Zero-Shot Action Recognition with Knowledge Enhanced Generative Adversarial Networks

Kaiqiang Huang, Luis Miralles-Pechuán and Susan Mckeever
Technological University Dublin, Grangegorman, Dublin, Ireland

Keywords: Human Action Recognition, Zero-Shot Learning, Generative Adversarial Networks.

Abstract: Zero-Shot Action Recognition (ZSAR) aims to recognise action classes in videos that have never been seen during model training. In some approaches, ZSAR has been achieved by generating visual features for unseen classes based on the semantic information of the unseen class labels using generative adversarial networks (GANs). Therefore, the problem is converted to standard supervised learning since the unseen visual features are accessible. This approach alleviates the lack of labelled samples of unseen classes. In addition, objects appearing in the action instances could be used to create enriched semantics of action classes and therefore, increase the accuracy of ZSAR. In this paper, we consider using, in addition to the label, objects related to that action label. For example, the objects 'horse' and 'saddle' are highly related to the action 'Horse Riding' and these objects can bring additional semantic meaning. In this work, we aim to improve the GAN-based framework by incorporating object-based semantic information related to the class label with three approaches: replacing the class labels with objects, appending objects to the class, and averaging objects with the class. Then, we evaluate the performance using a subset of the popular dataset UCF101. Our experimental results demonstrate that our approach is valid since when including appropriate objects into the action classes, the baseline is improved by 4.93%.

1 INTRODUCTION

The field of human action recognition (HAR) has drawn substantial attention from the computer vision research community. With the increase in the demand for HAR-based applications in real-world scenarios (e.g. video surveillance (Wang, 2013)), the low volume of available labelled action training data is a challenge when attempting to develop generalised and robust models. Due to the expensive labour work involved, it is difficult to build large-scale labelled video datasets of human actions by collecting adequate video instances with well-defined class-level annotations. Therefore, due to their reliance on the difficult-to-acquire large scale labelled datasets, it is not possible to scale supervised approaches for recognising actions (Huang et al., 2019b).

For this work, we consider the extreme case of data scarcity in the HAR field, termed Zero-Shot Action Recognition (ZSAR). ZSAR aims to recognise unseen action classes by training a classification model using knowledge transferred from other seen action classes with the assistance of semantic information (Mishra et al., 2018; Liu et al., 2019; Mandal et al., 2019). The common type of semantic information can be either annotations of class-level attributes (Liu et al., 2011; Wang and Chen, 2017) or word embedding (Mikolov et al., 2013). The semantic information is then used to embed both seen and unseen classes in fixed-sized and high-dimensional vectors. The general idea behind ZSAR is based on the fact that a human can recognise and detect unseen action classes by matching text-based descriptions of the class with previously learned visual concepts. For example, given the description 'an animal horse-like with black-and-white stripes', a human is able to detect 'zebra' as long as he understands what a horse looks like and what the pattern 'black-and-white stripe' looks like. Generalised ZSAR attempts to recognise video instances from both seen and unseen classes, whereas with conventional ZSAR, the test set contains unseen classes only. In this work, we are exploring a new approach so we consider conventional ZSAR initially. If the results of our approach are positive, we will apply it to generalised ZSAR in the future.

Early approaches to ZSAR largely depend on projection-based methods, which attempt to learn a

Huang, K., Miralles-Pechuán, L. and Mckeever, S.
Zero-Shot Action Recognition with Knowledge Enhanced Generative Adversarial Networks.
DOI: 10.5220/0010717000003063
In *Proceedings of the 13th International Joint Conference on Computational Intelligence (IJCCI 2021)*, pages 254-264
ISBN: 978-989-758-534-0; ISSN: 2184-2825

projection function to map the visual representations of seen classes to their corresponding semantic representations. Then, the learned projection function is used to recognise novel classes by measuring the similarity level between the prototype representations and the predicted representations of the video instances in the embedding space (Liu et al., 2011; Xian et al., 2016; Wang and Chen, 2017; Huang et al., 2021). The semantic representations from these works are generally obtained using only the action class label, which does not explicitly contain information about the objects that appear in the videos. However, a small number of papers proposed to build relationships between the objects and the action label to produce enriched semantics for representing class knowledge (Jain et al., 2015; Mettes and Snoek, 2017; Gao et al., 2019), achieving better performance than the approach of using only the class label.

It is noteworthy that the video instances of seen and unseen classes can be totally disjoint and their distributions are different, resulting in a large domain gap (termed domain shift). For example, both classes of 'horse' and 'pig' have the same attribute of 'tail', but their tails could look very different. Hence, the projection-based approaches learn a projection function using video instances from the seen classes without any adaptation, leading to the problem of domain shift (Fu et al., 2015). To alleviate the influence of domain shift, generative-based methods can learn a model (e.g. GANs) to synthesise visual features for unseen classes based on the relationships between the instances of seen classes and the semantic representations of both seen and unseen classes. A supervised model is then trained to produce predictions for a given test sample (Narayan et al., 2020).

