Anupam Guha aguha@cs.umd.edu
http://www.cs.umd.edu/~aguha • Department of Computer Science, University of Maryland, College Park, MD 20742

Research Statement

What Are We Talking About?

Humans are able to perceive the world, understand their perception via concepts, and communicate their perception of the world in a mutually intelligible manner to other humans, via natural language. In order to build effective intelligent agents which operate in the human world, and interact with humans, they too must have the capacity to not just understand the world, but to understand from human communication what is being referred to in a particular span of symbols. They should be able, like humans, to map these representations of concepts to what they observe. Furthermore, they should be able to use the knowledge obtained from reference chains in language to continuously refine their perception. My research goal, thus, is to investigate the problem of *what (visual) entity are we referring to in what span of text?* (Coreference Resolution), tie it to *how is it being represented?* (Embeddings), and then use this knowledge to help perceptual models decide *what is this?* (Detection). My past and current work is necessarily interdisciplinary and attempts to unify certain ideas in computational linguistics, computer vision, and deep learning.

The representations humans use in natural language to convey concepts about the world are neither direct nor trivial, and yet other humans are able to resolve with fluency, what span of text refers to what real world entity. In Linguistics, this is the problem of **Coreference Resolution**. Traditionally, the research on this is restricted as an NLP problem, and the datasets for coreference resolution are primarily obtained from journalistic sources, as newswire dominates NLP tasks. But humans neither refer to entities in language isolated from perception nor do they communicate always in clear references to entities the way trained journalists do. On the contrary, human language is complex and dirty, and human ability at sorting out references relies on both common sense and sophisticated world knowledge. In 2015, I investigated a new source of *hard and interesting* coreference problems [3], namely the quiz bowl trivia game. Quiz bowl questions, like the one in Figure 1 are structured as a series of sentences, each a clue towards the answer. There are multiple clusters of coreferent spans, one of which contains this answer. The machine learning models designed to solve coreference problems should be able to solve these. But the then state-of-the-art models [7] performed poorly on this dataset and my error analysis revealed that current models lack the ability to use the richness of world data and semantics effectively. In that work I proposed for the first time using distributional semantics (like word2vec and GloVe) [6] as a feature for my model, and that helped our relatively simple model beat sophisticated ones. **Word embeddings** does not just help us solve the text coreference problem, it helps us reason about complex vision problems.

[The Canadian rock band by [this name]] has released such albums as Take A Deep Breath, Young Wild and Free, and Love Machine and had a 1986 Top Ten single with Can't Wait For the Night. [The song by [this name]] is [the first track on Queen's Sheer Heart Attack]. [The novel by [this name]] concerns Fred Hale, who returns to town to hand out cards for a newspaper competition and is murdered by the teenage gang member Pinkie Brown, who abuses [the title substance]. [The novel] was adapted into [a 1947 film starring Richard Attenborough]; [this] was released in the US as Young Scarface. FTP, identify [the shared name of, most notably, [a novel by Graham Greene]].

Figure 1: An example quiz bowl question about the novel *Brighton Rock*. Every mention referring to the answer of the question has been marked; note the variety of mentions that refer to the same entity.

While computational linguistics and computer vision tasks are today treated as separate areas, solving coreference and allied problems in isolation from visual features is hard, especially for complex non-realistic representations requiring high level reasoning, say paintings and comics/cartoons which have proven resistant to current deep learning methods. How does a human infer that a blob of paint refers to the not-so-real world concept of an angel if said human has only looked at traditional representations of angels before? Problems like these confound the recent efforts made to create multi-modal representation of entities between these two domains. In 2016 one of my works was to recognise paintings [1] given text questions about them. In that work, even if we had annotated the objects in the paintings (for which we collected a dataset, an instance of which is Figure 2), they were not referred directly in the text. My work showed that it is possible to infer from the word vector space, learnt from enough data, not just the class of the object being talked about, but its visual properties and that these can be enough to understand the whole painting. For example, most if not all depictions of angels will occur near the top of the painting (flying) instead of the bottom, and unsurprisingly the word vector representation of angels has a smaller Euclidean distance in word-vector land from "top"-like words than "bottom"-like words.

