2021

Fairer Evaluation of Zero Shot Action Recognition in Videos

Kaiqiang Huang
d14122793@mytudublin.ie

Sarah Jane Delany
Technological University Dublin, sarahjane.delany@tudublin.ie

Susan Mckeever
Technological University Dublin, susan.mckeever@tudublin.ie

Recommended Citation

Huang, K.; Delany, S. and Mckeever, S. (2021). Fairer Evaluation of Zero Shot Action Recognition in Videos. In Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP, ISBN 978-989-758-488-6; ISSN 2184-4321, pages 206-215. DOI: 10.5220/0010324402060215

Fairer Evaluation of Zero Shot Action Recognition in Videos

Kaiqiang Huang, Sarah Jane Delany and Susan Mckeever
Technological University Dublin, Grangegorman, Dublin, Ireland
kaiqiang.huang@tudublin.ie, sarahjane.delany@tudublin.ie, susan.mckeever@tudublin.ie

Keywords: Human Action Recognition, Zero Shot Learning

Abstract: Zero-shot learning (ZSL) for human action recognition (HAR) aims to recognise video action classes that have never been seen during model training. This is achieved by building mappings between visual and semantic embeddings. These visual embeddings are typically provided via a pre-trained deep neural network (DNN). The premise of ZSL is that the training and testing classes should be disjoint. In the parallel domain of ZSL for image input, the widespread poor evaluation protocol of pre-training on ZSL test classes has been highlighted. This is akin to providing a sneak preview of the evaluation classes. In this work, we investigate the extent to which this evaluation protocol has been used in ZSL for human action recognition research work. We show that in the field of ZSL for HAR, accuracies for overlapping classes are being boosted by between 5.75% to 51.94% depending on the use of visual and semantic features as a result of this flawed evaluation protocol. To assist other researchers in avoiding this problem in the future, we provide annotated versions of the relevant benchmark ZSL test datasets in the HAR field: UCF101 and HMDB51 datasets - highlighting overlaps to pre-training datasets in the field.

1 INTRODUCTION

Over the last decade, the problem of how to identify human actions in videos (such as 'running') has been addressed via a variety of supervised learning-based methods (Feichtenhofer et al., 2016; Hara et al., 2018; Karpathy et al., 2014; Simonyan and Zisserman, 2014; Wang and Schmid, 2013). A limit of these methods, as per all supervised learning, is their inability to identify new classes outside of those in the model training data. Besides, collecting large-scale annotated human action datasets and retraining classification models can be extremely expensive. To address this problem, zero-shot learning (ZSL) has been used in both the image domain (Akata et al., 2013; Akata et al., 2015; Frome et al., 2013; Norouzi et al., 2013; Palatucci et al., 2009; Zhang and Saligrama, 2016) and the video domain (Jain et al., 2015; Liu et al., 2011; Qin et al., 2017; Wang and Chen, 2017b; Wang and Chen, 2017c; Xu et al., 2015; Xu et al., 2017; Xu et al., 2016) to classify novel instances which were not available in the training process. As shown in Fig.1, visual feature representations are extracted from action videos as visual embeddings, and a mapping function is learned between visual embedding and the corresponding semantic embedding in the training stage. In the ZSL testing phase, visual features for unseen testing instances are inputted to the learned mapping function, the semantic embedding is acquired and the action class with the closest distance by nearest neighbour search in the semantic embedding space is assigned (Junior et al., 2019). The issue we address in this paper is the poor evaluation practice of allowing classes used in the pre-training dataset to also be used as classes in the ZSL testing (supposed to be unseen) classes. These overlapping classes challenge the "unseen" nature of the ZSL test classes. In this paper, we answer the following questions (1) To what extent is this evaluation protocol applied in the ZSL for HAR research domain to date? (2) Does the overlap of classes in ZSL testing and pre-training classes boost ZSL accuracies, thus inflating ZSL results? In addition to reporting on the issue, we publish corrected non-overlapping versions of benchmarks ZSL test datasets in the HAR field for use by other researchers.

The rest of this paper is structured as follows. In Section 2, the background for zero-shot action recognition (ZSAR) is described, including visual and semantic embedding, and ZSL methods. In Section 3, we explain our methods for evaluating the impact of overlapping classes on benchmark pre-trained models and ZSL datasets in the HAR field. In Section 4, experimental implementations are explained in more

