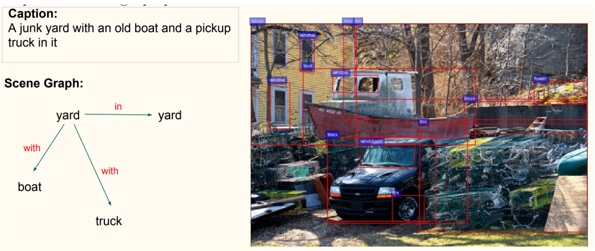
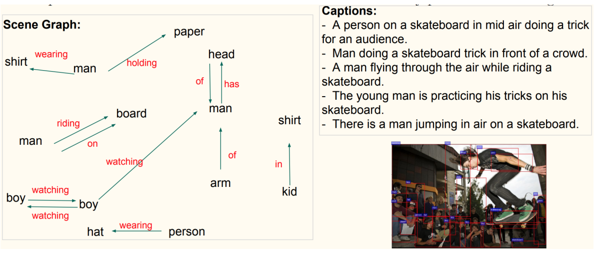
A Appendix

A.1 Failure Cases when Relying Solely on Captions



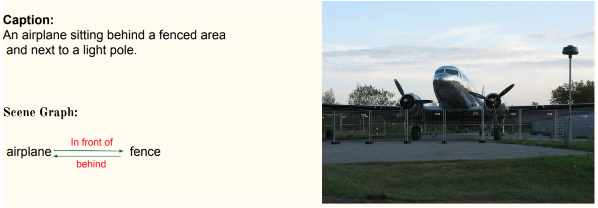
Caption:
A junk yard with an old boat and a pickup truck in it

Figure 3: Failure case where the scene graph parser makes errors



Captions:
- A person on a skateboard in mid air doing a trick for an audience.
- Man doing a skateboard trick in front of a crowd.
- A man flying through the air while riding a skateboard.
- The young man is practicing his tricks on his skateboard.
- There is a man jumping in air on a skateboard.

Figure 4: Failure case where all captions are insufficiently descriptive



Caption:
An airplane sitting behind a fenced area and next to a light pole.

Figure 5: Failure case where the caption does not capture all relationships

In this section, we identify the key failure cases when relying solely on captions. These failures are primarily due to scene graph parser errors, insufficient information present in captions, and the inability of captions to capture all relationships present in the image.

A.1.1 Scene Graph Parser Errors

Generally, the scene graph parser is as effective as using human-constructed scene graphs (Schuster et al., 2015). However, there exist cases where the scene graph generated from the caption by the Stanford Scene Graph Parser is incorrect. For instance, in Figure 3, the parser yields two "yard" nodes. However, we observe that the majority of the errors are caused by the two subsequently described issues.

A.1.2 Insufficient Caption Information

We find that captions describe far less information than actually present in the image. For example, in Figure 4, though there are multiple objects and relations in the image, none of the five captions are able to completely capture everything in the image.

A.1.3 Unable to Capture All Relationships

We find that captions don't adequately capture all relationships. For example, there are multiple relations such as beneath-behind and up-upward that are not correctly captured. In other cases, some relations actually present in the image are missing– one such example the inability to capture transitive relations. For example, in Figure 5, while the caption indicates that the light pole is next to the airplane, and that the airplane is behind a fence, the caption fails to capture the transitivity (i.e., that the light pole is behind the fence).

A.2 Correlation between Captions and Ground Truth

Our model formulation aims to pool information about subject-predicate-object triplets from the entire corpus of captions, and to use it to densely identify relations between entities in a single image. To validate whether the most common ground truth relation classes are actually present in the captions, we use the Stanford Scene Graph Parser to extract the predicates and compare their frequency counts with the ground truth relations. This correlation is demonstrated in Fig 6.

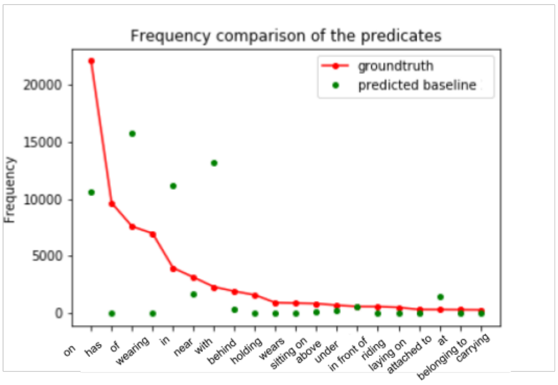


Figure 6: Correlation between the ground truth triplets and the triplets present in captions

