similarity as the retrieved node.

Evaluation and Discussion We use the SBERT model `paraphrase-multilingual-mpnet-base-v2` (Reimers and Gurevych, 2019, 2020), trained on over 50 languages using knowledge distillation on the task of sentence similarity. All languages covered in VisualSem are supported by the model. We directly use the $28,000$ held-out test glosses (2k per language) for model evaluation and show the results in Table 5. We note that retrieval results are in general very good according to Hits@k. Portuguese and Chinese queries have the best sentence retrieval performance according to this metric, whereas querying the model using Russian, Arabic, and Farsi has the worse performance. We note that mean ranks show high variances, which indicate that retrieved nodes could be noisy despite the good Hits@k scores. Surprisingly, mean ranks obtained with English and Russian queries are the highest (i.e., the lower the mean ranks, the better). We recommend that users use the sentence retrieval model with care when dealing with any of the low-quality languages.

[9]Models available in https://www.sbert.net/docs/pretrained_models.html#multi-lingual-models.

	Rank	Hits@k		
	mean (std) ↓	1 ↑	3 ↑	10 ↑
k-NN	$4,117$ $(16,705)$	10.0	16.5	25.6

Table 6: Image retrieval results on test images. We report Hits@k, which is the percentage of the time the correct node is retrieved in the top-k results (higher is better) and the mean rank of the correct node (lower is better).

4.2 Image retrieval

CLIP (Radford et al., 2021) is a discriminative model with a bi-encoder architecture trained on 400 million image-text pairs. It has one text and one image encoder and is trained to maximize the dot-product of correct image-sentence pairs, and at the same time to minimize the dot-product of random pairs. We use the pre-trained CLIP model `RN50x16` for image retrieval, which is the best released CLIP model at the time.[10]

We encode each image v_k in the validation and test sets ($k = 20,000$) using CLIP image encoder, and denote each image v_k's 768-dimensional vector representation generated with the image encoder by $\boldsymbol{v}_k$. We similarly encode all *training* English glosses g_j for all nodes in VisualSem using CLIP text encoder, which similarly computes a 768-dimensional vector representation $\boldsymbol{g}_j$ for each English gloss.

Evaluation and Discussion We use the encoding of each of the 20k images in the validation and test sets, and frame node retrieval using images as queries as an image-to-gloss ranking problem. We rank *training* English glosses in VisualSem ac-

[10]https://github.com/openai/CLIP/blob/main/clip/clip.py

cording to their cosine similarity with the input image, and use the node associated to the gloss with the highest cosine similarity as the retrieved node. When there are more than one English gloss associated to a node, we use the best ranking across all glosses as the ranking of the retrieved node. We report results in Table 6, and note that the quality of the image retrieval module is worse compared to the sentence retrieval. One of our plans for future work is to investigate how to improve VisualSem image retrieval module.

5 Related work

Knowledge bases (KBs) and KGs have a long and rich history, and many publicly available KBs exist and have been built for different purposes.[11] A seminal example is Cyc (Lenat et al., 1986), an early effort towards building a general-purpose knowledge base to store common sense facts and rules about how the world works. More recent examples of knowledge bases built with different purposes include WordNet (Miller, 1995; Bond and Paik, 2012), DBPedia (Auer et al., 2007), Wikidata (Vrandečić and Krötzsch, 2014), Freebase (Bollacker et al., 2008), YAGO (Rebele et al., 2016; Pellissier Tanon et al., 2020), ConceptNet (Speer and Havasi, 2012), ATOMIC (Sap et al., 2019), among many others (see Wang et al., 2017; Ji et al., 2020 for a detailed list of knowledge bases and algorithms). Our main goal is to design a KG to support research in vision & language, which none of the abovementioned KBs are designed to do.

Multi-modal knowledge bases. Two recently proposed multimodal knowledge bases are WN9-IMG (Xie et al., 2017) and FB15-IMG (Mousselly Sergieh et al., 2018): the former consists of a subset of entities and relations from WordNet and was built using the WN18 dataset proposed in Bordes et al. (2014), and additionally includes 10 images illustrating each entity; the latter is based on the FB15 dataset introduced by Bordes et al. (2013), which in turn consists of examples extracted from Freebase, and also includes 10 images per entity. Although these KBs include images, they are still restricted to a single data source and constrained in terms of the domains they encompass.

