Reformulating Zero-shot Action Recognition for Multi-label Actions (Supplementary Material)

Anonymous Author(s)
Affiliation
Address
email

1 AVA Dataset Evaluation

1.1 Extracting Video Clips

Since the AVA dataset consists of multiple actors within one video and ZSAR focuses only on the classification task, we extract clips centered on the ground-truth bounding boxes for each actor in the video. Standard video models expect frame dimensions with the same height and width, so we crop a square region around the actor and resize it to the network specific dimensions (112×112). We present some examples of AVA video frames with their annotations as well as the generated crops in Figure 1. This square crop can cause multiple actors to appear within one clip, as seen in the second example, but it ensures the aspect ratio of the person is not altered, which is necessary as this is the manner in which the video model is trained.

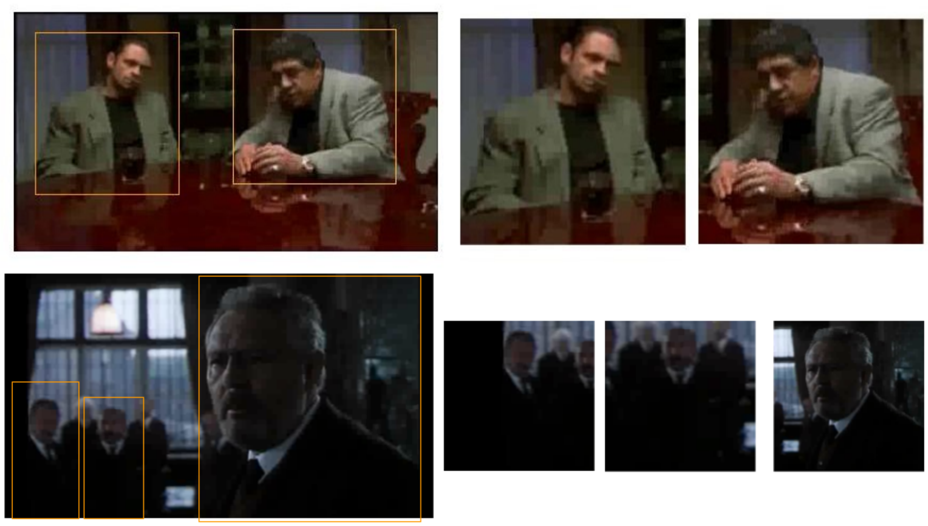


Figure 1: Example of original ground-truth bounding boxes (left) in the AVA dataset, with the cropped actors on the right.

1.2 Generating Multiple Predictions and Confidences

As previous methods for ZSAR tend to be designed for single-label action classification, we adjust these methods to generate multiple predictions along with prediction confidences. For PS-ZSAR

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

prediction confidences are obtained from the softmax probabilities output by our pair-wise similarity function. To obtain confidence scores from the method in Brattoli *et al.* [4], we apply a softmax operation on the inverse cosine distances between the video model's output and the semantic embeddings:

$$P(y|x) = \frac{\exp\left(-d\left(f_\theta\left(x\right), \psi\left(y\right)\right)/\gamma\right)}{\sum\limits_{y' \in \mathcal{U}} \exp\left(-d\left(f_\theta\left(x\right), \psi\left(y'\right)\right)/\gamma\right)}, \tag{1}$$

where d is the cosine distance. As the distances between embeddings tend to be small, we use a temperature parameter $\gamma \leq 1$ increase distances before being passed through the softmax. We find that selecting $\gamma = 0.1$ leads to best results.

To obtain multiple predictions from a given method there are several approaches. One trivial approach is to select the top-k predictions for a given sample. The main issue with this approach is that it may over-predict classes when k is too large or under-predict when k is too small. Another approach is to predict all classes, in which case the mAP evaluation would ignore most low-confidence predictions. This alleviates the issue of under-predicting, but will always over-predict. Finally, we can predict classes based on a confidence threshold, in the manner described in equation 3 of the main paper.

Table 1: mAP Results on AVA Dataset

	Top-1	Top-3	Top-5	No threshold	Threshold
Brattoli *et al.* [4] ($\gamma = 1$)	1.6	2.1	2.4	3.1	6.2
Brattoli *et al.* [4] ($\gamma = 0.1$)	1.6	2.1	2.3	3.3	6.4
Ours (word2vec)	1.6	3.0	3.4	6.4	6.5
Ours (sent2vec)	1.5	3.0	3.5	5.7	7.0

We present results for all approaches in Table 1. It shows that the use of thresholding on predicted probabilities leads to best results. Interestingly, only the top-1 predictions for both methods achieve similar performance, but when it is increased to top-5, the gap between mAP scores increases. This poor performance is due to the nearest neighbor classification which does not allow semantically dissimilar classes to be predicted confidently. On the other hand, our approach can have multiple dissimilar classes in the top-5 predictions.

