SysEng5212 /EE 5370 Project:Automated Image Interpretation via ConvNets & Recurrent Networks

Xiongming Dai

May 4,2018

1 Introduction

A quick glance at an image is sufficient for a human to point out and describe an immense amount of details about the visual scene [3]. However, this remarkable ability has proven to be an elusive task for our visual recognition models. The majority of previous work in visual recognition has focused on labeling images with a fixed set of visual categories and great progress has been achieved in these endeavors [10]. However, while closed vocabularies of visual concepts constitute a convenient modeling assumption, they are vastly restrictive when compared to the enormous amount of rich descriptions that a human can compose.

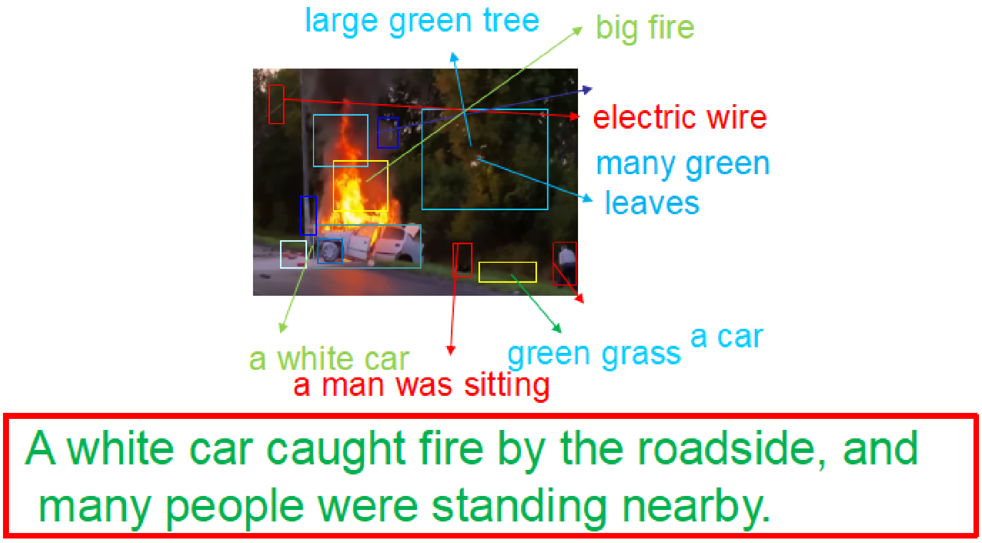
Some pioneering approaches that address the challenge of generating image descriptions have been developed [6]. However, these models often rely on hard-coded visual concepts and sentence templates, which imposes limits on their variety. Moreover, the focus of these works has been on reducing complex visual scenes into a single sentence, which we consider to be an unnecessary restriction.

In this project, we strive to take a step towards the goal of generating dense descriptions of images (Figure 1). The primary challenge towards this goal is in the design of a model that is rich enough to simultaneously reason about contents of images and their representation in the domain of natural language. Additionally, the model should be free of assumptions about specific hard-coded templates, rules or categories and instead rely on learning from the training data. The second, practical challenge is that datasets of image captions are available in large quantities on the internet [7], but these descriptions multiplex mentions of several entities whose locations in the images are unknown.

Our core insight is that we can leverage these large image-sentence datasets by treating the sentences as weak labels, in which contiguous segments of words correspond to some particular, but unknown location in the image. Our approach is to infer these alignments and use them to learn a generative model of descriptions. Concretely, our contributions are twofold:

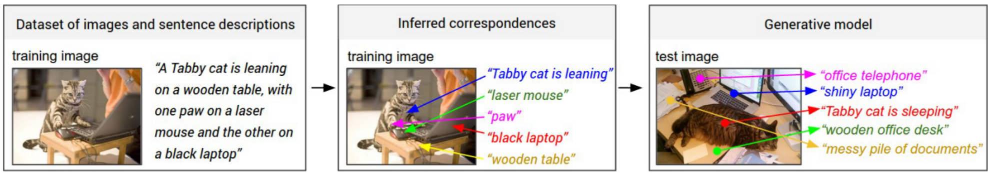
We develop a deep neural network model that infers the latent alignment between segments of sentences and the region of the image that they describe. Our model associates the two modalities through a common, multimodal embedding space and a structured objective. We validate the effectiveness of this approach on image-sentence retrieval experiments in which we surpass the state-of-the-art.

