Describing Images using Inferred Visual Dependency Representations

Desmond Elliott and **Arjen P. de Vries**
Information Access Group
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
elliott@cwi.nl, arjen@acm.org

Abstract

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-of-the-art multimodal deep neural network in images depicting actions.

1 Introduction

Humans typically write the text accompanying an image, which is a time-consuming and expensive activity. There are many circumstances in which people are well-suited to this task, such as captioning news articles (Feng and Lapata, 2008) where there are complex relationships between the modalities (Marsh and White, 2003). In this paper we focus on generating *literal* descriptions, which are rarely found alongside images because they describe what can easily be seen by others (Panofsky, 1939; Shatford, 1986; Hodosh et al., 2013). A computer that can automatically generate these literal descriptions, filling the gap left by humans, may improve access to existing image collections or increase information access for visually impaired users.

There has been an upsurge of research in this area, including models that rely on spatial relationships (Farhadi et al., 2010), corpus-based relationships (Yang et al., 2011), spatial and visual attributes (Kulkarni et al., 2011), n-gram phrase fusion from Web-scale corpora (Li et al., 2011), tree-substitution grammars (Mitchell et al., 2012), selecting and combining phrases from large image-description collections (Kuznetsova et al., 2012), using Visual Dependency Representations to capture spatial and corpus-based relationships (Elliott and Keller, 2013), and in a generative framework over densely-labelled data (Yatskar et al., 2014). The most recent developments have focused on deep learning the relationships between visual feature vectors and word-embeddings with language generation models based on recurrent neural networks or long-short term memory networks (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Mao et al., 2015; Fang et al., 2015; Donahue et al., 2015; Lebret et al., 2015). An alternative thread of research has focused on directly pairing images with text, based on kCCA (Hodosh et al., 2013) or multimodal deep neural networks (Socher et al., 2014; Karpathy et al., 2014).

We revisit the Visual Dependency Representation (Elliott and Keller, 2013, VDR), an intermediate structure that captures the spatial relationships between objects in an image. Spatial context has been shown to be useful in object recognition and naming tasks because humans benefit from the visual world conforming to their expectations (Biederman et al., 1982; Bar and Ullman, 1996). The spatial relationships defined in VDR are closely, but independently, related to cognitively plausible spatial templates (Logan and Sadler, 1996) and region connection calculus (Randell et al., 1992). In the image description task, explicitly modelling the spatial relationships between observed objects constrains how an image should be described. An example can be seen in Figure 1, where the training VDR identifies the defining relationship between the man and the laptop, which may be re-

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics
and the 7th International Joint Conference on Natural Language Processing, pages 42–52,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

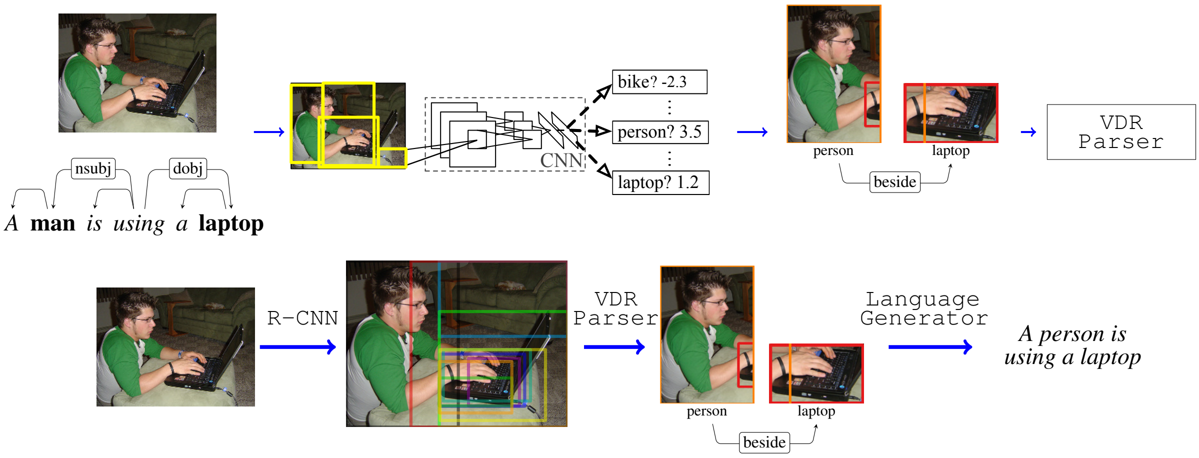


Figure 1: We present an approach to inferring VDR training data from images paired with descriptions (top), and for generating descriptions from VDR (bottom). Candidates for the subject and object in the image are extracted from the description. An object detector[1]searches for the objects and deterministically produces a training instance, which is used to train a VDR Parser to predict the relationships between objects in unseen images. When an unseen image is presented to the model, we first extract N-candidate objects for the image. The detected objects are then parsed into a VDR structure, which is passed into a template-based language generator to produce a description of the image.

alised as a “using”, “typing”, or “working” relationship between the objects.

