2021-12

Combining Text and Image Knowledge with GANs for Zero-Shot Action Recognition in Videos

Kaiqiang Huang
Technological University Dublin, d14122793@mytudublin.ie

Luis Miralles-Pechuán
Technological University Dublin, luis.mirlles@tudublin.ie

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

Recommended Citation

Huang, K., Miralles-Pechuán, L. and Mckeever, S. (2022). Combining Text and Image Knowledge with GANs for Zero-Shot Action Recognition in Videos. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* Volume 5: VISAPP, ISBN 978-989-758-555-5, pages 623-631. DOI: 10.5220/0010903100003124

Combining Text and Image Knowledge with GANs for Zero-Shot Action Recognition in Videos

Kaiqiang Huang, Luis Miralles-Pechuán and Susan Mckeever
School of Computing, Technological University Dublin, Central Quad, Grangegorman, Dublin, Ireland
kaiqiang.huang@tudublin.ie, luis.miralles@tudublin.ie, susan.mckeever@tudublin.ie

Keywords: Human Action Recognition, Zero-Shot Learning, Generative Adversarial Networks, Semantic Knowledge Source

Abstract: The recognition of actions in videos is an active research area in machine learning, relevant to multiple domains such as health monitoring, security and social media analysis. Zero-Shot Action Recognition (ZSAR) is a challenging problem in which models are trained to identify action classes that have not been seen during the training process. According to the literature, the most promising ZSAR approaches make use of Generative Adversarial Networks (GANs). GANs can synthesise visual embeddings for unseen classes conditioned on either textual information or images related to the class labels. In this paper, we propose a Dual-GAN approach based on the VAEGAN model to prove that the fusion of visual and textual-based knowledge sources is an effective way to improve ZSAR performance. We conduct empirical ZSAR experiments of our approach on the UCF101 dataset. We apply the following embedding fusion methods for combining text-driven and image-driven information: averaging, summation, maximum, and minimum. Our best result from Dual-GAN model is achieved with the *maximum* embedding fusion approach that results in an average accuracy of 46.37%, which is improved by 5.37% at least compared to the leading approaches.

1 INTRODUCTION

Over the last decade, the problem of Human Action Recognition (HAR) has been addressed by a variety of supervised learning approaches. For example, identifying whether a video belongs to a given trained class (e.g. *Jumping*) (Wang and Schmid, 2013). Recently, challenging research problem termed Zero-Shot Action Recognition (ZSAR) has been studied to recognise video instances of unseen classes (i.e. not used during the training process) by transferring semantic knowledge from the seen classes to the unseen ones in the HAR field.

Most approaches in the early research stage to achieving ZSAR have used projection-based methods. The methods learn a projection function to map the visual embedding of seen classes to their corresponding semantic embeddings. For example, a projection function can be used to map the visual feature of the *Running* class to the Word2Vec embedding of the *Running* class label. The learned projection function is then applied to recognise novel unseen classes by measuring a similarity-based metric between the ground-truth embeddings and the predicted embeddings on the testing videos (Liu et al., 2011; Xian et al., 2016; Huang et al., 2021a). However, the video samples of seen and unseen classes are totally different. Therefore, the projection-based approaches without developing any adaptation techniques between seen and unseen classes can lead to the problem of largely variational mismatching during the test phase. To mitigate this problem, recent ZSAR approaches have introduced a key approach for synthetic data generation called Generative Adversarial Networks (GANs) which is a natural candidate for the zero-shot learning task involving new unseen classes. ZSAR approaches using GANs aim to synthesise visual embeddings of unseen classes based on their corresponding semantic embeddings to mitigate the discrepancy between seen and synthesised data. After the synthesised data is generated for unseen classes, a classifier is trained with the real seen and the synthesised unseen data in a fully-supervised fashion to make predictions for a given test sample (Mandal et al., 2019; Narayan et al., 2020; Huang et al., 2021b).

