A Robust Approach to Open Vocabulary Image Retrieval with Deep Convolutional Neural Networks and Transfer Learning

Vishakh Padmakumar[1], Rishab Ranga[2], Srivalya Elluru[3] and Sowmya Kamath S[4]
Department of Information Technology,
National Institute of Technology Karnataka, Surathkal, Mangalore 575025 INDIA
{vishakhpadmakumar, rishab.ranga1996, srivalya.elluru}@gmail.com, sowmyakamath@nitk.edu.in

***Abstract*—Enabling computer systems to respond to conversational human language is a challenging problem with wide-ranging applications in the field of robotics and human computer interaction. Specifically, in image searches, humans tend to describe objects in fine-grained detail like color or company, for which conventional retrieval algorithms have shown poor performance. In this paper, a novel approach for open vocabulary image retrieval, capable of selecting the correct candidate image from among a set of distractions given a query in natural language form, is presented. Our methodology focuses on generating a robust set of image-text projections capable of accurately representing any image, with an objective of achieving high recall. To this end, an ensemble of classifiers is trained on ImageNet for representing high-resolution objects, Cifar 100 for smaller resolution images of objects and Caltech 256 for challenging views of everyday objects, for generating category-based projections. In addition to category based projections, we also make use of an image captioning model trained on MS COCO and Google Image Search (GISS) to capture additional semantic/latent information about the candidate images. To facilitate image retrieval, the natural language query and projection results are converted to a common vector representation using word embeddings, with which query-image similarity is computed. The proposed model when benchmarked on the RefCoco dataset, achieved an accuracy of 68.8%, while retrieving semantically meaningful candidate images.**

I. Introduction

The power of sight and visual recognition is an integral ability of human beings. Over the past decade, researchers have tried programming this ability into machines, making object recognition an crucial task in computer vision applications, specifically robotics. Another ability which is inherent in human beings is the ability to communicate and process language. The problem of making machines understand conversational natural language and perceive objects around them when described, is a core aspect of artificial intelligence. This has significant application in real-world systems like self-driven cars which perceive their surroundings to navigate safely and robot assistants capable of performing simple tasks.

The Open Vocabulary Object Retrieval (OVOR) problem tries to address both these issues, firstly, understanding a query which is in natural language form and secondly, retrieving the most appropriate image for the given query using latent image information extracted using computer vision techniques. When people talk about an object, they use rich and descriptive natural language, instead of a merely using basic nouns. For instance, while referring to a box, humnas often include rich context information like "*red cereal box*" while referring to it. Thus, image retrieval techniques that simply use basic grouping objects into categories often fail, as the context information is not considered. Hence, more robust techniques that can capture inherent semantic information like '*red*' and '*cereal*' in the given query to retrieve most appropriate candidate images become crucial.

In this paper, we propose a modular Open Vocabulary Object Retrieval approach, with five image-text projections each aimed at extracting one particular kind of information from the candidate images. The basic premise for dealing with the wide scope for ambiguity in natural language querying is the assumption that priority is to be given to the nouns, verbs and adjectives in the given natural language query. Thus, each of our projections is tailored to capture one of these parts from the natural language query. We define three category based projections to extract nouns or objects in the images and two instance based projections for verbs and adjectives. These are selected in such a manner so as to ensure maximum coverage of all potentially available information from the image as people may choose to describe the image in many different ways. The rest of this paper is organized as follows: Section II presents a discussion on existing research works in the field. Section III presents an in depth look at the proposed methodology for obtaining each individual image-text projection for the purpose of context-aware image retrieval. Experimental results are presented in Section IV, followed by concluding remarks and possible future improvements.

II. Related Work

Query based image retrieval is a much research problem, the most common applications of which are image search engines like those provided by Google or Bing. These suffer from significant limitations and low recall/precision when descriptive, conversational style natural language is used for querying or when very specific requirements are provided by the user for image retrieval. Deng et al [1] addressed the problem of image retrieval by representing given images as a "histogram of SIFT Codewords" [18] and performing classification using the k-Nearest Neighbors (kNN) algorithm [17]. They experimented with large-scale data, but the kNN algorithm could not effectively handle with the volume and diversity of large datasets. Their experiments underscored the challenges in large-scale image recognition, paving the way for more complex classification algorithms and deep learning approaches.

