Mark Yatskar
Research Statement

My research is motivated by a desire to create intelligent machines that have a shared understanding of the world with us (e.g. when a person asks a machine to “cut vegetables,” it does the right action), are able to communicate with us (e.g. when a person asks “where is the vegetable stock?”, it finds the right object and responds) and can ultimately be used in everyday environments with us. My work looks at the interplay between natural language and visual perception for the purpose of building such a machine.

I study how language can be used to structure visual perception.

I am designing new machine learning approaches that enable tight coupling between how people express themselves in language, how machine behavior is specified and how machines ultimately express themselves back to people. Such an agenda will, if successful, massively expand the scope of what machines can understand about the world. Yet if machines learn from people and their language, we risk allowing them to inherit human bias. Therefore, in addition to my main theme, I have been actively researching where this is happening and how to mitigate it. I have three active directions:

• How can language be used as a scaffold to accelerate visual intelligence? [1; 2; 3; 4]
• How can machines communicate about the world in natural language? [5; 6; 7; 8; 9]
• How do we specify machine behavior without inheriting human bias? [10; 11]

This research is multi-disciplinary, spanning the areas of natural language processing, computer vision, and fairness in machine learning. The following sections provide a high level summary of my recent advances in these three areas and then highlight the key challenges we must address to make further progress.

Language as a Scaffold for Perception  Natural language provides a rich signal for what people can perceive about the visual world. The most basic atoms of language, such nouns and verbs do not only provide a basis for communication but also provides descriptions and labels of what objects and actions people recognize in the world. But language is not made of words in isolation: words in a vocabulary have a relationships to each other as do words in a single sentence. For example, the phrase “cutting with a blade” not only implies the existence of “cutting” events and “blade” objects but also that blades can be “tools” for cutting. My work has shown such relationships also hold a rich signal for what people can perceive about the world and can be used to build large scale grounded common sense repositories and broad coverage recognition resources for computer vision.

Pairwise relationships between nouns and verbs, in context, provide an implicit structure for understanding the visual world. Consider the example in Figure 1, where the same action is taking place, “pouring”, and the same objects are involved, “man”, “beer”, “glass”, but people rarely indicate the same thing is happening in the two images. The key difference is that in one image the beer is ending up in the glass (the destination of the pouring) while in other the beer is leaving the glass (the source of the pouring). This classification of how entities participate in events is called Semantic Role Labeling (SRL) [12] and is an important aspect of how we use language. SRL is a common problem in natural language processing in the context of sentences and it’s associated resources, such as FrameNet [13], can be used to provide labels
for these images. This is the basis of a formalism I introduced called Situation Recognition [1], which is the first structured representation of events in images. The structure corresponds to people’s natural notions of event structure in the world because it resembles sentence structure, allowing for large scale annotation and wider semantic coverage of events than previously possible. Beyond coverage, prediction of actions, entities, and their shared interaction together improves over predicting any one in isolation. This work has been featured in New York Times.

Another example of relationships between words that can be used to discover visual information taxonomic “is-a” relationships, such as “fork is-a utensil.” Consider the case where we have a dense sampling of images over a small number of object types, such as is commonly found in object detection resources such MS-COCO [14]. By observing the geometric layout of objects, we can directly estimate which objects are, for example, found “on” another. But this will provide low coverage: few objects types can be covered with sufficient density to accurately estimate many types of relationships between them. On the other hand, “is-a” relationships, such as those found in WordNet [15], can be used as pivot to create a large scale repository of commonly occurring visual relationships, as can be seen in Figure 2. This allows us to compile accurate knowledge which is physical, not commonly covered in text, more exhaustive than what people can usually produce, at quantities several orders of magnitude larger than otherwise [3].

While relationships between words, as expressed in language, can allow us to achieve massive coverage of concepts in the visual world, methods trained for prediction in this setting face significant challenges. Any one particular concept is likely to occur infrequently during training, making standard methods that depend on hundreds of examples for any one desired output, fail. Instead, I designed models focused on composition that are able effectively put together pieces that they have rarely observed to create novel output [2]. Such methods achieve higher data-efficiency by explicitly creating general purpose composition functions that estimate what a combination of outputs would look like without allowing parameters specific to that combination. Another problem that commonly occurs in such large scale settings is that there are long range dependencies between predictions: one prediction may obligate prediction many pieces. Methods require careful model structure that allows dependencies between relevant elements to inform each-other without significant numbers of intervening, unrelated predictions [4].

**Communicating about the World in Natural Language**  Tightly coupling how information is represented in a machine and in language (i.e. by having the machine store facts about the world in natural language statements) allows the machine to communicate more easily. My work in this direction has been in both grounded and classical natural language processing settings. For example, in the task of converting images to descriptions (i.e. given an image of a dog jumping, describe the dog and the circumstances in a sentence), I have shown that natural language representations of image content can be used to construct captions that cover more types of content present in captions [7]. Furthermore, others have shown that the representations of events and entities I introduced above (Situation Recognition) can allow systems to more accurately generate image captions [16].

Beyond grounded problems, I have also considered natural language processing problems, such as question answering. I introduced a dataset of teacher-student style dialogs, where a student asks a teacher about the contents of a Wikipedia article and the teacher must respond with text found in the article [9]. For example, a student can ask about the origin of the Daffy Duck character, such as “when was he introduced?”
and a teacher must respond by selecting some text from an article about Daffy Duck’s origin, such as “first appeared in Porky’s Duck Hunt.” Such dialogs are not wholly natural because people engaged in dialog normally adapt the information they know to conversation (i.e. add “he” to the response above). Importantly, the information content of the reply is represented in natural language (an extraction from Wikipedia text), and this simplification has allowed us to create a large scale datasets for robust training and evaluation. Even for this constrained setting, existing methods fail because they face significant challenges: they need start conversations with general content, explore details as the student expresses interest, be able to say they don’t know, and avoid unnecessarily repeating information. Finally, I have also studied converting graph representations of meaning (i.e. predicate argument structure connected in a graph) to natural language statements, creating one of the best performing methods [8] and introducing one of the earliest methods for producing simple language from complex language [5].

