Zero-Shot Activity Recognition with Verb Attribute Induction

Rowan Zellers and **Yejin Choi**
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
{rowanz,yejin}@cs.washington.edu

Abstract

In this paper, we investigate large-scale zero-shot activity recognition by modeling the visual and linguistic attributes of action verbs. For example, the verb "salute" has several properties, such as being a light movement, a social act, and short in duration. We use these attributes as the internal mapping between visual and textual representations to reason about a previously unseen action. In contrast to much prior work that assumes access to gold standard attributes for zero-shot classes and focuses primarily on object attributes, our model uniquely learns to infer action attributes from dictionary definitions and distributed word representations. Experimental results confirm that action attributes inferred from language can provide a predictive signal for zero-shot prediction of previously unseen activities.

1 Introduction

We study the problem of inferring action verb attributes based on how the word is defined and used in context. For example, given a verb such as "swig" shown in Figure 1, we want to infer various properties of actions such as *motion dynamics* (moderate movement), *social dynamics* (solitary act), *body parts* involved (face, arms, hands), and *duration* (less than 1 minute) that are generally true for the range of actions that can be denoted by the verb "swig".

Our ultimate goal is to improve *zero-shot learning* of activities in computer vision: predicting a previously unseen activity by integrating background knowledge about the conceptual properties of actions. For example, a computer vision system may have seen images of "drink" activities during training, but not "swig". Ideally, the system should infer the likely visual characteristics of "swig" using world knowledge implicitly available in dictionary definitions and word embeddings.

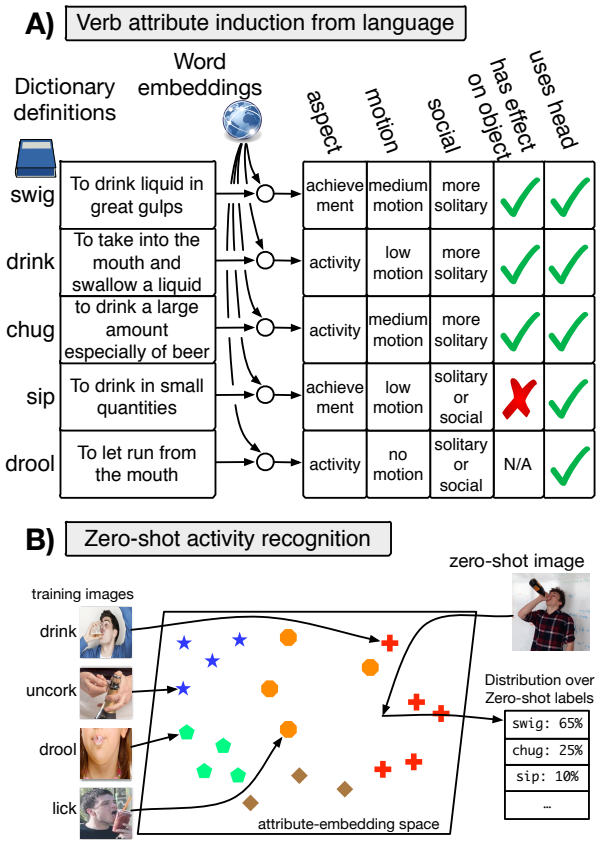

Figure 1: An overview of our task. Our goal is twofold. **A:** we seek to use use distributed word embeddings in tandem with dictionary definitions to obtain a high level understanding of verbs. **B:** we seek to use these predicted attributes to allow a classifier to recognize a broader set of activities than what was seen in training time.

However, most existing literature on zero-shot learning has focused on object recognition with only a few notable exceptions (see Related Work

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 955–967
Copenhagen, Denmark, September 7–11, 2017.

in Section 8). There are two critical reasons: *object attributes*, such as color, shape, and texture, are conceptually straightforward to enumerate. In addition, they have distinct visual patterns which are robust for current vision systems to recognize. In contrast, *activity attributes* are more difficult to conceptualize as they involve varying levels of abstractness, which are also more challenging for computer vision as they have less distinct visual patterns. Noting this difficulty, Antol et al. (2014) instead employ cartoon illustrations as intermediate mappings for zero-shot dyadic activity recognition. We present a complementary approach: that of tackling the abstractness of verb attributes directly. We develop and use a corpus of verb attributes, using linguistic theories on verb semantics (e.g., aspectual verb classes of Vendler (1957)) and also drawing inspiration from studies on linguistic categorization of verbs and their properties (Friedrich and Palmer, 2014; Siegel and McKeown, 2000).

In sum, we present the first study aiming to recover general action attributes for a diverse collection of verbs, and probe their predictive power for zero-shot activity recognition on the recently introduced imSitu dataset (Yatskar et al., 2016). Empirical results show that action attributes inferred from language can help classifying previously unseen activities and suggest several avenues for future research on this challenging task. We publicly share our dataset and code for future research.[1]

2 Action Verb Attributes

We consider seven different groups of action verb attributes. They are motivated in part by potential relevance for visual zero-shot inference, and in part by classical literature on linguistic theories on verb semantics. The attribute groups are summarized below.[2] Each attribute group consists of a set of attributes, which sums to $K = 24$ distinct attributes annotated over the verbs.[3]

[1] Aspectual Classes We include the aspectual verb classes of Vendler (1957):

- *state*: a verb that does not describe a changing situation (e.g. "have", "be")
- *achievement*: a verb that can be *completed* in a short period of time (e.g. "open", "jump")
- *accomplishment*: a verb with a sense of completion over a longer period of time (e.g. "climb")
- *activity*: a verb without a clear sense of completion (e.g. "swim", "walk", "talk")

[2] Temporal Duration This attribute group relates to the aspectual classes above, but provides additional estimation of typical time duration with four categories. We categorize verbs by best-matching temporal units: seconds, minutes, hours, or days, with an additional option for verbs with unclear duration (e.g., "provide").

