Can You Spot the Semantic Predicate in this Video?

Christopher Reale, Claire Bonial, Heesung Kwon and **Clare R. Voss**
U.S. Army Research Lab, Adelphi, Maryland 20783
christopher.reale@gmail.com
{claire.n.bonial.civ, heesung.kwon.civ, clare.r.voss.civ}@mail.mil

Abstract

We propose a method to improve human activity recognition in video by leveraging semantic information about the target activities from an expert-defined linguistic resource, VerbNet. Our hypothesis is that activities that share similar event semantics, as defined by the semantic predicates of VerbNet, will be more likely to share some visual components. We use a deep convolutional neural network approach as a baseline and incorporate linguistic information from VerbNet through multi-task learning. We present results of experiments showing the added information has negligible impact on recognition performance. We discuss how this may be because the lexical semantic information defined by VerbNet is generally not visually salient given the video processing approach used here, and how we may handle this in future approaches.

1 Introduction

Human activity recognition is a crucial component of comprehensive, multimodal event detection and identification as well as a prerequisite for more complex tasks, such as establishing timelines from video and video caption generation. In this work, we attempt to improve the performance of activity recognition in video by considering the event semantics associated with the activity types. Yatskar et al. (2016) have done this for images by leveraging a large dataset that is thoroughly labeled with the activity performed in each image, the actors involved, and the roles the actors play in the activity. While their method works well, it comes at the cost of obtaining and labeling the dataset. In this work, we eschew the use of an expensive labeled dataset and instead leverage the lexical semantic information found in VerbNet (VN) (Kipper et al., 2008) and only a small amount of manually annotated data.

Our hypothesis is that activities that share similar event semantics will be more likely to share some visual components. To begin to explore this hypothesis, we must first select the type and specificity of semantics that should be coupled with an activity. Here, we use VN to obtain semantic representations in the form of composed predicates, which apply to classes of verbs denoting more or less similar event types. The semantic representations therefore provide a level of generalization over somewhat distinct events. This extra information comes at the activity level rather than the sample level, and thus provides relatively weak supervision. Nonetheless, we feel our hypothesis intuitively holds promise and, if supported, would enable efficient improvement in activity recognition with less training data. Specifically, the ability to detect similar visual components across an event type could allow for generalizing from the recognition of one activity type (e.g., baseball pitch) to another that is semantically similar (e.g., throw discus).

2 Related Work

Human activity recognition is a heavily researched problem in computer vision. The goal is to determine the activity being performed in a video (e.g., walking, playing piano). This can be challenging due to the large range of appearances that videos of a given activity can take on. The problem is usually formulated as a classification task where each target video must be classified as one of a list of potential activities based on features extracted from its frames.

Proceedings of the Workshop on Events and Stories in the News, pages 55–60
Santa Fe, New Mexico, USA, August 20, 2018.

In addition to visible features extracted from videos, it is common to leverage external information from text, audio, or image data sets. Much work has been done investigating the relationship between text and imagery, though it has mainly focused on still images rather than videos. Recent work in this area has been spurred by the increasing availability of novel datasets, such as the visual question-answering research of Antol et al. (2015), which makes use of a dataset of open-ended natural language questions about images. On the video side, Motwani and Mooney (2012) use text mining and object recognition to help with activity recognition. Vondrick et al. (2016) leverage text captions to try to assign intent to human actions in video. To our knowledge, no work has been done to examine the relationship between event semantics in text-based lexical resources and videos.

