Out-of-Distribution Detection for Generalized Zero-Shot Action Recognition

Devraj Mandal*[1], Sanath Narayan*[2], Saikumar Dwivedi[3], Vikram Gupta[3], Shuaib Ahmed[3],
Fahad Shahbaz Khan[2], and Ling Shao[2]
[1]Indian Institute of Science, Bangalore [2]Inception Institute of Artificial Intelligence, UAE
[3]Mercedes-Benz R&D India, Bangalore
[1]devrajm@iisc.ac.in [2]firstname.lastname@inceptioniai.org
[3]firstname.lastname@daimler.com

Abstract

Generalized zero-shot action recognition is a challenging problem, where the task is to recognize new action categories that are unavailable during the training stage, in addition to the seen action categories. Existing approaches suffer from the inherent bias of the learned classifier towards the seen action categories. As a consequence, unseen category samples are incorrectly classified as belonging to one of the seen action categories. In this paper, we set out to tackle this issue by arguing for a separate treatment of seen and unseen action categories in generalized zero-shot action recognition. We introduce an out-of-distribution detector that determines whether the video features belong to a seen or unseen action category. To train our out-of-distribution detector, video features for unseen action categories are synthesized using generative adversarial networks trained on seen action category features. To the best of our knowledge, we are the first to propose an out-of-distribution detector based GZSL framework for action recognition in videos. Experiments are performed on three action recognition datasets: Olympic Sports, HMDB51 and UCF101. For generalized zero-shot action recognition, our proposed approach outperforms the baseline [33] with absolute gains (in classification accuracy) of 7.0%, 3.4%, and 4.9%, respectively, on these datasets.

1. Introduction

Zero-shot learning (ZSL) is a challenging problem, where the task is to classify images or videos into new categories that are unavailable during the training stage. Generalized zero-shot learning (GZSL), introduced in [34], differs from ZSL in that the test samples can belong to the seen or unseen categories. The task of GZSL is therefore harder than ZSL due to the inherent bias of the learned classifier towards the seen categories. In this paper, we focus on the problem of generalized zero-shot action recognition in videos and treat ZSL as a special case of GZSL.

Most existing approaches [14, 12, 31, 6] tackle the problem of action recognition in videos in a fully-supervised setting. In such a setting, all the action categories that occur during testing are known *a priori*, and instances from all action categories are available during training. However, the fully-supervised problem setting is unrealistic for many real-world applications (*e.g*., automatic tagging of actions in web videos), where information regarding some action categories is not available during training. Therefore, in this work we tackle the problem of action recognition under zero-shot settings.

Contrary to action recognition in videos, extensive research efforts have been dedicated to zero-shot image classification. Most earlier ZSL approaches are based on attribute mapping [2, 15]. On the other hand, a few recent works [10, 18] tackle the problem in a transductive manner, by assuming access to the full set of unlabelled testing data. This helps in decreasing the domain shift problem, in ZSL, caused due to disjoint categories in training and testing. Similar transductive strategies have also been explored for action recognition in videos [36, 24] to reduce the bias towards seen action categories. However, these approaches require unlabelled testing data for fine-tuning the parameters. Further, the bias still exists due to the similar treatment of both seen and unseen categories (see Fig. 1(a)). Instead, we propose a GZSL framework to separate the classification step for the seen and unseen action classes by introducing an out-of-distribution (OD) detector. As a result, the inherently-learned bias towards the seen classes in the action classifier is reduced (see Fig. 1(b)).

In our approach, the out-of-distribution (OD) detector is learned to produce a non-uniform distribution with an emphasis (peaks) for seen categories and a uniformly distributed output for the unseen categories. This is achieved

∗Equal Contribution. [Code at https://github.com/naraysa/gzsl-od]
[1]Work done partially during internship at Mercedes-Benz R&D India.

by utilizing an entropy loss to train our OD detector for maximizing the entropy of the output for unseen action category features. During inference, the entropy of the detector's output is compared to a specified threshold for determining whether the test feature belongs to a seen or unseen action category. Consequently, the test feature is dynamically routed to either of the two classifiers explicitly trained over seen and unseen classes, respectively, for final classification. Entropy loss has previously been used [30] to train generative adversarial networks [11] (GAN) for image synthesis, in both unsupervised and semi-supervised settings. However, to the best of our knowledge, we are the first to propose the use of entropy loss in the construction of an OD detector for generalized zero-shot action recognition.

The proposed OD detector requires features from both seen and unseen action classes to avoid an assumption on the prior data distribution. However, unseen action features are not available during training. Thus, we propose to synthesize unseen action features, to train our OD detector, by adapting a conditional Wasserstein GAN [4] (WGAN) with additional terms: cosine embedding and cycle-consistency losses. The additional loss terms aid in improving the feature generation process for a diverse set of action categories. In our work, both the generator and discriminator of the WGAN are conditioned on the category-specific auxiliary descriptions, called *class-embeddings* or *attributes*[1], to synthesize class-specific action features. Consequently, our OD detector and the two action classifiers (seen and unseen) are trained using real and synthesized features from seen and unseen categories, respectively.

Contributions: We introduce a novel generalized zero-shot action recognition framework based on an out-of-distribution (OD) detector. Our OD detector is designed to reduce the effect of the inherent bias towards the seen action classes generally present in the standard GZSL framework. To synthesize unseen features for our OD detector training, we adapt the conditional Wasserstein GAN with additional loss terms. To the best of our knowledge, we are the first to introduce a GZSL action recognition framework based on an OD detector trained using real features from seen action categories and synthesized features from unseen action classes. Our OD detector efficiently discriminates the semantically similar seen and unseen action categories, leading to improved action classification. Our approach sets a new state-of-the-art for generalized zero-shot action recognition on three benchmarks.