[3] Motion Dynamics This attribute group focuses on the energy level of motion dynamics in four categories: no motion ("sleep"), low motion ("smile"), medium ("walk"), or high ("run"). We add an additional option for verbs whose motion level depends highly on context, such as 'finish.'

[4] Social Dynamics This attribute group focuses on the likely social dynamics, in particular, whether the action is usually performed as a solitary act, a social act, or either. This is graded on a 5-part scale from least social (-2) to either $(+0)$ to most social $(+2)$

[5] Transitivity This attribute group focuses on whether the verb can take an object, or be used without. This gives the model a sense of the implied action dynamics of the verb between the agent and the world. We record three variables: whether or not the verb is naturally transitive on a person ("I hug her" is natural), on a thing ("I eat it"), and whether the verb is intransitive ("I run").

[6] Effects on Arguments This attribute group focuses on the effects of actions on agents and other arguments. For each of the possible transitivities of the verb, we annotate whether or not it involves *location change* ("travel"), *world change* ("spill"), *agent or object change* ("cry") , or *no visible change* ("ponder").

[7] Body Involvements This attribute group specifies prominent body parts involved in carrying out the action. For example, "open" typically involves "hands" and "arms" when used in a physical sense. We use five categories: head, arms, torso, legs, and other body parts.

[1] Available at http://github.com/rowanz/verb-attributes

[2] The full list is available in the supplemental section.

[3] Several of our attributes are categorical; if converted to binary attributes, we would have 40 attributes in total.

Action Attributes and Contextual Variations In general, contextual variations of action attributes are common, especially for frequently used verbs that describe everyday physical activities. For example, while "open" typically involves "hands", there are exceptions, e.g. "open one's eyes". In this work, we focus on stereotypical or prominent characteristics across a range of actions that can be denoted using the same verb. Thus, three investigation points of our work include: (1) crowd-sourcing experiments to estimate the distribution of human judgments on the prominent characteristics of everyday physical action verbs, (2) the feasibility of learning models for inferring the prominent characteristics of the everyday action verbs despite the potential noise in the human annotation, and (3) their predictive power in zero-shot action recognition despite the potential noise from contextual variations of action attributes. As we will see in Section 7, our study confirms the usefulness of studying action attributes and motivates the future study in this direction.

Relevance to Linguistic Theories The key idea in our work that action verbs project certain expectations about their influence on their arguments, their pre- and post-conditions, and their implications on social dynamics, etc., relates to the original Frame theories of Baker et al. (1998a). The study of action verb attributes are also closely related to formal studies on verb categorization based on the characteristics of the actions or states that a verb typically associates to (Levin, 1993), and cognitive linguistics literature that focus on causal structure and force dynamics of verb meanings (Croft, 2012).

3 Learning Verb Attributes from Language

In this section we present our models for learning verb attributes from language. We consider two complementary types of language-based input: dictionary definitions and word embeddings. The approach based on dictionary definitions resembles how people acquire the meaning of a new word from a dictionary lookup, while the approach based on word embeddings resembles how people acquire the meaning of words in context.

Overview This task follows the standard supervised learning approach where the goal is to predict K attributes per word in the vocabulary $\mathcal{V}$. Let $x_v \in \mathcal{X}$ represent the input representation of a word $v \in \mathcal{V}$. For instance, x_v could denote a word embedding, or a definition looked up from a dictionary (modeled as a list of tokens). Our goal is to produce a model $f : \mathcal{X} \rightarrow \mathbb{R}^d$ that maps the input to a representation of dimension d. Modeling options include using pretrained word embeddings, as in Section 3.1, or using a sequential model to encode a dictionary, as in Section 3.2.

Then, the estimated probability distribution over attribute k is given by:

$$\hat{y}_{v,k} = \sigma(\mathbf{W}^{(k)} f(x_v)). \quad (1)$$

If the attribute is binary, then $\mathbf{W}^{(k)}$ is a vector of dimension d and σ is the sigmoid function. Otherwise, $\mathbf{W}^{(k)}$ is of shape $d_k \times d$, where d_k is the dimension of attribute k, and σ is the softmax function. Let the vocabulary $\mathcal{V}$ be partitioned into sets $\mathcal{V}_{train}$ and $\mathcal{V}_{test}$; then, we train by minimizing the cross-entropy loss over $\mathcal{V}_{train}$ and report attribute-level accuracy over words in $\mathcal{V}_{test}$.

Connection to Learning Object Attributes This problem has been studied before for zero-shot object recognition, but there are several key differences. Al-Halah et al. (2016) build the 'Class-Attribute Association Prediction' model (CAAP) that classifies the attributes of an object class from its name. They apply it on the Animals with Attributes dataset, a dataset containing 50 animal classes, each described by 85 attributes (Lampert et al., 2014). Importantly, these attributes are concrete details with semantically meaningful names such as "has horns" and "is furry". The CAAP model takes advantage of this, consisting of a tensor factorization model initialized by the word embeddings of the object class names as well as the attribute names. On the other hand, *verb attributes* such as the ones we outline in Section 2, are technical linguistic terms. Since word embeddings principally capture common word senses, they are unsuited for verb attributes. Thus, we evaluate two versions of CAAP as a baseline: one where the model is preinitialized with GloVe embeddings (Pennington et al., 2014) for the attribute names (CAAP-pretrained), and one where the model is learned from random initialization (CAAP-learned).

3.1 Learning from Distributed Embeddings

One way of producing attributes is from distributed word embeddings such as word2vec

(Mikolov et al., 2013). Intuitively, we expect similar verbs to have similar distributions of nearby nouns and adverbs, which can greatly help us in zero-shot prediction. In our experiments, we use 300-dimensional GloVe vectors trained on 840B tokens of web data (Pennington et al., 2014). We use logistic regression to predict each attribute, as we found that extra hidden layers did not improve performance. Thus, we let $f^{emb}(x_v) = \mathbf{w}_v$, the GloVe embedding of v, and use Equation 1 to get the distribution over labels.

