A Generative Approach to Zero-Shot and Few-Shot Action Recognition

Ashish Mishra[*,1] , Vinay Kumar Verma[†,1], M Shiva Krishna Reddy[*], Arulkumar S[*]
Piyush Rai[†] and Anurag Mittal[*]
[*]Indian Institute of Technology Madras [†]Indian Institute of Technology Kanpur
{vkverma,piyush}@cse.iitk.ac.in, {mishra,shiva,aruls,amittal}@cse.iitm.ac.in

Abstract

We present a probabilistic framework for zero-shot action recognition where some of the possible action categories do not occur in the training data. This problem is of significant importance in Computer Vision since it is virtually impossible to collect training data for all the possible action classes. We propose a novel generative approach to handle this problem. Our model assumes that the parameters of the probability distribution representing any action in the visual space can be expressed as a linear combination of a set of basis vectors where the combination weights are given by the attributes of the action class. These basis vectors can be learned solely using labeled data from the known (i.e., previously seen) action classes, and can then be used to predict the parameters of the probability distributions of unseen action classes. We consider two settings: (1) Inductive setting, where we use only the labeled examples of the seen action classes to predict the unseen action class parameters; and (2) Transductive setting which further leverages unlabeled data from the unseen actions, facilitating domain adaptation. While traditional methods model only one of the two settings, we propose a very simple generative model that seamlessly combines the two in a stage-wise manner. We evaluate our model in both standard(disjoint) and generalized zero-shot learning settings.

Our framework also naturally extends to few-shot action recognition where a few labelled examples from unseen classes are provided. Our experiments on the standard datasets (UCF101, HMDB51 and Olympic) show significant performance improvement with and without domain adaptation in both standard(disjoint) and generalized settings.

1. Introduction

Action Recognition is an important problem in Computer Vision in which knowledge about a sequence of actions is learned from a large collection of video clips. It is a challenging task due to the inherent variability in actions, non-deterministic occlusion patterns, abrupt changes in illumination, cluttered dynamic background, and noisy videos. Knowledge about an action is inferred usually by learning from the labelled data in a supervised manner. Even as more complex models are being built, it is a common observation that the number of categories of actions is progressively increasing (for example, KTH has 6 categories while Olympic, HMDB and UCF datasets have 16, 51, and 101 categories, respectively). Consequently, annotating videos of this growing number of categories can be a very cumbersome task and hence restricts the scalability of supervised learning for a large number of categories.

To circumvent this problem, Zero-Shot Learning (ZSL) of actions has been actively pursued. In the conventional Action Recognition framework, only those classes present in the training data can be recognized by the model during the test phase. In Zero-Shot Learning, however, the model is expected to recognize and categorize action classes that did not appear in the training phase at all. The information about the unseen classes is provided via other modalities such as language in the form of textual descriptions, *word2vec* [22] or human annotated attributes. Essentially, the model has to learn to recognize the unseen classes based on the knowledge acquired from the data instances of the seen classes. Zero-shot learning typically categorized in two settings. First one is standard setting, in which both the seen and unseen classes are disjoint($Y_{tr} \cap T_{te} = \Phi$), which is not true in real world. Hence the generalized zero-shot(GZSL) setting has been proposed in which seen and unseen classes may occur during test time [19, 23]. The generalized zero-shot setting is harder than standard setting(disjoint setting) since the models are biased towards seen classes at training time. Most of the proposed methods consider disjoint train and the test classes. The generalized zero-shot setting relaxes this constraint.

Most zero-shot learning methods learn a linear compatibility mapping from the image space to the semantic space. Since mapping is learned from seen classes which

[1]Both authors have equal contribution.

are disjoint with unseen classes, it can lead to a domain shift problem [7, 32] between visual space and semantic space. Due to domain shift, learning an efficient mapping becomes challenging. Another challenge to Zero-Shot Action Recognition arises from the Hubness problem[4]. The hubness problem occurs in the embedding space when a few instances occur in the neighbourhood of a large number of other instances regardless of their class labels. Hence, a significant number of data instances can be potentially misclassified. For reducing the Hubness problem, many techniques have been proposed such as Domain Adaptation based learning[10], Manifold based regression model[5] but no significant improvement has been observed. In the proposed method, we present a simple generative approach that efficiently handles the above-mentioned shortcomings in the existing approaches. The complete architecture of our model is shown in figure 1. We tackle the hubness phenomenon [4] by learning a set of basis vectors in the visual space. The parameters of the distributions of each seen class are represented by a linear combination of the basis vectors weighted by the attributes of the corresponding class. The parameters of an unseen class distribution are then computed via a weighted combination of the learned basis vectors, with weights being the attributes of the respective unseen class. The loss function (More details in Methodology section) is formulated in such a way that the weighted combination of the basis vectors of the seen classes are close to the parameters obtained by the maximum likelihood estimate over the data. We add an additional regularizer to encourage reconstruction ability of the attribute vectors from the parameters of the seen class conditional densities so as to minimize the information loss. Our proposed model has a closed-form solution.

