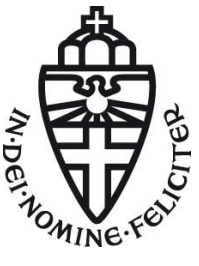
Radboud University Nijmegen



AI Master Thesis

Augmenting a multimodal Recurrent Neural Network with textual context for automatically generating image descriptions

Author:
Flip van Rijn
(s4050614)

Supervisor:
Dr. F. Grootjen

External supervisor Dedicon:
R. Versteeg

SOW-MKI91 AI Master Thesis
Artificial Intelligence
Faculty of Social Sciences
Radboud University Nijmegen

Abstract

1 Introduction

People with a visual impairment are not able to read text or reliably interpret visual input. For this target group special types of books, such as braille or audio books, are made to enable these people to still perform the same task but then via a different medium.

Similarly, user interfaces of programs on computers are often enhanced in such a way that screen readers (text-to-speech software) can easily read out what is being displayed on the screen. Examples of these enhancements are alternative texts for images or reorganizing the layout such that only relevant information is given to the screen reader. However, in a study about the frustrations of screen reader users on the web Lazar et al. [13] aggregated a list of causes of frustration from 100 blind users. This list showed that the often used enhancements, such as alternative text for images, are not always used on websites. The top causes of frustrations include (among others) the layout that causes out of order auditory output from the screen reader, poorly designed forms and no alternative text for images.

The study about the frustrations of screen reader users on the web shows that one cannot rely on websites to provide a user friendly experience, such as a well formatted layout or an additional description of the images. In collaboration with the foundation Dedicon, which is involved in producing braille and audio books, this thesis will mainly focus on enhancing images with relevant textual descriptions. A few examples of what Dedicon already takes care of are the layout of magazines, which will influence the reading order of the screen readers, or school books that are manually edited to make them more accessible for visual impaired people by changing assignments that refer to images. Currently, important images are manually enhanced or replaced with text and this thesis takes a step into automating this process by augmenting current state-of-the-art image annotation techniques with textual context. Multiple fields within artificial intelligence, such as computer vision, machine learning and linguistics, can be used to help the end-user with providing a computer generated description of an image.

The research question that will be answered in this thesis is twofold. The primary question is: 'Does textual context in which an image is present contribute to the performance of automatically generating a description?'. However, the main emphasis lies on the secondary question: 'If so, how can textual context be integrated into a multimodal Recurrent Neural Network model?'

In Section 2 the methods of the study is described in more detail. The results of this study are depicted in Section 3 and the last two sections, Section 4 and 5, discuss these results and draw a general conclusion from them.

The next section provides more clarity of the concept textual context that is used throughout this thesis. The following sections give more background information on the current state-of-the-art in the computer vision field concerning object detection and classification as well as the latest models regarding image description generation.

1.1 Textual context

Before going into more detail of previous studies, first some clarification on the definition of textual context that will be used in the rest of this thesis.

In linguistics context refers to the commonality of implicit information between sentences. In [5] this is exemplified with the following pair of sentences:

> *"John got a new job on Monday. He got up early, shaved, put on his best suit and went to the interview."*

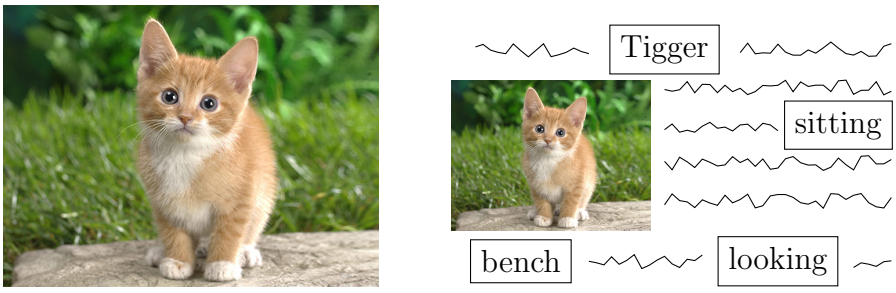
Here the common information is the temporal information about the day that is explicitly available in the first sentence, whereas the second sentence does not state this information. However, the temporal information is still implicitly available in the second sentence due to context.

A similar process seems to be involved around textual context in combination with images. Often an image is surrounded by text that is related to one or more objects or even relations between objects in the image. This textual context could give an additional semantic meaning which cannot be distilled from having solely the image. Examples of such additional meaning are names of objects (e.g. people, animals), place or resolving ambiguity (e.g. partial objects) in an image. Textual context versus no textual context of an arbitrary image is depicted in Figure 1. Without the surrounding text (Figure 1a), the only sensible text that can describe the image could involve the words {*red cat, sitting, stone, grass, looking*}.

