VisualSem

A high-quality knowledge graph for vision and language

Houda Alberts, Ningyuan (Teresa) Huang, Yash R. Deshpande, **Yibo Liu**,
Kyunghyun Cho, Clara Vania, Iacer Calixto
Rond Consulting, NL New York University ILLC, University of Amsterdam
Amazon, UK Johns Hopkins University, USA.

MRL 2021

Outline

- Introduction and Motivation
  - The VisualSem Knowledge Graph (KG)
  - Facts and Statistics
- Approach: Building the KG
- Topic model: Analysing the KG
- Using the KG
  - Sentence Retrieval
  - Image Retrieval
- Conclusions and Remarks

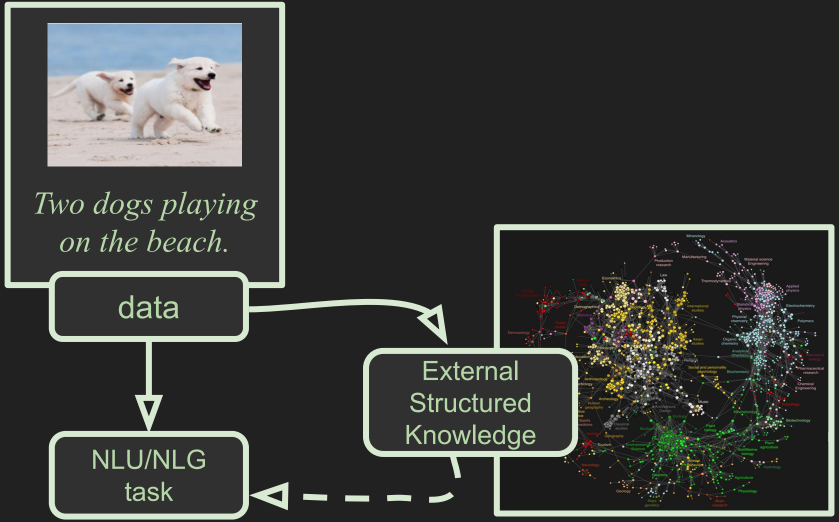
Introduction and Motivation

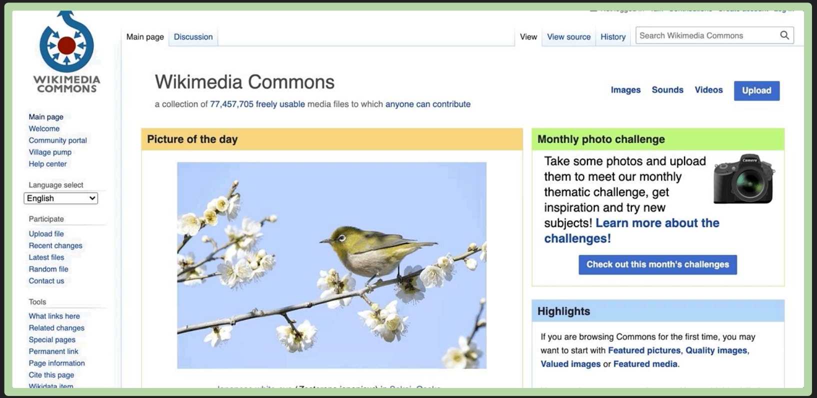
Introduction and Motivation

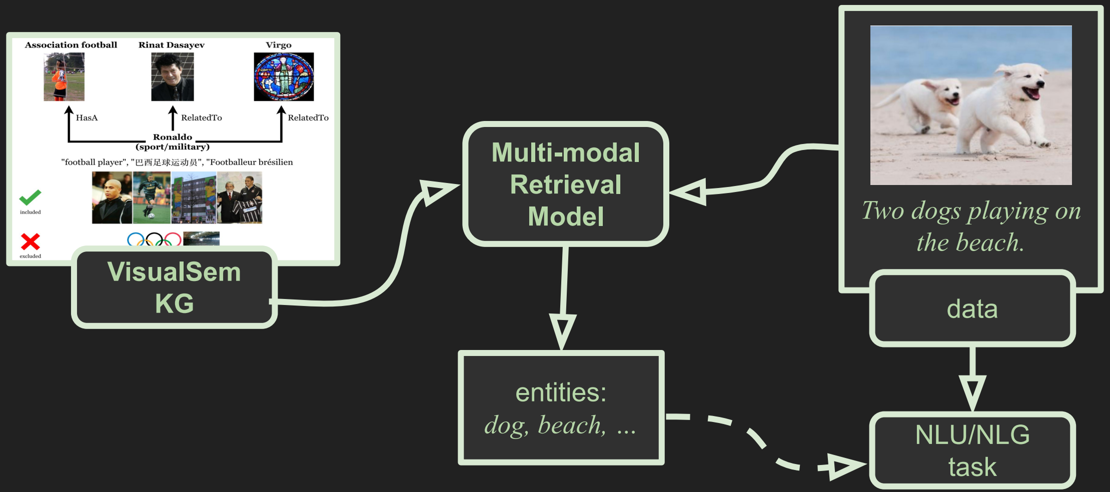
VisualSem is built to enable ***(vision-and-) language*** models that can efficiently access ***external structured knowledge*** repositories.