We introduce a multimodal Recurrent Neural Network architecture that takes an input image and generates its description in text. Our experiments show that the generated sentences significantly outperform retrieval based baselines, and produce sensible qualitative



A white car caught fire by the roadside, and many people were standing nearby.

Figure 1: Motivation/Concept Figure: Our model treats language as a rich label space and generates descriptions of image regions.



"A Tabby cat is leaning on a wooden table, with one paw on a laser mouse and the other on a black laptop"

Figure 2: Overview of our approach. A dataset of images and their sentence descriptions is the input to our model(left). Our model first infers the correspondences and then learns to generate novel descriptions.

predictions. We then train the model on the inferred correspondences and evaluate its performance on a new dataset of region-level annotations.

2 Approach

The ultimate goal of our model is to generate descriptions of image regions. During training, the input to our model is a set of images and their corresponding sentence descriptions (Figure 2). We first present a model that aligns sentence snippets to the visual regions that they describe through a multimodal embedding. We then treat these correspondences as training data for a second, multi- modal Recurrent Neural Network model that learns to generate the snippets.

2.1 Learning to align visual and language data

Our alignment model assumes an input dataset of images and their sentence descriptions. Our key insight is that sentences written by people make frequent references to some particular, but unknown location in the image. For example, in Figure 2, the words "Tabby cat is leaning" refer to the cat, the words "wooden table" refer to the table, etc. We would like to infer these latent correspondences, with the eventual goal of later learning to generate these snippets from image regions. We build on the approach of learning to ground dependency tree relations to image regions with a ranking objective. Our contribution is in the use of bidirectional recurrent neural network to compute word representations in the sentence, dispensing of the need to compute dependency trees and allowing unbounded interactions of words and their context in the sentence. We also substantially simplify their objective and show that both modifications improve ranking performance.

We first describe neural networks that map words and image regions into a common, multimodal embedding. Then we introduce our objective, which learns the embedding representations so that semantically similar concepts across the two modalities occupy nearby regions of the space.

Representing images We observe that sentence descriptions make frequent references to objects and their attributes. Thus, we follow the method of Girshick et al. [4] to detect objects in every image with a Region Convolutional Neural Network (RCNN). The CNN is pre-trained on ImageNet [6] and finetuned on the 200 classes of the ImageNet Detection Challenge [45]. Following Karpathy et al. [5], we use the top 19 detected locations in addition to the whole image and compute the representations based on the pixels Ib inside each bounding box as follows:

$$v = W_m[CNN_{\theta_c}(I_b)] + b_m$$

where $CNN(I_b)$ transforms the pixels inside bounding box I_b into 4096-dimensional activations of the fully connected layer immediately before the classifier. The CNN parameters θ_c contain approximately 60 million parameters. The matrix W_m has dimensions $h \times 4096$, where h is the size of the multimodal embedding space (h ranges from $1000 - 1600$ in our experiments). Every image is thus represented as a set of h-dimensional vectors $\{v_i | i = 1, ..., 20\}$.

Representing sentences To establish the inter-modal relationships, we would like to represent the words in the sentence in the same h- dimensional embedding space that the image regions occupy. The simplest approach might be to project every individual word directly into this embedding. However, this approach does not consider any ordering and word context information in the sentence. An extension to this idea is to use word bigrams, or dependency tree relations as previously proposed [5]. However, this still imposes an arbitrary maximum size of the context window and requires the use of Dependency Tree Parsers that might be trained on unrelated text corpora.

To address these concerns, we propose to use a Bidirectional Recurrent Neural Network (BRNN) [11] to compute the word representations.

$$x_t = W_w I_t \tag{1}$$

$$e_t = f(W_e x_t + b_e) \tag{2}$$

$$h_t^f = f(e_t + W_f h_{t-1}^f + b_f) \tag{3}$$

$$h_t^b = f(e_t + W_b h_{t+1}^b + b_b) \tag{4}$$

$$s_t = f(W_d(h_t^f + h_t^b) + b_d) \tag{5}$$

Here, I_t is an indicator column vector that has a single one at the index of the t-th word in a word vocabulary. The weights W_w specify a word embedding matrix that we initialize with 300-dimensional word2vec [8] weights and keep fixed due to overfitting concerns. However,