2 RareAct Evaluation

RareAct is a dataset compiled from rarely co-occurring nouns and verbs such as "microwave show" or "blend phone". It is meant to be "an evaluation dataset notably meant to be used to evaluate models trained on the HowTo100M dataset" [38]. We use RareAct in our work to evaluate how well zero-shot methods can deal with action classes which are extremely different from those seen during training. In the RareAct work [38], the authors propose different metrics (mWAP and mSAP). However, we evaluate our method using the top-1 and top-5 accuracy since the purpose of this work is to create a strong zero-shot classifier rather than learn a joint visual-textual model from a large-scale instructional dataset (i.e. HowTo100M).

3 Evaluation on UCF-101 and HMDB datasets using Random seeds

In Bratolli *et al.* [4], one of the primary methods of evaluation involves generating 10 different testing sets from UCF101 and HMDB by randomly choosing half of the classes. This is the standard evaluation in all works prior to [2], since lack of access to the Kinetics-700 dataset made testing on the full UCF-101 or HMDB dataset infeasible. However, we find that this metric is problematic as the results are dependant almost entirely on the random seed (implemented with numpy's rand package) used to choose which classes to test with. To illustrate this issue, we use 10k random seeds, and report the results in Table 2. The results for Brattoli *et al.* are obtained from the publicly available code and model weights. When results are averaged over all 10k seeds PS-ZSAR outperforms Brattoli *et al.*. Furthermore, we find that our method achieves higher accuracy on 58.3% of the seeds on UCF-101 and 76.5% of the seeds on HMDB .

Table 2: Evaluation on 50% of the UCF-101 and HMDB classes over 10k random seeds. Reported are the mean and standard deviation ($\mu \pm \sigma$).

	UCF101	HMDB
Brattoli *et al.*[4]	39.3 ± 4.3	25.1 ± 4.4
PS-ZSAR (ours)	40.1 ± 3.8	27.3 ± 4.0

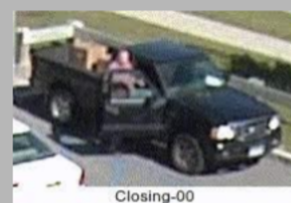
Dataset	**UCF101 Class**	**MEVA Class**
Class Name	BaseballPitch	person_opens_car_door
Encoder Input	“Baseball” “Pitch”	“A person opening the door to a vehicle. The only necessary track in this event is the vehicle. The vehicle door is not independently annotated from the vehicle. This event often overlaps with entering/exiting; however, can be independent or absent from these events.”
Example		

Figure 2: **Example of the natural language descriptions of MEVA classes versus simple class names of UCF101 classes.** Note also for MEVA videos are captured through surveillance camera, and thus actions are lower resolution, as well as less visually apparent.

As our results show, Bratolli *et al.* [4] scores an average of 39.3 on UCF101 and 25.1 on HMDB. However, their reported results are 48.0 and 32.7 respectively, nearly two standard deviations above the mean. In the interest of reporting the most comparable results despite the drawbacks of this evaluation method, we searched for a seed that resulted in their method achieving as close to their reported scores as possible. In the main paper, we then reported our accuracy on that same seed: 49.2 and 33.8 for UCF-101 and HMDB respectively. As this evaluation protocol (i.e. selecting only 10 splits with 50% of the classes) can lead to noisy results, we argue future ZSAR should be evaluate on the entirety of UCF-101 and HMDB.

4 MEVA Dataset Activity Descriptions

Contrary to conventional video datasets which use class names to generate semantic embeddings, the MEVA dataset contain natural language descriptions of the action classes. For example, the action *carrying* has the description "A person carrying an object up to half the size of the person, where the person’s gait has not been substantially modified. The object may be carried in either hand, with both hands, or on one’s back" and the action *falling* has the description "A person falling by either (1) losing one’s balance and possibly collapsing, or (2) moving downward from a higher to a lower level." These lengthy descriptions allow the ZSAR method to learn a richer semantic embedding which is useful for classifying surprise activities.