3 Background to VerbNet

VerbNet,[1] based on the verb classification of Levin (1993), groups verbs into classes according to their compatibility with certain "diathesis alternations" or syntactic alternations (e.g., *She loaded the wagon with hay* vs. *She loaded hay into the wagon*). Although the groupings are primarily syntactic, the classes do share semantic features as well, since, as Levin posited, the syntactic behavior of a verb is largely determined by its meaning.[2] VN makes the shared semantics of a class explicit by including a semantic representation for each usage example demonstrating a characteristic diathesis alternation of a class. For example, in the Throw class, the following semantic representation would apply to this usage example:

Ex.:"Maddox pitched the ball into the field."
Roles: Agent Verb Theme Destination
Semantic Predicates:
CONTACT(during(E0), Agent, Theme)
EXERT_FORCE(during(E0), Agent, Theme)
not(CONTACT(during(E1), Agent, Theme))
MOTION(during(E1), Theme)
LOCATION(end(E1), Theme, Destination)
not(LOCATION(start(E1), Theme, Destination))
CAUSE(Agent, E1)

This representation is intended to break the event down into smaller semantic elements, given as the predicates (in caps). The predicates are organized with respect to the time of the event ('E'); thus they can apply during, at the start, or at the end of an event. The above representation can be paraphrased as expressing that Maddox (Agent) is in contact with and exerts force (E0) on the ball (Theme); he then releases (is not in contact with) the ball and the ball is in motion (E1); the ball's location at the end of the motion event is the field (Destination), where it was not located at the start of the event; Maddox causes this event as the Agent. Notice that although this representation captures many of the salient semantic components of a throwing event, it may not capture the salient visual aspects of a throwing events.

The numbered classes in VN are organized into a shallow taxonomy. Classes with shared semantic elements, and accordingly with shared subsets of the same semantic predicates, begin with the same class number. For example, Throw-17.1 and Pelt-17.2 form one "meta-class."

4 Data and Annotations

For video data, we chose to use the benchmark UCF101 dataset (Soomro et al., 2012) because of its wide variety of activities in comparison to other datasets. We then compared the coverage of the 101 activities in UCF to the types of events represented in VN. We selected four types of events of interest that had some overlap between UCF and VN: events involving motion with a vehicle (VN meta-class 51.4.X), throwing events (VN 17.X), hitting events (VN 18.X), and human group motion events (VN 51.3.2). Intuitively, we felt that each of these events had some clear visual properties associated with the semantics of the type (i.e. fairly clear, distinct image-schemas (Lakoff, 1990))

[1] VerbNet version 3.2 is used: https://verbs.colorado.edu/verb-index

[2] Levin's hypothesis continues to be debated, but efforts to crowdsource empirical evidence of the presence and saliency of the semantics in VN are promising (Hartshorne et al., 2014; Hartshorne et al., 2013).

One linguist and author of this paper, experienced with VN annotation, annotated each of the 101 UCF activity categories with an indication of whether or not the semantics of one of the four types was present in that activity. This was done by first completing a thorough review of the semantic representations found in the (meta-)classes for a given event type.[3] Then, a sample of 10-12 videos from each UCF activity category were observed to get a sense of the nature of actions included in an activity, and the variability of the clips included under a single category.[4] For each UCF category, one of the following indications was given: "Yes" the semantic elements of an event type are present in this activity, "No" the semantics are not present in the activity, or "maybe" the semantics are present in some videos of the activity but not in all. For example, video clips of the UCF activity Biking all include the semantics of motion with a vehicle, none include the semantics of throwing or hitting events, and some clips may include the semantics of group motion in video clips that show a group of cyclists. Additional examples are included in Table 1.[5]

UCF Activity	**Vehicle Motion**	**Throwing**	**Hitting**	**Group Motion**
ApplyEyeMakeup	no	no	no	no
Bowling	no	yes	maybe	no
Drumming	no	no	yes	no
Surfing	yes	no	no	no
MilitaryParade	no	no	no	yes
SkateBoarding	yes	no	no	no
Total "Yes"/"Maybe"/"No"	12/1/88	10/8/83	14/12/75	3/4/94

Table 1: Sample of annotation examples showing which UCF activities correspond to which event type. The "maybe" indicates some Bowling video clips that include the ball hitting pins. Annotations were completed for all 101 UCF activities, only six of these are shown here.