We additionally experiment with retrofitted embeddings, in which embeddings are mapped in accordance with a lexical resource. Following the approach of Faruqui et al. (2015), we retrofit embeddings using WordNet (Miller, 1995), Paraphrase-DB (Ganitkevitch et al., 2013), and FrameNet (Baker et al., 1998b).

3.2 Learning from Dictionary Definitions

We additionally propose a model that learns the attribute-grounded meaning of verbs through dictionary definitions. This is similar in spirit to the task of using a dictionary to predict word embeddings (Hill et al., 2016).

BGRU encoder Our first model involves a Bidirectional Gated Recurrent Unit (BGRU) encoder (Cho et al., 2014). Let $x_{v,1:T}$ be a definition for verb v, with T tokens. To encode the input, we pass it through the GRU equation:

$$\vec{\mathbf{h}}_t = \text{GRU}(x_{v,t}, \vec{\mathbf{h}}_{t-1}). \quad (2)$$

Let $\overleftarrow{\mathbf{h}}_1$ denote the output of a GRU applied on the reversed input $x_{v,T:1}$. Then, the BGRU encoder is the concatenation $f^{bgru} = \vec{\mathbf{h}}_T \| \overleftarrow{\mathbf{h}}_1$.

Bag-of-words encoder Additionally, we try two common flavors of a Bag-of-Words model. In the standard case, we first construct a vocabulary of 5000 words by frequency on the dictionary definitions. Then, $f^{bow}(x_v)$ represents the one-hot encoding $f^{bow}(x_v)_i = [i \in x_v]$, in other words, whether word i appears in definiton x_v for verb v.

Additionally, we try out a Neural Bag-of-Words model where the word embeddings in a definition are averaged (Iyyer et al., 2015). This is $f^{nbow}(x_{v,1:T}) = \frac{1}{|T|} \sum_{t=1}^{T} f^{emb}(x_{v,t})$.

Dealing with multiple definitions per verb One potential pitfall with using dictionary definitions is that there are often many defnitions associated with each verb. This creates a dataset bias since polysemic verbs are seen more often. Additionally, dictionary definitions tend to be sorted by relevance, thus lowering the quality of the data if all definitions are weighted equally during training. To counteract this, we randomly oversample the definitions at training time so that each verb has the same number of definitions.[4] At test time, we use the first-occurring (and thus generally most relevant) definition per verb.

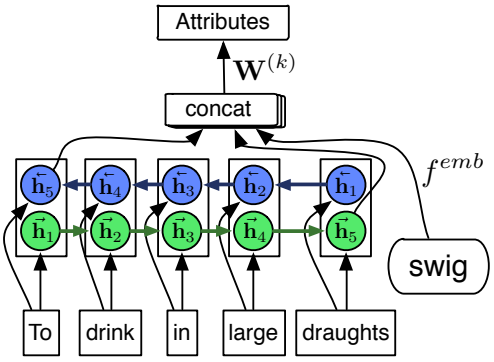

Figure 2: Overview of our combined dictionary + embedding to attribute model. Our encoding is the concatenation of a Bidirectional GRU of a definition and the word embedding for that word. The encoding is then mapped to the space of attributes using parameters $\mathbf{W}^{(k)}$.

3.3 Combining Dictionary and Embedding representations

We hypothesize that the two modalities of the dictionary and distributional embeddings are complementary. Therefore, we propose an early fusion (concatenation) of both categories. Figure 2 describes the GRU + embedding model – in other words, $f^{BGRU+emb} = f^{BGRU} \| f^{emb}$. This can likewise be done with any choice of definition encoder and word embedding.

4 Zero-Shot Activity Recognition

4.1 Verb Attributes as Latent Mappings

Given learned attributes for a collection of activities, we would like to evaluate their performance at describing these activities from real world images in a zero-shot setting. Thus, we consider several models that classify an image's label by pivoting through an attribute representation.

[4] For the (neural) bag of words models, we also tried concatenating the definitions together per verb and then doing the encoding. However, we found that this gave worse results.

Overview A formal description of the task is at follows. Let the space of labels be $\mathcal{V}$, partitioned into $\mathcal{V}_{train}$ and $\mathcal{V}_{test}$. Let $z_v \in \mathcal{Z}$ represent an image with label $v \in \mathcal{V}$; our goal is to correctly predict this label amongst verbs $v \in \mathcal{V}_{test}$ at test time, despite never seeing any images with labels in $\mathcal{V}_{test}$ during training.

Generalization will be done through a lookup table $\mathbf{A}$, with known attributes for each $v \in \mathcal{V}$. Formally, for each attribute k we define it as:

$$\mathbf{A}^{(k)}_{v',i} = \begin{cases} 1 & \text{if attribute } k \text{ for verb } v' \text{ is } i \\ -1 & \text{otherwise} \end{cases} \quad (3)$$

For binary attributes, we need only one entry per verb, making $\mathbf{A}^{(k)}$ a single column vector. Let our image encoder be represented by the map $g : \mathcal{Z} \to \mathbb{R}^d$. We then use the linear map in Equation 1 to produce the log-probability distribution over each attribute k. The distribution over labels is then:

$$P(\cdot|z_v) = \underset{v'}{\text{softmax}} \left(\sum_k \mathbf{A}^{(k)} \mathbf{W}^{(k)} g(z_v) \right) \quad (4)$$

where $\mathbf{W}^{(k)}$ is a learned parameter that maps the image encoder to the attribute representation. We then train our model by minimizing the cross-entropy loss over the training verbs $\mathcal{V}_{train}$.

