

# Attributes as Semantic Units between Natural Language and Visual Recognition

Marcus Rohrbach

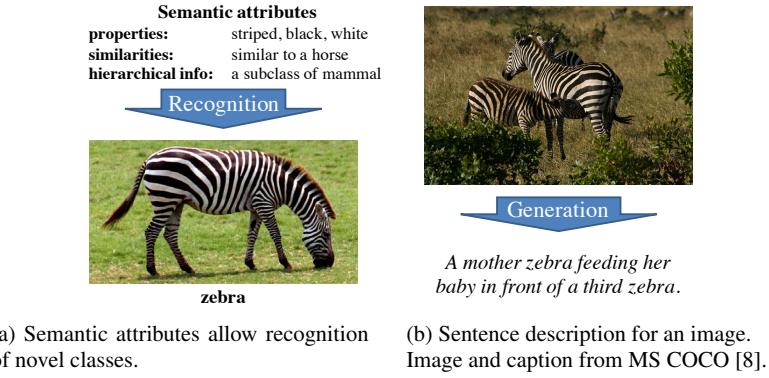
**Abstract** Impressive progress has been made in the fields of computer vision and natural language processing. However, it remains a challenge to find the best point of interaction for these very different modalities. In this chapter we discuss how attributes allow us to exchange information between the two modalities and in this way lead to an interaction on a semantic level. Specifically we discuss how attributes allow using knowledge mined from language resources for recognizing novel visual categories, how we can generate sentence descriptions about images and video, how we can ground natural language in visual content, and finally, how we can answer natural language questions about images.

## 1 Introduction

Computer vision has made impressive progress in recognizing large number of objects categories [83], diverse activities [92], and most recently also in describing images and videos with natural language sentences [91, 89] and answering natural language questions about images [48]. Given sufficient training data these approaches can achieve impressive performance, sometimes even on par with humans [28]. However, humans have two key abilities most computer vision systems lack. On the one hand humans can easily generalize to novel categories with no or very little training data. On the other hand, humans can rely on other modalities, most notably language, to incorporate knowledge in the recognition process. To do so humans seem to be able to rely on compositionality and transferability, which means they can break up complex problems into components, and use previously learned components in other (recognition) tasks. In this chapter we discuss how attributes can form such components which allow to transfer and share knowledge, incorporate external linguistic knowledge, and decompose the challenging problems of visual

---

Marcus Rohrbach  
UC Berkeley EECS and ICSI, Berkeley, USA



**Fig. 1** Examples for textual descriptions and visual content.

description and question answering into smaller semantic units, which are easier to recognize and associate with textual representation.

Let us first illustrate this with two examples. Attribute descriptions given in the form of hierarchical information (*a mammal*), properties (*striped, black, and white*), and similarities, (*similar to a horse*), allow humans to recognize a visual category, even if they never observed this category before. Given this description in form of attributes most humans would be able to recognize the animal shown in Fig. 1(a) as a *zebra*. Furthermore, once humans know that Fig. 1(a) is a *zebra*, they can describe what it is doing within a natural sentence, even if they never saw example images with captions of zebras before (Fig. 1b). A promising way to handle these challenges is to have compositional models which allow interaction between multi-modal information at a semantic level.

One prominent way to model such a semantic level are semantic attributes. As the term “*attribute*” has a large variety of definitions in the computer vision literature we define for the course of this chapter as follows.

**Definition 1.** An *attribute* is a semantic unit, which has a visual and a textual representation.

The first part of this definition, the restriction to a semantic unit is important to discriminate attributes from other representations, which do not have human interpretable meaning, such as image gradients, bag of (visual) words, or hidden representations in deep neural networks. We will refer to these as *features*. Of course for a specific feature, one can try to find or associate it with a semantic meaning or unit, but typically it is unknown and once one is able to identify such a association, one has found a representation for this semantic attribute. The restriction to a semantic unit allows to connect to other sources of information on a semantic level, i.e. a level of meaning. In the second part of the definition we restrict it to semantic units which

can be both represented textually and visually.<sup>1</sup> This is specific for this chapter as we want to exploit the connection between language and visual recognition. From this definition it should also be clear that attributes are not distinct from objects, but rather that objects are also attributes, as they obviously are semantic and have a textual and visual representation.

In this chapter we discuss some of the most prominent directions where language understanding and visual recognition interact. Namely how knowledge mined from language resources can help visual recognition, how we can ground language in visual content, how we can generate language about visual content, and finally how we can answer natural language questions about images, which can be seen as a combination of grounding the question, recognition, and generating an answer. It is clear that these directions cannot cover all potential interactions between visual recognition and language. Other directions include generating visual content from language descriptions [e.g. 102, 45] or localizing images in text i.e. to find where in a text an image is discussed. In the following we first analyze challenges for combining visual and linguistic modalities; afterwards we provide an overview of this chapter which includes a discussion how the different sections relate to each other and to the idea of attributes.

### ***1.1 Challenges for combining visual and linguistic modalities***

One of the fundamental differences between the visual and the linguistic modality is the level of abstraction. The basic data unit of the visual modality is a (photographic) image or video which always shows a specific instance of a category, or even more precisely a certain instance for a specific viewpoint, lighting, pose, time etc. For example Fig. 1(a) shows one specific instance of the category *zebra* from a side view, eating grass. In contrast to this, the basic semantic unit of the linguistic modality are words (which are strings of characters or phonemes for spoken language, but we will restrict ourselves to written linguistic expressions in this chapter). Although a word might *refer* to a specific instance, the word, i.e. the string, always *represents* a category of objects, activities, or attributes, abstracting from a specific instance. Interestingly this difference, instance versus category level representation, is also what defines one of the core challenges in visual recognition and is also an important topic in computational linguistics. In visual recognition we are interested in defining or learning models which abstract over a specific image or video to understand the visual characteristic of a category. In computational lin-

---

<sup>1</sup> There are attributes / semantic units, which are not visual but textually, e.g. smells, tastes, tactile sensory inputs, and ones which are visual but not textual, which are naturally difficult to describe in language, but think of many visual patterns beyond *striped* and *dotted*, for which we do not have name, or the different visual attributes between two people or faces which humans can clearly recognize but which might be difficult to put into words. We also like to note that some datasets such as Animals with Attributes [44] include non-visual attributes, e.g. *smelly*, which might still improve classification performance as they are correlated to visual features.

guistics, when automatically parsing a text, we frequently face the inverse challenge of trying to identify intra and extra linguistic references (co-reference resolution / grounding<sup>2</sup>) of a word or phrase. These problems arise because words typically represent concepts rather than instances and because anaphors, synonyms, hypernyms, or metaphorical expressions are used to refer to the identical object in the real world.

Understanding that the visual and linguistic modalities have different levels of abstraction is important when trying to combine both modalities. In Section 2 we use linguistic knowledge at category rather than instance level for visual knowledge transfer, i.e. we use linguistic knowledge at the level where it is most expressive that is at level of its basic representation. In Section 3, when describing visual input with natural language, we put the point of interaction at a semantic attribute level and leave concrete realization of sentences to a language model rather than inferring it from the visual representation, i.e. we recognize the most important components or attributes of a sentence, which are activities, objects, tools, locations, or scenes and then generate a sentence based on these. In Section 4 we look at a model which grounds phrases which refer to a specific instance by jointly learning visual and textual representations. In Section 5 we answer questions about images by learning small modules which recognize visual elements which are selected according to the question and linked to the most important components in the questions, e.g. questions words/phrases (*How many*), nouns, (*dog*) and qualifiers (*black*). By this composition in modules or attributes, we create an architecture, which allows learning these attributes, which link visual and textual modality, jointly across all questions and images.

## 1.2 Overview and outline

In this chapter we explain how linguistic knowledge can help to recognize novel object categories and composite activities (Section 2), how attributes help to describe videos and images with natural language sentences (Section 3), how to ground phrases in images (Section 4), and how compositional computation allows for effective question answering about images (Section 5). We conclude with directions for future work in Section 6.

All these directions have in common that attributes form a layer or composition which is beneficial for connecting between textual and visual representations. In Section 2, for recognizing novel object categories and composite activities, attributes form the layer where the transfer happens. Attributes are shared across known and novel categories, while information mined from different language resources is able to provide the associations between the know categories and attributes at training time to learn attribute classifiers and between the attributes and novel categories at test time to recognize the novel categories.