5 Experiment

We use the two-stream convolutional network approach of Simonyan and Zisserman (2014) as a baseline model.[6] In this model, two neural networks are trained to classify videos. The first is trained on the raw frames and the second on optical flow features extracted from the frames. Both networks are trained to classify the activities from small portions of the video (the visible network is trained on single frames, while the motion network is trained on five-frame segments). To test a video, 25 equally spaced frames are passed through the visual network and 25 equally spaced five-frame segments are passed through the motion network. The video is then classified as the activity with the highest average probability.

We alter the approach of Simonyan and Zisserman (2014) by injecting information from VN with multi-task learning (Caruana, 1993). Multi-task learning is the process of simultaneously training for several objectives. In our case, we train the networks to classify the VN class or meta-class the activities belong to in addition to the categories of the activities themselves.

5.1 Network and Training Details

We use the pre-trained (other than the final layer, which is randomly initialized) version of AlexNet (Krizhevsky et al., 2012) for the frame network. For the optical flow network, we use the structure of CNN-M 2048 of Chatfield et al. (2014) with random initialization. The networks are very similar in that both have five convolution layers followed by three fully connected layers. The main difference between the two (beyond initialization method) is that CNN-M 2048 is wider (more hidden nodes per layer).

[3]Note that we are focused on finding the presence or absence of particular semantic predicates (e.g., is there MOTION? is there CONTACT?), as opposed to certain participants or semantic roles. We are focusing on the latter in ongoing work.

[4]UCF includes 13320 videos, about 130 videos/activity.

[5]All annotations can be made available upon request.

[6]Many state-of-the-art activity recognition methods are offshoots of the two-stream approach of Simonyan and Zisserman (2014). For example, Feichtenhofer et al. (2016) experiment with early fusion of the two streams, and Wang et al. (2016) incorporate object detectors into the algorithm.

In order to incorporate information from VN, we create four extra tasks for the networks to solve. Each task is to determine whether or not a video belongs to one of the four chosen VN event types. We formulate this with a separate logistic regression loss for each event type. During training, for each event type, we treat activities labeled "yes" as positive samples, activities labeled "no" as negative samples, and ignore activities labeled "maybe."

All tasks share the first seven layers of the network (Conv1, Conv2, Conv3, Conv4, Conv5, FC6, and FC7). The last layer of the network is a fully connected (FC) layer that serves as a linear classifier for a given task, so it cannot be shared amongst them. Thus each objective has its own eighth (FC8) layer. See Figure 1 for a visual representation of the layout.

Figure 1: Multitask learning layout: the orange shapes represent loss functions, the blue shapes represent the network, and the green shape represents the input data. Sections outlined with dotted lines denote the additional multitask learning components.

We train the networks using the standard back-propagation algorithm and mini-batch stochastic gradient descent. We train the raw frame network for a total of 20,000 iterations. We start the learning rate at .01 and divide it by 10 after 14,000 iterations. Due to the random initialization, we must train the optical flow network for longer, and thus train it for 110,000 iterations. We start the learning rate at .01, and divide it by 10 after 50,000 and 100,000 iterations. When training both networks, we set the momentum to .9, the batch size to 256, and the weight decay to .0001. We use dropout in the first two fully connected layers (FC6 and FC7) of both networks with dropout rate of .5 for the frame network and .75 for the flow network.

5.2 Results

We ran two sets of experiments; we report results for the networks individually (as opposed to their fused performance) in order to examine which of the two networks (one trained on frames, the other on optical flow) is most affected by our method. For all experiments, we report the classification accuracy (i.e. percentage of videos matched to correct activity out of 101 choices) of the trained networks on the 3,783 test videos from split one of the of the UCF data set.