However, when a context (Figure 1b) is introduced this changes the meaning of the image, since the red cat now has a name as well as the object which the cat sits on top of and more sensible words can be used {*Tigger, sitting, bench, grass, looking*}. Even though this is a toy example, this illustrates the importance of context especially for visually impaired people where a general description might not be informative enough.

Thus, more formally textual context w.r.t. this thesis is:

> *Words or sentences that provide explicit information that can be used to support or alter the implicit information of an image.*



(a) (b)

Figure 1: Figure 1a shows an arbitrary image without further textual context, whereas an illustration of that same image place within textual context is depicted in Figure 1b. Candidate keywords are highlighted in the rest of the scribbled context.

1.2 Background

Multiple studies use context to categorize [20, 1], segment [23] or describe [17] objects or images each with their own application of context. For categorization the context in which the detected objects are present is useful to eliminate out of place labeled objects. Entirely different, for describing and segmenting objects or images context of lower-level features is used to improve the task at hand.

In light of textual context, Mao et al. [17] harness the idea of novel concepts to improve generated image descriptions for the Novel Visual Concept learning from Sentences (NVCS) task. The core idea here is to extend a pre-trained base model with the capability to update certain concepts, that are already trained on a large dataset, with novel concepts. These novel concepts, which are image-sentence pairs, are included in a small dataset. With limited amount of learning, their model can improve the base model. Throughout the paper an example is used involving the novel Harry Potter concept *quidditch*; the base model would describe an image depicting such a concept with the following sentence:

"A group of people playing a game of soccer"

After a learning session of 100 image-sentence pairs involving quidditch, the same example image now results in:

"A group of people is playing quidditch with a red ball"

The approach of NVCS can be seen as augmenting the generalized model with novel concepts and falls in the definition given in Section 1.1. Here, presenting novel concepts can be seen as providing explicit information of what is depicted on an image. The upside of their approach is that original concepts are not disturbed. However, this process is not one-shot learning, since multiple image-sentence pairs are used. Furthermore, in order to improve other concepts that are incorrectly described, a dataset per concept is required.

Literature shows many studies involving object recognition and classification [1, 9] using Convolutional Neural Networks (CNNs). In these studies objects ranging from inanimate to animate in images are classified, which results in a label that describes to which class each object belongs to. One of the more recent studies [9] claims a performance that even surpasses human performance on classifying images from the ImageNET dataset [22]. While this study focuses on the image classification by labeling images, other studies [16, 18, 27, 3] have focused on the generation of sentences from images by combining a computer vision approach with linguistic models.

For description generation often two types of approaches are used in the literature. The first type is connecting the grammar of a sentence in a description to an object or a relation between objects [11, 3]. Models that use this approach generate sentences for the description that are following the syntactically correctness of the language grammar. The description generation model that is mentioned in [18] follows this first approach. It is trained on the Flickr datasets which consists of images and descriptions. In order to generate meaningful sentences, the model uses co-occurrence statistics to compute the probability distribution within a noun phrase. Furthermore, the characteristics of visually descriptive text are inspected to determine what generally the structure is of this type of text. These statistics are then used in the model

along with the computer vision input (number of objects, labels) to generate novel sentences.

The second type of approach is using probabilistic machine learning to learn the probability density over multimodal input such as text and images. These models also generate sentences for the description, but are not necessarily according to a grammar. This results into more expressive sentences, but may contain less sound grammatical structures. The models in [16, 10, 11] are according to this second approach and the authors describe the model which consists of a multimodal Recurrent Neural Network. What this network makes multimodal network is the multimodal layer. This layer connects the word representations layer with the image feature extraction network that is finally combined into a multimodal feature vector.

In this taxonomy of approaches for description generation, this thesis fits into the latter approach where multimodal input is used to learn a probability density over image-description pairs.

1.3 Base model

The base model that will be used is based on the model by Karpathy and Fei-Fei where a CNN is combined with either a Recurrent Neural Network [10] or a Long-Short Term Memory (LSTM) network [26]. Both approaches use a CNN by Simonyan and Zisserman [24] to extract features from an image and a model which can integrate the image features and a sequence of words.

When an LSTM model is compared to an RNN model an LSTM model tends to outperform the other model on sequence tasks such as generating sentences from images or translation [26].

In the next sections the CNN is explained in more detail as well as the LSTM approach for combining the image and text.