---

<sup>2</sup> *co-reference* is when two or more words refer to the same thing or person within text, while *grounding* looks at how words refer to things outside text, e.g. images.

When describing images and videos (Section 3), we first learn an intermediate layer of attribute classifiers, which are then used to generate natural language descriptions. This intermediate layer allows us to reason across sentences at a semantic level and in this way to build a model which generates consistent multi-sentence description. Furthermore, we discuss how such an attribute classifier layer allows us to describe novel categories where no paired image-caption data is available.

When grounding sentences in images, we argue that it makes sense to do this on a level of phrases are rather full sentences, as phrases form semantic units, or attributes, which can be well localized in images. Thus, in Section 4 we discuss how we localize short phrases or referential expressions in images.

In Section 5 we discuss the task of visual question answering which connects these previous sections, as one has to ground the question in the image and then predict or generate an answer. Here we show how we can decompose the question into attributes which are in this case small neural network components, which are composed in a computation graph to predict the answer. This allows us to share and train the attributes across questions and images, but build a neural network which is specific for a given question.

The order of the following sections weakly follows the historic development, where we start with work which appeared at the time when attributes started to become popular in computer vision [43, 18]. And the last section on visual question answering, a problem which requires more complex interactions between language and visual recognition, has only recently become a topic in the computer vision community [48, 4].

## 2 Linguistic knowledge for recognition of novel categories

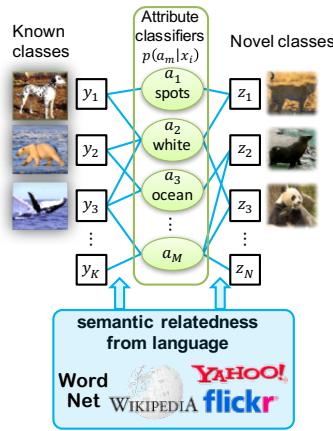
While supervised training is an integral part of building visual, textual, or multi-modal category models, more recently, knowledge transfer between categories has been recognized as an important ingredient to scale to a large number of categories as well as to enable fine-grained categorization. This development reflects the psychological point of view that humans are able to generalize to novel<sup>3</sup> categories with only a few training samples [56, 6]. This has recently gained increased interest in the computer vision and machine learning literature, which look at zero-shot recognition (with no training instances for a class) [44, 17, 58, 59, 22, 53, 21], and one- or few-shot recognition [85, 6, 61]. Knowledge transfer is particularly beneficial when scaling to large numbers of classes where training data is limited [53, 21, 70], distinguishing fine-grained categories [19, 13], or analyzing compositional activities in videos [22, 72].

Recognizing categories with no or only few labeled training instances is challenging. In this section we first discuss how we can build attribute classifiers using

---

<sup>3</sup> We use “novel” throughout this chapter to denote categories with no or few labeled training instances.

**Fig. 2** Zero-shot recognition with the Direct Attribute Prediction model [43] allows recognizing unseen classes  $z$  using an intermediate layer of attributes  $a$ . Instead of manually defined associations between classes and attributes (cyan lines), Rohrbach et al. [69] reduce supervision by mining object-attribute association from language resources, such as Wikipedia, WordNet, and image or web search.



only category-labeled image data and different language resources which allow recognize novel categories (Section 2.1). And then, to further improve this transfer learning approach, we discuss how to additionally integrate instance similarity and labeled instances of the novel classes if available (Section 2.2). Furthermore we discuss what changes have to be made to apply similar ideas to composite activity recognition (Section 2.3).

## 2.1 Semantic relatedness mined from language resources for zero-shot recognition

Lampert et al. [43, 44] propose to use attribute based recognition to allow recognizing unseen categories based on their object-attribute associations. Their Direct Attribute Prediction (DAP) model is visualized in Fig. 2. Given images which are labeled with known category labels  $y$  and object-attribute associations  $a_m^y$  between categories and attributes, we can learn attribute classifier  $p(a_m|x_i)$  for an image  $x_i$ . This allows to recognize novel categories  $z$  if we have associations  $a_m^z$ .

To scale the approach to a larger number of classes and attributes, Rohrbach et al. [69, 73, 70] show how these previously manual defined attribute associations  $a_m^y$  and  $a_m^z$  can be replaced with associations mined automatically from different language resources. Table 1(a) compares several language resources and measures to estimate semantic relatedness to determine if a class should be associated with a specific attribute. Yahoo Snippets [7, 73], which computes co-occurrence statistics on summary snippets returned by search engines, shows the best performance of all single measures. Rohrbach et al. [73] also discuss several fusion strategies to get more robust measures by expanding the attribute inventory with clustering and combining several measures, which can achieve performance on par with manually defined associations (second last versus last line in Table 1a).

Language Resource	Measure	in	AUC		images:	test	AUC + train cls*
WordNet [20], path	Lin measure [46]	[69]	60.5				
Yahoo Web, hit count [54]	Dice coef. [11, 82]	[69]	60.4				
Flickr Img, hit count [69]	Dice coef. [11, 82]	[69]	70.1				
Yahoo Img, hit count [69]	Dice coef. [11, 82]	[69]	71.0				
Wikipedia [69]	ESA [23, 98]	[69]	69.7				
Yahoo Snippets [7]	Dice/Snippets [73]	[73]	76.0				
Yahoo Img	Expanded attr.	[73]	77.2				
Combination	Classifier fusion	[73]	75.9				
Combination	Expanded attr.	[73]	79.5				
manual [43]		[73]	79.2				

(a) Attribute-based zero-shot recognition.

(b) Attributes versus direct-similarity, reported in [73].

**Table 1** Zero-shot recognition on AwA dataset [43]. Results for different language resources to mine association. Trained on 92 images per class, mean area under the ROC curve (AUC) in %.

As an alternative to attributes, Rohrbach et al. [69] also propose to directly transfer information from most similar classes which does not require an intermediate level of attributes. While this achieves higher performance when the test set only contains novel objects, in the more adversarial settings, when the test set also contains images from the known categories, the direct similarity based approach significantly drops in performance as can be seen in Table 1(b).

Rohrbach et al. [70] extend zero-shot recognition from the 10 unseen categories in the AwA dataset to a setting of 200 unseen ImageNet [9] categories. One of the main challenges in this setting is, that there are no pre-defined attributes on this dataset available. Rohrbach et al. propose to mine part-attributes from WordNet [20] as ImageNet categories correspond to WordNet synsets. Additionally, as the known and unknown classes are leaf nodes of the ImageNet hierarchy, inner nodes can be used to group leaf nodes, similar to attributes. Also, the closest known leaf node categories can transfer to the corresponding unseen leaf category.

An alternative approach is DeViSE [21] which learns an embedding into a semantic skip-gram word-space [55], trained on Wikipedia documents. Classification is achieved by projecting an image in the word-space and taking the closest word as label. Consequently this also allows for zero-shot recognition.

Table 2 compares the different approaches. The hierarchical variants [70] performs best, also compared to DeViSE [21] which relies on more powerful CNN [42] features. Further improvements can be achieved by metric learning [53]. As a different application, Mrowca et al. [57] show how such hierarchical semantic knowledge allows to improve large scale object detection not just classification. While the WordNet hierarchy is very reliable as it was manually created, the attributes are restricted to part attributes and the mining is not as reliably. To improve in this challenging setting, we discuss next how one can exploit instance similarity and few labeled examples if available.

**Object - Attribute Associations**

Yahoo Img	71.0	73.2 ( +2.2 )
Classifier fusion	79.5	78.9 ( -0.6 )
manual	79.2	79.4 ( +0.2 )

#### Direct Similarity

Yahoo Img	79.9	76.4 ( -2.5 )
Classifier fusion	75.9	72.3 ( -3.6 )

\* Effect of adding images from known classes in the test set as distractors/negatives.