In this work, we propose a Dual-GAN approach based on the VAEGAN model (Narayan et al., 2020) that fuses two semantic embeddings obtained from different knowledge sources (i.e. text and image)

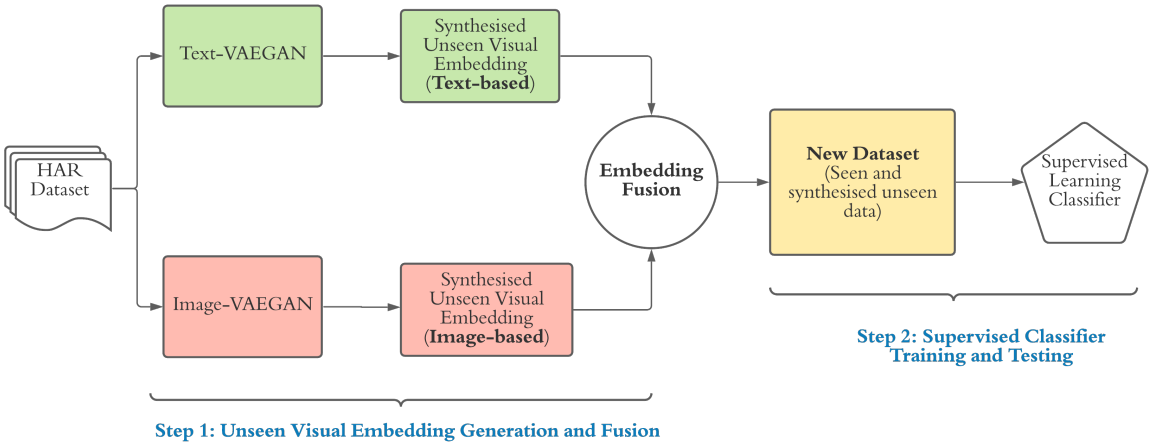

Figure 1: High-level perspective of the pipeline for the proposed Dual-GAN approach based on VAEGAN model.

for the ZSAR task. For our experiments, we used a commonly-used benchmark dataset in the HAR field named UCF101. In our paper, we answer the following two research questions: (1) Can *image-based* semantic embeddings, which have not been applied to the GAN-based model yet, have a higher ZSAR performance than the existing approaches based on *text-based* semantic embeddings? and (2) Can our proposed Dual-GAN approach incorporating two knowledge sources get higher accuracy than a Single-GAN approach (using either text or image)? Our main hypothesis is that *combined* semantic embeddings produced from two knowledge sources (e.g. text and image) that contain complementary information could improve the ZSAR performance in the GAN-based framework.

We summarise our contributions as follows:

1. We investigate two different knowledge sources (i.e. texts and images) that can be used to represent semantic meaning for action classes.
2. We propose a Dual-GAN approach based on the VAEGAN model to generate high-quality visual embeddings for unseen classes by fusing semantic embeddings obtained from two knowledge sources (i.e. texts and images). The fusion methods include averaging, summation, maximum and minimum.
3. Our Dual-GAN model outperforms the existing ZSAR approaches that use a GAN-based approach to synthesising unseen class representations. To the best of our knowledge, there are no previous works that employ a method of combining semantic embeddings derived from two different knowledge sources in the context of the GAN-based framework.

The rest of this paper is structured as follows. In Section 2, we provide a literature review of various approaches for the ZSAR. In Section 3, we introduce our proposed Dual-GAN approach based on the VAEGAN model using two knowledge sources for ZSAR. In Section 4, we describe the methodology, which includes the process of collecting images and feature fusion methods. In Section 5, we explain the experimental configurations and implementations in more detail. In Section 6, we show the results and key 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 related literature on the approaches in the early stage of the ZSAR research, as well as on the generative approaches based on GANs. In addition, we summarise the existing works that propose different types of semantic embedding, especially in the GAN-based framework.

In the early stage of research on ZSAR, several works (Xu et al., 2015; Li et al., 2016) proposed projection functions to map from a visual representation of video instances to a semantic representation of the class prototype that the video belongs to (i.e. typically an embedding space of a class label). These learned projection functions encode the relationship between visual embeddings and semantic embeddings using seen data. The learned projection function is then used to recognise new unseen classes by measuring the likelihood between the ground-truth and the predicted semantic representations of the video instances

in the embedding space. However, classes with similar semantic knowledge may have large variations in the visual space. For example, both action classes of *Diving* and *Swimming* have the same description such that *outdoor activity* and *has water*, but their video samples would look very different since *Diving* and *Swimming* have quite different body movements. Therefore, building a high-accuracy projection function is a big challenge, which may cause ambiguity in the visual-semantic mapping due to the large variation in the visual embedding.

