

Disambiguating Visual Verbs

Spandana Gella, Frank Keller, and Mirella Lapata

Abstract—In this article, we introduce a new task, visual sense disambiguation for verbs: given an image and a verb, assign the correct sense of the verb, i.e., the one that describes the action depicted in the image. Just as textual word sense disambiguation is useful for a wide range of NLP tasks, visual sense disambiguation can be useful for multimodal tasks such as image retrieval, image description, and text illustration. We introduce a new dataset, which we call VerSe (short for **Verb Sense**) that augments existing multimodal datasets (COCO and TUHOI) with verb and sense labels. We explore supervised and unsupervised models for the sense disambiguation task using textual, visual, and multimodal embeddings. We also consider a scenario in which we must detect the verb depicted in an image prior to predicting its sense (i.e., there is no verbal information associated with the image). We find that textual embeddings perform well when gold-standard annotations (object labels and image descriptions) are available, while multimodal embeddings perform well on unannotated images. VerSe is publicly available at <https://github.com/spandanagella/verse>.

Index Terms—Computer vision, Distributed representations, Natural Language Processing

1 INTRODUCTION

ACTION recognition, the task of identifying the actions depicted in videos or still images, is a widely studied problem in computer vision. Several applications stand to benefit from the ability to recognize actions, such as image description generation, image/video retrieval, surveillance, and a variety of systems involving human-computer interaction. The bulk of existing work has focused on video data, where motion and temporal information provide cues for recognizing actions. The absence of such cues renders the task more challenging in still images. Nevertheless, attempts to recognize actions in images can be broadly grouped into (a) action classification (AC), which aims to label an image with a verb phrase, typically a combination of a verb and its object (e.g., *play baseball*, *ride horse*), while assuming that such labels are mutually exclusive [1], [2], [3], [4], [5]; (b) human object interaction (HOI) recognition, which aims to identify all possible interactions between a human and an object in an image; co-occurring actions (e.g., *hold bicycle* and *ride bicycle*) can in principle be modeled since images receive multiple labels [6], [7], [8]; and (c) visual semantic role labeling (VSRL), which identifies the roles actors and objects play in the activity or situation depicted in the image [9], [10]. Figure 1 illustrates each of these tasks and how they relate to each other.

However, none of these action recognition tasks considers the ambiguity that arises when verbs are used as labels. For example, the verb *play* has multiple meanings in different contexts: participate in sport, play musical instrument, or engage in playful activity (see Figure 2). Moreover, action labels consisting of verb-object pairs may miss important generalizations, e.g., the fact that *ride horse* and *ride elephant* both evoke the same verb semantics, namely *ride animal*. Existing action labels also miss generalizations across verbs, e.g., the fact that *fix bike* and *repair bike* are semantically equivalent, in spite of the use of different verbs. These observations strongly suggest that actions should be analyzed at the level of *verb senses*, similarly to how they are studied in natural language processing.

In this article, we therefore propose the new task of visual verb sense disambiguation (VSD), which aims to label an image with a verb sense taken from a lexical database (see Figure 1). We explore two VSD scenarios: (1) given an image and a verb,

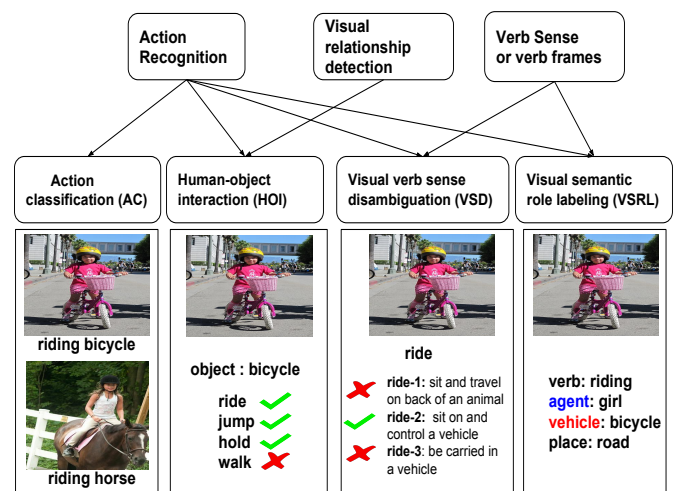


Fig. 1: Categorization of action recognition tasks in images.



Fig. 2: Visual sense ambiguity: three of the senses of the verb *play*: play sport, play instrument, children play.

assign the correct sense of the verb, i.e., the one that describes the action depicted in the image; and (2) given an image, predict a verb and its corresponding sense to correctly describe the action in the image. We present VerSe, a new dataset that augments existing multimodal datasets (COCO and TUHOI) with sense labels. VerSe contains 3,510 images, each annotated with one of 90 verbs, as well as the verb sense realized in the image according to the OntoNotes sense inventory [11].

For our first scenario, we explore both unsupervised and supervised disambiguation methods. We focus in particular on how to best represent word senses for visual disambiguation, and explore the use of textual, visual, and multimodal embeddings. Textual embeddings for a given image can be constructed over

• The authors are with the Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom.

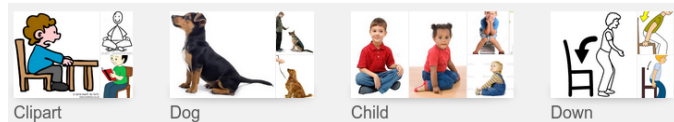


Fig. 3: Google Image Search trying to disambiguate *sit*. All clusters pertain to the sit down sense, other senses (baby sit, convene) are not included.

object labels or image descriptions, which are available as gold-standard in the COCO and TUHOI datasets, or can be computed automatically using object detectors and image description models. Our results show that textual embeddings perform best when gold-standard textual annotations are available, while multimodal embeddings perform best when automatically generated object labels are used. Interestingly, we find that automatically generated image descriptions result in inferior performance. For our second scenario, we predict the verbs depicted in an image using multilabel classification algorithms, which can operate on bounding boxes from an image or on the full image. Our results show that multiple instance learning (MIL), which takes inputs of positive and negative bounding boxes for every label, performs better than a multilabel CNN architecture.

In the remainder of this article, we first present an overview of related work. We then introduce the VerSe dataset and describe our annotation procedure. Next, we provide the details of our disambiguation and verb prediction models. Experimental results and discussion conclude the article.

2 RELATED WORK

Sense Disambiguation Visual sense disambiguation is related to word sense disambiguation (WSD), a canonical task in natural language processing. The aim in WSD is to identify the intended meaning (sense) of a word in its *textual context*. Reliable WSD has been argued to improve a range of NLP applications, including information retrieval, information extraction, machine translation, content analysis, and lexicography (see [12] for an overview).

There is an extensive literature on WSD for nouns, verbs, adjectives, and adverbs. Most of these approaches rely on lexical databases and sense inventories such as WordNet [13] or OntoNotes [11]. Unsupervised WSD approaches often rely on distributional representations, computed over the target word and its context [14], [15], [16]. Most supervised approaches use sense annotated corpora to extract linguistic features of the target word (context words, part-of-speech tags, collocation features), which are then fed into a classifier to disambiguate test data [17]. Recently, features based on sense-specific semantic vectors learned using large corpora and a sense inventory have been shown to achieve state-of-the-art results for supervised WSD [18], [19].

In a multimodal setting (e.g., newspaper articles with photographs), *visual context* is also available and can be used for sense disambiguation in multimodal tasks such as image retrieval. As an example, consider the output of Google Image Search for the query *sit*, shown in Figure 3: the search engine recognizes that the verb has multiple senses and tries to cluster relevant images. However, the result does not capture the polysemy of the verb well, and would clearly benefit from visual sense disambiguation.

In the existing literature, VSD has been attempted only for nouns (e.g., *apple* can mean fruit or computer¹). Sense discrimination for web images was introduced in Loeffel et al. [20], who

used spectral clustering over multimodal features from images and web text. Saenko et al. [21] employ sense definitions from a dictionary to learn a latent LDA space over senses, which is then used to construct sense-specific classifiers by exploiting the text surrounding an image.

