Inferring the Why in Images

Hamed Pirsiavash[*] **Carl Vondrick**[*] **Antonio Torralba**
Massachusetts Institute of Technology
{hpirsiav,vondrick,torralba}@mit.edu

Abstract

Humans have the remarkable capability to infer the motivations of other people's actions, likely due to cognitive skills known in psychophysics as the theory of mind. In this paper, we strive to build a computational model that predicts the motivation behind the actions of people from images. To our knowledge, this challenging problem has not yet been extensively explored in computer vision. We present a novel learning based framework that uses high-level visual recognition to infer why people are performing an actions in images. However, the information in an image alone may not be sufficient to automatically solve this task. Since humans can rely on their own experiences to infer motivation, we propose to give computer vision systems access to some of these experiences by using recently developed natural language models to mine knowledge stored in massive amounts of text. While we are still far away from automatically inferring motivation, our results suggest that transferring knowledge from language into vision can help machines understand why a person might be performing an action in an image.

1 Introduction

When we look at the scene in Fig.1a, we can accurately recognize many evident visual concepts, such as the man sitting on a sofa in a living room. But, our ability to reason extends beyond basic recognition. Although we have never seen this man outside of a single photograph, we can also confidently explain *why* he is sitting (because he wants to watch television).

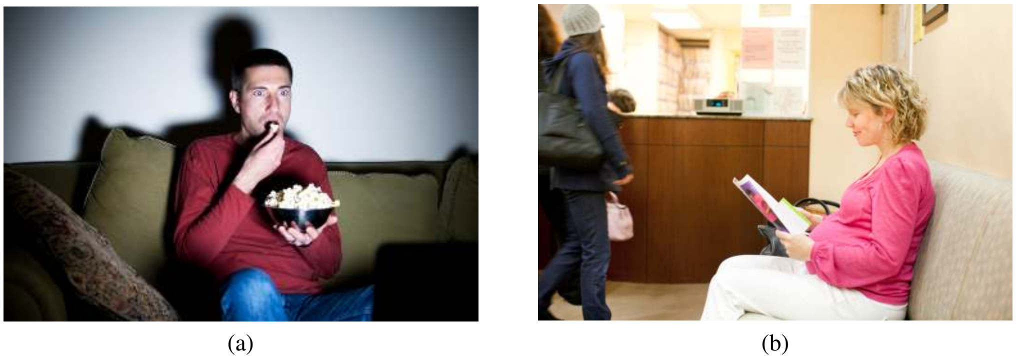

(a)
(b)

Figure 1: **Why are the people sitting on sofas?** Although we have never met them before, we can infer that the man on the left is sitting because he wants to watch television while the woman on the right intends to see the doctor. In this paper, we introduce a framework that automatically infers why people are performing actions in images by learning from visual data and written language.

[*] denotes equal contribution

Humans may be able to make such remarkable inferences partially due to cognitive skills known as the theory of mind [34]. Psychophysics researchers hypothesize that our capacity to reliably infer another person's motivation stems from our ability to impute our own beliefs to others [2, 30] and there may even be regions of the brain dedicated to this task [29]. If we ourselves were sitting on a sofa in a living room holding popcorn, then we would do this likely *because* we wanted to watch television. The theory of mind posits that, since we would want to watch television, we assume others in similar situations would also want the same.

In this paper, we seek to computationally deduce the motivation behind people's actions in images. To our knowledge, inferring why a person is performing an action from images in the wild has not yet been extensively explored in computer vision. This task is, unsurprisingly, challenging because it is unclear how to operationalize the reasoning behind the theory of mind in a machine. Moreover, people's motivations can often be outside of the visible image, either spatially as in Fig.1a or temporally as in Fig.1b.

We present a framework that takes the first strides towards automatically inferring people's motivations. Capitalizing on the theory of mind, we are able to instruct a crowd of workers to annotate why people are likely undertaking actions in photographs. We then combine these labels with state-of-the-art image features [18, 12] to train data-driven classifiers that predict a person's motivation from images. However, mid-level visual features alone may not be sufficient to automatically solve this task. Humans are able to rely on a lifetime of experiences: the reason we expect the man in Fig.1a to want to watch television is because we have experienced the same situation ourselves.