1.3.1 Image features

The CNN part of the multimodal model is designed by Simonyan and Zisserman. They participated in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2014 as the Visual Geometry Group (VGG) with a very deep convolution network and finished with the first and second place. Prior to their work, CNNs mainly were shallow and consisted of only a few layers and convolution filters with large receptive fields were used. However, Simonyan and Zisserman took a different approach by lowering the size of the convolution filters to 3×3 in all layers and by increasing the number of weight layers to 16 or 19. These architectures were not only able to achieve state-of-the-art accuracy on the classification and localization tasks, but were also able to generalize to other datasets.

The last layer of the model has 1000 output units that represent the 1000 categories on which the model has been trained. The activation of those units is the probability density for the categories to be present in the image.

In order to integrate this model in the overall multimodal model, the classification layers are stripped off to expose the rectified linear unit (ReLU) in order to get the more rich internal representation of 4096 abstract features. Together with the text this is the input for the LSTM model.

1.3.2 Combining image and text

Vinyals et al. use an LSTM model to generate sentences from images. The main component of such a model is the memory cell which contains information at every time step about the inputs that have been given up until this point of time. Furthermore, three binary gates govern the behavior of the model by telling the cell to forget the current value, read its input and outputting its new value. A general overview of such a model is shown in Figure 2 where each part itself and the relation with other parts are depicted. The $\odot$ nodes indicate a multiplication with the gate values.

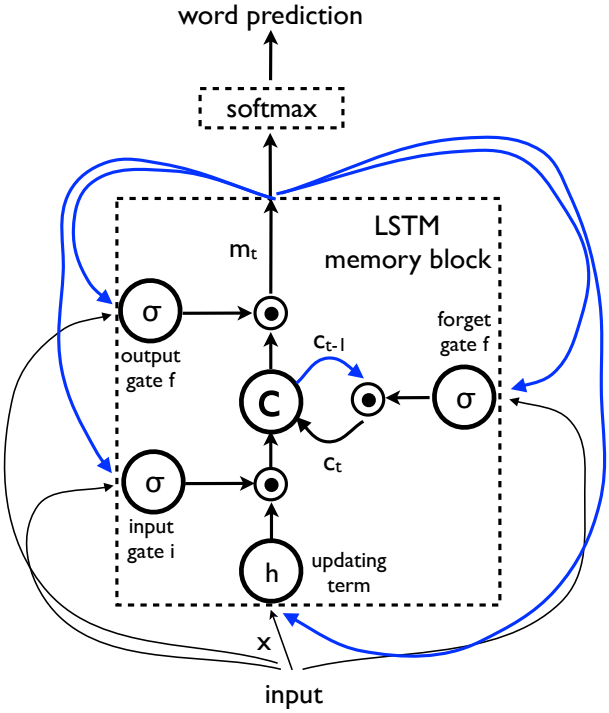


Figure 2: Illustration of an LSTM model adopted from [26]: Memory cell C which is controlled by three gates σ_{output}, σ_{input} and σ_{forget}. The blue lines indicate the recurrent connections from the output m_{t-1} back to the gates at time t.

The image I is fed into the model as input x only at the initial step. After this step the sequence of words in the ground truth sentence $S = (S_1, \ldots, S_N)$ is the input of the model. To mark the start and the ending of the sentence, S_0 and S_{N+1} are special tokens. The words in the sentence are encoded as a 1-of-K binary vector, where K is the number of words in the entire vocabulary (training set) and there is only a single 1 on the index of the word in the vocabulary.

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t) \tag{1}$$

The output of the model is a probability distribution p_t over all words in the vocabulary. Finally, the loss function formalized in Equation 1 is minimized taking into account all the parameters of the model, the image features from the CNN and the word encodings.

2 Methods

2.1 Dataset

This consists of images, their descriptions and the textual context in which these images occur. Many datasets, such as ImageNET [21] and MSCOCO [15], exist of which many only consist of images and their description. Though, next to images and their annotations, in this approach the dataset also needs to include the context in which the image is in. In the literature the MSCOCO and/or Flickr8K/Flickr30K dataset are used for various image retrieval tasks as well as for image annotation tasks. In order to compare the models in the literature with the model in the thesis and since this dataset contains the required components to develop and train a model for automatic image description generation, either one of these datasets can be used.

However, the textual context still is lacking in either of these datasets. Another dataset – ImageCLEF [6], consisting of 500,000 images from various sources on the internet – does have some representation of textual context, since this dataset also contains the webpages on which the images are hosted. This can be used as the context for the images.

