Knowledge Engineering with Image Data in Real-World Settings

Margaret Warren[a], David A. Shamma[b], and Patrick Hayes[a]

[a] *Institute for Human and Machine Cognition, Pensacola, Florida USA.*
[b] *Centrum Wiskunde en Informatica (CWI), Amsterdam, The Netherlands*

Abstract

We report on experiences in adding ML-trained visual recognition modules to a human-oriented image semantic annotation tool which creates RDF descriptions of images and scene contents. We conclude that ML cannot replace expert humans but can aid them in various ways, some unexpected. Semantic markup systems can be to designed to align human and machine blind spots. Finally, we briefly outline directions for future work.

Keywords 1

human-centered, knowledge engineering, image annotation, AI, ML, computer vision, HCI

1. Introduction

We have a mature semantic markup system for images that allows subject-matter expert users to construct RDF knowledge graphs as image annotations, intended for use in domains where objects and relationships are specialized and require expertise to identify. With a view to improving the functionality of this system, we recently extended it by adding modern pre-trained visual classifiers and object recognition software to automate bounding box creation and suggest classification labels for objects in the image. While this automation has its advantages, principally to speed through the rapid localization of items in the photo, we see the addition of automatic vision systems as a technique for assisting rather than replacing human annotation.

2. Structured Relationship Annotations

There is no shortage of tooling for annotating images with object bounding boxes that enclose specific classes for training. Much of the work on this class of tools seeks to speed up the task of drawing or specifying the points around the target object. [4] With simple boxes, relationships became important for scene understanding; for example, knowing a coffee cup is on a table has a specific relationship (in this case: "is on") which provides more information than simply knowing an image contains both objects. Annotations of relationships also bring a distinctly new set of tooling from just bounding box labeling. These visual relationships are principally represented in the 2016 Visual Genome [8] project, which contains over 100,000 images with 3.8 million object instances and 2.3 million relationships. Beyond the Visual Genome project's overall scale, the relationships included in the dataset are dense where a plurality of relationships can exist between the same set of objects. However, while mapped to Wordnet Synsets,

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, Palo Alto, California, USA, March 22-24, 2021.
EMAIL: mwarren@ihmc.us (A. 1); aymans@acm.org (A. 2); phayes@ihmc.us (A. 3)
ORCID: 0000-0002-6680-2431 (A. 1); 0000-0003-2399-9374 (A. 2); 0000-0002-6639-9187 (A. 3)

Visual Genome lacks commonly used structures, like RDF or OWL, in its representations and provides no tooling for creating annotations.

In contrast, the ImageSnippets[2] tool was designed to experiment with ways to produce structured semantic image markup by allowing users with minimal training to create machine-readable, ontology-based image and scene descriptions in the RDF syntax. Image descriptions, referred to as semantic markup or image graphs, are created as RDF triple stores which use the image identifier as root, and a core Lightweight Image Ontology (LIO) vocabulary of 11 relations, allowing users to quickly describe a variety of relations between objects and the scene, including the level of importance of an object to a scene, whether an object is in the foreground, background or has some other function and several other relations [6]. Users can also use the tool to add additional properties, describe spatial relations between objects, or engineer new ontologies, including those with OWL-type structures based on image contents. Scene objects and image properties are mapped to entities found in DBpedia, Wikidata, and other publicly accessible linked-data corpora, or custom-created if no existing concept can be found. Entity lookup is semi-automatic but guided by users, using an intuitive interface.