In our work, we propose to explore and improve the GAN-based approach (Narayan et al., 2020) by using additional object-based information as enhanced semantics for the ZSAR task. For our investigation, we use a subset of the UCF101 dataset, which is a common-used dataset for benchmarking in the HAR field. We answer the research question - Can object-based information be used as extra semantic information to improve the performance in the GAN-based framework for the ZSAR task? We summarise our contributions as follows. (1) We improve recognition accuracies for unseen classes by adding extra object-based knowledge into the GAN-based framework. (2) We perform empirical evaluations to investigate three methods for injecting object-based knowledge into the GAN-based framework. The methods are:

- Replacing: the semantic representations of action classes are replaced with the corresponding object-based representations.
- Appending: the object-based representations are appended to the semantic representations of action classes.
- Averaging: both class-based and object-based representations are averaged.

(3) We created a subset of the UCF101 dataset and selected multiple objects for each class to empirically evaluate models, including the selected objects for each action class.

The rest of this paper is structured as follows. In Section 2, we introduce an overview of ZSAR, including a review of the related works and the categorisation of the different approaches. In Section 3, we present the GAN-based framework and explain how the ZSAR task is performed. In Section 4, we describe the proposed methodology, which includes object-based information using a subset of the UCF101 dataset. We also define the metrics for evaluating its performance. In Section 5, we explain the experimental setups and their implementations in more detail. In Section 6, we show the results and findings. Finally, in Section 7, we conclude the paper and propose a few ideas for future work.

2 RELATED WORK

In this section, we review the relevant literature on early approaches for ZSAR as well as on generative approaches based on GANs.

Early works on the ZSAR depend on projection-based methods (Xu et al., 2015; Li et al., 2016; Xu et al., 2016; Xu et al., 2017; Wang and Chen, 2017). They learned a projection function that models the relationship between visual features and semantic features for seen classes, where these visual features are typically extracted from a deep neural network. The learned projection function is then used to recognise a novel class by measuring the likelihood between its true semantic representation and the predicted semantic representation of the video instances in the embedding space. It can be expected that classes with similar semantic descriptions contain similar vectors in the semantic embedding space. However, classes with similar attributes-based semantic knowledge may have large variations in the visual space. For example, both action classes of *walking* and *running* have the semantic description of *outdoor activity*, but their video instances could seem very different since *walking* and *running* have a very different pace and outfit. Therefore, building a high-accuracy projection function is a challenge, which may cause

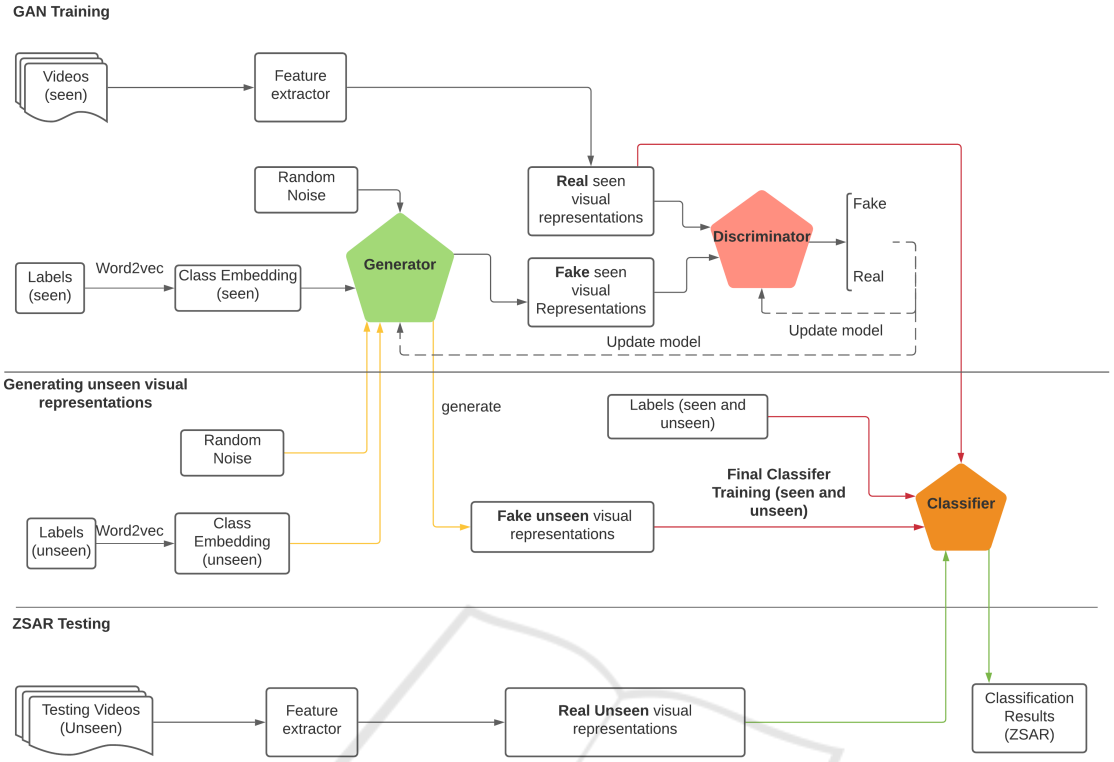

Figure 1: Pipeline of the GAN-based framework for Zero-Shot Action Recognition based on the approach (Xian et al., 2018).

ambiguity in the visual-semantic projection due to the large variation in the visual embedding space.

