# CAPRI 2017

Workshop on

## Computational Aspects of Pattern Recognition and Computer Vision with Neural Systems

May 18th 2017
Anchorage, Alaska, USA

**Organized under**
The International Joint Conference on Neural Networks (IJCNN 2017)

AGH University of Science and Technology
Signal Processing Group
Al. Mickiewicza 30
30-059 Krakow, Poland

# Automated Image Captioning Using Nearest-Neighbors Approach Driven by Top-Object Detections

Karan Sharma Arun CS Kumar Suchendra M. Bhandarkar
Department of Computer Science, The University of Georgia
Athens, Georgia 30602–7404, USA
{karan@uga.edu, aruncs@uga.edu, suchi@cs.uga.edu}

**Abstract.** The significant performance gains in deep learning coupled with the exponential growth of image and video data on the Internet have resulted in the recent emergence of automated image captioning systems. Two broad paradigms have emerged in automated image captioning, i.e., generative model-based approaches and retrieval-based approaches. Although generative model-based approaches that use the recurrent neural network (RNN) and long short-term memory (LSTM) have seen tremendous success in recent years, there are situations in automated image captioning for which generative model-based approaches may not be suitable and retrieval-based approaches may be more appropriate. However, retrieval-based approaches are known to suffer from a computational bottleneck with increasing size of the image/video database. With an aim to address the computational bottleneck and speed up the retrieval process, we propose an automated image captioning scheme that is driven by top-object detections. We surmise that by detecting the top objects in an image, we can prune the search space significantly and thereby greatly reduce the time for caption retrieval. Our experimental results show that the time for image caption retrieval can be reduced without suffering any loss in accuracy.

**Keywords:** Automated image captioning, top-object detection, image retrieval, $k$-nearest-neighbor search

## 1 Introduction

Automated image captioning, i.e., the problem of describing in words the situation captured in an image, is known to be challenging for several reasons. The recent significant performance gains in deep learning coupled with the exponential growth of image and video data on the Internet have resulted in the emergence of *automated* image captioning systems. Two broad paradigms have emerged in the field of automated image captioning, i.e., generative model-based approaches [3], [5], [9], [11], [17] and retrieval-based approaches [1]. Although generative model-based approaches that use the recurrent neural network (RNN) and long short-term memory (LSTM) have seen tremendous success in recent years, there are situations for which retrieval-based approaches may be better suited. Examples of such situations include:

(1) Situations wherein the training sets are dynamically changing. To keep up with the increasing pace of visual data being constantly uploaded on the Internet, computer

vision practitioners face a challenging task of training models that are capable of adapting to constantly changing datasets or reducing the size of the datasets. By reducing the size of the datasets, one runs the risk of discarding useful data resulting in the learning of simplistic models. Adaptive models have the added overhead of requiring constant training or retraining as the underlying datasets change over time. Moreover, adaptive models need to deal with the problem of *concept drift*, i.e., situations where the statistical properties of the target variable or concept, which the model is trying to predict, change over time in unforeseen ways, especially when the new data being uploaded is significantly different from previously observed data. In contrast, retrieval-based approaches, modeled on nearest-neighbor search, do not entail the overhead of constant retraining of models since one can store all the images in the dynamically changing dataset in a database.

(2) Situations wherein one needs to deploy an automated image captioning system with the goal of simultaneously reducing system development time and CPU execution time. Nearest-neighbor approaches lend themselves easily to rapid implementation and deployment since they have very few tunable hyperparameters compared to other approaches. Hence retrieval-based approaches based on nearest-neighbor search are naturally preferred in rapid prototyping situations. However, the potential downside of retrieval-based approaches is that nearest-neighbor search can be exhaustive if one has to perform all possible comparisons between the query image and the database entries. Traditionally, techniques such as locality sensitive hashing (LSH) have been used to speed up nearest-neighbor search. However, effective use of LSH requires the proper tuning of several hyperparameters in order to achieve accurate results. In contrast, the proposed approach has very few tunable hyperparameters and hence a much less computationally intensive hyperparameter tuning phase.