Convolutional Neural Network (CNN) Encoder Our image encoder is a CNN with the Resnet-152 architecture (He et al., 2016). We use weights pretrained on ImageNet (Deng et al., 2009) and perform additional pretraining on ImSitu using the classes $\mathcal{V}_{train}$. After this, we remove the top layer and set $g(x_v)$ to be the 2048-dimensional image representation from the network.

4.1.1 Connection to other attribute models

Our model is similar to those of Akata et al. (2013) and Romera-Paredes and Torr (2015) in that we predict the attributes indirectly and train the model through the class labels.[5] It differs from several other zeroshot models, such as Lampert et al. (2014)'s Direct Attribute Prediction (DAP) model, in that DAP is trained by maximizing the probability of predicting each attribute and then multiplies the probabilities at test time. Our use of joint training allows the recognition model to directly optimize class-discrimination rather than attribute-level accuracy.

[5]Unlike these models, however, we utilize (some) categorical attributes and optimize using cross-entropy.

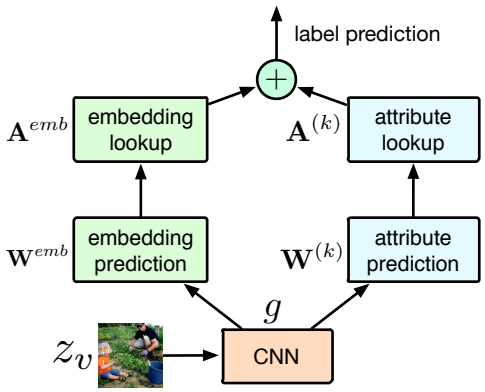

Figure 3: Our proposed model for combining attribute-based zero-shot learning and word-embedding based transfer learning. The embedding and attribute lookup layers are used to predict a distribution of labels over $\mathcal{V}_{train}$ during training and $\mathcal{V}_{test}$ during testing.

4.2 Verb Embeddings as Latent Mappings

An additional method of doing zero-shot image classification is by using word embeddings directly. Frome et al. (2013) build DeVISE, a model for zero-shot learning on ImageNet object recognition where the objective is for the image model to predict a class's word embedding directly. DeVISE is trained by minimizing

$$\sum_{v' \in \mathcal{V}_{train} \setminus \{v\}} \max\{0, .1 + (\mathbf{w}_{v'} - \mathbf{w}_v)\mathbf{W}^{emb} g(z_v)\}$$

for each image z_v. We compare against a version of this model with fixed GloVe embeddings $\mathbf{w}$.

Additionally, we employ a variant of our model using only word embeddings. The equation is the same as Equation 4, except using the matrix $\mathbf{A}^{emb}$ as a matrix of word embeddings: i.e., for each label v we consider, we have $\mathbf{A}^{emb}_v = \mathbf{w}_v$.

4.3 Joint prediction from Attributes and Embeddings

To combine the representation power of the attribute and embedding models, we build an ensemble combining both models. This is done by adding the logits before the softmax is applied:

$$\underset{v'}{\text{softmax}} \left(\sum_k \mathbf{A}^{(k)} \mathbf{W}^{(k)} g(z_v) + \mathbf{A}^{emb} \mathbf{W}^{emb} g(z_v) \right)$$

A diagram is shown in Figure 3. We find that during optimization, this model can easily overfit, presumably by excessive coadaption of the embedding and attribute components. To solve this,

we train the model to minimize the cross entropy of three sources independently: the attributes only, the embeddings only, and the sum, weighting each equally.

Incorporating predicted and gold attributes We additionally experiment with an ensemble of our model, combining predicted and gold attributes of $\mathcal{V}_{test}$. This allows the model to hedge against cases where a verb attribute might have several possible correct answers. A single model is trained; at test time, we multiply the class level probabilities $P(\cdot|z_v)$ of each together to get the final predictions.

5 Actions and Attributes Dataset

To evaluate our hypotheses on action attributes and zero-shot learning, we constructed a dataset using crowd-sourcing experiments. The *Actions and Attributes* dataset consists of annotations for 1710 verb templates, each consisting of a verb and an optional particle (e.g. "put" or "put *up*").

We selected all verbs from the ImSitu corpus, consisting of images representing verbs from many categories (Yatskar et al., 2016), then extended the set using the MPII movie visual description dataset and ScriptBase datasets, (Rohrbach et al., 2015; Gorinski and Lapata, 2015). We used the spaCy dependency parser (Honnibal and Johnson, 2015) to extract the verb template for each sentence, and collected annotations on Mechanical Turk to filter out nonliteral and abstract verbs. Turkers annotated this filtered set of templates using the attributes described in Section 2. In total, 1203 distinct verbs are included. The templates are split randomly by verb; out of 1710 total templates, we save 1313 for training, 81 for validation, and 316 for testing.

To provide signal for classifying these verbs, we collected dictionary definitions for each verb using the Wordnik API,[6] including only senses that are explicitly labeled "verb." This leaves us with 23,636 definitions, an average of 13.8 per verb.

6 Experimental Setup

BGRU pretaining We pretrain the BGRU model on the Dictionary Challenge, a collection of 800,000 word-definition pairs obtained from Wordnik and Wikipedia articles (Hill et al., 2016); the objective is to obtain a word's embedding given one of its definitions. For the BGRU model, we use an internal dimension of 300, and embed the words to a size 300 representation. The vocabulary size is set to 30,000 (including all verbs for which we have definitions). During pretraining, we keep the architecture the same, except a different 300-dimensional final layer is used to predict the GloVe embeddings.

Following Hill et al. (2016), we use a ranking loss. Let $\hat{\mathbf{w}} = \mathbf{W}^{emb} f(x)$ be the predicted word embeddings for each definition x of a word in the dictionary (not necessarily a verb). Let $\mathbf{w}$ be the word's embedding, and $\widetilde{\mathbf{w}}$ be the embedding of a random dictionary word. The loss is then given by:

$$L = \max\{0, .1 - \cos(\mathbf{w}, \hat{\mathbf{w}}) + \cos(\mathbf{w}, \widetilde{\mathbf{w}})\}$$

After pretraining the model on the Dictionary Challenge, we fine-tune the attribute weights $\mathbf{W}^{(k)}$ using the cross-entropy over Equation 1.