However, existing knowledge bases:

- cover limited domains,
- span multiple domains but are noisy,
- are typically hard to integrate into neural language pipelines!







To fill this gap, we release:

1. **VisualSem**: a high-quality knowledge graph (KG) which includes nodes with:
   - *multilingual* glosses
   - multiple illustrative images
   - *visually relevant* relations
2. A neural **multi-modal retrieval model**
   - use images or sentences as inputs to retrieve entities from the KG
   - can be integrated into (neural) model pipelines

VisualSem vs. Other multimodal KGs

- Diverse data sources

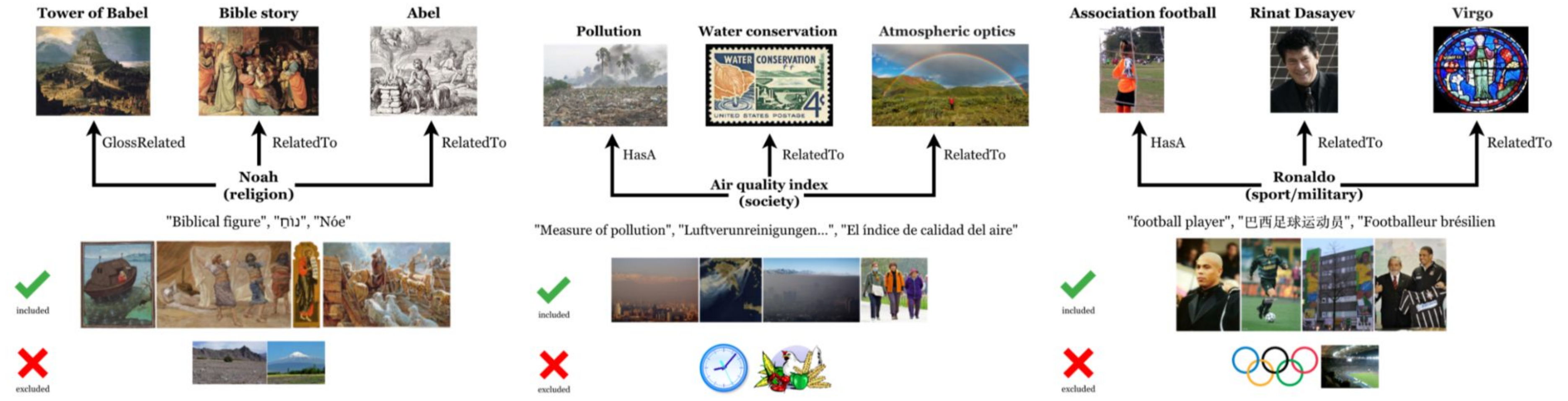
  Nodes are linked to Wikipedia articles, WordNet synsets, and (when available) high-quality images from ImageNet.

- High-quality images
  - We tackle noisy images by applying multiple filtering steps.
- Easy to integrate in neural pipelines
  - Code & retrieval models are publicly available at https://github.com/iacercalixto/visualsem/.

Our main contributions are:

1. 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 multiple curated sources.
2. We build an image filtering pipeline to obtain a clean set of images associated to each concept in VisualSem.
3. 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.

Example nodes



(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).