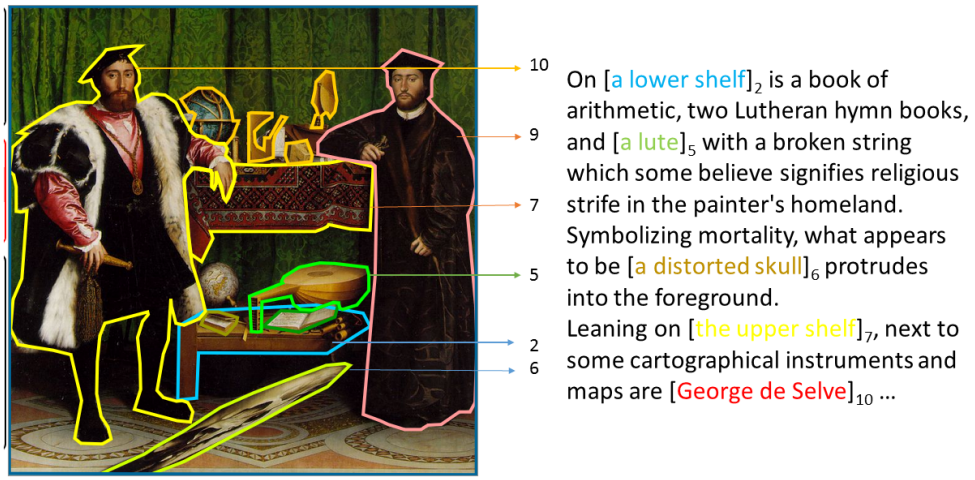

Figure 2: A painting example from the dataset we created.

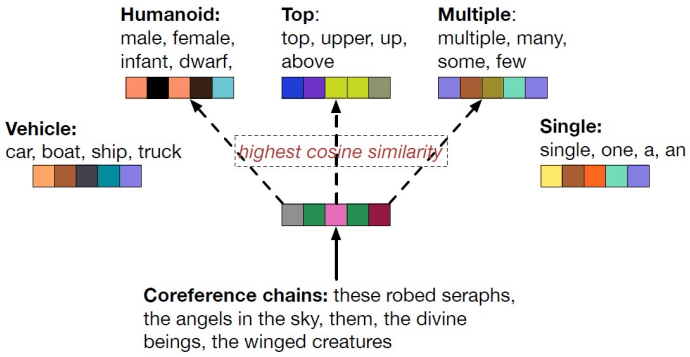

Figure 3: How word embeddings can infer visual properties.

More recently, I've worked with another kind of non-realistic image representation, namely comics, in which text and images need to work together to convey meaning. While my work on this [9] is ongoing, current results demonstrate how hard it is to solve these kinds of images. Look at Figure 3 to see how complex visual properties can be obtained from text. One should note here that creating and annotating new vision corpora is expensive whereas text depicting vision is aplenty and cheap.

Conversely, solving human vision in isolation from the *conceptual scaffolding* of language is hard as well. I have worked in the past at using grammar like structures to understand human actions in vision [4, 5] but the role of language in human vision is more than *structural*. Humans constantly use language to create the context of visual objects, not directly as an ontology, but indirectly, through references in text, till the scaffold built up is robust enough to aid even the hardest of vision problems. This context is fed back by the brain in a top-down fashion [8], to be used to refine the *quick and dirty* visual perception that happens unconsciously, after which the human actually *observes*.

The analogous idea of sending a top down signal obtained from "coreference resolution land" in language into a deep network learning vision categories to *regularise* it helped me win the **Qualcomm Innovation Fellowship, 2016**. This idea, as shown in Figure 4, which is also the basis of one of my two current projects can be demonstrated in this statement, *it is highly unlikely to find a giraffe standing in a kitchen*. So if your deep network is detecting the following object categories in a scene: knife, countertop, giraffe, plate, cupboard, one thing is not like the others, and your network should suppress that particular interpretation of the corresponding signals at a lower level so the category changes. The visual properties of giraffes are inferred from easily available text corpora in the manner described above in the paintings work. This part of my current work is called the language driven visual reasoner. This kind of controlled top-down regularisation of neural networks is completely alien to current deep learning architectures and takes inspiration from the human vision mechanism. Currently, while I'm limiting the process to refining categories, in the

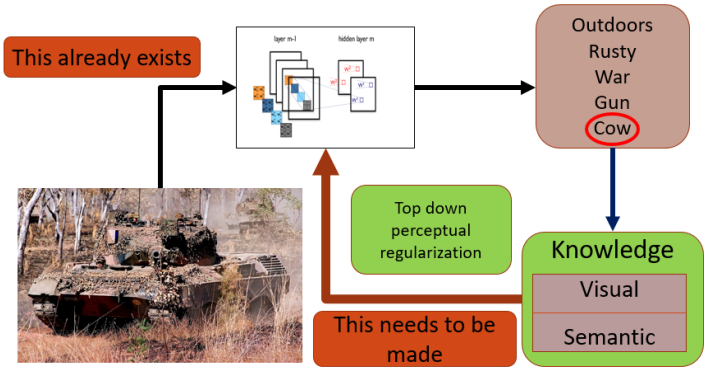

Figure 4: We learn from entity references present in language corpora that one of these things is not like the others

future I plan on extending this to unseen categories, or zero-shot learning [10].