The main limitation of previous research on VDR has been the reliance on gold-standard training annotations, which requires trained annotators. We present the first approach to automatically inferring VDR training examples from natural scenes using only an object detector and an image description. Ortiz et al. (2015) have recently presented an alternative treatment of VDR within the context of abstract scenes and phrase-based machine translation. Figure 1 shows a detailed overview of our approach. At training time, we learn a VDR Parsing model from representations that are constructed by searching for the subject and object in the image. The description of an unseen image is generated using a template-based generation model that leverages the VDR predicted over the top-N objects extracted from an object detector.

We evaluate our method for inferring VDRs in an image description experiment on the Pascal1K (Rashtchian et al., 2010) and VL2K data sets (Elliott and Keller, 2013) against two models: the bi-directional recurrent neural network (Karpathy and Fei-Fei, 2015, BRNN) and MIDGE (Mitchell et al., 2012). The main finding is that the quality of the descriptions generated by our method depends on whether the images depict an action. In the VLT2K data set of people performing actions, the performance of our approach is comparable to the BRNN; in the more diverse Pascal1K dataset, the BRNN is substantially better than our method. In a second experiment, we transfer the VDR-based model from the VLT2K data set to the Pascal1K data set without re-training, which improves the descriptions generated in the Pascal1K data set. This suggests that refining how we extract training data may yield further improvements to VDR-based image description.

The code and generated descriptions are available at `http://github.com/elliottd/vdr/`.

2 Automatically Inferring VDRs

The Visual Dependency Representation is a structured representation of an image that explicitly models the spatial relationships between objects. In this representation, the spatial relationship between a pair of objects is encoded with one of the following eight options: above, below, beside, opposite, on, surrounds, infront, and behind. Previous work on VDR-based image description has relied on training data from expert human annotators, which is expensive and difficult to scale to other data sets. In this paper, we describe an approach to automatically inferring VDRs using only an object detector and the description of an image. Our aim is to define an automated version

[1] The image of the R-CNN object detector was modified with permission from Girshick et al. (2014).

Relation	Definition
Beside	The angle between the subject and the object is either between 315° and 45° or 135° and 225°.
Above	The angle between the subject and object is between 225° and 315°.
Below	The angle between the subject and object is between 45° and 135°.
On	More than 50% of the subject overlaps with the object.
Surrounds	More than 90% of the subject overlaps with the object.

Table 1: The cascade of spatial relationships between objects in VDR. We always use the last relationship that matches. These definitions are mostly taken from (Elliott and Keller, 2013), except that we remove the 3D relationships. Angles are defined with respect to the unit circle, which has 0° on the right. All relations are specific with respect to the centroid of the bounding boxes.

of the human process used to create gold-standard data (Elliott and Keller, 2013).

An inferred VDR is constructed by searching for the subject and object referred to in the description of an image using an object detector. If both the subject and object can be found in the image, a VDR is created by attaching the detected subject to the detected object, given the spatial relationship between the object bounding boxes. The spatial relationships that can be applied between subjects and objects are defined in the cascade defined in Table 1. The set of relationships was reduced from eight to six due to the difficulty in predicting the 3D relationships in 2D images (Eigen et al., 2014). The spatial relation selected for a pair of objects is determined by applying each template defined in Table 1 to the object pair. We use only the final matching relationship, although future work may consider applying multiple matching relationships between objects.

Given a set of inferred VDR training examples, we train a VDR Parsing Model with the VDR+IMG feature set using only the inferred examples (Elliott et al., 2014). We tried training a model by combining the inferred and gold-standard VDRs but this lead to an erratic parsing model that would regularly predict flat structures instead of object–object relationships. One possibility for this behaviour is the mismatch caused by removing the infront and behind relationships in the inferred training data. Another possible explanation is the gold-standard data contains deeper and more complex structures than the simple object–object structures we infer.

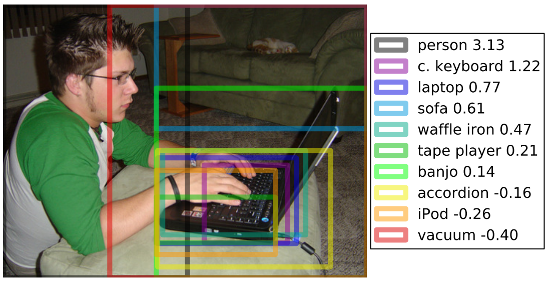


Figure 2: An example of the most confident object detections from the R-CNN object detector.

