Using Syntax for Referring Expression Recognition

Volkan Cirik, Taylor Berg-Kirkpatrick, Louis-Philippe Morency
{vcirik,tberg,morency}@cs.cmu.edu
Language Technologies Institute, School of Computer Science, Carnegie Mellon University

http://github.com/volkancirik/groundnet

1 - Motivation

Referring expression recognition is the task of identifying the object in an image referred to by a natural language expression.

- What is the right way to use syntax?
- Does syntax actually help?

2 - GroundNet

- Syntax-based modular dynamic neural network approach for identifying both the target and supporting objects for referring expression recognition.
- For each instance, a computation graph of neural modules is composed based on the parse tree of the referring expression.

3 – Experimental Setup

- Google-Ref (Mao et al. 2016) benchmark consisting of 26K images with 104K annotations.
- **New annotations** for measuring the localization accuracy of supporting objects. We annotated 2400 instances where 1023 of them have at least one supporting object bounding box.

Prediction of GroundNet

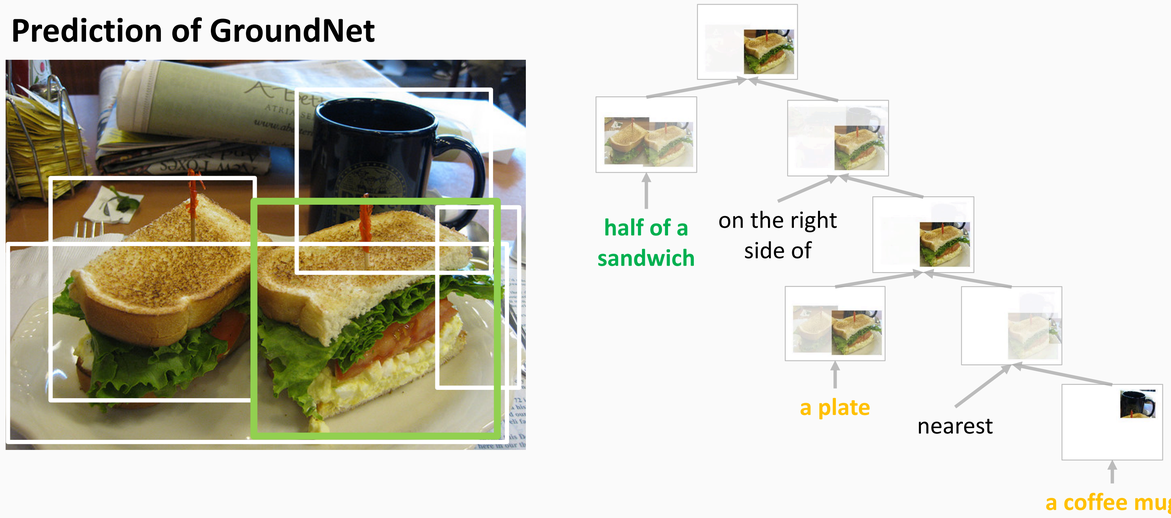
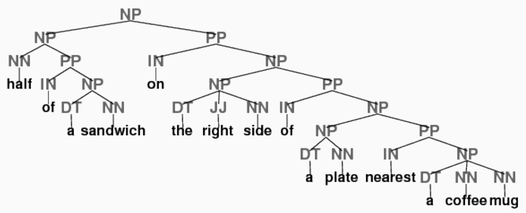
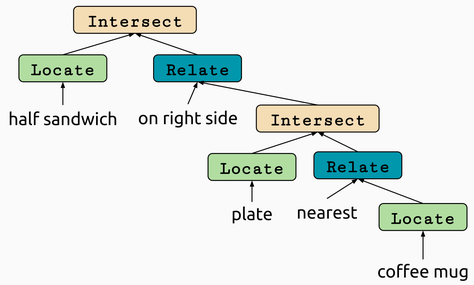


Figure 1. An example referring expression *"**half of a sandwich** on the right side of **a plate** nearest **a coffee mug**"*. GroundNet localizes both the target object (***half of a sandwich)*** and supporting objects (***a plate***, ***a coffee mug***).

