Computer vision for image understanding: A comprehensive review

Jácome-Galarza Luis-Roberto[1[0000-0002-2886-3372]], Realpe-Robalino Miguel-Andrés[1[0000-0001-8711-5596]], Chamba-Eras Luis-Antonio[2[0000-0003-3069-9628]], Viñán-Ludeña Marlon-Santiago[3[0000-0003-2692-5899]] and Sinche-Freire Javier-Francisco[3[0000-0001-9631-0449]]

[1] Escuela Superior Politécnica del Litoral. Facultad de Ingeniería en Electricidad y Computación, CIDIS, 09-01-5863 Guayaquil, Ecuador
[2] Universidad Nacional de Loja, Grupo de Investigación en Tecnologías de la Información y Comunicación (GITIC), Carrera de Ingeniería en Sistemas, Loja, Ecuador
[3] Universidad Nacional de Loja, Carrera de Ingeniería en Sistemas, Loja, Ecuador
lrjacome@unl.edu.ec, mrealpe@fiec.espol.edu.ec, lachamba@unl.edu.ec, marlon.vinan@unl.edu.ec, javier.sinche@unl.edu.ec

Abstract. Computer Vision has its own Turing test: Can a machine describe the contents of an image or a video in the way a human being would do? In this paper, the progress of Deep Learning for image recognition is analyzed in order to know the answer to this question. In recent years, Deep Learning has increased considerably the precision rate of many tasks related to computer vision. Many datasets of labeled images are now available online, which leads to pre-trained models for many computer vision applications. In this work, we gather information of the latest techniques to perform image understanding and description. As a conclusion we obtained that the combination of Natural Language Processing (using Recurrent Neural Networks and Long Short-Term Memory) plus Image Understanding (using Convolutional Neural Networks) could bring new types of powerful and useful applications in which the computer will be able to answer questions about the content of images and videos. In order to build datasets of labeled images, we need a lot of work and most of the datasets are built using crowd work. These new applications have the potential to increase the human machine interaction to new levels of usability and user's satisfaction.

Keywords: Computer Vision, Deep Learning, Image Understanding, CNN, Scene Recognition, Object classification.

1 Introduction

Computer vision is a field of computer science that works on enabling computers to see, identify and process images in the same way that human vision does [1]; that indeed is a very complex task. With the exponential growth of the computing power and the increased number of cameras that are installed all over the cities, now it is possible to build automated systems that accomplish with computer vision tasks [2].

However, complex tasks (HIT = Human Intelligence Task [3]) are still under investigation. Among these difficult tasks, we have image understanding in which the computer is able to describe the image in a similar way a human being would do it. Due to its complexity and its potential, it is a worth researching area for new applications. Table 1 presents a list of applications of Image Understanding [4], in which its contribution may change deeply those areas.

Table 1. Image Understanding Applications [4].

Area	Example
Inspection tasks	- Checking the results of casting processes for impurities. - Screening of medical images, screening of plant samples.
Remote Sensing	- Cartography. - Monitoring of traffic along roads, docks, and at airfields. - Management of land resources such as water, forestry, soil.
Making Computer Power More Accessible	- Management information systems that have communication channels wider than systems that are work by typing or pointing. - Document readers, design aids for architects, engineers.
Military Applications	- Tracking moving objects, automatic navigation based on passive sensing, target acquisition and range finding.
Aids for the Partially Sighted	- Systems that read a document and say what was read. - Automatic "guide dog" navigation systems.

1.1 Theoretical framework

Event recognition. An event can be seen as a semantically meaningful human activity, taking place within a selected environment and containing a number of necessary objects. It can also be defined as a descriptive interpretation of the visual world for the blind. In the other hand, for best understanding of images we can use the 5Ws questions: who, where, what, when and how. With event recognition, 3 of the 5 questions can be answered [5]: what? - The event label, where? - The scene environment label, who? - A list of the object categories.

It can be said that event recognition is composed of scene recognition + classification. The SUN dataset [6] (Scene Understanding) is an example of event recognition effort in which all the pictures are organized in hierarchical categories.

Scene recognition. In scene recognition, algorithms learn global statistics of the scene categories. In order to get better results and depending of the application, it is necessary to distinguish indoor scenes from outdoor scenes [7].

Fine-grained recognition. Is the task of distinguishing between visually very similar objects such as the species of a bird, the breed of a dog or the model of an aircraft[8].