We propose to give computer vision systems access to many of the human experiences by mining the knowledge stored in massive amounts of text. Using state-of-the-art language models [10] estimated on billions of webpages [4], we are able to acquire common knowledge about people's experiences, such as their interactions with objects, their environments, and their motivations. We model these signals from written language in concert with computer vision using a framework that, once trained on data from the crowd, deduces people's motivation in an image. While we are still a long way from incorporating theory of mind into a computer system, our experiments indicate that we are able to automatically predict motivations with some promising success. By transferring knowledge acquired from text into computer vision, our results suggest that we can predict why a person is engaging in an action better than a simple vision only approach.

This paper makes two principal contributions. First, we introduce the novel problem of inferring the motivations behind people's actions to the computer vision community. Since humans are able to reliably perform this task, we believe this is an interesting problem to work on and we will publicly release a new dataset to facilitate further research. Second, we propose to use common knowledge mined from the web to improve computer vision systems. Our results suggest that this knowledge transfer is beneficial for predicting human motivation. The remainder of this paper describes these contributions in detail. Section 2 introduces our model based on a factor graph with vision and written language potentials. Section 3 conducts several experiments designed to evaluate the performance of our framework. Finally, section 4 offers concluding remarks.

1.1 Related Work

Motivation in Vision: Perhaps the most related to our paper is work that predicts the persuasive motivation of the photographer who captured an image [14]. However, our paper is different because we seek to infer the motivation of the person *inside* the image, and not the motivation of the photographer.

Action Prediction: There have been several works in robotics that predicts a person's imminent next action from a sequence of images [31, 23, 15, 8, 17]. In contrast, we wish to deduce the motivation of actions in a single image, and not what will happen next. There also has been work in forecasting activities [16, 32], inferring goals [35], and early event detection [11], but they are interested in predicting the future in videos while we wish to explain the motivations of actions of people in images in the wild. As shown on Fig.1, the motivation can be outside the image either in space or time. We believe insights into motivation can help further progress in action prediction.

Action Recognition: There is a large body of work studying how to recognize actions in images. We refer readers to excellent surveys [27, 1] for a full review. Our problem is related since in some cases the motivation can be seen as a high-level action. However, we are interested in understanding the motivation of the person engaging in an action rather than the recognizing the action itself. Our work complements action recognition because we seek to infer *why* a person is performing an action.

Common Knowledge: There are promising efforts in progress to acquire common sense for use in computer vision tasks [36, 5, 7]. In this paper, we also seek to put common knowledge into computer vision, but we instead attempt to extract it from written language.

Language in Vision: The community has been incorporating natural language into computer vision over the last decade to great success, from generating sentences from images [19], producing visual models from sentences [37, 33], and aiding in contextual models [26, 21] to name a few. In our work, we seek to mine language models trained on a massive text corpus to extract common knowledge and use it to assist computer vision systems.

2 Inferring Motivations

In this section, we present our learning framework to predict why people perform actions. We begin by discussing our dataset that we use for training. Then, we describe a vision-only approach that estimates motivation from mid-level image features. Finally, we introduce our main approach that combines knowledge from written language with visual recognition to infer motivations.

2.1 Dataset

On the surface, it may seem difficult to collect training data for this task because people's motivations are private and not directly observable. However, we are inspired by the observation that humans have the remarkable cognitive ability to think about other people's thinking [2, 30]. We leverage this capability to instruct crowdsourced workers to examine photographs of people and predict their motivations, which we can use as training data. We found that workers were consistent with each other on many images, suggesting that these labels may provide some structure that allows us to learn to predict motivation.

We assembled a dataset of images in the wild so that we could train our approach. Using the images from PASCAL VOC 2012 [9] containing a person, we instructed workers on Amazon Mechanical Turk to annotate each person with their action, the object with which they are interacting, the scene, and their best prediction of the motivation. We asked workers to only enter verbs for the motivation. To ensure quality, we repeated the annotation process five times with a disjoint set of workers and kept the annotations where workers agreed. After merging similar words using WordNet [24], workers annotated a total of 79 unique motivations, 7 actions, 43 objects, and 112 scenes on 792 images. We plan to release this dataset publicly to facilitate further research.