2. Related Work

ZSL and GZSL have gained considerable attention in recent years since they can deal with challenging real-world problems, such as automatic tagging of images and videos with new categories previously unseen during training. Earlier approaches [2, 15, 16] for ZSL in images were based on direct or indirect attribute mapping between instances and their class attributes. Alternatively, several more recent works [26, 7, 1] determine the unseen classes based on the weighted combination of seen classes. In GZSL, obtaining realistic and discriminative training data for unseen classes to overcome the classifier's bias towards the seen classes is a challenge. Synthesizing visual features of unseen instances through an embedding-based matrix mapping to convert the ZSL problem to a typical supervised problem was explored in [20, 21]. Approaches such as [5, 33, 9] have used different variants of GANs [11] to generate synthetic unseen class features for the task of GZSL. Similar to [33, 9], we adapt the conditional WGAN [4] in our framework for generalized zero-shot action recognition.

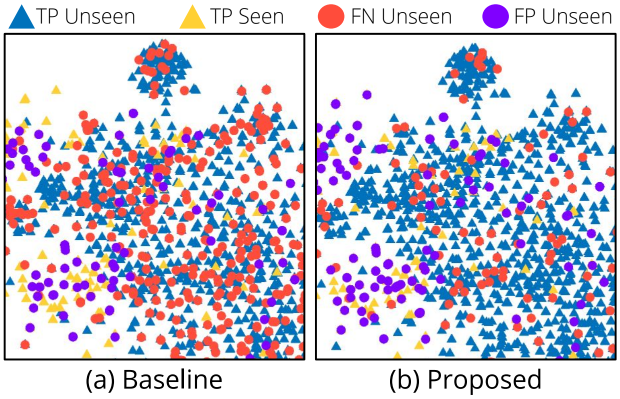

Figure 1. Illustration of the bias reduction achieved by the proposed framework on a random test split of the HMDB51 dataset. On the left: t-SNE scatter plot for baseline generalized zero-shot action recognition framework [33]. On the right: t-SNE scatter plot for our approach based on an OD detector. Action categories are grouped into seen and unseen classes for illustration. The baseline GZSL [33] incorrectly classifies several unseen category features (denoted by 'FN Unseen') into seen action categories. Our approach significantly reduces the bias towards seen categories, resulting in accurate action recognition. Best viewed in color.

In contrast to image classification, the problem of ZSL and GZSL for action recognition in videos has received less attention. Existing works pose the problem of ZSL and GZSL action recognition in the transductive setting, where unlabelled test data is also used during training [36, 13, 24]. A generative approach using Gaussians was used to synthesize unseen class data in [24], where each action is represented as a probability distribution in the visual space. These works do not treat seen and unseen action classes separately, as proposed in this work. Further, these methods use unlabelled real features from the unseen classes to rectify the bias of the learned parameters towards the seen classes. Unlike these approaches, we do not use any unlabelled real features from unseen action classes in the

[1] Both these terms are used interchangeably in this work

training stage of our model. In [37], action recognition under ZSL was addressed using a Fisher vector representation of traditional features and two-stream deep features with GloVE [27] class embedding. However, the more challenging problem of GZSL action recognition was not addressed. A one-to-one comparison using different features, such as C3D [31], I3D [6], also remains unexplored in the context of GZSL in these approaches.

Out-of-distribution detectors [17, 8] have been investigated in the context of image classification via cross-dataset evaluation. In [17], instances that appear to be at the boundary of the data manifold were used as out-of-distribution examples during training while [8] used the misclassified in-distribution samples as a proxy for out-of-distribution samples to calibrate the detector. However, in our approach, no such prior data distribution assumptions are made. Further, these detectors [17, 8] consider in-distribution samples from one image classification dataset and out-of-distribution samples from a different dataset, while our detector aims to distinguish between the seen and unseen class features of the same dataset.

Our approach: Different to the aforementioned works, an out-of-distribution detector is trained, with entropy loss, using GAN generated features of unseen action categories (as out-of-distribution samples) to recognize whether a feature sample belongs to either the seen or unseen group. Our method assumes no prior data distribution of the seen and unseen categories. The GAN itself is trained using the real features of seen categories, conditioned on the associated class-attributes of seen classes. During inference, based on the out-of-distribution detector's decision, features from a test instance are input to one of the two classifiers explicitly trained over seen and unseen action categories, respectively.

3. Proposed Approach

The proposed framework for GZSL is detailed in this section. The framework is divided into two parts: synthetic video feature generation for unseen classes using GANs (Sec. 3.1) and out-of-distribution (OD) classifier learning (Sec. 3.2). The illustration of the overall pipeline is shown in Fig. 2.

Let $\mathcal{S} = \{(x, y, e(y)|x \in \mathcal{X}, y \in \mathcal{Y}^s, e(y) \in \mathcal{E}\}$ be the training set for seen classes, where $x \in \mathbb{R}^{d_x}$ denotes the spatio-temporal CNN features, y denotes the class labels in $\mathcal{Y}^s = \{y_1, \ldots, y_S\}$ with S seen classes and $e(y) \in \mathbb{R}^{d_e}$ denotes the category-specific embedding that models the semantic relationship between the classes. Additionally, $\mathcal{U} = \{(u, e(u)|u \in \mathcal{Y}^u, e(u) \in \mathcal{E}\}$ is available during training, where u is a class from a disjoint label set $\mathcal{Y}^u = \{u_1, \ldots, u_U\}$ of U labels, and the corresponding videos or features are not available. The task in GZSL is to learn a classifier $f_{gzsl} : \mathcal{X} \to \mathcal{Y}^s \cup \mathcal{Y}^u$. Using the OD detector, this task can be reformulated into learning 3 classifiers: the out-of-distribution classifier $f_{od} : \mathcal{X} \to \{0, 1\}$ and the seen and unseen classifiers $f_s : \mathcal{X} \to \mathcal{Y}^s$ and $f_u : \mathcal{X} \to \mathcal{Y}^u$, respectively. The classifier f_{od} will determine if the feature is an in-distribution or out-of-distribution feature and route it to either f_s or f_u to determine the class.