An alternate approach for improving image retrieval task accuracy is to reformulate the user query itself. Arandjelovic et al [3] used Google Image Search (GIS) to find potential query strings that match images, which were used to rank candidate dataset images. A major drawback was that they used a small dataset and very few queries, and their method could not scale well. Their work was improved upon by Chatfield and Zisserman [4] who used GIS and potential query strings as labels for images for dynamically training machine learning classifiers. With this, images without any associated annotation across various categories can be retrieved, by 'appropriating' annotation information from Google. Though they were able to achieve improved performance, they limited their experiments to a dataset with just 20 different classes. Farrel et al [6] and Philbin et al [7] devised techniques for capturing the co-occurrences between image regions and tags, as the correlation between them is often unknown. However, as this requires training images with corresponding caption text, several such approaches are limited to category-level tags and are unable to handle instance-level tags. Grangier and Bengio [8] formalized the retrieval task and introduced learning procedure using kernel based classifiers where the learning objective was based on the final retrieval performance.

Early approaches to text representation centered around a sparse vector representation akin to a bag-of-words model. The main drawback here was the loss of word ordering in the document due to which the semantics of the words themselves cannot be captured. To account for the first drawback, different n-gram models were used, that retain some ordering information but suffer as tuples or triples generated in the n-grams are very unique and hence over-fitting is common. The big leap forward was Mikolov et al's work [24] in defining the concept of *word embeddings*, where different words are represented as points in an n-dimensional semantic space, based on shallow neural networks trained on large-scale text corpus. These are trained from text obtained from Google News across a year with an RNN to retain ordering of the words. Since the origins of these representations remain opaque on account of the nature of deep learning models, an alternative statistical method was proposed by Pennington et al [25], who used the probability of word co-occurrence in the corpus as the main metric for obtaining the word vectors..

Vast strides were made in the field of object recognition with the advent of deep learning based models first proposed by Krizhevsky et al [12], who designed a deep convolutional neural network with dense & max-pooling layers and the use of dropout with regularization for classifying ImageNet data, winning the ILSVRC-2012 competition. Their model was trained on 1.2 million high resolution images and remains one of the most extensive works in the field. Thus, Open Vocabulary Object Retrieval has become a topic of interest for several researchers. An alternative approach to mapping objects to predefined object categories is to define a scoring function for comparing a given text query against the varied representations of an image in a "weighted open-vocabulary text space" [14]. This means, a set of functions that can capture the essence of a given image into a sparse vector of words is to be chosen for optimal representation. Each function then produces a sparse representation of the image depending on the information extracted by that function. Each of these functions are divided into a set of classifiers (ImageNet or Visual Object recognition DCN) building upon works like Krizhevsky et al [12] and a set of large image mapping databases like GIS [4]. These are category and instance based projections which are combined to get one representation which is matched with the query to give the best result. To obtain text representation from classification results for comparison against the given query, WordNet based query expansion [10] is also used.

From the earlier discussion, it can be observed that existing approaches use a combination of instance-level projections and category-level projections, which is then used for computing similarity to the query. Category-level classification is done by training models on standard datasets and their fixed label sets. Instance level classification focuses on matching query-image on large image databases, where query expansion techniques can be employed for improving the retrieval results. Vector representations at the sentence and word level are also known to have comparable performance for different tasks and hence experiments are required to derive the best method that can outperform these techniques.

One aspect of deep learning which has not been incorporated yet into Open Vocabulary Object Retrieval is that of Transfer Learning. The basis of this is that, especially for computer vision tasks it is possible that some semantic information such as shape and color transcend the retrieval task at hand. Lampert et al [13] used transfer learning effectively to perform attribute based classification and improve the accuracy on classes with lower representative images. Shin et al [15] leveraged the concept of transfer learning for medical image classification, and comprehensively noted that models pre-trained with ImageNet managed to outperform CNNs trained from scratch for the given task. We intend to explore this avenue in developing a model that can outperform the state-of-the-art models in OVOR.