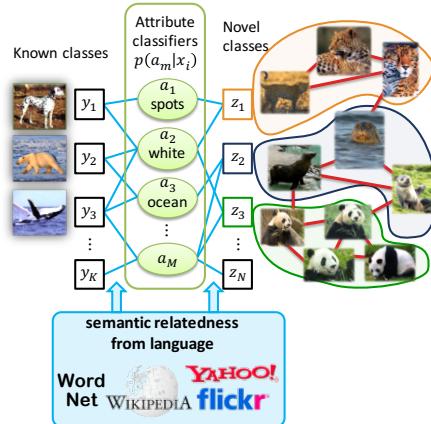
Approach/Language resource	in	Top-5 Error
<b>Hierarchy</b>		
leaf WordNet nodes	[69]	72.8
inner WordNet nodes	[69]	66.7
all WordNet nodes	[69]	65.2
+ metric learning	[53]	64.3*
<b>Part Attributes</b>		
Wikipedia	[69]	80.9
Yahoo Holonyms	[69]	77.3
Yahoo Image	[69]	81.4
Yahoo Snippets	[69]	76.2
all attributes	[69]	70.3
<b>Direct Similarity</b>		
Wikipedia	[69]	75.6
Yahoo Web	[69]	69.3
Yahoo Image	[69]	72.0
Yahoo Snippets	[69]	75.5
all measures	[69]	66.6
<b>Label embedding</b>		
DeViSe	[21]	68.2*

**Table 2** Large scale zero-shot recognition results. Flat error in % and hierarchical error in brackets.

\*Note that [53, 21] report on a different set of unseen classes than [69].

Transferring knowledge from known categories to novel classes is challenging as it is difficult to estimate visual properties of the novel classes. Approaches discussed in the previous section can not exploit instance similarity or few labeled instances, if available. The approach *Propagated Semantic Transfer* (PST) [74] combines four ideas to jointly handle the challenging scenario of recognizing novel categories. First, PST transfers information from known to novel categories by incorporating external knowledge, such as linguistic or expert-specified information, e.g., by a mid-level layer of semantic attributes as discussed in Section 2.1. Second, PST exploits the manifold structure of novel classes similar to unsupervised learning approaches [94, 80]. More specifically it adapts the graph-based Label Propagation algorithm [101, 100] – previously used only for semi-supervised learning [14] – to zero-shot and few-shot learning. In this transductive setting information is propagated between instances of the novel classes to get more reliable recognition as visualized with the red graph in Fig. 3. Third, PST improves the local neighborhood in such graph structures by replacing the raw feature-based representation with a semantic object- or attribute-based representation. And forth, PST generalizes from zero- to few-shot learning by integrating labeled training examples as certain nodes in its graph based propagation. Another positive aspect of PST is that attribute or category models do not have to be retrained if novel classes are added which can be an important aspect e.g. in a robotic scenario.

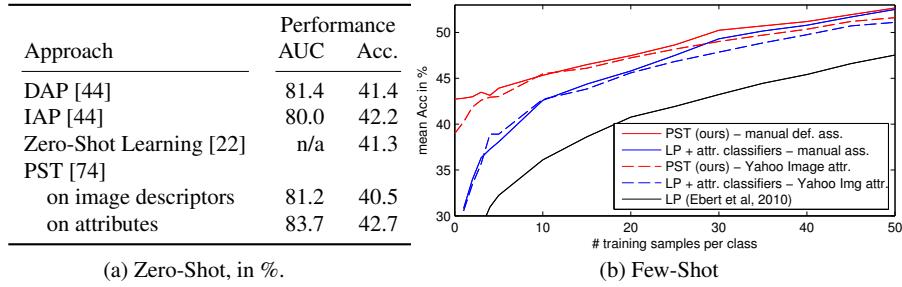
**Fig. 3** Recognition of novel categories. The approach *Propagated Semantic Transfer* [74] combines knowledge transferred via attributes from known classes (left) with few labeled examples in graph (red lines) which is build according to instance similarity.



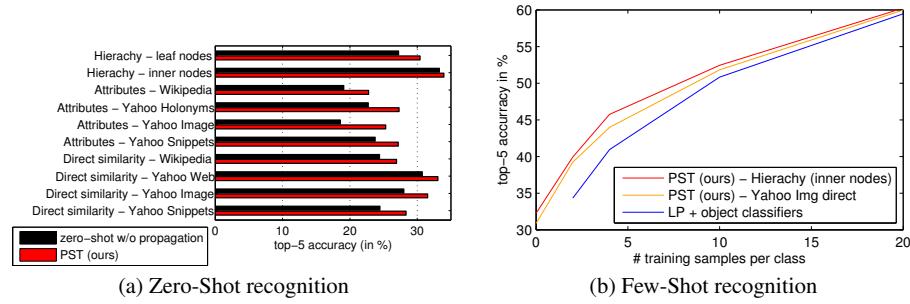
## 2.2 Propagated semantic transfer

Fig. 4 shows results on the AwA [43] dataset. We note that in contrast to the previous section the classifiers are trained on all training examples, not only 92 per class. Fig. 4(a) shows zero-shot results, where no training examples are available for the novel or in this case unseen classes. The table compares PST with propagating on a graph based on attribute-classifier similarity versus image descriptor similarity and shows a clear benefit of the former. This variant also outperform DAP and IAP [44] as well as Zero-Shot Learning [22]. Next we compare PST in the few-shot setting, i.e. we add labeled examples per class. In Fig. 4(b) we compare PST to two label propagation (LP) baselines [14]. We first note that PST (red curves) seamlessly moves from zero-shot to few-shot, while traditional LP (blue and black curves) needs at least one training example. We first examine the three solid lines. The black curve is the best LP variant from Ebert et al. [14] and uses similarity based image features. LP in combination with the similarity metric based on the attribute classifier scores (blue curves) allows to transfer knowledge residing in the classifier trained on the known classes and gives a significant improvement in performance. PST (red curve) additionally transfers labels from the known classes and improves further. The dashed lines in Fig. 4(b) provide results for automatically mined associations between attributes and classes from language resources. It is interesting to note that these automatically mined associations achieve performance very close to the manual defined associations (dashed vs. solid).

Fig. 5 shows results on the classification task with 200 unseen ImageNet categories. In Fig. 5(a) we compare PST to zero-shot without propagation presented as discussed in Section 2.1. For zero-shot recognition PST (red bars) improves performance over zero-shot without propagation (black bars) for all language resources and transfer variants. Similar to the AwA dataset, PST also improves over the LP-baseline for few-shot recognition (Fig. 5b). The missing LP-baseline on raw features is due to the fact that for the large number of images and high dimensional features



**Fig. 4** Zero-shot results on AwA dataset. Predictions with attributes and manual defined associations. Adapted from [74].



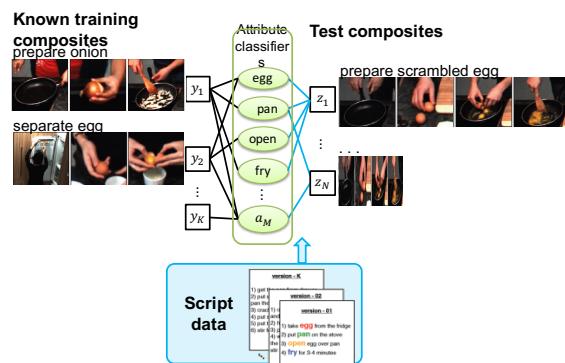
**Fig. 5** Results on 200 unseen classes of ImageNet. Adapted from [74].

the graph construction is very time and memory consuming if not infeasible. In contrast, the attribute representation is very compact and thus computational tractable even with a large number of images.

### 2.3 Composite activity recognition with attributes and script data

Understanding activities in visual and textual data is generally regarded as more challenging than understanding object categories due to the limited training data, challenges in defining the extend of an activity, and the similarities between activities [62]. However, long-term composite activities can be decomposed in shorter fine-grained activities [72]. Consider for example the composite cooking activities *prepare scrambled egg* which can be decomposed in attributes of fine-grained activities (e.g. *open*, *fry*), ingredients (e.g. *egg*), and tools (e.g. *pan*, *spatula*). These attributes can then be shared and transferred across composite activities as visualized in Fig. 6 using the same approaches as for objects and attributes discussed in the previous section. However, the representations, both on the visual and on the language side have to change. Fine-grained activities and associated attributes are

**Fig. 6** Recognizing composite activities using attributes and script data.



visually characterized by fine-grained body motions and low inter-class variability. In addition to holistic features [92], one consequently should exploit human pose-based [71] and hand-centric [77] features. As the previously discussed language resources do not provide good associations between composite activities and their attributes, Rohrbach et al. [72] collected textual description (*Script data*) of these activities with AMT. From this script data associations can be computed based on either the frequency statistics or, more discriminate, by term frequency times inverse document frequency (tf\*idf).