Recently, advanced generative-based methods have been used to synthesise visual embeddings of unseen classes according to their semantic embeddings. Some authors (Xian et al., 2018) proposed a conditional Wasserstein GAN (WGAN) model using classification loss to synthesise visual embeddings of unseen classes. The visual embeddings of the unseen classes are then synthesised using a trained conditional WGAN and used together with the real visual embeddings of seen classes to train a discriminative classifier in a fully-supervised manner. There are other authors (Mandal et al., 2019; Narayan et al., 2020; Mishra et al., 2020) who also apply extra components to enforce a cycle-consistency constraint on the reconstruction of the semantic embeddings during training. The extra components assist to produce a higher quality generator to synthesise semantically consistent visual embeddings of unseen classes. Although these generative-based methods show promising results for the ZSAR task, they still struggle to generate higher quality and more satisfying visual embeddings of unseen classes since the generated unseen data is directly used to train a supervised-based classifier along with seen data.

Also, as mentioned in Section 1, if we can obtain richer and more representative knowledge incorporated into the semantic embedding of the actions, intuitively we should improve downstream ZSAR accuracy when identifying unseen classes. The authors of the paper (Wang and Chen, 2017) enhanced the word vectors of the label by collecting and modelling textual descriptions of action classes. The contextual information (e.g. textual descriptions related to action classes) would remove the ambiguity of the semantics to some extent in the original word vectors of action labels. For example, the class *Haircut* has a description that 'A hairstyle, hairdo, or haircut refers to the styling of hair, usually on the human scalp'. Sometimes, this could also mean an editing of facial or body hair. In that same work (Wang and Chen, 2017), the authors also proposed a method to collect images related to the action labels for representing visually discriminative semantic embedding. However, the work only evaluated the proposed semantic embeddings in a project-based approach, not on the GAN-based one. Similarly, the authors (Hong et al., 2020) proposed a description text dataset whose definition was taken from the official Wikipedia website for the UCF101 action dataset and evaluated it in the GAN-based model.

Table 1: Dataset used for evaluations.

Dataset	**#Class**	**#Instances**	**Seen/Unseen Proportion**
UCF101	101	13320	51/50

3 APPROACH

In this section, we explain our Dual-GAN approach for Zero-Shot Action Recognition and how it fuses semantic embeddings from two knowledge sources: text and images, shown in Fig. 1.

As shown in Fig. 1, the high-level perspective of the pipeline for the proposed Dual-GAN approach contains two steps. Step 1 aims to synthesise the visual embeddings of unseen classes conditioned on the corresponding semantic embeddings obtained from two different knowledge sources (i.e. texts and images) through the two VAEGAN components: Text-VAEGAN and Image-VAEGAN. After that, the outputs of both image-driven and the text-driven unseen visual embeddings are combined by a fusion operation (e.g. averaging) to form a new dataset that contains the original seen data and the synthesised unseen data along with their respective labels. Step 2 focuses on training a classifier in a supervised learning fashion with the new dataset generated in the previous step. It is noted that the generator of each VAEGAN component is only trained with seen data (i.e. video instances and labels). Each VAEGAN component is able to synthesise semantically visual embeddings conditioned on a semantic embedding (e.g. either the Word2Vec of the action label or the image-based representation of the action label) without having access to any video instances of the unseen classes.

To expand the high-level pipeline described above, we implemented the VAEGAN component with a similar structure to the work proposed in (Narayan et al., 2020) and shown in Fig. 2. To keep this paper self-contained, we describe the VAEGAN component, which recently yielded promising results for the ZSAR task, in more detail. As mentioned in Section 1, GANs can synthesise visual embeddings that are close to the distribution of real instances, but they can suffer from an issue termed mode collapse (Arjovsky and Bottou, 2017), which leads to the prob-