Facts and Statistics

	# 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).

Glosses

- 1,342,764 glosses in total
- average 14.9 glosses per node

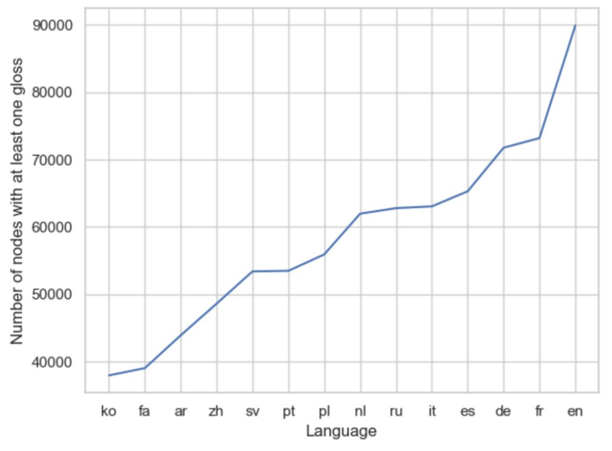
Covering 14 different languages

Arabic, Chinese, Dutch, English, Farsi, French, German, Italian, Korean, Polish, Portuguese, Russian, Spanish, and Swedish.

Why these languages?

- Representative of diverse scripts and linguistic families
- Cover a high number of nodes in the KG

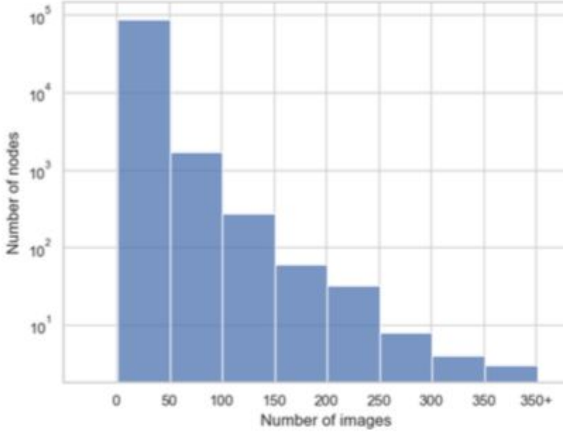
Highest (Lowest) coverage: English (Korean): 89, 896 nodes have at least one English (37, 970 at least one Korean) gloss.



(c) Number of nodes with at least one gloss in each language.

Images

- 938,100 in total
- Average 10.4 images per node, similarly to WN9-IMG and FB15-IMG



(a) Number of images per node.

Relations

- 13 relation types
- Imbalanced: `related-to` 82%, alleviated by our filtering approach

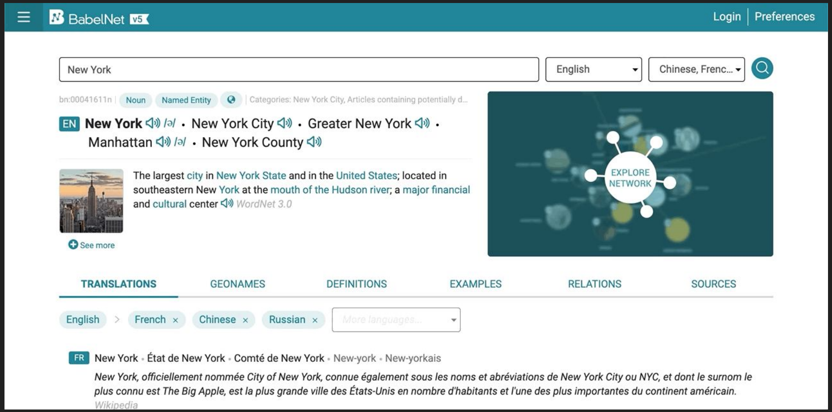
```
is-a
has-part
related-to
used-for
used-by
subject-of
receives-action
made-of
has-property
gloss-related
synonym
part-of
located-at
```

Approach: Building the KG

Approach

- Extract nodes, relations, and images using BabelNet* v4.0

- Criteria for selecting nodes & relations: nodes and relations with a strong *visual component*



* 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.

Relation types

- We choose relation types from previous work: Cui et al. (2018)*.
- We use 13 out of the 15 proposed types:

`is-a`, `has-part`, `related-to`, `used-for`, `used-by`, `subject-of`, `receives-action`, `made-of`, `has-property`, `gloss-related`, `synonym`, `part-of`, *and* `located-at`.

We have no examples of types `depicts` and `also-see` in the nodes we select from BabelNet.

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.