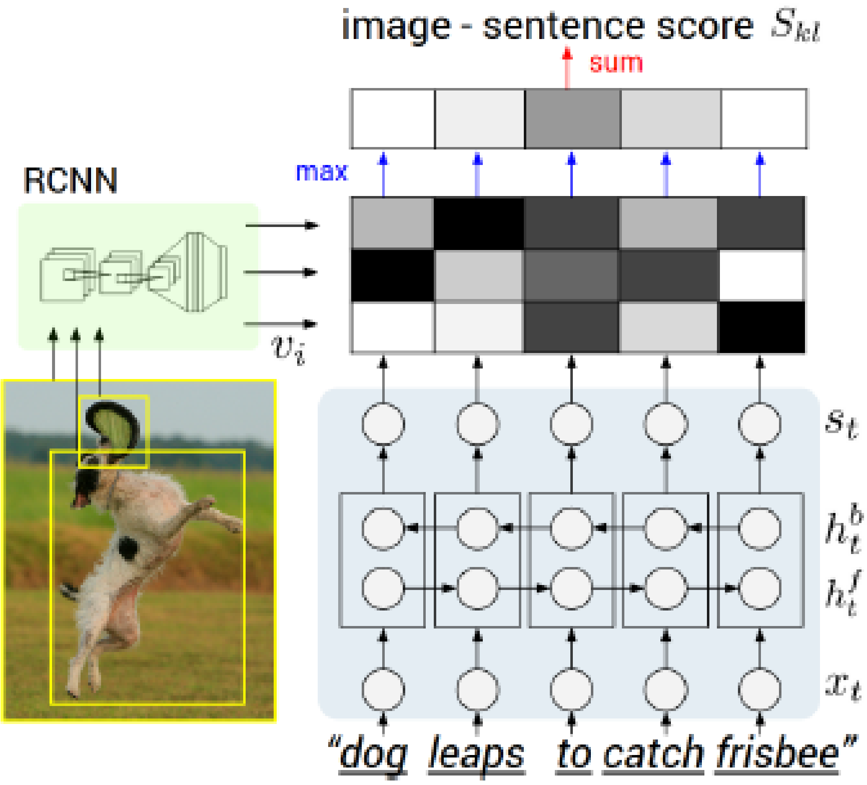


Figure 3: Diagram for evaluating the image-sentence score S_{kl}. Object regions are embedded with a CNN (left). Words (enriched by their context) are embedded in the same multimodal space with a BRNN(right). Pairwise similarities are computed with inner products (magnitudes shown in grayscale) and finally reduced to image-sentence score

in practice we find little change in final performance when these vectors are trained, even from random initialization. Note that the BRNN consists of two independent streams of processing, one moving left to right (h_t^f) and the other right to left (h_t^b) (see Figure 3 for diagram). The final h-dimensional representation s_t for the t-th word is a function of both the word at that location and also its surrounding context in the sentence. Technically, every s_t is a function of all words in the entire sentence, but our empirical finding is that the final word representations (st) align most strongly to the visual concept of the word at that location (I_t). We learn the parameters W_e, W_f , W_b, W_d and the respective biases b_e, b_f, b_b, b_d. A typical size of the hidden rep- resentation in our experiments ranges between 300-600 dimensions. We set the activation function f to the rectified linear unit (ReLU), which computes $f : x \to \max(0, x)$.

Alignment objective We have described the transformations that map every im- age and sentence into a set of vectors in a common h- dimensional space. Since the supervision is at the level of entire images and sentences, our strategy is to formulate an image-sentence score as a function of the individual region- word scores. Intuitively, a sentence-image pair should have a high matching score if its words have a confident support in the image. The objective encourages aligned image-sentences pairs to have a higher score, which is shown as follows:

$$S_{kl} = \sum_{t \in g_l} \sum_{i \in g_k} \max(0, v_i^T s_t) \tag{6}$$

$$C(\theta) = \sum_k [\sum_l \max(0, S_{kl} - S_{kk} + 1) + \sum_l \max(0, S_{lk} - S_{kk} + 1)] \tag{7}$$

Where g_k is the set of image fragments in image k and g_l is the set of sentence fragments in sentence l. The indices k, l range over the images and sentences in the training set. Together with their additional Multiple Instance Learning objective, this score carries the interpretation that a sentence fragment aligns to a subset of the image regions whenever the dot product is positive. $C(\theta)$ denotes the structured loss.