Our main contributions are as follows: (1) We provide a probabilistic generative approach for zero-shot learning (ZSL) where each action class is represented by a simple Gaussian distribution; (2) We show that our approach although simple generalizes well to the unseen classes in the inductive setting and improves over the state-of-the-art; (3) We show that our approach can be easily generalized to the transductive setting where unlabeled data from unseen classes are available at training time. (4) Since our approach is generative, so we can generate unseen class examples by using the parameters $\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c^2$ which helps the classifier in the generalized zero-shot setting which is much harder than the standard(disjoint) setting. Hence our model outperforms in the generalized setting with significant improvement over all the state-of-art methods. (5) We also extend the model to the case where a few examples of each class are available; and through extensive experimentation on three benchmark datasets, we show that our simple approach gives significant performance gains in all three settings over the state-of-the-art methods.

2. Methodology

For the zero-shot action recognition setting, we denote the total number of seen action classes by S and the total number of unseen action classes by U. We take a generative classification based approach to the action recognition problem where we assume that the data instances of each action class (seen/unseen) c are generated by a distribution $p(\mathbf{x}|\boldsymbol{\theta_c})$. Without loss of generality, and for simplicity of exposition, we will assume these distributions to be Gaussians (note that our approach can be used with other distributions as well). In the Gaussian case, the parameters $\boldsymbol{\theta_c}$ consist of the mean vector $\boldsymbol{\mu_c} \in \mathbb{R}^D$ and a diagonal covariance matrix $\boldsymbol{\sigma_c^2}$. Where $\boldsymbol{\sigma_c^2} \in \mathbb{R}^D$ is vector of diagonal covariance. In the zero-shot learning setting, we also assume that each seen/unseen class has a class attribute vector $\mathbf{a_c} \in \mathbb{R}^K$, either provided by a human expert or as the WORD2VEC embedding of the name of the action.

Given labeled data from the seen classes, it is straightforward to estimate the parameters $\boldsymbol{\mu_c}, \boldsymbol{\sigma_c}$ using Maximum Likelihood Estimation (MLE) or Maximum-a-Posteriori (MAP) estimation. For example, using MLE, the mean is estimated as $\boldsymbol{\mu_c} = \frac{1}{N_c}\sum_{i=1}^{N_c} \mathbf{x_i}$ and $\boldsymbol{\sigma_c} = diag(\frac{1}{N_c}\sum_{i=1}^{N_c}(\mathbf{x_i} - \boldsymbol{\mu_c})(\mathbf{x_i} - \boldsymbol{\mu_c})^\top)$ where N_c denotes the number of labeled examples from class c. Here $diag$ represent diagonal of the covariance matrix, since instead of full covariance, we are interested only in the diagonal covariance.

However, this approach cannot be used to estimate the parameters $\boldsymbol{\theta_c}(c = S+1, ..., S+U)$ of unseen classes due to unavailability of labeled data corresponding to unseen classes. To resolve this problem, we parametrize the $\boldsymbol{\theta_c} = (\boldsymbol{\mu_c}, \boldsymbol{\sigma_c^2})$ as a function of the class attributes $\mathbf{a_c}$, i.e., $\boldsymbol{\theta_c} = f(\mathbf{a_c})$ and learn the function f using the labeled data instances of seen classes in visual feature space. Once learned, the function f can be used to predict $\boldsymbol{\theta_c}$ for all the unseen class actions $c = S+1, \ldots, S+U$ using their respective class attributes.

One simple choice of f is a linear mapping from the class attributes $\mathbf{a_c}$ to the class parameters $\boldsymbol{\theta_c}$. In the Gaussian case, for the mean μ_c, such a linear function f can be model as

$$\boldsymbol{\mu_c} = f_\mu(\mathbf{a_c}) \quad s.t \quad \mathbf{a_c} = f'_\mu(\boldsymbol{\mu_c}), \text{where } c = 1 \ldots S \quad (1)$$

where f and ''inverse" f' are linear functions defined as:

$$f_\mu(\mathbf{a_c}) = \mathbf{W}_\mu \mathbf{a_c} \quad \text{and} \quad f'(\mu_c) = \mathbf{W_\mu^T} \boldsymbol{\mu_c}, \text{where } c = 1 \ldots S \quad (2)$$

Here $\boldsymbol{\mu}_c \in \mathbb{R}^D$ and $\mathbf{W}_\mu \in \mathbb{R}^{D \times K}$. $\mathbf{W}_\mu = [\mathbf{w}_{\mu_1}, \mathbf{w}_{\mu_2}, .., \mathbf{w}_{\mu_K}]$ is a set of *learned* basis vectors in the visual space, each column vector $\mathbf{w}_{\mu_1} \in R^D$ represents a basis vector. Given the empirical estimates of $\hat{\boldsymbol{\mu}}_c, c = 1, \ldots, S$, we can use $(\mathbf{a_c}, \hat{\boldsymbol{\mu}}_c)$ as "training data" to learn a