2.1 Linguistic Processing

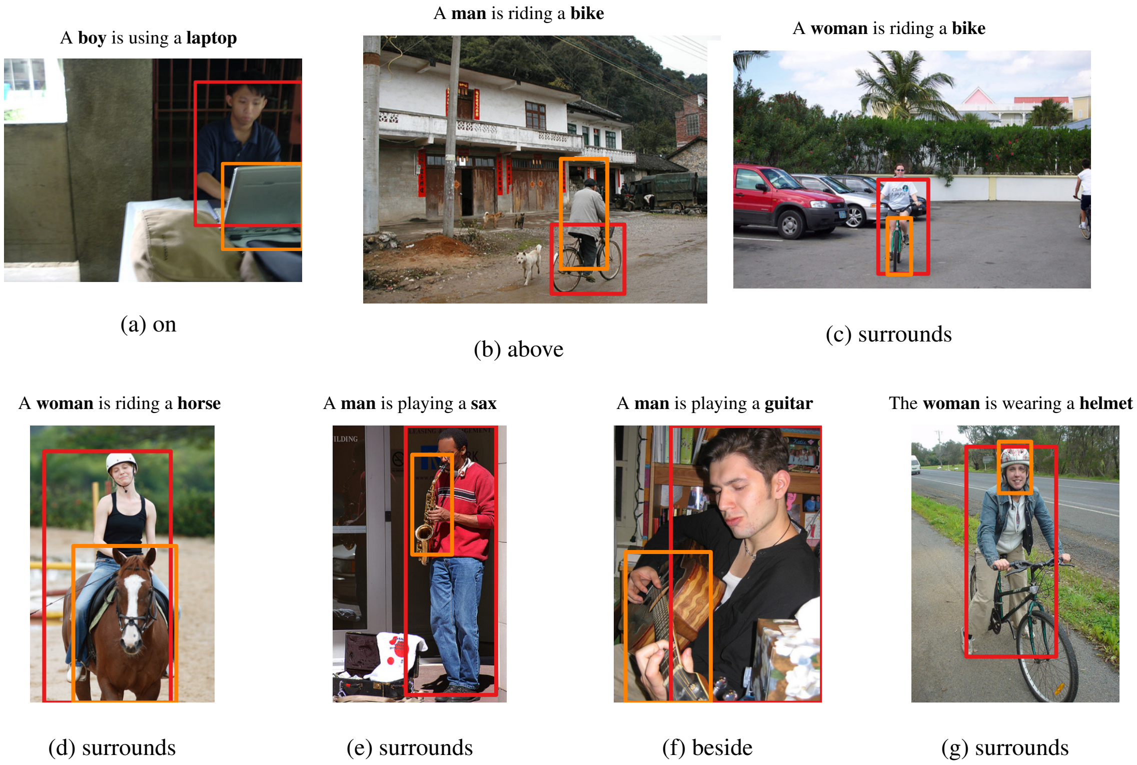
The description of an image is processed to extract candidates for the mentioned objects. We extract candidates from the `nsubj` and `dobj` tokens in the dependency parsed description[2]. If the parsed description does not contain both a subject and an object, as defined here, the example is discarded.

2.2 Visual Processing

If the dependency parsed description contains candidates for the subject and object of an image, we attempt to find these objects in the image. We use the Regions with Convolutional Neural Network features object detector (Girshick et al., 2014, R-CNN) with the pre-trained `bvlc_reference_ilsrvc13` detection model implemented in Caffe (Jia et al., 2014). This object detection model is able to detect 200 different types of objects, with a mean average precision of 31.4% in the ImageNet Large-Scale Visual Recognition Challenge[3] (Russakovsky et al., 2014). The output of the object detector is a bounding box with real-valued confidence scores, as shown in

[2] The descriptions are Part-of-Speech tagged using the Stanford POS Tagger v3.1.0 (Toutanova et al., 2003) with the `english-bidirectional-distsim` pre-trained model. The tagged descriptions are then Dependency Parsed using Malt Parser v 1.7.2 (Nivre et al., 2007) with the `engmalt.poly-1.7` pre-trained model.

[3] The state-of-the-art result for this task is 37.2% using a Network in Network architecture (Lin et al., 2014a); a pre-trained detection model was not available in the Caffe Model Zoo at the time of writing.


A **boy** is using a **laptop**
A **man** is riding a **bike**
A **woman** is riding a **bike**
A **woman** is riding a **horse**
A **man** is playing a **sax**
A **man** is playing a **guitar**
The **woman** is wearing a **helmet**

(a) on

(b) above

(c) surrounds

(d) surrounds

(e) surrounds

(f) beside

(g) surrounds

Figure 3: Examples of the object detections and automatically inferred VDR. In each example, the object detector candidates were extracted from the description and the VDR relationships were determined by the cascade in Table 1. Automatically inferring VDR allows us to learn differences in spatial relationships from different camera viewpoints, such as people riding bicycles.

Figure 2. The confidence scores are not probabilities and can vary widely across images.

The words in a description that refer to objects in an image are not always within the constrained vocabulary of the object labels in the object detection model. We increase the chance of finding objects with two simple back-offs: by lemmatising the token, and transforming the token into its WordNet hypernym parent. If the subject and the object can be found in the image, we create an inferred VDR from the detections, otherwise we discard this training example.

Figure 3 shows a collection of automatically inferred VDRs. One of the immediate benefits of VDR, as a representation, is that we can easily interpret the structures extracted from images. An example of helpful object orientation invariance can be seen in 3 (b) and (c), where VDR captures the two different types of spatial relationships between people and bicycles that are grounded in the verb “riding”. This type of invariance is useful and it suggests VDR can model interacting objects from various viewpoints. We note here the similarities between automatically inferred VDR and Visual Phrases (Sadeghi and Farhadi, 2011). The main difference between these models is that VDR is primarily concerned with object–object interactions for generation and retrieval tasks, whereas Visual Phrases were intended to model person–object interactions for *activity recognition*.

