Object Localization in Images Using Natural Language Query

Subhashree Radhakrishnan [PID: subha]

ABSTRACT

Image retrieval is an important problem in computer vision where the system automatically returns an image that is closely related to the query. Referral expression in particular points out to a person or object in an image with the help of spatial coordinates and visual attributes. This project aims to localize the query through a bounding box in an image based on a referral expression. Towards this end a new multi-modal embedding architecture is proposed that maps the encoded image and text features to a common space where the positive pair of bounding box and the query are closer. In this work, a MLP and an order embedding is learnt to map a natural language query and a bounding box in an image. Several experiments are conducted to encode the text features such as LSTM, word2vec and skip thought vectors and the performance of the model to retrieve boxes is evaluated. The quantitative and qualitative results are observed and reported.

KEYWORDS

Referral Expressions, Image Retrieval, Multi-modal Embedding

ACM Reference Format:
Subhashree Radhakrishnan [PID: subha]. 2017. Object Localization in Images Using Natural Language Query. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

One of the seminal applications in computer vision and machine learning is Content based image retrieval where the image is retrieved from a database based on content image or query. This is especially important in search engines and recommender systems in supermarkets etc. Recent works leveraging deep architectures for image retrieval are mostly limited to using a pre-trained network as local feature extractor. Most eïňĂorts have been devoted towards designing image representations suitable for image retrieval on top of those features. Localization in images using Referral expressions is a constrained problem of Image Retrieval where the particular object/ person being referred to is localized within an image. This especially has applications in digital editing through queries, instructions to robots for directing the robot to perform actions through its camera sensor.

Image Retrieval. Image retrieval is the method of finding the closest image to a given textual query by it's semantic meaning. Generally the methods used for ranking in retrieval are, a ranking function is learned through a recurrent neural network [10], metric learning [4], correlation analysis [8] and other methods [3, 7]. It was shown in [2] that a probabilistic image captioning model such as LRCN can also be used as an image retriever by using the probability of the query text sequence conditioned on the image generated by image captioning model as a score for retrieval.

Image Captioning. Image captioning methods take an input image and generate a text caption describing it. Recently, methods based on recurrent neural networks [10, 17] have shown to be effective on this task. LRCN [2] is one of these recent successful methods and involves a two-layer LSTM network with the embedded word sequence and image features as input at each time step. Image captioning can be thought of as the reverse process of Referral Expression based localization in Images.

2 RELATED WORK

One of the initial works in natural language object retrieval was by Hu et al.[6] where a Spatial Context Recurrent ConvNet (SCRC) model is proposed to score the offline generated proposals that encodes both the spatial and global image features. This work was followed by [9] which additionally considers context regions by penalizing if the same expression is generated from two proposals. This work was further improvised by [19] to encode relative attributes where the average over pooled difference in CNN features between target and context proposal is considered along with spatial coordinates of target region. [12] aims in generating two regions as output by maximizing likelihood conditioned on two regions. [5] adds an additional relationship module to their work in[6] and propose a Compositional Modular Network which grounds subject, object and the relationship of the query in images. Video Search/ Video Retrieval is a closely associated problem to ours in that it returns top ranked videos based on a search query and addresses the question 'when'. The seminal work in video retrieval was [16] that proposed Video Google which retrieves a video of a given object input using a bag of words model.[13] performed video retrieval by first retrieving web images based on the query and projected images, videos and sentence features onto an embedding.

3 DATASETS

In this work, Refer It COCO dataset is used which was built on MSCOCO by additionally annotating with referring expressions [18]. This dataset was collected in a two-player game, where the ïňĄrst player writes a referring expression. The second player is shown only the image and expression and has to click on the correct object described by the expression. If the click lies within the target object region, both sides get points and their roles switch. Using the same game interace, the authors further collected RefCOCO

Conference'17, July 2017, Washington, DC, USA
© 2017 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and RefCOCO+ datasets on MSCOCO images[31]. The RefCOCO and RefCOCO+ datasets each contain 50,000 referred objects with 3 referring expressions on average. The dataset contains a total of 230k annotated descripions. The training and the test