2.2 Vision Only Model

Given an RGB image x, a simple method can try to infer the motivation behind the person's action from only mid-level image features. Let $y \in \{1 \dots M\}$ represent a possible motivation for the

Relationship	Query to Language Model
action + object + motivation	action the object in order to motivation action the object to motivation action the object because pronoun wants to motivation
action + object + scene	action the object in a scene in a scene, action the object
action + scene + motivation	action in a scene in order to motivation action in order to motivation in a scene action because pronoun wants to motivation in a scene

Table 1: We show some examples of the third-order queries we make to the language model. We combinatorially replaced tokens with words from our vocabulary to score the relationships between concepts. The second-order queries (not shown) follow similar templates.

person. We use a linear model to predict the most likely motivation:

$$\underset{y \in \{1, \ldots, M\}}{\text{argmax}} \quad w_y^T \phi(x) \tag{1}$$

where $w_y \in \mathbb{R}^D$ is a classifier that predicts the motivation y from image features $\phi(x) \in \mathbb{R}^D$. We can estimate w_y by training an M-way linear classifier on annotated motivations. In our experiments, we use this model as a baseline.

2.3 Incorporating Common Knowledge

While we found modest success with the vision only model, it lacks the common knowledge from human experiences that makes people reliable at inferring motivation. In this section, we strive to give computers access to some of this knowledge by mining written language.

Parameterization: In order to incorporate high level information, let $y_i \in \{1 \ldots M_i\}$ be a type of visual concept, such as objects or scenes, for $i \in \{1 \ldots N\}$. We assign each visual concept y_i to one of the M_i vocabulary terms from our dataset. Our formulation is general to the types of visual concepts, but for simplicity we focus on a few: we assume that y_1 is the motivation, y_2 is the action, y_3 is an object, and y_4 is the scene.

Language Potentials: We captialize on state-of-the-art natural language models to score the relationships between concepts. We calculate the log-probability $L_{ij}(y_i, y_j)$ that the visual concepts y_i and y_j are related by querying a language model with sentences about those concepts. Tab.1 shows some of the sentence templates we used as queries. In our experiments, we query a 5-gram language model estimated on billions of web-pages [4, 10] to form each $L(\cdot)$.

Scoring Function: Given the image x, we score a possible labeling configuration y of concepts with the model:

$$\begin{aligned} \Omega(y; w, u, x, L) = &\sum_i^N w_{y_i}^T \phi_i(x) \\ &+ \sum_i^N u_i L_i(y_i) + \sum_{i<j}^N u_{ij} L_{ij}(y_i, y_j) + \sum_{i<j<k}^N u_{ijk} L_{ijk}(y_i, y_j, y_k) \end{aligned} \tag{2}$$

where $w_{y_i} \in \mathbb{R}^{D_i}$ is the unary term for the concept y_i under visual features $\phi_i(\cdot)$, and $L(y_i, y_j, y_k)$ are potentials that scores the relationship between the visual concepts y_i, y_j, and y_k. The terms $u_{ijk} \in \mathbb{R}$ calibrate these potentials with the visual classifiers. Our model forms a third order factor graph, which we visualize in Fig.2.

Note that, although ideally the unary and binary potentials would be redundant with the trinary language potentials, we found including the binary potentials and learning a weight u for each improved results. We believe this is the case because the binary language model potentials are not true marginals of the trinary potentials as they are built by a limited number of queries. Moreover, by learning extra weights, we increase the flexibility of our model, so we can weakly adapt the language model to our training data.

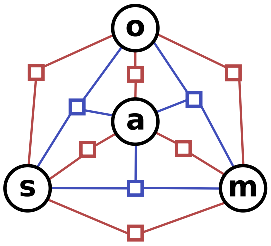


Figure 2: **Factor Graph Relating Concepts:** We visualize the factor graph for our model. Note the unary factors are not shown for simplicity. a refers to action, o for object, s for scene, and m for motivation. The binary potentials are red and the trinaries are blue for visualization purposes. We omitted the scene-object-motivation factor because it was combinatorially too large.

