Towards a Fair Evaluation of Zero-Shot Action Recognition using External Data

Alina Roitberg ✉, Manuel Martinez, Monica Haurilet, and Rainer Stiefelhagen

Karlsruhe Institute of Technology, Karlsruhe, Germany
{firstname.lastname}@kit.edu

Abstract. Zero-shot action recognition aims to classify actions not previously seen during training. This is achieved by learning a visual model for the seen *source* classes and establishing a semantic relationship to the unseen *target* classes *e.g.* through the action labels. In order to draw a clear line between *zero-shot* and conventional *supervised* classification, the *source* and *target* categories must be disjoint. Ensuring this premise is not trivial, especially when the source dataset is external. In this work, we propose an evaluation procedure that enables fair use of external data for zero-shot action recognition. We empirically show that external sources tend to have actions excessively similar to the target classes, strongly influencing the performance and violating the zero-shot premise. To address this, we propose a corrective method to automatically filter out too similar categories by exploiting the pairwise intra-dataset similarity of the labels. Our experiments on the HMDB-51 dataset demonstrate that the zero-shot models consistently benefit from the external sources even under our realistic evaluation, especially when the source categories of internal and external domains are combined.

Keywords: Action Recognition, Zero-Shot Learning

1 Introduction

Human activity recognition has a long list of potential applications, ranging from autonomous driving and robotics to security surveillance [7,12,8]. Knowledge transfer from external sources is crucial for using such models in practice, as they are especially sensitive to the amount of training data due to the 3-D convolution kernels leading to a higher number of parameters [2].

Intersection of vision and language allows us to generalize to new actions without any visual training data through Zero-Shot Learning (ZSL) [15]. ZSL connects a visual model trained on a dataset of known (*source*) classes to the unknown (*target*) classes through the high-level semantic descriptions of an action, e.g., the action label. The description is often represented by a word embedding model (e.g. *word2vec*[5,4]), previously trained on web data. Such ZSL methods would first compute the word vector by mapping a visual representation of a new instance to the common semantic space and then assign it to one of the previously unseen categories by selecting the category with the closest semantic representation (Figure 1).

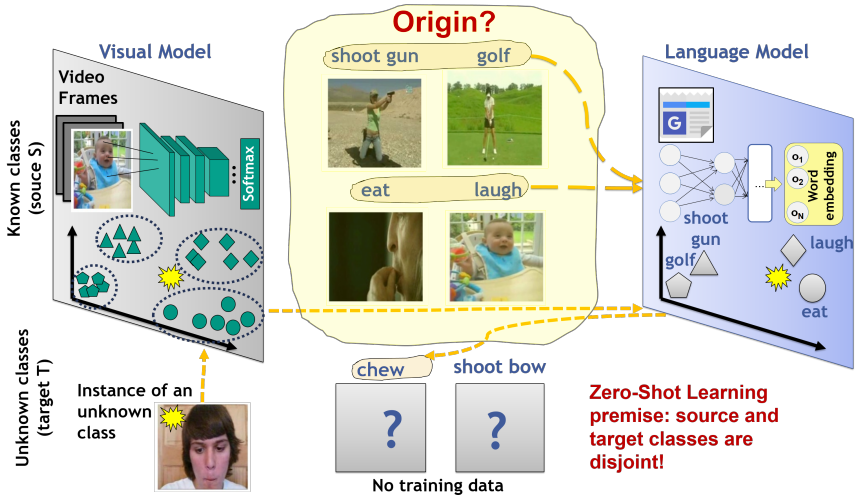

Fig. 1: Zero-shot action recognition paradigm: instances of the new ***target*** classes are recognized without any training data by linking visual features learned from the known ***source*** categories with a language-based representation of the action labels. Our work demonstrates, that the zero-shot premise of disjoint *source* and *target* categories may be violated when using external datasets for training.