4 METHOD AND ARCHITECTURE

The method employs a multi-modal embedding to map the visual and text features to low dimensional embedding space where the corresponding image, the bounding box in the image and the referral expression is close. The base architecture used is shown in **??**. As shown in the figure, the network has two modules which extract the visual features and the text features respectively. The visual features are extracted by using pre-trained VGG and the out feature vector is of size 4096. The network used was VGG-16 net [15] trained on ILSVRC-2012 dataset [14] .The hyperparameters chosen were dropout of 0.5, learning rate of 0.1 and nesterov momentum of 0.9. These features were concatenated with the ground truth bounding box coordinates. The coordinates are features representative the location of the object/ person referred to in the image. These features were concatenated with the CNN VGG features extracted with the same CNN architecture mentioned from the entire image to which the bounding box image corresponds to. This is required for the contextual information around a referred object or person. The final feature size of visual features is 21988. The model was trained for 30 epochs. The visual features are projected to the same size as textual features by passing through a fully connected layer.

The text features are extracted using three methods namely skip-thought, word2vec and GLOVE. Both these features are passed through a siamese network [1]. Simaese network consists of two identical networks. The advantage of siamese networks is it learns to differentiate between two given inputs instead of classifying the input. This is advantageous because the system should not only learn the correspondance between the right bounding box image and the query but should also learn that the negative examples of bounding boxes in the same image do not belong to the given query. Finally the embedding was trained with contrastive loss. The initial model was overfitting and hence a dropout of 0.2 was used.

As it can be seen in 3, the overall contrastive loss decreases as the training proceeds and after few epochs.

5 EXPERIMENTS

Data Preparation. The ReferIt COCO dataset had the image, bounding box and the corresponding query annotation. The total number of training samples were 99179. A train/test split of 80 - 20% was used. Since the siamese network requires positive and negative pairs, I had manually generated the negative examples. Every query was set with a random bounding box annotation from the same image to which the query belonged to. Every positive pair was assigned a label 1 and negative pair was assigned a label 0.

5.1 Different Text Encoding Methods

Skip-thought. The skip-thought model is trained to reconstruct the surrounding sentences to map sentences that share semantic and syntactic properties into similar vectors. The skip-thought model was used to encode the sentences directly to 40988 vector.

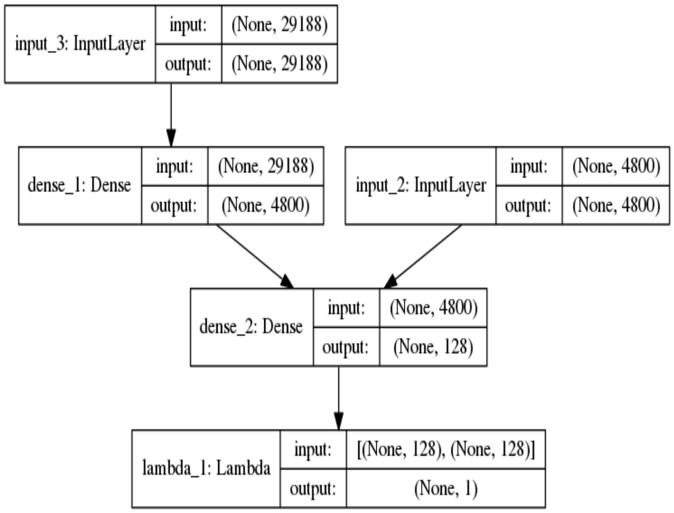


Figure 1: The Model Architecture of siamese network to learn embedding between image and query features.

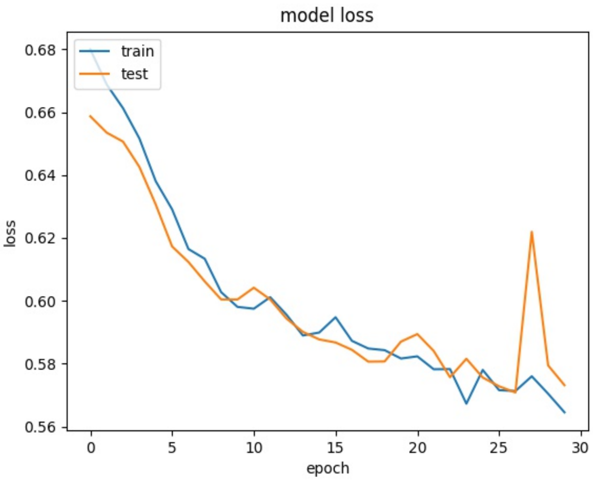


Figure 2: The model Loss plot of Siamese Network with contrastive loss.