3.1. Generating unseen class features

Given the training data of seen classes, $\mathcal{S}$, the goal is to synthesize features belonging to unseen classes, $\tilde{x}$, using the class attributes, $e(u)$. To this end, a generative adversarial network (GAN) is learned using the seen class features, x, and the corresponding class embedding, $e(y)$. A GAN [11] consists of a generator G and a discriminator D, which compete against each other in a two player minimax game. In the context of generating video features, D attempts to accurately distinguish real-video features from synthetically generated features, while G attempts to fool the discriminator by generating video features that are semantically close to real features. Since we need to synthesize features specific to unseen action categories, we use the conditional GAN [23] by conditioning both G and D on the embedding, $e(y)$. A conditional generator $G : \mathcal{Z} \times \mathcal{E} \to \mathcal{X}$ takes a random Gaussian noise $z \in \mathcal{Z}$ and a class embedding $e(y) \in \mathcal{E}$. Once the generator is learned, it is used to synthesize the video features of unseen classes, u, by conditioning on the unseen class embedding, $e(u)$. Further, we use the Wasserstein GAN [4] for the proposed framework due to its more stable training and recent success in [33, 9] for zero-shot image classification.

A conditional WGAN [4], conditioned on the embedding $e(y)$, is learned to synthesize the video features $\tilde{x}$, given the corresponding class embedding, $e(u)$. The conditional WGAN loss is given by

$$\begin{aligned}\mathcal{L}_{WGAN} = \mathbb{E}[D(x, e(y))] - \mathbb{E}[D(\tilde{x}, e(y))] - \\ \alpha\mathbb{E}[(||\nabla_{\hat{x}} D(\hat{x}, e(y))||_2 - 1)^2]\end{aligned} \tag{1}$$

where $\tilde{x} = G(z, e(y))$, $\hat{x}$ is a convex combination of x and $\tilde{x}$, α is the penalty coefficient and $\mathbb{E}$ is the expectation. The first two terms approximate the Wasserstein distance in equation 1, with the third term being the penalty for constraining the gradient of D to have unit norm along the convex combination of real and generated pairs. Additionally, we expect the generated features to be sufficiently discriminative such that the class embedding that generated them can be reconstructed back using the same features [38]. To this end, similar to [9], a decoder is used to reconstruct the class embedding $e(y)$ from the synthesized features $\tilde{x}$. Hence, a cycle-consistency loss is added to the loss formulation, which is given by,

$$\mathcal{L}_{cyc} = \mathbb{E}[||\hat{e}(y) - e(y)||_2] \tag{2}$$

where $\hat{e}(y)$ is the reconstructed embedding. Further, the synthesized features of a particular class y_i should be sim-

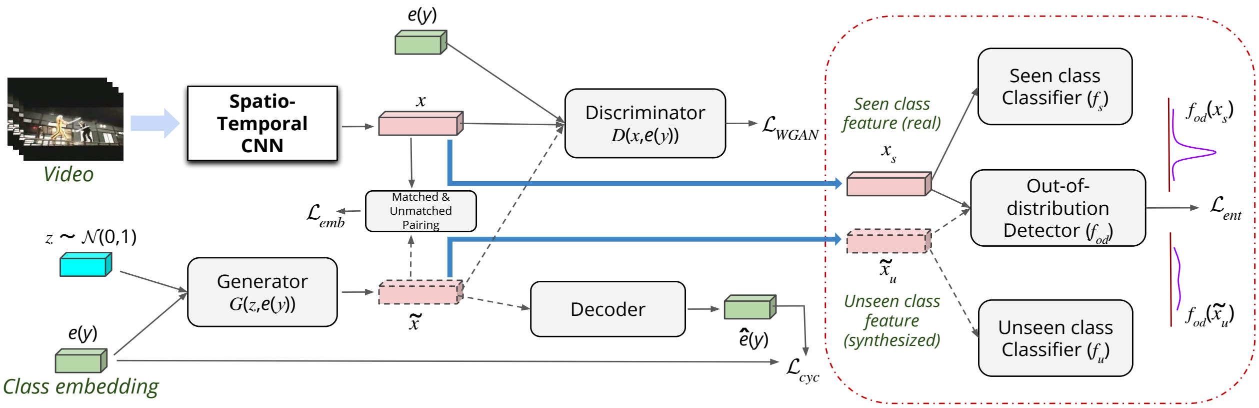

Figure 2. Illustration of the proposed GZSL approach: A conditional WGAN is trained to synthesize video features $\tilde{x}$, conditioned on the class embedding $e(y)$ via the losses $\mathcal{L}_{WGAN}$, $\mathcal{L}_{cyc}$ and $\mathcal{L}_{emb}$. A spatio-temporal CNN computes the real features x for the seen class videos. During post-training, the generator, conditioned on the unseen class embedding $e(u)$, synthesizes unseen class features $\tilde{x}_u$, which, along with real features x_s, are used to learn the three classifiers f_{od}, f_s and f_u. The expected outputs of f_{od} for seen and unseen class features are also portrayed. Cuboids with dashed borders denote synthesized features. Dashed arrows indicate their corresponding path.