Human Bias in Machine Learning  Learning from human authored datasets carries the risk of embedding human bias into models. For example, facial recognition software often works better for people of European decent because they appear more often in internet photographs, making its deployment for policing efforts extremely problematic. Recent work has shown other instances where this happens [17; 18], and that such biases can have significant impact on the users of systems [19]. These effects can be multi-faceted: impacting how people are perceived [20] or simply by working poorly for some people [21]. My work in this area has focused on such problems in vision and language understanding, with the intention of creating artificial intelligence systems that we can have confidence will work well for a diverse set of people.

It is generally assumed that biased training data creates biased algorithms, but in fact, the situation is significantly worse. In my work, I have shown that models trained on biased data actually amplify training set biases, exacerbating existing problems of misrepresentations [10]. In particular, grounded visual data, such as the images seen in Figure 3, is commonly sourced from the internet. These collections contain significantly more males than other types of people and depict many of their subjects in stereotyped ways [17]. When models are trained on this data, gender biased associations. For example training images depicting cooking were originally associated with woman over men by a ratio of 2:1. After training, a model evaluated on similarly distributed data associated woman with cooking by a ratio of 4:1.

A minimal fairness goal for systems trained on biased training data is to not make stereotyped associations worse than the original data. When finding ways to mitigate such effects, two requirements are usually in tension: accuracy and fairness. Outputting random predictions at the correct frequency satisfies our fairness goal, but makes our system practically useless. Surprisingly, it is possible to reduce this bias amplification without suffering any loss in model performance. To achieve our fairness goal, our work introduces Reducing Bias Amplification (RBA), a general purpose technique for encouraging structured prediction models to respect training time statistics at test time. RBA is able to reduce bias amplification by 50% on two types of prediction tasks without any significant loss in prediction performance. This work was awarded the Best Paper Award at EMNLP and featured in Wired.

An ideal solution for removing bias might be to simply collect bias free datasets but such collection is impractical at large scale. To get closer to this goal, I have been exploring ways of generating automatic
modifications to training data to neutralize observable bias in data [11]. The effort involves two related avenues: (1) modifying naturally collected training data in ways that do not harm underlying task performance and (2) creating small scale diagnostics datasets that are sufficiently controlled that we can attribute variations in performance to bias. As an initial effort, we have considered co-reference resolution, the task of indicating that two phrases in text refer to the same thing. Given this task, we can show that methods preferentially associate male dominated professions (i.e. doctor) with male pronouns and similarly for female dominated professions. Furthermore, by automatically changing the gender of entities in training data (i.e. changing “mother” to “father”), such that all male entities also participate in female contexts and vice-versa, we are able to entirely eliminate such bias.

Future I plan to continue my focus on combining elements on natural language into perception. I will develop new algorithms and datasets for problems in language and vision, while carefully controlling/studying the social impacts of such approaches. There are a number of open challenges that are important to study, including:

- **Language as Scaffold for Search** Language is a tool that people use to teach and guide each other to learn new skills. I would like to explore the degree to which the same could be used to guide machines. For example, if a robot were tasked with retrieving an apple in the sink from the kitchen, it might struggle to ever guess that first step in this process to go to a kitchen. Instead, natural language directions can be used to break down this task into manageable goals: (1) go to the kitchen, (2) find the sink, (3) take the apple. This would build on existing reinforcement learning paradigms: models would combine standard exploration methods with high-level instruction following and verification of intermediate goals.

- **New Forms of Generalization** Learning methods are especially strong when the patterns required for a test examples have also been observed in a large number of training examples. Yet after demonstrating strong performance on such tasks models have proven extremely brittle. This observation motivated me to co-organize the Workshop of New Forms of Generalization to encourage discussion on how we can find new ways to challenge our models. The workshop was attended by over 200 people and I continue to be very interested in alternative setting for evaluating machine learning models. These include forcing models to produce output they have never observed during training, adversarial testing splits of data, and combining datasets to achieve domain-independent models.

- **Understanding Large Semantic Spaces** As we push machines to recognize increasingly more concepts in the world, standard assumptions about how the concepts relate to each-other will become violated. For example, learning methods commonly assume that semantic categories are mutually exclusive (i.e cat is not a dog), but in reality many concepts have relationships, such as part-of (i.e. hand and person), is-a (i.e. animal and cat) or co-occurs (i.e dolphin and water). Such relationships pose challenges for learning methods that need to be overcome to be able to recognize as much as humans can in the world.

- **Adversarial Debiasing** Model bias is generally defined with respect to some set of predictions on a dataset but this is can be brittle because a model may be coincidentally behaving well on a sample. Instead, I propose we measure bias directly with respect to model internals, so we could verify that the representations in a model are bias free. This could be done by training an auxiliary model that, given a model representation, tries to predict whether the model is aware of a protected variable in the input. Such a definition of bias may allow us to more robustly argue more that models are bias free.

The methods I am developing can be applied beyond language and vision and bias is a problem throughout all areas of machine learning. I am excited to explore such directions potentially in collaborations others.
References


