Object Localization via Natural Language Expression

Alp Aribal
TUM Informatics
alp.aribal@tum.de

Hakan Sirin
TUM Informatics
hakan.sirin@tum.de

Abstract

On top of related work, we designed a novel model that generates two embeddings, one for the image and one for the query, and calculates a similarity score between them. Our model takes an image, a bounding box inside the image and an expression as input, and outputs a similarity score between the expression and the object inside the bounding box. In combination with a bounding box generation scheme, our model is capable of localizing object described by free text.

1. Introduction

In this project, we address the problem of object localization using queries in natural language. The problem can be defined as identifying a certain object inside the image by using a natural language description. An example of a natural language query could be "white car on the right". It is important the note that the descriptions are not restricted in any way. Hence, the structure, content or the spelling of these descriptions are controlled.

This task differs from classical localization tasks as the number of classification categories are unknown and the network additionally needs to interpret relational words in natural language such as "next to" or "on the right".

2. Related Work

[3] takes an image, a text query, and a set of candidate bounding boxes (generated by EdgeBox [5]) as input, and returns a score for each candidate. Features are extracted from both the whole image and the candidate using two different CNNs. Three LSTMs are used in order to capture local (candidate), global (whole image) and language (query) aspects of the problem. In the end, the probability distribution of the next word in the query is calculated with the input of the image, the candidate part of the image, spatial information of the candidate and the text query so far. The final likelihood of the candidate is defined as the multiplication of the probability of each individual word given the same conditions and also all previous words.

[2] incorporates a fully convolutional approach. Spatial features extracted from the image and an encoded expression generated from the query are processed by a fully convolutional classification layer and a segmentation output is generated where, instead of a bounding box, the relevant object mask is predicted.

3. Our Work

3.1. Model Inputs and Outputs

The model takes as input an image, a bounding box, and a natural language expression which is pre-encoded using the universal sentence encoder [1]. The image and the image region inside the bounding box is encoded into an image embedding. The NL expression is similarly encoded into a query embedding. These two embeddings are forced to lie on the same space. Model returns the cosine similarity of these two embeddings.

At training time, our ground truths are the Generalized Intersection over Union (GIoU) [4], and the model is trained with a hinge loss.

At test time, prior to model forward pass, candidate bonding boxes are generated using the Edge boxes method [5]. Then, each candidate box is passed into the forward pass and cosine similarity score is calculated. The candidate box with the highest similarity score is returned as predicted bounding box.

3.2. Model Structure

The model constitutes of two legs. One leg of the network generates the image embedding. To do so, first, features for the full image and the cropped image are extracted using pre-trained VGG16s. These features are stacked after a fully connected layer. A second fully connected layer is applied to decrease the dimensions. The resultant vector is stacked with the normalized bounding box coordinates. These coordinates are important for distinguishing directions and locations, e.g. left, right, below. The stacked vector is passed through a fully connected network to produce the image embedding.

The second uses a fully connected network very similar to the one in the first leg in order to project the encoded sentence into the shared space. This new vector is called the query embedding.

An important distinction of our model from a siamese network is that we use different fully connected networks in the two legs, as the inputs of the legs lie in different spaces.

The two embeddings generated by the two legs are compared using cosine similarity. Figure 1 demonstrates the model structure.

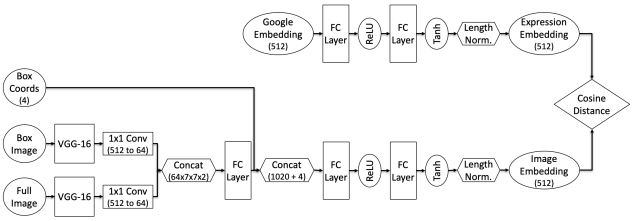


Figure 1: Our model structure

4. Results

As can be seen in Figure 2, the model was trained until the validation loss stagnated.

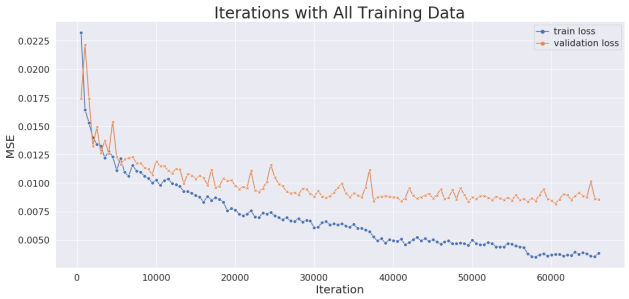
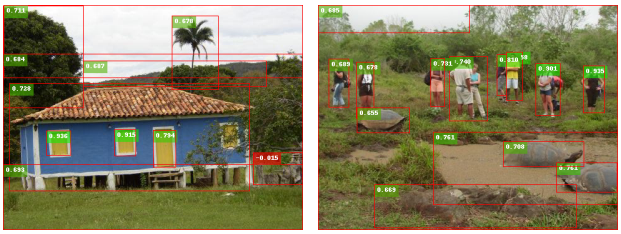