Object category recognition. Classifying images can be defined as a collection of regions, describing only their appearance and ignoring their spatial structure. For object categorization there are generative and discriminative models. Image similarity metrics are also used for object recognition, for example, distance metrics: Dssd, Dwarp, Dshift [9]. Object classification can be binary or multiclass classification.

BOW (Bag of Words). Bag of visual words is a vector of occurrence counts of a vocabulary of local image features. A codebook represents an image as sequence of appearance words. BOW can be treated as a supervised or unsupervised task [10]; the scene classification is a supervised task and the object discovery is unsupervised.

1.2 Related works

Many related works refer the use of Convolutional Networks for analyzing visual imagery, because CNN identify parts of objects in its convolution stages. [11]. For using CNN we need large amount of training data. Table 2 enumerates examples of popular specialized Image Datasets that are used for training a Deep Learning model; we have selected them in order to highlight the diverse fields of application.

Table 2. Specialized image databases for machine learning training.

Database	Description	Task	Content
VOC12	Pascal VOC 2012	Object Image Classification / Segmentation (20 classes)	Person, animals, vehicles, indoor objects
MIT67	MIT 67 Indoor Scenes	Scene Image Classification (67 categories)	Indoor places
VOC11s	PASCAL VOC 2011	Object Category Segmentation / Action classification (10 action classes + other)	Person, animals, vehicles, indoor objects
200Birds	UCSD-Caltech 2011-200 Birds dataset	Fine-grained Recognition (200 categories)	Bird species
102Flowers	Oxford 102 Flowers	Fine-grained Recognition (102 categories)	Flower categories
H3Datt	H3D poselets Human 9 Attributes	Attribute Detection (150 categories)	Poselets for person
LFW	Labeled Faces in the Wild	Metric Learning / Face recognition	Famous people in different poses
Oxford5k	Oxford 5k Buildings Dataset	Instance Retrieval (11 landmarks)	Oxford landmarks
Paris6k	Paris 6k Buildings[12]	Instance Retrieval	Paris landmarks
Sculp6k	Oxford Sculptures	Instance Retrieval (10	Sculptures by Hen-

	Dataset	objects to detect)	ry Moore and Auguste Rodin
Holidays	INRIA Holidays Scenes Dataset	Instance Retrieval (500 image groups)	INRIA personal holiday photos
UKB	Uni. of Kentucky Retrieval Benchmark Dataset	Instance Retrieval (2550 classes)	Animals, plants, household objects
IAPR TC-12	IAPR TC-12 Benchmark	Image captioning in English, German and Spanish	Sports and actions, people, animals, cities, landscapes
Flickr 30k	30k Flickr images with captions	Image captioning	Images of different types of objects
MSCOCO	Microsoft Common Objects in Context	Object detection, segmentation, image captioning	Person, animals, vehicles, furniture

Other interesting image datasets are [13] for video surveillance, human health monitoring, human pose, etc., in [14] they have 22.210 fully annotated images with objects and many with parts, in [15] there is a database of 360 degrees panoramas, in [16] there is a database of 400,000 spoken captions for natural images (Places 205 dataset).

In the other hand, in [2] they highlight the use of deep and wide networks like VGGNet [17] and GooLeNet [18].

A very interesting work is [19]; they study human behavior-recognition algorithms to understand transit scenes. They present datasets and implementation details. They highlight that there is still a big gap in analytical skills between a typical security guard and state of the art in image processing algorithms.

There are also many image understanding reviews that are applied to medicine, like [20], in which they introduce the state of the art in image understanding for iris biometrics. They present datasets, applications and conditions that may affect the iris.

Moreover, in [21], they present a survey for video tracking. It consists from simple window tracking to complex models that learn shape and dynamics.

After doing this preliminary review we come up with a research question: The use of image understanding may assist workers to get a better performance compared with traditional methods? What are popular tools for working with image understanding? What limitations are there in the field of image understanding?

2 Materials and Methods

In order to make this paper, we worked in two parts: documental and practical. For the documental part, we included mostly the papers that work with Deep Learning because this technology has given the best results in many computer vision tasks [11]. We also gave importance to papers that use Natural Language Processing methods because it is necessary to generate the text that describe the images. Finally, our last criterion for selecting the relevant papers was its topic. We chose papers in the fields

of Medicine, urban traffic, outdoor and indoor places, human actions and common object detection. In the other hand, we excluded papers that work in other fields like Agriculture, industry and other specialized areas, because we had to limit the depth of the study, considering that image understanding may be applied to any field of knowledge. In order to get the latest scientific information, we looked for academic databases like IEEE Xplore, Google Scholar, Research Gate, etc.; for all these academic databases we used the search string "image understanding". In the other hand, we found papers from 10 or 15 years ago that best describe the basic theory behind computer vision techniques and papers from recent years which describe the latest techniques.