The other project I'm working on currently [2] goes beyond singular entities in text. It started in collaboration with Mohit Iyyer in 2015 and dealt then with creative language understanding. The work used a deep autoencoder to understand the relationship between two characters in a novel as a sequence of exemplar topics, which significantly outperformed current topic models. I extended this work in my summer research at Comcast under Dr. Ferhan Ture, to infer not just relationships between pairs of entities in text, but all kinds of properties expressed as sequential lists, as shown in Figure 5, connected together in a hierarchical topic model, learnt by a combination of deep learning and interpretable dictionary learning. I am currently working on a model that will take this and add visual features to it if the data has corresponding text and image frames, and thus it will be possible to tell, for example, which movie frame sequences contain the particular sequence of topics **village: war: sadness**, all obtained in an unsupervised manner from a collection of movies and their scripts.

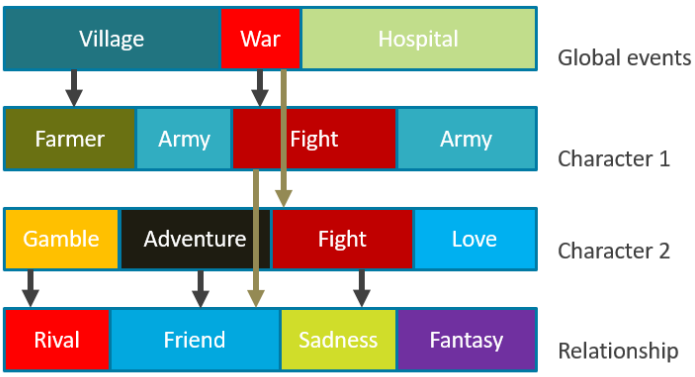

Figure 5: Sequences from a movie, described as connected lists of topics

In the long term I plan on using this to extend my language driven visual reasoner to cache even more abstruse relations, like *it is more likely for sad scenes in movies to be dull coloured*. Humans are able to infer such complex relations with their lifetime's experience of language, vision, world knowledge, and also transmitted historical and cultural intelligence and the intelligent agents we build should be able to use similar constructs to inform and refine their deep learning. I am interested in not just extending the problem of referring entities (and events, relationships, actions) to modalities other than language and static images, for example video, but also use these complex references to refine multiple kinds of perception and reasoning at multiple levels of granularity. Current deep learning architectures are either feedforward like CNNs where representations only go from lower to higher level of abstraction, or the signal even if it goes back in a directed cycle, like RNNs, has no capability for top-down control and regularisation. My work attempts to not just create better machine learning models for these problems, it also will fundamentally improve how we incorporate knowledge into machine learning, and how we unite the symbol with the signal.

References

1. **Anupam Guha**, Mohit Iyyer, and Jordan Boyd-Graber. "A Distorted Skull Lies in the Bottom Center..." Identifying Paintings from Text Descriptions, *North American Association for Computational Linguistics (NAACL) Human-Computer QA Workshop*, (2016)
2. Mohit Iyyer, **Anupam Guha**, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. Feuding Families and Former Friends: Unsupervised Learning for Dynamic Fictional Relationships, *North American Association for Computational Linguistics (NAACL)*, (2016).
3. **Anupam Guha**, Mohit Iyyer, Danny Bouman, and Jordan Boyd-Graber. Removing the Training Wheels: A Coreference Dataset that Entertains Humans and Challenges Computers, *North American Association for Computational Linguistics (NAACL)*, (2015).
4. Yezhou Yang, **Anupam Guha**, Cornelia Fermüller, and Yiannis Aloimonos. Manipulation Action Tree Bank: A Knowledge Resource for Humanoids, *IEEE/RAS International Conference on Humanoid Robots*, (2014).
5. **Anupam Guha**, Yezhou Yang, Cornelia Fermüller, and Yiannis Aloimonos. Minimalist Plans for Interpreting Manipulation Actions, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (2013).
6. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffery Dean. Distributed Representations of Words and Phrases and their Compositionality, *Advances in neural information processing systems (NIPS)*, (2013).
7. Greg Durrett and Dan Klein. Easy Victories and Uphill Battles in Coreference Resolution, *Empirical Methods in Natural Language Processing (EMNLP)*, (2013).
8. Charles D. Gilbert and Wu Li. Top-down influences on visual processing, *Nature Reviews Neuroscience*, (2013).
9. Mohit Iyyer, Varun Manjunatha, **Anupam Guha**, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daumé III, and Larry Davis. The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives, *arXiv*, (2016).
10. Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-Shot Learning Through Cross-Modal Transfer, *Advances in neural information processing systems (NIPS)*, (2013).