An exception is BabelNet (Navigli and Ponzetto, 2012), which is a very large KG that combines varied sources—such as Freebase, WordNet and Wikipedia in several languages, among many others—and includes over 54 million images mined mostly from Wikipedia and ImageNet. BabelNet can be seen as a high coverage KB, and in its version v4.0 has more than 15.7 million concepts in 284 languages. At times, images linked in BabelNet may have poor quality (Calabrese et al., 2020), e.g., blurred photos, images loosely related to the concept they illustrate, uninformative images such as icons, flags, or rendered graphs (see Figure 1 for examples). We build an *image filtering* pipeline that filters out noisy images linked in BabelNet, and we source the remaining high-quality images in VisualSem.

All the aforementioned KBs have in common the fact they are mostly *"entity-centric"*: nodes denote *concepts* and are associated with multimodal information. This is in contrast to "vision-centric" KBs such as Visual Genome (Krishna et al., 2017) and Visual KB (Zhu et al., 2015), where each instance in the dataset is an image with a *scene graph* representation to support visual query tasks. Besides visual queries, multimodal KBs have been designed for conversational reasoning (Moon et al., 2019) and node classification (Bloem et al., 2021). VisualSem is an entity-centric KG designed to support tasks at the intersection of vision and language.

6 Conclusions and Future work

We present the VisualSem knowledge graph with $\sim$ 90k nodes, over 1.3M glosses in up to 14 diverse languages, and around 938k curated and cleaned images illustrating nodes' concepts. VisualSem bridges a gap in the available resources to train grounded models of language, and is designed to be useful in vision and language research. We release neural entity retrieval models that accept text and image inputs. Sentences in any of the $\sim$ 50 languages supported by multilingual Sentence BERT can be used to retrieve entities from the KG, and we evaluate retrieval using all 14 languages in VisualSem in Table 5. Images can also be used to retrieve nodes using the state-of-the-art CLIP model, and visual entity retrieval is evaluated in Table 6. This allows researchers to easily integrate VisualSem into their (neural) model pipelines, and we encourage its use not just in tasks that involve vision and language, but across all sorts of language understanding and/or generation tasks, in *grounding* and/or in *data augmentation* settings.

[11] We do not differentiate between knowledge bases and knowledge graphs for the purposes of this work and use both terms interchangeably.

Future Work We will use VisualSem sentence/image retrieval mechanisms and gloss and image features for data augmentation in NLP tasks, e.g. word sense disambiguation and named entity recognition, and on vision and language tasks, e.g. image captioning and visual question answering. The reason for these tasks is that they all could intuitively benefit from the added knowledge, be it visual, textual, or multi-modal. We will also investigate how to improve the quality of the image and sentence retrieval modules, both central to the KG impact on current neural-based models. Finally, we plan to improve and grow the KG, which is under active development. We will particularly gauge whether there is interest from the research community in increasing its coverage to include other parts of, or the entirety of, Wikipedia, for example.

Acknowledgements

IC has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 838188. KC is partly supported by Samsung Advanced Institute of Technology (Next Generation Deep Learning: from pattern recognition to AI) and Samsung Research (Improving Deep Learning using Latent Structure). KC also thanks Naver, eBay, NVIDIA, and NSF Award 1922658 for support. CV's work on this project at New York University was financially supported by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program) and Samsung Research (under the project *Improving Deep Learning using Latent Structure*) and benefitted from in-kind support by the NYU High-Performance Computing Center. This material is based upon work supported by the National Science Foundation under Grant No. 1922658. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Houda Alberts and Iacer Calixto. 2020. Imagifilter: A resource to enable the semi-automatic mining of images at scale. *arXiv preprint*.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Peter Bloem, Xander Wilcke, Lucas van Berkel, and Victor de Boer. 2021. kgbench: A collection of knowledge graph datasets for evaluating relational and multimodal machine learning. In *Eighteenth Extended Semantic Web Conference - Resources Track*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue. 64–71.

Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2014. A semantic matching energy function for learning with multi-relational data. *Mach. Learn.*, 94(2):233–259.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2787–2795, Red Hook, NY, USA. Curran Associates Inc.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020. Fatality killed the cat or: BabelPic, a multimodal dataset for non-concrete concepts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4680–4686, Online. Association for Computational Linguistics.