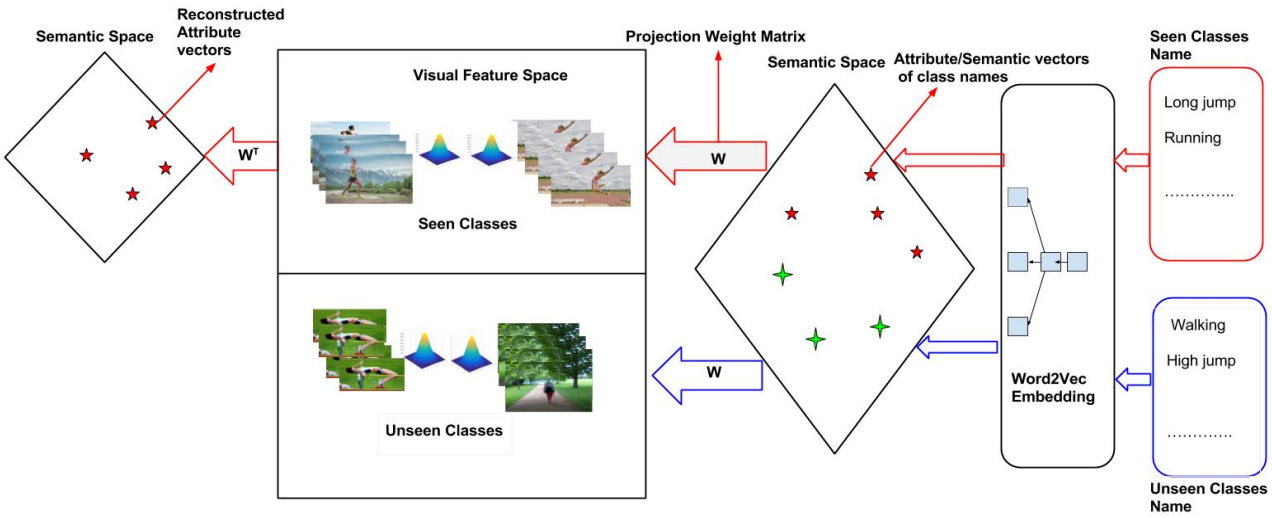

Figure 1. **Proposed Model**: Each class attribute is projected to the visual space, In the visual space each class is represented by a Gaussian distribution. To avoid information loss, a reconstruction regularizer is added.

regression model f parametrized by $\mathbf{W}_\mu$ that maps $\mathbf{a_c}$ to $\hat{\boldsymbol{\mu}}_{\boldsymbol{c}}$.

Note that the model in Eq. 2 is akin to an autoencoder [12] since

$$\boldsymbol{\mu}_c = \mathbf{W}_\mu \mathbf{a_c} = \mathbf{W}_\mu \mathbf{W}_\mu^\top \boldsymbol{\mu_c} \tag{3}$$

Once the basis vectors $\mathbf{W}_\mu, \mathbf{W}_\sigma$ (which define the function f) are learned, we can use them to estimate the data distribution parameters (e.g., $\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2$) of the unseen classes. Equation 3 shows the use of the basis vectors ($\mathbf{W}_\mu$) learned from seen class labeled data instances to estimate the distribution mean $\boldsymbol{\mu}_c$ of unseen classes.

2.0.1 Linear Regression

Given the labeled data from seen classes $c = 1, ..., S$, we can estimate their class distribution parameters using MLE. We can then learn the functions f_μ and f_{σ^2} using training data of the form $(\mathbf{a_c}, \boldsymbol{\mu_c})_{c=1}^S$ and $(\mathbf{a_c}, \boldsymbol{\sigma_c^2})_{c=1}^S$. In the linear regression approach $\boldsymbol{\mu}_c = f_\mu(\mathbf{a_c})$ and $\boldsymbol{\sigma}_c^2 = f_\sigma(\mathbf{a_c})$, we assume the functions f_μ and f_σ to be linear projections with weight matrices, $\mathbf{W}_\mu$ and $\mathbf{W}_\sigma$, making this problem equivalent to the following regression problem:

$$\boldsymbol{\mu}_c = \mathbf{W}_\mu \mathbf{a_c} \quad \text{s.t.} \quad \mathbf{a_c} = \mathbf{W}_\mu^\mathbf{T} \boldsymbol{\mu_c}$$

$$\rho_c = \log \boldsymbol{\sigma}_c^2 = \mathbf{W}_{\sigma^2} \mathbf{a_c} \quad \text{s.t.} \quad \mathbf{a_c} = \mathbf{W}_{\boldsymbol{\sigma^2}}^\mathbf{T} \boldsymbol{\sigma_c^2}$$

The projection matrices $\mathbf{W}_\mu$ and $\mathbf{W}_\sigma$ can be easily learned using a (regularized) least squares regression problem with training data $(\mathbf{a_c}, \boldsymbol{\mu_c})_{c=1}^S$ and $(\mathbf{a_c}, \boldsymbol{\sigma_c^2})_{c=1}^S$. These problems have simple closed form solution and we omit the equations here for brevity. We give details equations for the nonlinear case, as shown below.

2.0.2 Nonlinear Regression

For the non-linear regression, we first map the attributes $\{\mathbf{a_c}\}_{c=1}^S$ to the kernel space using the kernel function k which is defined as a nonlinear mapping ϕ. Using the Representer theorem [[27]], we can re-formulate the regression problem in kernel space as given in equation 4. Note that instead of computation the $\phi(\mathbf{a_c})$ explicitly, we have to compute only the dot product $\phi(\mathbf{a_c})^T \phi(\mathbf{a_{c'}}) = k(c, c')$ for the non-linear mapping of the two class c and c'. Let $\mathbf{K}$ be the kernel matrix of $S \times S$ dimension containing pairwise similarities of the attributes of the seen classes, M be the matrix containing the mean of all seen classes, then the nonlinear model f_{μ_c} for c^{th} class is obtained by:

$$\min_{\mathbf{W}_\mu} ||\mathbf{M} - \mathbf{W}_\mu \mathbf{K}||_F^2 + \lambda_\mu ||\mathbf{W}\mu||_2^2 \quad \text{s.t.} \quad \mathbf{K} = \mathbf{W}_\mu^* \mathbf{M} \tag{4}$$