III. Proposed Methodology

The OVOR task to be performed can be reduced to choosing the best match among a set of candidate images C: $\{C_1, C_2, C_3 \ldots C_k\}$ to a query q from set Q. This is done by computing a set of projections R for each C_i w.r.t. q and then choosing the best fit among the candidate images using the results of each projection in R. The set of projections employ object recognition classifiers (with transfer learning across datasets for better performance) for category level projections and large image matching databases like Google Image Search for instance level projections.

The high-level overview of the proposed approach is depicted in Fig. 1. Here, each candidate image is processed separately for transforming it into image-text projections and comparing it against the user query so as to select the best image from the set. The image-text projections are divided into *category based* and *instance based* approaches. Category

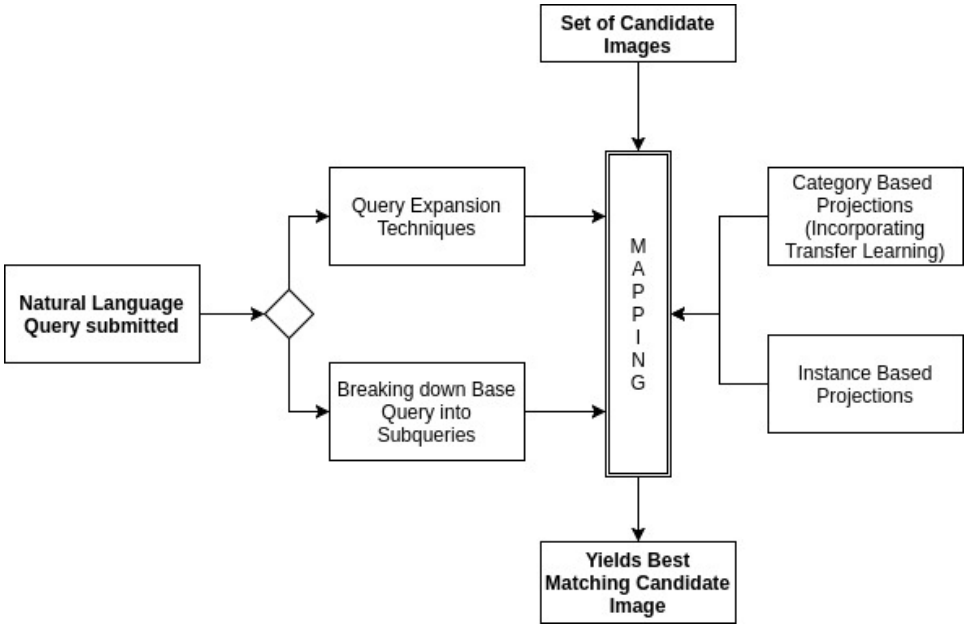

Figure 1. Proposed OVOR methodology

based projections aim to put the candidate image into a set of predefined categories. Instance based projections try to obtain more specific information about the image itself and to describe it accordingly. Each model is designed as per a bottom up methodology as each projection is individually generated, and finally all generated projections are employed for the task of OVOR.

Our approach makes use of three category based projections and two instance based projections. Fig. 2 illustrates the various projections and the process of mapping them to a given natural language query. The first projection uses the Caffenet model trained on ImageNet for the classification of larger good quality images. The second projection is generated by AlexNet trained on Cifar100 dataset is used to accurately classify small-sized RGB images while the third is generated by AlexNet trained on Caltech256 dataset, which contains complex, rotated and obscured images. Also, Google Image Search is used for generating a reverse image search projection that captures information not recognized by category-based projection (like brand names). Finally, the image captioning model consisting of a LSTM trained on MS COCO dataset is used to capture the actions i.e verbs w.r.t contents of a given image.