For the practical part of this study, we used the github (public projects repository) to get the code that allows us to work with image understanding. We also followed the Kaggle's tutorials for deep learning in order to have a better knowledge of the key concepts of this technology. We use a machine with the following features: Alienware with Intel Core i7 processor, CPU 2.80 GHz, 64 bits architecture, 16GB of RAM memory, NVIDIA GeForce GTX 1070 graphic card. Table 3 describes the experiments that we conducted.

Table 3. Image understanding in different IDE with datasets.

IDE	Library	Dataset	Task
Matlab	Deep Learning toolbox	AlexNet	Object recognition
Matlab	Deep Learning toolbox	VGG16, VGG19	Object recognition
RStudio	Keras	MNIST	Digit recognition
RStudio	Keras	Fashion MNIST	Clothes classification
RStudio	Keras	ImageNet [22]	Object recognition

We also tried platform solutions available online for image understanding. Table 4 describes them.

Table 4. Platform solutions for image understanding.

Solution	Vendor	Task
Watson [23]	IBM	Image captioning
Caption bot [24]	Microsoft	Image captioning
"Dog breed prediction with Keras" Tensorflow kernel [25]	Kaggle	Dog breed recognition

3 Results

The most relevant works for image understanding are presented in the following lines.

In [26], they present a Bayesian hierarchical model to learn and recognize natural scenes with the advantage that the learning model needs minimal human intervention.

In [27], a visual dictionary in which the nouns of the English language are arranged and related by their semantic meaning [28] is introduced.

In [29], they use Deep Learning to build a model for scene recognition. They also have an online demo where an image can be uploaded and the algorithm describes the content of that image [30].

In [31], they present a model which is able to answer questions about the content of an image. The model uses Long Short-Term Memory (LSTM) to extract the question representation, a Convolutional Neural Network (CNN) to extract the visual representation and LSTM to store the linguistic context in an answer; they use a fusing component to combine the information of the other 3 components. They also show the FM-IQA dataset with 150,000 images and 310,000 question-answer pairs. The model is able to answer questions like "Where is the cat?" giving the answer "On the table"

In [32], they propose neural networks and visual semantic embeddings to predict answers to simple questions about images. They also present a question generation algorithm which converts image descriptions into Question-Answer form.

In [33], they introduce a Region Proposal Network (RPN) that shares full-image convolutional features that enable low-cost region proposals. The RPN is a fully-convolutional network that simultaneously predicts object bounds and object scores at each position. They use RPN and the Fast R-CNN algorithm for training the convolutional features.

Scene labeling is a challenging computer vision task which requires the use of local discriminative features and global context information. In [34], they adopt a deep recurrent convolutional neural network (RCNN) for this task. They use the backpropagation through Time algorithm for an easy and simple training.

In [35], they present a multimodal Recurrent Neural Network (m-RNN) for generating sentence descriptions in order to explain the contents of images. The model consists of a deep RNN for sentences and a deep CNN for images. The model is validated on IAPR TC-12, Flickr 8K and Flickr 30K datasets.

In [36], they introduce an encoder-decoder pipeline that learns a multimodal joint embedding space with images and text. The encoder allows ranking images and sentences while the decoder generates descriptions of images. They also use LSTM to encode sentences.

In [37], they present a model that generates natural language descriptions of images and their regions. They use Convolutional Neural Networks for image regions, bidirectional Recurrent Neural Network architecture to generate descriptions of image regions. The experiments are done with Flickr 8k, Flickr 30k and MSCOCO datasets.

In [38], they trained a large, deep convolutional neural network to classify 1.3 million high-resolution images in the Large Scale Visual Recognition Challenge LSVRC-2010 ImageNet dataset into 1,000 different classes. The neural network has 60 million parameters and 500,000 neurons and 5 convolutional layers, some are connected to max-pooling layers, and 3 fully-connected layers with a final 1000-way softmax. They use the dropout method to reduce overfitting.

In [39], they address the task of learning new visual concepts, and their interactions with other concepts, from few images with sentence descriptions. Their method is able to conceive the semantic meaning of the new words and add them into its word dictionary, so the new concepts will be used to describe images.