Eq 4 shows our main objective function. Here the first term can be interpreted as learning an optimal weight matrix that projects the attribute space to the visual space using the kernel regression. The second term ensures that we can reconstruct the attribute vector from the visual space and acts as a regularization term. Experimentally we find that minimizing the reconstruction error gives the better generalizability to the proposed method. Therefore instead of learning the different weights for the reconstruction, we use the same shared weights. Therefore we have a new constraint:

$$\mathbf{W}_\mu^* = \mathbf{W}_\mu^T$$

Therefore the complete objective can be written as:

$$\mathbf{W}_{\mu}^{*} = \underset{\mathbf{W}_{\mu}}{\operatorname{argmin}} ||\mathbf{M} - \mathbf{W}_{\mu}\mathbf{K}||_F^2 + \lambda_{\mu}||\mathbf{W}\mu||_2^2 + \lambda_1||\mathbf{K} - \mathbf{W}_{\mu}^{\mathbf{T}}\mathbf{M}||_F^2 \quad (5)$$

For solving the Eq.[5] we are provide the proper optimization method in the below section.

2.0.3 Optimization

Using the trace property, we have $Tr(\mathbf{K}) = Tr(\mathbf{K}^T)$ and $Tr(\mathbf{W}_{\mu}^T\mathbf{M}) = Tr(\mathbf{M}^T\mathbf{W}_{\mu})$. Therefore equation 5 can be written as:

$$\mathbf{W}_{\mu}^{*} = \underset{\mathbf{W}_{\mu}}{\operatorname{argmin}} ||\mathbf{M} - \mathbf{W}_{\mu}\mathbf{K}||_F^2 + \lambda_{\mu}||\mathbf{W}\mu||_2^2 + \lambda_1||\mathbf{K}^T - \mathbf{M}^T\mathbf{W}_{\mu}||_F^2 \quad (6)$$

Taking the derivative of eq(6) and equating to zero we have.

$$\mathbf{M}\mathbf{M}^T\mathbf{W}_{\mu} + \mathbf{W}_{\mu}\lambda_1\mathbf{K}\mathbf{K}^T + \lambda_{\mu}\mathbf{W}_{\mu} = (1+\lambda_1)\mathbf{M}\mathbf{K}^T \quad (7)$$

$$\mathbf{M}\mathbf{M}^T\mathbf{W}_{\mu} + \mathbf{W}_{\mu}(\lambda_1\mathbf{K}\mathbf{K}^T + \lambda_{\mu}) = (1+\lambda_1)\mathbf{M}\mathbf{K}^T \quad (8)$$

The above equation is in the form of:

$$\mathbf{A}\mathbf{W} + \mathbf{W}\mathbf{B} = \mathbf{C} \quad (9)$$

This is a well known equation in the linear algebra which can be solved using the Bartels-Stewart algorithm [8] efficiently. Furthermore, a simple $Matlab$ function is available for solving this. The above equation is known as Sylvester equation where:

$$\mathbf{A} = \mathbf{M}\mathbf{M}^T \quad (10)$$

$$\mathbf{B} = \lambda_1\mathbf{K}\mathbf{K}^T + \lambda_{\mu} \quad (11)$$

$$\mathbf{C} = (1+\lambda_1)\mathbf{M}\mathbf{K}^T \quad (12)$$

In similar way, the nonlinear model $f_{\sigma_i^2}$ is obtained by solving:

$$\mathbf{W}_{\sigma^2}^{*} = \underset{\mathbf{W}_{\sigma^2}}{\operatorname{argmin}} ||\mathbf{R} - \mathbf{W}_{\sigma^2}\mathbf{K}||_F^2 + \lambda_{\sigma^2}||\mathbf{W}_{\sigma^2}||_2^2 + \lambda_2||\mathbf{K} - \mathbf{W}_{\sigma^2}^T\mathbf{R}||_F^2 \quad (13)$$

$$\mathbf{R}\mathbf{R}^T\mathbf{W}_{\sigma^2} + \mathbf{W}_{\sigma^2}(\lambda_2\mathbf{K}\mathbf{K}^T + \lambda_{\sigma^2}) = (1+\lambda_2)\mathbf{R}\mathbf{K}^T \quad (14)$$

The above equation is also in the form of $\mathbf{A}\mathbf{W} + \mathbf{W}\mathbf{B} = \mathbf{C}$

$$\mathbf{A} = \mathbf{R}\mathbf{R}^T \quad (15)$$

$$\mathbf{B} = \lambda_2\mathbf{K}\mathbf{K}^T + \lambda_{\sigma^2} \quad (16)$$

$$\mathbf{C} = (1+\lambda_2)\mathbf{R}\mathbf{K}^T \quad (17)$$

Given the learned parameters $\mathbf{W}_{\mu_c}$ and $\mathbf{W}_{\sigma_c^2}$, the parameters of data distribution for unseen classes $c = S+1, \ldots, S+U$ are estimated as:

$$\boldsymbol{\mu}_c = \mathbf{W}_{\mu}\mathbf{k}_{\mathbf{c}}, \quad \& \quad \boldsymbol{\sigma}_c^2 = \exp(\rho_c) = \exp(\mathbf{W}_{\sigma^2}\mathbf{k}_{\mathbf{c}}) \quad (18)$$

Where $\mathbf{k}_{\mathbf{c}} = [\mathbf{k}(\mathbf{a}_{\mathbf{c}}, \mathbf{a}_{\mathbf{1}}), ..., \mathbf{k}(\mathbf{a}_{\mathbf{c}}, \mathbf{a}_{\mathbf{S}})]$ denotes an $S \times 1$ vector of kernel-based similarities of the class attribute vectors of the unseen class c with the class attribute vectors of all the seen classes.

