A Multimodal Benchmark and Improved Architecture for Zero Shot Learning

Keval Doshi, Amanmeet Garg, Burak Uzkent, Xiaolong Wang, Mohamed Omar
Amazon Prime Video
{kcdos, amanmega, burauzke, xiaowanf, omarmk}@amazon.com

In this supplementary, we provide additional details regarding the implementation and extend the ablation studies from the manuscript.

1. Additional Implementation Details

1.1. Multiscale Video Transformer

To extract the visual representations from videos, we leverage the multiscale video transformer (MVITv2) architecture proposed in [8]. To have a fair comparison with existing approaches and avoid a direct overlap with unseen classes, we follow the training protocol proposed in [2] and remove the classes that have a semantic relatedness (SR) score of less than 0.05 with respect to UCF and ActivityNet classes. This results in 105 overlapping classes, and 595 non-overlapping classes. We train the MVITv2 model on these 595 classes (K595) to compare with the approaches in Table 4 & 5. On the other hand, [11] proposes training on Sports-1M dataset, which does not have the majority of videos available. To circumvent this issue, we directly use the visual features extracted in their implementation[1].

[1]https://github.com/ExplainableML/TCAF-GZSL

2. Additional Ablation Studies

2.1. Sensitivity analysis of sigma

We show the sensitivity to sigma values in Fig. 2. We observe that the performance is not very sensitive to the value of sigma.

2.2. Comparison to video based approaches

While there has been limited progress in multimodal zero-shot learning, there are several existing works that only leverage the video modality for zero-shot action recognition. In Table 1, we compare the performance using the video branch of the proposed MZST model to recent methods. For a fair comparison, we follow the training and evaluation protocol along with the splits discussed in [2]. All the methods are evaluated by randomly splitting the dataset in half and averaging the results over 10 trials. We can clearly observe that the proposed approach outperforms existing approaches by 4.18% on the UCF dataset and 2.3% on the ActivityNet dataset. This demonstrates the effectiveness of the multiscale representation learning leveraged by the MZST architecture.

Method	UCF	ActivityNet
DataAug [15]	18.3	-
InfDem [13]	17.8	-
Bidirectional [14]	21.4	-
TARN [1]	19	-
Action2Vec [5]	22.1	-
OD [9]	26.9	-
CLASTER [4]	46.4	-
DASZL [7]	48.9	-
GGM [9]	20.3	-
PS-ZSAR (662 classes) [6]	49.2	-
E2E (605 classes) [2]	44.1	26.6
ViSET-96(505 classes) [3]	45.6	35.8
MZST-V (Ours)	**49.78**	**38.1**

Table 1. Comparison with the state-of-the-art video-only methods on standard benchmark datasets.

2.3. Train/Test Splits

Due to the lack of a zero-shot evaluation set, in early video zero-shot literature the datasets were *randomly* split to create train-test sets, leading multiple train/test splits. On the other hand, recent approaches [10, 11] have proposed specific splits such that the test classes do not overlap with the pretraining dataset classes, therefore creating multiple splits is no longer possible. So we use the train-test splits proposed in AVCA.

2.4. Few-Shot setting

We follow GGM's setting [12] and fine-tune the MLP layer of our model on few-samples from each unseen class, and then evaluate the few-shot performance. The performance is averaged over 5 trials.