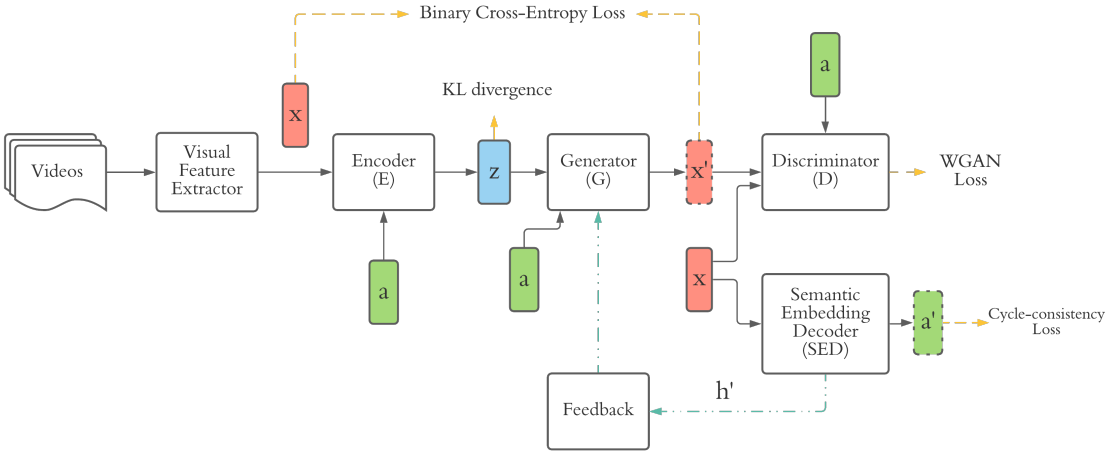

Figure 2: The details of one VAEGAN component (Huang et al., 2021b).

lem of having low diversity of synthesised visual embedding.

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 to transform the input from the latent variable. The generation of unseen visual embedding with VAE gives more stable outputs than with GANs (Verma et al., 2018). Hence, the architecture of the VAEGAN component combines the advantages of VAE and GAN by assembling the decoder of the VAE and the generator of the GAN to ultimately synthesise semantically consistent visual representations.

As shown in Fig. 2, the real visual embedding of seen classes x extracted from a deep neural network 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 the 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 process. The *SED* inputs either x or x'and reconstructs the semantic embedding a', which is trained using a cycle-consistency loss.

Table 2: The details of knowledge sources and semantic embeddings.

Semantics	Source	Embedding	Dimensions
Labels	Text	Word2Vec	300
Descriptions	Text	Word2Vec	300
Collected Images	Image	GoogleNet	1024
Collected Images	Image	ResNet101	2048

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 the supplementary information can assist to improve the quality of the visual representation synthesis and reduce ambiguity and misclassification among action classes.

The key approach to achieving ZSAR is to transfer semantic knowledge containing enriched and discriminative information from seen action classes to unseen action classes. Semantic embedding derived from multiple knowledge sources can potentially deliver better discriminative representation than only using a single source (Xiang et al., 2021). In this paper, we propose two improvements for ZSAR. First, we believe it is possible to improve the ZSAR performance by introducing a combination of text-based descriptions and images to represent semantic embed-

Table 3: Experimental configurations for comparing text-driven semantic embedding to image-driven semantic embedding in the Single-GAN model.

Dataset	Knowledge Source	Semantic Embedding
UCF101	Text (baseline)	Action Class Word2Vec
	Text	Description Word2Vec
	Image	GoogLeNet
	Image	ResNet101

ding for the corresponding action class. Therefore, we use two GANs rather than one, and we combine the generated features of each GANs by generating a new array that is calculated applying the following methods: average, maximum, minimum, or summation. Second, for extracting textual features, we employ an approach that uses textual descriptions for the action rather than the action class label itself. Intuitively, a textual description should contain more informative and contextual semantic meaning than just the class label. For the visual information, we use images related to the action class that provide enriched visual cues for representing the semantic meaning.

4 METHODOLOGY

In this section, we describe our methodology to perform the ZSAR task based on the proposed Dual-GAN model on the UCF101. We also introduce the method for collecting images for each action class and the method for extracting visual-based and text-based semantic embeddings in more detail.

Dataset We select the UCF101 (Soomro et al., 2012) dataset that is widely used as benchmark to evaluate the ZSAR performance. The details of the dataset is described in Table 1. Followed by the works (Mandal et al., 2019; Narayan et al., 2020), we use the same split for model training and evaluation. Each dataset has 30 independent splits and each split is randomly generated by keeping the same seen/unseen proportion so that all splits contain different seen and unseen classes for training and test. In other words, some classes are seen classes in one split, but these classes can be unseen ones in other splits.

