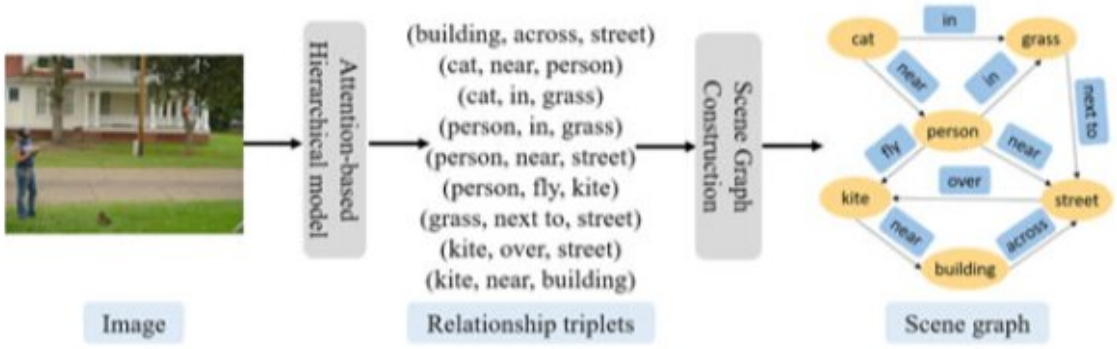
A hierarchical RNN-based model to predict scene graphs for images

April 8 2019, by Ingrid Fadelli



Overall procedure of scene graph prediction proposed in the recent paper. Credit: Gao et al.

Researchers at Shanghai University have recently developed a new approach based on recurrent neural networks (RNNs) to predict scene graphs from images. Their approach includes a model made up of two attention-based RNNs, as well as an entity localization component.

Over the past decade or so, researchers in the field of artificial intelligence (AI) have developed a variety of automatic tools to manage, analyze and retrieve digital images. To represent the content of images, traditional approaches typically use keywords or multi-view features. However, relying on either features or keywords often leads to a limited

understanding of images, failing to provide comprehensive knowledge about them.

To address these shortcomings, a few years ago, a team of researchers at Stanford University, Max Planck Institute for Informatics, Yahoo Labs and Snapchat proposed the use of a 'scene graph,' a type of data structure for describing visual concepts in an image. Scene graphs can store the description of a scene depicted in images as a structured graph in which nodes represent object information and edges provide predictions between two nodes.

These structured representations can help users to manage digital images. However, predicting a scene graph is often challenging, as it requires effective tools to recognize objects, as well as their attributes and interactions between them.

While there are several existing approaches to predict scene graphs, most of these have substantial limitations. In their study, the researchers at Shangai University set out to develop a neural network-based model to predict scene graphs from a visual attention-oriented perspective.

"A scene graph provides a powerful intermediate knowledge structure for various visual tasks, including semantic image retrieval, image captioning, and visual question answering," the researchers wrote in their paper, which was published on Wiley Online Library. "In this paper, the task of predicting a scene graph for an image is formulated as two connected problems, i.e. recognizing the relationship triplets, structured as , and constructing the scene graph from the recognized relationship triplets."

The approach devised by this team of researchers has two key components, one aimed at recognizing what they call 'relationship triplets' and the other at constructing a scene graph. To recognize

relationship triplets, the researchers used a model comprised of two attention-based RNNs in a hierarchical organization.

"The first network generates a topic vector for each relationship triplet, whereas the second network predicts each word in that relationship triplet given the topic vector," the researchers explained in their paper. "This approach successfully captures the compositional structure and contextual dependency of an image and the relationship triplets describing its scene."

Once this RNN-based model has extracted relevant information from an image, the second component of their approach uses this data to construct scene graphs. For this step, the researchers used an entity localization approach, which can determine the graph's structure using the attention information available. In addition to these two components, the researchers used an algorithm to clarify the process through which their approach converts the generated relationship triplet information into a scene graph.

Their approach was evaluated using the popular visual genome (VG) dataset and the visual relationship dataset (VRD). For the purpose of their study, the researchers annotated the images in these datasets with a set of triplets, labeling each subject and object pair with location information.

"The results of experiments on two popular datasets demonstrate that the hierarchical recurrent approach from the visual-attention-oriented perspective inside our model has a distinct improvement in results over baseline models," the researchers wrote. "In future work, we plan to enrich the scene graph with high-level semantics and more diversified attributes."

More information: Wenjing Gao et al. A hierarchical recurrent

approach to predict scene graphs from a visual-attention-oriented perspective, *Computational Intelligence* (2019). DOI: 10.1111/coin.12202

Justin Johnson et al. Image retrieval using scene graphs, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015). DOI: 10.1109/CVPR.2015.7298990