2.4 Inference

Predicting motivation then corresponds to calculating the most likely configuration y given an image x and learned parameters w and u over the factor graph:

$$y^* = \underset{y}{\text{argmax}}\ \Omega(y; w, u, x, L) \tag{3}$$

For both learning and evaluation, we require the K-best solutions, which can be done efficiently with approximate approaches such as K-best MAP estimation [3, 20] or sampling techniques [28, 25]. However, we found that, in our experiments, it was tractable to evaluate all configurations with a simple matrix multiplication, which gave us the exact K-best solutions in less than a second on a desktop computer.

2.5 Learning

We wish to learn the parameters w for the visual features and u for the language potentials using training data of images and their corresponding labels, $\{x^n, y^n\}$. Since our scoring function in Eqn.2 is linear on the model parameters $\theta = [w; u]$, we can write the scoring function in the linear form $\Omega(y; w, u, x, L) = \theta^T \psi(y, x)$. We want to learn θ such that the labels matching the ground truth score higher than incorrect labels. We adopt a max-margin structured prediction framework:

$$\begin{aligned} &\underset{\theta, \xi^n \geq 0}{\text{argmin}}\ \frac{1}{2} ||\theta||^2 + C \sum_n \xi^n \\ &\text{s.t.} \quad \theta^T \psi(y^n, x^n) - \theta^T \psi(h, x^n) \geq \Delta(y^n, h) - \xi^n \quad \forall_n, \forall_h \end{aligned} \tag{4}$$

The linear constraints state that the score for the correct label y^n should be larger than that of any other hypothesized label h^n by at least $\Delta(y^n, h^n)$. We use a standard 0-1 loss function for $\Delta(\cdot, \cdot)$ that incurs a penalty if any of the concepts do not match the ground truth. This optimization is equivalent to a structured SVM and can be solved by efficient off-the-shelf solvers [13]. In training, we iterate on the examples and alternate on (1) applying the model to collect the most violating constraints, and (2) updating the model by solving the QP problem in Eqn.4. The constraints are found by inferring K-best solutions of Eqn.2.

3 Experiments

In this section, we evaluate our framework's performance at inferring motivations against a vision-only baseline. We first describe our evaluation setup, then we present our results.

3.1 Experimental Setup

We designed our experiments to evaluate how well we can predict the motivation of people from our dataset. We assumed the person-of-interest is specified since the focus of this work is not person

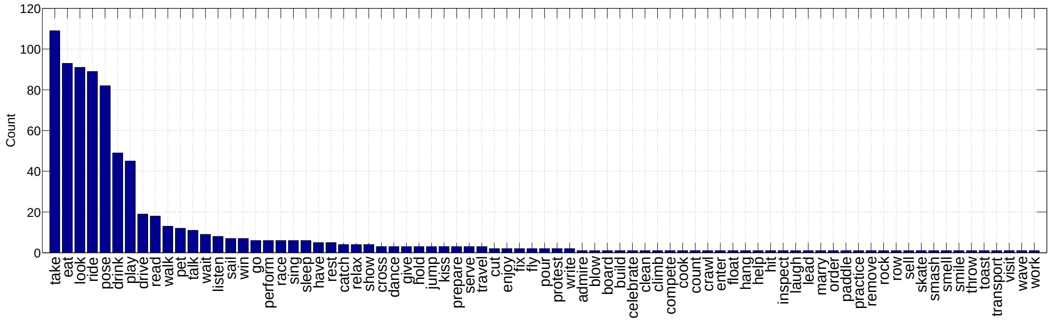


Figure 3: **Statistics of Motivations Dataset:** We show a histogram of frequencies of the motivations in our dataset. There are 79 different motivations with a long tail distribution making their prediction challenging.