Table 3 shows results on the MPII Cooking 2 dataset [76]. Comparing the first column (holistic Dense Trajectory features [92]) with the second, shows the benefit of adding the more semantic hand-[77] and pose-[71] features. Comparing line (1) with line (2) or (3) shows the benefit of representing composite activities with attributes as this allows sharing across composite activities. Best performance is achieved with 57.4% mean AP in line (6) when combining compositional attributes with the Propagated Semantic Transfer (PST) approach (see Section 2.2) and Script data to determine associations between composites and attributes.

### 3 Image and video description using compositional attributes

In this section we discuss how we can generate natural language sentences describing visual content, rather than just giving labels to images and videos as discussed in the previous section. This intriguing task has recently received increased attention in computer vision and computational linguistics communities [89, 90, 91] and has a large number of potential applications including human robot interaction, image and video retrieval, and describing visual content for visually impaired people. In this section we focus on approaches which decouple the visual recognition and the sentence generation and introduce an intermediate semantic layer, which can be seen a layer of attributes (Section 3.1). Introducing such a semantic layer has several advantages. First, this allows to reason across sentences on a semantic level, which is, as we will see, beneficial for multi-sentence description of videos (Sec-

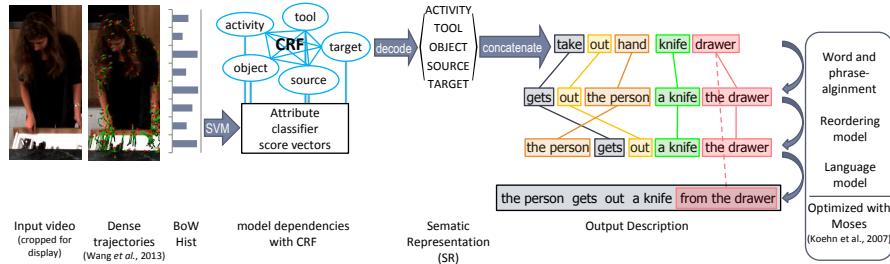
Attribute training on:	All Composites		Disjoint Composites	
	[92]	[92, 77, 71]	[92]	[92, 77, 71]
<b>With training data for composites</b>				
<i>Without attributes</i>				
(1) SVM	39.8	41.1	-	-
<i>Attributes on gt intervals</i>				
(2) SVM	43.6	52.3	32.3	34.9
<i>Attributes on automatic segmentation</i>				
(3) SVM	49.0	56.9	35.7	34.8
(4) NN	42.1	43.3	24.7	32.7
(5) NN+Script data	35.0	40.4	18.0	21.9
(6) PST+Script data	54.5	57.4	32.2	32.5
<b>No training data for composites</b>				
<i>Attributes on automatic segmentation</i>				
(7) Script data	36.7	29.9	19.6	21.9
(8) PST + Script data	36.6	43.8	21.1	19.3

**Table 3** Composite cooking activity classification on MPII Cooking 2 [76], mean AP in %. Top left quarter: fully supervised, right column: reduced attribute training data, bottom section: no composite cooking activity training data, right bottom quarter: true zero shot. Adapted from [76].

tion 3.2). Second, we can show that when learning reliable attributes, this leads to state-of-the-art sentences generation with high diversity in the challenging scenario of movie description (Section 3.3). Third, this leads to a compositional structure which allows describing novel concepts in images and videos (Section 3.4).

### 3.1 Translating image and video content to natural language descriptions

To address the problem of image and video description, Rohrbach et al. [75] propose a two-step translation approach which first predicts an intermediate semantic attribute layer and then learns how to translate from this semantic representation to natural sentences. Figure 7 gives an overview of this two-step approach for videos. First, a rich semantic representation of the visual content including e.g. object and activity attributes is predicted. To predict the semantic representation a CRF models the relationships between different attributes of the visual input. And second, the generation of natural language is formulated as a machine translation problem using the semantic representation as source language and the generated sentences as target language. For this a parallel corpus of videos, annotated semantic attributes, and textual descriptions allows to adapt statistical machine translation (SMT) [39] to translate between the two languages. Rohrbach et al. train and evaluate their ap-



**Fig. 7** Video description. Overview of the two-step translation approach [75] with an intermediate semantic layer of attributes (SR) for describing videos with natural language. From [68].

proach on the videos of the MPII Cooking dataset [71, 72] and the aligned descriptions from the TACoS corpus [62]. According to automatic evaluation and human judgments, the two-step translation approach significantly outperforms retrieval and n-gram-based baseline approaches, motivated by prior work. This similarly can be applied to image description task, however, in both cases it requires an annotated semantic attribute representation. In Sections 3.3 and 3.4 we discuss how we can extract such attribute annotations automatically from sentences. An alternative approach is presented by Fang et al. [16] who mine visual concepts for image description by integrating multiple instance learning [52]. Similar to the work presented in the following, Wu et al. [95] learn an intermediate attribute representation from the image descriptions. Captions are then generated solely from the intermediate attribute representation.

### 3.2 Coherent multi-sentence video description with variable level of detail

Most approaches for automatic video description, including the one presented above, focus on generating single sentence descriptions and are not able to vary the descriptions' level of detail. One advantage of the two-step approach with an explicit intermediate layer of semantic attributes is that it allows to reason on this semantic level. To generate coherent multi-sentence descriptions, Rohrbach et al. [64] extend the two-step translation approach to model across-sentence consistency at the semantic level by enforcing a consistent topic, which is the prepared dish in the cooking scenario. To produce shorter or one-sentence summaries, Rohrbach et al. select the most relevant sentences on the semantic level by using tf\*idf (term frequency times inverse document frequency). For an example output on the TACoS Multi-Level corpus [64] see Figure 8. In order to fully automatically do multi-sentence description, Rohrbach et al. propose a simple but effective method based on agglomerative clustering to perform automatic video segmentation. The most important component of good clustering is the similarity measure and it turns out that

**Detailed:**

A man took a cutting board and knife from the drawer. He took out an orange from the refrigerator. Then, he took a knife from the drawer. He juiced one half of the orange. Next, he opened the refrigerator. He cut the orange with the knife. The man threw away the skin. He got a glass from the cabinet. Then, he poured the juice into the glass. Finally, he placed the orange in the sink.

**Short:**

A man juiced the orange. Next, he cut the orange in half. Finally, he poured the juice into a glass.

**One sentence:**

A man juiced the orange.

**Fig. 8** Coherent multi-sentence descriptions at three levels of detail, using automatic temporal segmentation. See Section 3.2 for details. From [64].

the semantic attribute classifiers (see Fig. 7) are very well suited for that in contrast to Bag-of-Words dense trajectories [93]. This confirms the observation made in Section 2.2 that attribute classifiers seem to form a good space for distance computations.

To improve performance, Donahue et al. [12] show that the second step, the SMT-based sentence generation, can be replaced with a deep recurrent network to better model visual uncertainty, but still relying on the multi-sentence reasoning on the semantic level. On the TACoS Multi-Level corpus this achieves 28.8% BLEU@4, compared to 26.9% [64] with SMT and 24.9% with SMT without multi-sentence reasoning [75].

### 3.3 Describing movies with an intermediate layer of attributes

Two challenges arise, when extending the idea presented above to movie description [67], which looks at the problem how to describe movies for blind people. First, and maybe more importantly, there are no semantic attributes annotated as on the kitchen data, and second, the data is more visually diverse and challenging. For the first challenge, Rohrbach et al. [67] propose to extract attribute labels from the description to train visual classifiers to build a semantic intermediate layer by relying on a semantic parsing approach of the description. To additionally accommodate the second challenge of increased visual difficulty, Rohrbach et al. [66] show how to improve the robustness of these attributes or “Visual Labels” by three steps. First, by distinguishing three semantic groups of labels (verbs, objects and scenes) and using corresponding feature representations for each: activity recognition with dense trajectories [92], object detection with LSDA [31], and scene classification with

	SMT [67] S2VT [88] Visual labels [66] Reference	Someone is a man, someone is a man. Someone looks at him, someone turns to someone. Someone is standing in the crowd, a little man with a little smile. Someone, back in elf guise, is trying to calm the kids.
	SMT [67] S2VT [88] Visual labels [66] Reference	The car is a water of the water. On the door, opens the door opens. The fellowship are in the courtyard. They cross the quadrangle below and run along the cloister.
	SMT [67] S2VT [88] Visual labels [66] Reference	Someone is down the door, someone is a back of the door, and someone is a door. Someone shakes his head and looks at someone. Someone takes a drink and pours it into the water. Someone grabs a vodka bottle standing open on the counter and liberally pours some on the hand.