Glove. Glove is a count based model for encoding sentences. They first construct a large matrix of co-occurrence information, i.e. for each "word" (the rows), it counts how frequently we see this word in some context in a large corpus. It uses the hidden representation of the LSTM units to encode the sentence vector. The number of hidden units used were 512*2*2. Glove is an unsupervised technique that learns a lower representation through reconstruction loss.

Word2vec. Word2vec is a predictive model that learns a vector representation of the sentence by learning to predict the context/target words. It is a feed forward network that is optimized using gradient descent.

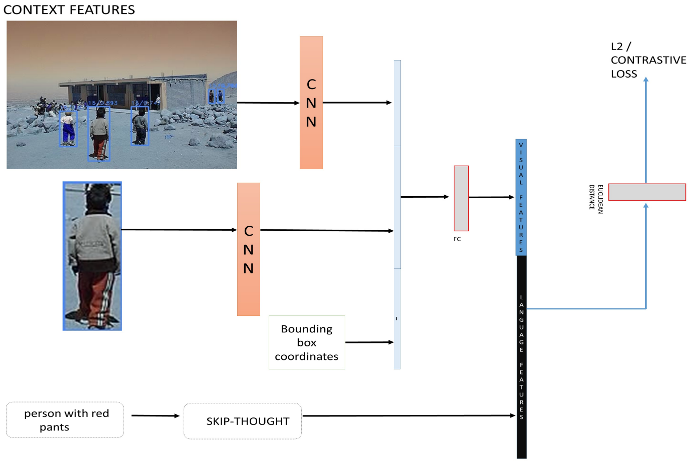


Figure 3: The overview figure for training for image based localization using queries.

5.2 Different Losses

The two losses that will be compared are Contrastive loss and L2 loss.

contrastive loss. The contrastive loss decreases the distance between a positive example of query and bounding-box and increase the distance between a negative pair of query and bounding box in the same image. The output distance of the Fully connected layer is directly used for L2 loss.

$$L_{contrastive} = \frac{1}{N}[(1 - y_{true})(D_s) + y_{true}(max(0, D_d))]$$

In the above equation, Ds is the Distance between positive pair of ground truth human tube and referral expression from same image and Dd is the Distance between negative pair of human tube and referral expression from same image. Ytrue is 1 if positive pir and 0 if negative pair.

L2 loss. L2 loss is the Euclidean distance between the image and query features after being forwarded through siamese network, that is output from the model and that value is directly backpropogated.

5.3 Different Embedding

The order embedding [11] was additionally experimented instead of MLP and it resulted in improvement of performance. This embedding learns an ordered representation and is the state of the art in the image-caption retrieval problem.
Testing During testing, an offline method of single shot detector was used that generated bounding boxes for the objects present in the image. Every box will was scored for its distance with the query provided using our trained models. The bounding box image with smallest distance was retrieved.

6 RESULTS AND EVALUATION

The observation of the various experiments conducted are summarized in 1. The reported metrics for the retrieval is Recall@1. This metric measures if atleast one of the retrieved bounding boxes is right. A box is assigned positive label if the distance with the query is less than 0.5. Thus during testing, the visual and the query features are extracted separately. Then the trained embedding is used to compute the distance between the query and the candidate bounding boxes generated by SSD.

Quantitative. In this we consider the application that retrieves the relevant images given a natural language query. The metric used was Recall@1. As it can be seen from the results, in general contrastive loss performs better than the L2 loss. This is likely because contrastive loss is a better regularizer and additionally trains the model to increase the distance between negative examples. The skip-thought vector outperformed the other two embeddings and this is because the skip thought model uses an encoder decoder to learn representations of sentences instead of being trained on individual words. Hence the skip thought vector learns a representation that has more contextual information. Glove and Word2vec performed almost similar because both of them learn a word based representation considering few context words. The only difference is word2vec is predictive model and glove is an unsupervised technique.

Context. The context features include concatenated CNN features of the whole image. This information is especially useful when the query comprises of information of the spatial position of the person or the object for e.g. The person on the left side. Thus context features give the necessary global information. When these features were removed during training of MLP embedding, the recall dropped to 54.83 as expected.