2.2 Multimodal Recurrent Neural Network for generating descriptions

In this section we assume an input set of images and their textual descriptions. These could be full images and their sentence descriptions, or regions and text snippets, as inferred in the previous section. The key challenge is in the design of a model that can predict a variable-sized sequence of outputs given an image. In previously developed language models based on Recurrent Neural Networks (RNNs) [13], this is achieved by defining a probability distri bution of the next word in a sequence given the current word and context from previous time steps. We explore a simple but effective extension that additionally conditions the generative process on the content of an input image. More for mally, during training our Multimodal RNN takes the image pixels I and a sequence of input vectors $(x_1, ..., x_T)$. It then computes a sequence of hidden states $(h_1, ..., h_t)$ and a sequence of outputs $(y_1, ..., y_t)$ by iterating the following recurrence relation for $t = 1$ to T :

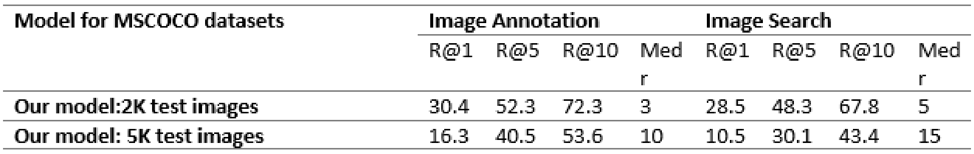
$$b_v = W_{hi}[CNN_{\theta_c}(I)] \tag{8}$$

$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + I(t = 1) \odot b_v) \tag{9}$$

$$y_t = soft\max(W_{oh}h_t + b_o) \tag{10}$$

Where $W_{hi}, W_{hx}, W_{hh}, W_{oh}, x_i$ and b_h, b_o are learnable parameters, and $CNN_{\theta_c}(I)$ is the last layer of a CNN. The output vector y_t holds the (unnormalized) log probabilities of words in the dictionary and one additional dimension for a special END token. Note that we provide the image context vector b_v to the RNN only at the first iteration, which we found to work better than at each time step. In practice we also found that it can help to also pass both b_v, $(W_{hx}x_t)$ through the activation function. A typical size of the hidden layer of the RNN is 512 neurons.

RNN training. The RNN is trained to combine a word (x_t), the previous context (h_{t1}) to predict the next word (y_t). We condition the RNN's predictions on the image information (b_v) via bias interactions on the first step. The training proceeds as follows (refer to Figure 4): We set $h_0 = 0$, x_1 to a special START vector, and the desired label y_1 as the first word in the sequence. Analogously, we set x_2 to the word vector of the first word and expect the network to predict the second word, etc. Finally, on the last step when x_T represents the last word, the target label is set to a special END token. The cost function is to maximize the log probability assigned to the target labels (i.e. Softmax classifier).



Model for MSCOCO datasets	Image Annotation				Image Search			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Our model:2K test images	30.4	52.3	72.3	3	28.5	48.3	67.8	5
Our model: 5K test images	16.3	40.5	53.6	10	10.5	30.1	43.4	15

Figure 4: Image-Sentence ranking experiments results. Med r is the median rank(low is good)

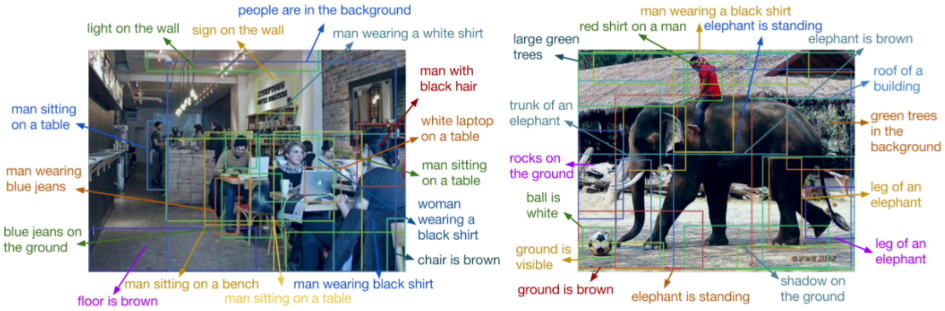


Figure 5: Example alignments predicted by our model.

3 Performance Results

Datasets In this project, We use MSCOCO [7] datasets in our experiments. These datasets contain 123,000 images and each is annotated with 5 sentences using Amazon Mechanical Turk. we use 5,000 images for both validation and testing.
Data Preprocessing We convert all sentences to lower- case, discard non-alphanumeric characters. We filter words to those that occur at least 5 times in the training set, which results in 8791 words for MSCOCO datasets.

3.1 Image-Sentence Alignment Evaluation

We first investigate the quality of the inferred text and image alignments with ranking experiments. We consider a withheld set of images and sentences and retrieve items in one modality given a query from the other by sorting based on the image-sentence score S_{kl}. We report the median rank of the closest ground truth result in the list, which measures the fraction of times a correct item was found among the top K results. The result of these experiments can be found in Figure 4, and example retrievals in Figure 5. We now highlight some of the takeaways.

