A Large Scale Multi-Label Action Dataset for Video Understanding

Mathew Monfort[1], Kandan Ramakrishnan[1], Dan Gutfreund[23], Aude Oliva[1]
[1] MIT CSAIL, [2] IBM Research, [3] MIT-IBM Watson AI Lab

Abstract

The world is inherently multi-label. Even when restricted to the space of actions, multiple things and events often happen simultaneously and a single label is commonly insufficient for adequately explaining the full meaning of an event. To develop methods reaching human-level understanding of dynamical events, we need to capture the complex nature of our environment. Here, we present a multi-label extension to the Moments in Time Dataset which includes annotation of multiple actions in each video. We perform a baseline analysis and compare recognition results, class selectivity, and network robustness of a temporal relation network (TRN) trained on both single-label Moments in Time and the proposed multi-label extension.

Introduction

While most action recognition datasets are characterized by having a single label per video, each clip often conveys multiple actions and a diversity of events unfolding at the same time which can be seen in the examples given in Figure 2. Additionally, a single description of an action can be ambiguous. For example, "running" can refer to a person jogging on the side of the road or an engine "running". The action label "running" would be correct in both of these instances, however it is clear that more information is needed to properly describe the activity space and increase the boundary between these two videos. Using a different, more specific, label than running might address this problem, however, this more specific label would likely ignore the semantic relationship between both instances of the action which contains rich information for making further abstraction and analogies. To address these issues, we have extended a video dataset, the Moments in Time dataset [6], to contain multiple action labels per video. This allows for a richer understanding of an event unfolding in a few seconds, as well as capture different levels of semantic hierarchy for each action happening in a video.

Due to the large coverage and high level labels provided, the Moments in Time dataset constitutes an ideal platform to build a richer semantic space of concepts. Here we outline the process we used to annotate additional actions for the videos in the dataset and present results on models trained using the proposed multi-label extension to the Moments in Time Dataset.

Related work

Visual understanding is a dynamic process which has inspired many teams to produce large scale datasets like Sports-1M [2], ActivityNet [1], Kinetics [3], and Moments in Time [1] which have provided deep learning models with enough data to learn to recognize actions in videos. Several architectures capitalizing on different sources of information have been proposed to deal with video streams such as two-stream CNNs [8]which separately process optical flow and RGB frames and 3D convolutional networks [10]. More recently, Temporal Relation Networks (TRN) [11] have been proposed to more explicitly learn temporal dependencies between frames.

Multi-label Moments

Annotation

We follow the same annotation pipeline used to annotate the original labels for the Moments in Time dataset. This includes using Amazon Mechanical Turk for crowd sourcing where each worker is presented with the video-verb pair and asked to press a Yes or No key responding if the action is happening in the scene. For more details we refer to the Moments in Time paper [6]. For each video existing in the Moments in Time dataset we generate new action candidates that we present to workers for annotation. The difficulty of this task is centered on choosing candidate actions that are likely to return positive responses when presented to workers for annotation.

Generating action candidates

We generate candidates actions for each video using a few different methods. The first consists of using WordNet [5] relationships and choosing actions that are closely linked in the provided semantic graph (with the addition of a few hand chosen relationships). We decided to restrict our vocabulary to the original Moments in Time vocabulary in order to simplify this process which allowed for us to choose the 5 most closely linked candidate actions to the original action label provided by the dataset. Similarly, we also choose the 5 most similar action candidates to the original label using Word2Vec [4] similarity scores, and finally the 5 actions with the highest probability given an svm ensemble model which combines spatial, temporal, and auditory information to form a single action prediction. This model was trained on the single label dataset and is provided by the authors of the original work [6]. Using both WordNet and Word2Vec allows us to identify closely related action candidates as well as labels for multiple layers of an actions semantic hierarchy (e.g. running -> exercising), and the neural network model allows for background, or seemingly unrelated but co-occurring, action candidates to be generated. Once a candidate is generated, we pass it through at least 2 rounds of annotation to obtain human consensus on its presence in the video. We repeated