A. Category Based Projections

The ultimate goal of the proposed approach is to be able to deal with images of various types of objects taken from multiple, different views. So, it is important to be able to identify an object from any kind of image. We have chosen to train our projections on three different datasets - ImageNet, Cifar 100 and Caltech 256. These datasets were chosen as they cover varied objects, specifically, household objects belonging to different classes, but each differ slightly in the kind of images represented. While ImageNet contains high resolution images of size 227x227 pixels, Cifar 100 is used for smaller images 32x32 size RGB images and Caltech 256 contains challenging views of the objects. We used these three datasets together so that any kind of image of an object should at least have one good representation from our system. Category-based projection models include CaffeNet models trained on Cifar-100 and Caltech-256 datasets. We incorporated Deep Convolutional Neural Networks (Deep-CNNs) for object classification and transfer learning across datasets to improve the accuracy. We used Caffe for the implementation of the Deep-CNN on account of its ease of use and modular functionality.

1) Deep CNN Model for Small-sized Images: The Cifar-100 dataset consists of small-sized images of size 32x32. For every forward pass on the Deep-CNN model, two separate labels are generated for each image, one from 20 coarse labels given and one from 100 fine labels provided. In the training set, each image is labeled with a coarse and fine label, so these are trained separately. The proposed model consists of two sets of convolutional layers, then two sets of fully connected layers followed by two separate softmax layers to predict coarse and fine labels. For use as a projection, both labels are important as they provide additional information which can be used to retrieve the correct image. The network architecture employed for this projection was *conv(64,4)-pool(2)-conv(128, 4)-pool(3)-ip(256)-softmax*. The model was inspired by AlexNet[12] due to its two sets of successive convolutions of increasing size as this has been known to perform the task of object detection very well. During experimental validation, the model's performance was comparable to that of the current state-of-the-art, HD-CNN developed by Yan et al [24]. As we have separate coarse and fine labels, we use both for comparing with the query, resulting in coarse label accuracy