Instead of either using class-level labels or manual-annotated attributes as external knowledge in the ZSAR task, other works (Jain et al., 2015; Mettes and Snoek, 2017; Gao et al., 2019; Xing et al., 2021) used the relationships between objects and actions as additional knowledge to improve the ZSAR performance. Intuitively, objects that appear in a video help to identify what the video is about. This leads to an approach in which the semantics of the action class can be obtained not only from the class label but also from the objects that appear in the video. The work (Jain et al., 2015) aims to recognise action classes using semantic embedding that are extracted from a skip-gram model for the object categories detected in videos. Action class labels are associated with objects with a convex combination function of action and object affinities. The authors of the paper (Mettes and Snoek, 2017) built a spatial-aware embedding by incorporating word embeddings, box locations for actors and objects, as well as spatial relations to generate action tubes. This method builds an embedding to represent the relationships among actions, objects, and their interactions, providing enriched semantic representation. Advancing on this approach, the work (Gao et al., 2019) built an end-to-end ZSAR framework based on a structured knowledge graph, which can jointly model the relationships between action-attribute, action-action, and attribute-attribute. All of these works attempt to build the relationships between objects detected in videos and action class labels to represent enriched semantics and that can be applied in the ZSAR task for better performance.

Recently, generative-based methods are used to synthesise visual features for unseen classes based on their semantic representations. In the GAN context, most approaches were originally proposed for the task of zero-shot image classification (ZSIC). GANs can be used for synthesising fake unseen samples. The authors of the paper (Xian et al., 2018) proposed a conditional Wasserstein GAN (WGAN) model using classification loss to synthesise visual features for unseen classes. Specifically, the conditional WGAN is learned using visual features of seen classes and their corresponding semantic embeddings. The visual features of the unseen classes are then synthesised using the trained conditional WGAN and used together with the real visual features from seen classes to train a classifier in a supervised manner. The overall pipeline of the GAN-based approach to achieve the ZSAR task is shown in Fig. 1. Other works (Felix et al., 2018; Huang et al., 2019a; Mandal et al., 2019; Narayan et al., 2020; Mishra et al., 2020) also apply

auxiliary components to enforce a cycle-consistency constraint on the reconstruction of the semantic embeddings during training. The auxiliary components help to produce a higher quality generator to synthesise semantically consistent visual features for unseen classes. However, these approaches only embedded the action class label to train GANs and did not consider any semantics related to objects that may appear in video instances. In this work, we propose to include object knowledge as further conditioning information. We add external knowledge into the semantic space to represent the semantics for the action classes. We expect this object-based enriched semantics to improve the ZSAR performance on the GAN-based framework.

3 APPROACHES

In this section, we first explain how GANs are used for ZSAR. We then present a specific GAN-based framework used in our empirical evaluations.

The general GAN-based framework shown in Fig. 1 was originally proposed for the ZSIC by (Xian et al., 2018). The main objective of using a GAN is to generate visual representations for unseen classes due to the lack of unseen instances. There are three stages in the framework, which are GAN training, generating unseen visual representations, and ZSAR testing. The GAN consists of a generator G and a discriminator D. D attempts to distinguish real visual representations from the generated ones, and G attempts to fool the discriminator by generating visual representations that are similar to the real ones. The GAN here is also trained with the condition of word embeddings for a seen action class label (e.g. word2vec), which enables a generator to synthesise accurate visual representations that are representative of that class. Once the GAN has been trained, G is applied to synthesise visual representations for unseen action classes, using the corresponding semantic embedding along with random noise. A discriminative classifier is then trained with real seen visual representations, synthesised unseen visual representations, and all the seen and unseen class labels in a fully supervised fashion as shown in the middle layer called 'Generating unseen visual representations' of Fig. 1. In the ZSAR testing stage, the learned discriminative classifier is used to predict a label by inputting the real visual representation of the testing action video. It is noted that the GAN-based framework is only trained with visual and semantic representations from seen action classes. But,it has the ability to synthesise semantically visual representations conditioned on a class-specific semantic vector (e.g. word2vec) without having access to any video instances of the unseen classes.

In our work, we follow the GAN-based approach named TF-VAEGAN (Narayan et al., 2020) and this work extended evaluations from the ZSIC to the ZSAR. To keep this paper self-contained, we introduce the TF-VAEGAN framework in more detail, which recently delivered impressive ZSAR results. The process of training a GAN is a key part of the framework and it is illustrated in Fig. 2. As discussed, GANs can synthesise visual representations that are close to the distribution of real instances, but they could suffer from an issue called mode collapse (Arjovsky and Bottou, 2017), which leads to the problem of having low diversity of generated visual representations. In other words, mode collapse could produce visual representations without large variations in the end. While similar to GANs, variational autoencoders (VAEs) (Kingma and Welling, 2013) are another generative model that employs an encoder to represent the input as a latent variable with a Gaussian distribution assumption and a decoder that reconstructs the input from the latent variable.