In general, VSD for nouns is a relatively straightforward task that can be solved with the help of an object detector [22], [23]. This is helped by resources such as ImageNet [24], a large image database containing 1.4 million images for 21,841 noun senses and organized according to the WordNet hierarchy. However, we are not aware of any previous work on VSD for verbs, and no ImageNet for verbs exists. Not only image retrieval would benefit from VSD, but also other multimodal tasks that have recently received a lot of interest, such as automatic image description [25] and visual question answering [26].

Action Recognition As mentioned in the introduction, our work relates to a variety of action recognition tasks. To elucidate key aspects of VSD and differences from previous approaches, we provide an overview of commonly used datasets for action recognition in Table 1. We observe that the number of verbs covered in these datasets is often smaller than the number of action labels reported (see columns #V and #L) and in many cases the action labels involve an object reference. A few of the first action recognition datasets (e.g., Ikizler [1] and Willow [27]) were taken from the sports domain, aiming to capture variation in human poses for actions such as *tennis serve* and *cricket bowling*. As a result, they contain images exhibiting diversity in camera view point, background, and resolution. Further datasets were created based on the intuition that object information helps in modeling action recognition [34], [35], using mutually exclusive labels such as *ride horse* or *ride bike*.

The limitations of the early datasets (small size, domain specificity, and the use of ad-hoc labels) have been recently addressed in a number of broad-coverage resources that are large scale and use linguistically-motivated labels [7], [10], [32]. Often these datasets use existing linguistic resources such as VerbNet [36], WordNet, and FrameNet [37] to classify verbs. This allows for a more general, semantically motivated treatment of verbs and verb phrases, and also takes into account the fact that not all verbs are depictable. For example, abstract verbs such as *presume* and *acquire* are not depictable, while other verbs have both depictable and non-depictable senses: *play* is non-depictable in *play with emotions*, but depictable in *play an instrument* and *play a sport*. A few datasets have been based on Microsoft Common Objects in Context (COCO; [38]), a dataset that consists of over 120k images with extensive annotations, including labels for 91 object categories and five descriptions per image. Although COCO was not created with action recognition in mind, it is possible to use the verbs present in the descriptions to annotate actions and their semantic roles [9], [32].

It is important to note that verb sense ambiguity is ignored in almost all existing action recognition datasets (and corresponding tasks). This misses important generalizations: for instance, the actions *ride horse* and *ride elephant* represent the same sense of *ride* and thus share visual, textual, and conceptual features. On the other hand, *play tennis* and *play guitar* share the same verb but represent different senses. We address this issue by creating VerSe, a dataset with explicit sense labels. VerSe is built on top of TUHOI (the Trento Universal Human-Object Interaction dataset; [6]) and COCO. The former dataset contains 10,805 images covering 2,974 actions. Action categories were crowdsourced, each image was

1. Throughout this paper we denote senses in sans serif font.

Dataset	Task	#L	#V	Obj	Images	Sen	Des	CIn	ML	Resource	Example Labels
Ikizler [1]	AC	6	6	0	467	N	N	Y	N	—	run, walk
Sports Dataset [2]	AC	6	6	4	300	N	N	Y	N	—	tennis serve, cricket bowling
Willow [27]	AC	7	6	5	986	N	N	Y	Y	—	ride bike, take photograph
PPMI [3]	AC	24	2	12	4.8k	N	N	Y	N	—	play guitar, hold violin
Stanford 40 Actions [5]	AC	40	33	31	9.5k	N	N	Y	N	—	cut vegetables, ride horse
PASCAL 2012 [28]	AC	11	9	6	4.5k	N	N	Y	Y	—	ride bike, ride horse
89 Actions [29]	AC	89	36	19	2k	N	N	Y	N	—	ride bike, fix bike
MPII Human Pose [30]	AC	410	—	66	40.5k	N	N	Y	N	—	ride car, hair styling
TUHOI [6]	HOI	2974	—	189	10.8k	N	N	Y	Y	—	sit on chair, play with dog
BU101 Dataset [31]	AC	101	68	—	23.8k	N	N	Y	N	—	horse race, play violin
COCO-a [32]	HOI	—	140	80	10k	N	Y	Y	Y	VerbNet	walk bike, hold bike
Google Images [33]	AC	2880	—	—	102k	N	N	N	N	—	riding horse, riding camel
HICO [7]	HOI	600	111	80	47k	Y	N	Y	Y	WordNet	ride#v#1 bike; hold#v#2 bike
VCOCO-SRL [9]	VSRL	—	26	48	10k	N	Y	Y	Y	—	verb: hit; instrument: bat; object: ball
imSitu [10]	VSRL	—	504	11k	126k	Y	N	Y	N	FrameNet WordNet	verb: ride; agent: girl#n#2 vehicle: bike#n#1; place: road#n#2
VerSe (Ours)	VSD	163	90	—	3.5k	Y	Y	Y	N	OntoNotes	ride.v.01, play.v.02

TABLE 1: Comparison of existing action recognition datasets according to various subtasks (AC stands for action classification, HOI for human object interaction recognition, VSRL for visual semantic role labeling, and VSD for visual verb sense disambiguation); #L denotes the number of action labels in the dataset; #V denotes the number of verbs covered in the dataset; Obj indicates the number of objects annotated; Sen indicates whether sense ambiguity is explicitly handled; Des indicates whether image descriptions are included; CIn denotes whether the dataset has been manually verified; ML indicates the possibility of multiple labels per image; Resource indicates whether a linguistic resource was used to label actions.

labeled by multiple annotators with a description in the form of a verb or a verb-object pair. The main drawback of TUHOI is that 1,576 out of 2,974 action categories occur only once, limiting its usefulness for VSD. Although COCO contains no explicit action annotation, verbs and verb phrases can be extracted from the descriptions. (But note that only about half of the COCO images depict actions.)

The recently created HICO (Humans Interacting with Common Objects) dataset is conceptually similar to VerSe. It consists of 47,774 images annotated with 111 verbs and 600 human-object interaction categories. Unlike other existing datasets, HICO uses sense-based distinctions: actions are denoted by sense-object pairs, rather than by verb-object pairs. HICO does not aim for complete coverage of senses: it restricts itself to a single sense of a verb (with the exceptions of a couple of verbs), which means that HICO is not suitable for verb sense disambiguation.

The COCO-a dataset [32] was created by identifying verbs that are visual and detectable in images.² This strategy meant that synonyms or related verbs were not included in the dataset, and also polysemous uses of verbs were excluded. The authors cross-checked the verbs they selected against the verbs used in the COCO image descriptions. This resulted in a total of 140 visual verbs being covered in COCO-a.

Another recent dataset is imSitu [10], which includes a large number of images and annotates each image with a verb and its semantic frames taken from FrameNet [37]. Each semantic frame includes a frame label (e.g., gardening), the frame elements (e.g., agent, tool), and the location (e.g., outdoors). The frame annotation by definition determines the sense of a verb. However, when imSitu was designed, it was decided not to include polysemous verbs, so for example the verb *play* is not in the dataset. Because all the verbs in the dataset only have one sense, imSitu cannot be used for visual sense disambiguation.

2. The selection criteria included that a 6–8 year old child should be able to distinguish the visual verbs.

3 THE VERSE DATASET

In this section we describe how VerSe was created. As mentioned earlier, it is based on COCO and TUHOI, covers 90 verbs, and contains 3,518 images. VerSe serves two main purposes: (1) to show the feasibility of annotating images with verb senses (rather than verbs or actions); (2) to function as test bed for evaluating automatic visual sense disambiguation methods.

Verb Selection Action recognition datasets often use a limited number of verbs in a given domain (see Table 1). We instead sampled verbs from COCO descriptions and TUHOI verb phrases (e.g., *sit on chair*), which we use in lieu of descriptions. We extracted all verbs from all descriptions in the two datasets and selected those with more than one sense in the OntoNotes dictionary [11]. This procedure resulted in 148 verbs in total (94 from COCO and 133 from TUHOI).

Depictability Annotation A verb can have multiple senses, but not all of them are depictable, e.g., senses describing cognitive and perception processes are not depictable. Consider the verb *touch* whose make physical contact sense is depictable, whereas the affect emotionally sense is not depictable. We therefore first annotated the senses of a verb as depictable or non-depictable. Amazon Mechanical Turk (AMT) workers were presented with the definitions of all the senses of a verb, along with examples, as given by OntoNotes [11]. An example for this annotation is shown in Figure 4. We used OntoNotes instead of WordNet, as WordNet senses are very fine-grained and potentially make depictability and sense annotation harder (see below). Granularity issues with WordNet for text-based WSD are well documented [12].