2.3 Building a Language Model

We build a language model using the subjects, verbs, objects, and spatial relationships from the successfully constructed training examples. The subjects and objects take the form of the object detector labels to reduce the effects of sparsity. The verbs are found as the direct common verb parent of the subject and object in the dependency parsed sentence. We stem the verbs using *morpha*, to reduce sparsity, and inflect them in a generated description with +ing using *morphg* (Minnen et al., 2001). The spatial relationship between the subject and object region is used to help constrain language generation to produce descriptions, given observed spatial contexts in a VDR.

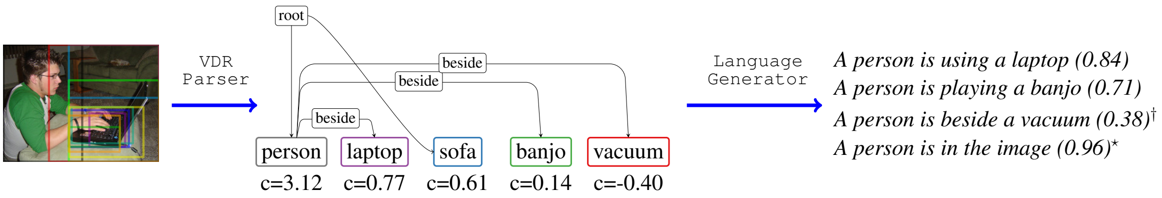

VDR Parser
root
beside
beside
beside
person
laptop
sofa
banjo
vacuum
c=3.12
c=0.77
c=0.61
c=0.14
c=-0.40
Language Generator
A person is using a laptop (0.84)
A person is playing a banjo (0.71)
A person is beside a vacuum (0.38)†
A person is in the image (0.96)⋆

Figure 4: An overview of VDR-constrained language generation. We extract the top-N objects from an image using an object detector and predict the spatial relationships between the objects using a VDR Parser trained over the inferred training data. Descriptions are generated for all parent–child subtrees in the VDR, and the final text has the highest combined corpus and visual confidence. †: only generated is there are no verbs between the objects in the language model; ⋆: only generated if there are no verbs between any pairs of objects in the image.

3 Generating Descriptions

The description of an image is generated using a template-based language generation model designed to exploit the structure encoded in VDR. The language generation model extends Elliott and Keller (2013) with the visual confidence scores from the object detector. Figure 4 shows an overview of the generation process.

The top-N objects are extracted from an image using the pre-trained R-CNN object detector (see Section 2.2 for more details). We remove non-maximal detections with the same class label that overlap by more than 30%. The objects are then parsed into a VDR structure using the VDR Parser trained on the automatically inferred training data. Given the VDR over the set of detected objects, we generate all possible descriptions of the image that can be produced in a depth-first traversal of the VDR. A description is assigned a score that combines the corpus-based evidence and visual confidence of the objects selected for the description. The descriptions are generated using the following template:

DT **head** is V DT **child**.

In this template, **head** and **child** are the labels of the objects that appear in the head and child positions of a specific VDR subtree. V is a verb determined from a subject-verb-object-spatial relation model derived from the training data descriptions. This model captures statistics about nouns that appear as subjects and objects, the verbs between them, and spatial relationships observed in the inferred training VDRs. The verb v that satisfies the V field is determined as follows:

$$v = \arg\max_{v} p(v|head, child, spatial) \qquad (1)$$

$$\begin{aligned} p(v|head,&child, spatial) = \\ &p(v|head) \cdot p(child|v, head) \cdot \qquad (2) \\ &p(spatial|child, v, head) \end{aligned}$$

If no verbs were observed between a particular object–object pair in the training corpus, V is filled using a back-off that uses the spatial relationship label between the objects in the VDR.

The object detection confidence values, which are not probabilities and can vary substantially between images, are transformed into the range [0,1] using $sgm(conf) = \frac{1}{1+e^{-conf}}$. The final score assigned to a description is then used to rank all of the candidate descriptions, and the highest-scoring description is assigned to an image:

$$\begin{aligned} score(head, v,&child, spatial) = \\ &p(v|head, child, spatial) \cdot \qquad (3) \\ &sgm(head) \cdot sgm(child) \end{aligned}$$

If the VDR Parser does not predict any relationships between objects in an image, which may happen if all of the objects have never been observed in the training data, we use a back-off template to generate the description. In this case, the most confidently detected object in the image is used with the following template:

A/An **object** is in the image.

The number of objects N objects extracted from an unseen image is optimised by maximising the sentence-level Meteor score of the generated descriptions in the development data.

4 Experiments

We evaluate our approach to automatically inferring VDR training data in an automatic image description experiment. The aim in this task is to generate a natural language description of an image, which is evaluated directly against multiple reference descriptions.