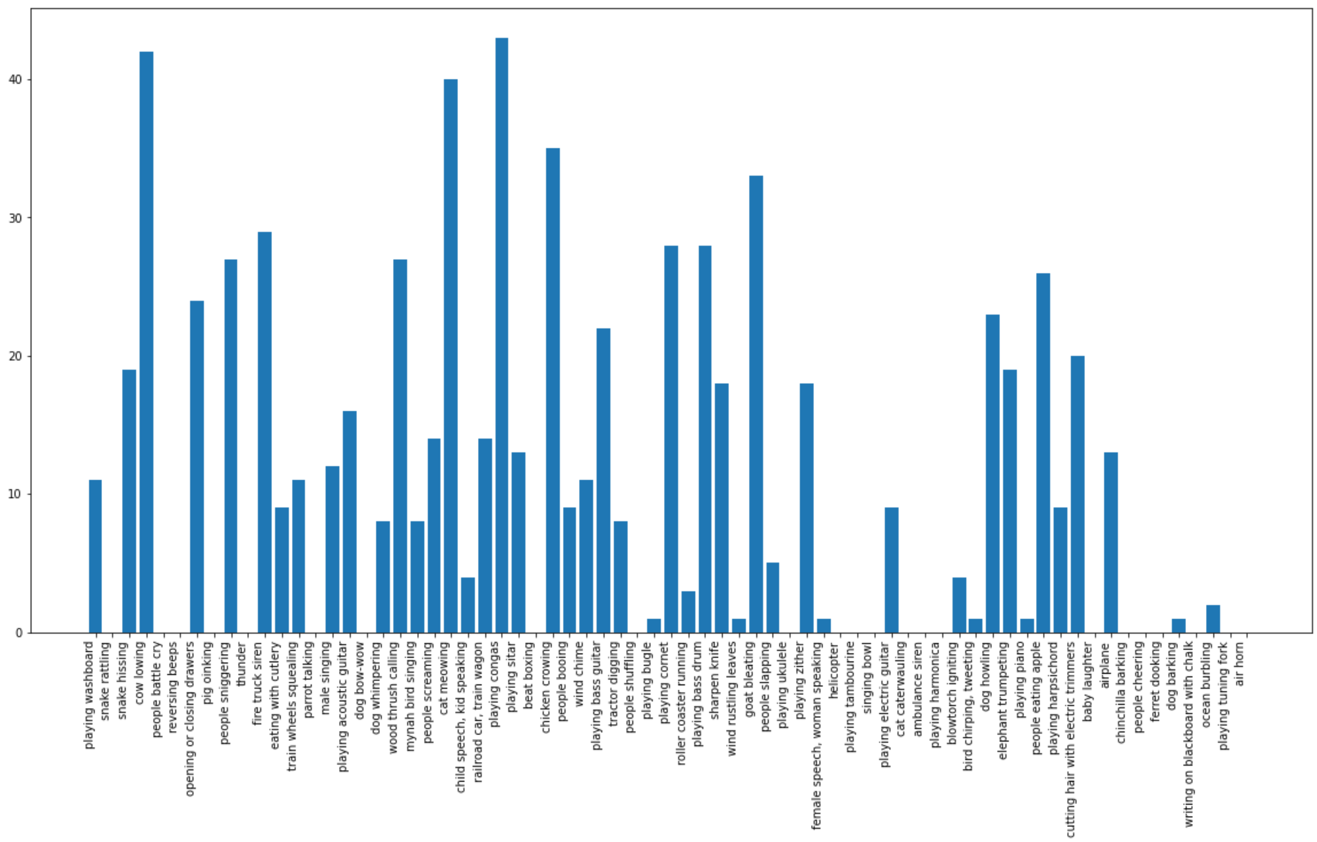

Figure 1. Correct predictions per class on the test split of VGG-Sound dataset.

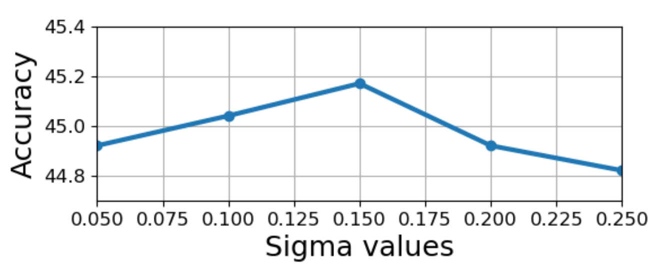

Figure 2. Sensitivity to sigma values

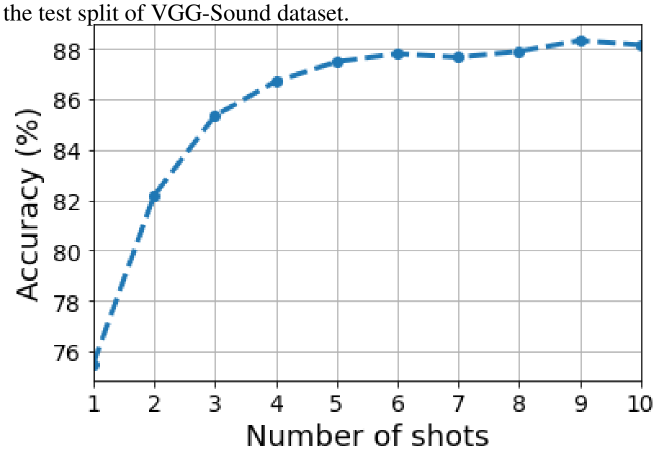

Figure 3. Few-Shot Performance

3. Limitations

In Fig. 1, we demonstrate the performance of the proposed MZST approach on the test split of the VGG-Sound dataset proposed in [10]. While MZST is able to predict several classes correctly, there are a few classes where the total number of correct predictions are zero. For example, *baby laughing* and *people giggling* have similar sounds, but to differentiate them requires a fine grained object level knowledge between a baby and a person. However, such object level information is generally missing in video models which are used to infer the appearance of an object.

References

[1] 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. 1

[2] Biagio Brattoli and others. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *CVPR*, 2020. 1

[3] Keval Doshi and Yasin Yilmaz. Zero-shot action recognition

with transformer-based video semantic embedding. *arXiv preprint arXiv:2203.05156*, 2022. 1

[4] Gowda et al. Claster: Clustering with reinforcement learning for zero-shot action recognition. *arXiv preprint arXiv:2101.07042*, 2021. 1

[5] Meera Hahn, Andrew Silva, and James M Rehg. Action2vec: A crossmodal embedding approach to action learning. *arXiv preprint arXiv:1901.00484*, 2019. 1

[6] Alec Kerrigan, Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Reformulating zero-shot action recognition for multi-label actions. *Advances in Neural Information Processing Systems*, 34:25566–25577, 2021. 1

[7] Tae Soo Kim, Jonathan Jones, Michael Peven, Zihao Xiao, Jin Bai, Yi Zhang, Weichao Qiu, Alan Yuille, and Gregory D Hager. Daszl: Dynamic action signatures for zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1817–1826, 2021. 1

[8] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 1

[9] 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. 1

[10] Otniel-Bogdan Mercea et al. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *CVPR*, 2022. 1, 2

[11] Otniel-Bogdan Mercea, Thomas Hummel, A Koepke, and Zeynep Akata. Temporal and cross-modal attention for audio-visual zero-shot learning. In *European Conference on Computer Vision*, pages 488–505. Springer, 2022. 1

[12] Ashish Mishra et al. A generative approach to zero-shot and few-shot action recognition. In *WACV*, 2018. 1

[13] Alina Roitberg, Ziad Al-Halah, and Rainer Stiefelhagen. Informed democracy: voting-based novelty detection for action recognition. *arXiv preprint arXiv:1810.12819*, 2018. 1

[14] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *International Journal of Computer Vision*, 124(3):356–383, 2017. 1

[15] 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. 1