Making interaction with AI systems more natural with textual grounding

December 8 2017

Two women wearing hats covered in flowers are posing. Credit: IBM

In an upcoming oral presentation at the 2017 Neural Information Processing Systems (NIPS) Conference, our teams from the University of Illinois at Urbana-Champaign and IBM Research AI have proposed a new supervised learning algorithm to solve a well-known problem in AI called textual grounding.

Imagine you wanted to ask someone to hand you an object. You might say, "Please hand me the blue pen on the table to your left."

That's how we, humans, communicate with each other: describing scenes and objects in natural language. However, teaching an AI system to execute on this command has been challenging historically. AI systems may recognize an object such as the blue pen and the table, but may not understand which table if there is more than one. The missing puzzle piece has been how to teach a system to connect, or ground, text to an object in a given image or scene—often within a very specific region of its visual field that includes many other objects – and how to do so accurately.

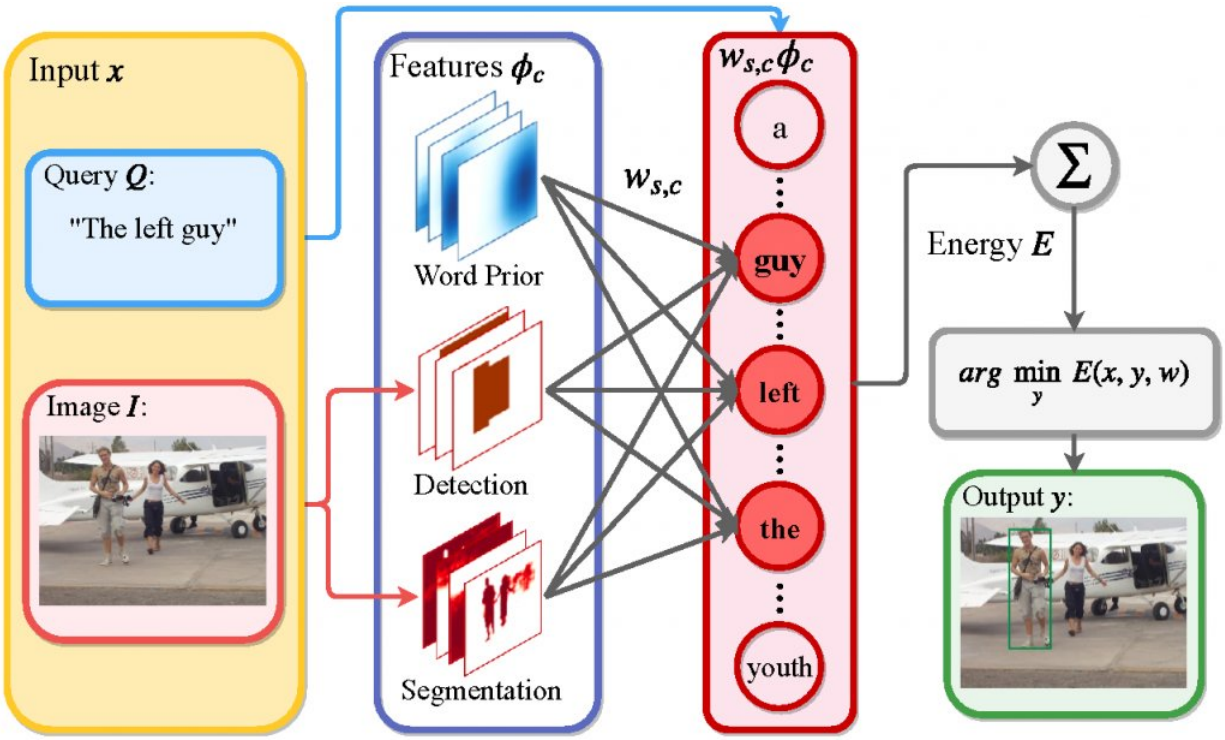
Equipped with various sensors, a machine can now easily capture details about its surroundings by recording images (or even videos) and voices. But to make sense out of these recordings for natural interaction with people, a machine needs to associate statements with images. Textual grounding solves the problem of associating text phrases (e.g., obtained from voices by a speech recognition engine) with image regions. In other words, for each named object (such as "the blue pen" and "the table to your left") from the text phrases, we need to identify a region in the image where the named object is located (so that the system knows where to get them for you).

It's easy to see that textual grounding has many potential applications. The above human interaction example is just a simple illustration.

Our algorithm achieved state-of-the-art results on two widely used datasets: 53.97 percent accuracy on the Flickr 30K Entities dataset (versus then state-of-the-art 50.89 percent), and 34.7 percent accuracy on the ReferItGame dataset (versus then state-of-the-art 26.93 percent). The figure below shows one example of the outputs of our algorithm.

What's most exciting about this research is not so much the improvement of the resulting numbers (though it's still an important metric), but the elegance of the proposed solution. The following figure shows an overview of our proposed solution.

Different from many existing deep neural network based methods, where features are extracted through end-to-end training but the meaning is hard to interpret, we propose a hybrid approach by combining a set of explicitly extracted features (we call features "score maps") and a structured support vector machine (SVM). The feature's score maps are extensible so that we can easily incorporate any new features into our algorithm. In the NIPS paper, we choose a number of easy-to-obtain features such as word priors from the input queries, region geometric preferences, and other deep neural network derived "image concepts" such as semantic segmentations, object detections and pose-estimations.



Overall model structure of the proposed solution. Credit: IBM

In most existing models, inference requires relatively straightforward matrix-vector multiplication given a set of region proposals. In our hybrid model, inference involves solving an energy minimization which searches all possible bounding boxes for the best fitting one.

To address the energy minimization problem, we adopt a subwindow search algorithm with branch-and-bound which makes the end-to-end training of our hybrid model computationally feasible (as training involves solving the energy minimization problem multiple times). We also define a proper energy function with an easy-to-compute bound on the objective function to help solve the problem efficiently and remove the need for a set of "region proposals" which are used by most existing textual grounding techniques.

We see an impact on the quality of textual grounding and also observe improved interpretability. One manifestation of the interpretability is a word embedding like representation of query words, where each embedding element is directly related to features' score maps (or image concepts) we have extracted explicitly. The usefulness of such embeddings can be illustrated by computing the cosine similarity between pairs of word vectors, which in turn shows that words close to each other are also semantically related (and grouped). For example, as shown in the following figure, because "cup," "drink" and "coffee" are semantically close to each other, their similarity in the embedding space is much higher than their similarity to other unrelated words.

Our future research plans include: (1) linking image features and words for improved interpretability, and (2) incorporating structural

information (like the structural outputs shown in this work) explicitly into the model whenever possible. We acknowledge that there have been new results on textual grounding in the literature since our initial submission of this work. We will continue our textual grounding research motivated by the goal to improve human computer interaction.

Provided by IBM