Image Collection We apply a similar strategy to collect images to the one proposed by (Wang and Chen, 2017) in which the following steps are followed. First, we consider the action labels as the keywords to search related images by the image search engines (i.e. Google Image Source). [1] For example, we use the keyword *Playing YoYo* for searching images for the class *YoYo*. Then, after collecting the images, we remove the irrelevant and small-size images for each class. As a result, we obtain 15,845 images (157 images per class on average).

Visual & Semantic Embeddings To extract real visual embedding x in Fig. 2, we adopted the off-the-shelf I3D model for visual feature extraction provided by (Mandal et al., 2019). I3D was originally proposed by (Carreira and Zisserman, 2017) and it contains RGB and Inflated 3D networks to generate appearance and flow features from the *Mixed_5c* layer. For each video instance, the outputs from the *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 features are concatenated to represent a video with an 8192-dimensional vector.

We produce four types of semantic embedding a that can be used to condition the VAEGAN as shown in Fig. 2. The summary of semantic embedding is given in Table 2. The semantic embedding of action labels is extracted by Word2Vec. 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. The text-based description per class are provided by the work (Wang and Chen, 2017), motivated by the fact that a class label is not adequate to represent the complex concepts in human actions. The idea is that each label is transformed into a description of that label and then we use Word2vec to represent each word of that description. Then, we simply average all the generated arrays by Word2vec, which also delivers a 300-dimensional vector for each class.

To extract features for collected images, we apply two off-the-shelf models: GoogLeNet (Szegedy et al., 2015) and ResNet101 (He et al., 2016) which were both pre-trained on the ImageNet dataset. The average pooling layer that is before the last fully connected layer is used as the deep image features for both pre-trained models. Finally, all the extracted image features are averaged for each action class.

Embeddings Fusion As shown the Step 1 in Fig. 1, we aim to synthesise and combine different visual em-

[1] Image scraping tool is available at `https://github.com/Joeclinton1/google-images-download.git`

Table 4: Comparing our results to the TF-VAEGAN.

Dataset \ Model	TF-VAEGAN (Narayan et al., 2020)	Single-GAN (**ours**)
UCF101	41.00%	38.42%

Table 5: Results from the Single-GAN approach for UCF101 dataset. Acc denotes mean average accuracy and Std denotes standard deviation. W2V denotes Word2Vec.

Dataset	**Semantic Embedding**	**Acc**	**Std (%)**
UCF101	Action Class W2V	28.02%	3.04%
	Description W2V	29.09%	2.61%
	GoogLeNet	44.35%	2.87%
	ResNet-101	**45.87%**	3.42%

beddings for unseen classes using various knowledge sources in the proposed Dual-GAN approach. We have considered four methods to fuse the pseudo unseen visual embeddings conditioned by the text-based and the image-based knowledge sources that are averaging, summation, maximum and minimum. For averaging, we calculate the mean of the unseen visual embedding from the text-based semantic knowledge source and the unseen visual embedding from the image-based semantic knowledge source. For summation, the same position of each element for both synthesised unseen visual embeddings is summed up. For maximum, the larger value in each position between two synthesised visual embeddings is selected. Similarly, for minimum, the smaller value in each position is selected. All four embedding fusion methods will be empirically evaluated on the dataset using the proposed Dual-GAN approach.

Evaluation Metrics Class accuracy is a standard metric in the ZSAR field. To represent the performance of the methodologies, we use the average per-class accuracies introduced by the work (Xian et al., 2017). The mean per-class accuracy averaged over 30 independent splits will be reported along with the standard deviation.

5 EXPERIMENTS

In this section, we present the experimental configurations for comparing our proposed Dual-GAN approach that incorporates two knowledge sources (i.e. texts and images) with other state-of-the-art methodologies. The implementations are then described in detail.

Experiments and Baseline For answering the first research question described in Section 1, we aim to investigate whether the synthesised visual embeddings conditioned on the image-driven knowledge source can lead to better ZSAR accuracies than those from the text-driven knowledge source using a Single-GAN model. The Single-GAN model follows only one line of the Dual-GAN pipeline (using either Text-VAEGAN or Image-VAEGAN depending on which knowledge source is used) without the process of embedding fusion illustrated in Fig. 1. Table 3 shows that two text-driven knowledge sources (i.e. class label and description) and two image-driven knowledge sources (i.e. GoogLeNet and ResNet101) will be evaluated for each dataset. As the baseline, we use the Word2Vec of action class label to represent the semantic embedding for the UCF101.