Qualitative. The visualization of the retrieved bounding box for the given query using our trained model is shown in 5 and 5. The model localizes queries spatially accurately. However there were few failure cases wherein the SSD failed to predict few classes that were present in the query. This could be attributed to the fact that

Visual Features	Text Features	Loss	Recall@1
CNN	Skip-Thought	Contrastive Loss	76.57
CNN	Skip-Thought	Contrastive Loss(without context features)	54.83
CNN	Skip-Thought	L2	49.4
CNN	Glove	Contrastive Loss	66.7
CNN	Glove	L2	37.8
CNN	word2vec	Contrastive Loss	68.3
CNN	word2vec	L2	42.8
CNN	Skip-Thought	Order Embedding	79.9

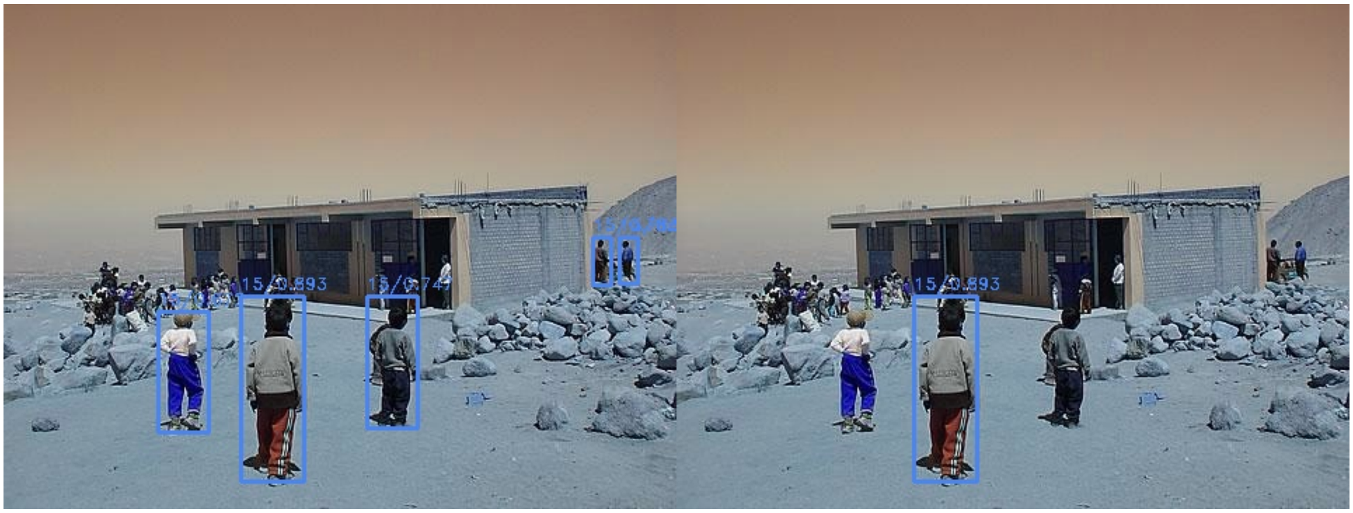
Table 1: Recall of Bounding box retrieval in an image given a query

SSD was trained separately. Similarly, the bounding box was not regressed when identifying multiple people/ objects.

7 CONCLUSION AND FUTURE SCOPE

Thus, a Convolutional-recurrent based model was proposed and implemented that maps text and image features to an embedding.The model learns a common representation between images and query and inturn localizes the referred object/ person of the query in the image. The experiments conducted gives a comprehensive analysis of the performance of word2vec, glove and skip-thought vectors in their effectiveness in encoding a sentence. It was observed that skip-thought vectors performed well compared to other models. Also, it was observed that the context features(the CNN features of the entire image) helped in the retrieval of the bounding box when concatenated with the visual features. Thus a model that learns an embedding space for text and image was implemented which additionally learns the spatial location in the image. The reported maximum recall@1 is 79.9.