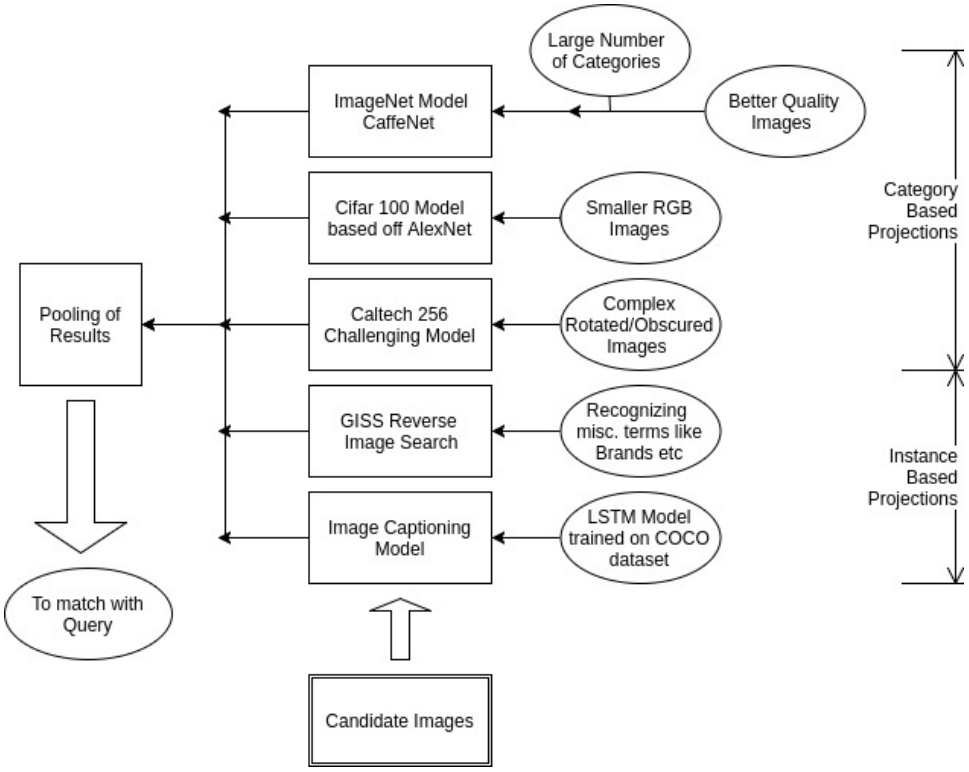

Figure 2. Mapping projections to a given natural language query

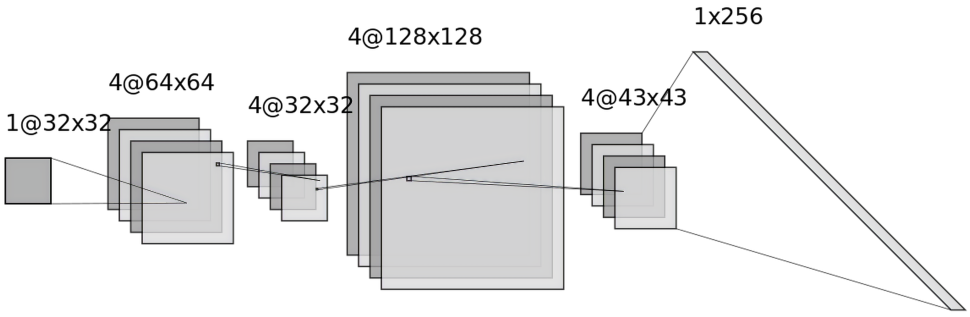

Figure 3. Architecture of the Deep CNN used for Cifar-100 Dataset

of 77% while that achieved with fine labels is 63.16%.

2) Transfer Learning Model for Challenging Images: The Caltech-256 dataset contains images with certain additional constraints like obstruction and rotation. We selected images such that both obstruction as well as rotation was included, across 256 different categories each with roughly 50 images. It is to be noted that the task itself is quite challenging with a benchmark accuracy of just 34.1% as reported by Griffin et al [21] and 34.02% achieved by Yang et al [22]. We experimented with models based on LeNet and AlexNet, but our own design had an optimum performance of around 28%. The final architecture that yielded the best results is based on LeNet and transfer learning on these weights is employed to aid the training phase. The architecture is *conv(20,5)-pool(2)-conv(50,5)-pool(2)-fc(500)-relu-fc(100)-softmax*. Since the accuracy of the model is low, we maintain a high softmax

threshold to classify images into a particular category, which is to a certain extent analogous to accepting only entries where the model is 'sure' of its result. The reasoning behind retaining this model was that, often, difficult images cannot easily classified by standard models which might possibly be correctly labeled by this model. The ImageNet model we used was obtained from Caffe Model Zoo and is used as a projection directly on account of the extensive training and testing it was subjected to.

B. Instance Based Projections

To recognize any details of a given image that do not fall into any of the general categories identified by our classification models, Instance Based projections are used. An Image Captioning model along with Google Reverse Image search was used in tandem to capture those details of a given image, which were not captured by the Category Based Projection method. The image captioning model is an encoder CNN model that uses VGG16 [16] as the image encoder. It is implemented using the sequential API of Keras. In addition, a LSTM (Long Short Term Memory) network trained on the MS COCO dataset is used for the task of generating captions. The LSTM network takes the image vector and partial captions at the current epoch as input and generates the next most probable word as output. Image captioning helps to capture the descriptive information accurately.

In addition to generating image captions, Google Reverse Image search was used to help improve the classification by capturing additional latent context information like brand names etc, which is otherwise difficult to obtain. Python's Beautiful Soup library is used to to scrape the corresponding query for the image uploaded. To create an instance of Google Chrome so that query can be effectively scraped, Chrome Driver and an automation tool Selenium is used. The actual process is detailed in Algorithm III-B. For image captioning, a LSTM model trained on MS COCO dataset, is used for generating natural language descriptions that can be used as image captions. The image features obtained from the CNN model pretrained on ImageNet are fed into the image captioning LSTM network to generate a sentential description of the image in valid English.