4.1 Models

We compare our approach against two state-of-the-art image description models. MIDGE generates text based on tree-substitution grammar and relies on discrete object detections (Mitchell et al., 2012) for visual input. We make a small modification to MIDGE so it uses all of the top-N detected objects, regardless of the confidence of the detections[4]. BRNN is a multimodal deep neural network that generates descriptions directly from vector representations of the image and the description (Karpathy and Fei-Fei, 2015). The images are represented by the visual feature vector extracted from the FC_7 layer of the VGG 16-layer convolutional neural network (Simonyan and Zisserman, 2015) and the descriptions are represented as a word-embedding vector.

4.2 Evaluation Measures

We evaluate the generated descriptions using sentence-level Meteor (Denkowski and Lavie, 2011) and BLEU4 (Papineni et al., 2002), which have been shown to have moderate correlation with humans (Elliott and Keller, 2014). We adopt a jack-knifing evaluation methodology, which enables us to report human–human results (Lin and Och, 2004), using MultEval (Clark et al., 2011).

4.3 Data Sets

We perform our experiments on two data sets: Pascal1K and VLT2K. The Pascal1K data set contains 1,000 images sampled from the PASCAL Object Detection Challenge data set (Everingham et al., 2010); each image is paired with five reference descriptions collected from Mechanical Turk. It contains a wide variety of subject matter drawn from the original 20 PASCAL Detection classes. The VLT2K data set contains 2,424 images taken from the trainval 2011 portion of the PASCAL Action Recognition Challenge; each image is paired with three reference descriptions, also collected from Mechanical Turk. We randomly split the images into 80% training, 10% validation, and 10% test.

	VLT2K		Pascal1K	
	Meteor	BLEU	Meteor	BLEU
VDR	16.0	14.8	7.4	9.0
BRNN	18.6	23.7	12.6	16.0
-genders	16.6	17.4	12.1	15.1
MIDGE	5.5	8.2	3.6	9.1
Human	26.4	23.3	21.7	20.6

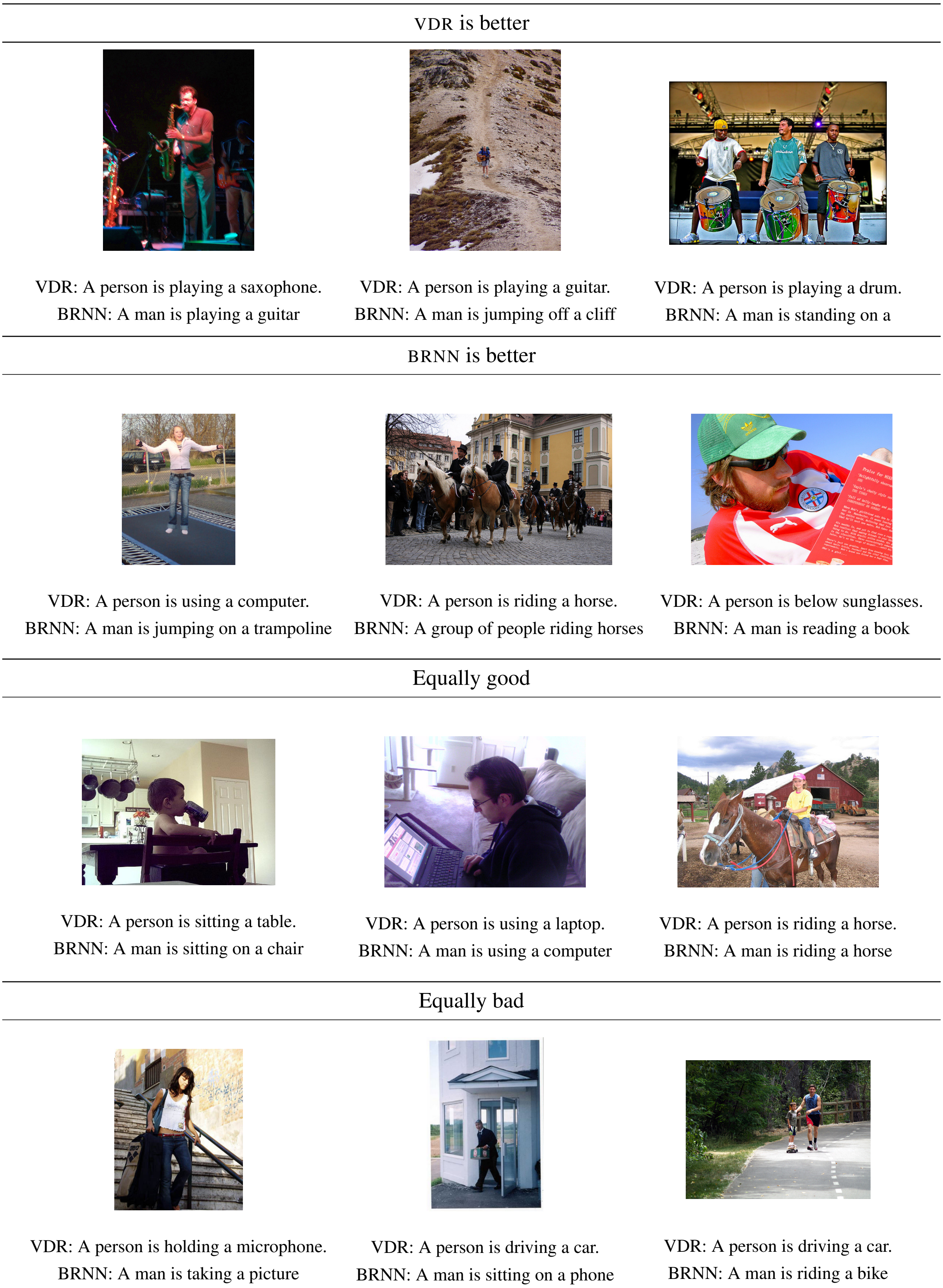
Table 2: Sentence-level evaluation of the generated descriptions. VDR is comparable to BRNN when the images exclusively depict actions (VLT2K). In a more diverse data set, BRNN generates better descriptions (Pascal1K).