5 Method Limitations

We analyse how PS-ZSAR performs on the UCF-101 dataset to understand the limitations of the approach. We find that ZSAR methods achieve strong performance on certain classes, while many classes tend to be ignored and not predicted. We present 10 classes on which our method achieves 0% in Table 3. PS-ZSAR tends to predict classes which are visually similar to the target class. For instance, videos with the "Jump Rope" and "Jumping Jack" actions tend to be predicted as "Handstand Pushups" since all three actions involve similar motions (i.e. repetitive up and down motions). This is a limitation for not only our approach, but most ZSAR approaches. For example, Bratolli *et al.* [4] achieve 0% accuracy on 36 classes and PS-ZSAR achieves 0% accuracy on 22 classes. We believe solving this problem would be an interesting avenue for future work.

Table 3: Ten classes which PS-ZSAR performs worst on in the UCF-101 dataset. We include the class name, the accuracy, and the class predicted for most videos of the given class.

Class Name	Most Predicted
Jump Rope	Handstand Pushups
Jumping Jack	Handstand Pushups
Hula Hoop	Tai Chi
YoYo	SalsaSpin
Front Crawl	Breast Stroke
Bowling	Basketball
Parallel Bars	Trampoline Jumping
Playing Daf	Head Massage
Playing Violin	Playing Flute
Pole Vault	Trampoline Jumping

References

[1] Valter Estevam, Helio Pedrini, and David Menotti. Zero-shot action recognition in videos: A survey. *Neurocomputing*, 439:159–175, 2021.

[2] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9436–9445, 2018.

[3] Meera Hahn, Andrew Silva, and James M Rehg. Action2vec: A crossmodal embedding approach to action learning. *arXiv preprint arXiv:1901.00484*, 2019.

[4] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4613–4623, 2020.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] Ioannis Alexiou, Tao Xiang, and Shaogang Gong. Exploring synonyms as context in zero-shot action recognition. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4190–4194. IEEE, 2016.

[7] Xun Xu, Timothy M Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *European Conference on Computer Vision*, pages 343–359. Springer, 2016.

[8] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, 123(3):309–333, 2017.

[9] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *International Journal of Computer Vision*, 124(3):356–383, 2017.

[10] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017.

[11] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9985–9993, 2019.

[12] Xingyu Chen, Xuguang Lan, Fuchun Sun, and Nanning Zheng. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *European Conference on Computer Vision*, pages 572–588. Springer, 2020.

[13] William Thong and Cees GM Snoek. Bias-awareness for zero-shot learning the seen and unseen. *arXiv preprint arXiv:2008.11185*, 2020.

[14] Jiamin Wu, Tianzhu Zhang, Zheng-Jun Zha, Jiebo Luo, Yongdong Zhang, and Feng Wu. Self-supervised domain-aware generative network for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12767–12776, 2020.

[15] Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Zheng-Jun Zha, and Yongdong Zhang. Domain-aware visual bias eliminating for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12664–12673, 2020.

[16] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019.

[17] Chuang Gan, Ming Lin, Yi Yang, Gerard Melo, and Alexander G Hauptmann. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[18] AJ Piergiovanni and Michael S Ryoo. Learning shared multimodal embeddings with unpaired data. *CoRR*, 2018.

[19] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 374–390, 2018.

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

[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[22] Qiang Qiu, Zhuolin Jiang, and Rama Chellappa. Sparse dictionary-based representation and recognition of action attributes. In *2011 International Conference on Computer Vision*, pages 707–714. IEEE, 2011.

[23] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *CVPR 2011*, pages 3337–3344. IEEE, 2011.

[24] Chuang Gan, Ming Lin, Yi Yang, Yueting Zhuang, and Alexander G Hauptmann. Exploring semantic inter-class relationships (sir) for zero-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[25] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1576–1585, 2018.

[26] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8776–8786, 2020.

[27] He Huang, Yuanwei Chen, Wei Tang, Wenhao Zheng, Qing-Guo Chen, Yao Hu, and Philip Yu. Multi-label zero-shot classification by learning to transfer from external knowledge. *arXiv preprint arXiv:2007.15610*, 2020.

[28] Qian Wang and Ke Chen. Multi-label zero-shot human action recognition via joint latent ranking embedding. *Neural Networks*, 122:1–23, 2020.

[29] Mahdi Naser Moghadasi and Yu Zhuang. Sent2vec: A new sentence embedding representation with sentimental semantic. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4672–4680. IEEE, 2020.

[30] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[31] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[33] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

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

[37] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[38] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Rareact: A video dataset of unusual interactions. *arXiv preprint arXiv:2008.01018*, 2020.

[39] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3, 2008.

[40] Alina Roitberg, Manuel Martinez, Monica Haurilet, and Rainer Stiefelhagen. Towards a fair evaluation of zero-shot action recognition using external data. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[41] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.

[42] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1060–1068, January 2021.

[43] Kitware inc, the multiview extended video with activities (meva) dataset.

[44] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

[45] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.