Zero-shot with the imSitu dataset We build our image-to-verb model on the newly introduced imSitu dataset, which contains a diverse collection of images depicting one of 504 verbs. The images represent a variety of different semantic role labels (Yatskar et al., 2016). Figure 4 shows examples from the dataset. We apply our attribute split to the dataset and are left with 379 training classes, 29 validation classes, and 96 test classes.

Zero-shot activity recognition baselines We compare against several additional baseline models for learning from attributes and embeddings. Romera-Paredes and Torr (2015) propose "Embarassingly Simple Zero-shot Learning" (ESZL), a linear model that directly predicts class labels through attributes and incorporates several types of regularization. We compare against a variant of Lampert et al. (2014)'s DAP model discussed in Section 4.1.1. We additionally compare against DeVISE (Frome et al., 2013), as mentioned in Section 4.2. We use a Resnet-152 CNN finetuned on the imSitu $\mathcal{V}_{train}$ classes as the visual features for these baselines (the same as discussed in Section 4.1).

Additional implementation details are provided in the Appendix.

[6] Available at http://developer.wordnik.com/ with access to American Heriatge Dictionary, the Century Dictionary, the GNU Collaborative International Dictionary, Wordnet, and Wiktionary.

		acc-macro	acc-micro	Body	Duration	Aspect	Motion	Social	Effect	Transi.
	most frequent class	61.33	75.45	76.84	76.58	43.67	35.13	42.41	84.97	69.73
Emb	CAAP-pretrained	64.96	78.06	84.81	72.15	50.95	46.84	43.67	85.21	71.10
Emb	CAAP-learned	68.15	81.00	86.33	76.27	52.85	52.53	**45.57**	88.29	75.21
Emb	GloVe	66.60	79.69	85.76	75.00	50.32	48.73	43.99	86.52	75.84
Emb	GloVe + framenet	67.42	80.79	86.27	76.58	49.68	50.32	44.94	88.19	75.95
Emb	GloVe + ppdb	67.52	80.75	85.89	76.58	51.27	50.95	43.99	88.21	75.74
Emb	GloVe + wordnet	68.04	81.13	86.58	76.90	54.11	50.95	43.04	**88.34**	**76.37**
Dict	BGRU	66.05	79.44	85.70	76.90	51.27	48.42	40.51	86.92	72.68
Dict	BoW	62.53	77.61	83.54	76.58	48.42	35.76	36.39	86.31	70.68
Dict	NBoW	65.41	78.96	85.00	76.58	52.85	42.41	43.35	86.87	70.78
D+E	NBoW + GloVe	67.52	80.76	**86.84**	75.63	53.48	51.90	41.77	88.03	75.00
D+E	BoW + GloVe	63.15	77.89	84.11	**77.22**	49.68	34.81	38.61	86.18	71.41
D+E	BGRU + GloVe	**68.43**	**81.18**	86.52	76.58	**56.65**	**53.48**	41.14	88.24	**76.37**

Table 1: Results on the text-to-attributes task. All values reported are accuracies (in %). For attributes where multiple labels can be selected, the accuracy is averaged over all instances (e.g., the accuracy of "Body" is given by the average of accuracies from correctly predicting Head, Torso, etc.). As such, we report two ways of averaging the results: microaveraging (where the accuracy is the average of accuracies on the underlying labels) and macroaveraging (where the accuracy is averaged together from the groups).

7 Experimental Results

7.1 Predicting Action Attributes from Text

Our results for action attribute prediction from text are given in Table 1. Several examples are given in the supplemental section in Table 3. Our results on the text-to-attributes challenge confirm that it is a challenging task for two reasons. First, there is noise associated with the attributes: many verb attributes are hard to annotate given that verb meanings can change in context.[7] Second, there is a lack of training data inherent to the problem: there are not many common verbs in English, and it can be difficult to crowdsource annotations for rare ones. Third, any system must compete with strong frequency-based baselines, as attributes are generally sparse. Moreover, we suspect that were more attributes collected (so as to cover more obscure patterns), the sparsity would only increase.

Despite this, we report strong baseline results on this problem, particularly with our embedding based models. The performance gap between embedding- and definition-based models can possibly be explained by the fact that the word embeddings are trained on a very large corpus of real-world *examples* of the verb, while the definition is only a single high-level representation meant to be understood by someone who already speaks that language. For instance, it is likely difficult for the definition-based model to infer whether a verb is transitive or not (Transi.), since definitions might assume commonsense knowledge about the underlying concepts the verb represents. The strong performance of embedding models is further enhanced by using retrofitted word embeddings, suggesting an avenue for improvement on language grounding through better representation of linguistic corpora.

We additionally see that both joint dictionary-embedding models outperform the dictionary-only models overall. In particular, the BGRU+GloVe model performs especially well at determining the aspect and motion attributes of verbs, particularly relative to the baseline. The strong performance of the BGRU+GloVe model indicates that there is some signal that is missing from the distributional embeddings that can be recovered from the dictionary definition. We thus use the predictions of this model for zero-shot image recognition.

Based on error analysis, we found that one common mode of failure is where contextual knowledge is required. To give an example, the embedding based model labels "shop" as a likely solitary action. This is possibly because there are a lack of similar verbs in $\mathcal{V}_{train}$; by random chance, "buy" is also in the test set. We see that this can be partially mitigated by the dictionary, as evidenced by the fact that the dictionary-based models label "shop" as in between social and solitary. Still, it is a difficult task to infer that people like to "visit stores in search of merchandise" together.

7.2 Zero-shot Action Recognition

Our results for verb prediction from images are given in Table 2. Despite the difficulty of predicting the correct label over 96 unseen choices,

[7] As such, our attributes have a median Krippendorff Alpha of $\alpha = .359$.