The generation of unseen visual representations with VAEs can be achieved with more stable outputs than with GANs (Verma et al., 2018). Hence, the TF-VAEGAN framework combines the advantages of VAEs and GANs by assembling the decoder of the VAE and the generator of the GAN to synthesise semantically consistent visual representations, achieving impressive classification ZSAR results. In Fig. 2, the real visual representations of seen action classes x extracted from a deep neural network (e.g. I3D model (Carreira and Zisserman, 2017)) along with the semantic embeddings a are the input to the encoder E. The output of E is the latent code z that compresses the information from visual representations x, optimised by the Kullback-Leibler divergence. The random noise and semantic embeddings a are the input of the generator G that synthesises the visual representation x', and the synthesised visual representations x'and real visual representations x are compared using a binary cross-entropy loss.

The discriminator D takes either x or x'along with the corresponding semantic embeddings a as input, and determines whether the input is real or synthesised. The WGAN loss is applied to the output of D to distinguish between the real and the synthesised visual representations. Additionally, both the semantic embedding decoder *SED* and the feedback module F improve the process of visual representation synthesis and reduce ambiguities among action classes during the zero-shot classification. The *SED* inputs

either x or x'and reconstructs the semantic embedding a', which is trained by a cycle-consistency loss.

The feedback module F transforms the latent embedding of *SED* and puts it back to the latent representation of G which can refine x'to achieve an enhanced visual representation synthesis. It is worth noting that the generator G transforms the semantic embeddings to visual representations, while *SED* transforms the visual representations to semantic embeddings. Consequently, the G and the *SED* include supplementary information regarding visual representation and also the supplementary information can assist to improve the quality of visual representation synthesis and reduce ambiguity and misclassification among action classes.

In our work, we enrich the semantic embeddings a by incorporating object-based semantic embedding with different inclusion techniques such as replacing, appending and averaging rather than only using simple word embedding of action labels. By doing this, we expect to achieve better performance based on the TF-VAEGAN framework for the ZSAR task.

4 METHODOLOGY

In this section, we describe our methodology to perform the ZSAR task using the TF-VAEGAN framework with the inclusion of object-based semantics on a subset of a benchmark dataset (i.e. UCF101). We also describe in more detail the applied visual and semantic embeddings.

Dataset. The videos in our selected dataset only contains a single action, so we are not dealing with the complexities of either action transitions or multiple actions in the same instance (Ye et al., 2010). In our work, we choose the UCF101 (Soomro et al., 2012) dataset, which has 101 action categories with a total of 13320 videos. To apply object-based semantic information, we select objects for each action class. First, we selected 10 seen classes for training and 10 unseen classes for testing. Then, for each seen class, we manually selected three objects that appear in the videos. Also, we manually reviewed all the objects for seen classes to ensure that the objects we picked are linked to the actions recommended by ConceptNet (Speer et al., 2017). [1] For unseen classes, we found suitable objects solely using ConceptNet, as video instances of unseen action classes are not available during the training process. Hence, only an external knowledge base (i.e. ConceptNet) is used to find out objects for unseen action classes, which can avoid breaking the premise of zero-shot learning (i.e. seen and unseen classes are disjoint). For example, the objects for the action class of *Shaving Beard* are *beard*, *face* and *shaver*. The action class of *Biking* has the objects *bicycle*, *helmet* and *wheel*. In a nutshell, we choose 10 seen classes and 10 unseen classes from the UCF101 dataset along with the selection of three objects for each class as shown in Table 1.

Table 1: The information of action classes and their objects for our generated subset.

Seen Classes	Objects
Archery	arrow,bow,bracer
Basketball Shoot	basketball,hoop,backboard
Brushing Teeth	teeth,toothpaste,mouth
Diving	pool,water,springboard
Mopping Floor	mop,floor,mphead
Shaving Beard	beard,face,shaver
Surfing	wave,surfboard,surf
Typing	finger,keyboard,monitor
Walking With Dog	leash,dog,road
Writing On Board	board,mark pen,finger

Unseen Classes	Objects
Biking	bicycle,helmet,wheel
Blow Dry Hair	hair,dryer,head
Blowing Candles	candle,light,table
Golf Swing	golf,golf club,grass
Haircut	hair, hand,scissors
Horse Riding	horse,saddle,race course
Rafting	raft,river,paddle
Soccer Penalty	soccer,goal,pitch
Table Tennis Shot	ping pong,table tennis bat, table
Tennis Swing	tennis ball, tennis racket, tennis net

Visual & Semantic Embeddings. To produce real visual representations x in Fig. 2, we adopted the off-the-shelf I3D model for visual feature extraction provided by (Mandal et al., 2019), which is the most common approach. I3D was originally proposed by (Carreira and Zisserman, 2017) and contains RGB and Inflated 3D networks to generate appearance and flow features from *Mixed_5c* layer. For each video instance, the outputs from *Mixed_5c* layer for both networks are averaged through a temporal dimension, pooled in the spatial dimension, and then flattened to obtain a 4096-dimensional vector for appearance and flow features. In the end, both appearance and flow

[1] ConceptNet is a knowledge base that connects words and phrases of natural language with labelled relationships. Its knowledge was collected from many resources, such as WordNet(Bond and Foster, 2013) and DBPedia(Auer et al., 2007).