OntoNotes lists 921 senses for our 148 target verbs. For each sense, three AMT workers selected all depictable senses. The majority label was used as the gold-standard for subsequent experiments. This resulted in a 504 depictable senses. Inter-annotator agreement (ITA) as measured by Fleiss’ Kappa

Verb type	Examples	Verbs	Images	Senses	Depct	ITA
Motion	run, walk, jump, etc.	39	1812	10.76	5.79	0.680
Non-motion	sit, stand, lay, etc.	51	1698	8.27	4.86	0.636

Fig. 4: Example item for depictability and sense annotation: sense definitions and examples (in blue) for the verb *touch*.

Verb type	Examples	Verbs	Images	Senses	Depct	ITA
Motion	run, walk, jump, etc.	39	1812	10.76	5.79	0.680
Non-motion	sit, stand, lay, etc.	51	1698	8.27	4.86	0.636

TABLE 2: Overview of VerSe dataset divided into motion and non-motion verbs; Depct: depictable senses; ITA: inter-annotator agreement.

was 0.645.

Sense Annotation We then annotated a subset of the images in COCO and TUHOI with verb senses. An image was assigned the verb that occurs most frequently in the descriptions for that image (for TUHOI, the descriptions are verb-object pairs, see above). Although multiple verbs can be applicable in a given image, we only annotated the most frequently occurring verb. Perhaps not surprisingly, we observed that the distribution of verbs and their corresponding images is Zipfian: there are many verbs represented by a few images, and a few verbs represented by a large number of images. For sense annotation, we selected only verbs for which either COCO or TUHOI contained five or more images, resulting in a set of 90 verbs (out of the total 148). All images for these verbs were included, resulting in a dataset of 3,528 images: 2,340 images for 82 verbs from COCO and 1,188 images for 61 verbs from TUHOI (some verbs occur in both datasets).

These image-verb pairs formed the basis for sense annotation. AMT workers were presented with the image and all the depictable OntoNotes senses of the associated verb. The workers had to choose the sense of the verb that was instantiated in the image (or “none of the above”, in the case of irrelevant images). Annotators were given sense definitions and examples, as in the depictability annotation (see Figure 4). For every image-verb pair, five annotators performed the sense annotation task. A total of 157 annotators participated, reaching an inter-annotator agreement of 0.659 (Fleiss’ Kappa). Out of 3,528 images, we discarded 18 images annotated with “none of the above”, resulting in a set of 3,510 images covering 90 verbs and 163 senses. Number of images per verb sense varied from 1 – 100. We present statistics of our dataset in Table 2; we group the verbs into motion verbs and non-motion verb using Levin verb classes [39].

4 VISUAL VERB SENSE DISAMBIGUATION

For our disambiguation task, we assume we have a set of images I , and a set of polysemous verbs V and each image $i \in I$ is paired with a verb $v \in V$. For example, Figure 2 shows different images paired with the verb *play*. Every verb $v \in V$, has a set of senses $\mathcal{S}(v)$, described in a dictionary \mathcal{D} . Now, given an image i paired with a verb v , our task is to predict the correct sense $\hat{s} \in \mathcal{S}(v)$,

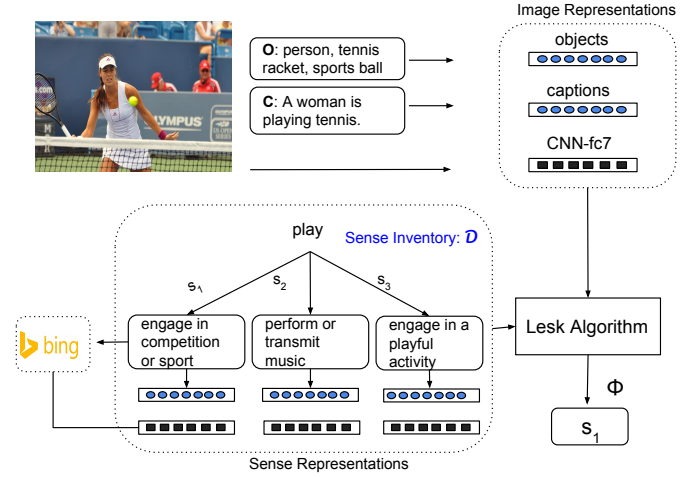


Fig. 5: Schematic overview of the visual sense disambiguation model.

i.e., the sense that is depicted by the associated image. In Figure 2, the correct sense for the first image is participate in sport, for the second one it is play an instrument, and so on.

The disambiguation task can be performed in a supervised manner, using samples of images, verbs, and their manually annotated senses. In this case, a classifier is used to assign each verb its appropriate sense based on evidence from contextual features extracted from the accompanying image or any textual information available. While this approach often achieves high accuracy, adequately large sense labeled data sets are difficult to obtain across languages and sense inventories. We therefore also explore an unsupervised approach which requires no sense annotated training data (we use the sense annotations in the VerSe dataset only for evaluation). For unsupervised sense disambiguation, we propose a new variant of the Lesk algorithm [40], a well-known approach to text-based WSD, which relies on the calculation of the word overlap between the sense definitions and the context in which a word occurs. The algorithm uses the following scoring function to disambiguate the sense of a verb v :

$$\hat{s} = \arg \max_{s \in \mathcal{S}(v)} \Phi(s, v, \mathcal{D}) = |\text{context}(v) \cap \text{definition}(s, \mathcal{D})| \quad (1)$$

Here, $\text{context}(v)$ is the set of words that occur close to the target word v and $\text{definition}(s, \mathcal{D})$ is the set of words in the definition of sense s in dictionary \mathcal{D} .

In our case, $\text{context}(v)$ is the image i associated with v . We create a representation for a given image (the vector \mathbf{i}), which can be text-based (using the object labels and descriptions for i), visual, or multimodal. Similarly, we create text-based, visual, and multimodal representations (the vector \mathbf{s}) for every sense s of a verb. Based on the representations \mathbf{i} and \mathbf{s} (detailed below), we score senses as:³

$$\hat{s} = \arg \max_{s \in \mathcal{S}(v)} \Phi(s, v, i, \mathcal{D}) = \mathbf{i} \cdot \mathbf{s} \quad (2)$$

An overview of our method is given in Figure 5. The various image representations (visual, textual, and multimodal) also serve as features in the supervised setting. In that setting, there is no need to represent senses; the sense are simply labels the classifier

3. Taking the dot product of two normalized vectors is equivalent to using cosine as similarity measure. We experimented with other similarity measures, but cosine performed best.

learns to predict. In the following, we will describe in more detail how we obtain image and sense representations.

4.1 Image Representations

Visual Modality Creating a visual representation \mathbf{i}^c of an image i is straightforward. We used the VGG 16-layer architecture (VGGNet) trained on 1.2M images of the 1,000 class ILSVRC 2012 object classification dataset, a subset of ImageNet [41]. This CNN model has a top-5 classification error of 7.4% on ILSVRC 2012. We used the publicly available reference model implemented using CAFFE [42] to extract the output of the fc7 layer, i.e., a 4,096 dimensional vector \mathbf{c}_i , for every image i . We use this vector as our image representation.

Textual Modality We also explore the possibility of representing the image indirectly, viz., through text associated with it in the form of object labels (O) or image descriptions (C), as shown in Figure 5. We experiment with two different forms of textual annotation: gold-standard (GOLD) annotation, where object labels and descriptions are provided by human annotators, and predicted (PRED) annotation, where state-of-the-art object recognition and image description generation systems are applied to the image.

GOLD object annotations are provided with the two datasets we use. Images sampled from COCO are annotated with one or more of 91 object categories. Images from TUHOI are annotated with one more of 189 object categories. PRED object annotations were generated using the same VGG 16-layer CNN object recognition model that was used to compute visual representations. Only object labels with an object detection threshold $t > 0.2$ were used.

