Inferring visual semantic similarity with deep learning and Wikidata: Introducing *imagesim-353*⋆

Finn Årup Nielsen and Lars Kai Hansen

Cognitive Systems, DTU Compute, Technical University of Denmark

Abstract. Aiming at multi-modal knowledge representation we construct a dataset with pairs of digital photos of objects. We manually score image pairs for semantic object similarity. A pre-trained ImageNet-based deep neural network predicts the objects and we use the output to estimate the similarity between two images. With a linkage between the neural network and Wikidata, we augment the model and incorporate knowledge graph information into the similarity measure. We compare the machine-based predicted similarity with the human-based semantic similarity.

1 Introduction

At the interface between machine learning and knowledge graphs lies interesting avenues of research and the combination of the two techniques may yield increased task performance. For instance, state-of-the-art results were obtained with a combination of word embedding models and the ConceptNet knowledge graph on a classical word similarity task [14].[1] Knowledge graphs may also be used in connection with machine learning models handling images, where deep convolutional neural networks represent state-of-the-art. Several works have integrated these models and knowledge graphs, e.g., for image classification, object detection or video classification using WordNet [8,15] or ConceptNet [4,17]. Linking the ImageNet dataset [2], through WordNet synsets to the Wikidata knowledge graph [10], the latter could also be used as a resource in computer vision systems. Work on using machine learning to populate Wikidata with quality images is ongoing [12].

In purely text based semantics several word similarity datasets exist, see, e.g., [5,13,14]. Similarity may also be computed on the semantic/synset level. For instance, NLTK implements a range of similarity measures for WordNet synsets based on the WordNet graph [1]. A method for computing similarity also exists for pairs of Wikidata items [9]. Semantic similarity has also been considered for images, e.g., [3,16], and the *Image Similarity Data* dataset exists with triplets of images scored for similarity.[2]

[1] For a list of the state-of-the-art see https://aclweb.org/aclwiki/WordSimilarity-353_Test_Collection_(State_of_the_art).

[2] https://sites.google.com/site/imagesimilaritydata/

In the following, we describe the construction of a semantic similarity dataset (*imagesim-353*) and the combination of a pre-trained deep learning neural network used together with the Wikidata knowledge graph as a means for machine-based visual semantic similarity estimation.

2 Method

2.1 Constructing a visual semantic similarity dataset

Our inspiration for constructing a visual semantic similarity data is the wordsim-353 dataset [5] with 353 pairs of words scored for similarities by humans with a value between 0 and 10 (The lowest value is 0.23, viz. 'king' and 'cabbage'). The dataset has 437 different words and these words are mostly nouns, both common and proper nouns (e.g., monk, drug, proton vs. Freud, FBIS, OPEC) as well as concrete and abstract nouns (e.g., CD, tiger, cemetery vs. category, endurance, recommendation). We set up the following requirements for the images:

1. Must be a color photo of reasonable quality in the JPEG format, neither a drawing nor a gray-scale photo.
2. Must be of a sufficiently large size to input to a neural network. Keras' NASNetLarge model uses images with size 331-by-331, and may be the largest images among the ImageNet-based image classifiers, so the photo should be larger than this size.
3. Should display one single type of recognizable object (understood in a broad sense, i.e. both man-made and natural entities) in the central location of the photo. The object should not necessarily be among the ImageNet challenge categories.
4. No photos of images, paintings nor other forms of depictions. This is to avoid the the ambiguity of what is in the image, e.g., would a photo of a relief display a relief or what the relief displays? Mirrors where neither included.
5. Should not display a person.
6. Should not contain elements that require very specific cultural knowledge of what the object is.
7. Should be freely licensed from Wikimedia Commons.

To collect the images, we used the "Random file" MediaWiki facility in the Wikimedia Commons wiki at `https://commons.wikimedia.org/wiki/Special:Random/File`, and iterated until a suitable image was identified. Using the Wikidata Query Service, we also queried for images that was associated with a Wikidata item linked to a ImageNet WordNet synset with a SPARQL query [10]. Based on the returned list of images, we added relevant images to our list. We also included a few different images of the same object: the leaning tower in Pisa. We collected a total of 353 images. They had varying sizes.

We sought to select images and pair them so the range of similarities was roughly equally represented. The images were not selected equally frequent. In the pairing, 163 images were used once, 108 twice, while one image was used 11 times.

2.2 Human rating of similarity

We constructed a small Flask web application running on the local computer for scoring the similarity of 353 image object pairs. For scoring the similarity, we stressed the difference between similarity and relatedness [7][3] as well as the difference between conceptual and perceptual similarity, see, e.g., [13]. Our target is conceptual similarity, i.e., image pairs should neither be scored for relatedness nor perceptual similarity. We scored images with discrete integer values between 0 (no similarity) and 10 (completely similar). So far the dataset is only scored by a single human (FÅN).