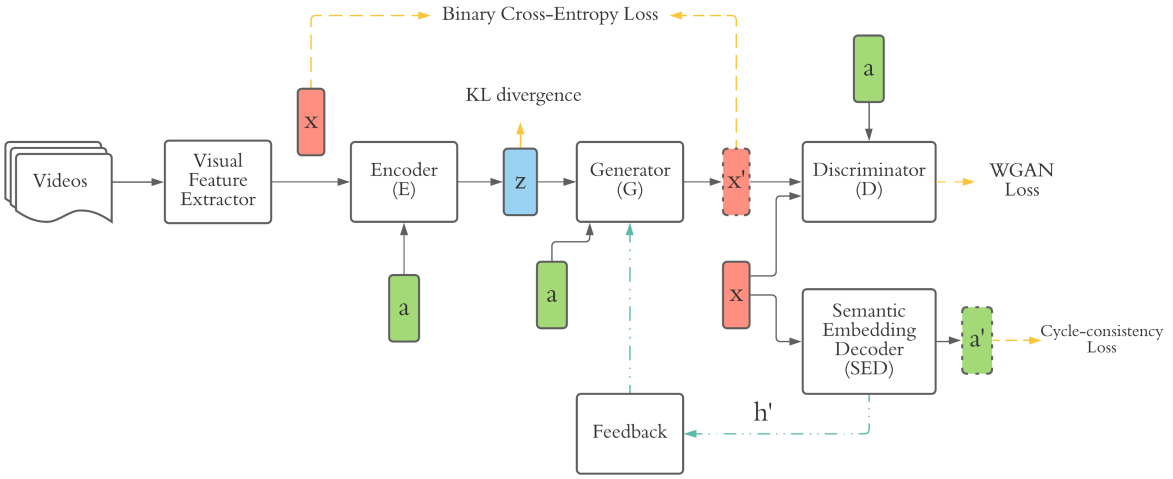

Figure 2: The overall architecture of TF-VAEGAN framework.

features are concatenated to represent a video with an 8192-dimensional vector.

It has been recently shown that the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) presents state-of-the-art results in a wide variety of natural language processing tasks, such as question answering and natural language inference. However, in the literature of ZSAR, Word2Vec has been extensively used to produce semantic embedding for action labels. The aim of this paper is to justify whether the ZSAR performance can be improved by including object-based information and therefore, we used the common approach to compare our performance with other researchers (Xu et al., 2015; Li et al., 2016; Xu et al., 2016; Xu et al., 2017; Wang and Chen, 2017). In the same way, we implemented the I3D CNNs rather than the latest CNN models as they are much more popular in the literature. Therefore, for producing the class-specific semantic representations a in Fig. 2, Word2Vec (Mikolov et al., 2013), which is built upon a skip-gram model that was pre-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 on our generated subset. Furthermore, for all the selected objects, each of their names is represented as a 300-dimensional vector using Word2Vec. For this subset of the UCF101 dataset, all the action class labels and object names have the same type and size of word embedding.

The Inclusion of Object Semantics. In previous GAN-based ZSAR works, the component of class embedding a in Fig. 1 is a simple word embedding (i.e. Word2Vec) for each action label. The input does not contain any object-related semantics. As discussed earlier, we aim to include objects into the class embedding a as enhanced knowledge to improve ZSAR performance, evaluating on the GAN-based framework (i.e. TF-VAEGAN). We have considered three approaches to add object-based semantics that are replacing, appending, and averaging. For replacing, the embedding of the class label is directly substituted by either the embedding of the one corresponding object or by more objects. For appending, the object embeddings are concatenated to the corresponding action class embedding to form a new high dimensional vector, whose length relates to the number of objects to be appended. For example, the new embedding would be a 600-dimensional vector if one object embedding is appended to action class embedding. For averaging, we calculate the mean of the embedding of the action class label and the embedding of the corresponding objects as a new 300-dimensional vector, which also contains semantic information for both class and objects within a small-size vector. All three object-inclusion approaches will be empirically evaluated on our generated subset using the modified TF-VAEGAN framework.

Evaluation Metrics. Class accuracy is a standard metric in the ZSAR field. We use the average per-class accuracies defined in the following equation (Xian et al., 2017):

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

5 EXPERIMENTS

In this section, we explain the experimental configurations for evaluating the TF-VAEGAN framework on

Table 2: Experimental configurations for the inclusion of **one** object. Note that, *Rep.*, *App.* and *Avg.* denote the inclusion approaches of replacing, appending and averaging, respectively. *obj1*, *obj2* and *obj3* denote the first, second and third objects being selected for an *action* class, respectively.