Since the literature mainly uses MSCOCO and/or the Flickr datasets and the results of this thesis has to be compared to earlier research, the MSCOCO dataset is used. This dataset is constructed with photos from the Flickr website. However, only the direct link to the photo is provided. Therefore, the Flickr API is used to retrieve the title and the description provided by the author of the photo.

However, the model of this thesis will be used by Dedicon and data that they have might not represent the data that is used by MSCOCO. First of all the language (English vs. Dutch) is different which will have a big impact if this model were to be trained on the MSCOCO dataset and then tested on the Dedicon dataset. Furthermore, the images in the Dedicon dataset might be of an entirely different category compared with the images in the MSCOCO dataset. Therefore, only one of the datasets can be used to both train and test the model on. Once a working prototype of the model has been created and trained, the model is tested on a subset of the dataset that is used.

2.2 Experimental setup

The specifications of the computer that is used during the experiments are as follows: NVIDIA Quadro K2200 GPU with 4GB GDDR5 memory, 16 core Intel Xeon E5-1660 3Ghz CPU and 32GB of RAM.

2.3 Image pre-processing

Regularly words in a description correspond with a region in the image. This information can be used to improve the novel description generation of images. A first step in pre-processing would be to localize objects in the image which then result in bounding-boxes that describe regions of interest. Of the state-of-the-art methods that recognize objects – such as exhaustive search [28, 4] and selective search [23] – selective search by Sande repurposes segmentation for object recognition. Selective search is a much faster method that prefers approximate over exact object localization, has a high

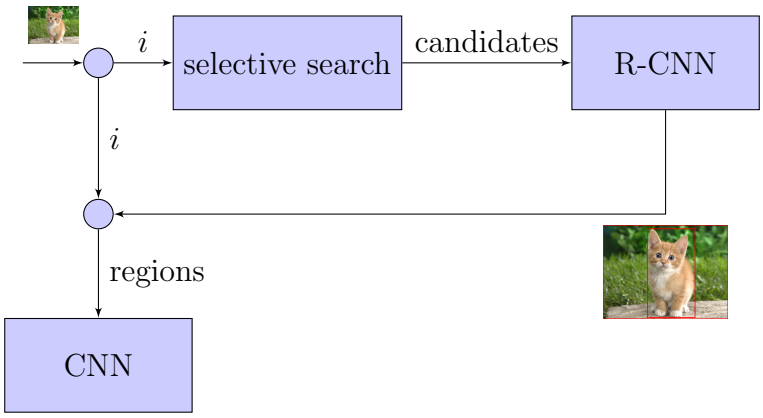


Figure 3: Image pre-processing pipeline. Each image $i \in I$ is processed (in batches or individually) by this pipeline resulting in regions of interest in each image.

recall and permits the use of more expensive features such as bag-of-words. With this method several candidate bounding boxes are generated per image.

The next step is using these candidates as an input for a Regional Convolutional Neural Network (R-CNN) [7]. In essence, the purpose of the R-CNN is to score each candidate using a localization CNN which is pre-trained on the ImageNet dataset. The network receives two inputs: a batch of images and a list of regions of interest. The output of the network is a class posterior probability distribution and bounding-box offsets relative to the candidates.

For the CNN the three best networks – ordered from smallest to largest – are the CaffeNet network [8] (equivalent to the AlexNet network), the VGG-CNN-M-1024 network [2] and the VGG16 network [24]. Due to hardware limitations, the third network was not able to be loaded into the GPU memory. Therefore, the second network is used instead which differs in width compared to CaffeNet. The R-CNN is trained on the 81 classes that are included in the MSCOCO dataset using the VGG-CNN-M-1024 network.

The result of the image pre-processing step are the top 19 detected regions of interest in addition to the whole image. Thus, the representation of an image in bounding box b is:

$$r_i = CNN(I_b) \tag{2}$$

where the CNN converts the sub-image I_b into a 4096-dimensional feature vector. The image pre-processing pipeline is depicted in Figure 3.

2.4 Sentence pre-processing

3 Results

1. BLUE [19]
2. CIDEr-D [25]

3. Meteor [12]
4. ROUGE [14]

4 Discussion

5 Conclusion

Bibliography

[1] Peter Carbonetto, Nando de Freitas, Kobus Barnard, N Freitas, and Kobus Barnard. A statistical model for general contextual object recognition. In *Computer Vision-ECCV 2004*, pages 350–362. Springer, 2004.

[2] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[3] 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 *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6314 LNCS, pages 15–29, 2010. ISBN 364215560X. doi: 10.1007/978-3-642-15561-1_2.