		Baseline (Vision Only)	Our Method (With Language)
Given Ideal Detectors for:	Action+Object+Scene	13	10
	Action+Object	12	11
	Object+Scene	15	12
	Action+Scene	19	13
	Object	19	13
	Action	18	15
	Scene[1]	37	18
Fully Automatic		23[2]	**15**

Table 2: **Evaluation of Median Rank:** We compare our approach to baselines with the median rank of the correct motivation. Lower is better with 1 being perfect. Chance is 39. Since the distribution of motivations is non-uniform, we normalize the rank so that all categories are weighted equally. We show results when different combinations of visual concepts are given to reveal room for future improvement.

detection. Fig.3 shows a histogram of the frequency of motivations from our dataset. We split the images of our dataset into equal training and testing partitions. We computed features from the second to last layer in the AlexNet convolutional neural network trained on ImageNet [18, 12] due to their state-of-the-art performance on other visual recognition problems. Due to memory constraints, we reduced the dimensionality of the features to the top 100 principal components. We trained both our model and the baselines using cross validation to estimate hyperparameters, and we report results on the held-out test set. Since to our knowledge we are the first to address the problem of inferring motivation in computer vision, we compare against a vision-only baseline and chance.

3.2 Quantitative Evaluation

We evaluate our approach on an image by finding the rank of a ground truth motivation in the max-marginals on all states for the motivation concept y_1. This is equivalent to the rank of ground truth motivation in the list of motivation categories, sorted by their best score among all possible configurations. We show the median rank of our algorithm and the baseline across our dataset in Tab.2. Our results in the last row of the table suggest that incorporating knowledge from written language can improve accuracy over a naive vision-only approach. Moreover, our approach is significantly better than chance, suggesting that our learning algorithm is able to capitalize on structure in the data.

For diagnostic purposes, the top of Tab.2 shows the performance of our approach versus the baseline if we had ideal recognition systems for each visual concept. In order to give the vision-only baseline access to other visual concepts, we concatenate its features with a ground truth object bank [22]. Our results suggest that if we had perfect vision systems for actions, objects, and scenes, then our model would moderately improve, but it would still not solve the problem. In order to improve performance further, we hypothesize integrating additional visual cues such as human gaze and clothing will yield better results, motivating work in high-level visual recognition.

To demonstrate the importance of the trinary language model potentials, we trained our model with only binary and unary language potentials. The model without trinary potentials obtained a degraded median rank of 18, suggesting that the trinary potentials are able to capture beneficial knowledge in written language.

We compare the accuracy of our approach versus the number of top retrieved motivations in Fig.4 where we consider an image correct if our model predicts the ground truth motivation within the set of top retrievals. Interestingly, when the top number of retrievals is small, our fully automatic method with imperfect vision (solid red curve) only slightly trails the baseline with ideal detectors (dashed blue curve), suggesting that language models may have a strong effect when combined with

[2] Note that given ideal scene classifiers, we obtain worse performance than the automatic approach. We believe this is the case because our model overfits to the scene.

[2] While the rest of this baseline uses Crammer and Singer's multiclass SVM [6], we found a one-vs-rest strategy worked better for the fully automatic baseline (median rank 30 for Crammer and Singer, and median rank 23 for one-vs-rest). We report the better baseline in the table for fair comparison.

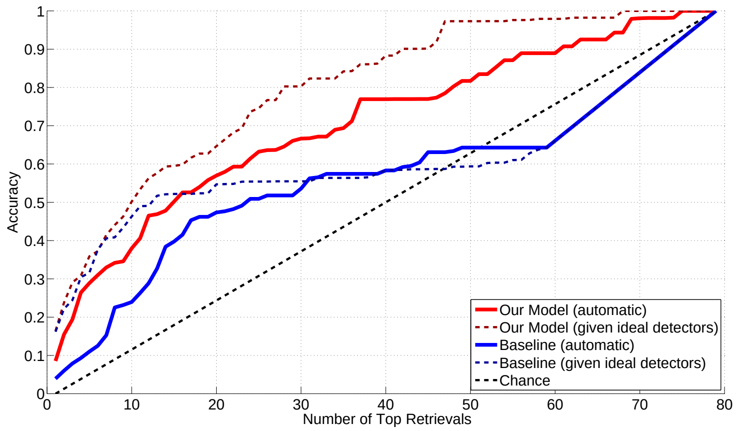


Figure 4: **Accuracy vs. Number of Retrievals:** We plot the number of retrieved motivations versus the accuracy of retrieving the correct motivation. Higher is better. Our approach does better than the baseline in all cases. Notice how the baseline flattens at 50 retrievals. This happens due to the long tail distribution of our dataset: the baseline struggles to identify motivations that do not have large amounts of training data. Our approach appears to use language to obtain reasonable performance even in the tail.

current vision methods. We partially attribute this gain due to the common knowledge available in written language.