Experiments	Object Inclusion	Semantic Embedding
1	No (baseline)	action
2	Rep.	obj1
3	Rep.	obj2
4	Rep.	obj3
5	App.	action, obj1
6	App.	action, obj2
7	App.	action, obj3
8	Avg.	AVG(action, obj1)
9	Avg.	AVG(action, obj2)
10	Avg.	AVG(action, obj3)

our generated subset with the three aforementioned operations (i.e. replacing, appending and averaging) to include object-based information as enhanced knowledge. The ZSAR implementations are then described in more detail.

Experiments and Baseline. Table 2 shows ten experimental configurations conducted to perform ZSAR with different inclusion approaches with only one object. The object inclusion approaches are replacing (**Rep.**), appending (**App.**) and averaging (**Avg.**). As the baseline, the object-based knowledge is not used in Experiment 1, but only knowledge related to the action class. From Experiment 2 to Experiment 10, a single object is included to produce the object-involved semantic embedding a, shown in Fig. 2. The results from Experiment 2 to Experiment 10 are compared to the baseline to see whether the object inclusion can improve the ZSAR performance. Notably, the visual features of seen classes (i.e. x) are represented by I3D and the synthesised visual features for unseen classes (i.e. x') are the same as I3D in size. Also, the semantic representations of action class labels and object names are extracted using Word2Vec to form the different types of semantic embedding (i.e. a) with different object inclusion approaches.

For further investigations, Table 3 shows eight experimental configurations including multiple objects as semantic embedding to perform the ZSAR evaluations with the approaches of *App.* and *Avg.* We ignore the approach of *Rep.* in the cases of including multiple objects because of its poor performance in single-object inclusion evaluations, which will be shown in the next section.

Table 3: Experimental configurations for the inclusion of **multiple** objects.

Exp.	Object Inclusion	Semantic Embedding
11	App.	action, obj1, obj2
12	App.	action, obj1, obj3
13	App.	action, obj2, obj3
14	App.	action, obj1, obj2, obj3
15	Avg.	AVG(action, obj1, obj2)
16	Avg.	AVG(action, obj1, obj3)
17	Avg.	AVG(action, obj2, obj3)
18	Avg.	AVG(action, obj1, obj2, obj3)

Implementation Details. Similar to the TF-VAEGAN approach (Narayan et al., 2020), the structures of discriminator D, encoder E, and generator G are developed as fully connected networks in two layers along with 4096 hidden units. Also, the semantic embedding decoder *SED* and the feedback module F have the same structure as D, E and G. Leaky ReLU is used for each activation function, except in the output of G, where a sigmoid activation is applied to calculate the binary cross-entropy loss. The whole framework is trained using an Adam optimiser with 10^{-4} learning rate. The final discriminative classifier is a single-layer fully connected network with equal output units to the number of unseen classes. We directly apply the same hyper-parameters as the TF-VAEGAN work, such as α, β and σ are set to 10, 0.01 and 1, respectively. As in the work (Xian et al., 2019), α is the coefficient for weighting the WGAN loss, β is a hyper-parameter for weighting the decoder reconstruction error in the semantic decoder embedding *SED*, and σ is used in the feedback module F to control the feedback modulation. The gradient penalty coefficient λ is initially set to 10 for training a WGAN in the framework. Additionally, each configuration is run 5 times.

6 RESULTS & ANALYSIS

In this section, we present and discuss the experimental results for the configurations shown in Table 2 and Table 3. For each configuration, we measure the mean average accuracy across 10 unseen classes over 5 repetitions. We compare the results obtained in the experiments with object inclusion to the results from the baseline experiment without any object knowledge.

Table 4 shows the mean average accuracy from the baseline and other experiments with a single object inclusion, and also the comparison results in percent-

Table 4: Experimental results for the baseline without object knowledge and other experiments with **single** object inclusion and the comparison results are shown in percentage. Note that, MAA denotes the mean average accuracy and STD denotes the standard deviation.

Exp.	Object Inclusion	MAA / STD	Compared to baseline
1	Baseline	47.06% / 0.02	\
2	Rep. obj1	43.69% / 0.03	-3.37%
3	Rep. obj2	17.08% / 0.04	-29.98%
4	Rep. obj3	42.88% / 0.02	-4.18%
5	App. obj1	51.96% / 0.02	4.90%
6	App. obj2	31.33% / 0.02	-15.73%
7	App. obj3	48.55% / 0.01	1.49%
8	Avg. obj1	51.00% / 0.01	3.94%
9	Avg. obj2	36.28% / 0.03	-10.78%
10	Avg. obj3	50.56% / 0.03	3.50%

age against the baseline. As can be seen, the approach of *Rep.* (Experiment 2-4) yields poor performance against the baseline regardless of which object is incorporated as there is no semantic knowledge of the action class included which could have improved the quality of visual feature synthesis of the generator. To this point, the approach of *Rep.* is not considered for the experiments with multiple objects inclusion.