ilar to the real features of the same class and dissimilar to the features of other classes y_j (for $j \neq i$). To this end, we first pair the real and synthesized features in a mini-batch to generate matched (same classes) and unmatched (different classes) pairs. Then, we minimize and maximize the distance between the matched and unmatched features, respectively, using the cosine embedding loss, as given by,

$$\mathcal{L}_{emb} = \mathbb{E}_m[1 - cos(x, \tilde{x})] + \mathbb{E}_{um}[\max(0, cos(x, \tilde{x}))] \quad (3)$$

where the respective expectations are over the matched (m) and unmatched (um) pair distributions. While the other losses ($\mathcal{L}_{WGAN}$ and $\mathcal{L}_{cyc}$) train the network by emphasizing the similarity between real and generated features of a particular class, the embedding loss also trains the network by emphasizing how the generated features of an action class should be dissimilar to the other class features. The final objective for training the GAN, using β and γ as hyper-parameters for weighting the respective losses, is given by

$$\min_G \max_D \mathcal{L}_{WGAN} + \beta\mathcal{L}_{cyc} + \gamma\mathcal{L}_{emb} \quad (4)$$

3.2. Out-of-distribution detector for unseen class

An out-of-distribution detector is proposed for differentiating between the features belonging to the seen classes and those belonging to unseen classes. After training the GAN using the training data $\mathcal{S}$, the generator (G) is used to synthesize features, $\tilde{x} = G(z, e(u))$, for the unseen categories $u \in \mathcal{Y}^u$. A training set of generated features, $\tilde{\mathcal{U}} = \{(\tilde{x}, u, e(u))\}$, is obtained by generating sufficient features for all the unseen action categories.

The real features of the seen classes, x_s and the generated features of the unseen classes, $\tilde{x}_u$, are used to train the out-of-distribution detector. Approaches in [17, 8] learn an OD detector with a prior data distribution assumption of the seen class features. However, using generated samples of the unseen classes can help to better learn the boundaries between the seen and unseen categories, without assuming any prior data distribution. The detector is a fully-connected network with the dimension of the output layer the same as the number of seen classes, S. As shown in Sec. 4.2, a binary classifier is insufficient to learn this task due to the complex boundaries between the many seen and unseen classes. Instead of attempting to directly predict whether the input is from a seen or unseen class, we use the concept of entropy to learn an embedding that projects the features of the seen and unseen classes far apart in the *entropy space*. The network is trained with entropy loss, $\mathcal{L}_{ent}$, as given by

$$\mathcal{L}_{ent} = \mathbb{E}_{x \sim \mathcal{S}}[H(p_s)] - \mathbb{E}_{\tilde{x} \sim \mathcal{U}}[H(\tilde{p}_u)] \quad (5)$$

where $H(p)$=$-\sum_i p[i] \log(p[i])$ is the entropy of p, and p_s=$f_{od}(x_s)$ and $\tilde{p}_u$=$f_{od}(\tilde{x}_u) \in \mathbb{R}^S$ are the predictions of the network for the seen and unseen features x_s and $\tilde{x}_u$, respectively. Further, a negative log-likelihood term $N(p_s)$=$-\log(p_s[y_s])$, where y_s is the class label of x_s, is added to Eq. 5 for faster convergence. This type of loss formulation models the output of the network such that its entropy is minimum and maximum for the input features of seen and unseen classes, respectively. The higher the entropy, the higher the uncertainty. Thus, a seen class feature input will have a non-uniformly distributed output (with an emphasis on seen classes). Similarly, an unseen class feature will have a near-uniform distribution as its output. The expected output of the classifier, f_{od}, for the seen and unseen class features is illustrated in the far-right side of Fig. 2.

Seen and unseen classifiers: Alongside the OD detector training, we also train two separate classifiers, one for the seen classes and one for the unseen classes. The two classifiers f_s and f_u are trained on real features of seen classes x_s and generated features of unseen classes $\tilde{x}_u$, respectively. During inference, the test video is passed through a spatio-temporal CNN to compute the real features x_{test} and then sent to the OD detector. If the entropy of the output $f_{od}(x_{test})$ is less than a threshold ent_{th}, the feature x_{test} is passed through the seen-classes classifier f_s in order to predict the label of the test video. If the entropy of $f_{od}(x_{test})$ is greater than ent_{th}, then the label is predicted using the unseen-classes classifier f_u. In ZSL, where the test samples are restricted to belonging to unseen classes, only the unseen-classes classifier f_u is required to predict the category of the video. In summary, the OD detector separates the classification of seen and unseen categories and reduces the bias towards seen categories.

4. Experiments

4.1. Experimental setup

Video features: Two types of video features, I3D [6] (Inflated 3D) and C3D [31] (Convolution 3D), designed for generic action recognition, are used for evaluation. The appearance and flow I3D features are extracted from the *Mixed_5c* layer output of the RGB and flow I3D networks, respectively. Both networks are pretrained on the Kinetics dataset [6]. For an input video, the *Mixed_5c* output of both networks are averaged across the temporal dimension and pooled by 4 in the spatial dimension and then flattened to obtain a vector, of size 4096, representing the appearance and flow features, respectively. The appearance and flow features are concatenated to obtain video features of size 8192. We use the C3D model, pre-trained on the Sports-1M dataset [12], to extract the C3D features for representing the actions in a video. A video is divided into non-overlapping 16-frame clips and the mean of the *fc6* layer outputs, of size 4096, is taken as the video feature for the action.