2.3 Machine-based similarity

We use the Keras[4] Python deep learning framework to analyze the images. There are various models implemented in Keras which are pre-trained on the ImageNet image recognition dataset. They yield a 1000-dimensional output representing probabilities over 1000 predefined and fixed classes. We use the ResNet50 model [6] and the included image loading function that resamples the images to a shape that fits with the input of the neural network. We also use the Keras ResNet50 preprocessing function before the image data is feed to the model.

From the output of the neural network for the i'th image, $\mathbf{y}_i$, we compute the correlation coefficient between image pairs, $\mathbf{y}_i$ and $\mathbf{y}_j$, i.e., using the entire distributed representation of an image, rather than just the most probable class label. To explore the information content in the class probabilities for the non-dominant classes we also compute the correlation coefficient from the element-wise logarithm of the output vectors, $\ln(\mathbf{y}_i) = \tilde{\mathbf{y}}_i$ and $\ln(\mathbf{y}_j) = \tilde{\mathbf{y}}_j$.

As a form of (low) baseline measure we compute the similarity between two images as the correlation coefficient between pixel values. We expect this method to perform poorly, but better than chance.

To incorporate a knowledge graph into the similarity computation, we query the Wikidata Query Service with a SPARQL query to obtain a list of properties for each Wikidata item that is linked to ImageNet via a WordNet synset URI.

```
SELECT
  ?item
  (SAMPLE(?synset_) AS ?synset)
  (GROUP_CONCAT(?pid) AS ?properties)
WITH {
  SELECT ?item ?uri WHERE {
    ?item wdt:P2888 ?uri .
    FILTER STRSTARTS(STR(?uri),
      "http://wordnet-rdf.princeton.edu/wn30/")
  }
} AS %items
WHERE {
```

[3] https://www.cl.cam.ac.uk/~fh295/simlex.html
[4] https://keras.io/

```
  INCLUDE %items
  ?item ?p [] .
  ?property wikibase:directClaim ?p .
  BIND(SUBSTR(STR(?p), 37) AS ?pid)
  BIND(CONCAT('n', SUBSTR(STR(?uri), 39, 8)) AS ?synset_)
}
GROUP BY ?item
```

With this at hand, we build a bag-of-properties matrix, $\mathbf{W}$, where each row corresponds to one of the 1000 ImageNet classes (i.e., the ResNet50 outputs and `?item` variable in the SPARQL query) and each column corresponds to a specific Wikidata property. The element values of the matrix are set to the count of the number of times a Wikidata property is used for a specific class. The bag-of-properties matrix is scaled via the *tfidf* transformer in scikit-learn [11] and columns with a count of only zero or one in the matrix are excluded. This scaled matrix, $\tilde{\mathbf{W}}$, is used to project the output of the neural network, e.g., $\tilde{\mathbf{Y}}\tilde{\mathbf{W}} = \mathbf{Z}$, where $\tilde{\mathbf{Y}}$ is the full set of logarithm-transformed output from all 353 images.

To compare the machine-based similarity score with the human ground truth, we use Spearman's correlation, — the standard measure for word similarity [5].

3 Results

We find a Spearman correlation between human similarity scores and machine-based similarity score of 0.62 for correlation coefficients on untransformed neural network output and on 0.67 for correlation coefficients on logarithm-transformed output. The simple baseline on correlation of image (color) pixel values yields 0.12. We did not generally see an improvement when we projected the neural network output through the transformed bag-of-properties matrix, obtaining a resulting matrix on 192 columns. For instance, with $\tilde{\mathbf{Y}}\tilde{\mathbf{W}} = \mathbf{Z}$, a Spearman correlation of 0.60 showed a deterioration in performance. However, combining the logarithm-transformed output matrix with a projection, $[\ln(\mathbf{Y}), \ln(\mathbf{Y}\tilde{\mathbf{W}})]$, to a total of 1192 columns provided a slight improvement in performance to a Spearman correlation of 0.70. Figure 1 shows the swarmplot between the human scoring and similarities computed with this 1192-dimensional space.

4 Discussion

Objects in the natural world are usually "surrounded by context" which make it difficult to dissociate the object from the context. For instance, liquids such as coffee are usually found in a container, such as a coffee mug. For a viewer (human or machine), it may be ambiguous whether a photo of coffee is a photo of coffee or a coffee mug. Another example is a bridge over a river: Is it the bridge or the river that is the object? The distinction between similarity and relatedness is harder to maintain for these cases. Another problem for the determination of a ground truth similarity may be the recognizability of the objects and the