Davide Colla, Enrico Mensa, Daniele P. Radicioni, and Antonio Lieto. 2018. Tell me why: Computational explanation of conceptual similarity judgments. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*, pages 74–85, Cham. Springer International Publishing.

P. Cui, S. Liu, and W. Zhu. 2018. General knowledge embedded image representation learning. *IEEE Transactions on Multimedia*, 20(1):198–207.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

D. Eastlake and P. Jones. 2001. Rfc3174: Us secure hash algorithm 1 (sha1).

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. 2020. A survey on knowledge graphs: Representation, acquisition and applications. *arXiv preprint arXiv:2002.00388*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, and et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Doug Lenat, Mayank Prakash, and Mary Shepherd. 1986. Cyc: Using common sense knowledge to overcome brittleness and knowledge acquistion bottlenecks. *AI Mag.*, 6(4):65–85.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. In *Advances in Neural Information Processing Systems*, volume 33, pages 18470–18481. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM 2018)*, page to appear. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Thomas Pellissier Tanon, Gerhard Weikum, and Fabian Suchanek. 2020. Yago 4: A reason-able knowledge base. In *The Semantic Web*, pages 583–596, Cham. Springer International Publishing.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Thomas Rebele, Fabian M. Suchanek, Johannes Hoffart, Joanna Asia Biega, Erdal Kuzey, and Gerhard Weikum. 2016. Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *International Semantic Web Conference*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 3027–3035. AAAI Press.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3679–3686, Istanbul, Turkey. European Languages Resources Association (ELRA).

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Q. Wang, Z. Mao, B. Wang, and L. Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 3140–3146. AAAI Press.

Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. 2015. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*.

A Examples

We now show some example data in VisualSem to illustrate what kind of information is available in a node, including metadata, as well as what images are kept and what are filtered out. We select five topics (i.e., country, people, physics, biology, space) in the topic model discussed in Section 3.2 and show one example node in each, including: **multilingual glosses** in up 14 languages (we show a maximum of one gloss per language per node to avoid clutter); **images** linked to the node, including the ones filtered out by our procedure illustrated in Section 2; **relations** to other nodes in the KG (we show a maximum of three connected nodes to avoid clutter). We illustrate the five node examples in Figures 4–8, and we use these examples to support a discussion on the quality of the data.

First, we note that information described in a node's glosses can be repetitive across languages (e.g., in Figure 6 the glosses in English and Portuguese are translations of one another), but often can also be complementary (e.g., again in Figure 6 the gloss in French includes novel information such as the mention to leguminous crops, which does not appear neither in English nor Portuguese).

We also note that the removed images are most of the time non-photographic and/or noisy, in accord with our motivation. Images such as the map or the sketch in Figure 4, or the excluded images in Figures 7 or 8, are not *descriptive* of the node. However, a limited number of relevant images are sometimes excluded by our procedure (e.g., the field with animals in Figure 6), and conversely sometimes images that are tangentially related to a node are kept (e.g., the magnet in Figure 8).