Network architecture: The generator G is a three-layer fully-connected (FC) network with an output layer dimension equal to the size of the video feature. The hidden layers are of size 4096. The decoder is also a three-layer FC network with an output size equal to the class-embedding size and a hidden size equal to 4096. The discriminator D is a two-layer FC network with the output size equal to 1 and a hidden size equal to 4096. The individual classifiers f_s and f_u are single-layer FC networks with an input size equal to the video feature size and output sizes equal to the number of seen and unseen classes, respectively. The OD detector f_{od} is a three-layer FC network with output and hidden layer sizes equal to the number of seen classes and 512, respectively. The parameters β and γ are set to 0.01 and 0.1, respectively, for all the datasets. The threshold value ent_{th} is chosen to be the average of the prediction entropies of the seen class features in the training data. All the modules are trained using the Adam optimizer with a 10^{-4} learning rate.

Dataset	#Videos	#Class	Split (Seen / Unseen)
Olympic Sports	783	16	8/8
HMDB51	6766	51	26/25
UCF101	13320	101	51/50

Table 1. Datasets used for evaluation

Datasets: Three challenging video action datasets (Olympic Sports [25], HMDB51 [14] and UCF101 [29]), widely used as benchmarks for GZSL and ZSL, are used for evaluating the performance of the proposed technique. The details of the three datasets are given in Tab. 1. The mean per-class accuracy averaged over 30 independent test runs is reported along with the standard deviation. Each test run is carried out on a random split of the seen and unseen classes in the dataset. For GZSL, we also report the mean accuracy for the seen classes, mean accuracy of the unseen classes and the harmonic mean of the two. For the GZSL setting, the test data consists of all the videos belonging to unseen classes and a random subset of 20% videos from seen class categories.

Class-embedding: We use two types of class-embedding to semantically represent the classes: the human annotated attributes and *word vectors* [22]. The UCF101 and Olympic Sports datasets also have manually-annotated class attributes of sizes 40 and 115, respectively. A skip-gram model, trained on the news text corpus provided by Google, is used to generate the action class-specific word vector representations of size 300 using the action category names as input. The HMDB51 dataset does not have any associated manual attributes.

4.2. Baseline comparison

The proposed framework is compared with the baseline by evaluating on the generalized zero-shot action recognition task using I3D concatenated features. Since our GAN framework for synthesizing features also uses the WGAN [4], we choose f-CLSWGAN [33], originally designed for zero-shot image classification, as the baseline. The performance comparison for the three datasets is shown in Tab. 2. We also compare our GZSL framework with and without the OD detector (denoted as CEWGAN-OD and CEWGAN, respectively, in Tab. 2). Further, to quantify the effectiveness of our OD detector, we also combine CEWGAN with a binary OD classifier, OD$_{bin}$. The classification accuracy for the seen and unseen categories and their harmonic mean are denoted by s, u and H , respectively.

The proposed OD detector (OD$_{ent}$) always outperforms the binary OD detector (OD$_{bin}$) (see Tab. 2), proving that a binary classifier is not sufficient for learning the task. The OD$_{bin}$ requires generated features for seen and unseen classes to achieve reasonable performance and it still fares, generally, worse than CEWGAN. It only yields better re-

		Embed	Olympic Sports			HMDB51			UCF101		
			s	u	H	s	u	H	s	u	H
(a)	f-CLSWGAN* [33]	w2v	66.0	35.5	46.1	52.6	23.7	32.7	74.8	20.7	32.4
		manual	72.3	51.1	59.9	-	-	-	83.9	30.2	44.4
(b)	$\mathcal{L}_{WGAN} + \mathcal{L}_{cyc} + \mathcal{L}_{emb}$ (**Ours: CEWGAN**)	w2v	67.6	36.5	47.4	51.7	24.9	33.6	73.7	21.8	33.7
		manual	73.7	52.3	61.1	-	-	-	80.2	31.7	45.5
(c)	(b) + OD$_{bin}$	w2v	71.6	38.1	49.8	36.7	24.1	29.1	62.4	19.2	29.4
		manual	72.1	56.9	63.6	-	-	-	67.4	28.2	39.8
(d)	(b) + OD$_{ent}$ (**Ours: CEWGAN-OD**)	w2v	73.2	41.8	**53.1**	55.6	26.8	**36.1**	75.9	24.8	**37.3**
		manual	71.5	61.6	**66.2**	-	-	-	76.7	36.4	**49.4**

Table 2. Comparison of proposed approach with the baseline f-CLSWGAN* [33] (* - adapted implementation) using concatenated I3D features for GZSL action recognition. CEWGAN-OD and CEWGAN denote the proposed framework with and without the out-of-distribution (OD) detector, respectively. OD$_{bin}$ and OD$_{ent}$ denote the binary classifier and proposed OD detectors, respectively. Higher is better. Manual attributes are not available for HMDB51. s, u and H denote the accuracy for seen and unseen classes and their harmonic mean, respectively. CEWGAN outperforms the baseline f-CLSWGAN on all datasets. Integrating OD$_{ent}$ with CEWGAN achieves further gains.

sults than CEWGAN for the Olympic Sports dataset. The main reason is that Olympic Sports has only eight seen and unseen classes. Hence, it is easier to separate the corresponding test features. As the number of classes increases, OD$_{bin}$ fails to accurately separate the seen and unseen category features.

Importantly, we see that the proposed GAN (CEWGAN) performs better than the baseline approach (f-CLSWGAN) on all combinations of datasets and attributes. Integrating the proposed OD detector (OD$_{ent}$) with CEWGAN further improves the performance across datasets. Average gains of 7.0%, 3.4%, and 4.9% (in terms of accuracy) are achieved over f-CLSWGAN [33] for the Olympic Sports, HMDB51 and UCF101 datasets, respectively, using *word2vec*. Achieving a considerable gain on a difficult dataset, such as HMDB51, shows the promise of our framework for generalized zero-shot action recognition.