3.2 Generated Descriptions: Fulframe evaluation

We now evaluate the ability of our RNN model to describe images and regions. We first trained our Multimodal RNN to generate sentences on full images with the goal of verifying that the model is rich enough to support the mapping from image data to sequences of words. For these full image experiments we use the more powerful VGGNet image features [12]. We report the BLEU [9], METEOR [2] and CIDEr [14] scores computed with the coco-caption code. Each method evaluates a candidate sentence by measuring how well it matches a set of five reference sentences written by humans.

Figure 6 and figure 7 show that the Multimodal RNN confidently outperforms this retrieval method. Hence, even with 113,000 train set images in MSCOCO the retrieval approach is inadequate. Additionally, the RNN takes only a fraction of a second to evaluate per image.

4 Conclusion

We introduced a model that generates natural language descriptions of image regions based on weak labels in form of a dataset of images and sentences, and with very few hard- coded assumptions. Our approach features a novel ranking model that aligned parts of visual and language modalities through a common, multimodal embedding. We showed that this model provides state of the art performance on image-sentence ranking experiments. Second, we described a Multimodal Recurrent Neural Network architecture that generates descriptions of visual data. We evaluated its performance on both fullframe and region-level experiments and showed that in both cases the Multimodal RNN outperforms retrieval baselines.

5 Reference

[1]J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei- Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.

[2]M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, 2014.

[3]L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. What do we perceive in a glance of a real-world scene? Journal of vision, 7(1):10, 2007.

[4]R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich fea- ture hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[5]A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment em- beddings for bidirectional image sentence mapping. arXiv preprint arXiv:1406.5679, 2014.

[6]G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In CVPR, 2011.

[7]T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ra- manan, P. Dolla´r, and C. L. Zitnick. Microsoft coco: Com- mon objects in context. arXiv preprint arXiv:1405.0312,2014.

[8]T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, 2013.

[9]K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics, 2002.

[10]O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2014.

[11]M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. Signal Processing, IEEE Transactions on, 1997. [47]K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

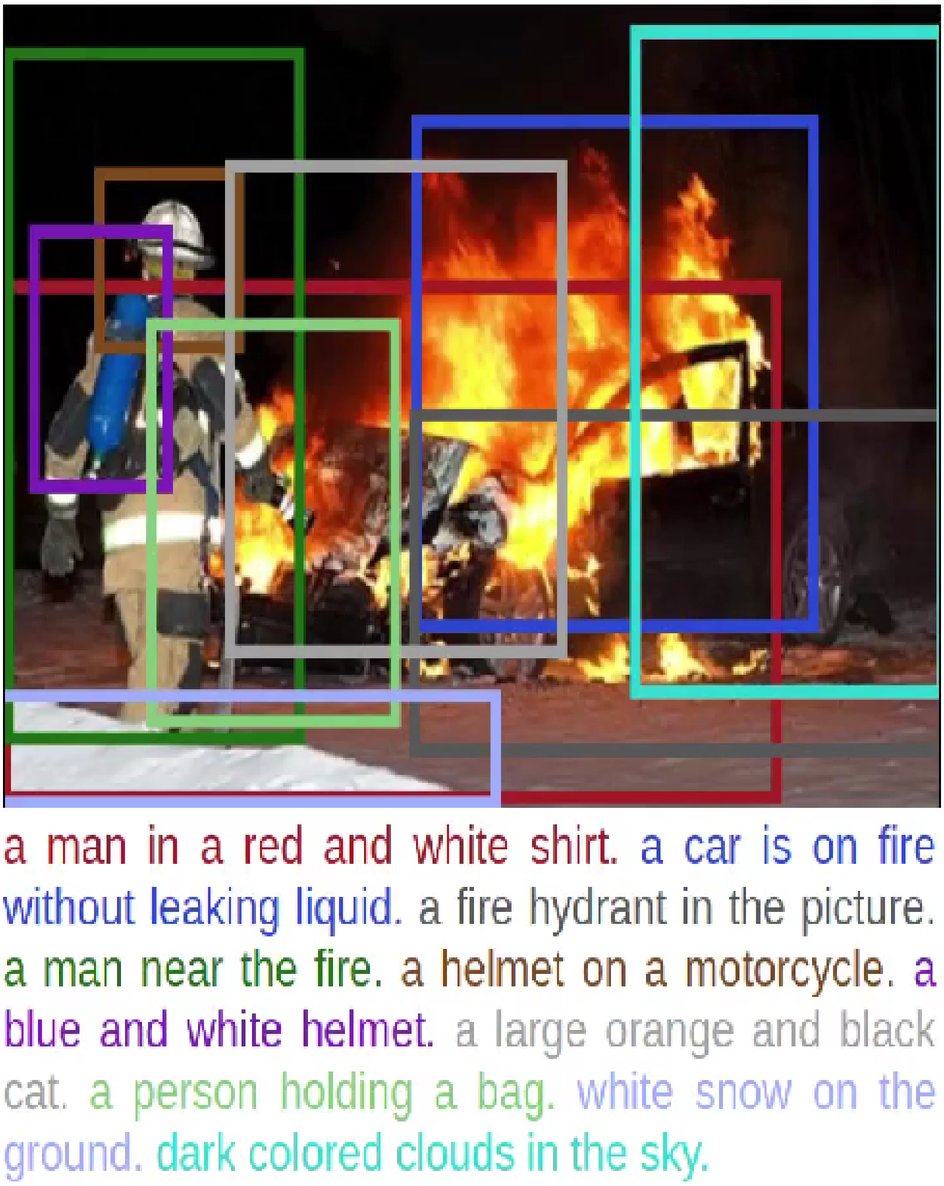
[12]K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[13]I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In ICML, 2011.