ZSL for action recognition gained popularity over the past few years, usually dividing the dataset into *seen* categories for training and *unseen* categories for evaluation [16,9,11,17,14]. Recent emergence of large-scale action recognition datasets has led to an increasing interest in the field of domain adaptation, where the model trained on a high amount of external data is classifying instances from a smaller, potentially application-specific dataset [18]. At the first glance, one would assume, that classifying data from a foreign source would be a harder problem because of the existing domain shift. However, recent works using data from origins other then the evaluation dataset for training of the visual recognition model, report extraordinary results in zero-shot action recognition, doubling the performance of the previous models focused on the inner-dataset split [18]. A single dataset would not contain the same activity twice. Action labels of an external dataset, on the other hand, possibly intersect with the test categories, violating the ZSL premise of assigning action classes **not** seen during training and turning the problem into supervised classification. We argue, that in order to draw the line between *zero-shot* and standard *supervised* recognition across different datasets, it is crucial to take a closer look at the similarity of action categories of source and target data and create a standardized evaluation procedure which eliminates the influence of such overlapping activities.

Contributions and Summary. This work aims to highlight the fact that cross-dataset evaluation of zero-shot action recognition is greatly influenced by the presence of overlapping activity classes in the source and target datasets. We quantitatively analyze the similarities of labels used for training (*source*) and testing (*target*) in the inner-dataset and cross-dataset setup and demonstrate, that *external* labels tend to have categories excessively similar to the unseen *target* classes, therefore violating the ZSL assumption of disjoint source and target categories. We propose a novel procedure that enables the use of *external* data for zero-shot action recognition settings in a fair way, by using the maximum *internal* semantic similarity within the *target* dataset to restrict the *external* classes. We evaluate our method on the HMDB-51 dataset, and show how using external data improves the performance of ZSL approaches, even in our more fair evaluation setting. Finally, we propose a novel *hybrid* ZSL regime, where the model is allowed to use all the internal labels and additional large-scale external data, consistently improving the zero-shot action recognition accuracy.

2 Fair transfer of external categories

Problem Definition. We define the zero-shot learning task as follows. Let $A = \{a_k\}_{k=1}^{K}$ be a set of K previously seen *source* categories. Given the set of previously unseen *target* categories $T = \{t_m\}_{m=1}^{M}$ and a new data sample X, our goal is to predict the correct action category $t \in T$ without having any training data (*i.e.* labeled samples) for this class. Since the core idea of ZSL is to recognize previously unseen visual categories, source labels and target labels are set to be strictly disjoint. This is known as the *zero-shot premise* and is formalized as: $A \cap T = \emptyset$.

2.1 Evaluation protocols for ZSL

Intra-dataset protocol. A common way to evaluate zero-shot learning approaches is to divide a dataset into seen and unseen categories. That is, while a subset of unseen categories is held out during training, both the source and target labels belong to the same dataset: $A = A_{intra}$. In this setting, source and target categories do not overlap, since well designed datasets contain no category duplicates.

Cross-dataset protocol. The main goal of zero-shot learning, however, is to apply knowledge from available data to tasks from a different domain where labeled data is difficult to obtain. This setting is evaluated by training and evaluating on a different datasets: $A = A_{cross}$. In that case, however, the zero-shot premise is not given by default. In the most extreme case, if $T \subset A$, no semantic transfer is needed.

Intra- and Cross- dataset protocol. Recently, several approaches in other computer vision areas have been presented that investigate ways of increasing the performance by mixing the available domain-specific datasets with large amounts of training data from external sources [10]. We transfer this paradigm to the zero-shot action recognition and formalize this ***hybrid*** evaluation regime as: $A = A_{\text{intra}} \cup A_{\text{cross}}$. Similarly to the previous setting, the zero-shot premise is not ensured.

2.2 Proposed protocol to incorporate external datasets

In the intra-class protocol, compliance of the zero-shot premise is given for granted, and generally well accepted by researchers [14,16,9]. However, when external datasets are involved, one has to ensure that the terms of ZSL are still met and the *source* and *target* categories are disjoint. For example, Zhu *et al.* [18] excludes classes from the training dataset whose category label overlaps with a tested label. This procedure would remove the action *brushing hair*, present in both ActivityNet [1] and Kinetics [2], since the label *brush hair* is present in the target classes from the HMDB-51 [3] dataset.

However, it is not trivial to determine if a source class should be excluded and eliminating direct category matches may not be enough. External datasets often contain slightly diverging variants or specializations of the target actions (*e.g.*, *drinking beer* and *drink*), leading to a much closer relation of source and target actions compared to the inner dataset protocol, even if the direct matches are excluded. We argue, that taking into account the similarity of source and target labels is a key element for evaluation of zero-shot action recognition when external sources datasets are used.