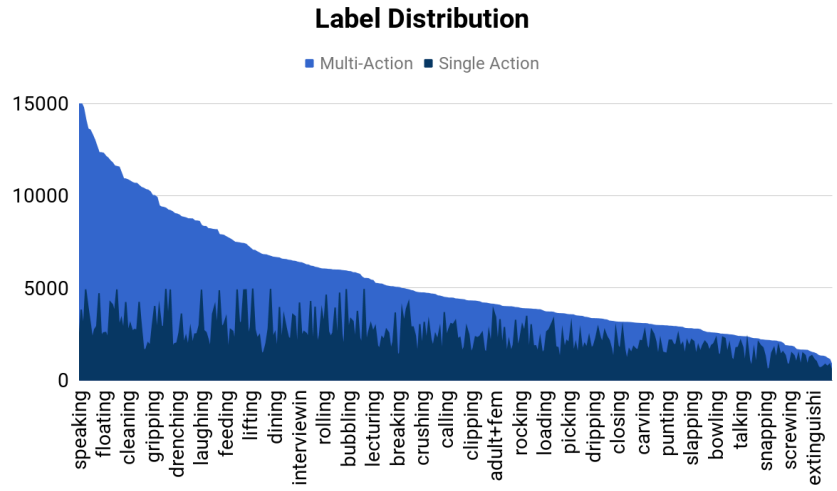


Figure 1: Comparison of the number of videos associated with each action class in the proposed multi-label training set to the single-label Moments in Time training set.

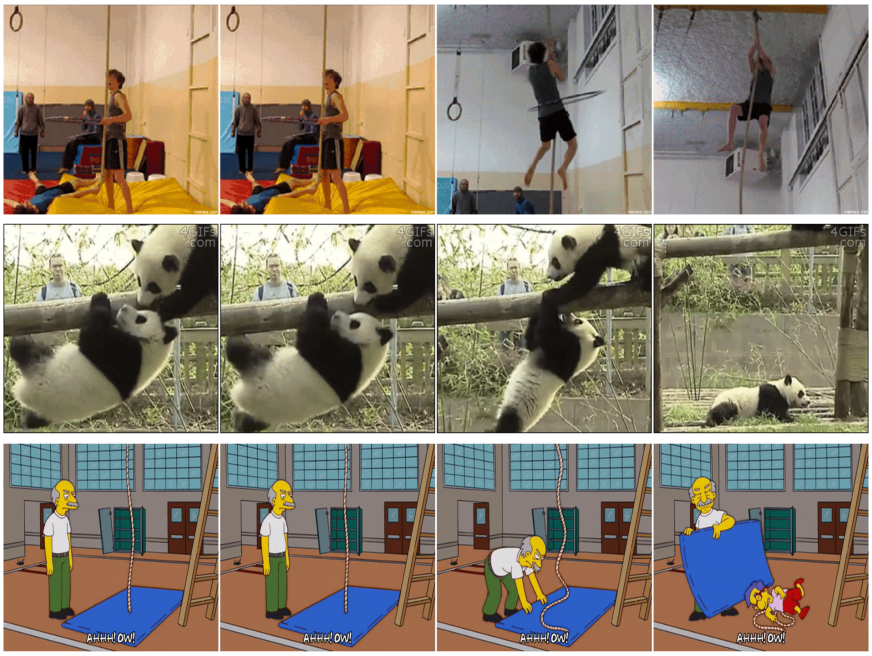


Figure 2: An example of frames from 3 videos from the Moments in Time dataset depicting mulitple actions. The top video shows someone hulahooping while climbing a rope, the middle video depicts a panda climbing a tree before fall to the ground, and the bottome video shows a child falling to the floor after a man pulls a mat away.

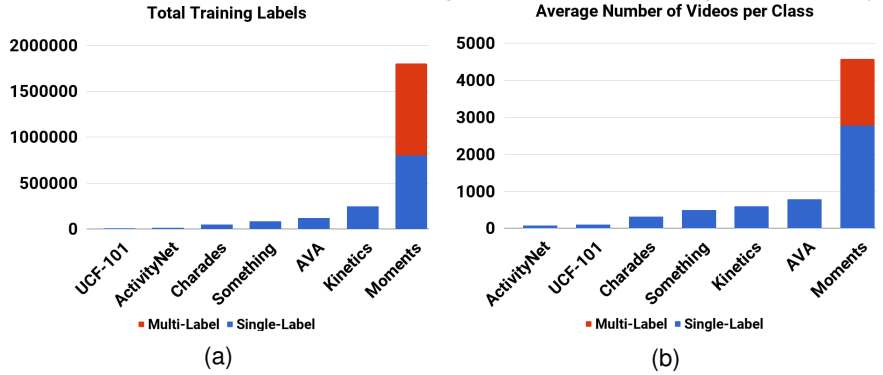


Figure 3: Comparison of the total number of action labels (a) and the average number of videos per class (b) in the proposed multi-label training set to the single-label training set as well as other existing video datasets.