To obtain GOLD image descriptions, we used the used human-generated descriptions that come with COCO. For TUHOI images, we generated descriptions of the form subject-verb-object tuples, where the subject is always *person*, and the verb-object pairs are the action labels that come with TUHOI. To obtain PRED descriptions, we generated three descriptions for every image using the state-of-the-art image description system of Vinyals et al. [43].⁴

We create a textual representation \mathbf{i}^t of image i using word2vec [44], a widely used model of word embeddings. Specifically, we obtain a vector for each object label and word in the image descriptions. An overall representation of the image is then computed by averaging these vectors over all labels, all content words in the description, or both. For our experiments we used the pre-trained 300 dimensional vectors available with the word2vec package (trained on part of the Google News dataset, about 100 billion words).

Modality Combination Apart from experimenting with separate textual and visual representations of images, it also makes sense to combine the two modalities into a multimodal representation. The simplest approach is a concatenation model which appends textual and visual features. More complex multimodal vectors can be created using methods such as Canonical Correlation Analysis (CCA; [45]) and Deep Canonical Correlation Analysis (DCCA; [46], [47]). CCA allows us to find a latent space in which the linear projections of text and image vectors are maximally correlated [48], [49]. DCCA can be seen as a non-linear version of CCA and has been successfully applied to the image description task [50], outperforming previous approaches, including kernel-based CCA.

4. We used Karpathy’s implementation, publicly available at <https://github.com/karpathy/neuraltalk>.

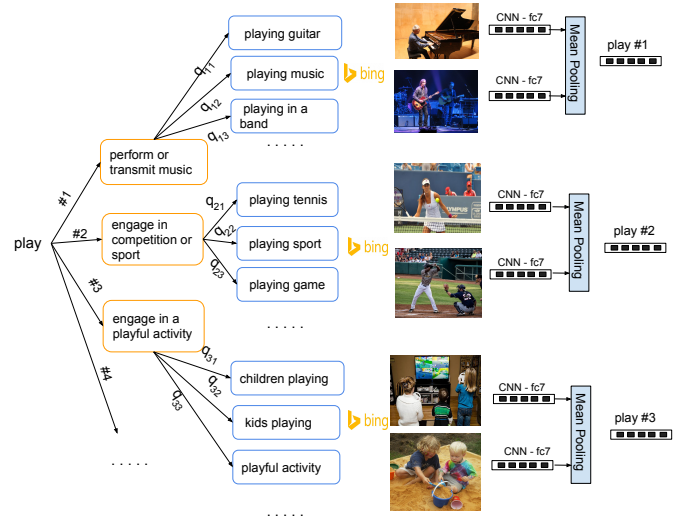


Fig. 6: Extracting visual sense representation for the verb *play*.

We use both CCA and DCCA to map the vectors \mathbf{i}^t and \mathbf{i}^c (which have different dimensions) into a joint latent space of $n = 300$ dimensions. We represent the projected vectors of textual and visual features for image i as $\mathbf{i}^{t'}$ and $\mathbf{i}^{c'}$ and combine them to obtain a multimodal representation \mathbf{i}^m as follows:

$$\mathbf{i}^m = \lambda \mathbf{i}^{t'} + (1 - \lambda) \mathbf{i}^{c'} \quad (3)$$

where λ is a parameter representing the relative importance of the textual and visual modalities.

4.2 Sense Representations

For unsupervised disambiguation, we must also obtain representations for verb senses (see Equation (2)). Analogously to image representations, we create a visual sense representation \mathbf{s}^c , a text-based sense representation \mathbf{s}^t , and one that combines both modalities.

Visual Modality Sense dictionaries typically provide sense definitions and example sentences, but no visual examples or images. For nouns, this is remedied by ImageNet [24], which provides a large number of example images for a subset of the senses in the WordNet noun hierarchy. However, no comparable resource is available for verbs (see Section 2).

In order to obtain visual sense representation \mathbf{s}^c , we therefore collected sense-specific images for the verbs in our dataset. For each verb sense s , three trained annotators were presented with the definition and examples from OntoNotes, and had to formulate a query $Q(s)$ that would retrieve images depicting the verb sense when submitted to a search engine. For every query q we retrieved images $I(q)$ using the Bing image search engine (for examples, see Figure 6). We used the top 50 images returned by Bing per query.

Images were converted into feature representations, using the output of the fc7 layer of VGGNet (same setup as in Section 4.1). To generate a visual representation for an individual sense \mathbf{s}^c , we perform mean pooling over the images obtained using the sense specific queries:

$$\mathbf{s}^c = \frac{1}{n} \sum_{q_j \in Q(s)} \sum_{i \in I(q_j)} \mathbf{c}_i \quad (4)$$

where n is the total number of images retrieved per sense s .

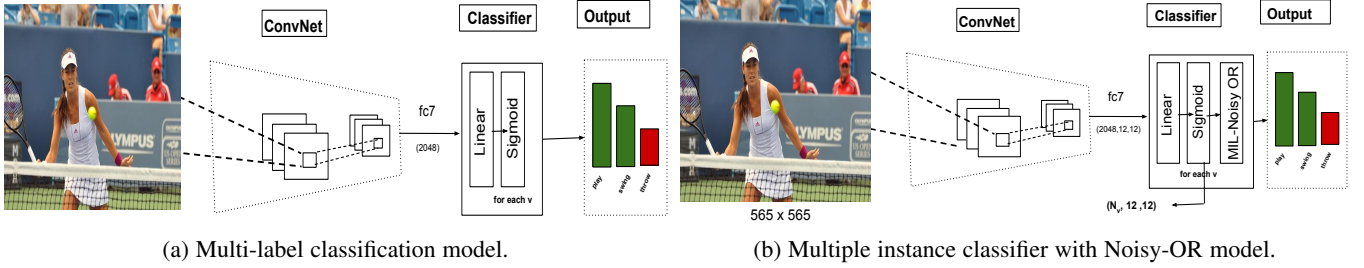


Fig. 7: Multi-label verb prediction classifiers.

Text-based Sense Representation We create a vector \mathbf{s}^t for every sense $s \in \mathcal{S}(v)$ of a verb v from its definition and the example usages provided in the OntoNotes dictionary \mathcal{D} . Again, we apply word2vec [44] to obtain a vector for every content word in the definition and examples of the sense and take the average of these vectors to compute an overall representation of the verb sense.

Modality Combination Visual and textual modalities for senses were combined as explained previously for images. We obtain a multimodal representation for sense s as follows:

$$\mathbf{s}^m = \lambda \mathbf{s}^v + (1 - \lambda) \mathbf{s}^c \quad (5)$$

where vectors \mathbf{s}^v and \mathbf{s}^c are projections of the visual and textual representations of sense s onto a joint latent space.

We use vectors $(\mathbf{i}^t, \mathbf{s}^t)$, $(\mathbf{i}^c, \mathbf{s}^c)$, and $(\mathbf{i}^m, \mathbf{s}^m)$ as described in Equation (2) to perform sense disambiguation.

5 VERB PREDICTION

So far we have focused on disambiguating verbs co-occurring with an image. In cases where images are not associated with textual information, it would be natural to first predict a verb representing the action depicted and then predict the verb sense (using the methods introduced in the previous sections). In the following, we describe two methods for predicting verbs given an image: (1) a multilabel CNN-based classification approach which simultaneously predicts all verbs associated with an image; and (2) a multiple instance learning approach which considers bags of positive and negative bounding boxes to decide which verb is compatible with the image.

5.1 Multilabel Classification

We trained a multilabel CNN to simultaneously predict all verbs depictable in a novel test image. Our vocabulary \mathcal{V} consists of the 250 most common verbs (including the 90 verbs in VerSe) in the descriptions of TUHOI, Flickr30k, and COCO datasets. We included Flickr30k as it has a more diverse distribution of verbs compared to COCO and the descriptions are action oriented [51].

We used a sigmoid cross entropy loss and optimized the ResNet 152-layer CNN architecture. We initialized the network weights with the publicly available CNN pretrained on ImageNet⁵ and finetuned it with our own verb labels. We used stochastic gradient descent with momentum set to 0.99 and a learning rate of $1e^{-5}$, i.e., lower than the original network to account for the sparsity of the labels in the training set. The network was trained with a batch size of one for three epochs. The CNN architecture for multilabel classification (MLC) is shown in Figure 7a.