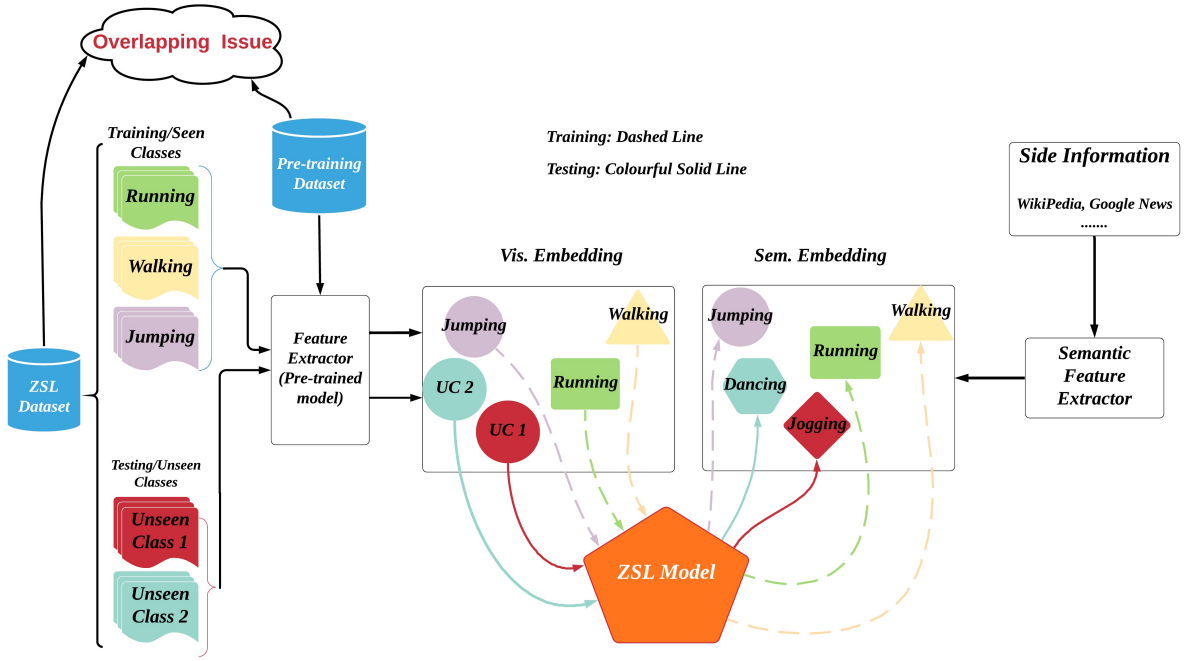

Figure 1: General ZSL Process for HAR. Note that, the colourful solid arrow lines indicate the ZSL training phrase and the dashed arrow lines indicate the ZSL testing phrase.

detail. In Section 5, the results and findings are demonstrated and explained. Finally, the conclusion is drawn in Section 6.

2 BACKGROUND & RELATED WORK

The aim of ZSL is to break away from the traditional limitation of supervised learning - to build a classifier that can identify classes outside those used in training. The ZSL training stage maps instances of trained classes to a semantic space, where the wider semantic space has representations of both training and new unseen classes. In the case of human action recognition, each instance is a video - each represented as a visual embedding that is extracted from a pre-trained model. During the ZSL training stage, visual embedding of instances are mapped to their equivalent semantic label embedding, as shown in Figure 1. This learned mapping is then used to map visual embeddings of new unseen zero-shot test classes into the semantic space for identification of the new class. Next, we look at both visual and semantic embedding in more detail in order to understand the background for our work.

Visual Embedding There are two principal approaches to generating embeddings for the visual representations of actions for ZSL (1) Hand-crafted features (Wang and Schmid, 2013) and (2) Deep features (Carreira and Zisserman, 2017; Hara et al., 2018; Tran et al., 2015). The principal approach to hand-crafted features, Improved Dense Trajectory (IDT) (Wang and Schmid, 2013) has been successfully applied to represent visual information for videos of actions in early ZSL works. It consists of four descriptors, computed through the tracked trajectories based on the movement of detected interest points in a video. More recently, visual representations have been obtained from pre-trained DNN-based models (so called "deep features") in the ZSL for action recognition field. As shown in Table 1, there are three deep feature visual representations used in the field, namely C3D, I3D and 3D-ResNet. C3D (Tran et al., 2015) is a deep 3-dimensional convolutional neural network (3D CNNs) that learns spatio-temporal representations for videos. Due to the highly computationally intense training in 3D CNNs, I3D (Carreira and Zisserman, 2017) was proposed. Instead of a single 3D network, two different 3D networks are used, containing RGB and flow as two architectural streams, to represent appearance and optical flow, respectively. A third pre-trained network, 3D-ResNet (Hara et al., 2018) was proposed by building a deeper 3D CNN to learn spatio-temporal informa-

Table 1: Visual features and pre-training datasets for human actions (Papers in bold indicate that overlap occurs between pre-training and ZSL test classes)

Visual Features	Feature Types	Pre-training Datasets	Papers Used
IDT	Hand Crafted	N/A	(Gan et al., 2015; Guadarrama et al., 2013) (Jain et al., 2015; Qin et al., 2017) (Xu et al., 2015; Wang and Chen, 2017c) (Xu et al., 2017; Xu et al., 2016) (Zhang and Peng, 2018)
C3D	Deep	Sports-1M	(Brattoli et al., 2020) **(Mandal et al., 2019; Mishra et al., 2020)** **(Mishra et al., 2018; Wang and Chen, 2017b)** **(Wang and Chen, 2017a; Tian et al., 2019)** **(Wang and Chen, 2017c; Zhang et al., 2018)**
I3D	Deep	Kinetics-400	**(Mandal et al., 2019; Narayan et al., 2020)** **(Piergiovanni and Ryoo, 2018)** (Roitberg et al., 2018)
3D-ResNet	Deep	Kinetics-400	**(Liu et al., 2020)**

tion from actions. As shown in Table 1, each of these three deep feature representations is underpinned by pre-trained models that have been trained on action datasets: Sports-1M for C3D, and Kinetics-400 for both I3D and 3D-ResNet. Consequently, where these pre-trained models are used to provide deep visual feature representations for action instances in videos, explicit or semantic overlaps between the pre-training classes with those of the zero-shot test classes should be avoided, if a ZSL approach is claimed. Table 1 also indicates which research works use these representations. Papers in bold indicate that some degree of overlap occurs between pre-training and ZSL test classes in the evaluation protocol. Note that IDT does not involve a pre-trained model so the issue of overlapping of pre-training and ZSL test classes is not relevant. There are two leading benchmark ZSL datasets in the ZSL action recognition field: (1) UCF101 (Soomro et al., 2012) and (2) HMDB51 (Kuehne et al., 2011). Our identification of overlapping classes between these two ZSL datasets the pre-training datasets is described further in Section 3.1.