Although retrieval-based approaches to automated image captioning have not been as successful as generative model-based approaches, the performance of retrieval-based approaches has been observed to be not very far behind that of RNN- and LSTM-based approaches when addressing the *Microsoft Common Objects in Context* (MS COCO) challenge. Therefore the obvious question arises - in situations (such as the ones described previously) where retrieval-based approaches are called for, how does one speed up the nearest-neighbor search procedure? To this end, we propose a variant of the nearest-neighbor search procedure to speed up image caption retrieval using top-object detections. Specifically, we use the detection of the most significant objects in an image (i.e., the top objects) to speed up the $k$-nearest-neighbor ($k$-NN) search for retrieval-based automated image captioning. Although, as noted previously, approaches such as LSH can used to accelerate the retrieval process, LSH entails a hyperparameter tuning procedure that is computationally complex and difficult to implement thereby calling for a significant expenditure of programmers' development time.

## 2 Related Work

**Automated image captioning:** Automated image captioning systems have grown in prominence owing, in large part, to the tremendous performance gains shown by deep

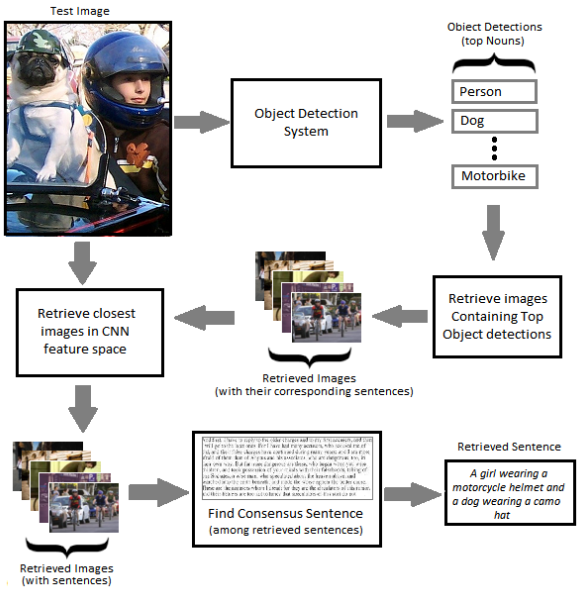


**Fig. 1.** Top-object detections used to drive $k$-NN search.

learning in recent times. Existing automated image captioning systems can be categorized as either generative model-based or retrieval-based. Generative model-based systems involve correctly identifying objects, verbs, adjectives, prepositions or visual phrases in an image and generating a caption from these or directly from the representation of the image [3], [5], [9], [11], [17]. Retrieval-based approaches [1], on the other hand, involve retrieving the most suitable caption from a database of captions and assigning it to an image. Currently, generative model-based approaches that use the RNN and LSTM have been shown to yield the best performance metrics in the context of automated image captioning; however retrieval-based approaches have also proved to be quite competitive in terms of performance. The generative model-based and retrieval-based paradigms are each suited for different kinds of situations and applications. However, in situations where retrieval-based approaches are more appropriate and successful, we propose a scheme to optimize and speed up the caption retrieval process by exploiting the top-object detections in the image.

## 3 Motivation

Although generative model-based approaches that use the RNN and LSTM are regarded as the state-of-the-art in automated image captioning, there are potential situations for which they may not be well suited and hence retrieval based captioning approaches may be called for. However, retrieval-based approaches to automated image captioning can be computationally intensive and slow especially when a query image is compared with all the images stored in the database. However, before we proceed to address the question of how to speed up retrieval-based approaches to automated image captioning,

we digress to answer an important related question, i.e., under what potential situations would retrieval-based approaches have an advantage over state-of-the-art generative model-based approaches that use the RNN and LSTM in the context of automated image captioning?

**Concept Drift:** Consider the problem of automated image captioning in situations where the underlying datasets are dynamically changing such as when visual data (in the form of images and videos) is being uploaded over the Internet at an extraordinary pace, both on popular online social media (OSM) platforms such as Facebook, Instagram, Snapchat, Google and Twitter, and on websites that contain more structured and specific information such as those dealing with news, sports, art, and technology. Many of the uploaded images and videos have some sort of textual information associated with them, typically in the form of tags, captions, and/or comments. Mining such a large data set is tremendously challenging for most computer vision practitioners. The constant pace of the dynamically changing dataset makes it incredibly difficult to learn reliable computer vision models. The standard assumption underlying most machine learning techniques is that the training data will be similar to the testing or querying data. However, in dynamic situations where the underlying data is continuously changing, it is especially hard to train reliable models. In this paper, we propose a retrieval-based model for automated image captioning for situations wherein the training datasets are very volatile and constantly changing.