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

Data Collection

Goal: high-quality, well-curated, and visually relevant

First: Set high-quality seeds as *node pool* (1,000 ImageNet classes used in the ILSVRC image classification competition*)

Steps: iteratively add nodes by following the steps below, until reaching *N* nodes (*N*=90,000)

1. **Retrieve neighbors:** retrieve *first-degree* neighbors;
2. **Validate images:** remove images that do not meet quality criteria;
3. **Filter nodes:** filter out nodes that do not meet criteria;
4. **Update pool:** accept top-k nodes among remaining nodes. (k=10)

* Russakovsky, O., Deng, J., Su, H. et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 115, 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y

Validate images

Apply 4 filters to validate images:

1. check if images are valid files ~ 6.3% invalid, e.g. an audio file
2. remove near-duplicate images remove using SHA1 hashing
3. train a binary ResNet classifier to remove non-photographic images
4. use OpenAI's pre-trained CLIP* model to remove images that do not minimally match any of the node's glosses remove irrelevant images

images: ~ 5.6 million --> ~ 5.3 million --> ~ 2.1 million --> ~ 1.5 million --> 938,100.

[1] [2] [3] [4]

*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.

Binary visual classifier

Many images are very noisy, so we filter out undesirable images.

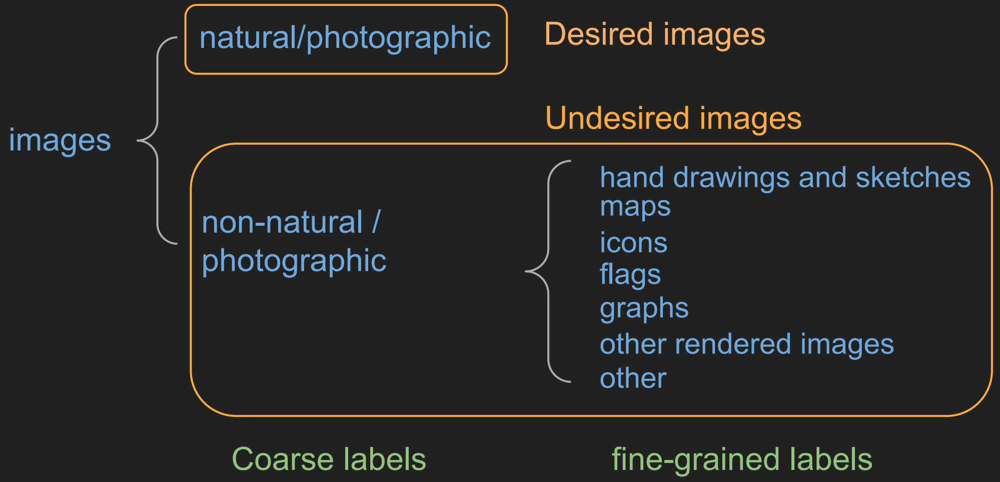
Dataset

*ImagiFilter** dataset: 6k training images.

Model training

Pretrained ResNet-152 + binary classifier fine-tuned on coarse labels.

images ~ 2.1 million --> 1.5 million



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

Remove unrelated images - CLIP model

CLIP:

- Trained on ~400M image-caption pairs
- Bi-encoder architecture (one text encoder, one image encoder)
- Trained to maximize (minimize) the dot-product of correct (random) image-sentence pairs

Use CLIP to match image with gloss:

- Encode all *English* glosses and encode images available for node
- Keep only the images with at least one dot-product > 0.5 with one of the English glosses

Node filtering

At the end of each iteration, filter out the nodes that do not satisfy:

- node has at least one image
- node contains relations with at least two different relation types

Update pool

Accept top-k nodes after sorting. Prioritize nodes:

- nodes with a larger number of images
- nodes that include as many diverse relation types as possible
- nodes that include more relations of the least frequent types

Topic model: Analyzing the KG

Topic Model

- Embedded Topic Model*, an extension of LDA
- Document -> node -> the combination a node's English glosses

Findings:

- KG tends to represent factual knowledge
- Concepts well covered in Wikipedia are also well covered in VisualSem

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).

AB. Dieng, FJR. Ruiz, and DM. Blei. 2020. Topic modeling in embedding spaces. TACL, 8:439–453.