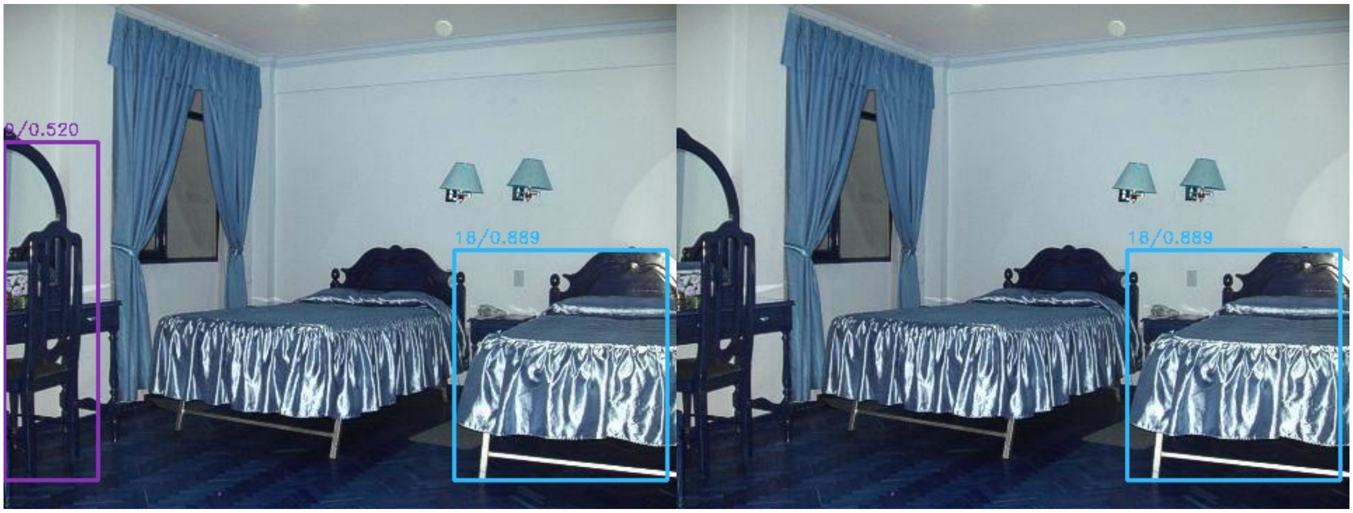
As future scope, the training can be done end to end wherein the visual features and the query features are not encoded offline. Thus the CNN and the Skip-thought model could be trained simultaneously along with the MLP reducing the distance between the query and the corresponding bounding box. This way the skip-thought vector would encode features unique to the individual bounding box, and the CNN would generate image features in accordance with the query. When the model is trained end to end, The visual features for the embedding can be encoded in an unsupervised way using an autoencoder. Since, the SSD could not identify all objects as it is pretrained separately, the SSD bounding box generation should be included in the end to end architecture instead of generating them offline. This way the parameters of the SSD will learn to regress bounding boxes according to query.



(a) Original Image after SSD Detection

(b) Localized Image

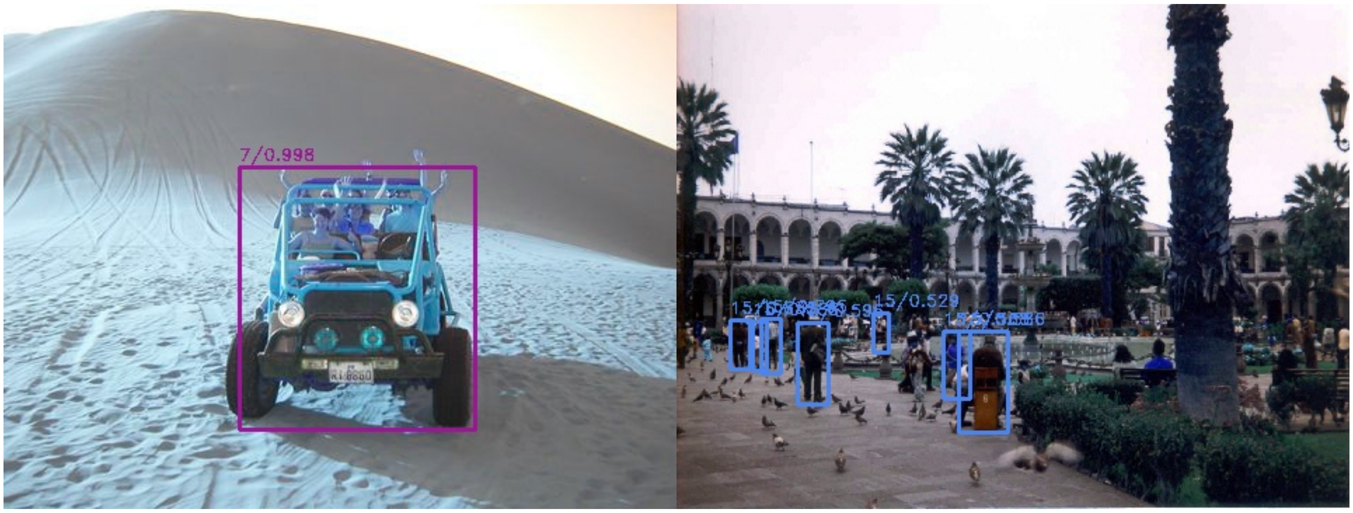
Figure 4: Success Case Query: Man with Red Pants