In [40], they investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. They evaluate very deep convolutional networks, with up to 19 weight layers. It is claimed that representation depth is beneficial for the classification accuracy.

In [41], they present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation in order to generate natural sentences that describe an image. This is challenging because a descriptor must capture objects and it has to express how the objects interact with each other; moreover the semantic knowledge has to be put in a language like English.

In [42], they introduce an attention based model that automatically learns to describe the content of an image. They train the model using backpropagation techniques and maximize the variational lower bound (ELBO Evidence Lower Bound). They evaluate the model with Flickr 8k, Flickr 30k and MSCOCO.

In [43], they propose the use of visual denotations of linguistic expressions to define denotational similarity metrics. To compute denotational similarities they construct a denotation graph (a subsumption hierarchy over constituents and their denotations, based on a large corpus of 30K images and 150K descriptive captions).

In [44], they present a dataset to improve object recognition models by placing the question of object recognition in the context of scene understanding. They label objects with pre-instance segmentations to aid precise object localization. The dataset contains pictures of 91 object types which are easily recognizable by a 4 year old, having a total of 2.5 million labeled instances in 328k images. They also use a Deformable Parts Model for improving the results of bounding box and segmentation. Finally, they use MSCOCO dataset for training the model.

In [45], they present a simple model that is able to generate descriptive sentences when they give a sample image. The model has a strong focus on the syntax of the descriptions. They train a bilinear model that learns a metric between image representation and phrases used to describe them. The model is then able to infer phrases from a given image sample. They also propose a simple language model that is able to produce relevant descriptions for a given test image using the inferred phrases.

In [46], they propose a method for automatically answering questions about images by bringing together recent advances from NLP and computer vision. They combine discrete reasoning with uncertain predictions by a multiworld approach that represents uncertainty about the perceived world in a Bayesian framework. The system is trained from question-answer pairs. Table 5 shows the use of the proposed method.

In [47] and [48], they propose a method for NLP is which they group words of a human language into a corpus that is indexed in vectors for representing similarity and semantic meaning. In this approach (word embeddings) complex meaning and association can be represented.

Table 5. The proposed multiword approach (I = Individual, S = Set)[46].

Type	Description	Template	Example
I	Counting and colors	How many {color}{object} are in {image_id}?	How many gray cabinets are in image 1?
I	Room type	Which type of the room is depicted in {image_id}?	Which type of the room is depicted in image 1?
I	Superlatives	What is the largest {object} in {image_id}?	What is the largest object in image 1?
S	Negations type 1	Which images do not have {object}?	Which images do not have sofa?
S	Negations type 2	Which images are not {room_type}?	Which images are not bedroom?
S	Negations type 3	Which images have {object} but do not have a {object}?	Which images have desk but not have a lamp?

In [49], they present the Neural Image Caption (NIC) model that generates natural sentences describing a model. It uses Convolutional Neural Networks for computer vision and Recurrent Neural Networks for language generating.

In [50], they present the Novel Visual Concept (NVC) dataset, in this project, the model learns novel visual concepts and their interactions, from a few images with sentences descriptions.

In [51], they propose a multimedia analysis framework to process video and text jointly for understanding events and answering queries. The model produces a parse graph that represents the compositional structures of spatial information (objects and scenes), temporal information (actions and events) and casual information (causalities between events and fluent) in the video and text. The knowledge is represented in a S/T/C-AOG graph (spatial-temporal-causal And-Or Graph).

In [52], they compare the accuracy of models in many tasks like object classification, segmentation, etc., using different image datasets like Pascal VOC 2007, MIT 67 Indoor Scenes, etc.

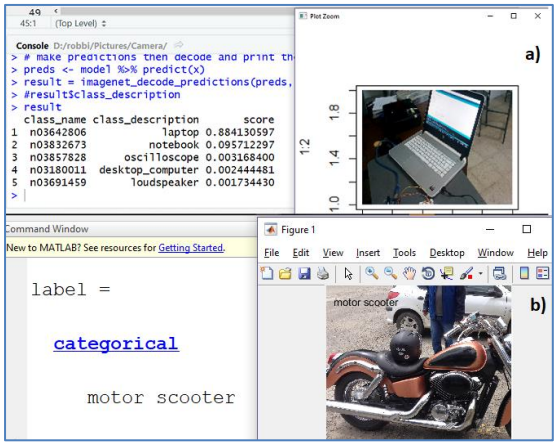

Fig.1. Tests of image classification. a) RStudio + Keras + imageNet b) Matlab + Deep Learning Toolbox + AlexNet

Finally, we tested some of the models using a) RStudio with keras package, b) Matlab with the Deep Learning Toolbox and AlexNet model, having good results as we can see in Figure 1.