Model	Attributes used atts(P)	atts(G)	GloVe	$v \in \mathcal{V}_{test}$ top-1	top-5
Random				1.04	5.20
DeVISE			✓	16.50	37.56
ESZL		✓		3.60	14.81
	✓			3.27	13.21
DAP		✓		3.35	16.69
	✓			4.33	17.56
Ours		✓		4.79	19.98
	✓			7.04	22.19
	✓	✓		7.71	24.90
			✓	17.60	39.29
		✓	✓	18.10	41.46
	✓		✓	16.75	40.44
	✓	✓	✓	**18.15**	**42.17**

Table 2: Results on the image-to-verb task. atts(P) refers to attributes predicted from the BGRU+GloVe model described in Section 3, atts(G) to gold attributes, and GloVe to GloVe vectors. The accuracies reported are amongst the 96 unseen labels of $\mathcal{V}_{test}$.

our models show predictive power. Although our attribute models do not outperform our embedding models and DeVISE alone, we note that our joint attribute and embedding model scores the best overall, reaching 18.10% in top-1 and 41.46% in top-5 accuracy when using gold attribute annotations for the zero-shot verbs. This result is possibly surprising given the small number of attributes ($K = 24$) in total, of which most tend to be sparse (as can be seen from the baseline performance in Table 1). We thus hypothesize that collecting more activity attributes would further improve performance.

We also note the success in performing zero-shot learning with predicted attributes. Perhaps paradoxically, our attribute-only models (along with DAP) perform better in both accuracy metrics when given predicted attributes at test time, as opposed to gold attributes. Further, we get an extra boost by ensembling predictions of our model when given two sets of attributes at test time, giving us the best results overall at 18.15% top-1 accuracy and 42.17% top-5. Interestingly, better performance with predicted attributes is also reported by Al-Halah et al. (2016): predicting the attributes with their CAAP model and then running the DAP model on these predicted attributes outperforms the use of gold attributes at test time. It is somewhat unclear why this is the case - possibly, there is some bias in the attribute labeling, which the attribute predictor can correct for.

In addition to quantitative results, we show some zero-shot examples in Figure 4. The examples show inherent difficulty of zero-shot action recognition. Incorrect predictions are often reasonably related to the situation ("rub" vs "dye") but picking the correct target verb based on attribute-based inference is still a challenging task.

Although our results appear promising, we argue that our model still fails to represent much of the semantic information about each image class. In particular, our model is prone to *hubness*: the overprediction of a limited set of labels at test time: those that closely match signatures of examples in the training set. This problem has previously been observed with the use of word embeddings for zero-shot learning (Marco and Georgiana, 2015) and can be seen in our examples (for instance, the over-prediction of "buy"). Unfortunately, we were unable to mitigate this problem in a way that also led to better quantitative results (for instance, by using a ranking loss as in DeVISE (Frome et al., 2013)). We thus leave resolving the hubness problem in zero-shot activity recognition as a question for future work.

8 Related Work

Learning attributes from embeddings Rubinstein et al. (2015) seek to predict McRae et al. (2005)'s feature norms from word embeddings of concrete nouns. Likewise, the CAAP model of Al-Halah et al. (2016) predicts the object attributes of concrete nouns for use in zero-shot learning. In contrast, we predict verb attributes. A related task is that of improving word embeddings using multimodal data and linguistic resources (Faruqui et al., 2015; Silberer et al., 2013; Vendrov et al., 2016). Our work runs orthogonal to this, as we focus on word attributes as a tool for a zero-shot activity recognition pipeline.

Zero-shot learning with objects Though distinct, our work is related to zero-shot learning of objects in computer vision. There are several datasets (Nilsback and Zisserman, 2008; Welinder et al., 2010) and models developed on this task (Romera-Paredes and Torr (2015); Lampert et al. (2014); Mukherjee and Hospedales (2016); Farhadi et al. (2010)). In addition,

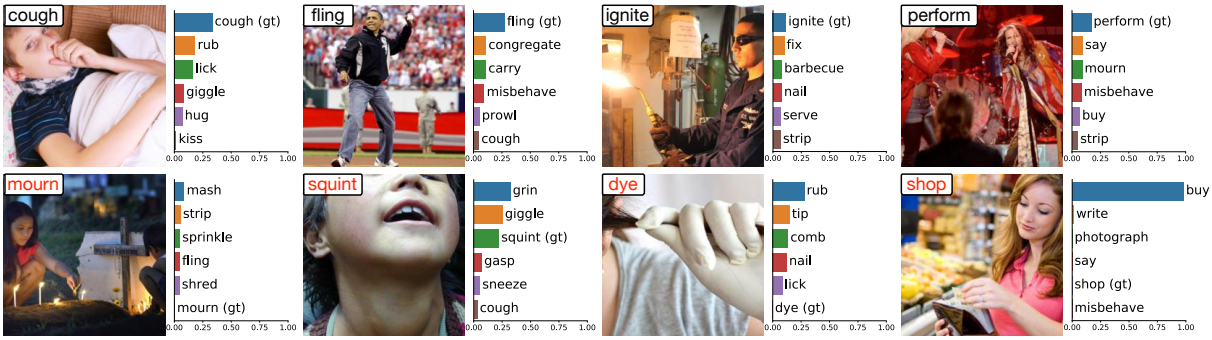

Figure 4: Predictions on unseen classes from our attribute+embedding model with gold attributes. The top and bottom rows show successful and failure cases respectively. The bars to the right of each image represent a probability distribution, showing the ground truth class and the top 5 scoring incorrect classes.

Ba et al. (2015) augment existing datasets with descriptive Wikipedia articles so as to learn novel objects from descriptive text. As illustrated in Section 1, action attributes pose unique challenges compared to object attributes, thus models developed for zero-shot object recognition are not as effective for zero-shot action recognition, as has been empirically shown in Section 7.