4.3. State-of-the-art comparison

In this section, a comparison of our proposed framework against the other approaches for the tasks of ZSL and GZSL in action recognition is given. Since our aim is reducing the bias of the classifier towards seen classes in generalized zero-shot action recognition, we first compare the GZSL performance (Tab. 3), and then the ZSL performance (Tab. 4), with the other approaches in literature. In both the tables, we report the performance of our approach trained using the I3D (appearance + flow) features. The performance of our approach using other features is given as an ablation study in Sec. 4.6.

GZSL performance comparison: The proposed out-of-distribution detector is applicable only in the GZSL framework. The comparison of our proposed approach with the other approaches on the GZSL task is reported in Tab. 3. The best results for each dataset and attribute combination are in boldface. The standard deviation from the mean is also reported. We see that the proposed approach, CEWGAN-OD, outperforms the other approaches (fewer approaches compared to the ZSL task) on all datasets. The results for CLSWGAN [33] are obtained by adapting the author's implementation for our GZSL action recognition task. This is denoted by the superscript '*' in Tab. 3. Both CLSWGAN and the proposed approach are trained using the I3D features. The best existing approach for GZSL action recognition, GGM [24], employs a generative approach to synthesize unseen class data and utilizes unlabelled real features (C3D) from the unseen classes to rectify the bias of the learned parameters towards seen classes. Particularly, for the UCF101 dataset and manual attributes combination, the proposed approach, CEWGAN-OD, achieves gains of 5.1% and 25.8% (in terms of accuracy) over the CLSWGAN [33] and GGM [24], respectively. Further, for the *word2vec* embedding, the proposed CEWGAN-OD achieves gains of 16% and 19.8% over the best existing approach, GGM [24], for the HMDB51 and UCF101 datasets, respectively.

ZSL performance comparison: In Tab. 4, the proposed approach trained using the I3D (appearance + flow) features is denoted by CEWGAN. Here, the suffix OD (used in Tab. 3) is dropped since the out-of-distribution detector is applicable only in the GZSL task. From Tab. 4, we see that our approach outperforms the other approaches in the zero-shot action recognition task for all combinations of datasets and attributes. The proposed approach, CEWGAN, in general, has less or comparable deviation as the other approaches. This shows that the proposed approach consistently improves across the splits. All the other approaches use either the *word2vec* or manually-annotated embedding (denoted by *w* and *m*, respectively) except MICC [37], which uses *GloVE* [27], an embedding similar to *word2vec*. The proposed approach using I3D features and the *word2vec* embedding has absolute gains of 6.6%,

Method		Olympics	HMDB51	UCF101
HAA [19]	*m*	49.4±10.8	-	18.7±2.4
SJE [3]	*w*	32.5±6.7	10.5±2.4	8.9±2.2
ConSE [26]	*w*	37.6±9.9	15.4±2.8	12.7±2.2
GGM [24]	*m*	52.4±12.2	-	23.7±1.2
	w	42.2±10.2	20.1±2.1	17.5±2.2
CLSWGAN* [33]	*m*	59.9±5.3	-	44.4±3.0
	w	46.1±3.7	32.7±3.4	32.4±3.3
Ours: CEWGAN-OD	*m*	**66.2±6.3**	-	**49.4±2.4**
	w	**53.1±3.6**	**36.1±2.2**	**37.3±2.1**

Table 3. GZSL performance comparison (in %) with existing approaches. *m* and *w* indicate manual attributes and *word2vec*, respectively. CLSWGAN* [33] (* - adapted implementation) and CEWGAN-OD denote the baseline and proposed approach, respectively, using I3D features. Higher is better. Best results for each embedding are in bold. Manual attributes are not available for HMDB51. CEWGAN-OD achieves an absolute gain of 5.0% over the baseline for UCF101, using manual attributes, and outperforms existing methods by a significant margin on all datasets.

Method		Olympics	HMDB51	UCF101
PST [28]	*m*	48.6±11	-	15.3±2.2
ST [35]	*w*	-	15±3	15.8±2.3
TZWE [36]	*m*	53.5±11.9	-	20.2±2.2
	w	38.6 ±10.6	19.1±3.8	18.0±2.7
Bi-dir [32]	*m*	-	-	28.3±1.0
	w	-	18.9±1.1	21.4±0.8
UDA [13]	*m*	-	-	13.2±0.6
MICC [37]	*g*	43.9±7.9	25.3±4.5	25.4±3.1
GGM [24]	*m*	57.9±14.1	-	24.5±2.9
	w	41.3±11.4	20.7±3.1	20.3±1.9
CLSWGAN* [33]	*m*	64.7±7.5	-	37.5±3.1
	w	47.1±6.4	29.1±3.8	25.8±3.2
Ours: CEWGAN	*m*	**65.9±8.1**	-	**38.3±3.0**
	w	**50.5±6.9**	**30.2±2.7**	**26.9±2.8**

Table 4. ZSL performance comparison (in %) with existing approaches. *m*, *g* and *w* indicate manual attributes, *GLoVE* and *word2vec*, respectively. CLSWGAN* [33] (* - adapted implementation) and CEWGAN denote the baseline and proposed approach, respectively, using I3D features. Higher is better. Best results for each embedding are in bold. Our approach achieves the state-of-the-art on all datasets.