2.1. Domain Adaptation in Transductive setting

One of the unique advantages of the proposed generative approach is that unlabeled data from unseen classes can be used to improve the parameters ($\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c$). In zero-shot learning, train and test data come from different domains. Hence it is very likely that parameters learned in the training, will not work well for the test data. This phenomenon is called domain shift. An illustrative view of the domain shift can be seen in figure 2. One way to overcome this issue is to use unlabeled data during test time to further fine tune the parameters. In the transductive setting, we assume that all the test data is given at once at the test time. Because of the availability of this data, we can infer more information about the unseen classes. In this work, we handle the domain shift problem by initializing the parameters $\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c$ using the learned basis vectors, which are then fine-tuned using the unlabeled data using the EM algorithm. From extensive experimentation, we show that this approach gives better performance.

The estimated distribution parameters of unseen classes $(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)_{c=S+1}^{S+U}$ can be further improved by using the unlabeled unseen classes data. In this Transductive setting, we use Expectation-Maximization(EM) based iterative procedure that updates the estimation of the distribution parameters for unseen classes. This procedure is equivalent to the GMM model which uses the unlabeled data $(\mathbf{x}_{\mathbf{n}})_{n=1}^{N_u}$ from unseen classes. This GMM has U mixture components, each corresponding to one unseen class and is initialized by the estimated parameters of unseen classes $(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)_{c=S+1}^{S+U}$ in the inductive setting. The procedure for Transductive setting is described stepwise below:

1. Let the initial estimate of the unseen class parameters be $\boldsymbol{\Theta} = (\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)_{c=S+1}^{S+U}$ where $\boldsymbol{\mu}_c = \mathbf{W}_{\mu}\mathbf{a}_{\mathbf{c}}$, $\boldsymbol{\sigma}_c^2 = \exp(\mathbf{W}_{\sigma^2}\mathbf{a}_{\mathbf{c}})$. Here $\mathbf{W}_{\mu}$ and $\mathbf{W}_{\sigma^2}$ are estimated from seen class data using equations 5, 13.
2. E Step: Infer the probabilities for each example $\mathbf{x}_{\mathbf{n}}$ belonging to each of the unseen classes $c = S+1, ..., S+U$ as $p(y_n = c|\mathbf{x}_{\mathbf{n}}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_{\mathbf{n}}|\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$
3. M Step: Use the inferred class labels to re-estimate $\Theta = (\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)_{c=S+1}^{S+U}$
4. Go to step 2 if not converged.

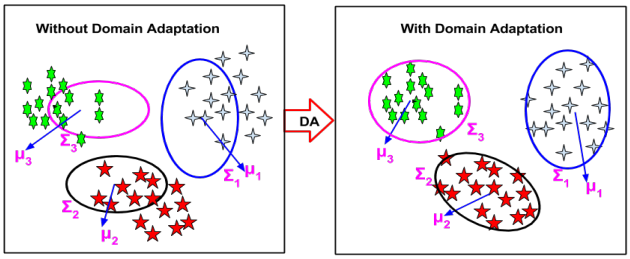

Figure 2. **Domain Adaptation illustrative example**: Each class attribute is projected to the visual space, In the visual space each class are represented by a distribution. Because the seen and unseen class are disjoint, there is a problem of domain shift.

2.2. Extension for Few-shot/one-shot action recognition

In few-shot action recognition, we have a small number of labeled examples for each of the unseen classes. Since our method assumes a Gaussian data distribution, we can extend our zero-shot action recognition method to few-shot/one-shot action recognition. We assume the initial estimate obtained using the previous approach as the prior. Due to conjugacy of the Gaussian, we can update the estimates $(\mu_c, \sigma_c^2)_{c=S+1}^{S+U}$ obtained from zero-shot action recognition method in a straightforward manner when such labeled data for unseen classes is provided. Given a small number of labeled data $(\mathbf{x_n})_{n=1}^{N_c}$ for unseen class c the parameters of this class can be directly updated as:

$$\boldsymbol{\mu}_c^{FS} = \frac{\boldsymbol{\mu} + \sum_{n=1}^{N_c} \mathbf{x_n}}{1 + N_c} \tag{19}$$

$$\boldsymbol{\sigma}_c^{2(FS)} = (\frac{1}{\boldsymbol{\sigma}_c^2} + \frac{N_c}{\boldsymbol{\sigma}^{2*}})^{-1} \tag{20}$$

where $\boldsymbol{\sigma}^{2*} = \frac{1}{N_c}\sum_{n=1}^{N_c}(\mathbf{x_n} - \boldsymbol{\mu}_c)^2$ denotes empirical variance of N_c observations from the unseen class c.

3. Related Work

ZSL can be viewed as an interplay of three subproblems: a visual representation of data instances (feature representation), semantic representation of all classes such as *word2vec* representation [22], and learning a function which establishes the relationship between visual representations and semantic representations of each class[16, 17].