5.2 Multiple Instance Learning

In addition to multilabel classification, we experimented with a weakly supervised model based on multiple instance learning (MIL; [52]) which has shown promising results in a variety of computer vision tasks including object detection [53], image description generation [54], scene classification [55], and action recognition [56], [57].

For each verb $v \in \mathcal{V}$, MIL samples sets of “positive” and “negative” bags of bounding boxes, where each bag corresponds to one image i . A bag b_i is positive if verb v is in image i ’s description, and negative otherwise. During training, instances within the positive bags are iteratively selected and the model is retrained using the updated positive labels. Compared to multilabel classification, which makes predictions considering the image as a whole, MIL is intuitively more appropriate for our task, since different parts of an image could represent different verbs.

We predict p_{ij}^v , the probability that a region j in image i corresponds to verb v , using a multi-layered convolutional neural network architecture which computes a logistic function on top of the last hidden layer (fc7; see [54] for more details):

$$p_{ij}^v = \frac{1}{1 + \exp(-(\mathbf{w}_v \phi(b_{ij}) + w_b))} \quad (6)$$

where $\phi(b_{ij})$ is the fc7 representation for image region j in image i , and \mathbf{w}_v, w_b are the weights and bias associated with verb v . We then use a noisy-OR version of MIL, where the probability of bag b_i depicting verb v is calculated from the probabilities of the individual instances in the bag:

$$p_i^v = 1 - \prod_{j \in b_i} (1 - p_{ij}^v) \quad (7)$$

Following previous work [54], we upsample images to 565 pixels and use a sliding window of 224×224 with a stride of 32. The noisy-OR version of MIL (Equation (7)) is implemented on top of 144 intermediate predictions p_{ij}^v (corresponding to each bounding box region b_{ij}) to compute a single probability p_i^v for each $v \in \mathcal{V}$. We use cross-entropy loss and optimize ResNet-152 (initialized with a CNN network pretrained on ImageNet) end-to-end with stochastic gradient descent. We use the same hyperparameter settings as in multilabel classification for three epochs. At test time, a novel image i is upsampled to 565 pixels to obtain the probability p_i^v for each verb $v \in \mathcal{V}$. The MIL architecture is shown in Figure 7b.

6 EXPERIMENTS

In the following, we report results for two sets of experiments. We first focus on visual sense disambiguation when the input to the system is an image and a verb associated with it and then move on to the more challenging task of detecting the verbs that are depicted in the image prior to predicting their senses.

5. <https://github.com/KaimingHe/deep-residual-networks#models>

6.1 Verb Sense Disambiguation

Table 3 summarizes the results of the unsupervised disambiguation method introduced in Section 4. We present results separately for motion and non-motion verbs in our gold-standard (GOLD) and predicted (PRED) settings. As explained earlier, we represent images and their senses by individual modalities (textual or visual) or their combination. To train the CCA and DCCA models, we use the text representations learned from image descriptions in the COCO and Flickr30k datasets as one view and the VGG-16 features from the respective images as the second view. We divide the data into train, test and development samples (using an 80/10/10 split). We use the trained models to generate the projected representations of text and visual features for the images in VerSe. Once the textual and visual features are projected, we merge them to get the multimodal representation. We experimented with two ways of combining visual and textual features projected via CCA or DCCA, namely interpolation (see Equations (3) and (5)) and concatenation.

To evaluate our proposed method, we compare against the first sense heuristic (FS), which defaults to the sense listed first in the dictionary (where senses are typically ordered by frequency). This is a strong baseline which is known to outperform more complex models in traditional text-based WSD. In VerSe we observe skew in the distribution of the senses and the first sense heuristic is as strong as it is on text. We further report the performance of the most frequent sense heuristic (MFS), which assigns the most frequently annotated sense for a given verb in VerSe. Note that MFS is supervised (as it requires sense annotated data to obtain the frequencies), so it should be regarded as an upper limit on the performance of the unsupervised methods we propose (as is also the case in unsupervised WSD for text [12]).

In the GOLD setting we find that for both types of verbs, textual representations based on image descriptions (C) outperform visual representations (CNN features). The text-based results compare favorably to the original Lesk algorithm (as described in Equation (1)), which performs at 30.7 for motion verbs and 36.2 for non-motion verbs in the GOLD setting. This improvement is clearly due to the use of word2vec embeddings.⁶ Note that CNN-based visual features alone perform better than gold-standard object labels alone in the case of motion verbs.

We also observed that adding visual features to textual features improves performance in some cases: multimodal features perform better than textual features alone both for object labels (CNN+O) and for image descriptions (CNN+C). However, adding CNN features to textual features based on both object labels and descriptions (CNN+O+C) results in a small decrease in performance. Furthermore, we note that CCA models outperform simple vector concatenation in case of GOLD setting for motion verbs, and overall DCCA performs considerably worse than concatenation. For CCA and DCCA we report the best performing scores achieved using weighted interpolation of textual and visual features with $\lambda = 0.5$.

When comparing to our baseline and upper limit, we find that all GOLD models which use descriptions-based representations (except DCCA) outperform the first sense heuristic for motion-verbs (accuracy 70.8), but not for non-motion verbs (accuracy 80.6). As expected, both motion and non-motion verbs perform significantly below the most frequent sense heuristic (accuracy

86.2 and 90.7 respectively), which provides an upper limit for unsupervised approaches.

We now turn to results obtained using object labels and image descriptions predicted by state-of-the-art automatic systems (PRED configuration). This is arguably a more realistic scenario, as it only requires images as input, rather than human-generated object labels and image descriptions (though object detection and image description systems are required instead). In the PRED setting, we find that textual features based on object labels (O) outperform both first sense heuristic and textual features based on image descriptions (C) in the case of motion verbs. Combining textual and visual features via concatenation improves performance for both motion and non-motion verbs. The overall best performance of 72.6 is obtained by combining CNN features and embeddings based on object labels and outperforms the first sense heuristic in case of motion verbs (accuracy 70.8). In the PRED setting for both classes of verbs the simpler concatenation model performs better than the more complex CCA and DCCA models. Note that for CCA and DCCA we report the best performing scores achieved using weighted interpolation of textual and visual features with $\lambda = 0.3$. Overall, our findings are consistent with the intuition that motion verbs are easier to disambiguate than non-motion verbs, as they are more depictable and likely to involve objects. This is also reflected in the higher inter-annotator agreement for motion verbs (see Table 2).

In order to better understand where the proposed unsupervised algorithm fails, we analyzed images that were disambiguated incorrectly. In the PRED setting, we observed that automatically generated image descriptions obtained lower scores compared to predicted object labels. The main reason for this is that the generated descriptions are often unrelated to the action depicted, whereas the object labels predicted by the CNN model are mostly topical and related to the image. This highlights that current image description systems still have clear limitations, despite high evaluation scores reported in the literature [43], [54]. Examples of images which were assigned incorrect senses are shown in Table 4 together with automatically generated descriptions and object labels.

We also investigated disambiguation performance in a supervised setting. Specifically, we trained logistic regression classifiers for sense prediction by dividing the images in VerSe into training and testing. To train the classifiers (one per verb), we selected verbs which have at least 20 images and at least two senses in VerSe.⁷ This resulted in 19 motion verbs and 19 non-motion verbs. The classifiers used textual (O, C) and visual (CNN) features, either in isolation or combined. Our results are summarized in Table 5; for comparison, we also report the scores of our unsupervised algorithm on the same set of verbs (in both GOLD and PRED settings).

We observe that supervised classifiers perform better than the first sense baseline (for both motion and non-motion verbs). In most cases multimodal features (CNN+C+O) outperform textual or visual features alone especially in the PRED setting, which is arguably the more realistic scenario. The features from PRED image descriptions show better results for non-motion verbs for both supervised and unsupervised approaches, whereas PRED object features show better results for motion verbs. We also find that supervised classifiers outperform the most frequent sense for

6. We also experimented with Glove vectors [58] but observed that word2vec representations consistently achieved better results than Glove vectors.

7. Few verbs such as *board*, *hang* only had one sense annotated in VerSe. Few other verbs have very skewed distribution of senses resulting in 5 or less number of images per sense. We ignore all such verbs.