Figure 2: Training and validation loss of the model

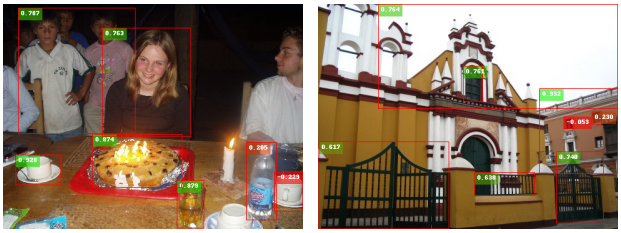
It was observed that the model is highly successful in understanding words that signify location such as left, right, top, and bottom. Figure 3 demonstrates some correct examples.

On the other hand, in complex queries where the object is described relative to another object, model accuracy drops. In some cases, high scores are returned for all items, showing a lack of distinction between them. It was also observed that the accuracy of the candidate boxes plays an important role in model accuracy. Model tends to predict lower scores when the bounding box covers more than just the object. Sometimes the correct object is just not identified. Figure 4 demonstrates some incorrect examples.

When compared to the reference work [3], the accuracy of our model is lower. Table 1 summarizes the accuracy of

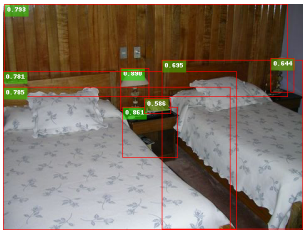


'yellow window left side' 'person farthest right'

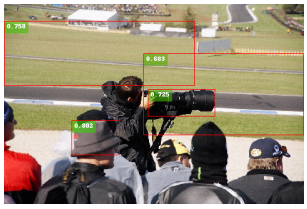


'cup left center' 'building right'

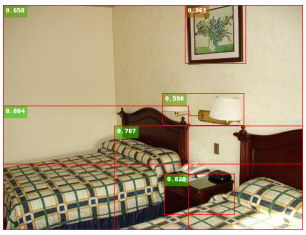
Figure 3: Successful examples where the object is correctly identified



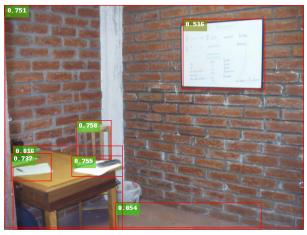
'lamp on right nightstand'



'camera'



'lamp'



'board'

Figure 4: Unsuccessful examples where an incorrect object is identified

the two models. For our model, accuracy values for different IoU thresholds are given. The accuracy in the table 1 is the percentage of cases where the top-1 scoring bounding box has bigger IoU with the ground truth object region than the selected threshold. We took this methodology also from the reference work [3] to get comparable results.

We further analyzed the learned shared embedding space in between different type of objects by plotting the embeddings to a lower dimension using a PCA in Figure 5. When the largest classes are analyzed in isolation, clear distinctions between the classes can be seen. Furthermore, similar

Table 1: Comparison of our model with reference work

IoU Threshold	Our Work	Reference Work [3]
30%	63.7	-
50%	61.1	-
70%	60.5	-
100%	60.0	72.7

classes tend to lie in similar regions. The image embeddings and query embeddings are seen to be closely located in this shared space. These findings validate that the model has indeed learned to match images and queries. It is important to note that we utilize a lot more than 2 dimensions and the first two principal components only capture 40 percent of the variance of the embeddings.

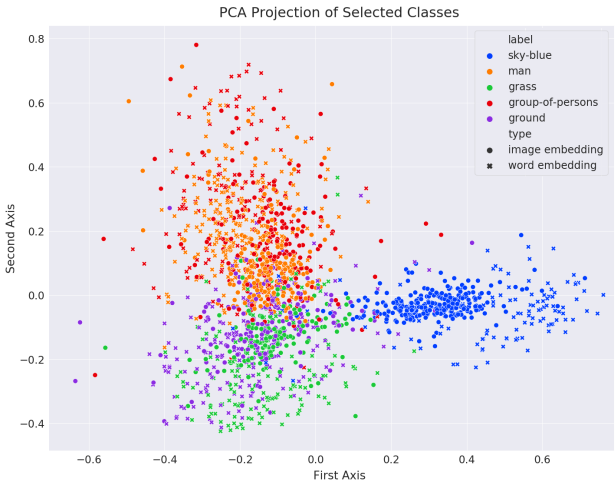


Figure 5: 2D projections of images and queries from the largest classes

5. References

References

[1] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Lyn Untalan Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cspedes, Steve Yuan, Chris Tar, Yun hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. In *In submission to: EMNLP demonstration*, Brussels, Belgium, 2018. In submission.

[2] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016.

[3] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016.

[4] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union. June 2019.

[5] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.