[14]R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. CoRR, abs/1411.5726, 2014.

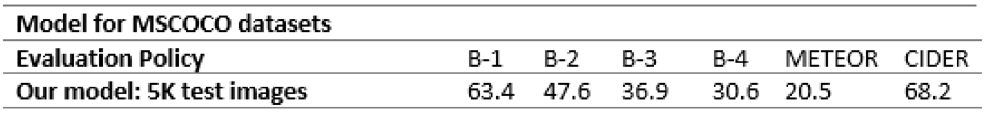
6 Appendix A: Explanation of Files Used in this Project

The appendix A is shown in Figure 8.



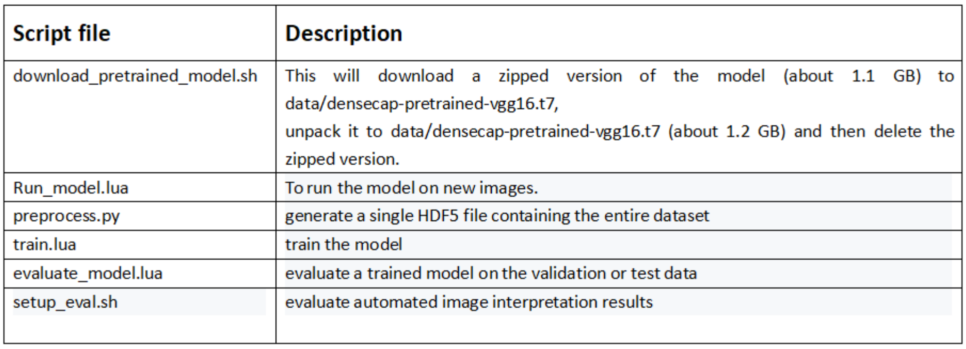
a man in a red and white shirt. a car is on fire without leaking liquid. a fire hydrant in the picture. a man near the fire. a helmet on a motorcycle. a blue and white helmet. a large orange and black cat. a person holding a bag. white snow on the ground. dark colored clouds in the sky.

Figure 6: Example sentences generated by the multimodal RNN for test images.



Model for MSCOCO datasets						
Evaluation Policy	B-1	B-2	B-3	B-4	METEOR	CIDER
Our model: 5K test images	63.4	47.6	36.9	30.6	20.5	68.2

Figure 7: Evaluation of full image predictions on 5,000 test images. B-n is BLEU score that uses up to n-grams. High is good in all columns.



Script file	Description
download_pretrained_model.sh	This will download a zipped version of the model (about 1.1 GB) to data/densecap-pretrained-vgg16.t7, unpack it to data/densecap-pretrained-vgg16.t7 (about 1.2 GB) and then delete the zipped version.
Run_model.lua	To run the model on new images.
preprocess.py	generate a single HDF5 file containing the entire dataset
train.lua	train the model
evaluate_model.lua	evaluate a trained model on the validation or test data
setup_eval.sh	evaluate automated image interpretation results

Figure 8: Listing of Project Files