One of the problems faced when dealing with dynamic datasets is the problem of *concept drift* where the function learned by a machine learning model is rendered not particularly useful for newly arriving data. For example, near Christmas, people tend to post more pictures or images of their activities around a Christmas-oriented theme on OSM sites. An existing machine learning model may not be in situation to automatically label or caption these images since it has not seen these images previously. One solution is to constantly retrain the existing model as the new data arrives or use models that are capable of adapting to new data. However, the constant retraining of models could pose significant and, in some cases, an impossibly high computational demand, especially in situations where images are being uploaded at a very rapid pace. Moreover, many adaptive models, in the interest of computational efficiency, subsample the data during retraining. The discarding of data could lead to the learning of overly simplistic models. The interested reader is referred to the work of Gama et al. [7] for a more detailed and comprehensive treatment of the concept drift problem.

For the reasons mentioned above, some of the most popular automated image captioning schemes, based on generative models that use the RNN and LSTM, are seriously disadvantaged in situations where a large proportion of the training data is in a state of constant flux. In such instances, the generative models will learn a classification or prediction function that could account for most cases, but may miss cases that occur only a few times. Moreover, the cases that occur infrequently may contain valuable information. For example, if the training set has millions of images, and only five instances of *Man is biting a dog*, the generative model may simply ignore this infrequent case during the training process, although the case may be of potential interest. Hence, for this reason and reasons described in previous paragraph, generative models are not well suited for image captioning under dynamically changing training datasets. However,

retrieval-based approaches, such as ones based on $k$-NN search do not suffer from such problems. It has been convincingly shown by Hays and Efros [8] that $k$-NN search is one of the most effective retrieval algorithms if one has a very large dataset. However, exhaustive $k$-NN search could be computationally very expensive. Although techniques such as LSH have been traditionally used to speed up $k$-NN search-based image retrieval [4], the hyperparameter tuning procedure needed to optimize the performance of LSH is non-trivial in terms of its computational complexity [4]. The situation is further complicated if we need to retune the LSH procedure in the face of constantly arriving new training data. Thus, retrieval-based automated image captioning techniques suffer from the same disadvantages as their generative model-based counterparts if the former use $k$-NN search optimized via LSH. In this paper, we propose a simple retrieval-based technique for automatic image captioning that is accurate, reliable and computationally efficient. The proposed technique is based on enhancing the $k$-NN search by exploiting the top-object detections in an image.

**Rapid Prototyping:** We use top-object detections to speed up the caption retrieval procedure during automated image captioning. Specifically, we use the detection of the most significant objects in an image (i.e., the top objects) to speed up the $k$-NN search for retrieval-based automated image captioning. We show top-object detection to be a preferable alternative to the more conventional retrieval-based automated image captioning methods that employ LSH to speed up the $k$-NN search. It is to be noted that although techniques such as LSH can be used to speed up $k$-NN search-based image retrieval, the hyperparameter tuning procedure needed to optimize the performance of LSH is non-trivial in terms of computational complexity [4], especially in the case of complex applications such as automated image captioning. Thus, complete automation of the LSH procedure for automated image captioning is a challenging task. Implementation and proper tuning of LSH also presents a significant expenditure of system development time, which is an important consideration in real-world situations where rapid prototyping is called for.

## 4 $k$-NN Search Driven by Top-Object Detections

Previously, Devlin et al. [1] have obtained good results for automated image captioning based on $k$-NN search-based image retrieval. Their approach determines the $k$-NN images by computing a measure of image similarity between the test/query image and each of the database images. The test/query image is then assigned the caption obtained by computing the consensus of the retrieved $k$-NN image captions. Performing an exhaustive search of the image database to retrieve the $k$-NN images using an image feature-based similarity metric is clearly not a scalable approach. We show that, in the context of automated image captioning, by detecting all objects in a test image, selecting the top-$n$ objects (where $n$ is a small number) and retrieving all images that contain at least one of these $n$ objects, one can achieve results comparable to those of $k$-NN retrieval via exhaustive search while simultaneously obtaining a significant speedup. Fig. 2 summarizes the proposed approach. We demonstrate our approach on the MS COCO dataset as a proof of concept. We believe the experimental results on the MS COCO dataset are transferable and generalizable to real-world dynamic datasets.