4.4 Results

Table 2 shows the results of the image description experiment. The main finding of our experiments is that the performance of our proposed approach VDR depends on the type of images. We found that VDR is comparable to the deep neural network BRNN on the VLT2K data set of people performing actions. This is consistent with the hypothesis underlying VDR: it is useful to encode the spatial relationships between objects in images. The difference between the models is increased by the inability of the object detector used by VDR to predict bounding boxes for three objects (cameras, books, and phones) crucial to describing 30% of the images in this data set. In the more diverse Pascal1K data set, which does not necessarily depict people performing actions, the deep neural network generates substantially better descriptions than VDR and MIDGE. The tree-substitution grammar approach to generating descriptions used by MIDGE does not perform well on either data set.

There is an obvious discrepancy between the BLEU4 and Meteor scores for the models. BLEU4 relies on lexical matching between sentences and thus penalises semantically equivalent descriptions. For example, identifying the gender of a person is important for generating a good description. However, object recognizers are not (yet) able to reliably achieve this distinction, and we only have a single recogniser for "persons". The BRNN generates descriptions with "man" and "woman", which leads to higher BLEU scores than our VDR model, but this is based on corpus statistics than the observed visual information. Me-

[4] In personal communication with Margaret Mitchell, she explained that the object confidence thresholds for MIDGE were determined by visual inspection on held-out data, which we decided was not feasible for 200 new detectors.



VDR is better

VDR: A person is playing a saxophone.
BRNN: A man is playing a guitar

VDR: A person is playing a guitar.
BRNN: A man is jumping off a cliff

VDR: A person is playing a drum.
BRNN: A man is standing on a

BRNN is better

VDR: A person is using a computer.
BRNN: A man is jumping on a trampoline

VDR: A person is riding a horse.
BRNN: A group of people riding horses

VDR: A person is below sunglasses.
BRNN: A man is reading a book

Equally good

VDR: A person is sitting a table.
BRNN: A man is sitting on a chair

VDR: A person is using a laptop.
BRNN: A man is using a computer

VDR: A person is riding a horse.
BRNN: A man is riding a horse

Equally bad

VDR: A person is holding a microphone.
BRNN: A man is taking a picture

VDR: A person is driving a car.
BRNN: A man is sitting on a phone

VDR: A person is driving a car.
BRNN: A man is riding a bike

Figure 5: Examples of descriptions generated using VDR and the BRNN in the VLT2K data set. Keen readers are encouraged to inspect the second image with a magnifying glass or an object detector.

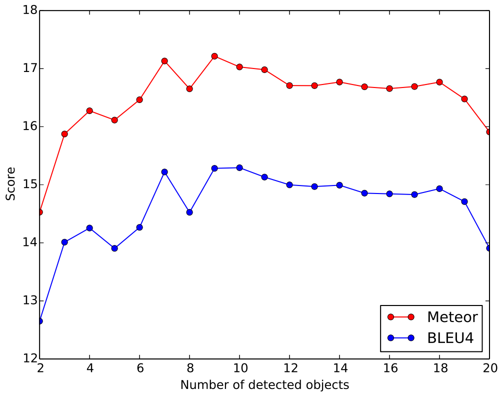


Figure 6: Optimising the number of detected objects against generated description Meteor scores for our model. Improvements are seen until eight objects, which suggests good descriptions do not always need the most confident detections.

teor is able to back-off from “man” or “woman” to “person” and still give partial credit to the description. If we replace the gendered referents in the descriptions generated by the BRNN, its performance on the VLT2K data set drops by 2.0 Meteor points and 6.3 BLEU points.

Figure 6 shows the effect of optimising the number of objects extracted from an image against the eventual Meteor score of a generated description in the validation data. It can be seen that the most confidently predicted objects are not always the most useful objects for generating descriptions. Interestingly, the quality of the descriptions does not significantly decrease with an increased number of detected objects, suggesting our model formulation is appropriately discarding unsuitable detections.

Figure 5 shows examples of the descriptions generated by VDR and BRNN on the VLT2K validation set. The examples where VDR generates better descriptions than BRNN are because the VDR Parser makes good decisions about which objects are interacting in an image. In the examples where the BRNN is better than VDR, we see that the multimodal RNN language model succeeds at describing intransitive verbs, group events, and objects not present in the R-CNN object detector. Both models generate bad descriptions when the visual input pushes them in the wrong direction, seen at the bottom of the figure.