3.3 Qualitative Results

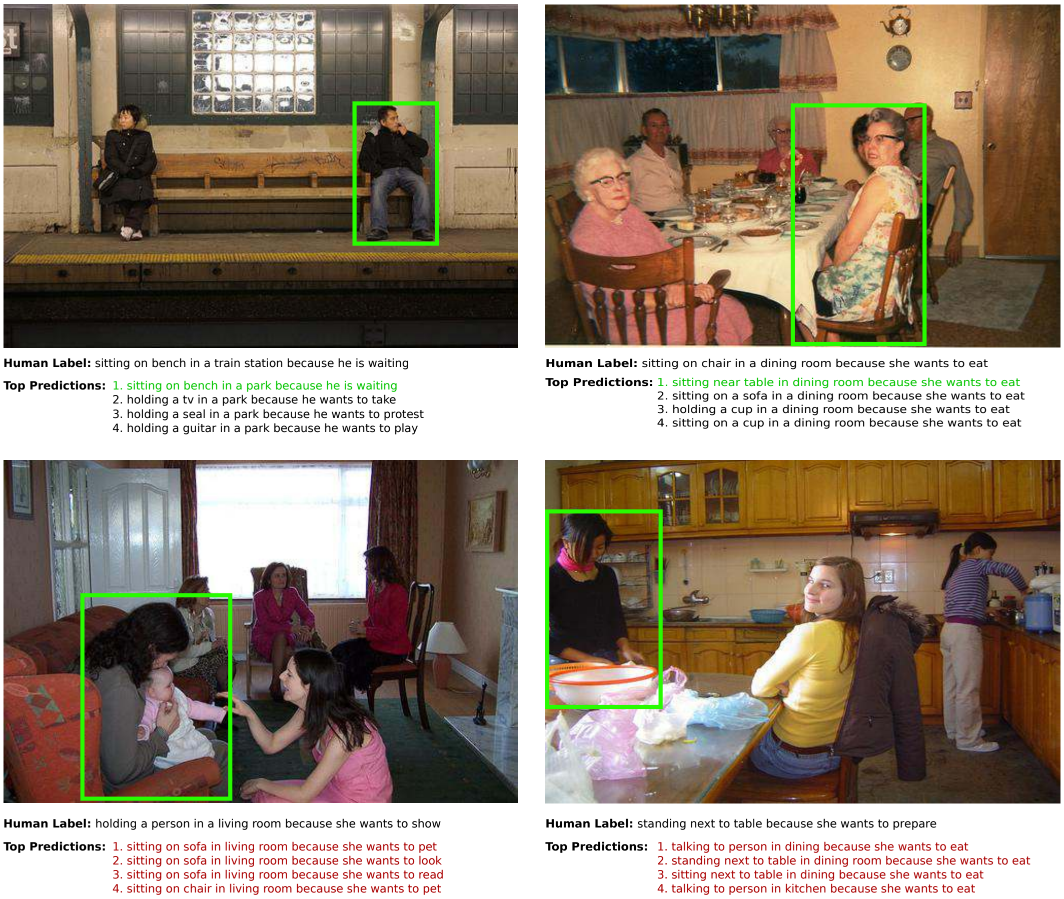
We show a few samples of successes and failures for our approach in Fig.5 and the baseline in Fig.6. We hypothesize that our model often produces more sensible failures because it leverages some of the common knowledge available in written language. For example, our framework incorrectly predicts that the woman is performing an action because she wants to eat, but this is reasonable because she is in the kitchen. However, the baseline can fail in unusual ways: for a woman sitting in a living room while reading, it predicts she wants to eat!

Since our method attempts to reason about many visual concepts, it can infer a rich understanding of images, such as predicting the action, object of interaction, and scene simultaneously with the motivation. Fig.5 suggests our system does a reasonable job at this joint inference. The language model in this case is acting as a type of contextual model [26, 21]. As our goal in this paper is to explore human motivations in images, we did not evaluate these other visual concepts quantitatively. Nonetheless, our results hint that language models might be a powerful contextual feature for computer vision.