For visual (or feature) representation of class instance, popular hand-crafted features such as HOG [3], HOF [2], ITF [33] were designed. However, the proven utility of deep features for many tasks such as Object Recognition [13, 28, 30], Object Detection [6] etc has made features from well performing CNNs such as [18], Two-Stream CNN[21], 3DCNN [9] ubiquitous for Action Recognition tasks including the zero shot setting. By using 3DCNN features in ZSL, a significant boost in accuracy has been observed [34]. Semantic representation of a class provides additional, complementary information to the visual features of the classes. Typically, two types of semantic representations have been widely used in the ZSL literature: attribute representations [15] and word vector representations [22]. Attribute representations are manually annotated vectors for each class based on the gesture and motion appearance of the objects in the video. Word Vector representations are automatically learned from a large amount of textual data (Wikipedia Corpus). Word2vec models have been used successfully for extracting semantic word vectors from class names [34, 10, 38]. The core step in ZSL is to find a function or projection matrix which can establish a relationship between visual space and semantic space in such a way that visual features of classes map close to their semantic features and vice versa. For example, we would like to have visual features of 'running' map close to semantic features of 'running' and far away from an unrelated action such as 'eating'.

Most methods for zero-shot learning are evaluated on image classification whereas only a few methods have been proposed for zero-shot action recognition in the literature [34, 38, 37, 24]. Most methods model either the inductive or the transductive setting. The most popular approach to ZSL is learning a linear compatibility between the visual and semantic space [1]. [26, 12] provide novel regularizations while learning a linear compatibility function. ESZSL [26] models the relationship between features and attributes as a linear compatibility function while explicitly regularizing the objective. UDA [11] uses a domain adaptation technique by using unlabeled data of unseen classes for better estimation of the parameters.

Recently [32] proposed a simple generative approach for zero-shot learning, without reconstruction regularizer from visual to attribute space. Their paper assume the data distributions are Gaussian. [12] proposed a semantic auto-encoder for zero-shot learning which introduced the reconstructability regularizer. This paper works only in the inductive setting and their approach is not generative. The generative approach of our paper is motivated by [32] with auto-encoder style regularizer proposed by [12]. [38] have proposed a transductive framework for zero-shot action recognition, which uses unlabeled unseen class data for training the model. In their work, they introduced a manifold-regularized regression and a data augmentation strategy to enhance the performance. They have also introduced a multi-task visual-semantic mapping for zero-shot action recognition. In addition, they used prioritized auxiliary data augmentation for domain adaptation and improved the mapping between visual and semantic spaces.

Because of the generative nature of our proposed approach, we can synthesize the data from unseen class based on attribute and train the classifier. This approach helps to reduce the baisness in the case of Generalize Zero-Shot Learning. The efficacy of the proposed approach for the GZSL as well as ZSL can be seen from the experiment on three standard datasets. On all standard datasets our approach has shown the state-of-art result. Here our assumption is data distributions are Gaussian, which can be the hard assumption. Recently [35] proposed an approach which can transform any data distribution to a unimodal Gaussian in the ZSL setup.

4. Experiments

Datasets and Settings: We evaluate our proposed method in three most challenging video action recognition data sets, UCF101 [29], HMDB51 [14] and Olympic [20]. In zero-shot action recognition, these three datasets have been widely used as bench mark. We report mean accuracy along with standard deviation on 30 independent test runs with random train/test class splits.

- **UCF101:** [29] is human action recognition data set with 101 different classes of actions and total of 13320 video clips. In our experiments, we split the classes into 51 seen and 50 unseen class respectively. ‘

- **HMDB51:** [14] is the one of the most challenging human action recognition dataset with 51 different classes of human actions and total number of 6766 video clips. Each class has more than 100 video clips. For the evaluation of our model, we perform a 26/25 split for seen and unseen classes respectively.

- **Olympic:** [20] This dataset has 783 videos from 16 different classes. Splits for seen and unseen classes is 8/8.

Dataset	#videos	#classes	seen/unseen	Attribute dim
UCF101	13320	101	51/50	115
HMDB51	6676	51	26/25	N/A
Olympic	783	16	8/8	40

Table 1. Dataset details and their train test split on all the three dataset used in our experiment.

Visual features: The quality of visual features directly affect the efficiency of the model. We use deep features as they have been shown to be successful in many computer vision tasks. In our experiments, we use the latest convolutional 3D(C3D) visual features provided by [31]. This model was pre-trained on the sports-1M dataset. We extract the outputs of fc6 layer for all segments similar to [31] and then averaged over the segments to form a 4096-dimensional video representation which is used as input visual features in our method.

Semantic representation: Two types of semantic representations are widely used : human labeled attributes [15] and automatically learned distributed semantic representations such as word vectors [22]. Word vector representation is learned automatically by a skip-gram model trained on the google news text corpus provided by Google. Each word is represented by a 300 dimensional vector. We experiment on both attribute and word2vec representations. For HMDB51 dataset, to the best of our knowledge, there is no publicly available attribute representations of the classes. Hence only word2vec is used for HMDB51. However, for UCF101 and Olympic datasets, 115 and 40 dimensional attribute vectors are available respectively [29, 20].

