Frame Annotation for Images

Daniel Lee, Julie Song, Yichen Yao, Jung Ho Yoon

Columbia University, Fu Foundation School of Engineering and Applied Science,
Computer Science Department

Abstract

Currently, it is possible to parse a sentence into semantic frames which can capture the sentence's semantic meaning. However, what if we wanted to capture the semantic frames of a visual scene, and not just a sentence? Image captioning models do this implicitly, as they understand the relationship between objects seen in an image and relate them using an appropriate verb. However, this project, *FrameNet Annotator for Images,* attempts to take in an image and explicitly label its semantic frames.

Introduction

Our project, *Frame Annotation for Images,* can take a simple image in, and label that image with its frame semantic parsing. To do this, we take in an image and use the CNN-RNN Image Captioning model to caption that image; we then feed that caption into SLING, a framework that can parse semantic frames. Simultaneously, we run ImageAI's object detector model on the input image to get all the objects and their positions. Then we run an NLTK word similarity algorithm to connect the image captions and semantic frames to the objects and their positions. Using this, we can output an image with labeled semantic frames.

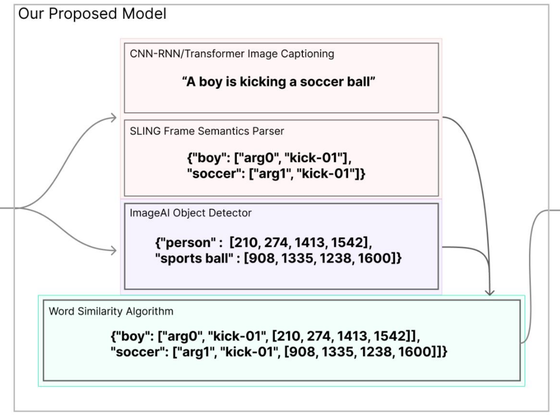
Method

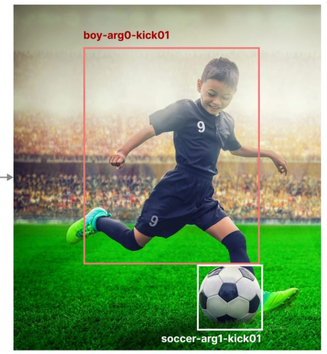
1. Image Captioning

- CNN for image encoding (InceptionV3)
- GRU or Transformer for caption generation resulted in similar description
- CNN-RNN architecture chosen for this demo

2. Frame Semantics Parser

- Used SLING, a frame semantics parser by Ringgaard et al. (2017)
- Takes text as input, and outputs a graph representation of the sentence, similar to those of AMR
- An older model, called "SEMPAR," was used instead of the current "CASPAR" model, as "CASPAR" focused on maximizing metric performance by dropping support for PropBank Framesets
- Structured sentence elements not fully accessible to SLING user; string parsing with regular expressions was therefore needed





3. Object Detector

- Used ImageAI model for object detection
- Perform object detection using YOLOv3 pre-trained model trained on the MS COCO dataset
- Returns objects, object positions, and percent probability certainty of detection

4. Word Similarity

- Used NLTK WordNet Wu-Palmer Similarity: which gives score denoting how similar two word senses are, based on the depth of the two senses
- Match similarity between image caption / semantic frame to object / object position

Results + Discussion

Figure 1: The semantic frames are labeled correctly.

Figure 2: The object detection model missed the "man-arg0", so the man who is holding the wine glass is not labeled.

Figure 3: The object detection model caught another person in the image, but not the main person eating, so the "person" is labeled in the wrong position.

Figure 4: The image captioning model mislabeled a "computer" as a "television"

Conclusion

Frame Annotation for Images can take in a simple image as its input and output an image with its semantic frames, allowing for easy visualization of the image's semantic meaning. We achieved this by putting together multiple off-the-shelf models.

Future Work

We faced a compatibility issue between the different modules used. Therefore, one future work involves addressing these environment discrepancies.

As seen in our results, bottlenecks were present at both the image captioning and the object detector. With a better model for both, the proposed pipeline will be able to better label the semantics of a given image. However, for the image captioning model, we faced some hardware constraints because the cloud GPU wasn't fast enough to train the large dataset. Using RTX 3080 allowed for no dropout. Hence, we faced some overfitting issue early in the training. Therefore, a better way to regularize training is necessary for the future work. Also, during the training, we realized that the model predicted no-verb captions. This is because the dataset contains no-verb captions. Future work could clean the dataset and remove captions with no verb.