4 Discussion

Inferring the motivations of people is a novel problem in computer vision. We believe that computers should be able to solve this problem because humans possess the capability to think about other people's motivations, likely due to cognitive skills known in psychophysics as the theory of mind. Interestingly, recent work in psychophysics provides evidence that there is a region in our brain devoted to this task [29]. We have proposed a learning-based framework to infer motivation with some success. We hope our contributions can help advance the field of image understanding.

Our experiments indicate that there is still significant room for improving machines' ability to infer motivation. We suspect that advances in high-level visual recognition can help this task. However, our results suggest that visual information alone may not be sufficient for this challenging task. We hypothesize that incorporating common knowledge from other sources can help, and our results imply that written language is one valuable source. We believe that progress in natural language processing can advance high-level reasoning in computer vision.



Human Label: sitting on bench in a train station because he is waiting

Top Predictions: 1. sitting on bench in a park because he is waiting
2. holding a tv in a park because he wants to take
3. holding a seal in a park because he wants to protest
4. holding a guitar in a park because he wants to play

Human Label: sitting on chair in a dining room because she wants to eat

Top Predictions: 1. sitting near table in dining room because she wants to eat
2. sitting on a sofa in a dining room because she wants to eat
3. holding a cup in a dining room because she wants to eat
4. sitting on a cup in a dining room because she wants to eat

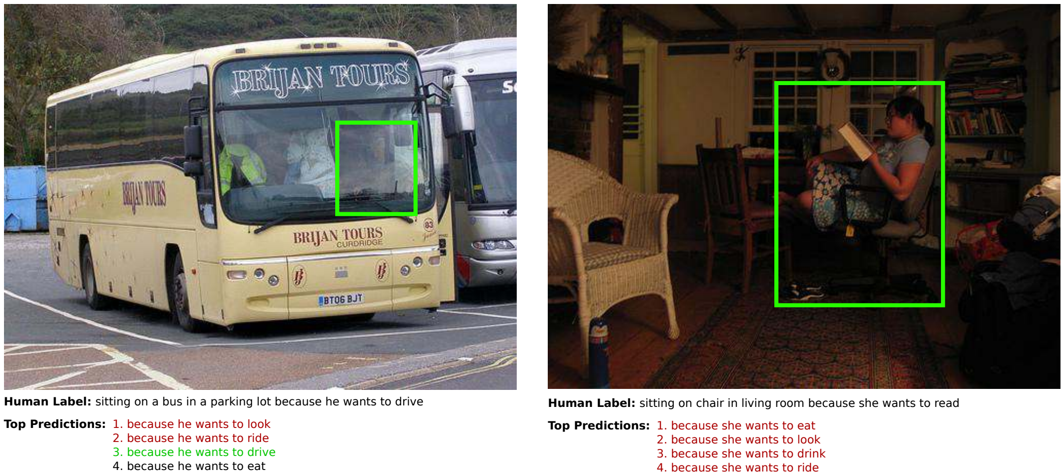
Human Label: holding a person in a living room because she wants to show

Top Predictions: 1. sitting on sofa in living room because she wants to pet
2. sitting on sofa in living room because she wants to look
3. sitting on sofa in living room because she wants to read
4. sitting on chair in living room because she wants to pet

Human Label: standing next to table because she wants to prepare

Top Predictions: 1. talking to person in dining because she wants to eat
2. standing next to table in dining room because she wants to eat
3. sitting next to table in dining because she wants to eat
4. talking to person in kitchen because she wants to eat

Figure 5: **Language+Vision Example Results:** Our framework is able to use a rich understanding of the images to try to infer the motivation behind people's actions. We show a few successes (top) and failures (bottom) of our model's predictions. The human label shows the ground truth by a worker on MTurk. The sentences shown are only to visualize results; our goal is not to generate captions. In many cases, the failures are sensible (e.g. for the bottom right woman in the kitchen, predicting she wants to eat), likely due to the influence of the language model.