Although the proposed approach involves tuning the parameters of a support vector machine (SVM)-based classifier for object detection/recognition, it is computationally much less expensive than the LSH hyperparameter tuning procedure used to optimize $k$-NN search and also yields readily to automation.

Although running various object (i.e., noun) detectors on the test/query image imposes a computational overhead, it is offset by the following considerations: (a) the space of objects (i.e., nouns) is bounded. Also, since objects are concrete entities, generating training sets for object detectors is not very difficult if one uses web-based data coupled with crowdsourcing, (b) sliding windows are not used during the object detection procedure, i.e., the entire test/query image is fed as input to the SVM-based object detector. The computational overhead of object detection in the test/query image is also offset by: (a) the resulting speedup over $k$-NN image retrieval via exhaustive search and, (b) savings in development time compared to the scenario wherein $k$-NN image retrieval is optimized using LSH. Additionally, the proposed approach also results in significant savings in CPU execution time as shown in Table 1.

**Complexity Analysis:** Given a set of objects $X = \{x_1, x_2, \ldots, x_n\}$, and a set of images $I = \{I_1, I_2, \ldots, I_m\}$, we make the following assumption regarding the dataset: Each object $x_i$ does not occur in more than $k$ images in the dataset where $k \ll m$. In real world datasets, especially in large datasets, it is expected that no single object category will dominate the images in the dataset. Even generic categories such as *person, car,* ..., would be expected to occur in a significantly small percentage of the total number of images in the dataset. Also, for a small subset $Y \subset X$ where no member of $Y$ occurs in more than $r$ images ($r \ll m$) in the dataset, the number of comparisons is bounded by $r \cdot |Y|$ resulting in a $O(r \cdot |Y|)$ time complexity. However, what if $r$ is a large number? We argue that in datasets that are sufficiently representative of real world, this will not be the case. For example, consider an image whose top detections are *person, dog, road,* and *building*. Intuitively, in a large dataset representative of many nouns and concepts in the world, we can expect that all the images that contain at least one entity from the set { *person, dog, road, building*} are far fewer than all the images in the dataset thus resulting in an order of magnitude reduction in search complexity.

**Retraining Event Analysis:** Assume an image dataset (with associated captions for each image) of size $N$ (i.e., $N$ is the number of data points). Assume this dataset is being constantly augmented with new incoming image data (and the associated captions). Assume that after every $w$ data points (i.e., images) are added to the dataset, there is a concept drift, that requires retraining of the model. In a traditional generative model-based system that uses an RNN, retraining will be needed in two situations after the addition of new data points to the existing dataset:

(a) Changes in concepts, where a concept is any word, which includes nouns, verbs, adjectives and so on. Assume that the concepts change at an average rate of $c$ concepts after $w$ new data points are introduced. Clearly, the space of concepts is far greater than the space of objects (i.e., nouns). Let $tr(c)$ denote the average number of training events required to account for the concept changes after a collection of $w$ new data points is added to the existing dataset.

(b) Changes in concept dependencies. The dependency between two words is a measure of how much a given word depends on the other word. For example, the word *eating*

is dependent on the words *person* and *food*. We need to retrain the model to learn such dependencies after a collection of $w$ new data points is introduced. Assume that the concept dependencies change an average rate of $d$ concept dependencies upon introduction of a collection of $w$ new data points. Again, based on our understanding of the real world, the space of these dependencies is significantly larger than the space of objects alone. Let $tr(d)$ denote the average number of training events required to account for the changes in concept dependencies after a collection of $w$ new data points is added to the existing dataset.

In contrast to a traditional generative model-based system, in the proposed approach, the training events will be required only when new objects are introduced at an average rate of $ob$ objects after $w$ new data points are added to the existing dataset. Clearly, the training events are bounded by the number of objects under consideration. Let $tr(ob)$ denote the average number of training events required after $w$ new data points are added to the existing dataset. Based on our knowledge of the real world and the above arguments, the training events in the proposed approach will be significantly fewer than the training events in a traditional generative model-based system (such as one that uses an RNN), i.e., $tr(ob) \ll tr(c) + tr(d)$.

