A Treebank of Visual and Linguistic Data

Desmond Elliott
School of Informatics
University of Edinburgh
d.elliott@ed.ac.uk

Frank Keller
School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

Abstract

The treebank is a new resource for researchers working at the intersection between vision and language. It will be a freely-available corpus of images and corresponding text for the development and evaluation of models for natural language generation, image annotation, and structure induction. The treebank differs from existing datasets because it contains syntactic representations of the data, which makes it applicable to a wider range of tasks. The images are provided in their original form, a set of gold-standard object annotations, and as gold-standard visual dependency graphs derived from the annotations and corresponding text. The annotations are made *with respect to* the corresponding text so they cover a wide range of object classes and are directly related to the image description. The visual dependency graphs are generated using a geometric dependency grammar, which defines how relations between pairs of objects can be generated. The text is provided in its original form and as a syntactic dependency tree, which is produced by a state-of-the-art parser. The treebank currently contains several hundred completely annotated pairs of data and is being scaled up to several thousand pairs.

1 Introduction

Modelling the relationships between visual and linguistic data has been studied for several tasks, including image annotation [1] and natural language generation [2]. The datasets used for these tasks have contained annotated images, such as the Corel Image Collection or the PASCAL Visual Object Classification dataset (VOC) [3], or images with descriptions such as the University of Illinois at Urbana Champaign PASCAL Sentence dataset (UIUC) [4]. However, none of these datasets contain structured representations of the visual or linguistic data, which would be valuable for developing structured models of multimodal data. The treebank is an attempt to address this by providing a collection of images, text, and structured representations of both. The treebank contains images from the VOC Action Recognition Taster task and image descriptions collected from untrained annotators on Amazon Mechanical Turk. The image descriptions are a pair of sentences which describe what is happening in the image and the context within which it is taking place. We also include a syntactic dependency representation of the linguistic data. The images have been annotated with labelled polygons guided by the corresponding image descriptions. A structured representation of the images has been created using these labelled polygons and a geometric dependency grammar to create a visual dependency tree. The remainder of this paper presents the images, the image descriptions, the annotations, and the visual dependency trees in more detail.

2 Corpus

The treebank is a collection of annotated images, with multiple descriptions for each image, and syntactic representations of both the images and the text.

2.1 Images

The treebank contains 2,000 images from the the 2011 PASCAL VOC Action Recognition Taster task. The aim of the action recognition task is to predict a labelled bounding box around a person performing an action in an image. The training and validation images contain gold-standard labelled bounding boxes around the person performing the action. The focus of our research is not action recognition but the images are interesting, from a linguistic perspective, because something is happening in them. Ten different actions are represented in the treebank: jumping, phoning, playing an instrument, reading, riding a bike, riding a horse, running, taking a photo, using a computer, and walking.

The image annotations were extended with object annotations derived from the image descriptions. Annotation was performed by trained annotators using descriptions collected from untrained workers on Mechanical Turk, as described in the next section. The annotators drew polygons around the objects visible in both the image and in the image descriptions using the LabelMe annotation tool [5]. An example of the object annotations can be seen in Figure 1(a). The additional image annotations provide a richer representation of the image data in the corpus, compared to the original PASCAL action class annotations.

2.2 Text

The only linguistic data provided with the original images is an action class label for each bounding box. We extended the linguistic annotations by collecting multiple descriptions of each image from untrained annotators on Mechanical Turk. These image descriptions also allow us to obtain an extended vocabulary of labels for image annotation, as outlined in the previous section. An example of the sentences obtained can be seen in Figure 1(a). We asked the annotators to describe an image in two sentences: the first sentence describes the action being performed, the person performing the action and all objects involved in the action; the second sentence describes any other objects in the image that are not directly involved in the action. Sixty-one self-selecting annotators took part in the task and annotated an average of 11.15 images ± 15.73. They were paid $0.04 per task and it took on average 58.63 seconds ± 57.25 to complete a single task, which equates to a payment of approximately $2.57/hour. The average length of a description was 18.96 words ± 5.67 words. Annotators were encouraged to describe fewere than than 30 images to encourage a diverse linguistic dataset. Post-processing of the descriptions removed approximately 30% of the descriptions due to spelling mistakes or incorrect descriptions.

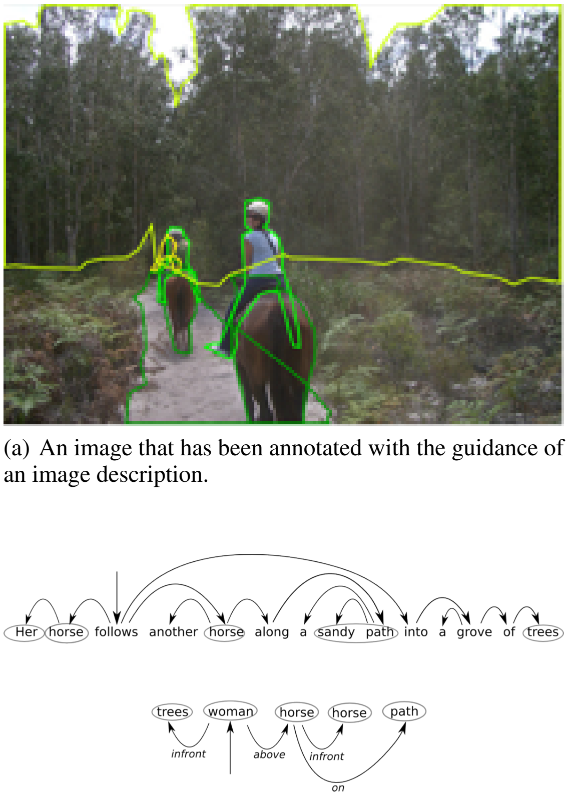
2.3 Dependency Trees

We represent the structure of the image descriptions as syntactic dependency trees, which describes the structure of a sentence in terms of the relationships between the words. The image descriptions were parsed using the MSTPaser [6]; the top of Figure 1(b) shows an an example of a parsed sentence.