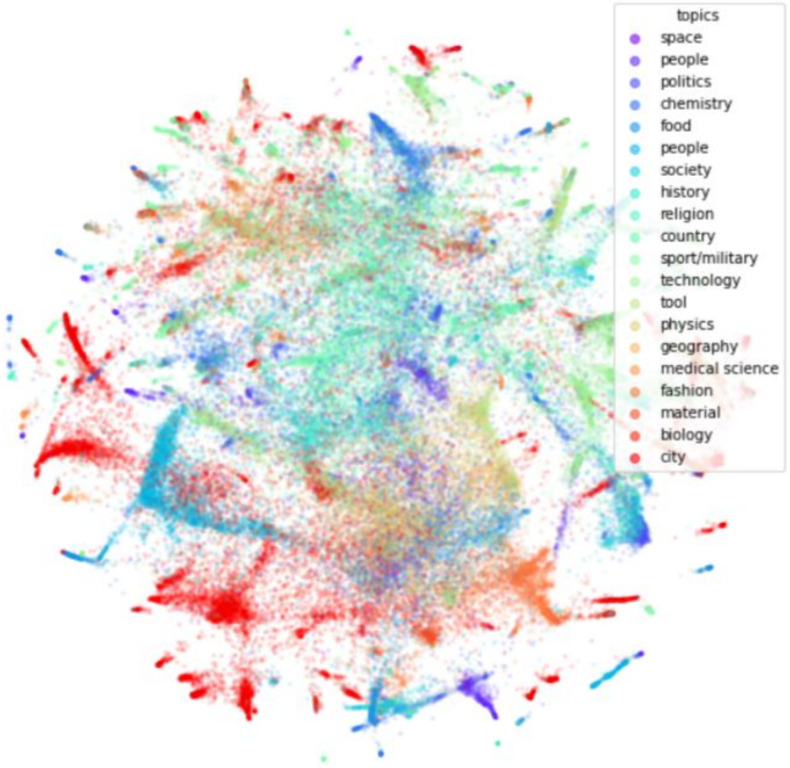


Figure 3: T-SNE plot for node embeddings where each node is represented as the average of its gloss embeddings. Topic assignments are used to colorise node embeddings, and topics are computed with the topic model described in Section 3.2 and Table 3.

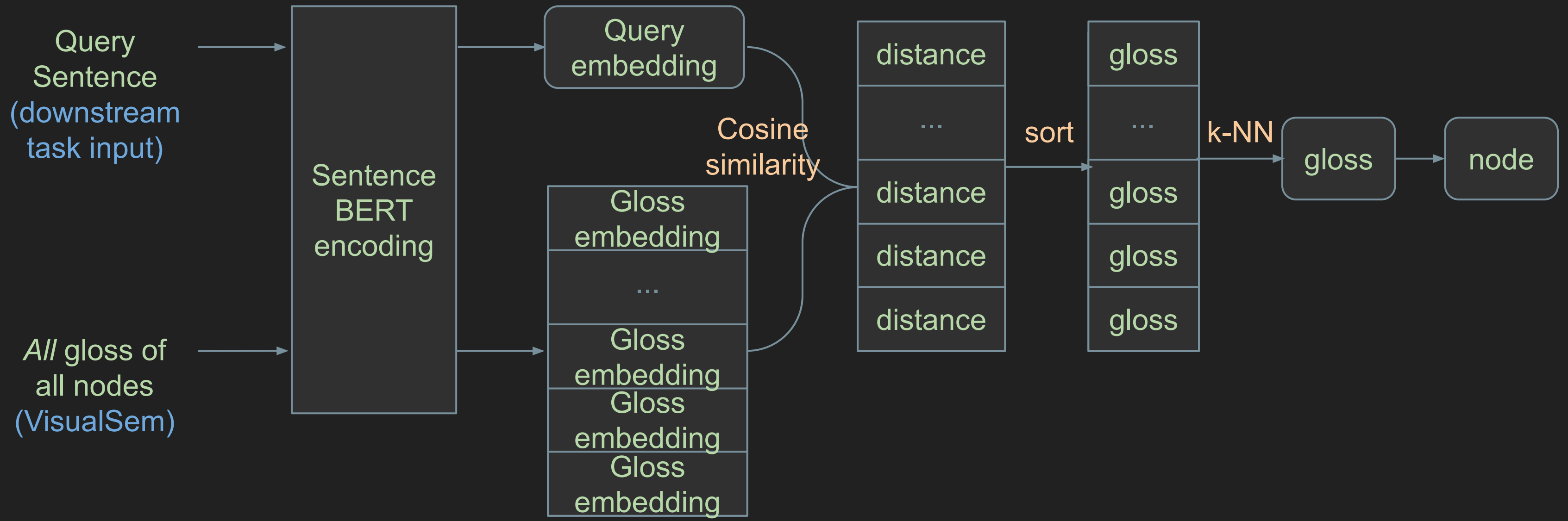
Node visualization

- To understand the topic distribution of the KG
- t-SNE to visualize average gloss embeddings of nodes
- Nodes are colored by the topic assigned to the node by topic model

Using the KG:

Sentence and Image Retrieval

Sentence retrieval



- **Frame the problem:** sentence-to-gloss ranking problem.
- **Model:** k-nearest neighbour (k-NN), encode all VisualSem glosses in training set as well as our query sentences using Sentence BERT* (`paraphrase-multilingual-mpnet-base-v2`)

*Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert networks. In: EMNLP/IJCNLP 2019.