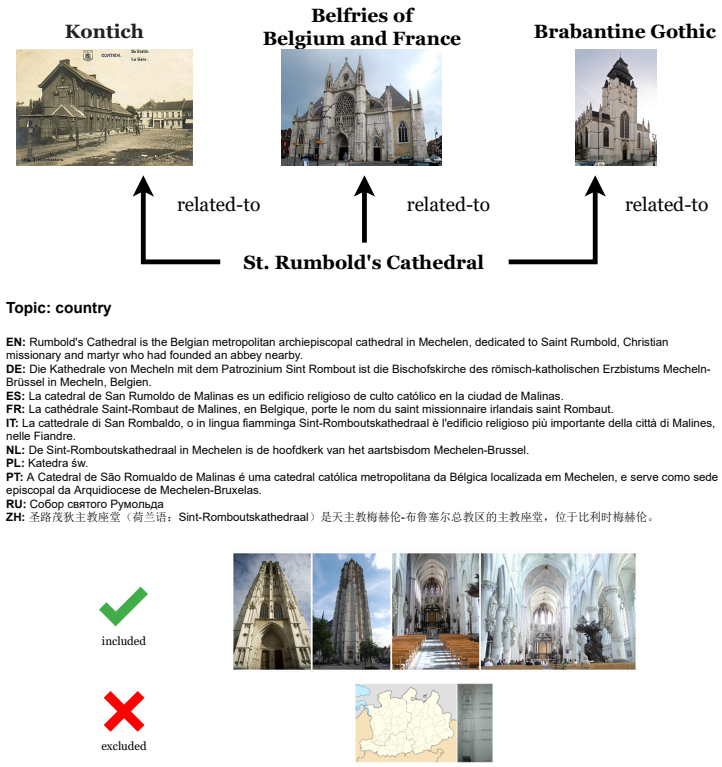


Figure 4: *St. Rumbold's Cathedral*. The node is connected to four other nodes (3 shown) with relation types `related-to` and `has-a` (not shown), and has a total of 23 images kept (4 shown).

Figure 5: *David Gilmour*. The node is connected to 10 other nodes (3 shown) with 2 relation types `related-to` and `has-a` (not shown), and has a total of 10 images kept (4 shown).

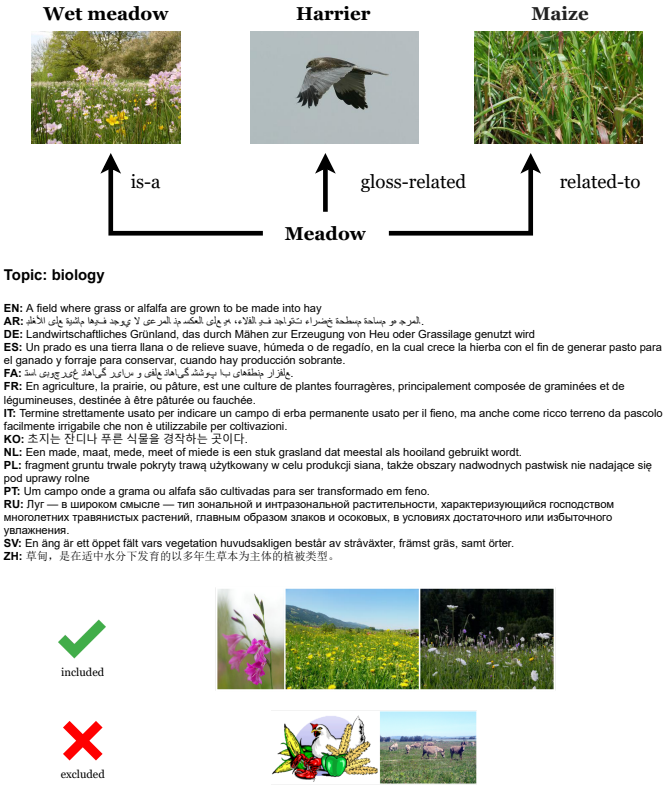


Figure 6: *Meadow*. The node is connected to 127 other nodes (3 shown) with relation types `related-to`, `gloss-related`, `is-a`, `has-a` (not shown), and has a total of 26 images kept (3 shown).

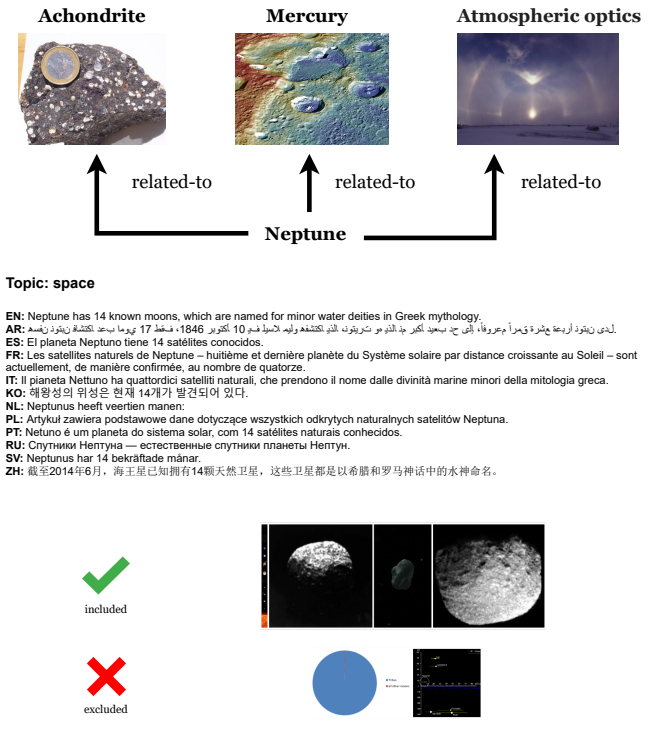
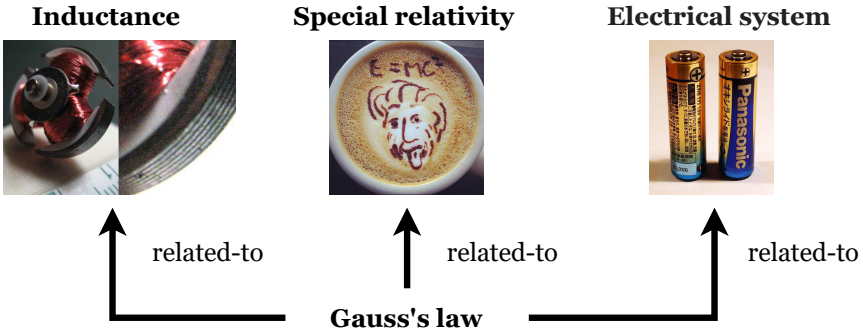