4 Discussion

Despite computer vision and image understanding don't get the same results as the human vision does, they have gotten an outstanding progress. Deep Learning models have achieved the best performance so far. Researchers can use pre-trained models like VGGNet or GooLeNet in order to skip the training process and start testing and using these models. For different problems, researchers also can benefit from transfer learning that is a technique that uses most of the layers of a pre-trained model but changes the last layers for a particular classification.

In order to answer our research question: "The use of image understanding may assist workers to get a better performance compared with traditional methods?", we can say that there are many fields like Medicine, security, traffic, industrial process that are experimenting with image understanding giving promising results that may benefit workers to perform much better and get brilliant results.

Popular tools for working with image understanding are the Python/R language in combination with Tensorfow library and Keras framework. Matlab with its deep learning toolbox is also an easy to start tool.

In the other hand, the limitations that we have with deep learning for image understanding are the need of a lot of data and computer power in order to train new models. To overcome these problems, we can build our own specialized datasets but this initiative takes a lot of work. To cope with the need of computer power, we can use cloud platforms like Microsoft cognitive services, IBM Watson, Google cloud services, Amazon cloud services, etc.

5 Conclusions

Describing content of an image is a complex and challenging task which joins computer vision and natural language processing techniques. In this context, for computer vision, researchers use Convolutional Neural Networks, and for Natural Language Processing they often use Recurrent Neural Networks and Long Short-Term Memory. We also found that the depth of layers is important for the accuracy of the classification. Researchers also use many computer vision techniques like segmentation, image classification, scene understanding, and color detection.

For training models we need a lot of human work, so many related works are based on crowd work. Among others the MSCOCO is a popular dataset. For image description, researchers train models with question-answer pairs. Models are able to infer phrases from images. So far, very limited questions can be asked to the model. For

designing the questions that the model can answer, they use templates with the actions that the model can detect.

Newcomers and developers can use pre-trained deep learning models like VGGNet or GooLeNet in order to perform tasks like object identification, image segmentation, image captioning, etc. They can also use transfer learning to easily adapt those models to new datasets. In the other hand we can use languages like Python, Matlab or R language in conjunction with libraries like Tensorflow, OpenCV, Keras to work with image understanding tasks. Finally, web portals like Kaggle, allow us to program with Python or R with Tensorflow in the cloud, we do not need to install them.

For future work, we plan to test some of the models that we described in the present work in order to figure out the best techniques for image understanding. In the other hand, our future goal is to train and test models and datasets of urban traffic applications.

References

1. Techopedia, https://www.techopedia.com/definition/32309/computer-vision, 2019/05/03
2. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826). (2016).
3. Eickhoff, C., & de Vries, A. How crowdsourcable is your task. In Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the fourth ACM international conference on web search and data mining (WSDM) (pp. 11-14). (2011).
4. Draper, R., Hunt D. Smart robots, A handbook of intelligent robotic system, (1985).
5. Li-Jia Li , Li Fei-Fei; What,where and who? Classifying events by scene and object recognition, Proceedings / IEEE International Conference on Computer Vision. IEEE International Conference on Computer Vision, (2007)
6. SUN dataset, https://groups.csail.mit.edu/vision/SUN/hierarchy.html, 2019/03/26
7. L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. What do we see in a glance of a scene? Journal of Vision, 7(1):10, 1–29, http://journalofvision.org/7/1/10/, doi:10.1167/7.1.10. 1, (2007).
8. Coursera, Université nationale de recherche, École des hautes études en sciences économiques. https://www.coursera.org/learn/deep-learning-in-computer-vision/home/welcome 2019/03/12
9. L. Fei-Fei, R. Fergus, and A. Torralba. Recognizing and learning object categories. Short Course CVPR, International Conference on Computer Vision, (2007).
10. Recognizing and Learning Object Categories course, http://people.csail.mit.edu/torralba/shortCourseRLOC/index.html, 2019/03/25
11. Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson; CNN Features off-the-shelf: an Astounding Baseline for Recognition, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 806-813 (2014).
12. The Paris Dataset, http://www.robots.ox.ac.uk/~vgg/data/parisbuildings/, 2019/03/24
13. VisLab – Computer and Robot Vision Laboratory, http://vislab.isr.ist.utl.pt/datasets/#hda, 2019/03/25
14. ADE20K dataset, http://groups.csail.mit.edu/vision/datasets/ADE20K/, 2019/03/26
15. SUN360 panorama database, http://people.csail.mit.edu/jxiao/SUN360/index_high.html, 2019/03/24