Algorithm 1 GISS Reverse Image Search

Input: Dataset images
Output: Corresponding best-fit Query set for an image

1: **for** images in dataset **do**
2: Create a Chrome Web driver instance
3: Send a POST request to upload image to Chrome instance
4: Parse generated query_string
5: return the *query_element*
6: **end for**

C. Query Representation

Next, the natural language user query is converted to a vector using the Word2Vec representation[24] which provides a 300 dimensional float vector that is trained to represent the word itself in a semantic space. The release provided as part of the work was extensively trained on Google News so it provides a comprehensive set of words. To combine multiple words of the query, the vectors are averaged to obtain a final vector representation of the query. To provide some additional information along with that given by the user, WordNet [35] hypernyms and synonyms are concatenated with the query for obtaining better and well-rounded descriptions. For instance, a user might provide a query such as '*kitten*' so including synonyms like '*cat*', '*feline*' and hypernyms like '*animal*' would help provide a more well-rounded caption before the conversion to vector form is performed.

D. Query-Image Matching

The overall objective is to retrieve closest image representations for a given NL query, for which the chosen text representation model should be highly effective in comparing generated image captions. Since no domain knowledge is available for most user queries, models that have been trained specifically for a particular task are ineffective. Our goal was to design a diverse and robust representation of the underlying image collection, hence, multiple models were trained on varied datasets for enabling diverse representations for different types of candidate images. The classification results of each model are aggregated to obtain the final candidate image list. The final layer of each model employed is a softmax probability layer, and we heuristically determined the thresholds to ensure that a model with less benchmark accuracy like Caltech 256 is given a higher softmax probability before the final label is selected. We used Pearson Correlation metric to evaluate the similarity between the query and projection outputs for identifying the best candidate image. The final image retrieval task was evaluated using metrics like recall, which gives details on how many retrievals were successfully completed as compared to the total number of retrievals. This is useful as it can give valuable insights into the performance of our model.

IV. Experimental Results

An experimental benchmarking of the proposed approach was performed with emphasis on end-to-end evaluation of entire framework, the results of which are discussed in this section. The proposed approach is evaluated on the ReferIt dataset [34], specifically the images from the RefCoco task which provides us with both images and corresponding captions. A set of candidate images is chosen and the caption of one of these is then passed to the system which ranks the images accordingly. This is then repeated for 1000 such sets and the *recall* metric is used to evaluate the model as a whole. Experiments on the effect of the performance in case of variations in size of the candidate set were also conducted to observe the performance of the model when varying degree of distractions (i.e. images with dissimilarity of different degree) are present. The results of these experiments are tabulated in Table II.

We observed that, given a candidate set of 2 images, our model is able to actively select the correct image with more than 70% accuracy. As the number of candidate images from which the model must select the best is increased, there is naturally a decrease in the accuracy. To benchmark the performance of our approach against that of state-of-the-art works, we used a standard dataset RefCoco. In this, the approach and goal adopted by Guadarrama et al [14] is slightly different, however, our approach compared favorably to their model, and was able to achieve the OVOR task well. Following this, we also evaluated our model against standards set by Cer et al [29] and Jainan et al [30] on the RefCoco and again our model outperformed these models. It should however be noted that the evaluation strategies of these works are slightly different, that is, their strategy focuses on choosing proposals within the image and evaluating recall based on how close those proposals are to the actual main object of interest. So, while a direct comparison is not entirely feasible, the underlying performance of the IR model cannot be understated. We provide these results as a standard to reinforce the difficulty of the various datasets chosen and the robust, end-to-end models that constitute the proposed OVOR system.

Table I
PERFORMANCE OF CATEGORY-BASED PROJECTION MODELS BENCHMARKED AGAINST STATE-OF-THE-ART WORKS

Dataset	Model	Accuracy (%)
Cifar 100	NiN [19]	62.32
	Stochastic Pooling [20]	62
	HD-CNN [23]	67.38
	Our Model	**64.16**
Caltech256	Griffin [19]	34.1
	Yang [20]	34.02
	Our Model	**28.57**

Table II
RECALL PERFORMANCE OF PROPOSED MODEL IN COMPARISON TO STATE-OF-THE-ART MODELS BASED ON SIZE OF CANDIDATE IMAGE SET

