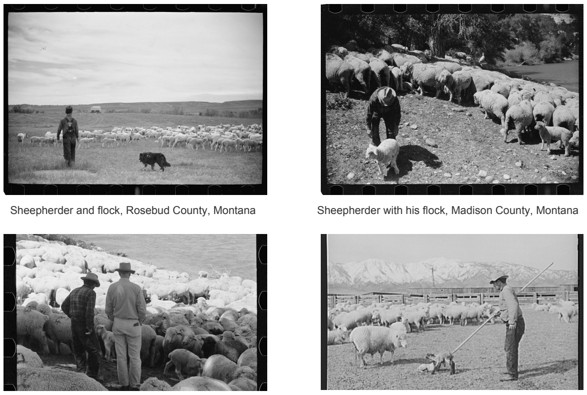
Deep-History-Geek: Automatic Description of Historical Image Photographs

Tianlu Wang, Xuwang Yin, Vicente Ordonez
Department of Computer Science, University of Virginia, Charlottesville, VA 22904
[tw8cb, xy4cm, vicente]@virginia.edu

Abstract

Generating image descriptions has been a popular task in computer vision in recent years. Popular image datasets for this task such as the MS-COCO dataset include image captions collected using crowdsourcing. However, these descriptions might not apply to other image domains where context plays a bigger role. We focus on this work in describing historical photographs obtained from the US Library of Congress [1]*. In our project we explore the use of an encoder-decoder framework where the input image is used as input to a convolutional neural network (CNN), and the output is fed onto a recurrent neural network (RNN) that translates visual features into language. We train our model on descriptions from the US Library of Congress, and compare its performance to a model trained on the MS-COCO Dataset using standard metrics. We show that using both in-domain images and text are crucial to obtain competitive performance on this task.*



Sheepherder and flock, Rosebud County, Montana

Sheepherder with his flock, Madison County, Montana

Rancher and sheepherder with a large flock, Madison County, Montana

Sheepherder with ewe and newborn twin lambs. Dangberg Ranch, Douglas County, Nevada

Figure 1. Here we show some sample images from the Historical Images Dataset that we are using in our work. Our objective is to generate these types of descriptions given these input images. The data is publicly available but we wrote a script to download both images, descriptions, and metadata.

1. Introduction

Describing images using natural language is a challenging task for computers. It requires identifying visual information from an image and then generating coherent language. These are both hard AI tasks that are the subject of increased interest in both Computer Vision and Natural Language processing communities. One recent popular approach consists in applying a convolutional neural network (CNN) on the input image, and then use the output representations of this model as input to a recurrent neural network (RNN) that learns to predict language as a sequence of words. This is popularly known as an encoder-decoder model where the encoder is the CNN and the decoder is an RNN. Previous works using a variation of this model include [8, 4, 3]. In our work we focus on historical images obtained from a publicly available dataset (see sample images in Figure 1). In contrast to these previous works we additionally use some of the metadata (year and location) as an additional context to guide the language generation process. We show that this greatly enhances the quality of results.

2. Related Work

Early work on image descriptions from photographs include [1, 5] where image descriptions are retrieved using a nearest-neighbor approach based on some intermediate image representation. More recent works include the use of neural networks trained in an end-to-end fashion [8, 4, 3]. Most of these previous work concentrate on image descriptions as a generic task, however we show in this work that a generic model doesn't translate well to specific domains such as the one presented in this report on historical images. To the best of our knowledge our work is one of the first to generate descriptions on historical images.

While there have been some previous work on historical images in multimedia [6, 7], most of the previous works have been for the task of image retrieval. We focus in our work instead on image descriptions which could also potentially help in the image retrieval task.