3. Experts Throughout the Loop

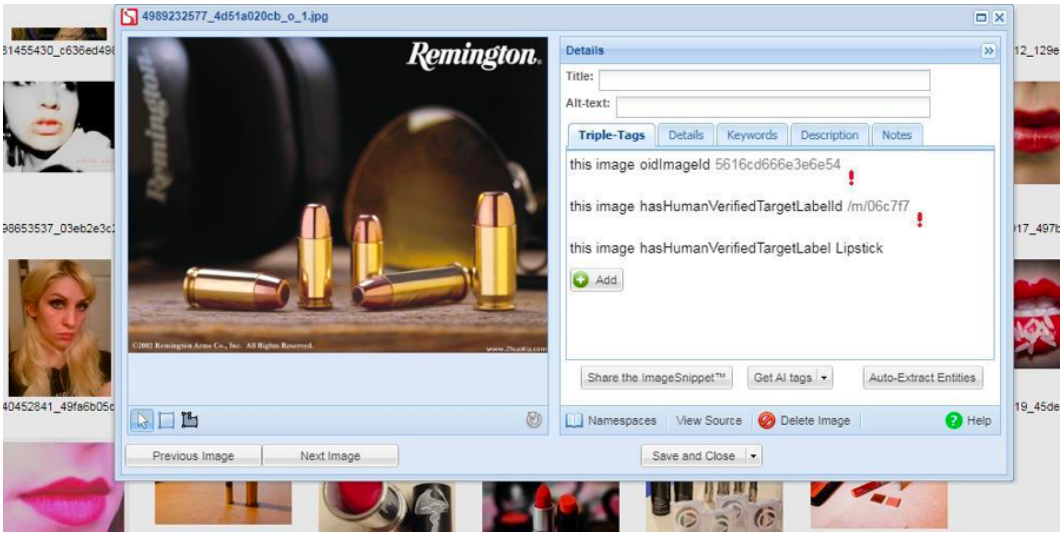

Figure 1 Remington bullets that the classifier predicted as lipstick were later verified incorrectly by a human editor.

Our annotation applications typically involve specialist domain knowledge and create structured data backed by formal ontologies. In many annotation settings, image classification and object recognition are error-prone even with human verification (see **Figure 1**). It is essential for the outputs of automated image classification and recognition tools to be evaluated by how they aid and support, rather than replace, expert human users. More, human expertise is necessary not only as a final verification check but throughout the entire process. [2] Our experiences in adding automated annotations have highlighted several findings. First, the utility of locating and isolating items of potential interest in complex images is useful, mostly independent of the predicted annotation label. Second, the predicted label of a targeted bounding box may be helpful as a base qualification at a high level of generality in a typical formal ontological classification. This can aid the human annotator by directing their attention to the relevant topic and guiding the search for formal concepts. In other words, even trivial detections have utility in expert domains. We note that image recognition and human visual abilities often complement each other in these situations when working

[2] Demonstration available at http://imagesnippets.com/.

quickly with complex or crowded images. For these reasons, we assert human experts must be involved throughout the annotation process's lifecycle when it comes to specialized domains.

4. Domain Example

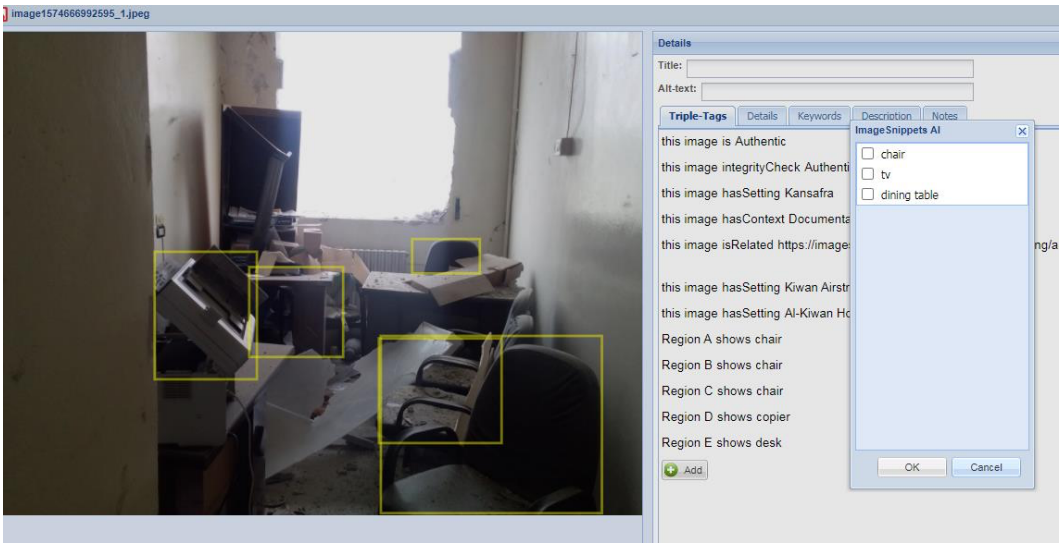

Figure 2 Annotating an image of a room in a hospital after an airstrike.