We propose a standardized procedure to decide whether an external class should be used or discarded when training the visual model. Our corrective method is based on the fact that zero-shot learning is well-defined for the intra-class protocol, *i.e.* thus all *source* categories of the intra-dataset split can always be used to train our model. We will remove a source category if its label is semantically too similar to any of the target categories by leveraging the maximum similarity observed inside the same dataset as a rejection threshold for categories of foreign origin. Formally, an external category $a_k \in A$ is allowed if and only if following condition is satisfied:

$$\forall t_m \in T,\ s(\omega(a_k), \omega(t_m)) \leqslant s_{th}. \tag{1}$$

The similarity threshold s_{th} corresponds to the maximum pairwise similarity between the source and target labels in the intra-class setting:

$$s_{th} = \max_{a_k \in A_{\text{intra}}, t_m \in T} s(\omega(a_k), \omega(t_m)). \tag{2}$$

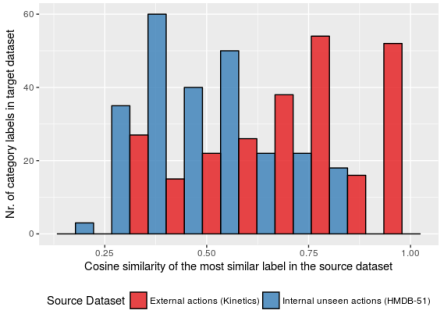

Fig. 2: Histogram of semantic similarities between all target labels and the most similar source label.

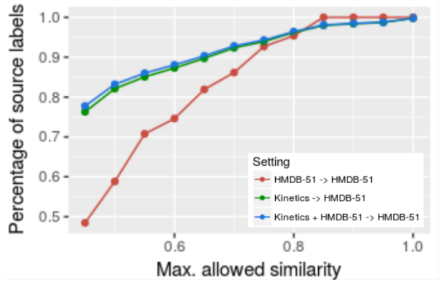

Fig. 3: Proportion of source labels allowed depending on the semantic similarity threshold s_{th}.

3 Experiments

Experimental setup. To evaluate our idea, we adapt an off-the shelf ZSL approach Convex Combination of Semantic Embeddings (ConSE) [6]. While ConSE has been used for zero-shot action recognition before [17], wheren the underlying visual model was based on dense trajectory features [13] encoded as Fisher Vector, we employ a model based on CNNs.

We denote the model for mapping an action label to the word vector representation as $\omega(\cdot)$ and the cosine similarity of the two word vectors as $s(\omega(a_i), \omega(a_j))$. In the next step, a word vector embedding for X is synthesized by taking a linear combination of the predicted probabilities and the semantic representation of source classes: $w^*(X) = \sum_{k=1}^{K} p(a_k|X)\omega(a_k)$. X will be classified to the target category whose semantic representation is most similar to the synthesized word embedding:

$$t_X^* = \underset{t_m \in T}{\operatorname{argmax}}\, s(\omega(t_m), w^*(X)).$$

As our visual recognition model, we use I3D [2], which is the current state-of-the-art method for action recognition. The model is trained using SGD with momentum of 0.9, and an initial learning rate of 0.005 for 100 epochs. To compute the word vectors embeddings of the action categories, we use the publicly available *word2vec* model trained on 100 billion words from Google News articles, which maps the input into a 300 dimensional semantic space [5].

We use HMDB-51 [3] as our target dataset, and we follow the zero-shot learning setup of Wang *et al.* [14]: we generate 10 random splits with 26 seen and 25 unseen categories each. As a foreign data source we use the Kinetics dataset [2], which covers 400 activity categories.