A.3 Failure to Ground Cluttered Scenes

One failure case for C-GEARD is shown in Figure 7. The main reason for this is the large num-

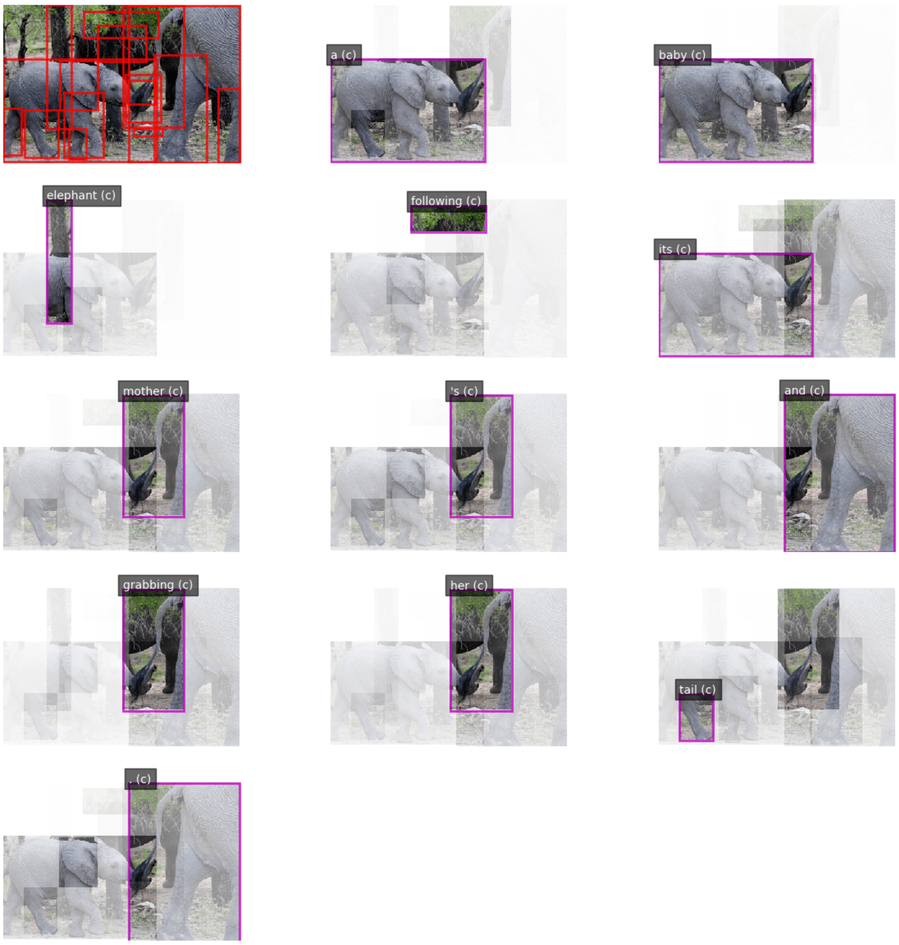


Figure 7: Examples of failures of entity and relation groundings generated by C-GEARD

ber of ground truth bounding boxes present in the image, which led to the model being unable to correctly capture the groundings.

A.4 Importance of ResNet Features

We tried two variants of extracting ResNet features given ground-truth bounding boxes. In the first, we used a fully convolutional approach, using the original object sizes. However, we observed extremely poor performance, and hypothesize that classification networks trained on ImageNet are tuned to ignore small objects. To resolve this, we resized objects so that their larger side is of size 224. We observed significantly better performance; consequently, all reported numbers use these features.

To validate that the benefits observed were due to the changed object feature representations, we trained a simple classifier using the 50 VG object classes with a linear layer (we tried other variants as well, but all other results obtained were comparable). We observed a substantial difference in the performance between these two variants: 45% accuracy vs 54% respectively.

A.5 Hyperparameters

We train the top-down attention model with entity attention dimension of 512, tanh non-linearity and batch size of 100. We used both the language model LSTM and the attention LSTM with 1000 hidden cells. The ResNet extracted object features were 2048 dimensional and the word embeddings were initialized to FastText embeddings of 300 dimensions. Finally, we train our model using an Adam optimizer and a learning rate of 0.0001 for 75 epochs. We train the relation classifier using a simple MLP with 2 hidden layers of 64 units each, with dropout of 0.5 using Adam optimizer and learning rate of 0.001 for 50 epochs.