4.9% and 1.5% (in terms of accuracy) over the best existing ZSL results on the Olympic Sports, HMDB51 and UCF101 datasets, respectively. Further, for the *word2vec* embedding, we observe that the proposed CEWGAN achieves gains of 1.2%, 1.1% and 1.1% over the CLSWGAN [33] for the same datasets, respectively. Generally, for both GZSL and ZSL tasks, using the same features but learning with manual attributes (instead of *word2vec*) results in better performance across different approaches.

4.4. Bias towards seen categories

Tab. 5 quantifies the bias reduction due to the proposed framework, CEWGAN-OD, for the three datasets, using the *word2vec* embedding. For this experiment, we consider all the features of unseen categories as one class and the remaining features from seen categories as another. A feature sample is said to be wrongly classified if the predicted class is not the same as the ground-truth class, regardless of whether the feature was classified as belonging to the correct category within each class or not. This allows us to quantify the bias reduction achieved by the standalone OD detector. We observe that CEWGAN-OD reduces the bias towards the seen categories and achieves better classification for the unseen class features. Specifically, the proposed CEWGAN-OD achieves gains of 6.2% and 10.1% over CEWGAN for the HMDB51 and UCF101 datasets, respectively, using the *word2vec* embedding.

	CEWGAN		**CEWGAN-OD**	
	SC	UC	SC	UC
Olympic Sports	68.2	72.3	**73.9**	**82.8**
HMDB51	66.7	82.5	**71.6**	**88.7**
UCF101	74.4	81.1	**76.5**	**92.2**

Table 5. Comparison of the bias towards seen classes, between the baseline (CEWGAN) and the proposed (CEWGAN-OD) frameworks on the three datasets using the *word2vec* embedding. SC, UC denote seen classes and unseen classes, respectively. Lower UC accuracy indicates higher bias towards seen categories. The proposed CEWGAN-OD achieves gains of 6.2% and 10.1% (classification accuracy) over the baseline CEWGAN for the unseen categories in the HMDB51 and UCF101 datasets, respectively.

Fig. 3 shows a comparison, in terms of the classification accuracy, between our two frameworks: CEWGAN and CEWGAN-OD. The comparison is shown for random test splits of HMDB51 and UCF101. The x-axis denotes the number of unseen class feature instances in a test split. The unseen class feature instances are sorted (high to low) according to the confidence scores of the respective classifiers (CEWGAN and CEWGAN-OD). The plot shows that integrating the proposed OD detector in the GZSL framework results in a significant improvement in performance for both datasets (denoted by green and red curves in Fig. 3). The number of unseen class feature instances incorrectly classified (into a seen class) is reduced with the integration of the proposed OD dectector. This improvement in classification performance for unseen action categories leads to a significant reduction in bias towards seen classes.

4.5. Transferring word representations

As mentioned previously in Sec. 4.1, manual attributes are not available for the HMDB51 dataset. While *word2vec* representations give a good measure of the semantic representations of the classes, learning with manual attributes always results in better performance, as can be seen from the results in Sec. 4.3 and 4.2. Here, we learn to generate the manual attributes from the *word2vec* embedding to show that using the transformed class embedding achieves bet-

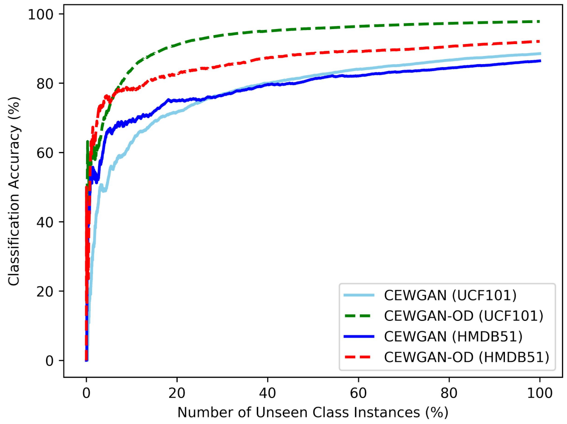

Figure 3. Classification accuracy (in %) comparison between the proposed GZSL frameworks (CEWGAN-OD and CEWGAN) for random test splits of HMDB51 and UCF101 datasets. X-axis denotes the number of unseen class instances (in %). For each framework, the unseen feature instances are sorted (in descending order) according to their respective classifier scores. Integrating the OD detector in the CEWGAN framework achieves higher classification accuracy (red and green lines) for both datasets. CEWGAN-OD decreases the bias towards seen classes. Best viewed in color.

ter generation of features, resulting in better performance compared to the *word2vec* embedding. We use the class embeddings of the UCF101 dataset to learn the transformation using a two-layer FC network. To generate a sufficient number of samples for training, the video features are concatenated with their respective *word2vec* and used as input. The trained model is then used to transform *word2vec* representations into manual attribute embeddings.

To comply with the ZSL paradigm of not using any video features from the unseen classes, we use the generated features for the HMDB51 unseen classes as input for the embedding transformation network. Here, the generator is learned using the *word2vec* embedding and the seen class features of the HMDB51 dataset. The learned attributes for HMDB51 are the same size as the manual attributes of UCF101, *i.e.*, 115. The performance of the proposed framework under ZSL and GZSL settings for the HMDB51 dataset using the transferred attributes (denoted by *m*) and different features is reported in Tab. 6. The results show that the transferred attributes for HMDB51 achieve better performance than the *word2vec*. Hence, synthesizing features using transferred attributes, for datasets without manually-annotated attributes, achieves better performance compared to synthesizing using the standard *word2vec* embedding.