	VLT → Pascal	
	Meteor	BLEU
VDR	7.4 → 8.2	9.1 → 9.2
BRNN	12.6 → 8.1	16.0 → 10.2

Table 3: Sentence-level scores when transferring models directly between data sets with no retraining. The VDR-based approach generates better descriptions in the Pascal1K data set if we transfer the model from the VLT2K data set.

4.5 Transferring Models

The main reason for the low performance of VDR on the Pascal1K data set is that the linguistic and visual processing steps (Section 2) discard too many training examples. We found that only 190 of the 4,000 description in the training data were used to infer VDRs. This was because most of the descriptions did not contain both a subject and an object, as required by our method. This observation led us to perform a second experiment where we transferred the VDR Parsing and Language Generation models between data sets. The aim of this experiment was to determine whether VDR simply cannot work on more widely diverse data sets, or whether the process we defined to replicate human VDR annotation was too strict.

Table 3 shows the results of the model transfer experiment. In general, neither model is particularly good at transferring between data sets. This could be attributed to the shift in the types of scenes depicted in each data set. However, transferring VDR from the VLT2K to the Pascal1K data set *improves* the generated descriptions from 7.4 → 8.2 Meteor points. The performance of BRNN substantially decreases when transferring between data sets, suggesting that the model may be overfitting its training domain.

4.6 Discussion

Notwithstanding the conceptual differences between multi-modal deep learning and learning an explicit spatial model of object–object relationships, two key differences between the BRNN and our approach are the nature visual input and the language generation models.

The neural network model can readily use the pre-softmax visual feature vector from any of the pre-trained models available in the Caffe Model

Zoo, whereas VDR is currently restricted to discrete object detector outputs from those models. The implication of this is that the VDR-based approach is unable to describe 30% of the data in the VLT2K data set. This is due to the object detection model not recognising crucial objects for three of the action classes: cameras, books, and telephones. We considered using the VGG-16 pretrained model from the ImageNet Recognition and Localization task in the RCNN object detector, thus mirroring the detection model used by the neural network. Frustratingly, this does not seem possible because none of the 1,000 types of objects in the recognition task correspond to a person-type of entity. One approach to alleviating this problem could be to use weakly-supervised object localisation (Oquab et al., 2014).

The template-based language generation model used by VDR lacks the flexibility to describe interesting prepositional phrases or variety within its current template. An n-gram language generator, such as the phrase-based approaches of (Ortiz et al., 2015; Lebret et al., 2015), that works within the constraints imposed by VDR structure may generate better descriptions of images than the current template.

5 Conclusions

In this paper we showed how to infer useful and reliable Visual Dependency Representations of images without expensive human supervision. Our approach was based on searching for objects in images, given a collection of co-occurring descriptions. We evaluated the utility of the representations on a downstream automatic image description task on two data sets, where the quality of the generated text largely depended on the data set. In a large data set of people performing actions, the descriptions generated by our model were comparable to a state-of-the-art multimodal deep neural network. In a smaller and more diverse data set, our approach produced poor descriptions because it was unable to extract enough useful training examples for the model. In a follow-up experiment that transferred the VDR Parsing and Language Generation model between data, we found improvements in the diverse data set. Our experiments demonstrated that explicitly encoding the spatial relationships between objects is a useful way of learning how to describe actions.

There are several fruitful opportunities for future work. The most immediate improvement may be found with broader coverage object detectors. It would be useful to search for objects using multiple pre-trained visual detection models, such as a 200-class ImageNet Detection model and a 1,000-class ImageNet Recognition and Localisation model. A second strand of further work would be to relax the strict mirroring of human annotator behaviour when searching for subjects and objects in an image. It may be possible to learn good representations using only the nouns in the POS tagged description. Our current approach strictly limits the inferred VDRs to transitive verbs; images with descriptions such as “A large cow in a field” or “A man is walking” are also a focus for future relaxations of the process for creating training data. Another direction for future work would be to use a n-gram based language model constrained by the structured predicted in VDR. The current template based method is limiting the generation of objects that are being correctly realised in images.

Tackling the aforementioned future work opens up opportunities to working with larger and more diverse data sets such as the Flickr8K (Hodosh et al., 2013), Flickr30K (Young et al., 2014), and MS COCO (Lin et al., 2014b) or larger action recognition data sets such as TUHOI (Le et al., 2014) or MPII Human Poses (Andriluka et al., 2014).

Acknowledgements

We thank the anonymous reviewers for their comments, and members of LaCo at ILLC and WILLOW at INRIA for comments on an earlier version of the work. We thank the Database Architectures Group and the Life Sciences Group at CWI for access to their NVIDIA Tesla K20 GPUs. D. Elliott is funded by an Alain Bensoussain Career Development Fellowship, A. P. de Vries is partially funded by COMMIT/.

References

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *CVPR ’14*, pages 3686–3693, Columbus, OH, US.

Moshe Bar and Shimon Ullman. 1996. Spatial Context in Recognition. *Perception*, 25(3):343–52.

Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. 1982. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177.