Testcase	Candidate Set size	Our Model	Other Works
1	2	68.8%	-
2	3	56.4%	52.35% [29]
3	4	54.2%	41.2% [30]
4	5	48.1%	40.39% [30]
5	6	41.4%	-

The success of our methodology is on account of the interplay of three factors. In terms of text representation, we make use of the Word2Vec model as opposed to a simple sparse vector matching. Incorporating semantic information into the query matching process positively contributes to the accuracy of the task. The second is the use of an image captioning model which captures information regarding verbs and actions in the query. Instead of limiting to just noun and adjective based projections, we gain a more complete perception about the image itself due to this. Lastly, while most works are dependent on ImageNet class hierarchy, our method employs a more robust approach considering inputs from multiple and varied image sources like Cifar 100 and Caltech 256, better enabling the proposed approach in dealing with complex and obscure object retrieval.

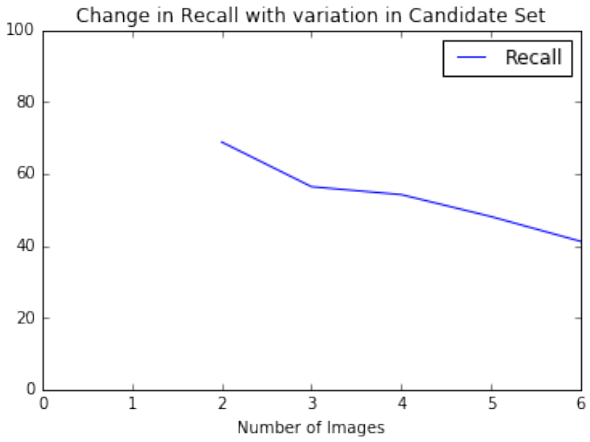

Figure 4. Variation of Recall values with Candidate set size

When evaluated against existing benchmarks on RefCoco, it can be seen that our model outperformed other IR Models consistently. Performance of state-of-the-art models for candidate image size 2 and 6 were not available, hence we were unable to compare our model in these cases. However, it can be seen that as the number of distraction images is increased, naturally the performance deteriorates, but our model still stayed ahead of existing benchmarks in recall performance. We also observed that as the number of distractions increases i.e. if the model needs to select the correct image from maybe 3 or 4 different candidate images, our model still was able to perform better that other contemporary works, which highlights the robustness of the proposed model. Rather than fitting a proposal to a particular class label, we attempt to create an end-to-end textual representation of all information in the images, to generate a diverse set of projections that ensures a good representation of even complex, obscure and low-resolution images.

V. CONCLUSION AND FUTURE WORK

In this work, a robust and comprehensive approach for addressing the problem of Open Vocabulary Vocabulary Image Retrieval was presented. The approach incorporates a well-rounded set of category based and instance based projections that were able to capture good representations of all possible candidate images, including complex obscure and rotated images. The task of open vocabulary object retrieval given a natural language query is addressed using all terms in the English language. A better captioning model would further improve the performance of the task as the model we are using involves a hand trained captioning model on MS COCO. Also the methodology can be fine tuned and applied to a subset of image data such as healthcare or any other such application.

ACKNOWLEDGEMENTS

We gratefully acknowledge the use of the facilities at the Department of Information Technology, NITK Surathkal, funded by Govt. of India's DST-SERB Early Career Research Grant (ECR/2017/001056) to the first author.

REFERENCES

[1] Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei. What Does Classifying More Than 10,000 Image Categories Tell Us? In ECCV, 2010.

[2] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. "Every picture tells a story: Generating sentences from images", In ECCV. Springer, 2010.

[3] Relja Arandjelovic and Andrew Zisserman, "Multiple queries for large scale specific object retrievaL", In BMVC,2012.

[4] Ken Chatfield and Andrew Zisserman, "Visor: Towardson-the-fly large-scale object category retrieval", In ACCV 2012, pages 432446. Springer, 2013.

[5] Ryan Farrell, Om Oza, Ning Zhang, Vlad I Morariu, Trevor Darrell, and Larry S Davis, "birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance", In ICCV, 2011.