For answering the second research question introduced in Section 1 about if two sources can work better than just one, we aim to investigate and evaluate which embedding fusion method is the best. The embedding fusion methods are averaging (**Avg.**), summation (**Sum.**), maximum (**Max.**) and minimum (**Min.**). The results from Dual-GAN experiments are compared to the results from the Single-GAN to investigate whether Dual-GAN can deliver better ZSAR performance than Single-GAN.

Implementation Similar to our last work (Huang et al., 2021b), the structures of discriminator D, encoder E, and generator G are designed as fully connected networks in two layers along with 4096 hidden units. 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 supervised-learning classifier is a single-layer fully connected network with equal output units to the number of unseen classes. We apply the same hyper-parameters as our last work and the work (Narayan et al., 2020), such as α, β and σ are set to 10, 0.01 and 1, respectively. As explained 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 GAN. All experiments were conducted on Google Colab that provides Tesla P100 GPU with 25 GB memory usage.

Table 6: A comparison of Dual-GAN model with different fusion methods for UCF101. Acc and Std denote mean average accuracy and standard deviation (in %), respectively. * denotes the best result among all cases.

Dual Semantic Embedding	Avg		Sum		Max		Min	
	Acc	Std	Acc	Std	Acc	Std	Acc	Std
Action Class Word2Vec & GoogLeNet	41.20%	3.21%	41.14%	3.17%	**41.84%**	3.22%	41.06%	3.19%
Action Class Word2Vec & ResNet101	41.29%	3.34%	41.05%	3.38%	**41.95%**	3.37%	41.24%	3.33%
Description Word2Vec & GoogLeNet	45.01%	2.78%	44.73%	2.71%	**45.59%**	2.77%	44.85%	2.66%
Description Word2Vec & ResNet101	45.58%	3.00%	45.57%	3.12%	**46.37% ***	3.10%	45.37%	3.00%

Additionally, the number of synthesised visual embeddings is a hyper-parameter in the experiments. Therefore, for efficiently conducting the experiments, we synthesised 400 visual embeddings for each unseen class for the UCF101, which can yield decent results within a reasonable time duration. Our code is available online, which is compatible with Pytorch 1.9.0 and CUDA 11.1 version [2].

6 RESULTS & ANALYSIS

In this section, we present and analyse the results of empirical experiments for all configurations described in Section 5. For each configuration, the mean average accuracy is reported along with the standard deviation.

Verification of Experimental Baseline Our first experimental run is to confirm that we have set up the TF-VAEGAN experimental pipeline correctly. We compare our results to the work (Narayan et al., 2020) that our model is built upon, using identical semantic embeddings. The result is shown in Table 4. For the UCF101, the annotated class-level attributes provided by the work (Liu et al., 2011) is used and our result is decreased by 2.58%. Note that, due to the scaling limit of using annotated attributes in other datasets, attribute-based semantic information will not be used for further experiments and comparisons.

Is Image Source better than Text Source? Table 5 shows the results of evaluating the text-based (i.e. action class and textual description) and image-based (GoogLeNet and ResNet101) semantic embeddings on our Single-GAN implementations. As can be seen, the Single-GAN results for the UCF101 are expected to our hypothesis as the image-based ResNet101 semantic embedding outperforms action class Word2Vec, description Word2Vec and image-based GoogLeNet by large margins of 17.85%, 16.78% and a small margin of 1.52%, respectively. The video instances from UCF101 have a clean background with single and centred actors, which can be accurately represented by either textual descriptions or relevant images. Moreover, ResNet101 can deliver a slight boost than GoogLeNet due to better model capability of generalisation. In addition, we suggest that using textual descriptions for action classes has the potential risk of reducing the model performance, which depends on how well representative video samples are.

Is the Dual-GAN approach better than the Single-GAN? As can be seen in Table 6, the *Max.* fusion method obviously surpasses others for all *Dual Semantic Embeddings* cases in the UCF101 where the *Max.* fusion of descriptions and ResNet101 delivers the best performance at 46.37%, which surpasses the baseline (i.e. action class Word2Vec in the Single-GAN model) by a large margin of 18.35%. We suggest that the textual descriptions used to represent the semantic embedding of the class has a positive impact on performing the ZSAR. Additionally, as shown in Fig. 3, the *Max.* also performs the best on average level.