[1] http://photogrammar.yale.edu/map/

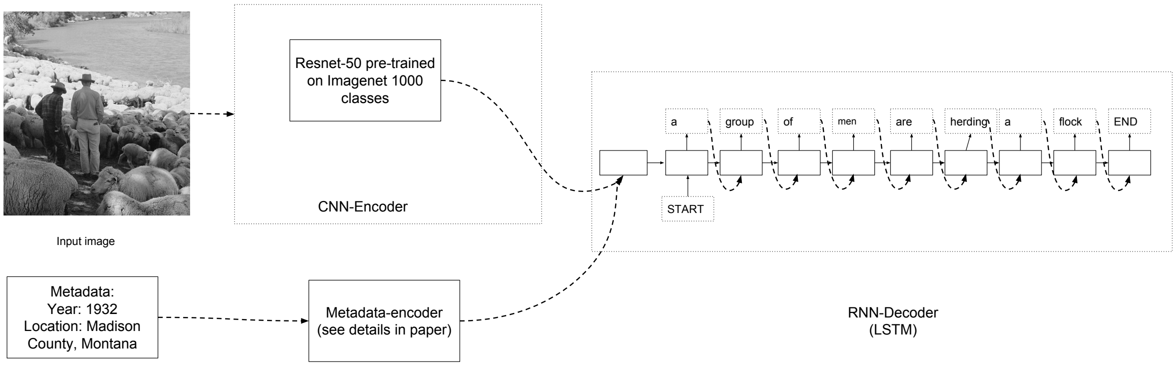

Input image
Resnet-50 pre-trained on Imagenet 1000 classes
CNN-Encoder
a
group
of
men
are
herding
a
flock
END
START
RNN-Decoder (LSTM)
Metadata: Year: 1932 Location: Madison County, Montana
Metadata-encoder (see details in paper)

Figure 2. Here we show a figure detailing our model components. The input is an image and a set of metadata. The output is a sequence of words.

Method	Train-data	BLEU	ROUGE	Perplexity
CNN+LSTM	MS-COCO	[result]	[result]	[result]
CNN+LSTM	HistoricImages	XX	XX	XX
CNN+LSTM+Context	HistoricImages + Year	XX	XX	XX
CNN+LSTM+Context	HistoricImages + Year + Location	XX	XX	XX

Table 1. Preliminary experimental results. XX marks pending results on planned experiments. You should include actual results.

3. Model

We propose a modification of the standard encoder-decoder framework where the encoder consists of a pre-trained convolutional neural network (CNN), and thee decoder consists of a recurrent neural network (CNN). The modification consists on allowing another input to the language model consisting of metadata information from the input images (year, and location). A similar model was proposed by [9] for e-commerce images that have additional categorical metadata. We show a schematic overview of our proposed model in Figure 2. Notice the extra inputs to the language model. In our experiments we used LSTMs for the recurrent networks.

4. Experiments and Results

We used pytorch to implement our model, and used the code from a publicly available implementation as our starting implementation[2]. We first trained the model on the standard MS-COCO dataset and run the model to predict captions directly on the HistoricImages dataset. We show the result of this experiment on the first row in Table 1. We used the 50-layer version of the ResNet architecture [2] as our base deep convolutional network pre-trained in the ILSVRC task on Imagenet. We train our full model using stochastic gradient descent with momentum 0.9 and a mini-batch size of 64. For all experiments, unless indicated otherwise, the learning rate was set to $1e-3$. The models were trained for a maximum of 60 epochs. We show in Figure 3 the loss over time when we trained this baseline model. The images were isotropically scaled and cropped to a 256x256 resolution and processed at a 224x224 resolution by performing random crops and horizontal flips during training, and a center crop at test time. We still have not finished trained our model in the HistoricImages dataset but we have a working implementation. Additionally we have pending to implement the use of year and location as additional context but we have clear direction about how to do this.

In our final report we will additionally include sample outputs from our model showing side-by-side input images with metadata, and output captions obtained from our final model. Given our preliminary experiments we are confident that training on historical images should produce better results.

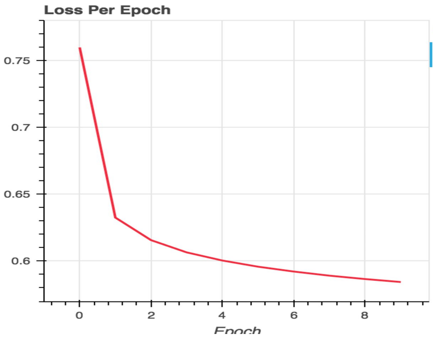


Figure 3. This figure is also included only for illustrative purposes and does not reflect a real experiment. You should include actual results of your experiments subject to the Honor Code at UVA.

[2] https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials

References

[1] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer Berlin Heidelberg, 2010.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[3] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[4] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[5] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, 2011.

[6] V.-W. Soo, C.-Y. Lee, J. J. Yeh, and C.-c. Chen. Using sharable ontology to retrieve historical images. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 197–198. ACM, 2002.

[7] B. Stvilia and C. Jörgensen. Member activities and quality of tags in a collection of historical photographs in flickr. *Journal of the American Society for Information Science and Technology*, 61(12):2477–2489, 2010.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] T. Yashima, N. Okazaki, K. Inui, K. Yamaguchi, and T. Okatani. Learning to describe e-commerce images from noisy online data. In *Asian Conference on Computer Vision (ACCV)*, 2016.