16. The Places Audio Caption Corpus, https://groups.csail.mit.edu/sls/downloads/placesaudio/index.cgi, 2019/03/25
17. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, (2014).
18. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9, (2015).
19. Candamo, J., Shreve, M., Goldgof, D. B., Sapper, D. B., & Kasturi, R. Understanding transit scenes: A survey on human behavior-recognition algorithms. IEEE transactions on intelligent transportation systems, 206-224, (2010).
20. Bowyer, K. W., Hollingsworth, K., & Flynn, P. J. Image understanding for iris biometrics: A survey. Computer vision and image understanding, 281-307, (2008).
21. Trucco, E., & Plakas, K. Video tracking: a concise survey. IEEE Journal of Oceanic Engineering, pages 520-529, (2006).
22. Imagenet large scale visual recognition challenge 2013 (ilsvrc2013); http://www.imagenet.org/challenges/LSVRC/2013/, 2019/03/13.
23. IBM Watson demonstration website, https://www.ibm.com/watson/services/visual-recognition/demo/#demo, 2019/05/10.
24. Microsoft Caption Bot, https://www.captionbot.ai/, 2019/05/10.
25. Kaggle's "Dog breed identification" kernel, https://www.kaggle.com/kerneler/starter-dog-breed-identification-0c8eb184-8, 2019/05/10.
26. L. Fei-Fei and P. Perona. A Bayesian hierarchy model for learning natural scene categories. CVPR, (2005)
27. A. Torralba, R. Fergus, W. Freeman; 80 million tiny images: a large dataset for non-parametric object and scene recognition, IEEE Trans on Pattern Analysis and Machine Intelligence, 30(11): 1958-1970, (2008)
28. Tiny Images dataset, http://groups.csail.mit.edu/vision/TinyImages/, 2019/03/01
29. Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, Aude Oliva; Learning Deep Features for Scene Recognition using Places Database, Advances in Neural Information Processing Systems (NIPS) 27, (2014).
30. Cross-Modal Places database, http://projects.csail.mit.edu/cmplaces/, 2019/02/23
31. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering, (2015).
32. M. Ren, R. Kiros, R. Zemel. Exploring Models and Data for Image Question Answering, Conference paper at NIPS, (2015).
33. S. Ren, K. He, R. Girshick, J. Sun. FasterR-CNN:Towards Real-Time Object Detection with Region Proposal Networks, (2016).
34. M. Liang, X. Hu, B. Zhang. Convolutional Neural Networks with Intra-layer Recurrent Connections for Scene Labeling, Proceeding NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. Pages 937-945 (2015)
35. J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. NIPS DeepLearning Workshop, 201, (2014)
36. R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. TACL, (2015).
37. A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, (2015).
38. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, (2012).

39. J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. arXiv preprint arXiv:1504.06692, (2015).
40. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, (2015).
41. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In CVPR, (2015).
42. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044, (2015).
43. P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In ACL, pages 479–488, (2014).
44. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft coco: Common objects in context. arXiv preprint arXiv:1405.0312, (2014).
45. R. Lebret, P. O. Pinheiro, and R. Collobert. Simple image description generator via a linear phrase-based approach. arXiv preprint arXiv:1412.8419, (2014).
46. M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In Advances in Neural Information Processing Systems, pages 1682–1690, (2014).
47. T.Mikolov, M.Karafi´at, L.Burget, J.Cernock`y,and S.Khudanpur. Recurrent neural network based language model. INTERSPEECH, pages 1045–1048, (2010).
48. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, pages 3111–3119, (2013).
49. Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge
50. J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. arXiv preprint arXiv:1504.06692, (2015).
51. K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu. Joint video and text parsing for understanding events and answering queries. MultiMedia, IEEE, 21(2):42–70, (2014).
52. Benchmark of Deep Learning Representations for Visual Recognition, http://www.csc.kth.se/cvap/cvg/DL/ots/, 2019/02/23.