Using GOLD annotations for objects and captions

	Images	FS	MFS	Textual			Vis	Concat (CNN+)			CCA (CNN+)			DCCA (CNN+)		
				O	C	O+C	CNN	O	C	O+C	O	C	O+C	O	C	O+C
Motion	1812	70.8	86.2	54.6	73.3	75.6	58.3	66.6	74.7	73.8	50.5	75.4	74.0	52.4	66.3	68.3
Non-Motion	1698	80.6	90.7	57.0	72.7	72.6	56.1	66.0	72.2	71.3	53.6	71.6	70.2	57.3	59.8	55.1

Using PRED annotations for objects and captions

	Images	FS	MFS	Textual			Vis	Concat (CNN+)			CCA (CNN+)			DCCA (CNN+)		
				O	C	O+C	CNN	O	C	O+C	O	C	O+C	O	C	O+C
Motion	1812	70.8	86.2	65.1	54.9	61.6	58.3	72.6	63.6	66.5	54.0	56.6	56.2	57.1	56.5	56.2
Non-Motion	1698	80.6	90.7	59.0	64.3	64.0	56.1	63.8	66.3	66.1	50.7	55.3	54.8	49.5	50.0	50.0

TABLE 3: Sense disambiguation scores for **gold-standard verbs**: accuracy scores for motion and non-motion verbs using different types of sense and image representations (O: object labels, C: image descriptions, CNN: image features, FS: first sense heuristic, MFS: most frequent sense heuristic). Model configurations that performed the best are shown in **bold**.




Image	Descriptions	Objects
	A man holding a nintendo wii game controller. A man and a woman playing a video game. A man and a woman are playing a video game.	person, bassoon, violin fiddle, oboe, hautboy
play: perform or transmit music , engage in competition		
	A woman standing next to a fire hydrant. A woman walking down a street holding an umbrella. A woman standing on a sidewalk holding an umbrella.	person, horizontal bar, high bar, pole
swing: move in a curve or arc , hang freely		
	A couple of cows standing next to each other. A cow that is standing in the dirt. A close up of a horse in a stable	arabian camel, dromedary, per- son
feed: give food , eat , be sustained on		

TABLE 4: Images assigned an incorrect sense (shown in red) in the PRED setting. Gold-standard senses are shown in blue.

Features	Motion verbs: 19, MFS: 76.1				Non-Motion Verbs: 19, MFS: 80.0				
	GOLD		PRED		GOLD		PRED		
	Sup	Unsup	Sup	Unsup	Sup	Unsup	Sup	Unsup	
FS	60.0	60.0	60.0	60.0	FS	71.3	71.3	71.3	71.3
O	82.3	35.3	80.0	43.8	O	79.1	48.6	78.2	46.0
C	78.4	53.8	69.2	41.5	C	79.1	53.9	77.3	61.7
O+C	80.0	55.3	70.7	45.3	O+C	79.1	66.0	77.3	55.6
CNN	82.3	58.4	82.3	58.4	CNN	80.0	55.6	80.0	55.6
CNN+O	83.0	48.4	83.0	60.0	CNN+O	80.0	56.5	80.0	52.1
CNN+C	82.3	66.9	82.3	53.0	CNN+C	80.0	56.5	80.3	60.0
CNN+O+C	83.0	58.4	83.0	55.3	CNN+O+C	80.0	59.1	80.0	55.6

TABLE 5: Accuracy scores for motion and non-motion verbs for supervised and unsupervised approaches using different types of sense and image representation features (O: object labels, C: image descriptions, CNN: image features, FS: first sense heuristic, MFS: most frequent sense heuristic). Configurations that perform the best are shown in **bold**.

motion verbs, whereas for non-motion verbs our scores match the most frequent sense heuristic.

6.2 Verb Prediction and Sense Disambiguation

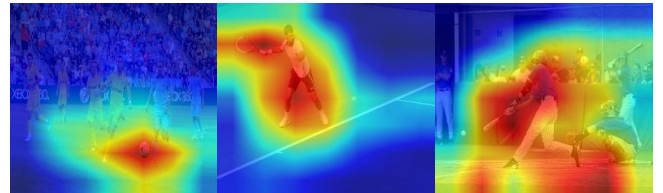
We measure verb prediction performance using both accuracy and mean average precision (mAP). If a verb is used in at least one of the gold-standard image descriptions, it is included as a positive instance; as a result, an image can have multiple gold-standard verb labels. Both MLC and MIL systems output a distribution of verbs given an image. We consider verbs with probability higher than a threshold $\tau = 0.2$ as positive predictions.

Verb type	Verbs	Images	Accuracy		mAP	
			MLC	MIL	MLC	MIL
Motion	39	1,812	46.96	50.60	35.81	41.47
Non-motion	51	1,698	34.82	37.47	31.12	35.27

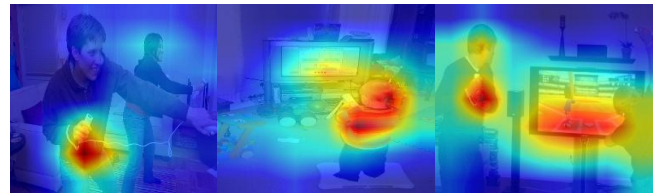
TABLE 6: Verb prediction accuracy and mAP on VerSe; MIL: Multiple Instance Learning; MLC: Multi-label classification.



(a) play instrument



(b) play sport



(c) children play video games

Fig. 8: Localizations for different senses of the verb *play*.

Table 6 summarizes the performance of MLC and MIL. As can be seen, MIL performs best both in terms of accuracy and mAP, across motion and non-motion verbs. Among motion verbs, the most accurately predicted ones were *drive*, *fly*, *ride*, *play*; for non-motion verbs *sit* and *hold* were most accurate. Figure 9 shows visualizations of different verbs detected in images, while Figure 10 shows examples of verbs predicted by the MIL and MLC models for three different images. In Figure 8 we also show the visualizations of different senses of the verb *play*, which indicate that depending on the sense of verb being depicted our models are

Using GOLD annotations for objects and captions

	Images	FS	MFS	Textual			Vis	Concat (CNN+)			CCA (CNN+)			DCCA (CNN+)		
				O	C	O+C	CNN	O	C	O+C	O	C	O+C	O	C	O+C
Motion	918	68.4	87.3	58.3	78.7	82.7	65.1	73.5	79.4	79.6	54.0	75.9	75.8	56.4	72.0	75.9
Non-Motion	637	83.8	92.3	63.7	78.1	80.5	58.7	73.3	76.9	76.7	59.6	73.4	70.1	61.9	63.1	61.2

Using PRED annotations for objects and captions

	Images	FS	MFS	Textual			Vis	Concat (CNN+)			CCA (CNN+)			DCCA (CNN+)		
				O	C	O+C	CNN	O	C	O+C	O	C	O+C	O	C	O+C
Motion	918	68.4	87.3	72.3	65.1	71.6	65.1	79.4	74.0	75.8	49.3	60.3	57.8	64.0	66.4	64.8
Non-Motion	637	83.8	92.3	65.7	77.3	76.2	58.7	70.0	74.4	74.2	49.6	59.1	59.1	54.0	53.0	54.6

TABLE 7: Sense disambiguation scores for **predicted verbs**: accuracy scores for motion and non-motion verbs using different types of sense and image representations (O: object labels, C: image descriptions, CNN: image features, FS: first sense heuristic, MFS: most frequent sense heuristic). Model configurations that performed the best are shown in **bold**.

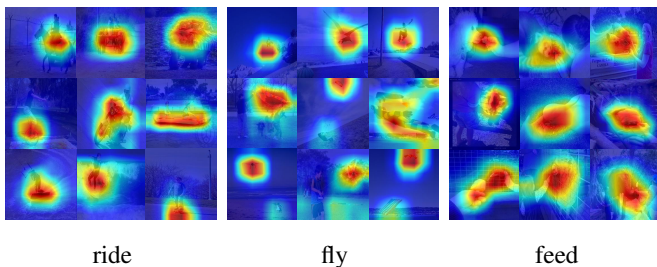


Fig. 9: Localizations for predicted verbs *ride*, *fly* and *feed*.