Hyper-parameters: We investigate the optimal hyper-parameters while training via grid search. In our proposed model, there are four hyper-parameters: λ_μ, λ_1 (refer equation 5) and λ_{σ^2}, λ_2 (refer equation 13) for estimating the projection matrix for mean and variance. The optimal values of hyper-parameters are chosen via cross validation on the seen classes. For cross validation, we randomly fix 1/4th of the seen classes as validation classes and conduct five trials on 30 random splits (same as [34]). For generalized zero-shot learning setting, generating a number of examples for unseen classes is also hyper-parameter which we find using cross-validation and observe best model performance for 200 examples.

Comparison with State-of-the-art ZSL Methods: In our first set of experiments, we evaluate our model for zero-shot action recognition with inductive and transductive setting and compare with a number of state-of-the-art methods.

Evaluation Metric: We have evaluated our model using 30 different splits into seen and unseen classes provided by [34] for UCF101(51/50), HMDB51(26/25) datasets. For Olympic dataset we generate 30 random splits for seen and unseen classes(8/8). We use the average accuracy for all 30 splits as the evaluation metric. For fair comparison, we run five such trials for 30 random splits and present the final accuracy with average and standard deviation.
For generalized zero-shot setting we have evaluated for 30 different splits as above and calculated the average accuracy for seen and unseen classes. The final evaluation of our model is on the harmonic mean of the average accuracy of seen and unseen classes, which is similar as [23, 19, 1].

Method	Embed	Olympic	UCF101	HMDB51
HAA [19]	A	46.1 ± 12.4	14.9 ± .8	N/A
DAP [16]	A	45.4 ± 12.8	14.3 ± 1.3	N/A
IAP [17]	A	42.3±12.5	12.8 ± 2	N/A
ST [36]	W	N/A	13.0±2.7	10.9±1.5
SJE [1]	W	28.6±4.9	9.9±1.4	13.3±2.4
SJE [1]	A	47.0±14.8	12.0±1.2	N/A
ESZSL [26]	W	39.6±9.6	15.0±1.3	18.5±2
UDA [11]	A	N/A	13.2±1.9	N/A
Bi-dir [34]	A	N/A	20.5±.5	N/A
Bi-dir [34]	W	N/A	18.9±.4	18.6±.7
Ours	A	**50.41±11.2**	**22.74±1.2**	N/A
Ours	W	34.12±10.1	17.33+1.1	**19.28±2.1**

Table 2. Results on inductive setting for standard zero shot learning setting(disjoint setting) for the action recognition. Here A represents the human annotated attribute vectors and W represents the *word2vec* embedding.

Inductive setting: In this setting, it is assumed that only the labeled data from the seen classes is available during training. Table 2 shows the experimental results in the inductive setting of the zero-shot action recognition problem. We assume that the train and test classes are disjoint. Note that this assumption is made for all the evaluation settings in this work. In this setting, we obtain an improvement of **3%** on the Olympic dataset. On the UCF-101 which is the most used dataset for zero shot action recognition, the proposed model outperforms state-of-the-art on attribute-based semantic representations. For HMDB dataset, the attribute vectors are not available. Hence, we present results only on word2vec embeddings. Our model outperforms the state-of-the-art for this dataset also.

Transductive setting: In the transductive setting, it is assumed that the unlabeled data of the unseen classes is also available at train time. Table 3 shows the performance of our model in the transductive setting. This unlabeled data acts as a source of extra information which helps domain adaptation to the unseen class. Our model outperforms the state-of-the-art in the Olympic and HMDB datasets. The performance on the UCF-101 dataset is slightly lower, where [34] has the best performance. However, we outperform them in the inductive setting.

Method	Embed	Olympic	UCF101	HMDB51
PST [25]	A	48.6±11	15.3 ±2.2	N/A
ST [36]	W	N/A	15.8±2.3	15.0±3
TZWE [37]	A	53.5±11.9	20.2±2.2	N/A
TZWE [37]	W	38.6±10.6	18.0±2.7	19.1 ±3.8
Bi-dir [34]	A	N/A	**28.3±1.0**	N/A
Bi-dir [34]	W	N/A	21.4±.8	18.9±1.1
UDA [11]	A	N/A	13.2±.6	N/A
Ours	A	**57.88±14.1**	24.48±2.9	N/A
Ours	W	41.27±11.4	20.25±1.9	**20.67±3.1**

Table 3. Results on transductive setting for the standard zero shot learning setting(disjoint setting) for the action recognition

Generalized zero-shot setting

In this setting, the test data may come from both seen and unseen classes. In this setting, from the seen classes, we separate 20% of the data for the testing and remaining 80% data is used as training data for calculating W_μ and W_{σ^2} which is used to predict the mean(μ) and variance(σ) for the unseen classes. One way to handle this setting is to assign each test data-point to the class whose estimated distribution gives the highest score. However, we notice that such an approach is biased towards seen classes since the model has not seen any unseen class examples. In our approach, we propose an attractive solution to this issue: we synthesize class instances of unseen classes using the $\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c^2$ which are obtained from the transductive setting approach; these class instances are called pseudo class instances for unseen classes. Here we generate 200 instances for each unseen classes. Since we now have labelled data for seen classes and pseudo labelled data for unseen classes, we train SVM classifier for labelled seen classes data and pseudo labelled data for unseen classes. We then pass the test data (unseen class data plus 20% seen class data) to the trained SVM classifier for classification. Table 4 presents the performance of our model in the generalized setting for zero-shot action classification which clearly shows that it significantly outperforms state-of-art on all the datasets.