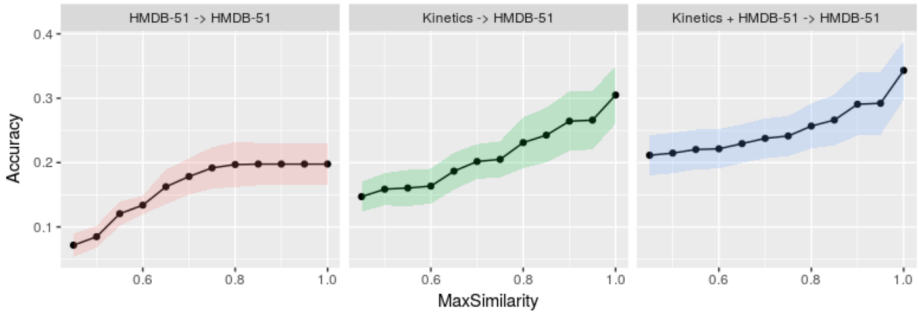

Fig. 4: Influence of source-target label similarity on ZSL performance. X-Axis denotes the semantic similarity threshold s_{th} above which source categories are excluded. Having similar classes in the seen and unseen sets strongly affects accuracy, an effect that is more pronounced when using external datasets.

Intra- and cross-dataset class similarity. First, we re-assure our assumption that labels of seen actions tend to be significantly closer to the unseen categories if they originate from an external dataset. Figure 2 shows the distribution of the maximum pairwise source-target similarity for each source label. We observe that actions from external dataset are far closer, often even identical, to the target classes dataset in comparison to the same dataset case. We explain this distribution by the nature of datasets design, as a single dataset does not contain duplicates or activities that are too close to each other.

Effect of the similar activities on the classification accuracy. Our next area of investigation is the influence of such analogue activities and external data on the classification results. We report the average and standard deviation of the recognition accuracy over the splits for different similarity thresholds s_{th} for restricting the target categories (Fig. 4 and Tbl. 1). Extending the model trained on the native data (intra-dataset) with external datasets (intra- and cross-dataset regimes) increases the accuracy by almost 15%, with 10% improvement observed when an external source is used alone (cross-dataset regime). Excluding direct matches (s_{th} of 0.95) leads to a performance decline of 4% for cross-dataset scenario, although only around 1% of external action categories are excluded (Fig. 3). In other words, only 1% of external action labels (which are extremely similar to the target) account for almost half of the cross-dataset performance boost.

The accuracy saturates at a similarity threshold of around 0.8 in the inner-dataset regime, as no duplicate activities are present (Fig. 3). Our evaluation procedure leverages this maximum inner-dataset similarity to effectively eliminate synonyms from external sources, while not influencing the inner-dataset performance. In our framework, the majority of the external dataset is kept 384.7 of 400. However, the influence of analogue activities is clearly tamed, leading to a performance drop from 34.77% to 25.67% for the inner- and cross-dataset pro-

Exclusion protocol	Source	# source labels	Accuracy	ZSL premise
n. a.	HMDB-51	26	19.92 (±3.3)	✓
Use all source labels	Kinetics	400	30.72 (±4.4)	–
	Kinetics+ HMDB-51	426	**34.77 (±4.5)**	–
Exclude exact labels	Kinetics	≈394.8	26.6 (±4.6)	–
	Kinetics+ HMDB-51	≈420.8	**29.22 (±4.9)**	–
Exclude similar labels **(ours)**	Kinetics	≈384.7	23.1 (±3.9)	✓
	Kinetics+ HMDB-51	≈410.7	**25.67 (±3.5)**	✓

Table 1: ZSL on HMDB-51 for different evaluation regimes with and without our corrective approach. Naively using external sources may not honor the ZSL premise.

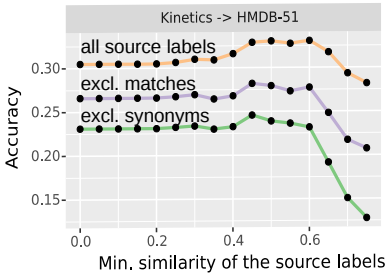

Fig. 5: Effect of eliminating unfamiliar concepts: source categories with similarity to the target labels below 0.4 hinder the performance

tocol. Still, using external data is very beneficial for the recognition results and using both internal and external data sources consistently outperforms single-source models. A clear standardized protocol for defining allowed external source classes without violating the ZSL rules, is a crucial step towards a more adequate model evaluation.