Semantic Label Embedding There are two main approaches to representing the semantic label space in ZSL for human action recognition - (1) attribute-based methods (Liu et al., 2011; Wang and Chen, 2017c) and (2) word embedding based methods (Mikolov et al., 2013). Attributed-based methods (i.e. annotated class-level attribute) involve the manual annotation of visual attributes to each action class. A Golf-Swing action class would include visual attributes such as single-leg motion and arm-over-shoulder motion for example. Word embedding is an alternative approach to attribute-based, removing the manual annotation requirement and it is widely used to represent semantics for action classes in recent work (Brattoli et al., 2020; Gao et al., 2019; Jain et al., 2015; Mandal et al., 2019; Roitberg et al., 2018; Wang and Chen, 2017c; Xu et al., 2015; Xu et al., 2016; Zhu et al., 2018).

ZSL Techniques for Human Action Recognition A comprehensive overview of ZSL approaches for human action recognition field is provided by the work (Junior et al., 2019). We briefly review them here. Early work for ZSL (Liu et al., 2011) makes use of human-annotated attributes as semantics to learn a mapping function from visual embedding to attributes embedding using an SVM model that can decide the significance of each attribute for each action class. LatEm (Xian et al., 2016) introduced a piece-wise linear compatibility as non-linearity to learn the relationship between both embeddings by the latent variables. And each latent variable is encoded for different visual properties of the input data. Instead of projecting the visual embedding to the semantic embedding, BiDiLEL (Wang and Chen, 2017c) projected both visual and semantic embeddings into a low-dimensional latent embedding in order to preserve the intrinsic relationships between them. The SynC approach (Changpinyo et al., 2016) learned a projection between semantic label embeddings and a classifier model space. In the classifier model space, a weighted bipartite graph is created by the training classes and a set of "phantom" classes that are synthesised to align semantic embedding and classifier model spaces by minimising distortion error. ConSE (Norouzi et al., 2013) is a two-stage approach: First, it learns the probability for a training action instance

belonging to one of the seen classes. Then, it projects the visual embedding into the semantic embedding by taking the convex combination of the top T most likely seen classes. Finally, the class with the closest cosine distance in the semantic embedding space is predicted as the unseen class result. Note that, the ZSL methods of ConSE, SynC and LatEm were originally proposed for ZSL image classification, but the latter work extended them to ZSL for human action recognition (Liu et al., 2019).

With regards to the issue of overlapping classes between pre-training and ZSL testing datasets, the work flagged the problem for ZSL for images (Xian et al., 2017). However, the issue has had limited recognition in the domain of ZSL for action recognition. The work by (Roitberg et al., 2018) proposed an evaluation procedure that enables fair use of external data by exploring how semantically similar classes in the ZSL training and test sets can impact on ZSL for human action recognition evaluation (i.e. Kinetics-400 for training and HMDB51 for testing). Similarly, work by (Brattoli et al., 2020) removing overlapped classes between ZSL training and testing classes for their ZSL work. However, neither work addresses the issue of overlap between pre-training model classes and ZSL test classes. In the next section, we explain our approach to quantifying the impact of the pre-train/test overlapping classes issue for ZSL human action recognition evaluation.

3 METHODS

We wish to measure whether classes that appear in both the pre-trained model training set and the ZSL test sets are obtaining boosted recognition accuracies. We will focus on the two leading ZSL datasets in the field, UCF101 and HMDB51. Classes in these two benchmark datasets that overlap with the classes from the two pre-training datasets (Sports-1M and Kinetics-400) are the root of our problem. First, we identify and tag these overlapping classes. An overlap is defined as two classes having the same name (i.e. *SkyDiving* and *Knitting* in ZSL dataset UCF101 and pre-training dataset Kinetics-400) or a semantically similar name (i.e. *JumpRope* in UCF101 and *Skipping rope* in Kinetics-400). For the benefit of other researchers, we publish these tagged benchmark datasets [1]. Having flagged the overlaps, we can now create ZSL training/test sets, where we can control the proportions of overlapping classes in our ZSL test sets, in order to evaluate their impact on ZSL accuracies. We will test accuracies using the three pre-trained models for producing visual deep features shown in Table 1. We also use a fourth baseline non-deep approach, IDT, which does not involve any pre-trained models (and therefore no overlapping issue). We will measure the ZSL test accuracies of both overlapping and non-overlapping classes, for three pre-trained models and one baseline approach.

3.1 Creating ZSL Training/Test Splits

The next step is to create suitable ZSL training/test splits for our experiments. The UCF101 dataset has 101 action categories with a total of 13320 videos collected from YouTube. The HMDB51 dataset contains 51 action categories with 6670 videos in total collected from commercial videos and YouTube. For our work, we need to divide our two ZSL datasets into training and testing splits. In terms of training/test split size, we note the previous approaches in ZSL for human action recognition (Wang and Chen, 2017c; Xu et al., 2015): (1) UCF101: 101 classes split 51/50 classes for ZSL training/test respectively. (2) HMDB51: 51 classes split into 26/25 classes for ZSL training/test respectively. These previous training/test splits were created by random selection of classes for training and test, without any controls of overlapping classes. We use these dataset split sizes for our work. As we know which classes in our ZSL datasets are overlapping with pre-training datasets (shown in Table 2), we can now create our own controlled ZSL test sets consisting of (1) genuinely unseen classes (not overlapping with pre-trained model - termed 'true' unseen classes (TUCs) and (2) problematic overlapping unseen classes (OUCs).