Zero-shot activity recognition In prior work, zero-shot activity recognition has been studied on video datasets, each containing a selection of concrete physical actions. The MIXED action dataset, itself a combination of three action recognition datasets, has 2910 labeled videos with 21 actions, each described by 34 action attributes (Liu et al., 2011). These action attributes are concrete binary attributes corresponding to low-level physical movements, for instance, "arm only motion," "leg: up-forward motion." By using word embeddings instead of attributes, Xu et al. (2017) study video activity recognition on a variety of action datasets, albeit in the transductive setting wherein access to the test labels is provided during training. In comparison with our work on imSitu (Yatskar et al., 2016), these video datasets lack broad coverage of verb-level classes (and for some, sufficient data points per class).

The abstractness of broad-coverage activity labels makes the problem much more difficult to study with attributes. To get around this, Antol et al. (2014) present a synthetic dataset of cartoon characters performing dyadic actions, and use these cartoon illustrations as internal mappings for zero-shot recognition of dyadic actions in real-world images. We investigate an alternative approach by using linguistically informed verb attributes for activity recognition.

9 Future work / Conclusion

Several possibilities remain open for future work. First, more attributes could be collected and evaluated, possibly integrating other linguistic theories about verbs, with more accurate modeling of context. Second, while our experiments use attributes as a pivot between language and vision domains, the effects of this could be explored more in future work. In particular, since our experiments show that unsupervised word embeddings significantly help performance, it might be desirable to learn data-driven attributes in an end-to-end fashion directly from a large corpus or from dictionary definitions. Third, future research on action attributes should ideally include videos to better capture attributes that require temporal signals.

Overall, however, our work presents a strong early step towards zero-shot activity recognition, a relatively less studied task that poses several unique challenges over zero-shot object recognition. We introduce new action attributes motivated by linguistic theories and demonstrate their empirical use for reasoning about previously unseen activities.

Acknowledgements We thank the anonymous reviewers, Mark Yatskar, Luke Zettlemoyer, and Yonatan Bisk, for their helpful feedback. We also thank the Mechanical Turk workers and members of the XLab, who helped with the annotation process. This work is supported by the National Science Foundataion Graduate Research Fellowship (DGE-1256082), the NSF grant (IIS-1524371), DARPA CwC program through ARO (W911NF-15-1-0543), and gifts by Google and Facebook.

References

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2013. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826.

Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen. 2016. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5975–5984.

Stanislaw Antol, C. Lawrence Zitnick, and Devi Parikh. 2014. Zero-Shot Learning via Visual Abstraction. In *ECCV*.

Jimmy Ba, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998a. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference*, pages 86–90, Montreal, Canada.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998b. 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*, pages 86–90. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

William Croft. 2012. *Verbs: Aspect and causal structure*. OUP Oxford. Pg. 16.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Ali Farhadi, Ian Endres, and Derek Hoiem. 2010. Attribute-centric recognition for cross-category generalization. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2352–2359. IEEE.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. Association for Computational Linguistics.

Annemarie Friedrich and Alexis Palmer. 2014. Automatic prediction of aspectual class of verbs in context. In *ACL (2)*, pages 517–523.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marco Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129. Curran Associates, Inc.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *HLT-NAACL*, pages 758–764.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076, Denver, Colorado. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Felix Hill, KyungHyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to Understand Phrases by Embedding the Dictionary. *Transactions of the Association for Computational Linguistics*, 4(0):17–30.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Mohit Iyyer, Varun Manjunatha, Jordan L Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *ACL (1)*, pages 1681–1691.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465.

Beth Levin. 1993. *English verb classes and alternations : a preliminary investigation*.

Jingen Liu, Benjamin Kuipers, and Silvio Savarese. 2011. Recognizing human actions by attributes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3337–3344. IEEE.

Angeliki Lazaridou Marco, Baroni and Dinu Georgiana. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. ACL.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behav Res Methods*, 37(4):547–559.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Tanmoy Mukherjee and Timothy Hospedales. 2016. Gaussian visual-linguistic embedding for zero-shot recognition. EMNLP.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on*, pages 722–729. IEEE.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A Dataset for Movie Description. *arXiv:1501.02530 [cs]*. ArXiv: 1501.02530.

Bernardino Romera-Paredes and Philip HS Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161.

Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Eric V Siegel and Kathleen R McKeown. 2000. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–628.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of semantic representation with visual attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–582, Sofia, Bulgaria. Association for Computational Linguistics.

Zeno Vendler. 1957. Verbs and Times. *The Philosophical Review*, 66(2):143–160.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *ICLR*.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. Caltech-ucsd birds 200.

Xun Xu, Timothy Hospedales, and Shaogang Gong. 2017. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, pages 1–25.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542.

A Supplemental

Implementation details

Our CNN and BGRU models are built in PyTorch[8]. All of our one-layer neural network models are built in Scikit-learn (Pedregosa et al., 2011) using the provided LogisticRegression class (using one-versus-rest if appropriate). Our neural models use the Adam optimizer (Kingma and Ba, 2014), though we weak the default hyperparameters somewhat.

Recall that our dictionary definition model is a bidirectional GRU with a hidden size of 300, with a vocabulary size of 30,000. After pretraining on the Dictionary Challenge, we freeze the word embeddings and apply a dropout rate of 50% before the final hidden layer. We found that such an aggressive dropout rate was necessary due to the small size of the training set. During pretraining, we used a learning rate of 10^{-4}, a batch size of 64, and set the Adam parameter ϵ to the default $1e^{-8}$. During finetuning, we set $\epsilon = 1.0$ and the batch size to 32. In general, we found that setting too low of an ϵ during finetuning caused our zero-shot models to update parameters too aggressively during the first couple of updates, leading to poor results.