Multi-Label Worker Results		
Return	Label 1	Label 2
0.93	pouring	cooking
0.89	surfing	flowing
0.85	rowing	boating
0.84	pedaling	bicycling
0.84	drumming	playing
0.84	flowing	descending
0.84	snuggling	cuddling
0.84	instructing	smiling
0.84	manicuring	grooming
0.83	washing	wetting
0.83	sweeping	cleaning
0.83	removing	working
0.82	shaving	grooming
0.80	shouting	performing
0.80	sleeping	resting
0.80	buttoning	dressing
0.80	rinsing	wetting
0.80	dropping	falling
0.80	welding	working
0.80	competing	playing+sports

Figure 4: Sample action relationships found from annotating new actions. The "return" reports the percentage of positive worker responses when presented with candidate action "Label 2" and a video already annotated with "Label 1".

this process (candidate generation and annotation) for each new action label annotated for each video. Figure 4 displays a sample of the positive return (percentage of "yes" responses) for presenting videos to workers with candidate action "Label 2" when they are already annotated with "Label 1".

Statistics and analysis

The training set of the Moments in Time Dataset consists of 802,244 annotated 3-second videos each with a single label from a set of 339 different verbs depicting an action or activity. Similarly the validation set contains 33,900 videos each with a single label.

Our multi-label variant of the Moments in Time training set consists of 1,800,047 labels where 507,362 videos are annotated with more than 1 label and 274,150 videos are annotated with 3 or more labels. In addition, we have expanded the validation set to include 99,472 labeled actions for the provided 33,900 videos. Figure 1 compares the distribution of the number of videos for each action class in our multi-label training set to the original single-label training set. Additionally, Figure 3 highlights the expansion of the labelset in comparison to other large-scale video datasets for action recognition.

Baseline results and analysis

To gather some preliminary baseline results on the proposed multi-label extension to Moments in Time, we decided to use Temporal Relation Networks (TRN) [11] as this was the best performing architecture on the original single-label version of Moments in Time. These networks were designed to explicitly learn the temporal dependencies between video segments that best characterize a particular action. This 'plug-and-play" module can model several short-range and long-range temporal dependencies simultaneously to classify actions that unfold at multiple time scales. In this paper, a 6 frame single scale TRN is trained using a ResNet18 [9] network as the base model. For the single-label network we used a softmax on the final layer and a cross entropy loss function. For the multi-label network we used a sigmoid on the final layer and a binary cross entropy loss function.

Dataset	mAP	Input Selectivity	Bottleneck Selectivity
Single	0.23	0.35	0.35
Multi	0.24	0.15	0.17

Table 1: Mean average precision (mAP) and selectivity results for the input to the TRN module (Input Selectivity) and the bottleneck layer of the TRN module (Bottleneck Selectivity) of a TRN trained using a Resnet18 base model trained on both the original single-label Moments in Time Dataset (single) and the proposed multi-label extension (multi). Results are evaluated on the multi-label version of the validation set.

Class selectivity

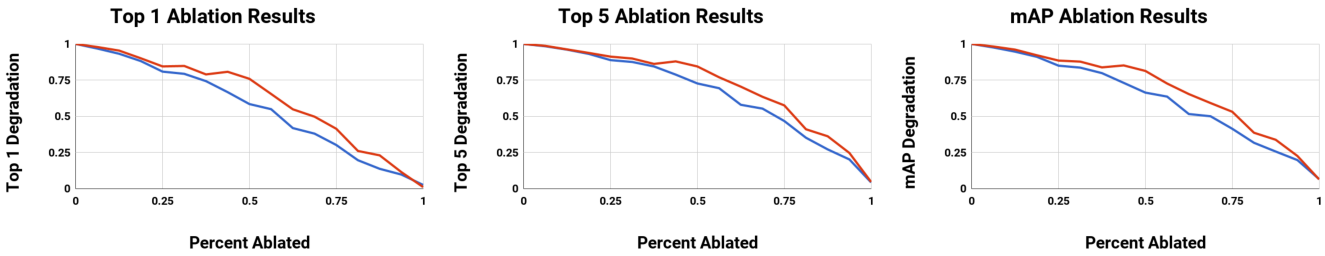
To further analyze the effect of training models using multiple labels we evaluate not just the recognition performance, using mean average precision (mAP), but also the mean class selectivity of both the feature layer used as input into the TRN module (input selectivity) and the bottleneck feature layer of the module itself (bottleneck selectivity). This is similar to recent work in analyzing class selectivity in neural networks for image recognition [7]. We calculate the mean selectivity for each layer,

$$selectivity = \frac{1}{N}\sum_{i=1}^{N}\frac{x_{max} - x_{mean}}{x_{max} + x_{mean}},$$