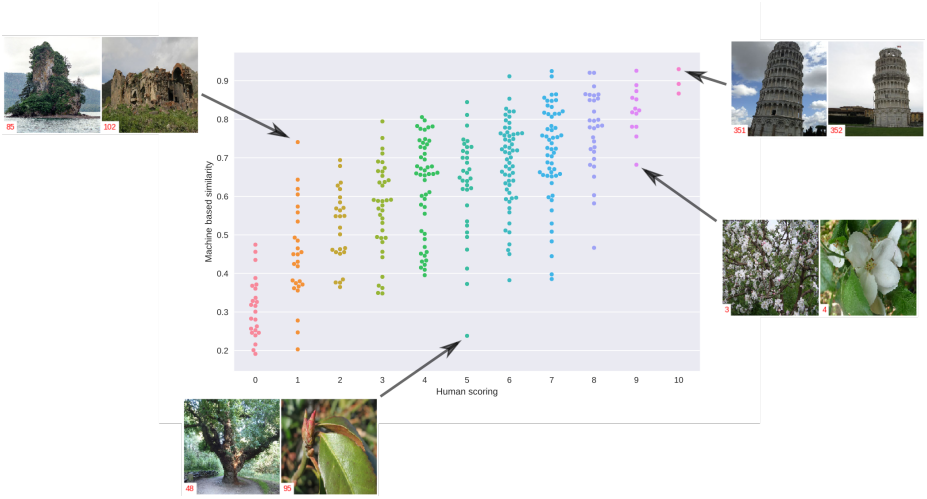

Fig. 1. Swarmplot for image similarities between human scoring and ResNet50 output correlation augmented with a Wikidata bag-of-properties projection. The images show four outlying pairs: Three where the human and machine similarities diverge (New Eddystone Rock/Arates monastery ruin, Klopstock's Oak/*Euonymus japonicus* buds and apple tree flowers) and one where they both score with high similarities (Leaning Tower of Pisa).

Creative Commons images by Jerzystrzelecki (CC BY), Z galstyan (CC BY-SA), Jerzy Opioła (CC BY-SA), Dahola (CC BY-SA), Rodrigo Pereira da S... (CC BY-SA) and SLCESAR (CC BY-SA). All images are from Wikimedia Commons resampled with Keras.

required detailed knowledge of the similarity between the objects depicted. Furthermore, what similarity should be assigned to an object photographed from different angles or at different times? There are probably no definite answer to this question, cf. the philosophic discussion of *temporal parts*.

Many extensions of this work are possible: Keras contains several other pretrained models apart from ResNet50. The evaluation of these models should be straightforward. The use of the last internal layers of the neural network—instead of just the output—could also be interesting to explore.

Our dataset is small, restricting the use of machine learning to optimize a model for similarity. Crowd-sourcing could create a larger dataset, Multiple human annotators of the similarity might also improve the quality of the human scoring and give an indication of the variability of the human visual semantic similarity scoring.

Acknowledgment: This research is funded by the Innovation Foundation Denmark through the DABAI project.

References

1. Bird, S., Klein, E., and Loper, E. Natural Language Processing with Python.
2. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Li, F.-F. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition* (June 2009).
3. Deselaers, T., and Ferrari, V. Visual and semantic similarity in ImageNet. *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (June 2011), 1777–1784.
4. Fang, Y., Kuan, K., Lin, J., Tan, C., and Chandrasekhar, V. Object Detection Meets Knowledge Graphs. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (August 2017), 1661–1667.
5. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. Placing search in context: the concept revisited. *ACM Transactions on Information Systems 20* (January 2002), 116–131.
6. He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition* (December 2015), 770–778.
7. Hill, F., Reichart, R., and Korhonen, A. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics 41* (August 2014), 665–695.
8. Marino, K., Salakhutdinov, R., and Gupta, A. The More You Know: Using Knowledge Graphs for Image Classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition* (July 2017).
9. Nielsen, F. Å. Wembedder: Wikidata entity embedding web service.
10. Nielsen, F. Å. Linking ImageNet WordNet Synsets with Wikidata. *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France* (April 2018).
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12* (October 2011), 2825–2830.
12. Redi, M. How we're using machine learning to visually enrich Wikidata. *Wikimedia Blog* (March 2018).
13. Smith, L. B., and Heise, D. Perceptual Similarity and Conceptual Structure. *Percepts, Concepts and Categories The Representation and Processing of Information* (December 1992), 233–272.
14. Speer, R., Chin, J., and Havasi, C. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (December 2016), 4444–4451.
15. Tang, Y., Wang, J., Gao, B., Dellandrea, E., Gaizauskas, R., and Chen, L. Large Scale Semi-Supervised Object Detection Using Visual and Semantic Knowledge Transfer. *2016 IEEE Conference on Computer Vision and Pattern Recognition* (June 2016).
16. Wang, J., song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., and Wu, Y. Learning Fine-grained Image Similarity with Deep Ranking. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (April 2014).
17. Yuan, F., Wang, Z., Lin, J., D'Haro, L. F., Jae, K. J., Zeng, Z., and Chandrasekhar, V. End-to-End Video Classification with Knowledge Graphs.