4.6. Comparison of video features

Here, we give a performance comparison of the different video features for the tasks of ZSL and GZSL. The features that are used for comparison are C3D, I3D$_a$ (appearance), I3D$_f$ (flow) and I3D$_{af}$ (appearance and flow). The features are evaluated on the HMDB51 and UCF101 datasets using both the manual attributes and *word2vec* embedding. The manual attributes for HMDB51 refer to the transformed attributes, as described in Sec. 4.5. The entire setup remains the same except for the input or output layers, which depend on the video feature dimensions. The results are reported in Tab. 6. In general, we see that the I3D$_a$ features perform better than the C3D and I3D$_f$ features. The I3D$_f$ features are still better than the C3D features, while the best performance is achieved when the appearance and flow features are combined. This is in line with the performance of the features in the task of fully-supervised action recognition, as noted in [6]. This also indicates that our framework can be used with new and improved features as and when they are designed and a corresponding improvement in GZSL action recognition can be expected. The results in Tab. 3 and 4 for CEWGAN-OD and CEWGAN, respectively, use the combined features, I3D$_{af}$.

Feature		**HMDB51**		**UCF101**	
		ZSL	GZSL	ZSL	GZSL
C3D	*m*	26.0	30.9	28.1	38.7
	w	24.2	29.1	21.5	32.0
I3D$_a$	*m*	30.8	36.1	33.9	44.3
	w	28.2	33.8	23.2	33.4
I3D$_f$	*m*	29.7	34.9	32.2	42.7
	w	27.4	32.0	22.7	32.6
I3D$_{af}$	*m*	**34.8**	**39.5**	**38.3**	**49.4**
	w	**30.2**	**36.1**	**26.9**	**37.3**

Table 6. Performance comparison of C3D, I3D appearance (I3D$_a$), I3D flow (I3D$_f$) and I3D appearance+flow (I3D$_{af}$) video features on the HMDB51 and UCF101 datasets. For HMDB51, *m* denotes the transferred attributes, as discussed in Sec. 4.5. Best results are in bold for both types of embedding. For every combination of feature and attribute, ZSL and GZSL denote the performance of CEWGAN and CEWGAN-OD, respectively.

5. Conclusion

In this work, we proposed a novel out-of-distribution detector integrated into the generalized zero-shot learning action recognition framework. An out-of-distribution detector was learned to detect unseen category features as out-of-distribution samples. It was trained using real and GAN-generated features from seen and unseen categories, respectively. The use of an out-of-distribution detector enabled the classification of the seen and unseen categories to be separated and hence, reduced the bias towards seen classes that is present in the baseline approaches. The approach was evaluated on three human action video datasets, using different types of embedding and video features. The proposed approach outperformed the baseline [33] in generalized zero-shot action recognition using *word2vec*, with absolute gains of 7.0%, 3.4% and 4.9% on the Olympic Sports, HMDB51 and UCF101 datasets, respectively.

References

[1] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. Multi-cue zero-shot learning with strong supervision. In *CVPR*, 2016. 2

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label embedding for attribute-based classification. In *CVPR*, 2013. 1, 2

[3] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. 7

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 2, 3, 5

[5] Maxime Bucher, Stphane Herbin, and Frdric Jurie. Generating visual representations for zero-shot classification. In *ICCV-TASK-CV*, 2017. 2

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1, 3, 5, 8

[7] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016. 2

[8] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 3, 4

[9] Rafael Felix, B. G. Vijay Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, 2018. 2, 3

[10] Yanwei Fu, Timothy M Hospedales, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *ECCV*, 2014. 1

[11] Ian Goodfellow, Jean PougetAbadie, Mehdi Mirza, Bing Xu, and David Warde-Farley. Generative adversarial nets. In *NIPS*, 2014. 2, 3

[12] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1, 5

[13] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015. 2, 7

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011. 1, 5

[15] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes. In *CVPR*, 2009. 1, 2

[16] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 2014. 2

[17] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *ICLR*, 2018. 3, 4

[18] Xin Li, Yuhong Guo, and Dale Schuurmans. Semi-supervised zero-shot classification with label representation learning. In *ICCV*, 2015. 1

[19] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 7

[20] Yang Long, Li Liu, Ling Shao, Fumin Shen, Guiguang Ding, and Jungong Han. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In *CVPR*, 2017. 2

[21] Y. Long, L. Liu, F. Shen, L. Shao, and X. Li. Zero-shot learning using synthesised unseen visual data with diffusion regularisation. *TPAMI*, 2017. 2

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 5

[23] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[24] Ashish Mishra, Vinay Kumar Verma, M Shiva Krishna Reddy, Arulkumar S, Piyush Rai, and Anurag Mittal. A generative approach to zero-shot and few-shot action recognition. In *WACV*, 2018. 1, 2, 6, 7

[25] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010. 5

[26] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zeroshot learning by convex combination of semantic embeddings. In *ICLR*, 2014. 2, 7

[27] Jeffrey Pennington, Richard Socher, and Christopher Manning. Zero-shot action recognition with error-correcting output codes. In *EMNLP*, 2014. 3, 6

[28] Marcus Rohrbach, Sandra Ebert, and Bernt Schiele. Transfer learning in a transductive setting. In *NIPS*, 2013. 7

[29] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint*, (arXiv:1212.0402), 2012. 5

[30] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *ICLR*, 2016. 2

[31] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 1, 3, 5

[32] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *IJCV*, 2017. 7

[33] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 1, 2, 3, 5, 6, 7, 8

[34] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017. 1

[35] Xun Xu, Timothy Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. In *ICIP*, 2015. 7

[36] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *IJCV*, 2017. 1, 2, 7

[37] Chenrui Zhang and Yuxin Peng. Visual data synthesis via gan for zero-shot video classification. In *IJCAI*, 2018. 3, 6, 7

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3