We wish to investigate ZSL accuracies for both TUCs and OUCs, to determine whether there is a general pattern of OUCs achieving higher ZSL accuracies across the various ZSL techniques. Table 2 shows the combinations of pre-training datasets and ZSL datasets that we use in our experiments. Looking at the first row as an example, for models pre-trained on Kinetics-400 and ZSL dataset UCF101 (which overlaps on 61 classes), we split UCF101 into class splits of 51 for ZSL training, and 50 for ZSL testing, as per previous ZSL split sizes. For our ZSL test set of 50 classes, we use equal proportions of TUCs and OUCs, using random selection from the pools of overlapping and non-overlapping classes. The remaining classes form our ZSL training set.

[1] https://github.com/kaiqiangh/fairer-eval-zsar

Table 2: Our splits and training/testing set for UCF101 and HMDB51

Index	Pre-training Datasets	Pre-trained Models	ZSL Datasets (total classes)	Overlapping Classes	Non-overlapping Classes	Training	Testing (OUC + TUC)
OS1	Kinetics-400	I3D 3D-ResNet	UCF101 (101)	61	40	51	50(25+25)
OS2	Sports-1M	C3D	UCF101 (101)	36	65	51	50(25+25)
OS3	Kinetics-400	I3D 3D-ResNet	HMDB51 (51)	28	23	26	25(12+13)
OS4	Sports-1M	C3D	HMDB51 (51)	17	34	26	25(12+13)

3.2 Evaluation Metrics

We aim to measure whether overlapping classes (OUCs) can achieve higher accuracies than non-overlapping classes (TUCs) in the ZSL test sets. Class accuracy is a standard metric in the ZSL field. We want to measure accuracies for each of OUC and TUC separately, so we average per-class accuracies as per the following equation (Xian et al., 2017) using TUC as the example:

$$ACC_{TUC} = \frac{1}{N_{TUC}} \sum_{C=1}^{N_{TUC}} \frac{\#\ correct\ \ predictions\ in\ Class\ C}{\#\ instances\ in\ Class\ C} \quad (1)$$

In addition to comparing absolute differences in class accuracies, we also want to be able to compare across ZSL methods and pre-trained models, so we need a normalised metric. Specifically, we wish to quantify the boost or gain in accuracy of overlapping classes in comparison to true unseen classes. We define an Overlap Gain metric (**OverlapGain**), which measures the difference between OUC and TUC relative to the average accuracy for the ZSL test set. The equation is defined as:

$$OverlapGain = \frac{ACC_{OUC} - ACC_{TUC}}{Average(ACC_{OUC}, ACC_{TUC}))} \quad (2)$$

A positive *OverlapGain* value indicates that the average accuracy for OUC is higher than TUCs, indicating a boost for overlapping classes.

4 EXPERIMENTAL SETUP

In this section, we explain the experimental configurations for testing combinations of pre-trained models and ZSL datasets across a variety of ZSL methods. We also explain the visual and semantic features used in experiments and the hyper-parameter settings used in our ZSL implementations.

4.1 Experiments and Baseline

Table 3 shows five experimental configurations we will carry out, based on testing three deep pre-trained models and the baseline approach, against the two benchmark ZSL datasets. Overlap class gain results from deep features (i.e C3D and I3D) will be compared to the IDT baseline. For each of these four experimental configurations, we test the leading ZSL methods described in Section 2. In the fifth experiment configuration, we are reproducing the results for one specific ZSL method, BiDiLEL (Wang and Chen, 2017c) as in this work, the actual ZSL training/ test splits were published with the work, allowing us to use both their randomly selected ZSL test splits and compare to our controlled ZSL test splits. This dataset split information is not available for other ZSL work, so we will do this comparison for BiDiLEL only. As per previous work (Wang and Chen, 2017c; Xu et al., 2017), we use 30 training/test iterations for each evaluation.

4.2 Visual and Semantic Embeddings

We adopt the off-the-shelf IDT and C3D provided by (Wang and Chen, 2017c), and I3D provided by (Mandal et al., 2019). Our baseline **IDT** feature representation approach (Wang and Schmid, 2013) contains four different descriptors: HoG, HoF, MBHx and MBHy. The video representations are generated by the fisher vector derived from a Gaussian mixture model with 256 components. To reduce the computational cost, PCA is applied to reduce dimensions to 3000 for each descriptor, and all descriptors are simply concatenated to obtain a final 12000-dimensional vector. For **C3D** (Tran et al., 2015), a video is divided into 16-frames segments and there is an overlap of eight frames on two consecutive segments. As a result, the *fc6* activation is first extracted for all the segments and then averaged to form a 4096-dimensional representation for a video. **I3D** (Carreira and Zisserman, 2017) contains RGB and Inflated 3D networks to generate appearance and flow features from

Table 3: Experimental Configurations. Att. denotes human-annotated attribute and WV denotes Word2Vec.