For the *App.* approach (Experiments 5-7), the comparison results against the baseline are 4.90%, -15.73% and 1.49% by incorporating *obj1*, *obj2* and *obj3*, respectively. Similarly, after evaluating the *Avg.* approach (Experiments 8-10), the comparison results are 3.94%, -10.78% and 3.50% when including *obj1*, *obj2* and *obj3*, respectively. It is worth noting that the inclusion of the second object (*obj2*) delivers the worst performance by a large margin of all approaches when compared to the baseline and moreover, there is an improvement when using *obj1* and *obj3* for both *App.* and *Avg.* approaches. In this regard, we observe that the selection of objects for each action class has a significant impact on the ZSAR evaluation in the TF-VAEGAN framework. We then check our generated subset to find out which exact *obj2* is selected for each action class. For example, we choose 'light' as *obj2* for the action class of 'Blowing Candles' and 'river' as *obj2* for the action class of 'Rafting'. These objects seem too general to be added to the corresponding action class, which indicates that the selected objects for an action class should be specifically related to the contents of the videos of that class to provide more precise knowledge to represent its semantic embedding. On the contrary, a boost compared to the baseline is gained by including *obj1* (e.g. Experiments 5 and 8) because *obj1* is specifically related to the corresponding action class. For example, the object 'hair' is selected as *obj1* for the action class 'Haircut', and also the object 'horse' as *obj1* belongs to the action class 'Horse Riding'. To this point, it is understood that including high-relevant objects into one action class can improve the performance in the ZSAR field.

Table 5: Experimental results for the baseline and other experiments including **multiple** objects inclusion and the comparison results are shown in percentage.

Exp.	Object Inclusion	MAA / STD	Compared to baseline
1	Baseline	47.06% / 0.02	\
11	App.(obj1, obj2)	43.62% / 0.02	-3.44%
12	App.(obj1, obj3)	50.40% / 0.01	3.34%
13	App.(obj2, obj3)	43.34% / 0.02	-3.72%
14	App. all obj.	45.63% / 0.02	-1.43%
15	Avg.(obj1, obj2)	44.37% / 0.01	-2.69%
16	Avg.(obj1, obj3)	51.69% / 0.02	4.63%
17	Avg.(obj2, obj3)	43.18% / 0.01	-3.88%
18	Avg. all obj.	43.10% / 0.01	-3.96%

As can be seen in Fig. 3, the inclusion approach of *Rep.* does not improve the ZSAR performance. We consider that removing class labels has a negative impact on representing semantic embedding given to the GANs. This is congruent with the results of *App.* and *Avg.*, which have better results than the *Rep.* and also than the baseline for *obj1*, *obj2* and *obj3*. Furthermore, the objects of the second column (i.e. *obj2*) do not improve the results but worsen them. We think the objects were not properly selected and in the future, we will explore different strategies for selecting objects to improve the baseline individually as well as when combined with other objects.

As further explorations, we conduct more experiments including multiple objects into one class for ZSAR evaluations (Experiment 11-18), and the results are shown in Table 5. For the *App.* approach, the results against the baseline are -3.44%, 3.34%, -3.72% and -1.43% by incorporating *obj1 + obj2*, *obj1 + obj3*, *obj2 + obj3* and *all objects*, respectively. It can be found that the performance boost is only gained when excluding *obj2*, which also delivers poor performance in the single-object inclusion experiments. Similarly, for the *Avg.* approach, the comparison results are -2.69%, 4.63%, -3.88% and -3.96% for the cases of *obj1 + obj2*, *obj1 + obj3*, *obj2 + obj3* and *all objects*, respectively. The inclusion of *obj2* yields poor performance regardless of which other objects to be included together, shown in Fig. 4. In other words, similar to experiments of using single object, *obj2* can pull down the ZSAR performance if it is included into semantic embedding in the framework. Additionally, the *Avg.* approach cannot dominantly outperform the *App.* approach in all cases, and vice versa. The standard deviation values for all the experiment results are under 0.04, indicat-

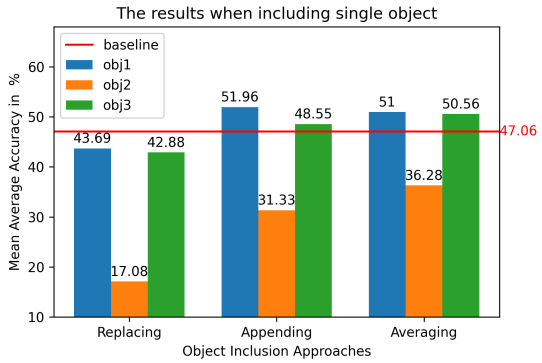

Figure 3: The results of mean average accuracy when including **single** object for three inclusion approaches.

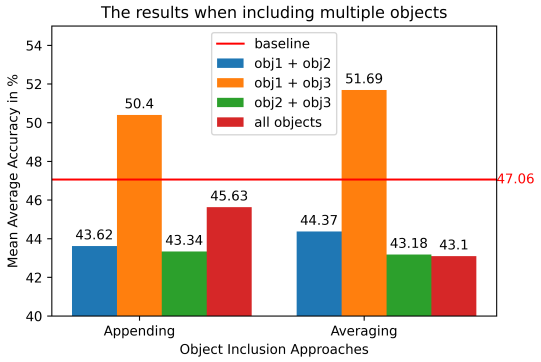

Figure 4: The results of mean average accuracy when including **multiple** objects for three inclusion approaches.

ing that the generation process of visual features for unseen classes is stable in the TF-VAEGAN and also the final discriminative classifier based on the supervised learning approach is learned consistently.