**Fig. 9** Qualitative results on the MPII Movie Description (MPII-MD) dataset [67]. The “Visual labels” approach [66] which uses an intermediate layer of robust attributes, identifies activities, objects, and places better than related work. From [66].

Places-CNN [99]. Second, training each semantic group separately, which removes noisy negatives. And third, selecting only the most reliable classifiers. While Rohrbach et al. use SMT for sentence generation in [67], they rely on a recurrent network (LSTM) in [66].

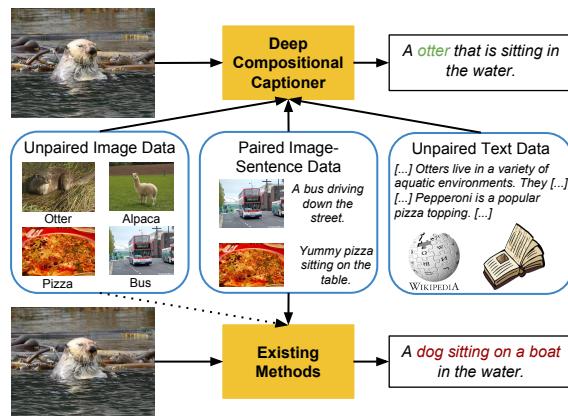
The Visual Labels approach outperforms prior work [88, 67, 96] on the MPII-MD [67] and M-VAD [86] dataset with respect to automatic and human evaluation. Qualitative results are shown in Fig. 9. An interesting characteristic of the compared methods is the size of the output vocabulary, which is 94 for [67], 86 for [88] (which uses an end-to-end LSTM approach without an intermediate semantic representation) and 605 for [66]. Although it is far lower than 6,422 for the human reference sentences, it clearly shows a higher diversity of the output for [66].

### 3.4 Describing novel object categories

In this section we discuss how to describe novel object categories which combines challenges discussed for recognizing novel categories (Section 2) and generating descriptions (Section 3.1). State-of-the-art deep image and video captioning approaches (e.g. [91, 50, 12, 16, 89]) are limited to describe objects which appear in caption corpora such as MS COCO [8] which consist of pairs of images and sentences. In contrast, labeled image datasets without sentence descriptions (e.g. ImageNet [10]) or text only corpora (e.g. Wikipedia) cover many more object categories.

Hendricks et al. [30] propose the Deep Compositional Captioner (DCC) to exploit these vision-only and language-only unpaired data sources to describe novel categories as visualized in Fig. 10. Similar to the attribute layer discussed in Section 3.1, Hendricks et al. extract words as labels from the descriptions to learn a “Lexical Layer”. The Lexical Layer is expanded by objects from ImageNet [10].

**Fig. 10** Describing novel object categories which are not contained in caption corpora (like otter). The Deep Compositional Captioner (DCC) [30] uses an intermediate semantic attribute or “lexical” layer to connect classifiers learned on unpaired image datasets (ImageNet) with text corpora (e.g. Wikipedia). This allows it to compose descriptions about novel objects without any paired image-sentences training data. Adapted from [29].

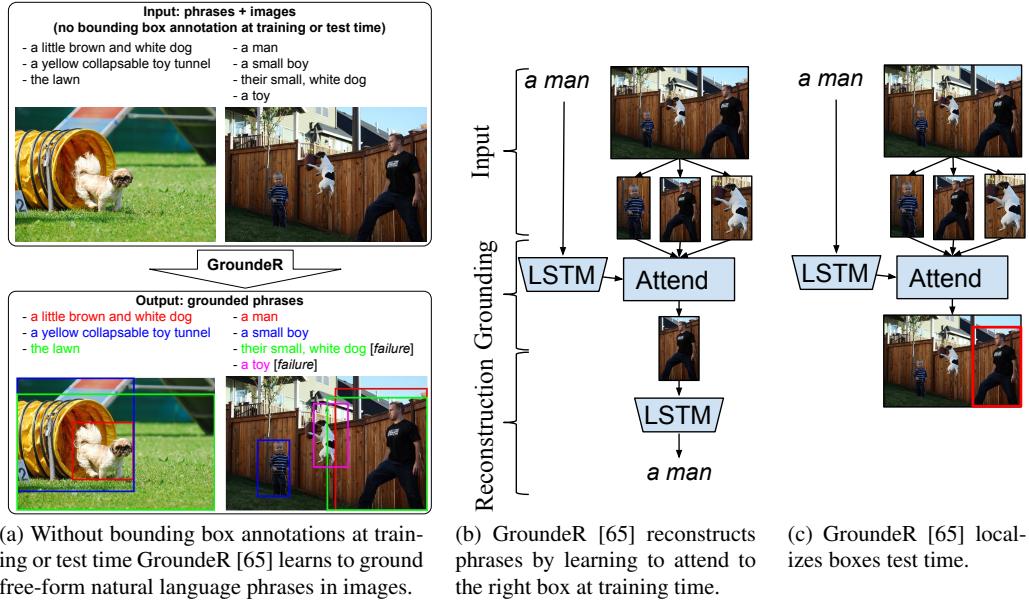


bird → toad	chocolate → tiramisu
 No Transfer: A close up of a person holding a cell phone. DCC: A <b>toad</b> is sitting on a table.	 No Transfer: A piece of cake with a fork and a fork. DCC: A <b>tiramisu</b> is sitting on a plate.
sheep → alpaca	fruit → persimmon
 No Transfer: A couple of cows sitting next to each other. DCC: A couple of <b>alpaca</b> standing next to each other in a field.	 No Transfer: A close up of a plate of food with a bowl of fruit. DCC: A close up of a plate of food with a bowl of <b>persimmon</b> .
vase → candelabra	tree → fig
 No Transfer: A white and black and white photo of a white and blue fire hydrant. DCC: A <b>candelabra</b> is sitting on a table.	 No Transfer: A close up of a plate of food on a table. DCC: A close up of a <b>fig</b> .
bird → dragonfly	giraffe → impala
 No Transfer: A small bird sitting on a green plant. DCC: A <b>dragonfly</b> with a green plant on a green plant.	 No Transfer: A brown and white cow standing in a field. DCC: A <b>impala</b> is standing in the dirt.
bird → cockatoo	bear → coyote
 No Transfer: A white bird standing on a white surface.. DCC: A white <b>cockatoo</b> sitting on a grass covered field.	 No Transfer: A couple of giraffe standing next to each other. DCC: A <b>coyote</b> is standing in the middle of a forest.

**Fig. 11** Qualitative results for describing novel ImageNet object categories. DCC [30] compared to an ablation without transfer. X → Y: known word X is transferred to novel word Y. From [29].

To be able to not only recognize but also generate the description about the novel objects, DCC transfers the word prediction model from semantically closest known word in the Lexical Layer, where similarity is computed with Word2Vec [55]. Interesting to note is, that image captioning approaches such as [91, 12] do use ImageNet data to (pre-) train the models (indicated with a dashed arrow in Fig. 10), but they do not make use of the semantic information but only the learned representation.

Fig. 11 shows several categories where there exist no captions for training. With respect to quantitative measures, compared to a baseline without transfer, DCC improves METEOR from 18.2% to 19.1% and F1 score, which measures the appear-

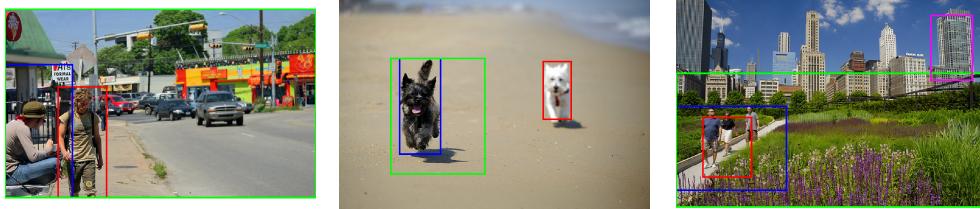


**Fig. 12** Unsupervised grounding by learning to associate visual and textual semantic units. From [65].

ance of the novel object, from 0 to 34.3%. Hendricks et al. also show similar results for video description.