Experiments	Data Splits Index	Pre-training Datasets	ZSL Datasets	Semantics	Visual Features	ZSL Methods
1	OS1	Kinetics-400	UCF101	Att. WV	IDT (Baseline) I3D 3D-ResNet	ConSE BiDiLEL LatEm SynC^{ovo} SynC^{struct}
2	OS2	Sports-1M	UCF101	Att. WV	IDT (Baseline) C3D	
3	OS3	Kinetics-400	HMDB51	WV	IDT (Baseline) I3D 3D-ResNet	
4	OS4	Sports-1M	HMDB51	WV	IDT(Baseline) C3D	
5	OS2	Sports-1M	UCF101	Att. WV	IDT(Baseline) C3D	BiDiLEL
	OS4		HMDB51	WV		

Mixed_5c layer. For each video instance, the outputs from *Mixed_5c* layer for both networks are averaged through the temporal dimension, pooled in the spatial dimension and then flattened to obtain a 4096-dimensional vector for appearance and flow features, respectively. In the end, both appearance and flow features are concatenated to represent video with 8192-dimensional vector. **3D-ResNet** (Hara et al., 2018) has been shown to achieve the impressive performance in deeper network style for generic action recognition, and ResNext-101 model achieved the optimal results on UCF101 and HMDB51. Each video is divided into 16 frames as a segment, and each frame is resized to 112×112, resulting in the size of $3channels \times 16frames \times 112pixels \times 112pixels$ with 50% possibility of horizontal flipping for each segment as input. Next, the processed video is fed into ResNext-101 model for feature extraction by taking the output of a 3D global average pooling layer, of size 2048, as final video representation.

For semantic embedding, we use both annotated attributes and Word2Vec. The UCF101 has 115 binary attributes, defined and provided by (Liu et al., 2011). As shown in experimental configurations 3 and 4, we do not evaluate annotated attributes as semantics for HMDB51 as there are no published attributes sources for this dataset. With regards to Word2Vec (Mikolov et al., 2013), a skip-gram model that was trained on a large-scale text corpus (i.e. Google News Dataset) is used to deliver a 300-dimensional vector for each action class label for both UCF101 and HMDB51, provided by (Wang and Chen, 2017c).

4.3 ZSL Implementation Details

We adopt off-the-shelf implementations for four leading ZSL methods in our experiments: ConSE[2], BiDiLEL[3], LatEm[4] and SynC[5]. In ConSE, the hyper-parameters of C and T are set to 0.1 and 5, respectively. In BiDiLEL, we do not tune any hyper-parameters, but directly use the optimal ones from the original work. As a consequence, there are three hyper-parameters α, k_G and d_y, indicating the trade-off factor applied to the regularisation for graph construction, the number of nearest neighbours for graph construction, and the dimensionality of the learned latent space, respectively, and their values are set to 10, 10 and 150 for α, k_G and d_y, respectively. Next, similar hyper-parameter settings for LatEm, the learning rate for SGD (θ), the number of epoch for SGD and the number of embeddings to learn (K) are set to 0.1, 150 and 10, respectively. Lastly, for SynC, the balance coefficient in the objective function to learn the base classifiers (λ) and a factor to model correlation between a real class and a phantom class by attributes (σ) are set to 2^{-10} and 1, respectively. In addition, all experiments are conducted in MATLAB.

[2] https://github.com/pujols/zero-shot-learning

[3] http://staff.cs.manchester.ac.uk/~kechen/BiDiLEL/

[4] https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/zero-shot-learning/latent-embeddings-for-zero-shot-classification

[5] https://github.com/pujols/zero-shot-learning

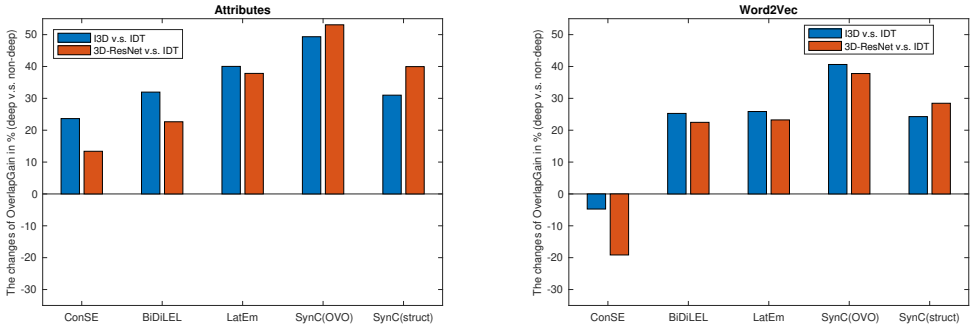

Figure 2: Configuration 1: the differences of *OverlapGain* in % by comparing deep features (I3D, 3D-ResNet) to baseline non-deep feature (IDT) with regard to the case of UCF101 with Kinetics-400. Left and right figures shows the results for Attributes and Word2Vec semantic embeddings approaches, respectively.

5 RESULTS AND DISCUSSION

In this section, we present and discuss the experimental results for the configurations shown in Table 3. For each configuration, we measure the difference in accuracies of overlapping (OUCs) versus true unseen classes (TUCs) for the pre-trained models in question. We compare this difference to that of the baseline approach (IDT).

5.1 Gains to Overlapping Unseen Classes: Experiments 1 - 4

For all figures in this section, the X-axis indicates the ZSL methods evaluated and the Y-axis indicates the differences of *OverlapGain*, described in Section 3, using a pre-trained model over the baseline non-deep feature i.e. $OverlapGain_{i3d} - OverlapGain_{idt}$.