## 5 Experimental Results

**Training:** For the purpose of training, we use 80 annotated object categories in the MS COCO dataset [10]. Binary SVM classifiers are trained for each of these 80 annotated categories using VGG-16 *fc-7* image features [13], and the SVMs are calibrated using Platt scaling. For the extraction of fc-7 features, Matconvnet package [15] is employed.

In addition, we store each training image in the MS COCO dataset and its accompanying sentences (5 sentences per image) in our database. We treat these sentences as ground truth captions for the corresponding training image. For testing purposes, we consider the MS COCO validation set consisting of close to 40,000 images.

**Testing:** For each test image in the MS COCO validation set, we run all the 80 object detectors on the test image. We select the top-$n$ objects from all the detected objects in the image. In our current implementation $n = 5$. The detected top objects are the ones that are deemed to possess the highest probability of occurrence in the image. The probability of occurrence of an object in the image is computed by mapping the classification confidence value generated by the SVM classifier for that object to a corresponding probability value using Platt scaling [12]. From the training dataset, we retrieve all images that contain at least one of the top-$n$ objects detected in the previous step, using the corresponding ground truth captions, i.e., a training image is retrieved if at least one of its associated ground truth captions contains a noun describing the object under consideration. In addition, for the purposes of retrieval, all the synonyms for certain words such as *person* (synonyms are man, woman, boy, girl, people, etc.) are taken into consideration. Using the cosine distance between the *fc-7* features of each retrieved image and the test image, we select the $k$-NN images for further processing.

In the current implementation we have chosen $k = 90$ as recommended by [1]. Since each of the $k$-NN images has 5 associated sentences (captions), we have a total of $5k$ potential captions for the test image. We determine the centroid of the $5k$ potential

captions and deem it to represent the consensus caption for the test image. The consensus caption is then assigned to the test image in a manner similar to [1]. The BLEU measure is used to evaluate the similarity (or distance) between individual captions and to determine the centroid of the $5k$ potential captions. We have also implemented image retrieval using exhaustive $k$-NN search [1] and compared the CPU execution time of the proposed approach with that of image retrieval using exhaustive $k$-NN search for 2000 random images .

**Results:** As shown in Table 1, the proposed image retrieval, using $k$-NN search driven by top-object detections, and the standard image retrieval, that employs exhaustive $k$-NN search, yield very similar results when the BLEU and CIDEr [16] similarity metrics are used to compare the retrieved captions.

**Table 1.** Comparison of image captioning results obtained using the proposed approach for image retrieval based on $k$-NN search driven by top-object detections (Obj-$k$-NN) and those obtained using conventional image retrieval based on exhaustive $k$-NN search (Exh-$k$-NN).

| | $BLEU_1$ | $BLEU_2$ | $BLEU_3$ | $BLEU_4$ | CIDEr | CPU time |
|---|---|---|---|---|---|---|
| Exh-$k$-NN | 65.6% | 47.4% | 34% | 24.7% | 0.70% | 2.5e+04s |
| Obj-$k$-NN | 64.6% | 46.2% | 32.8% | 23.6% | 0.68% | 1.17e+04s |

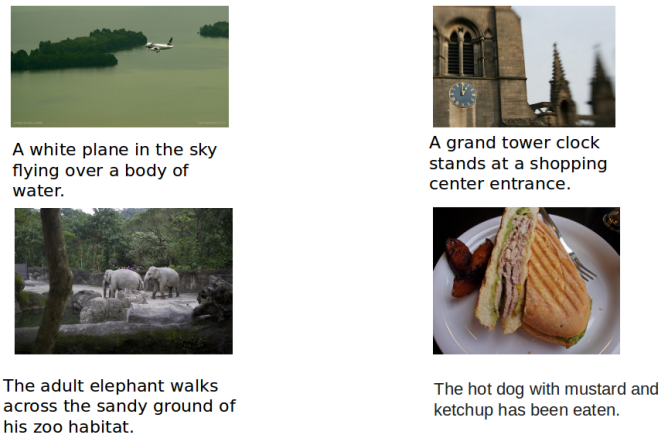


**Fig. 2.** Qualitative Results for nearest neighbor driven by top-object detections. Some captions retrieved accurately describe the image while others are partially correct.