## 4 Grounding text in images

In this section we discuss the problem of grounding natural language in images. Grounding in this case means that given an image and a natural language sentence or phrase, we aim to localize the subset of the image which corresponds to the input phrase. For example, for the sentence “*A little brown and white dog emerges from a yellow collapsible toy tunnel onto the lawn.*” and the corresponding image in Fig. 12(a), we want to segment the sentence into phrases and locate the corresponding bounding boxes (or segments) in the image. While grounding has been addressed e.g. in [40, 34, 5, 81], it is restricted to few categories. An exception are Karpathy et al. [36, 37] who aim to discover a latent alignment between phrases in text and bounding box proposals in the image. Karpathy et al. [37] ground dependency-tree relations to image regions using multiple instance learning (MIL) and a ranking objective. Karpathy and Fei-Fei [36] simplify the MIL objective to just the maximal scoring box and replace the dependency tree with a learned recurrent network. These approaches have unfortunately not been evaluated with respect to the grounding performance due to a lack of annotated datasets. Only recently two datasets were



A man walking by a sitting man on the street.

A white dog is following a black dog along the beach.

Three people on a walk down a cement path beside a field of wildflowers with skyscrapers in the background.

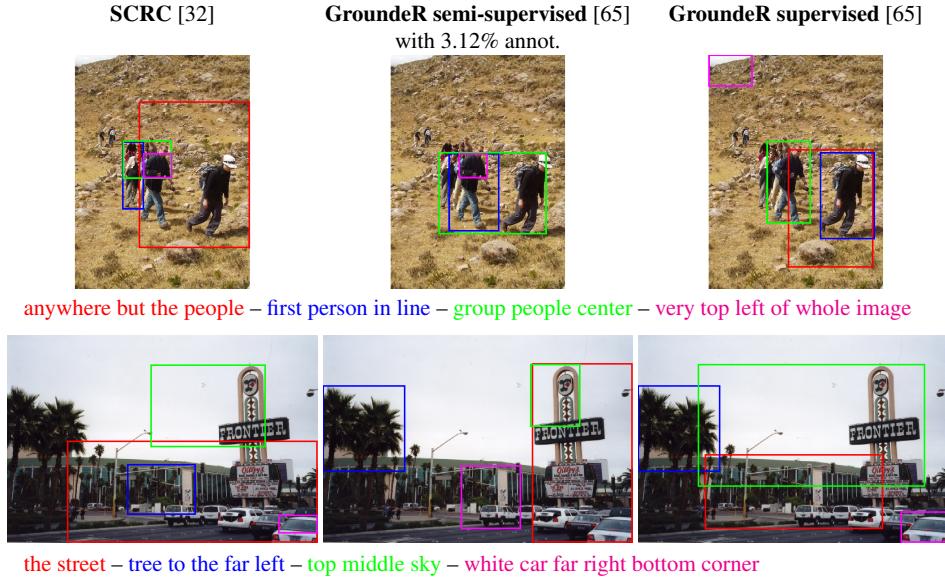
**Fig. 13** Qualitative results for GroundeR unsupervised [65] on Flickr 30k Entities [60]. Compact textual semantic units (phrases, e.g. “*a sitting man*”) are associated with visual semantic units (bounding boxes). Best viewed in color.

released: Flickr30k Entities [60] augments Flickr30k [97] with bounding boxes for all noun phrases present in textual descriptions and ReferItGame [38] has localized referential expressions in images. Even more recent, at the time of writing, efforts are being made to also collect grounded referential expressions for the MS COCO [47] dataset, namely the authors of ReferItGame are in progress of extending their annotations as well as longer referential expressions have been collected by Mao et al. [51]. Similar efforts are also made in the Visual Genome project [41] which provides densely annotated images with phrases.

In the following we focus on how to approach this problem and the first question is, where is the best point of interaction between linguistic elements and visual elements? Following the approaches in [36, 37, 60] a good way to this is to decompose both, sentence and image into concise semantic units or attributes which we can match to each other. For the data as shown in Figures 12(a) and 13, sentences can be split into phrases of typically a few words and images are composed into a larger number of bounding box proposals [87]. An alternative is to integrate phrase grounding in a fully-convolutional network, for bounding box prediction [35] or segmentation prediction [33]. In the following, we discuss approaches which focus on how to find the association between visual and linguistic components, rather than the actual segmentation into components. We first look at an unsupervised setting with respect to the grounding task, i.e. we assume that no bounding box annotations are available for training (Section 4.1), and then we show how to integrate supervision (Section 4.2). Section 4.3 discusses the results.

#### 4.1 Unsupervised grounding

Although many data sources contain images which are described with sentences or phrases, they typically do not provide the spatial localization of the phrases. This



**Fig. 14** Qualitative grounding results on ReferItGame Dataset [38]. Different colors show different referential expressions for the same image. Best viewed in color.

is true for both curated datasets such as MSCOCO [47] or large user generated content as e.g. in the YFCC 100M dataset [84]. Consequently, being able to learn from this data without grounding supervision would allow large amount and variety of training data. This setting is visualized in Fig. 12(a).

For this setting Rohrbach et al. [65] propose the approach GroundeR, which is able to learn the grounding by aiming to reconstruct a given phrase using an attention mechanism as shown in Fig. 12(b). In more detail, given images paired with natural language phrases (or sentence descriptions), but without any bounding box information, we want to localize these phrases with a bounding box in the image (Fig. 12c). To do this, GroundeR learns to attend to a bounding box proposal and, based on the selected bounding box, reconstructs the phrase (Fig. 12b). Attention means that the model predicts a weighting over the bounding boxes and then takes the weighted average of the features from all boxes. A softmax over the weights encourages that only one or a few boxes have high weights. As the second part of the model (Fig. 12b, bottom) is able to predict the correct phrase only if the first part of the model attended correctly (Fig. 12b, top), this can be learned without additional bounding box supervision. At test time we evaluate the grounding performance, i.e. whether the model assigned the highest weight to / attended to the correct bounding box. The model is able to learn these associations as the parameters of the model are learned across all phrases and images. Thus, for a proper reconstruction, the visual semantic units and linguistic phrases have to match, i.e. the models learns what certain visual phrases mean in the image.

Approach	Accuracy	Approach	Accuracy
<b>Unsupervised training</b>			
GroundeR (VGG-CLS) [65]	24.66	LRCN [12] (reported in [32])	8.59
GroundeR (VGG-DET) [65]	32.42	CAFFE-7K [27] (reported in [32])	10.38
<b>Semi-supervised training</b>			
GroundeR (VGG-CLS) [65]	33.02	GroundeR (VGG+SPAT) [65]	10.44
3.12% annotation	37.10	3.12% annotation	15.03
6.25% annotation	38.67	6.25% annotation	19.53
12.5% annotation		12.5% annotation	21.65
<b>Supervised training</b>			
CCA embedding [60]	25.30	Supervised training	
SCRC (VGG+SPAT) [32]	27.80	SCRC (VGG+SPAT) [32]	17.93
GroundeR (VGG-CLS) [65]	41.56	GroundeR (VGG+SPAT) [65]	26.93
GroundeR (VGG-DET) [65]	47.70		

(a) Flickr 30k Entities dataset [60]

(b) ReferItGame dataset [38]

**Table 4** Phrase grounding, accuracy in %. VGG-CLS: Pre-training the VGG network [79] for the visual representation on ImageNet classification data only. VGG-DET: VGG further fine-tuned for the object detection task on the PASCAL dataset [15] using Fast R-CNN [25]. VGG+SPAT: VGG-CLS + spatial bounding box features (box location and size).

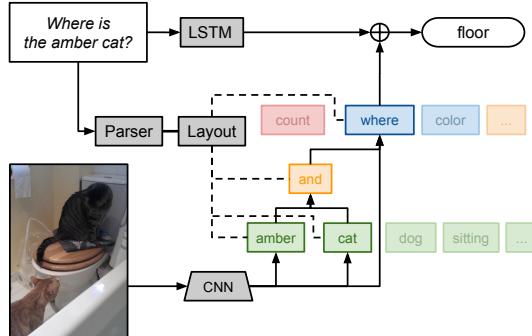
#### 4.2 Semi-supervised and fully supervised grounding

If grounding supervision (phrase bounding box associations) is available, GroundeR [65] can integrate it by adding a loss over the attention mechanism (Fig. 12b, “Attend”). Interestingly, this allows to provide supervision only for a subset of the phrases (semi-supervised) or all phrases (fully supervised).