JH Clark, Chris Dyer, Alon Lavie, and NA Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *ACL-HTL '11*, pages 176–181, Portland, OR, U.S.A.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *SMT at EMNLP '11*, Edinburgh, Scotland, U.K.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *CVPR '15*, Boston, MA, U.S.A.

David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *NIPS 27*, Lake Tahoe, CA, U.S.A, June.

Desmond Elliott and Frank Keller. 2013. Image Description using Visual Dependency Representations. In *EMNLP '13*, pages 1292–1302, Seattle, WA, U.S.A.

Desmond Elliott and Frank Keller. 2014. Comparing Automatic Evaluation Measures for Image Description. In *ACL '14*, pages 452–457, Baltimore, MD, U.S.A.

Desmond Elliott, Victor Lavrenko, and Frank Keller. 2014. Query-by-Example Image Retrieval using Visual Dependency Representations. In *COLING '14*, pages 109–120, Dublin, Ireland.

Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. 2010. The PASCAL Visual Object Classes Challenge. *IJCV*, 88(2):303–338.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From Captions to Visual Concepts and Back. In *CVPR '15*, Boston, MA, U.S.A.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: generating sentences from images. In *ECCV '10*, pages 15–29, Heraklion, Crete, Greece.

Yansong Feng and Mirella Lapata. 2008. Automatic Image Annotation Using Auxiliary Text Information. In *ACL '08*, pages 272–280, Colombus, Ohio.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *JAIR*, 47:853–899.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In *MM '14*, pages 675–678, Orlando, FL, U.S.A.

Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR '15*, Boston, MA, U.S.A.

Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *NIPS 28*, Montreal, Quebec, Canada.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *CVPR '11*, pages 1601–1608, Colorado Springs, CO, U.S.A.

Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. 2012. Collective Generation of Natural Image Descriptions. In *ACL '12*, pages 359–368, Jeju Island, South Korea.

Dieu-thu Le, Jasper Uijlings, and Raffaella Bernardi. 2014. TUHOI : Trento Universal Human Object Interaction Dataset. In *WVL at COLING '14*, pages 17–24, Dublin, Ireland.

Remi Lebret, Pedro O. Pinheiro, and Ronan Collobert. 2015. Phrase-based Image Captioning. In *ICML '15*, Lille, France, February.

Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *CoNLL '11*, pages 220–228, Portland, OR, U.S.A.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL '04*, pages 605–612, Barcelona, Spain.

Min Lin, Qiang Chen, and Shuicheng Yan. 2014a. Network In Network. In *ICLR '14*, volume abs/1312.4, Banff, Canada.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014b. Microsoft COCO: Common Objects in Context. In *ECCV '14*, pages 740–755, Zurich, Switzerland.

GD Logan and DD Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 492–592. MIT Press.

Junhua Mao, Wei Xu, Yi Yang, Yiang Wang, and Alan L. Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR '15*, volume abs/1412.6632, San Diego, CA, U.S.A.

Emily E. Marsh and Marilyn Domas White. 2003. A taxonomy of relationships between images and text. *Journal of Documentation*, 59(6):647–672.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daum. 2012. Midge : Generating Image Descriptions From Computer Vision Detections. In *EACL '12*, pages 747–756, Avignon, France.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):1.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks. In *CVPR '14*, pages 1717–1724, Columbus, OH, US.

Luis M. G. Ortiz, Clemens Wolff, and Mirella Lapata. 2015. Learning to Interpret and Describe Abstract Scenes. In *NAACL '15*, Denver, CO, U.S.A.

Erwin Panofsky. 1939. *Studies in Iconology*. Oxford University Press.

Kishore Papineni, Salim Roukos, Todd Ward, and WJ Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL '02*, pages 311–318, Philadelphia, PA, U.S.A.

DA Randell, Z Cui, and AG Cohn. 1992. A spatial logic based on regions and connection. In *Principles of Knowledge Representation and Reasoning*, pages 165–176.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *AMT at NAACL '10*, pages 139–147, Los Angeles, CA, U.S.A.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. 2014. ImageNet Large Scale Visual Recognition Challenge.

Mohammad A Sadeghi and Ali Farhadi. 2011. Recognition Using Visual Phrases. In *CVPR '11*, pages 1745–1752, Colorado Springs, CO, U.S.A.

Sara Shatford. 1986. Analysing the Subject of a Picture: A Theoretical Approach. *Cataloging & Classification Quarterly*, 6(3):39–62.

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR '15*, volume abs/1409.1, San Diego, CA, U.S.A.

Richard Socher, Andrej Karpathy, Q Le, C Manning, and A Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *TACL*, 2:207–218.

Kristina Toutanova, Dan Klein, and Christopher D Manning. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *HLT-NAACL '03*, pages 173–180, Edmonton, Canada.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR '15*, Boston, MA, U.S.A.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-Guided Sentence Generation of Natural Images. In *EMNLP '11*, pages 444–454, Edinburgh, Scotland, UK.

Mark Yatskar, Michel Galley, L Vanderwende, and L Zettlemoyer. 2014. See No Evil, Say No Evil: Description Generation from Densely Labeled Images. In **SEM*, pages 110–120, Dublin, Ireland.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.