The proposed approach is seen to yield significant gains in CPU execution time when compared to image retrieval using exhaustive $k$-NN search. Essentially, the proposed image retrieval technique based on $k$-NN search driven by top-object detections is observed to provide an attractive alternative to LSH for the purpose of speeding up

$k$-NN search-based image retrieval in the context of automated image captioning. As a proof of concept, the results of the proposed image retrieval technique based on $k$-NN search driven by top-object detections on the MS COCO dataset are fairly convincing. We believe that these results could be directly transferred to real-world datasets that are dynamically changing.

These results show that $k$-NN search driven by top-object detections, even though simple in concept, can provide significant gains in critical situations where the datasets are dynamically changing. This approach requires that we store all the training images along with their associated captions in the database. When dealing with real-world problems, we will store all the image instances in the database and retrieve the relevant images from the database using top-object driven $k$-NN search. There are three advantages to the proposed approach: We do not need to subsample the dataset by discarding any potentially useful information, we do not need to exhaustively search for the $k$-NN images, and we do not need to retrain the retrieval models in the face of changing information.

## 6 Conclusions

We have shown that retrieval-based approaches for automated image captioning could be made computationally more efficient if they are driven by top-object detections. The potential advantages of our approach are in situations where the underlying datasets are changing dynamically. In addition, the proposed approach needs much less parameter tuning when compared to the computationally intensive hyperparameter tuning associated with traditional LSH-based optimization of $k$-NN search. The proposed approach is a natural candidate for use in rapid prototyping conditions that also call for optimization of CPU time.

**Acknowledgment** The authors wish to thank Devi Parikh for her invaluable suggestions during this research.

## References

1. Devlin, J. et al. (2015). Exploring nearest-neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*.
2. Devlin, J. et al. (2015). Language models for image captioning: The quirks and what works. *Proc. ACL 2015*.
3. Donahue, J. et al. (2014). Long-term recurrent convolutional networks for visual recognition and description. *Proc. IEEE Conf. Comp. Vis. Patt. Recog.* (CVPR 2014).
4. Dong, W. et al. (2008). Modeling lsh for performance tuning. *Proc. ACM Conf. Info. & Know. Mgmt.*, October, pp. 669-678.
5. Fang, H. et al. (2014). From captions to visual concepts and back. *arXiv preprint arXiv:1411.4952*.
6. Farhadi, A. et al. (2010). Every picture tells a story: Generating sentences from images. *Proc. Eur. Conf. Comp. Vis.* (ECCV 2010), pp. 15-29.
7. Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys* (CSUR), Vol. 46(4), pp. 44.

8. Hays, J., and Efros, A. A. (2007). Scene completion using millions of photographs. *ACM Transactions on Graphics* (TOG), Vol. 26(3), pp. 4, August.
9. Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proc. IEEE Conf. Comp. Vis. Patt. Recog.* (CVPR 2015).
10. Lin, T. Y. et al. (2014). Microsoft COCO: Common objects in context. *Proc. Eur. Conf. Comp. Vis.* (ECCV 2014), pp. 740-755.
11. Mao, J. et al. (2014). Explain images with multimodal recurrent neural networks. *Proc. NIPS 2014*.
12. Platt, J. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, Vol.10(3), pp. 61-74.
13. Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Proc. Intl. Conf. Learn. Rep.* (ICLR 2014).
14. Slaney, M. et al. (2012). Optimal parameters for locality-sensitive hashing. *Proc. IEEE*, Vol. 100(9), pp. 2604-2623.
15. Vedaldi, A. and Lenc, K. (2015). MatConvNet-convolutional neural networks for MATLAB. *Proc. ACM Conf. Multimedia Systems* (MMSys 2015).
16. Vedantam, R. et al. (2015). Cider: Consensus-based image description evaluation. *Proc. IEEE Conf. Comp. Vis. Patt. Recog.* (CVPR 2014).
17. Vinyals, O. et al. (2014). Show and tell: A neural image caption generator. *Proc. IEEE Conf. Comp. Vis. Patt. Recog.* (CVPR 2014).
18. Yang, Y. et al. (2011). Corpus-guided sentence generation of natural images. *Proc. Conf. EMNLP*, pp. 444-454.