Figure 2 shows an example using images collected after an airstrike on a hospital where the user's goal is to engineer a knowledge graph from the accumulated evidence of war crimes. Beyond identifying objects like oxygen bottles, hospitals, classified aircraft, people, and damage type, one must also account for knowledge of the terrain and context. Further shown in the figure is the systems interface at the point where the user has called on the object detector, which has found a chair, tv, and dining table in the image of a hospital room after an airstrike. At this point, the user can decide whether each detected region should serve as a subject of a triple in the RDF annotation. If so, then—regardless of whether the object is accurately identified—the user can accept the region and either accept the object label provided by the computer vision system (in which case the detected object label is automatically mapped to correct DBpedia and Wikidata values which become an object of the triple) or simply ignore the offered label and instead manually insert a correct label value. In this example, a copier was located as a region of interest but misidentified as a television. However, even in this misidentified case, the vision detector plays a significant role in 'noticing' the object of interest and locating it in the image with a bounding box far more rapidly and reliably than a human user. The result is a correctly identified object, accurately located in the image by a synergistic collaboration between human expertise and ML-trained identification and classification, each strengthening each other's weakness. The user can then further adjust the triple by altering the relationship of the objects to the overall image using other terms in the LIO vocabulary, perhaps by specifying unique relationships such as '`desk` *`isUnder`* `wall`', or by adding context: '`image` *`hasSetting`* `Office`' or establishing scene relationships: '`this image` *`hasInBackground`* `motorcycle`.'

5. Future Work

To date, our work has primarily focused on integrating contemporary classifiers and detectors into a human-centered semantic annotation system. But this work has illuminated and suggested several new future pathways. By observing first-hand problems such as underspecification in the ML pipelines [1], we see

utility for semantic annotation methods to become part of environments where data excellence can be incentivized [5], and machine learning algorithms can be examined by people as part of internal AI auditing frameworks [9]. Future work will include the generation of expert-created training and test sets that can be fed back, using transfer/active learning methods, to create models trained to return increasingly more precise object suggestions, as well as the production of test sets for spatial relation scene graphs research [3] and the creation of adversarial training sets through the rapid human identification of machine blind spots in current image classifiers. Combined with the already apparent utility for the machine to find human blind spots, we feel we are working towards a beneficial synergy of blind spot alignment, visual learning, and knowledge representation.

6. Acknowledgments

Many thanks to Jonathan Dotan at Stanford and Stephen Honan at Hala Systems for providing the domain-related discussions and examples.

7. References

[1] A. D'Amour, K. Heller, D. Moldovan B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, and F. Hormozdiari, 2020. Underspecification presents challenges for credibility in modern machine learning. arXiv preprint arXiv:2011.03395.

[2] D. A. Shamma, L. Kennedy, J. Li, B. Thomee, H. Jin, and J. Yuan. 2016. Finding Weather Photos: Community-Supervised Methods for Editorial Curation of Online Sources. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (*CSCW '16*). Association for Computing Machinery, New York, NY, USA, 86–96. DOI: https://doi.org/10.1145/2818048.2819989

[3] D. Nunes, L. Ferreira, P. Santos, A. Pease, ″Representation and Retrieval of Images by Means of Spatial Relations Between Objects" AAAI Spring Symposium on Combining Machine Learning with Knowledge Engineering AAAI-MAKE (2019) http://ceur-ws.org/Vol-2350/paper7.pdf

[4] D. P. Papadopoulos, J. R. Uijlings, F. Keller, & V. Ferrari, (2017). Training object class detectors with click supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6374-6383).

[5] I. D. Raji, A. Smart., R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, 2020, January. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33-44).

[6] M. Warren, P. J. Hayes, "Bounding Ambiguity: Experiences with an Image Annotation System." 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing *SAD/CrowdBias@HCOMP*. (2018). http://ceur-ws.org/Vol-2276/paper5.pdf

[7] N. Sambasivan, S.Kapania, H. Highfill, D. Akrong, P. Paritosh, L.Aroyo. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21). Association for Computing Machinery, New York, NY, USA, To appear.

[8] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, J. Li, D. A. Shamma, M. Bernstein, F. Li, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations." *International Journal of Computer Vision* **123,** 32–73 (2017). https://doi.org/10.1007/s11263-016-0981-7

[9] Ramya Ramakrishnan, Ece Kamar, Besmira Nushi, Debadeepta Dey, Julie Shah, Eric Horvitz, 2019, "Overcoming Blind Spots in the Real World: Leveraging Complementary Abilities for Joint Execution." *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (01), 6137-6145