Fig. 10: Example verb predictions of MIL and MLC classifiers

localizing different aspects of the image. Finally, Table 8 provides examples of the best and worst performing verbs for MLC and MIL using average precision (AP). Although informative, AP is a pessimistic evaluation metric because we can not exhaustively annotate all possible verbs depicted in an image. Consider the case where our model predicts the verbs *stand*, *hold*, [*play*] for an image depicting a person playing tennis. The predictions are all correct, but AP would penalize us if those verbs are not in our gold-standard annotation.

To study in more detail the quality of the verb predictions, we conducted a human evaluation study. We presented the top 10 verbs predicted by the MIL classifier for a given image to Amazon Mechanical Turk workers and asked them to select those that apply. For this study, we sampled 640 images from VerSe across verbs and senses with 2–5 images per unique verb sense. For every image, we collected annotations from three workers. Overall, 54 workers took part in the study, with pair-wise inter-annotator agreement of 0.741.

Table 9 presents mean accuracy scores across all 640 images using human selected verbs as gold-standard labels. Specifically, we compute accuracy for every image based on (a) majority labels,

Verb	Count	MLC	MIL	Verb	Count	MLC	MIL
shoot	339	0.14	0.16	draw	985	50.37	63.27
drill	128	0.26	0.27	hit	6459	68.98	68.53
break	794	2.26	1.63	kick	1780	75.00	79.27
lift	980	3.89	3.98	paddle	1027	76.41	83.76
chase	745	4.35	5.05	fly	13395	80.90	85.19

TABLE 8: Average precision scores for individual verbs. Count refers to number of positive training instances. Verbs with the lowest and highest performance are shown.

	$\tau = 0.05$	$\tau = 0.1$	$\tau = 0.15$	$\tau = 0.2$	$\tau = 0.25$	$\tau = 0.3$
Majority	48.5	57.6	63.5	66.6	66.9	64.6
All	68.2	74.8	78.5	80.6	80.3	76.0

TABLE 9: Human evaluation accuracy scores for verb prediction labels. τ is the confidence threshold of verb predictions.

i.e., if at least two out of three annotators agreed that a particular verb is depicted in the image; and (b) all labels, i.e., if at least one annotator thought a particular verb is depicted in the image. The average number of verbs selected per image is 4.17 for majority labels and 6.18 for all labels. In Table 9 we present the accuracy scores against the gold-standard from the human annotation whilst we vary τ , the prediction confidence threshold. As can be seen, the best accuracies are achieved at $\tau = 0.2$ and $\tau = 0.25$. Overall, most verb predictions are considered appropriate by humans, even under the stricter majority label criterion.

Sense accuracy scores for predicted verbs are shown in Table 7. Again, scores are shown for motion and non-motion verbs separately. We report results for unsupervised methods, using the multiple instance learning approach to obtain verb predictions. Here, we only consider images for which the MIL system predicted the same verbs as in VerSe. These are 918 images compared to 1,812 in the full dataset. For this reason, we do not report supervised experiments in the predicted verb setting: there are not enough image-verb instances to train a supervised classifier. Also notice that even though several of the verbs predicted by the MIL system may be appropriate for VerSe images, we do not have sense annotations for them to perform either evaluation or training. Overall, sense disambiguation results for predicted verbs follow the same pattern as those obtained from observed verbs: motion verbs are easier to disambiguate than non-motion ones; in the GOLD setting best model performance is achieved with object labels and image descriptions combined, whereas in the PRED setting concatenation of CNN features with object labels yields best results.

7 CONCLUSIONS

In this article, we introduced the new task of visual verb sense disambiguation: given an image and a verb, identify the verb sense depicted in the image. We developed VerSe, a new dataset with verb sense annotation based on the COCO and TUHOI datasets. We evaluated supervised and unsupervised visual sense disambiguation models and demonstrated that both textual and visual information associated with an image can contribute to sense disambiguation. In an in-depth analysis of various image representations we showed that object labels and visual features extracted using state-of-the-art convolutional neural networks result in good disambiguation performance, while automatically generated image descriptions were shown to be less useful.

We also explored a second scenario for visual sense disambiguation, where we assumed that only the image is given, and both the verb and its sense need to be predicted. We conceptualized this as a two-stage process: First, we predicted verb labels using multi-instance learning or multilabel classification. Then, we disambiguated the predicted verbs using our sense disambiguation approach combining visual and textual features. We showed that the verbs predicted by this method agree well with human intuitions, and we also obtained good sense accuracy scores. Note that the second scenario differs from our first scenario in a crucial respect: we are able to predict multiple verbs per image, and each of these verbs can be associated with a different image region (if the multi-instance learning model is used). While a lot of images in our dataset only depict a single action, this is not always the case (e.g., the child in the rightmost image in Figure 10 is both sitting in the sand and holding a toy).

In this work, we explored visual sense disambiguation as a standalone task. We did not yet show that applications benefit from VSD; this is an important project for future work. An obvious example would be image search: recall Figure 3, which depicts a search result obtained with the verb *sit* as query. If the search engine had access to verb sense disambiguation for images, then it would be able to cluster the search results based on verb senses, rather than forming groups based on image or query similarity.

Other language/vision task that are also likely to benefit include image description and visual question answering. An image description system that has access to verb prediction and sense disambiguation can make sure that it outputs only descriptions that are compatible with the verb senses that are attested in the image it tries to describe. A simple re-ranking architecture could be used to implement this: We take an existing image description system, use it to generate a set of candidate descriptions for a given image, and then re-rank the descriptions based on the output of our verb prediction and VSD models. In a similar fashion, VSD could be used to re-rank the output of a visual question answering system (or the VSD scores could simply serve as a feature).

Another important area for future research is the connection between verb sense ambiguity and translation ambiguity. This rests on the observation that sense ambiguity in one language can manifest itself as ambiguity in lexical choice in another language. The English verb *ride*, for instance, can have the senses (1) sit on and control a vehicle (as in *ride bicycle*), or (2) sit and travel on the back of animal (as in *ride horse*). These two senses corresponds to two different lexical choices in German, viz., the verbs *fahren* (for sense 1) and *reiten* (for sense 2). In other words, we need to sense disambiguate the verb in order to translate it correctly. This observation is of practical importance, as machine translation systems often suffer from disambiguation errors such as this [59].

If the ambiguous verb occurs in a visual context, then we can apply the VSD methods developed in this article to the resolution of translation ambiguities as they occur in multilingual image description or crosslingual image retrieval. Again, this is something we would like to explore in future work (preliminary results for multilingual image description are presented in [60]).

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their helpful comments. The authors gratefully acknowledge the support of the European Research Council (Lapata: award 681760) and of the Leverhulme Trust (Keller: award IAF-2017-019).

REFERENCES

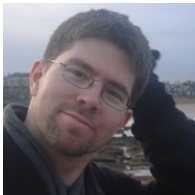
- [1] N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu, "Recognizing actions from still images," in *19th International Conference on Pattern Recognition (ICPR 2008)*, December 8-11, 2008, Tampa, Florida, USA, 2008, pp. 1–4. [Online]. Available: <http://dx.doi.org/10.1109/ICPR.2008.4761663>
- [2] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, 2009.
- [3] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 9–16.
- [4] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010. [Online]. Available: <http://dx.doi.org/10.1007/s11263-009-0275-4>
- [5] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1331–1338.
- [6] D.-T. Le, J. Uijlings, and R. Bernardi, *Proceedings of the Third Workshop on Vision and Language*. Dublin City University and the Association for Computational Linguistics, 2014, ch. TUHOI: Trento Universal Human Object Interaction Dataset, pp. 17–24. [Online]. Available: <http://aclweb.org/anthology/W14-5403>
- [7] Y. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "HICO: A benchmark for recognizing human-object interactions in images," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 1017–1025. [Online]. Available: <http://dx.doi.org/10.1109/ICCV.2015.122>
- [8] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *European Conference on Computer Vision*. Springer, 2016, pp. 852–869.
- [9] S. Gupta and J. Malik, "Visual semantic role labeling," *CoRR*, vol. abs/1505.04474, pp. 1–11, 2015.
- [10] M. Yatskar, L. Zettlemoyer, and A. Farhadi, "Situation recognition: Visual semantic role labeling for image understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 26-July 1, 2016*, 2016.
- [11] E. H. Hovy, M. P. Marcus, M. Palmer, L. A. Ramshaw, and R. M. Weischedel, "Ontonotes: The 90% solution," in *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA, 2006*, pp. 57–60. [Online]. Available: <http://acl.ldc.upenn.edu/N/N06/N06-2015.pdf>
- [12] R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 2, p. 10, 2009.
- [13] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990.
- [14] D. Lin, "Using syntactic dependency as local context to resolve word sense ambiguity," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997, pp. 64–71.
- [15] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll, "Finding predominant word senses in untagged text," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, pp. 279–286.