In the first set of experiments, we use all four VN classification objectives in addition to the primary UCF classification objective. We then train the networks with different values (0, 1, 2) for the relative weight given to the VN objectives. Table 2 compares the results when the VN objectives (All, All 2x Weight) are used and when they are not used (Baseline). When the weight is set to one, our method has a negligible effect on the performance compared to the baseline method (i.e. weight = 0). Increasing the weight to two causes the network performance to decrease.

In the second set of experiments, we train the networks with only one additional VN task at a time to see if some event types are more beneficial than others. All objectives were given the same weight in these experiments. In Table 2, we also compare the results of using one VN objective at a time (Vehicle Motion, Throwing, Hitting, Group Motion) to not using any of the VN objectives (Baseline). The extra VN tasks have little effect on the activity recognition performance compared to the baseline method.

We believe the main reason our method fails to provide significant improvement over the baseline is the lack of a relationship between the VN categories and visual appearance of the activities. For example, Bowling and BaseballPitch are both in the throwing category, but they are not necessarily visually similar—what portions of the actions are visually similar are likely not salient enough to be

VN Type	Frames	Optical Flow
All	67.14	77.07
All 2x Weight	61.33	76.47
Vehicle Motion	66.22	77.07
Throwing	66.32	77.27
Hitting	66.59	76.92
Group Motion	67.17	77.50
Baseline	67.08	77.13

Table 2: Performance comparison of our method to the baseline method (Baseline) which does not use any VN information. We evaluate our method using all four VN objectives at once (All), all at once but doubly weighted (All 2x Weight), and with one VN objective at a time (Vehicle Motion, Throwing, Hitting, Group Motion). All values denote the classification accuracy on the 3783 videos.

useful in the current video processing approach (see sample still images taken from the UCF101 dataset in Figure 2). Although our hypothesis was not supported using this approach, we have recourse to pursue this hypothesis in different ways. It may be that the semantic representation that we selected is adequate, but our video processing methodology is not, or that we need to select a different semantic representation more suited to the current video processing methodology.

Figure 2: Still images taken from video clips of Baseball Pitch and Bowling videos in UCF101 dataset (http://crcv.ucf.edu/data/UCF101.php). Although the event semantics of the two activities may share similarities, many visual aspects of the activities (e.g., the surroundings) are very different.

6 Conclusions & Future Work

Although a negative result, this is a notable finding: event-semantic similarity does not necessarily translate into visual similarity. In our next steps, we will ensure that the extra semantic information used to train our network is also visually salient. We plan to do this with a multi-task learning approach where, in addition to training the network to recognize the activities, we will also train it to recognize objects and entities that are frequently associated with the target activities. This will enable us to better recognize, for example, Bowling activities based on the recognition of a bowling ball and pins. We will leverage annotated text corpora to determine what objects most often fill a participant role slot, or semantic role, for the target activities. We will then use ImageNet (Deng et al., 2009) to train our systems to detect these objects.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Richard Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann.

K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*.

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.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941.

Joshua K Hartshorne, Claire Bonial, and Martha Palmer. 2013. The verbcorner project: Toward an empirically-based semantic decomposition of verbs. In *EMNLP*, pages 1438–1442.

Joshua K Hartshorne, Claire Bonial, and Martha Palmer. 2014. The verbcorner project: Findings from phase 1 of crowd-sourcing a semantic decomposition of verbs. In *ACL (2)*, pages 397–402.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

George Lakoff. 1990. The invariance hypothesis: is abstract reason based on image-schemas? *Cognitive Linguistics (includes Cognitive Linguistic Bibliography)*, 1(1):39–74.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Tanvi S Motwani and Raymond J Mooney. 2012. Improving video activity recognition using object recognition and text mining. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 600–605. IOS Press.

Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

C. Vondrick, D. Oktay, H. Pirsiavash, and A. Torralba. 2016. Predicting motivations of actions by leveraging text. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2997–3005, June.

Yifan Wang, Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. 2016. Two-stream sr-cnns for action recognition in videos. In *BMVC*, pages ***–***.

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.