Fig.2 shows the results for experimental configuration 1 - the increase in the differences in *OverlapGain* for deep features against the baseline non-deep feature (IDT) where we treat UCF101 as ZSL dataset and models pre-trained on Kinetics-400 (I3D and 3D-ResNet). Looking at the configuration on the left-hand side that uses the attributes as semantic space, the gains over the baseline range for the pre-trained deep models range from just over 10% (ConSE for 3D-ResNet) to just over 50% for SynC for 3D-ResNet). When using Word2Vec as shown on the right-hand graph, all pre-trained models have a positive gain for overlapping classes, in comparison to the non-deep baseline, with the exception of the ConSE method. Specifically, for the performances of using attributes, the boosts of 18.52%, 27.30%, 38.93%, 51.21% and 35.48% are gained when evaluating on deep features against non-deep feature for ConSE, BiDiLEL, LatEm, SynC^{ovo} and SynC^{struct}, respectively. Additionally, for the performances of using Word2Vec, ConSE obtains a decrease of -11.94%, but other methods still gain the boosts of 23.87%, 24.55%, 39.21% and 26.36%. As can be seen, attribute-based semantic spaces show a stronger boost to overlapping classes accuracies than Word2Vec. We suggest that human-annotated attributes are better for transferring semantic knowledge than Word2Vec in ZSL for human action recognition as attributes are fully supervised by human experts but Word2Vec is produced in an unsupervised way.

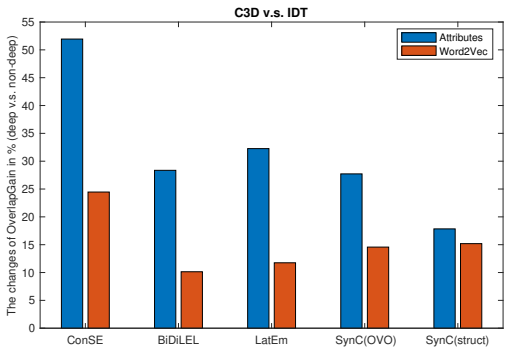

Figure 3: Configuration 2: the differences of *OverlapGain* in % regarding the case of UCF101 with Sports-1M, for pre-trained model C3D against baseline IDT along with attribute and Word2Vec as semantic embedding.

Fig.3 shows the results for experimental configuration 2, where pre-trained model C3D is used for generating deep visual representations, pre-trained on the Sports-1M dataset. For the ZSL methods of ConSE, BiDiLEL, LatEm, SynC^{ovo} and SynC^{struct}, the boost to overlapping class is higher than the baseline IDT by 51.94%, 28.36%, 32.26%, 27.72% and 17.84% when using attribute, and by 24.46%, 10.14%, 11.75%, 14.57% and 15.19% when using Word2Vec. Similar to experiment 1, the boost to overlap class accuracies is higher for attribute-based semantic spaces than Word2Vec.

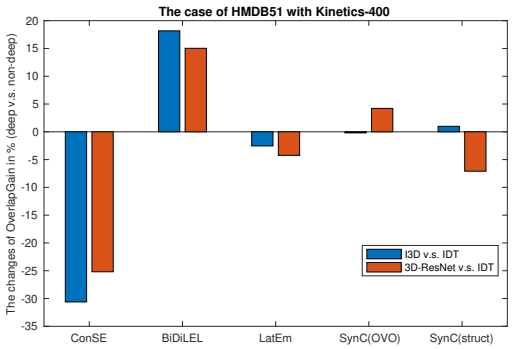

Figure 4: Configuration 3: the differences of *OverlapGain* in % regarding the case of HMDB51 with Kinetics-400.

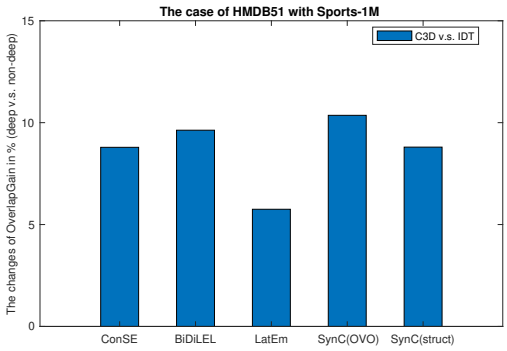

Figure 5: Configuration 4: the differences of *OverlapGain* in % regarding the case of HMDB51 with Sports-1M.

Fig.4 shows the results for experimental configuration 3. For ZSL dataset HMDB51 with Kinetics-400 for pre-training, only the BiDiLEL method shows a gain for overlapped classes against non-lapped over the baseline. From the perspective of ZSL methods, only BiDiLEL aims to project both visual and semantic embeddings to a low-dimensional latent embedding, but other methods aim to project visual to semantic embeddings (i.e. ConSE, SynC) or directly learn compatibility between embeddings (i.e. LatEm). We suggest that the influences by OUCs could be alleviated when predicting ZSL test classes in a low-dimensional space. Based on the observations in experiment 3, we find that it does not fully support our hypothesis. The differences of *OverlapGain* from deep to non-deep features are slightly negative for most ZSL methods, and clearly negative for ConSE, indicating that overlapping classes are not gaining any advantage over true unseen classes when using I3D and 3D-ResNet in this case. We note that the videos in the ZSL dataset HMDB51 and the Kinetics-400 are noisy, in comparison to the UCF101 and Sports-1M datasets. In HMDB51, the instances have non-clean backgrounds and videos contain multiple actors. Since this particular combination of ZSL dataset and pre-trained dataset does not exhibit a ZSL gain from having a sneak preview of unseen test classes, the results suggest there is insufficient overlap in a Kinetics-400 class video to the equivalent video class in HMDB51 to provide a boost.