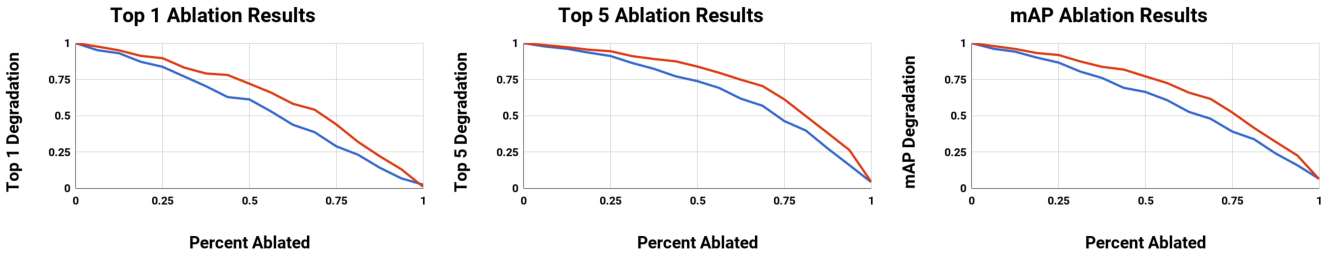
where N represents the number of neurons in the layer and x_{max} and x_{mean} are the maximum and average class probabilities when the activation of neuron i in the feature layer is set to 1 and the activation of all other neurons in the layer are set to 0. Table 1 displays mean average precision (mAP) results on both the original single-label Moments in Time Dataset and the proposed multi-label extension. All results were evaluated on the multi-label validation set.

Ablation

In addition to class selectivity, we performed an ablation analysis to examine the robustness of the learned feature representations of both single-label and multi-label networks. Recent work has proposed that networks that are less reliant on single directions achieve better generalization performance [7]. For this experiment,



(a) Results from ablating the input to the TRN module.



(b) Results from ablating the bottleneck layer of the TRN module.

Figure 5: Results of ablation analysis (averaged over 10 trials) on a resnet18 model trained on single-label Moments (blue) and one trained on multi-label Moments (red) using different performance metrics. The graphs show the improved robustness to neuron ablation of the network trained using multiple labels per video.

we randomly ablated (clamped activations to 0) an increasing percentage of the neurons in 2 different fully connected layers of the TRN module and compared the performance of the network (top-1, top-5, and mAP) to its original non-ablated score. Figure 5 shows the degradation of performance for the model trained on multiple labels (red) and the model trained on single labels (blue) as the number of ablated neurons are increased. The multi-label network suffers consistently less degradation to its performance as the number of ablated neurons increase compared to the single-label network. This, combined with the difference in class selectivity, suggests the model is learning a more distributed representation.

Conclusion

We present a multi-label extension to the Moments in Time Dataset in order to capture full spectrum of actions taking place in each video. This provides a large-scale collection for multi-label video understanding and action recognition covering a wide class of dynamic events that occur in 3-seconds and involve different types of agents (people, animals, objects, and natural phenomena). This dataset presents a novel and difficult task for the field of computer vision in that the labels correspond to different levels of abstraction and can capture multiple simultaneous events. Thus it will serve as a new challenge to develop models that can appropriately scale to the level of complexity and abstract reasoning that a human processes on a daily basis.

Acknowledgements: This work was supported by the MIT-IBM Watson AI Lab and IBM Research.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[2] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[3] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[5] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[6] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa M. Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding. *CoRR*, abs/1801.03150, 2018.

[7] Ari S. Morcos, David G.T. Barrett, Neil C. Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. In *ICLR*, 2018.

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

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *CVPR*, 2015.

[11] Bolei Zhou, Alex Andonian, and Antonio Torralba. Temporal relational reasoning in videos. *arXiv preprint arXiv:1711.08496*, 2017.