Human Label: sitting on a bus in a parking lot because he wants to drive

Top Predictions: 1. because he wants to look
2. because he wants to ride
3. because he wants to drive
4. because he wants to eat

Human Label: sitting on chair in living room because she wants to read

Top Predictions: 1. because she wants to eat
2. because she wants to look
3. because she wants to drink
4. because she wants to ride

Figure 6: **Vision-Only Example Results:** We show a success and failure for a simple vision-only model trained to predict motivation given just mid-level image features. The failures are frequently due to the lack of common knowledge. The baseline predicts the woman wants to eat or ride even though she is in a living room reading. Our full model uses language to suppress these predictions.

Acknowledgements: We thank Lavanya Sharan for important discussions, Bolei Zhou for help with the SUN database, and Kenneth Heafield and Christian Buck for help with transferring the 6 TB language model across the Atlantic ocean. Funding was provided by a NSF GRFP to CV, and a Google research award and ONR MURI N000141010933 to AT.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, page 16. 2

[2] C. L. Baker, R. Saxe, and J. B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 2009. 2, 3

[3] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse m-best solutions in markov random fields. In *ECCV*. 2012. 5

[4] C. Buck, K. Heafield, and B. van Ooyen. N-gram counts and language models from the common crawl. LREC, 2014. 2, 4

[5] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013. 3

[6] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2002. 6

[7] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014. 3

[8] J. Elfring, R. van de Molengraft, and M. Steinbuch. Learning intentions for improved human motion prediction. *Robotics and Autonomous Systems*, 2014. 2

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge. *IJCV*, 2010. 3

[10] K. Heafield. Kenlm: Faster and smaller language model queries. In *Statistical Machine Translation*, 2011. 2, 4

[11] M. Hoai and F. De la Torre. Max-margin early event detectors. In *CVPR*, 2012. 2

[12] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding, 2013. 2, 6

[13] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009. 5

[14] J. Joo, W. Li, F. F. Steen, , and S.-C. Zhu. Visual persuasion: Inferring communicative intents of images. In *CVPR*, 2014. 2

[15] R. Kelley, L. Wigand, B. Hamilton, K. Browne, M. Nicolescu, and M. Nicolescu. Deep networks for predicting human intent with respect to objects. In *International Conference on Human-Robot Interaction*, 2012. 2

[16] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*. 2012. 2

[17] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *RSS*, 2013. 2

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 6

[19] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011. 3

[20] E. L. Lawler. A procedure for computing the k best solutions to discrete optimization problems and its application to the shortest path problem. *Management Science*, 1972. 5

[21] D. Le, R. Bernardi, and J. Uijlings. Exploiting language models to recognize unseen actions. In *ICMR*, 2013. 3, 7

[22] L.-J. Li, H. Su, E. P. Xing, and F.-F. Li. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, volume 2, page 5, 2010. 6

[23] C. McGhan, A. Nasir, and E. Atkins. Human intent prediction using markov decision processes. In *Infotech@ Aerospace 2012*. 2012. 2

[24] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995. 3

[25] S. Nowozin. Grante: Inference and estimation for discrete factor graph model. 5

[26] M. Patel, C. H. Ek, N. Kyriazis, A. Argyros, J. V. Miro, and D. Kragic. Language for learning complex human-object interactions. In *ICRA*, 2013. 3, 7

[27] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010. 2

[28] J. Porway and S.-C. Zhu. C^ 4: Exploring multiple solutions in graphical models by cluster sampling. *PAMI*, 2011. 5

[29] R. Saxe, S. Carey, and N. Kanwisher. Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annu. Rev. Psychol.*, 2004. 2, 7

[30] R. Saxe and N. Kanwisher. People thinking about thinking people: the role of the temporo-parietal junction in theory of mind. *Neuroimage*, 2003. 2, 3

[31] D. Song, N. Kyriazis, I. Oikonomidis, C. Papazov, A. Argyros, D. Burschka, and D. Kragic. Predicting human intention in visual observations of hand/object interactions. In *ICRA*, 2013. 2

[32] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *CVPR*, 2012. 2

[33] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. 2009. 3

[34] H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 1983. 2

[35] D. Xie, S. Todorovic, and S.-C. Zhu. Inferring" dark matter" and" dark energy" from videos. In *ICCV*, 2013. 2

[36] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013. 3

[37] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *ICCV*, 2013. 3