- [16] S. Brody and M. Lapata, "Good neighbors make good senses: Exploiting distributional similarity for unsupervised wsd," in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2008, pp. 65–72.
- [17] Z. Zhong and H. T. Ng, "It makes sense: A wide-coverage word sense disambiguation system for free text," in *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, System Demonstrations*, 2010, pp. 78–83. [Online]. Available: <http://www.aclweb.org/anthology/P10-4014>
- [18] S. Rothe and H. Schutze, "Autoextend: Extending word embeddings to embeddings for synsets and lexemes," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, 2015, pp. 1793–1803. [Online]. Available: <http://aclweb.org/anthology/P/P15/P15-1173.pdf>
- [19] S. K. Jauhar, C. Dyer, and E. H. Hovy, "Ontologically grounded multi-sense representation learning for semantic vector space models," in *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, 2015, pp. 683–693. [Online]. Available: <http://aclweb.org/anthology/N/N15/N15-1070.pdf>
- [20] N. Loeff, C. O. Alm, and D. A. Forsyth, "Discriminating image senses by clustering with multimodal features," in *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. Association for Computational Linguistics, 2006, pp. 547–554. [Online]. Available: <http://aclweb.org/anthology/P06-2071>
- [21] K. Saenko and T. Darrell, "Unsupervised learning of visual sense models for polysemous words," in *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, 2008, pp. 1393–1400. [Online]. Available: <http://papers.nips.cc/paper/3389-unsupervised-learning-of-visual-sense-models-for-polysemous-words>
- [22] K. Barnard, M. Johnson, and D. Forsyth, "Word sense disambiguation with pictures," in *Proceedings of the HLT-NAACL 2003 workshop on Learning word meaning from non-linguistic data-Volume 6*. Association for Computational Linguistics, 2003, pp. 1–5.
- [23] X. Chen, A. Ritter, A. Gupta, and T. M. Mitchell, "Sense discovery via co-clustering on images and text," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 5298–5306.
- [24] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, 2009*, pp. 248–255. [Online]. Available: <http://dx.doi.org/10.1109/CVPRW.2009.5206848>
- [25] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikidler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, 2016.
- [26] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: visual question answering," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 2425–2433. [Online]. Available: <http://dx.doi.org/10.1109/ICCV.2015.279>
- [27] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations," in *BMVC 2010-21st British Machine Vision Conference*, 2010.
- [28] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015. [Online]. Available: <http://dx.doi.org/10.1007/s11263-014-0733-5>
- [29] D. T. Le, R. Bernardi, and J. Uijlings, "Exploiting language models to recognize unseen actions," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 2013, pp. 231–238.
- [30] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.
- [31] S. Ma, S. A. Bargal, J. Zhang, L. Sigal, and S. Sclaroff, "Do less and achieve more: Training cnns for action recognition utilizing action images from the web," *Pattern Recognition*, vol. 68, pp. 334–345, 2017.
- [32] M. R. Ronchi and P. Perona, "Describing common human visual actions in images," in *Proceedings of the British Machine Vision Conference (BMVC 2015)*. BMVA Press, September 2015, pp. 52.1–52.12.
- [33] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rossenberg, and L. Fei-Fei, "Learning semantic relationships for better action retrieval in images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1100–1109.
- [34] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.
- [35] N. Ikidler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *European conference on computer vision*. Springer, 2010, pp. 494–507.
- [36] K. K. Schuler, "Verbnets: A broad-coverage, comprehensive verb lexicon," Ph.D. dissertation, University of Pennsylvania, 2005.
- [37] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley framenet project," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 1998, pp. 86–90.
- [38] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *CoRR*, vol. abs/1504.00325, 2015. [Online]. Available: <http://arxiv.org/abs/1504.00325>
- [39] B. Levin, *English verb classes and alternations: A preliminary investigation*. University of Chicago Press, 1993.
- [40] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," in *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC 1986, Toronto, Ontario, Canada, 1986, 1986*, pp. 24–26. [Online]. Available: <http://doi.acm.org/10.1145/318723.318728>
- [41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [42] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014, 2014*, pp. 675–678. [Online]. Available: <http://doi.acm.org/10.1145/2647868.2654889>
- [43] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 3156–3164. [Online]. Available: <http://dx.doi.org/10.1109/CVPR.2015.7298935>
- [44] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: <http://arxiv.org/abs/1301.3781>
- [45] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004. [Online]. Available: <http://dx.doi.org/10.1162/0899766042321814>
- [46] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, 2013, pp. 1247–1255. [Online]. Available: <http://jmlr.org/proceedings/papers/v28/andrew13.html>
- [47] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "On deep multi-view representation learning," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2015, pp. 1083–1092. [Online]. Available: <http://jmlr.org/proceedings/papers/v37/wangb15.html>
- [48] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, 2014, pp. 529–545. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10593-2_35
- [49] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics (extended abstract)," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, 2015, pp. 4188–4192. [Online]. Available: <http://ijcai.org/papers15/Abstracts/IJCAI15-593.html>
- [50] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 3441–3450. [Online]. Available: <http://dx.doi.org/10.1109/CVPR.2015.7298966>

- [51] 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*, vol. 2, pp. 67–78, 2014.
- [52] O. Maron and T. Lozano-Pérez, “A framework for multiple-instance learning,” in *Advances in Neural Information Processing Systems 10*, M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds., 1997.
- [53] C. Zhang, J. C. Platt, and P. A. Viola, “Multiple instance boosting for object detection,” in *Advances in neural information processing systems*, 2005, pp. 1417–1424.
- [54] H. Fang, S. Gupta, F. N. Iandola, R. K. 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 1473–1482. [Online]. Available: <http://dx.doi.org/10.1109/CVPR.2015.7298754>
- [55] O. Maron and A. L. Ratan, “Multiple-instance learning for natural scene classification,” in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*, 1998, pp. 341–349.
- [56] F. Sener, C. Bas, and N. Ikizler-Cinbis, “On recognizing actions in still images via multiple features,” in *European Conference on Computer Vision*. Springer, 2012, pp. 263–272.
- [57] A. Mallya and S. Lazebnik, “Learning models for actions and person-object interactions with transfer to question answering,” in *European Conference on Computer Vision*. Springer, 2016, pp. 414–428.
- [58] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1532–1543. [Online]. Available: <http://aclweb.org/anthology/D/D14/D14-1162.pdf>
- [59] D. Vilar, J. Xu, L. F. d’Haro, and H. Ney, “Error analysis of statistical machine translation output,” in *Proceedings of LREC*, 2006, pp. 697–702.
- [60] S. Gella, R. Sennrich, F. Keller, and M. Lapata, “Image pivoting for learning multilingual multimodal representations,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Short Papers*, Copenhagen, 2017, pp. 2829–2835.



Spandana Gella is a PhD candidate at the School of Informatics, University of Edinburgh. Her research interests include weakly supervised learning of action recognition, scene understanding, multilingual multimodal learning, and image description.



Frank Keller is professor in the School of Informatics, University of Edinburgh. His background includes an undergraduate degree from Stuttgart University, a PhD from Edinburgh, and postdoctoral and visiting positions at Saarland University and MIT. His research focuses on how people solve complex tasks such as understanding language or processing visual information. His work uses experimental techniques and computational modeling to investigate reading, sentence comprehension, translation, and language generation, both in isolation and in a multimodal context.



Mirella Lapata is a professor in the School of Informatics, University of Edinburgh. Her research interests include machine learning techniques for natural language understanding, generation, and grounded language acquisition.