For supervised grounding, Plummer et al. [60] proposed to learn a CCA embedding [26] between phrases and the visual representation. The Spatial Context Recurrent ConvNet (SCRC) [32] and the approach of Mao et al. [51] use a caption generation framework to score phrases on a set of bounding box proposals. This allows to rank bounding box proposals for a given phrase or referential expression. Hu et al. [32] show the benefit of transferring models trained on full-image description datasets as well as spatial (bounding box location and size) and full-image context features. Mao et al. [51] show how to discriminatively train the caption generation framework to better distinguish different referential expression.

#### 4.3 Grounding results

In the following we discuss results on the Flickr 30k Entities dataset [60] and the ReferItGame dataset [38], which both provide ground truth alignment between noun phrases (within sentences) and bounding boxes. For the unsupervised models, the grounding annotations are only used at test time for evaluation, not for training. All



**Fig. 15** To approach visual question answering, Andreas et al. [2] propose to dynamically create a deep network which is composed of different “modules” (colored boxes). These “modules” represent semantic units, i.e. attributes, which link linguistic units in the question with computational units to do the corresponding visual recognition. Adapted from [1].

approaches use the activations of the second last layer of the VGG network [79] to encode the image inside the bounding boxes.

Table 4(a) compares the approaches quantitatively. The unsupervised variant of GroundeR reaches nearly the supervised performance of CCA [60] or SCRC[32] on Flickr 30k Entities, successful examples are shown in Fig. 13. For the referential expressions of the ReferItGame dataset the unsupervised variant of GroundeR reaches performance on par with prior work (Table 4b) and quickly gains performance when adding few labeled training annotation (semi-supervised training). In the fully supervised setting GroundeR improves significantly over state-of-the-art on both datasets, which is also reflected in the qualitative results shown in Fig. 14.

## 5 Visual question answering

Visual question answering is the problem of answering natural language questions about images, e.g. for the question “*Where is the amber cat?*” about the image shown in Fig. 15 we want to predict the corresponding answer *on the floor*, or just *floor*. This is a very interesting problem with respect to several aspects. On the one hand it has many applications, such visual search, human-robot interaction, and assisting blind people. On the other hand, it is also an interesting research direction as it requires to relate textual and visual semantics. More specifically it requires to ground the question in the image, e.g. by localizing the relevant part in the image (*amber cat* in Fig. 15), and then recognizing and predicting an answer based on the question and the image content. Consequently, this problem requires more complex semantic interaction between language and visual recognition than in previous sections, specifically, the problem requires ideas from grounding (Section 4) and recognition (Section 2) or description (Section 3).

	test-dev				test All
	Y/N	Num	Other	All	
LSTM	78.7	36.6	28.1	49.8	–
ATT+LSTM	80.6	36.4	42.0	57.2	–
NMN	70.7	36.8	39.2	54.8	–
NMN+LSTM	81.2	35.2	43.3	58.0	–
NMN+LSTM+FT	81.2	38.0	44.0	58.6	<b>58.7</b>

LSTM: a question-only baseline  
 ATT: single find+describe for all questions  
 NMN+LSTM: full model shown in Fig. 15  
 +FT: image features fine-tuned on captions [12]  
 NMN: ablation w/o LSTM

(a) Results from evaluation server of [4] in %.



*how many different lights in various different shapes and sizes?*   *what color is the vase?*   *is the bus full of passengers?*

(b) Answers from [1] (ground truth answers in parentheses).

**Fig. 16** Results on the VQA dataset [4]. Adapted from [1].

Most recent approaches to visual question answering learn a joint hidden embedding of the question and the image to predict the answer [49, 63, 24, 4] where all computation is shared and identical for all questions. An exception to this is proposed by Wu et al. [95], who learn an intermediate attribute representation from the image descriptions, similar to the work discussed in Sections 3.3 and 3.4. Interestingly, this intermediate layer of attributes allows to query an external knowledge base to provide additional (textual) information not visible in the image. The embedded textual knowledge base information is combined with the attribute representation and the hidden representation of a caption-generation recurrent network (LSTM) and forms the input to an LSTM-based question-answer encoder-decoder [49].

Andreas et al. [2] go one step further with respect to compositionality and propose to predict a compositional neural network structure from the questions. As visualized in Fig. 15 the question “*Where is the amber cat?*” is decomposed into network “modules” *amber*, *cat*, *and*, and *where*. These modules are semantic units, i.e. attributes, which connect most relevant semantic components of the questions (i.e. word or short phrases) with corresponding computation to recognize it in the image. These Neural Module Networks (NMN) have different types of modules for different types of attributes. Different types have different colors in Fig. 15. The *find[cat]* and *find[amber]* (green) modules take in CNN activations (VGG [79], last convolutional layer) and produce a spatial attention heatmap, while *combine[and]* (orange) combines two heatmaps to a single one, and *describe[where]* (blue) takes in a heatmap and CNN features to predict an answer. Note that the distinction between different types, e.g. *find* versus *describe*, which have different kind of computation and different instances, e.g. *find[cat]* versus *find[amber]*, which learn different parameters. All parameters are initialized randomly and only trained from question answer pairs. Interestingly, in this work attributes are not only distinguished with respect of their type, but also are composed with other attributes in a deep network, whose parameters’ are learned

end-to-end from examples, here question-answer pairs. In a follow up work, Andreas et al. [3] learn not only the modules, but also what the best network structure is from a set of parser proposals, using reinforcement learning.

In addition to NMN, Andreas et al. [2, 3] also incorporate a recurrent network (LSTM) to model common sense knowledge and dataset bias which has been shown to be important for visual question answering [49]. Quantitative results in Table 16(a) indicate that NMNs are indeed a powerful tool to question answering, a few qualitative results can be seen Fig. 16(b).

## 6 Conclusions

In this chapter we presented several tasks and approaches where attributes enable a connection of visual recognition with natural language on a semantic level. For recognizing novel object categories or activities, attribute can build an intermediate representation which allows incorporating knowledge mined from language resources or script data (Section 2). For this scenario we saw that semantic attribute classifiers additionally build a good metric distance space useful for constructing instance graphs and learning composite activity recognition models. In Section 3 we explained how an intermediate level of attributes can be used to describe videos with multiple sentences and at a variable level and allow describing novel object categories. In Section 4 we presented approaches for unsupervised and supervised grounding of phrases in images. Different phrases are semantically overlapping and the examined approaches try to relate these semantic units by jointly learning representations for the visual and language modalities. Section 5 discusses an approach to visual question answering which composes the most important attributes of a question in a compositional computation graph, whose parameters are learned end-to-end only by back-propagating from the answers.

While the discussed approaches take a step towards the challenges discussed in Section 1.1, there are many future steps ahead. While the approaches in Section 2 use many advanced semantic relatedness measures minded from diverse language resources they are not jointly trained on textual and visual modalities. Regneri et al. [62] and Silberer et al. [78] take a step in this direction by looking at joint semantic representation from the textual and visual modalities. Section 3 presents compositional models for describing videos, but it is only a first step towards automatically describing a movie to a blind person as humans can do it [67], which will require an even higher degree of semantic understanding, and transfer within and between modalities. Section 4 describes interesting ideas to grounding in images and it will be interesting to see how this scales to the size of the Internet. Visual question answering (Section 5) is an interesting emerging direction with many challenges as it requires to solve all of the above, at least to some extend.

**Acknowledgements** I would like to thank all my co-authors, especially those whose publications are presented in this chapter. Namely, Sikandar Amin, Jacob Andreas, Mykhaylo Andriluka, Trevor

Darrell, Sandra Ebert, Jiashi Feng, Annemarie Friedrich, Iryna Gurevych, Lisa Anne Hendricks, Ronghang Hu, Dan Klein, Raymond Mooney, Manfred Pinkal, Wei Qiu, Michaela Regneri, Anna Rohrbach, Kate Saenko, Michael Stark, Bernt Schiele, György Szarvas, Stefan Thater, Ivan Titov, Subhashini Venugopalan, and Huazhe Xu.

Marcus Rohrbach was supported by a fellowship within the FITWeltweit-Program of the German Academic Exchange Service (DAAD).