Representing the structure of visual data is an active area of research, but all prior approaches in the literature have used phrase-structure grammars [7, 8]. We propose that a phrase-structure grammar may not be suited to this problem because there is not a clear analogue between the linguistic theory behind representing words as phrases and representing image regions or pixels. If the structure of an image is represented as a dependency grammar then it means that we are not insisting on an order or a set of fixed phrase labels over the data. We define a visual dependency tree as a representation of the objects in an image using six geometric relations between pairs of annotated objects. It is created by starting with the subject of the image, which can usually be found near the centre, which forms the head of the first dependency relation. The child of this first dependency is the object of the action.

Annotators determine the relations between the objects based on a three-dimensional co-ordinate system centred on the centre of mass of the head. A line is drawn from this point to the centre of mass of the object in the child relation. The angle formed between the x-axis through the origin and a counter-clockwise rotation to the line between the centres of mass is used to guide the labelling of the dependency relation. These relations can be seen in Table 1. The remaining objects are added

‘‘A young woman on horseback wearing a white helmet, jeans, and a short-sleeved t-shirt smiles and turns to look behind her. Her horse follows another horse along a sandy path into a groove of trees.”



(a) An image that has been annotated with the guidance of an image description.

(b) The linguistic and visual dependency trees.

Figure 1: An example of the images, descriptions, and syntactic representations in the Visual Treebank. (a) shows an image which has been annotated with several objects beyond that of the PASCAL annotations using an image description; (b) shows the linguistic and visual dependency trees for an image-sentence pair.

to the dependency tree by considering which of the objects they interact with or are closest to. The bottom of Figure 1(b) shows an example of a visual dependency tree.

3 Applications

The treebank is an in-development resource for vision and language research. It can be used for many applications, such as generating image descriptions and labelling images; or for tasks such as image segmentation or modelling the structure of visual and linguistic data. For those interested in natural language generation, the annotated images and image descriptions provide a benchmark for systems to generate complex image descriptions. For image annotation research, the annotations cover a wide range of object classes and the descriptions could prove useful for models which exploit the surrounding context of an unlabelled image region; however, the treebank is not intended to replace standardised dataset. For the task of modelling the joint structure of linguistic and visual data, the treebank provides data to develop models which condition on the image annotations and visual dependency trees, or, alternatively, on the linguistic dependency tree.

Relation	Description & Example
X $\overrightarrow{on}$ Y	Most of the pixels of polygon X overlap with polygon Y. In Figure 1(a), the background horse is on the path.
X $\overrightarrow{beside}$ Y	If the angle between the centre of mass of X and the centre of mass of Y lies between 315° and 45° or 135° and 225° then X is **beside** Y. In Figure 1(a), the background horse is beside foreground lady.
X $\overrightarrow{above}$ Y	If the angle between X and Y lies between 45° and 135° then X is **above** Y. In Figure 1(a), the trees are above the foreground lady.
X $\overrightarrow{below}$ Y	If the angle between X and Y lies between 225° and 315° then X is **below** Y. In Figure 1(a), the foreground horse is below the foreground lady.
X $\overrightarrow{infront}$ Y	Another other relation could be applied but the Z-axis relationship between the objects is dominant. In Figure 1(a), the foreground lady could be **below** the trees, but if this image was presented in 3D, the lady would be **infront** of the trees.
X $\overrightarrow{around}$ Y	X almost completely surrounds Y.
X $\overrightarrow{opposite}$ Y	Similar to *beside*, but used when there is a substantial distance between X and Y.

Table 1: The Geometric Dependency Grammar defines seven relations between pairs of annotated polygons. All relations are considered with respect to the centre of a polygon.

4 Acknowledgements

The research presented in this paper was supported by the European Research Council under award number 203427 Synchronous Linguistic and Visual Processing. The authors would like to thank Konstantinos Tsagkaridis for validating the image descriptions and Richard Glassey for feedback on a draft of this paper.

References

[1] Victor Lavrenko, R. Manmatha, and Jiwoon Jeon. A model for learning the semantics of pictures. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, 2003.

[2] Yansong Feng and Mirella Lapata. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 831–839, Los Angeles, California, June 2010. Association for Computational Linguistics.

[3] Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[4] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[5] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, May 2008.

[6] Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[7] Amos J. Storkey and Christopher K. I. Williams. Image modeling with position-encoding dynamic trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):859–871, July 2003.

[8] Richard Socher, Cliff Chiung-Yu Lin, Andrew Ng, and Chris Manning. Parsing natural scenes and natural language with recursive neural networks. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning*, pages 129–136, New York, NY, USA, June 2011. ACM.