Figure 7: *Neptune*. The node is connected to 10 other nodes (3 shown) with relation types `related-to`, `is-a` (not shown), `has-a` (not shown), and has a total of 26 images kept (4 shown).



Topic: physics

EN: In physics, Gauss's law, also known as Gauss's flux theorem, is a law relating the distribution of electric charge to the resulting electric field.
AR: فـي الفيزياء، قانون جاوس الكهربي، المعروفة أيضاً باسم مبرهنة جاوس فـي التدفق الكهربي، والقانون يصف العلاقة بين توزيع الشحنة الكهربائية والمجال الكهربائي الناتج عنها.
DE: Das gaußsche Gesetz beschreibt in der Elektrostatik und Elektrodynamik den elektrischen Fluss durch eine geschlossene Fläche.
ES: En física la ley de Gauss, también conocida como teorema de Gauss, establece que el flujo de ciertos campos a través de una superficie cerrada es proporcional a la magnitud de las fuentes de dicho campo que hay en el interior de la misma superficie.
FA: قانون گاوس که با نام قضیه شار گاوس هم شناخته شده، قانونی در فیزیک است که ارتباط توزیع بار الکتریکی و میدان الکتریکی حاصل از آن را بیان میکند.
FR: En électromagnétisme, le théorème de Gauss permet de calculer le flux d'un champ électrique à travers une surface fermée contenant des charges électriques.
IT: Nella teoria dei campi vettoriali il teorema del flusso, noto anche come teorema di Gauss, afferma che i campi vettoriali radiali dipendenti dal reciproco del quadrato della distanza dall'origine hanno un flusso attraverso una qualunque superficie chiusa che dipende solo dalla carica in essa contenuta ed è indipendente dalla posizione interna delle cariche che lo generano.
KO: 가우스 법칙은 폐곡면을 통과하는 전기 선속이 폐곡면 속의 알짜 전하량에 비례한다는 법칙이다.
NL: De wet van Gauss geeft in de fysica de relatie weer tussen de elektrische flux door een gesloten oppervlak en de elektrische lading binnen het oppervlak.
PL: Prawo Gaussa dla elektryczności – prawo wiążące pole elektryczne z jego źródłem, czyli ładunkiem elektrycznym.
PT: A lei de Gauss é a lei que estabelece a relação entre o fluxo do campo elétrico através de uma superfície fechada com a carga elétrica que existe dentro do volume limitado por esta superfície.
RU: Теорема Гаусса — один из основных законов электродинамики, входит в систему уравнений Максвелла.
SV: I fysiken syftar Gauss lag på någon tillämpning av den generella matematiska satsen Gauss sats som ger sambandet mellan ytintegralen av något flöde, till exempel av vätska, som flödar ut ur en sluten yta och resultatet av källor som är inneslutna i den slutna ytan.
ZH: 高斯定律（Gauss' law）表明在闭合曲面内的电荷分佈與產生的電場之間的關係：

included

excluded

Figure 8: *Gauss' Law*. The node is connected to 6 other nodes (3 shown) with relation type `related-to`, and has a total of 5 images kept (3 shown).