For further investigations, we compare our best results to the existing approaches that follow the GAN-based framework on the UCF101 dataset, presented in Table 7. Our Dual-GAN model outperforms other approaches up to 5.37% for the UCF101. There is no doubt that fusing embeddings derived from different

[2] https://github.com/kaiqiangh/kg_gnn_gan

Table 7: A comparison of ZSAR performance among our best results and the existing approaches (generative-based) for the UCF101 dataset.

	GMM (Mishra et al., 2018)	CLSWGAN (Xian et al., 2018)	CEWGAN (Mandal et al., 2019)	f-VAEGAN (Xian et al., 2019)	TF-VAEGAN (Narayan et al., 2020)	Dual-GAN **(ours)**
UCF101	20.3%	37.5%	38.3%	38.2%	41.0%	**46.37%**

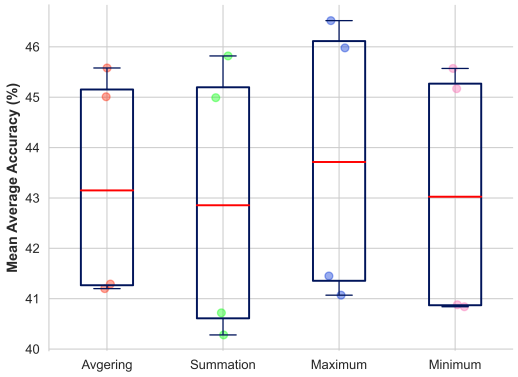

Figure 3: A comparison of Dual-GAN using different fusion methods in UCF101.

knowledge sources (i.e. texts and images) delivers a performance boost in the ZSAR. Note that, we do not re-implement and evaluate other approaches, but directly report the results from the work (Narayan et al., 2020).

As a result, we summarise our main findings as follows: (1) The image-driven semantic embedding is not absolutely better than the text-driven one, which depends on how the quality of video samples is. (2) All cases of using the Dual-GAN model outperform their counterpart cases of using Single-GAN since the fused semantic embedding obtained from two knowledge sources is capable of producing more representative semantics to the classes. (3) The *Max.* fusion method generally performs better than other methods in most cases. Additionally, the limitation of this work is that we do not fine-tune the proposed Dual-GAN model by optimising the hyper-parameters, such as the number of synthesised visual embeddings of unseen classes.

7 CONCLUSIONS

In this work, we have empirically evaluated the ZSAR performances using either text-driven or image-driven semantic embeddings related to the action classes in the GAN-based framework on UCF101. We also have investigated the impact of combining both text and image knowledge by applying different fusion methods (i.e. averaging, summation, maximum, minimum).

We have proven that applying the image-driven semantic embedding can deliver significant boosts against the text-driven one within a range between 15.26% (GoogLeNet against Description) and 17.85% (ResNet101 against Action Class) in the Single-GAN framework for UCF101. Furthermore, our proposed Dual-GAN model outperforms the baseline (i.e. action class in the Single-GAN model) by large margin of 18.35%, as well as against the existing GAN-based approaches improved by 5.37%.

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 other approaches to produce more enriched and meaningful semantic embedding that can also mitigate the problem of the semantic gap between classes and video samples. We are also planning to use other fusion methods such as concatenation or using two different classifiers and calculating the predicted class as a combination of both classifiers. Lastly, we plan to use other supervised methods such as Random Forest, Support Vector Machines, or Deep Learning to see if they are able to deliver better results.

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*.

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.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Pro-*

ceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.

Hong, M., Li, G., Zhang, X., and Huang, Q. (2020). Generalized zero-shot video classification via generative adversarial networks. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2419–2426.

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

Huang, K., Luis, Miralles-Pechuán, B., and Mckeever, S. (2021b). Zero-shot action recognition with knowledge enhanced generative adversarial networks. In *In Proceedings of the 13th International Joint Conference on Computational Intelligence*, pages 254–264.

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.

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*, 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*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

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

Wang, Q. and Chen, K. (2017). 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.

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.

Xiang, H., Xie, C., Zeng, T., and Yang, Y. (2021). Multi-knowledge fusion for new feature generation in generalized zero-shot learning. *arXiv preprint arXiv:2102.11566*.

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.