(a) Original Image after SSD Detection

(b) Localized Image

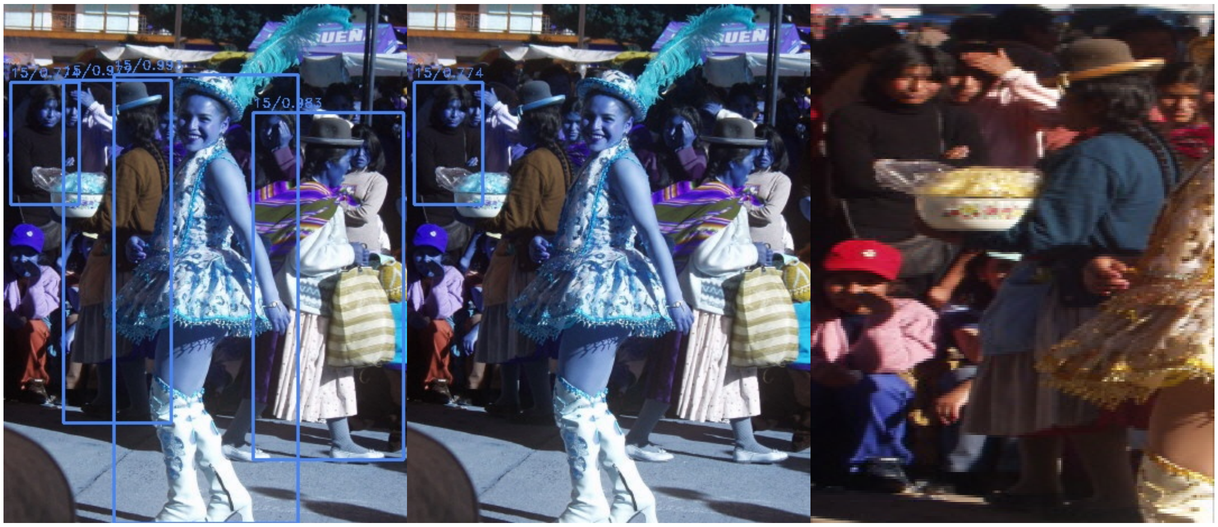
Figure 5: Success case Query: Bed on Left



(a) Query: Mountain on Left - Mountain Not Detected

(b) Right Palm Tree- Tree not detected

Figure 6: Failure Case as SSD did not identify the box



(a) Original Image after SSD Detection

(b) Localized Image by model

(c) Actual Ground Truth

Figure 7: Failure Case as could not localize when referred to multiple objects

REFERENCES

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*. Springer, 850–865.

[2] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.

[3] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129.

[4] Steven CH Hoi, Wei Liu, Michael R Lyu, and Wei-Ying Ma. 2006. Learning distance metrics with contextual constraints for image retrieval. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, Vol. 2. IEEE, 2072–2078.

[5] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2016. Modeling relationships in referential expressions with compositional modular networks. *arXiv preprint arXiv:1611.09978* (2016).

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

[7] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[8] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2014. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399* (2014).

[9] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 11–20.

[10] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632* (2014).

[11] Brian McFee and Gert Lanckriet. 2009. Partial order embedding with multiple kernels. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 721–728.

[12] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*. Springer, 792–807.

[13] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. 2016. Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision*. Springer, 651–667.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[15] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[16] Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *null*. IEEE, 1470.

[17] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. 2048–2057.

[18] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. 2015. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2461–2469.

[19] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*. Springer, 69–85.