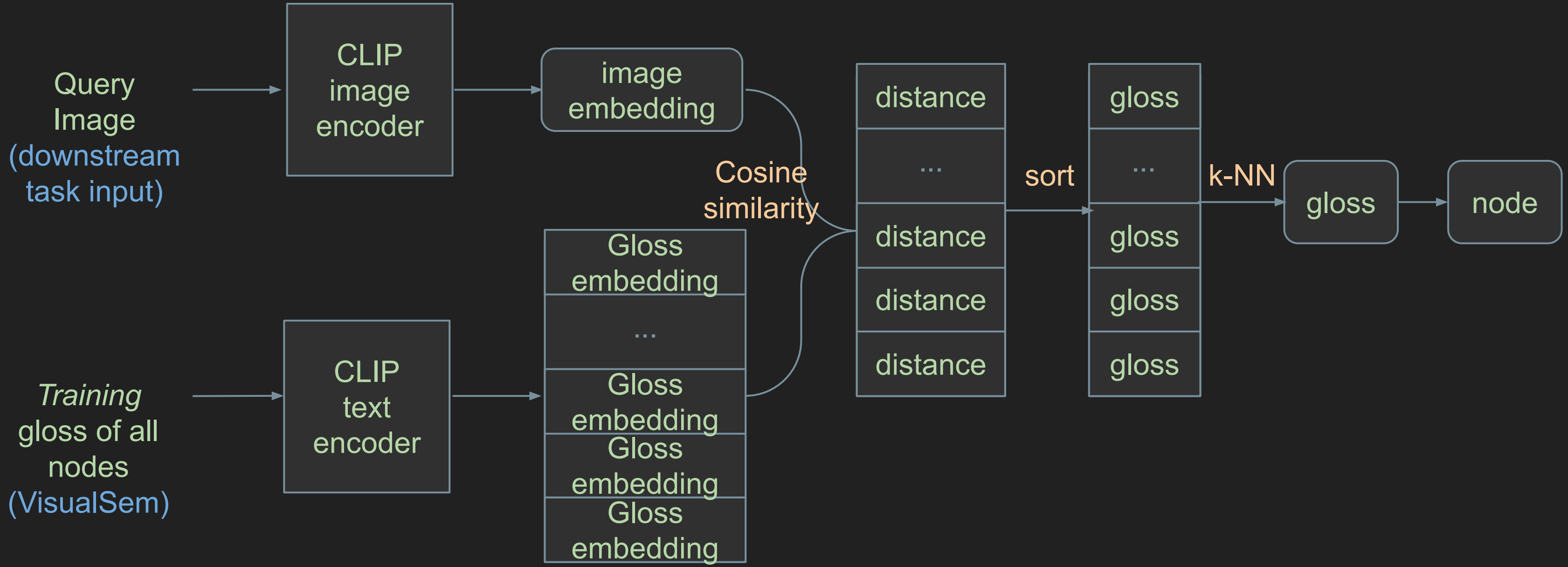
Sentence retrieval evaluation

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

- **Generally good results according to Hits@k**
- mean ranks show high variances -> retrieved nodes could be noisy
- mean ranks of English and Russian queries are the *highest* (the lower the mean ranks, the better)
- **We recommend that users use the sentence retrieval model with care when dealing with lower-quality languages.**

	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)

Image retrieval



- **Frame the problem:** image-to-gloss ranking problem.
- We use the pre-trained CLIP model `RN50x16`.

Image retrieval evaluation

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

The quality of the image retrieval is not as good as the sentence retrieval module.

One of our plans for future work is to investigate how to improve VisualSem image retrieval module.

	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).

Conclusions and final remarks

VisualSem knowledge graph: bridges a gap in resources to train grounded models of language, and is designed to be useful in vision and language research

Neural entity retrieval models that accept text and image inputs - allows for easy integration into (neural) model pipelines

Future work

1. Use VisualSem sentence/image retrieval mechanisms and gloss and image features for **data augmentation** in:

    - NLP tasks word sense disambiguation, named entity recognition
    - vision and language tasks image captioning, visual question answering

2. Improve the quality of the image & sentence retrieval modules

3. Improve and grow the KG (under active development)

Code

https://github.com/iacercalixto/visualsem

Contact

iacer.calixto@nyu.edu

Thank you for attending!

IC has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-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 \textit{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.