ERCIM "ALAIN BENSOUSSAN"
FELLOWSHIP PROGRAMME

ERCIM
European Research Consortium
for Informatics and Mathematics

Scientific Report

First name / Family name	Desmond Elliott
Nationality	British
Name of the *Host Organisation*	CWI
First Name / family name of the *Scientific Coordinator*	Arjen de Vries
Period of the fellowship	01/10/2014 to 30/09/2015

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

My scientific activities can be split into three main strands.

1. The first stage of my scientific activity focused on reducing the costs associated with human annotation for creating Visual Dependency Representations of images. It is relatively cheap to obtain multiple descriptions of an image from Amazon Mechanical Turk, but expensive to train annotators. We transitioned from human object annotations to automatic annotations using the convolutional neural network object detector (RCNN; Yia et al., 2014). We used the RCNN detector out-of-the-box with the Caffe framework and integrated this into the existing VDR-based image description pipeline. At training time, the image descriptions were used to restrict the application of the pre-trained object detectors. The output of the object detectors are labelled bounding boxes, from which we can automatically predict the dependencies in the Visual Dependency Representation, and thus produce semi-supervised training data. The semi-supervised training data could either be used to supplement the gold-standard data, or to train an image parser from scratch. We found best results using only the automatically predicted structures, and used this to perform end-to-end automatic image description. The result of this activity was published at the ACL.

2. The second activity was to learn about how recent developments in deep learning for Natural Language Processing and Computer Vision could apply to my research. To this end, I collaborated with postdocs from the University of Amsterdam and University of Cambridge on multilingual image description. We implemented novel algorithms that combined machine translation and image description in a single end-to-end model. The output of this activity is under review at ICLR.

3. Finally, I collaborated both within and outside the Information Access Group on collecting novel datasets of images in newspaper contexts. In a collaboration with Laura Hollink, Adriatik Bedjeti, Martin van Harmelen, we collected a dataset of images in online newspapers. In a collaboration with Martijn Kleppe, we collected a datasets of scanned images in Dutch historic newspapers. The outputs of these activities are under review at LREC.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

Published

1. D. Elliott and A. P. de Vries. 2015. Describing Images using Inferred Visual Dependency Representations. In Proceedings of the 53rd Annual Meeting of the Association of Computational Linguistics (ACL '15), Beijing, China.

The Visual Dependency Representation (VDR) is an explicit model of the spatial relationships between objects in an image. In this paper we present an approach to training a VDR Parsing Model without the extensive human supervision used in previous work. Our approach is to find the objects mentioned in a given description using a state-of-the-art object detector, and to use successful detections to produce training data. The description of an unseen image is produced by first predicting its VDR over automatically detected objects, and then generating the text with a template-based generation model using the predicted VDR. The performance of our approach is comparable to a state-ofthe-art multimodal deep neural network in images depicting actions.

Accepted

2. R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B, Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. To appear in Journal of Artificial Intelligence Research.

Automatic description generation from natural images is a challenging problem that has recently received a large amount of interest from the computer vision and natural language processing communities. In this survey, we classify the existing approaches based on how they conceptualize this problem, viz., models that cast description as either a generation, a transfer, or a retrieval problem. We provide a detailed review of existing models, highlighting their advantages and disadvantages. Moreover, we give an overview

of the benchmark image datasets and the evaluation measures that have been developed to assess the quality of machine-generated image descriptions. Finally, we extrapolate future directions in the area of automatic image description generation.

Submitted

3. D. Elliott, S. Frank, and E. Hasler. 2015. Multilingual Image Description with Neural Sequence Models. Submitted to International Conference on Learning Representations.

We introduce multilingual image description, the task of generating descriptions of images given data in multiple languages. This can be viewed as visually-grounded machine translation, allowing the image to play a role in disambiguating language. We present models for this task that are inspired by neural models for image description and machine translation. Our multilingual image description models generate target-language sentences using features transferred from separate models: multimodal features from a monolingual source-language image description model and visual features from an object recognition model. In experiments on a dataset of images paired with English and German sentences, using BLEU and Meteor as a metric, our models substantially improve upon existing monolingual image description models.

4. Laura Hollink, Adriatik Bedjeti, Martin van Harmelen, and Desmond Elliott. 2016. A corpus of images and text in online news. Submitted to Language Resources and Evaluation Conference.

5. Desmond Elliott and Martijn Kleppe. 2016. 1 Million Captioned Dutch Newspaper Images. Submitted to Language Resources and Evaluation Conference.

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

I attended the 2015 Conference of the Association of Computational Linguistics in Beijing, China, to present a paper co-authored with Arjen de Vries.

I also attended the 1st Integrating Vision and Language Summer School in Leuven, Belgium, where I gave a tutorial on Datasets and Evaluation Methods for Image Description.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

I visited Ivan Laptev and Josef Sivic at the WILLOW Research Group in March 2015 at INRIA Paris. I presented my Ph.D and early postdoc research on Visual Dependency Representations, gave a short tutorial on Natural Language Processing, and discussed possible ideas for language and vision research. I visited WILLOW again in June 2015 to

discuss further possibilities for collaboration. I am currently involved in loosely supervising a Ph.D student in the group.