Fig.5 shows the results for experimental configuration 4. For ZSL datasets HMDB51 with pre-training dataset Sports-1M, overlapping classes gain from 5.75% up to 10.36% across all ZSL methods using C3D against baseline IDT. To summarise, three of our four experimental configurations demonstrate a strong gain for overlapping classes, with the exception of some ZSL methods in configuration 3 (Kinetics400 pre-training and HMDB51 ZSL test).

5.2 Comparison to Previous Work: Experiment 5

In experimental configuration 5, we examine the impact of using equal split overlapping versus non-overlapping classes in our ZSL test set - versus results gained by using previous ZSL training/test in the field. We analysed the splits used previously by (Wang and Chen, 2017c; Xu et al., 2015) as previous splits (PS). In their work, We determined that the number of classes for OUC and TUC occupy 35.4% and 64.5% respectively in UCF101, and 32.67% and 67.33% in HMDB51. In our splits (OS), we have UCF101 (50%/50%), and 48% OUC and 52% TUC in HMDB51 - the latter mismatch is due to the number of testing classes being odd. The key point is that the previous splits from published works have a higher proportion of TUCs than ours, and thus we expect that ZSL class accuracies will be higher as a result of using more overlapping classes in our ZSL test sets. Table 4 shows the results as the mean of per-class average accuracies when using deep feature (C3D) and non-deep feature (IDT) along with different semantic representations under previous splits and our splits for both ZSL datasets.

As can be seen in Table 4, there is no significant difference between previous splits and our splits in UCF101 when we use the baseline IDT for visual representation. This is as expected as the concept of overlapping classes is not relevant here. In contrast, when using pre-trained model C3D, ZSL accuracies are higher using our splits than previous research splits, showing an accuracy increases of 11.22% and 3.70% for evaluating attributes and Word2Vec in UCF101, respectively, along with the 14.6% rise of the number of OUCs in our testing sets against previous splits.

Similarly for HMDB51, no significant difference

Table 4: Configuration 5: the comparison between previous splits and our splits for both UCF101 and HMDB51 when evaluating BiDiLEL. The results report average per-class accuracy in %. Note that, Att. denotes attribute and WV denotes Word2Vec.

Splits	Vis. Rep.	UCF101		HMDB51
		Att.	WV	WV
PS	IDT	16.6	15.4	16.4
OS	IDT	16.5	15.1	16.0
PS	C3D	20.5	18.9	18.6
OS	C3D	**22.8**	**19.6**	**20.6**

is found between previous splits and our splits when using IDT as expected. When using C3D pre-trained model, there is 10.75% boost gained for Word2Vec with the 15.33% rise of the number of OUCs in our testing sets from HMDB51 compared to previous splits. Summarising the results from experimental configuration 5, the greater the proportion of overlapping classes, the higher the ZSL class accuracies obtained, supporting the case that overlapping classes between pre-training and ZSL test sets are distorting ZSL results.

All our discussions so far have focused on the overlapping of ZSL test classes and pre-trained datasets as an issue. And indeed, the lack of allowance for this in evaluation results is an issue due to the unrecognised inflation of zero-shot capabilities. A counter-intuitive point to consider is that better accuracies for 'unseen' overlapping classes may highlight opportunities. If we can boost accuracies via the embedded knowledge in our pre-trained model, we can consciously factor suitable knowledge-rich pre-trained models into ZSL frameworks in order to provide a stronger basis for supporting the zero-shot case.

6 CONCLUSION

We have examined the extent and evaluation impact of zero-shot 'unseen' classes appearing in underlying pre-trained models for feature representation in ZSL human action recognition. Our work covered four leading ZSL techniques for the two benchmark ZSL datasets in the field: UCF101 and HMDB51. Three widely used pre-trained models were tested: C3D, I3D and 3D-ResNet. From the overall perspective, our results showed a gain to overlapping classes for the UCF101 dataset from 17.84% up to 51.94% by using attribute, and a gain from 10.14% up to 39.21% by using Word2Vec. Also, we note a less conclusive pattern for HMDB51 with lower gains for the overlapping classes in the case of HMDB51 with Sports-1M (the gains from 5.75% up to 10.36%). For HMDB51 with Kinetics-400, some ZSL methods showed no overlap gain, which may be linked to the noisy and complex natures of the instances. We publish the tagged version of the two benchmark ZSL datasets for use by researchers in the field who wish to avoid breaking the premise of ZSL.

As future work, we aim to focus on the task of generalised zero-shot learning (GZSL), which provides an additional challenging of including both seen and unseen classes in the ZSL test stage - a challenging problem of predictions biased towards seen classes. We also highlight an alternative view of the work here - the choice of suitable pre-trained models could support the planned boosting of zero-shot accuracies if carried out in a transparent and deliberate way.

ACKNOWLEDGEMENTS

This project is funded under the Fiosraigh Scholarship of Technological University Dublin.

REFERENCES

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2013). Label-embedding for attribute-based classification. In *Proceedings of CVPR*, pages 819–826.

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2015). Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438.

Brattoli, B., Tighe, J., Zhdanov, F., Perona, P., and Chalupka, K. (2020). Rethinking zero-shot video classification: End-to-end training for realistic applications. In *Proceedings of IEEE CVPR*, pages 4613–4623.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Changpinyo, S., Chao, W.-L., Gong, B., and Sha, F. (2016). Synthesized classifiers for zero-shot learning. In *Proceedings of CVPR*, pages 5327–5336.

Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *Proceedings of CVPR*, pages 1933–1941.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129.

Gan, C., Lin, M., Yang, Y., Zhuang, Y., and Hauptmann, A. G. (2015). Exploring semantic inter-class relationships (sir) for zero-shot action recognition. In *Proceedings of the National Conference on Artificial Intelligence*.

Gao, J., Zhang, T., and Xu, C. (2019). I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *Proceedings of the AAAI*, volume 33, pages 8303–8311.

Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., and Saenko, K. (2013). Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the IEEE ICCV*, pages 2712–2719.

Hara, K., Kataoka, H., and Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of CVPR*, pages 6546–6555.

Jain, M., van Gemert, J. C., Mensink, T., and Snoek, C. G. (2015). Objects2action: Classifying and localizing actions without any video example. In *Proceedings of the IEEE ICCV*, pages 4588–4596.

Junior, V. L. E., Pedrini, H., and Menotti, D. (2019). Zero-shot action recognition in videos: A survey. *arXiv preprint arXiv:1909.06423*.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of IEEE CVPR*, pages 1725–1732.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE.

Liu, H., Yao, L., Zheng, Q., Luo, M., Zhao, H., and Lyu, Y. (2020). Dual-stream generative adversarial networks for distributionally robust zero-shot learning. *Information Sciences*, 519:407–422.

Liu, J., Kuipers, B., and Savarese, S. (2011). Recognizing human actions by attributes. In *CVPR 2011*, pages 3337–3344. IEEE.

Liu, K., Liu, W., Ma, H., Huang, W., and Dong, X. (2019). Generalized zero-shot learning for action recognition with web-scale video data. *WWW*, 22(2):807–824.

Mandal, D., Narayan, S., Dwivedi, S. K., Gupta, V., Ahmed, S., Khan, F. S., and Shao, L. (2019). Out-of-distribution detection for generalized zero-shot action recognition. In *Proceedings of CVPR*, pages 9985–9993.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mishra, A., Pandey, A., and Murthy, H. A. (2020). Zero-shot learning for action recognition using synthesized features. *Neurocomputing*.

Mishra, A., Verma, V. K., Reddy, M. S. K., Arulkumar, S., Rai, P., and Mittal, A. (2018). A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE Winter Conference on WACV*, pages 372–380. IEEE.

Narayan, S., Gupta, A., Khan, F. S., Snoek, C. G., and Shao, L. (2020). Latent embedding feedback and discriminative features for zero-shot classification. *arXiv preprint arXiv:2003.07833*.

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G. S., and Dean, J. (2013). Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*.

Palatucci, M., Pomerleau, D., Hinton, G. E., and Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. In *Advances in NIPS*, pages 1410–1418.

Piergiovanni, A. and Ryoo, M. S. (2018). Learning shared multimodal embeddings with unpaired data. *CoRR*.

Qin, J., Liu, L., Shao, L., Shen, F., Ni, B., Chen, J., and Wang, Y. (2017). Zero-shot action recognition with error-correcting output codes. In *Proceedings of the IEEE Conference on CVPR*, pages 2833–2842.

Roitberg, A., Martinez, M., Haurilet, M., and Stiefelhagen, R. (2018). Towards a fair evaluation of zero-shot action recognition using external data. In *ECCV*, pages 0–0.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in NIPS*, pages 568–576.

Soomro, K., Zamir, A. R., and Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tian, Y., Kong, Y., Ruan, Q., An, G., and Fu, Y. (2019). Aligned dynamic-preserving embedding for zero-shot action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of IEEE ICCV*, pages 3551–3558.

Wang, Q. and Chen, K. (2017a). Alternative semantic representations for zero-shot human action recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 87–102. Springer.

Wang, Q. and Chen, K. (2017b). Multi-label zero-shot human action recognition via joint latent embedding. *arXiv preprint arXiv:1709.05107*.

Wang, Q. and Chen, K. (2017c). Zero-shot visual recognition via bidirectional latent embedding. *IJCV*, 124(3):356–383.

Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., and Schiele, B. (2016). Latent embeddings for zero-shot classification. In *Proceedings of CVPR*, pages 69–77.

Xian, Y., Schiele, B., and Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on CVPR*, pages 4582–4591.

Xu, X., Hospedales, T., and Gong, S. (2015). Semantic embedding space for zero-shot action recognition. In *2015 IEEE ICIP*, pages 63–67. IEEE.

Xu, X., Hospedales, T., and Gong, S. (2017). Transductive zero-shot action recognition by word-vector embedding. *IJCV*, 123(3):309–333.

Xu, X., Hospedales, T. M., and Gong, S. (2016). Multi-task zero-shot action recognition with prioritised data augmentation. In *ECCV*, pages 343–359. Springer.

Zhang, B., Hu, H., and Sha, F. (2018). Cross-modal and hierarchical modeling of video and text. In *Proceedings of ECCV*, pages 374–390.

Zhang, C. and Peng, Y. (2018). Visual data synthesis via gan for zero-shot video classification. *arXiv preprint arXiv:1804.10073*.

Zhang, Z. and Saligrama, V. (2016). Zero-shot learning via joint latent similarity embedding. In *proceedings of CVPR*, pages 6034–6042.

Zhu, Y., Long, Y., Guan, Y., Newsam, S., and Shao, L. (2018). Towards universal representation for unseen action recognition. In *Proceedings of CVPR*, pages 9436–9445.