The limitations of our work are described as follows: (1) We manually selected objects that appear in videos, along with manual-checking in the ConceptNet, for each action class. This approach does not scale well if the number of considered action classes along with their selected object increases. (2) We only considered evaluating our object-inclusion approach on the TF-VAEGAN framework which could be a limitation, leading to finding out the less conclusive points.

All discussions so far are focused on model performance with object inclusion as enhanced knowledge in the framework. There is no doubt that including high-relevant objects for action classes can improve the ZSAR results in the TF-VAEGAN framework. We also believe that the approach of object inclusion can be applied to the generative-based model, which can be GANs, VAE and the combination of both GAN and VAE. Additionally, our results in the experiments show that it is possible to improve the performance of ZSAR by adding related objects to the classes. Other investigations have better results because they use more elaborated frameworks. However, in order to understand better what kind of objects improved the baseline the most to refine the object selection, it was more suitable not to select a complex framework.

7 CONCLUSION

In this piece of research, we have investigated the impact on the evaluations of zero-shot action recognition by incorporating object-based knowledge in the TF-VAEGAN framework. We generated a subset from the UCF101 dataset by manually selecting objects that appear in videos for each action class, and then evaluated the framework on this subset. Furthermore, we evaluated and discussed the different approaches of object inclusion (i.e. *Replacing*, *Appending* and *Averaging*). We have proved that adding objects to the class labels is a feasible strategy to improve the results when *Appending* and *Averaging* approaches are applied. We have also seen that the objects can improve but also can hinder the performance of ZSAR. Our approach can also be applied to generalised ZSAR. Lastly, we concluded that a performance boost can be achieved by incorporating high-relevant and specific objects as enhanced semantic knowledge into the corresponding action classes in the generative-based method for the ZSAR task.

As future work, we aim to investigate generalised ZSAR which is a more challenging task that tests both seen and unseen classes together in the classification stage. Also, we will explore strategies for selecting objects in a better way. In addition, more approaches to include the object-based information as enriched knowledge will be explored and examined, such as extracting semantic knowledge from an existing knowledge graph (i.e. ConceptNet) for action classes. Furthermore, we expect that this performance could be improved by applying optimisation techniques to the hyper-parameters values. We also think that using more recent and effective modules in the pipeline such as BERT or 3D ResNets (Kataoka et al., 2020) will boost the performance of our approach.

ACKNOWLEDGEMENTS

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

REFERENCES

Arjovsky, M. and Bottou, L. (2017). Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Felix, R., Reid, I., Carneiro, G., et al. (2018). Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37.

Fu, Y., Hospedales, T. M., Xiang, T., and Gong, S. (2015). Transductive multi-view zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2332–2345.

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 conference on artificial intelligence*, volume 33, pages 8303–8311.

Huang, H., Wang, C., Yu, P. S., and Wang, C.-D. (2019a). Generative dual adversarial network for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 801–810.

Huang, K., Delany, S.-J., and McKeever, S. (2019b). Human action recognition in videos using transfer learning. In *Proceedings of IMVIP 2019: Irish Machine Vision & Image Processing, Technological University Dublin, Dublin, Ireland*.

Huang, K., Delany, S. J., and McKeever, S. (2021). Fairer evaluation of zero shot action recognition in videos. In *VISIGRAPP (5: VISAPP)*, pages 206–215.

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 international conference on computer vision*, pages 4588–4596.

Kataoka, H., Wakamiya, T., Hara, K., and Satoh, Y. (2020). Would mega-scale datasets further enhance spatiotemporal 3d cnns? *arXiv preprint arXiv:2004.04968*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, Y., Hu, S.-h., and Li, B. (2016). Recognizing unseen actions in a domain-adapted embedding space. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4195–4199. IEEE.

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.

Mettes, P. and Snoek, C. G. (2017). Spatial-aware object embeddings for zero-shot localization and classification of actions. In *Proceedings of the IEEE international conference on computer vision*, pages 4443–4452.

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*, 390:117–130.

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. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 479–495. Springer.

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*.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Verma, V. K., Arora, G., Mishra, A., and Rai, P. (2018). Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4281–4289.

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

Wang, X. (2013). Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19.

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., Lorenz, T., Schiele, B., and Akata, Z. (2018). Feature generating networks for zero-shot learning. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5542–5551.

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.

Xian, Y., Sharma, S., Schiele, B., and Akata, Z. (2019). f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10275–10284.

Xing, M., Feng, Z., Su, Y., Peng, W., and Zhang, J. (2021). Ventral & dorsal stream theory based zero-shot action recognition. *Pattern Recognition*, 116:107953.

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

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

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

Ye, J., Coyle, L., Mckeever, S., Dobson, S., Ye, J., Coyle, L., Mckeever, S., and Dobson, S. (2010). Dealing with activities with diffuse boundaries. In *In Proceedings of the Workshop on How*.