[4] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[5] Ovidiu Fortu and Dan Moldovan. Identification of textual contexts. *Modeling and Using Context*, pages 169–182, 2005. ISSN 03029743. doi: 10.1007/11508373_13. URL `http://link.springer.com/chapter/10.1007/11508373_13`.

[6] Andrew Gilbert, Luca Piras, Josiah Wang, Fei Yan, Emmanuel Dellandrea, Robert Gaizauskas, Mauricio Villegas, and Krystian Mikolajczyk. Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In *CLEF2015 Working Notes*, {CEUR} Workshop Proceedings, Toulouse, France, September 2015. CEUR-WS.org.

[7] Ross Girshick. Fast R-CNN. *arXiv preprint arXiv:1504.08083*, 2015.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, U C Berkeley, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Cvpr'14*, pages 2–9, 2014. ISSN 10636919. doi: 10.1109/CVPR.2014.81. URL `http://arxiv.org/abs/1311.2524`.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv preprint arXiv:1502.01852*, 2015.

[10] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*, 2014.

[11] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *Advances in Neural Information Processing Systems*, pages 1–9, 2014.

[12] Michael Denkowski Alon Lavie. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. *ACL 2014*, page 376, 2014.

[13] Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. What frustrates screen reader users on the web: A study of 100 blind users. *International Journal of human-computer interaction*, 22(3):247–269, 2007.

[14] C Y Lin. Rouge: A package for automatic evaluation of summaries. *Proceedings of the workshop on text summarization branches out (WAS 2004)*, pages 25–26, 2004. ISSN 00036951. URL `papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85`.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. *arXiv preprint arXiv ...*, cs.CV, 2014. doi: 10.1007/978-3-319-10602-1_48.

[16] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

[17] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Learning like a Child: Fast Novel Visual Concept Learning from Sentence Descriptions of Images. *arXiv preprint arXiv:1504.06692*, 2015.

[18] Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Alyssa Mensch, Alex Berg, Xufeng Han, Tamara Berg, Oregon Health, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics, 2012. ISBN 978-1-937284-19-0.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wj Zhu. BLEU: a method for automatic evaluation of machine translation. *... of the 40Th Annual Meeting on ...*, (July):311–318, 2002. ISSN 00134686. doi: 10.3115/1073083.1073135.

[20] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *Proceedings of the IEEE International Conference on Computer Vision*, 2007. ISBN 978-1-4244-1631-8. doi: 10.1109/ICCV.2007.4408986.

[21] Olga Russakovsky and Li Fei-Fei. Attribute learning in large-scale datasets. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6553 LNCS, pages 1–14, 2012. ISBN 978-3-642-35748-0. doi: 10.1007/978-3-642-35749-7_1.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and Others. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014.

[23] Kea Van De Sande. Segmentation as selective search for object recognition. *IEEE International Conference on Computer Vision*, (2):1879–1886, 2011. ISSN 1550-5499. doi: 10.1109/ICCV.2011.6126456.

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

[25] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. *arXiv preprint arXiv:1411.5726*, 2014.

[26] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.

[27] Yezhou Yang, Ching Lik Teo, Hal Daume, and Yiannis Aloimonos. Corpus-Guided Sentence Generation of Natural Images. In *Proceedings of EMNLP*, pages 444–454, 2011. ISBN 1937284115.

[28] Long Zhu, Yuanhao Chen, Alan Yuille, and William Freeman. Latent hierarchical structural learning for object detection. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1062–1069. IEEE, 2010.

Appendices

A Software dependencies

- Caffe, http://caffe.berkeleyvision.org/

B Object classes

These are the 81 object classes in the MSCOCO dataset and are listed in Table 1.

Table 1: List of all 81 classes in the MSCOCO dataset.

Class
background
person
bicycle
car
motorcycle
airplane
bus
train
truck
boat
traffic light
fire hydrant
stop sign
parking meter
bench
bird
cat
dog
horse
sheep
cow
elephant
bear
zebra
giraffe
backpack
umbrella
handbag
tie
suitcase
frisbee
skis
snowboard
sports ball
kite
baseball bat
baseball glove
skateboard
surfboard
tennis racket
bottle

Class
wine glass
cup
fork
knife
spoon
bowl
banana
apple
sandwich
orange
broccoli
carrot
hot dog
pizza
donut
cake
chair
couch
potted plant
bed
dining table
toilet
tv
laptop
mouse
remote
keyboard
cell phone
microwave
oven
toaster
sink
refrigerator
book
clock
vase
scissors
teddy bear
hair drier
toothbrush
blank