Syntax → Computation Graph

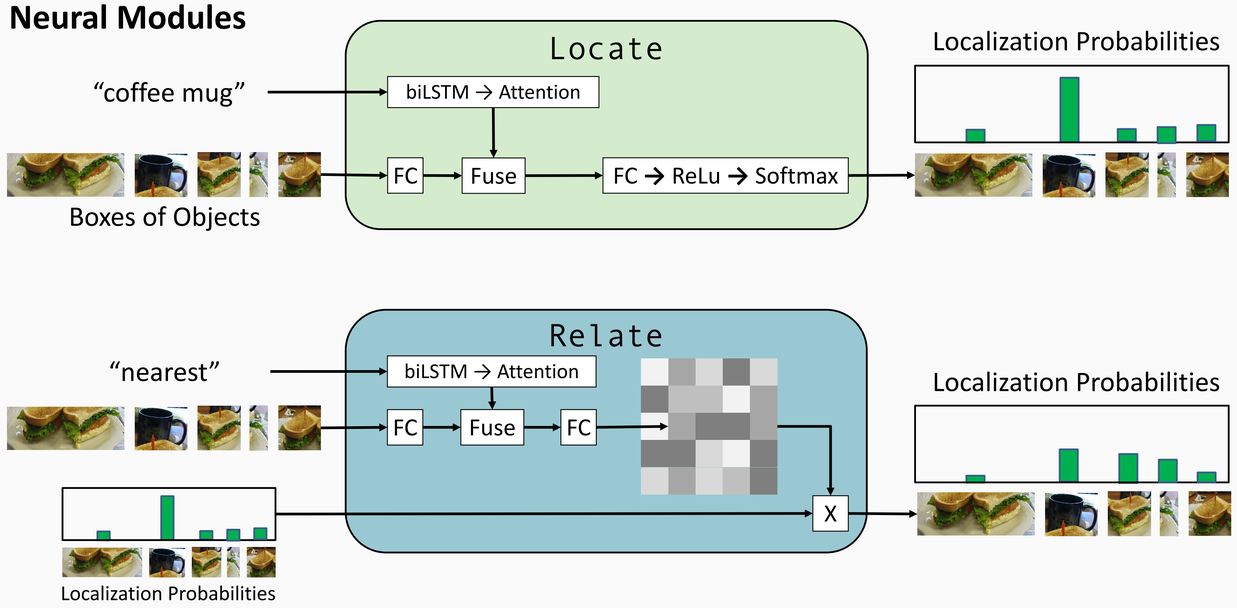


Step 1. Parsing the referring expression



Step 2. Generating a computation graph

Neural Modules



4 – Results

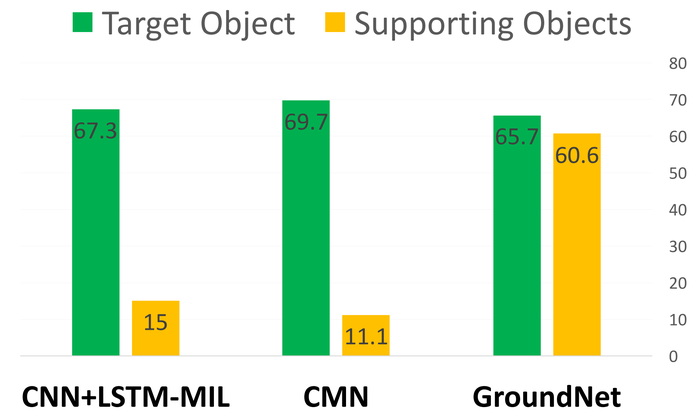


Figure 2. Localization accuracies of the state-of-the-art

GroundNet effectively integrates syntax to achieve the balance between accurately identifying both the target object and supporting objects.

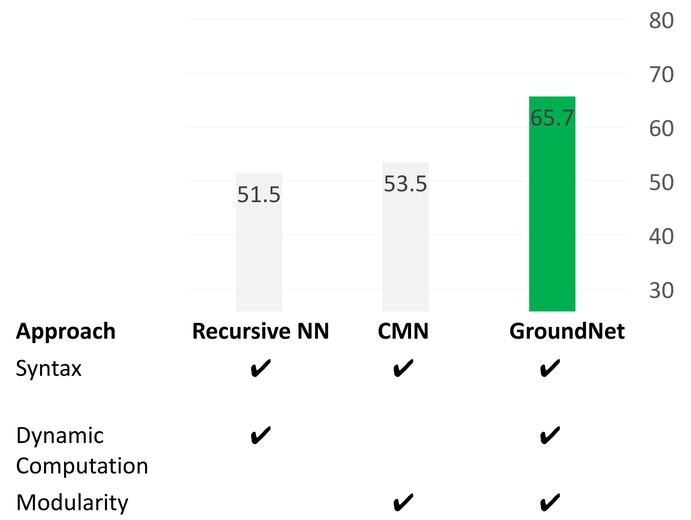


Figure 3. Localization accuracies of syntax-based approaches

Dynamic computation **and** modularity are two necessary ingredients for an accurate syntax-based model.