[6] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic,and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching", In CVPR, 2007.

[7] Sergio Guadarrama ,Erik Rodner , Kate Saenko , Ning Zhang , Ryan Farrell , Jeff Donahue and Trevor Darrell, "Open-vocabulary Object Retrieval", UC berkeley.

[8] David Grangier and Samy Bengio, "A discriminative kernel-based model to rank images from text queries", PAMI 2007

[9] J. Krapac, M. Allan, J. Verbeek, and F. Jurie. "Improving web search results using query-relative classifiers" In CVPR, 2010.

[10] Yiming Liu, Dong Xu, and Ivor W. Tsang. "Using large-scale web data to facilitate textual query based retrieval of consumer photos". In ACM-Multimedia COnference, 2009.

[11] Aurelien Lucchi and Jason Weston, "Joint image and word sense discrimination for image retrieval". In ECCV, 2012.

[12] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[13] Lampert, Christoph H., Hannes Nickisch, and Stefan Harmeling. "Learning to detect unseen object classes by between-class attribute transfer." Computer Vision and Pattern Recognition, CVPR 2009.

[14] Guadarrama, Sergio, et al. "Open-vocabulary Object Retrieval." Robotics: science and systems. Vol. 2. No. 5. 2014

[15] Shin, Hoo-Chang, et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning." IEEE transactions on medical imaging 35.5 (2016): 1285-1298.

[16] Simonyan, K. and Zisserman, A., 2014. "Very deep convolutional networks for large-scale image recognition", ICLR 2015, arXiv preprint arXiv:1409.1556.

[17] Hwang, Wen-Jyi, and Kuo-Wei Wen. "Fast kNN classification algorithm based on partial distance search." Electronics letters 34, no. 21 (1998): 2062-2063.

[18] Scovanner, Paul, Saad Ali, and Mubarak Shah. "A 3-dimensional sift descriptor and its application to action recognition.", 15th ACM international conference on Multimedia. ACM, 2007.

[19] Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." arXiv preprint arXiv:1312.4400 (2013)

[20] Zeiler, Matthew D., and Rob Fergus. "Stochastic pooling for regularization of deep convolutional neural networks." arXiv preprint arXiv:1301.3557 (2013).

[21] Griffin, Gregory, Alex Holub, and Pietro Perona. "Caltech-256 object category dataset." (2007).

[22] Yang, Jianchao, et al. "Linear spatial pyramid matching using sparse coding for image classification." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[23] Yan, Zhicheng, et al. "HD-CNN: hierarchical deep convolutional neural networks for large scale visual recognition.", IEEE International Conference on Computer Vision. 2015.

[24] Mikolov, Tomas, Chen et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781, 2013

[25] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation.", 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[26] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International Conference on Machine Learning. 2014.

[27] Wieting, John, et al. "Towards universal paraphrastic sentence embeddings." arXiv preprint arXiv:1511.08198 (2015).

[28] Daniel Cer, Mona Diab, Eneko Agirre, Iigo Lopez-Gazpio, and Lucia Specia (2017) SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2017)

[29] Jianan, et al. "Deep Attribute-preserving Metric Learning for Natural Language Object Retrieval.", 2017 ACM on Multimedia Conference. ACM, 2017.

[30] Wu, Fan, Zhongwen Xu, and Yi Yang. "An End-to-End Approach to Natural Language Object Retrieval via Context-Aware Deep Reinforcement Learning." arXiv preprint arXiv:1703.07579 (2017).

[31] Kazemzadeh, Sahar, et al. "ReferItGame: Referring to Objects in Photographs of Natural Scenes." EMNLP 2014.

[32] Yu, Licheng, et al.. "Modeling Context in Referring Expressions." ECCV 2016.

[33] LeCun, Y. (2015). LeNet-5, convolutional neural networks. URL: http://yann.lecun.com/exdb/lenet, 20.

[34] Kazemzadeh, S., Ordonez, V., Matten, M., & Berg, T. (2014). "Referitgame: Referring to objects in photographs of natural scenes.", 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 787-798).

[35] Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.