For our CNN models, we pretrained the Resnet 152 (initialized with imagenet weights) on the

[8] pytorch.org

training classes of the imSitu dataset, using a learning rate of 10^{-4} and $\epsilon = 10^{-8}$. During fine-tuning, we dropped the learning rate to 10^{-5} and set $\epsilon = 10^{-1}$. We also froze all parameters except for the final resnet block, and the linear attribute and embedding weights. We also found L2 regularization quite important in reducing overfitting, and we applied regularization at a weight of 10^{-4} to all trainable parameters.

Full list of attributes

The following is a full list of the attributes. In addition to the attributes presented here, we also crowdsourced attributes for the emotion content of each verb (e.g., happiness, sadness, anger, and surprise). However, we found these annotations to be skewed towards "no emotion", since most verbs do not strongly associate with a specific emotion. Thus, we omit them in our experiments.

(1) Aspectual Classes: one attribute with 5 values:
   (a) State
   (b) Achievement
   (c) Accomplishment
   (d) Activity
   (e) Unclear without context

(2) Temporal Duration: one attribute with 5 values:
   (a) Atemporal
   (b) On the order of seconds
   (c) On the order of hours
   (d) On the order of days

(3) Motion Dynamics: One attribute with 5 values:
   (a) unclear without context
   (b) No motion
   (c) Low motion
   (d) Medium motion
   (e) High motion

(4) Social Dynamics: One attribute with 5 values:
   (a) solitary
   (b) likely solitary
   (c) solitary or social
   (d) likely social
   (e) social

(5) Transitivity: Three binary attributes:
   (a) Intransitive: 1 if the verb can be used intransitively, 0 otherwise
   (b) Transitive (person): 1 if the verb can be used in the form "<someone>", 0 otherwise
   (c) Transitive (object): 1 if the verb can be used in the form "<verb> something", 0 otherwise

(6) Effects on Arguments: 12 binary attributes
   (a) Intransitive 1: 1 if the verb is intransitive and the subject moves somewhere
   (b) Intransitive 2: 1 if the verb is intransitive and the external world changes
   (c) Intransitive 3: 1 if the verb is intransitive, and the subject's state changes
   (d) Intransitive 4: 1 if the verb is intransitive, and nothing changes
   (e) Transitive (obj) 1: 1 if the verb is transitive for objects and the object moves somewhere
   (f) Transitive (obj) 2: 1 if the verb is transitive for objects and the external world changes
   (g) Transitive (obj) 3: 1 if the verb is transitive for objects and the object's state changes
   (h) Transitive (obj) 4: 1 if the verb is transitive for objects and nothing changes
   (i) Transitive (person) 1: 1 if the verb is transitive for people and the object is a person that moves somewhere
   (j) 'Transitive (person) 2: 1 if the verb is transitive for people and the external world changes
   (k) Transitive (person) 3: 1 if the verb is transitive for people and if the object is a person whose state changes
   (l) Transitive (person) 4: 1 if the verb is transitive for people and nothing changes

(7) Body Involements: 5 binary attributes
   (a) Arms: 1 if arms are used
   (b) Head: 1 if head is used
   (c) Legs: 1 if legs are used
   (d) Torso: 1 if torso is used
   (e) Other: 1 if another body part is used

	Model	Definition	Social	Aspect	Energy	Time	Body part
shop	GT	To visit stores in search of merchandise or bargains	likely social	accomplish.	high	hours	arms,head
	embed		likely solitary	activity	medium	minutes	
	BGRU		solitary or social	activity	medium	minutes	
	BGRU+		solitary or social	activity	medium	minutes	
mash	GT	To convert malt or grain into mash	likely solitary	activity	high	seconds	arms
	embed		likely solitary	activity	medium	seconds	arms
	BGRU		solitary or social	achievement	medium	seconds	arms
	BGRU+		likely solitary	activity	high	seconds	arms
photograph	GT	To take a photograph of	solitary or social	achievement	low	seconds	arms,head
	embed		solitary or social	accomplish.	medium	minutes	arms
	BGRU		solitary or social	achievement	medium	seconds	arms
	BGRU+		solitary or social	unclear	low	seconds	arms
spew out	GT	eject or send out in large quantities also metaphorical	solitary or social	achievement	high	seconds	head
	embed		likely solitary	achievement	medium	seconds	
	BGRU		solitary or social	achievement	high	seconds	arms
	BGRU+		likely solitary	achievement	medium	seconds	
tear	GT	To pull apart or into pieces by force rend	likely solitary	achievement	low	seconds	arms
	embed		solitary or social	achievement	medium	seconds	arms
	BGRU		solitary or social	achievement	high	seconds	arms
	BGRU+		solitary or social	achievement	high	seconds	arms
squint	GT	To look with the eyes partly closed as in bright sunlight	likely solitary	achievement	low	seconds	head
	embed		likely solitary	achievement	low	seconds	head
	BGRU		likely solitary	achievement	low	seconds	head
	BGRU+		likely solitary	achievement	low	seconds	head
shake	GT	To cause to move to and fro with jerky movements	solitary or social	activity	medium	seconds	
	embed		likely solitary	achievement	medium	seconds	arms
	BGRU		likely solitary	activity	medium	seconds	
	BGRU+		likely solitary	activity	medium	seconds	
doze	GT	To sleep lightly and intermittently	likely solitary	state	none	minutes	head
	embed		likely solitary	achievement	medium	seconds	
	BGRU		likely solitary	achievement	low	seconds	
	BGRU+		likely solitary	activity	low	seconds	
writhe	GT	To twist as in pain struggle or embarrassment	solitary or social	activity	high	seconds	arms,torso
	embed		likely solitary	activity	medium	seconds	
	BGRU		likely solitary	activity	medium	seconds	arms
	BGRU+		likely solitary	activity	medium	seconds	

Table 3: Example sentences and predicted attributes. Due to space constraints, we only list a few representative attributes and verbs.