## References

- [1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Deep compositional question answering with neural module networks. *arXiv preprint arXiv:1511.02799*, 2015.
- [2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2016.
- [4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015.
- [5] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research (JMLR)*, 3:1107–1135, 2003.
- [6] E. Bart and S. Ullman. Single-example learning of novel classes using representation by similarity. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2005.
- [7] H.-H. Chen, M.-S. Lin, and Y.-C. Wei. Novel association measures using web search with double checking. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2006.
- [8] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [10] J. Deng, A. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *European Conference on Computer Vision (ECCV)*, 2010.
- [11] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
- [12] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for

- visual recognition and description. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [13] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering Localized Attributes for Fine-grained Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [14] S. Ebert, D. Larlus, and B. Schiele. Extracting Structures in Image Collections for Object Recognition. In *European Conference on Computer Vision (ECCV)*, 2010.
- [15] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.
- [16] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [17] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [18] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [19] R. Farrell, O. Oza, V. Morariu, T. Darrell, and L. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *International Conference on Computer Vision (ICCV)*, 2011.
- [20] C. Fellbaum. *WordNet: An Electronical Lexical Database*. The MIT Press, 1998.
- [21] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Conference on Neural Information Processing Systems (NIPS)*, 2013.
- [22] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(2):303–316, 2014.
- [23] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.
- [24] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *Conference on Neural Information Processing Systems (NIPS)*, 2015.
- [25] R. Girshick. Fast R-CNN. In *International Conference on Computer Vision (ICCV)*, 2015.
- [26] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision (ECCV)*, 2014.

- [27] S. Guadarrama, E. Rodner, K. Saenko, N. Zhang, R. Farrell, J. Donahue, and T. Darrell. Open-vocabulary object retrieval. In *Robotics: science and systems*, 2014.
- [28] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision (ICCV)*, 2015.
- [29] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. *arXiv preprint arXiv:1511.05284v1*, 2015.
- [30] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [31] J. Hoffman, S. Guadarrama, E. Tzeng, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. LSDA: Large scale detection through adaptation. In *Conference on Neural Information Processing Systems (NIPS)*, 2014.
- [32] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [33] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. *arXiv preprint arXiv:1603.06180*, 2016.
- [34] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [35] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [36] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [37] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Conference on Neural Information Processing Systems (NIPS)*, 2014.
- [38] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- [39] P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.
- [40] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [41] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual

- genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016.
- [42] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Conference on Neural Information Processing Systems (NIPS)*, 2012.
- [43] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [44] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(3):453–465, 2014.
- [45] C. Liang, C. Xu, J. Cheng, W. Min, and H. Lu. Script-to-movie: A computational framework for story movie composition. *Multimedia, IEEE Transactions on*, 15(2):401–414, 2013.
- [46] D. Lin. An information-theoretic definition of similarity. In *International Conference on Machine Learning (ICML)*, 1998.
- [47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.
- [48] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Conference on Neural Information Processing Systems (NIPS)*, 2014.
- [49] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *International Conference on Computer Vision (ICCV)*, 2015.
- [50] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *International Conference on Learning Representations (ICLR)*, 2015.
- [51] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [52] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *Conference on Neural Information Processing Systems (NIPS)*, 1998.
- [53] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Metric Learning for Large Scale Image Classification: Generalizing to New Classes at Near-Zero Cost. In *European Conference on Computer Vision (ECCV)*, 2012.
- [54] R. Mihalcea and D. I. Moldovan. A method for word sense disambiguation of unrestricted text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1999.
- [55] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Conference on Neural Information Processing Systems (NIPS)*, 2013.
- [56] Y. Moses, S. Ullman, and S. Edelman. Generalization to novel images in upright and inverted faces. *Perception*, 25:443–461, 1996.

- [57] D. Mrowca, M. Rohrbach, J. Hoffman, R. Hu, K. Saenko, and T. Darrell. Spatial semantic regularisation for large scale object detection. In *International Conference on Computer Vision (ICCV)*, 2015.
- [58] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. In *Conference on Neural Information Processing Systems (NIPS)*, 2009.
- [59] D. Parikh and K. Grauman. Relative attributes. In *International Conference on Computer Vision (ICCV)*, 2011.
- [60] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision (ICCV)*, 2015.
- [61] R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng. Self-taught learning: Transfer learning from unlabeled data. In *International Conference on Machine Learning (ICML)*, 2007.
- [62] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding Action Descriptions in Videos. *Transactions of the Association for Computational Linguistics (TACL)*, 2013.
- [63] M. Ren, R. Kiros, and R. Zemel. Image question answering: A visual semantic embedding model and a new dataset. In *Conference on Neural Information Processing Systems (NIPS)*, 2015.
- [64] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. In *Proceedings of the German Conference on Pattern Recognition (GCPR)*, 2014.
- [65] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. *arXiv preprint arXiv:1511.03745*, 2015.
- [66] A. Rohrbach, M. Rohrbach, and B. Schiele. The long-short story of movie description. *Proceedings of the German Conference on Pattern Recognition (GCPR)*, 2015.
- [67] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [68] M. Rohrbach. *Combining visual recognition and computational linguistics: linguistic knowledge for visual recognition and natural language descriptions of visual content*. PhD thesis, Saarland University, 2014.
- [69] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps Where - and Why? Semantic Relatedness for Knowledge Transfer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [70] M. Rohrbach, M. Stark, and B. Schiele. Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

- [71] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [72] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. In *European Conference on Computer Vision (ECCV)*, 2012.
- [73] M. Rohrbach, M. Stark, G. Szarvas, and B. Schiele. Combining language sources and robust semantic relatedness for attribute-based knowledge transfer. In *Proceedings of the European Conference on Computer Vision Workshops (ECCV Workshops)*, volume 6553 of *LNCS*, 2012.
- [74] M. Rohrbach, S. Ebert, and B. Schiele. Transfer Learning in a Transductive Setting. In *Conference on Neural Information Processing Systems (NIPS)*, 2013.
- [75] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *International Conference on Computer Vision (ICCV)*, 2013.
- [76] M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision (IJCV)*, 2015.
- [77] A. Senina, M. Rohrbach, W. Qiu, A. Friedrich, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. *arXiv preprint arXiv:1403.6173*, 2014.
- [78] C. Silberer, V. Ferrari, and M. Lapata. Models of semantic representation with visual attributes. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013.
- [79] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.
- [80] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering Object Categories in Image Collections. In *International Conference on Computer Vision (ICCV)*, 2005.
- [81] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [82] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.*, 5:1–34, 1948.
- [83] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [84] B. Thomee, B. Elizalde, D. A. Shamma, K. Ni, G. Friedland, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

- [85] S. Thrun. Is learning the n-th thing any easier than learning the first. In *Conference on Neural Information Processing Systems (NIPS)*, 1996.
- [86] A. Torabi, C. Pal, H. Larochelle, and A. Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070v1*, 2015.
- [87] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013.
- [88] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence – video to text. *arXiv preprint arXiv:1505.00487v2*, 2015.
- [89] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence – video to text. In *International Conference on Computer Vision (ICCV)*, 2015.
- [90] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2015.
- [91] 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.
- [92] H. Wang and C. Schmid. Action recognition with improved trajectories. In *International Conference on Computer Vision (ICCV)*, 2013.
- [93] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- [94] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000.
- [95] Q. Wu, C. Shen, A. v. d. Hengel, P. Wang, and A. Dick. Image captioning and visual question answering based on attributes and their related external knowledge. *arXiv preprint arXiv:1603.02814*, 2016.
- [96] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. *arXiv preprint arXiv:1502.08029v4*, 2015.
- [97] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2:67–78, 2014.
- [98] T. Zesch and I. Gurevych. Wisdom of crowds versus wisdom of linguists - measuring the semantic relatedness of words. *Natural Language Engineering*, 16(1):25–59, 2010.
- [99] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. In *Conference on Neural Information Processing Systems (NIPS)*, 2014.

- [100] D. Zhou, O. Bousquet, T. N. Lal, Jason Weston, and B. Schölkopf. Learning with Local and Global Consistency. In *Conference on Neural Information Processing Systems (NIPS)*, 2004.
- [101] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *International Conference on Machine Learning (ICML)*, 2003.
- [102] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *International Conference on Computer Vision (ICCV)*, 2013.