Context of previous work. In this work, our goal is to highlight the ambiguities which arise when external datasets come into play in zero-shot action recognition and we do not aim at state-of the art performance. The vast majority of evaluated methods has used the inner-dataset split, *e.g.* a similar ConSE model employed by [17] which reaches 15.0%, while our model with underlying deep showes an improvement of 19.92%. The state-of-the-art approach using inner-dataset evaluation achieves 22.6% [9], while the recent work of Zhu *et al.* [18] reports highly impressive results of 51.8% employing an external data source (ActivityNet). We want to note, that our model also consistently outperforms state-of-the-art which uses inner-dataset split only. However, we find that systematic elimination of synonyms is crucial for a fair comparison, as we do not know, which actions were allowed in the setting of [18] and we show, that few analogue actions might lead to a clear performance boost.

Eliminating too unfamiliar concepts for better domain adaptation. As a side-observation, we have found that using an additional **lower bound** on the similarity of the external and target categories leads to a performance increase of around 2% for every evaluation setting (Fig. 5). In other words, unfamiliar concepts act as a distractor for the purposes of ZSL.

4 Conclusions

Current machine learning methods based on CNNs benefit immensely from having a high amount of data. Hence, it is sensible to integrate external datasets within the context of zero-shot learning to improve its performance. However, blindly using external datasets may break the zero-shot learning premise, *i.e.* that source and target categories should not overlap. In this work, we have proposed an objective metric that defines which source categories may constitute a synonym of a target category. By pruning these categories from the source set, we honor the zero-shot learning premise. We evaluate this approach in the context of action recognition, and show that adding external data still helps considerably to improve the accuracy of zero-shot learning, even after removing all the similar categories from the source datasets.

Acknowledgements

The research leading to this results has been partially funded by the German Federal Ministry of Education and Research (BMBF) within the PAKoS project.

References

1. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Conference on Computer Vision and Pattern Recognition (2015)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Conference on Computer Vision and Pattern Recognition (2017)
3. Kuehne, H., Jhuang, H., Stiefelhagen, R., Serre, T.: Hmdb51: A large video database for human motion recognition. In: High Performance Computing in Science and Engineering, pp. 571–582. Springer (2013)
4. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
5. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (2013)
6. Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S., Dean, J.: Zero-shot learning by convex combination of semantic embeddings. arXiv preprint arXiv:1312.5650 (2013)
7. Ohn-Bar, E., Trivedi, M.M.: Looking at humans in the age of self-driving and highly automated vehicles. IEEE Transactions on Intelligent Vehicles **1**(1), 90–104 (2016)
8. Poppe, R.: A survey on vision-based human action recognition. Image and vision computing **28**(6), 976–990 (2010)
9. Qin, J., Liu, L., Shao, L., Shen, F., Ni, B., Chen, J., Wang, Y.: Zero-shot action recognition with error-correcting output codes. In: Conference on Computer Vision and Pattern Recognition (2017)
10. Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G., He, K.: Data distillation: Towards omni-supervised learning. arXiv preprint arXiv:1712.04440 (2017)

11. Roitberg, A., Al-Halah, Z., Stiefelhagen, R.: Informed Democracy: Voting-based Novelty Detection for Action Recognition. In: British Machine Vision Conference (BMVC). Newcastle upon Tyne, UK (September 2018)
12. Roitberg, A., Somani, N., Perzylo, A., Rickert, M., Knoll, A.: Multimodal human activity recognition for industrial manufacturing processes in robotic workcells. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. pp. 259–266. ACM (2015)
13. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: International Conference on Computer Vision (2013)
14. Wang, Q., Chen, K.: Zero-shot visual recognition via bidirectional latent embedding. International Journal of Computer Vision **124**(3), 356–383 (2017)
15. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. arXiv preprint arXiv:1703.04394 (2017)
16. Xu, X., Hospedales, T., Gong, S.: Semantic embedding space for zero-shot action recognition. In: International Conference on Image Processing (2015)
17. Xu, X., Hospedales, T., Gong, S.: Transductive zero-shot action recognition by word-vector embedding. International Journal of Computer Vision pp. 1–25 (2017)
18. Zhu, Y., Long, Y., Guan, Y., Newsam, S., Shao, L.: Towards universal representation for unseen action recognition. In: Conference on Computer Vision and Pattern Recognition (2018)