Few-shot action recognition: We also experiment our method for the few shot action recognition and present the results. Here only a small number of examples for each of the unseen classes are available during training. Our generative model provides a simple way to update the parameters of the class distribution using equation 19, 20 . It is clear

Method	Embed	Olympic	UCF101	HMDB51
HAA [19]	A	49.4 ± 10.8	18.7 ± 2.4	N/A
SJE [1]	W	32.5±6.7	8.9±2.2	10.5±2.4
ConSE [23]	W	37.6 ± 9.9	12.7 ± 2.2	15.4± 2.8
Ours	A	**52.41±12.2**	**23.74±1.2**	N/A
Ours	W	**42.23±10.2**	**17.45±2.2**	**20.10±2.1**

Table 4. Results on the transductive setting for generalized zero-shot learning setting for the action recognition. Here A represents the human annotated attribute vectors and W represents the *word2vec* embedding.

from the Table 5 that availability of the few data points of the unseen classes significantly improves the performance which is now comparable to that of supervised learning. Note that we do not assume any unlabeled data from the unseen classes in this setting. We test our model with varying number of examples of each unseen classes. The plot of accuracy with respect to the number of samples per class is shown in Figure 3. To the best of our knowledge, this is the first work reporting results in this setting on video datasets.

Previous approaches formulate the few shot learning problem (in the image classification domain) in a purely supervised framework. Whereas, our model takes advantage of transfer learning from semantic space along with supervision.

Dataset	2 points	3 points	4 points	5 points
UCF101	68.78±3.3	73.49±2.2	76.51±2.1	78.68±1.8
HMDB51	42.10±3.6	47.54±3.3	50.34±3.4	52.58±3.1
Olympic	73.20±7.4	75.35±7.3	80.21 ±7.24	83.81±7.11

Table 5. Results on inductive setting for few/one shot learning for the action recognition

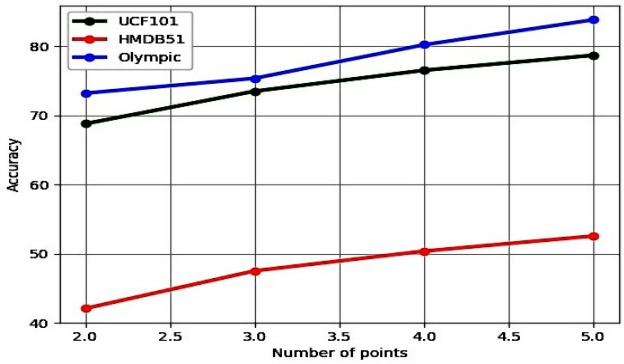

Figure 3. Accuracy vs number of data points for few-shot learning

5. Conclusion

A simple probabilistic framework for the zero-shot action recognition problem is presented in this work. The proposed approach performs well in both the inductive and transductive setting for the standard (disjoint) and generalized zero-shot learning. The unseen data are synthesized utilizing the generative nature of the proposed approach to overcome the unbiased learning of classifier. This handles the problem of GZSL and solves it successfully. In addition, another approach is proposed for domain shift problem using domain adaptation in transductive setting. This approach yields a closed form solution for the parameters to make it fast and easy to implement. Experimental results are shown to achieve state-of-the-art performance. The proposed method also generalizes to few-shot action recognition setting, achieving comparable results to *fully supervised* learning using only few-data.

Acknowledgments: Thanks to Vismay who always gives his support in discussion and experiment. Special thanks to Prof. Hema A Murthy for her valuable feedback. Vinay acknowledges support from Visvesvaraya Ph.D. fellowship.

References

[1] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015. 5, 6, 7, 8

[2] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *computer vision and pattern recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1932–1939. IEEE, 2009. 5

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 5

[4] G. Dinu, A. Lazaridou, and M. Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014. 2

[5] Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2635–2644, 2015. 2

[6] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 5

[7] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007. 2

[8] A. Jameson. Solution of the equation ax+xb=c by inversion of an m*m or n*n matrix. *SIAM Journal on Applied Mathematics*, 16(5):1020–1023, 1968. 4

[9] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013. 5

[10] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised

domain adaptation for zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2452–2460, 2015. 2, 5

[11] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2452–2460, 2015. 5, 7

[12] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345*, 2017. 3, 5

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 5

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011. 6

[15] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009. 5, 6

[16] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 5, 7

[17] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 5, 7

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[19] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3337–3344. IEEE, 2011. 1, 6, 7, 8

[20] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3337–3344. IEEE, 2011. 6

[21] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib. Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *arXiv preprint arXiv:1703.10667*, 2017. 5

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 1, 5, 6

[23] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 1, 6, 8

[24] J. Qin, L. Liu, L. Shao, F. Shen, B. Ni, J. Chen, and Y. Wang. Zero-shot action recognition with error-correcting output codes. In *Proc. CVPR*, 2017. 5

[25] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *Advances in neural information processing systems*, pages 46–54, 2013. 7

[26] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015. 5, 7

[27] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. 2001. 3

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[29] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 5

[31] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 6

[32] V. K. Verma and P. Rai. A simple exponential family framework for zero-shot learning. *arXiv preprint arXiv:1707.08040*, 2017. 2, 5

[33] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013. 5

[34] Q. Wang and K. Chen. Zero-shot visual recognition via bidirectional latent embedding. *arXiv preprint arXiv:1607.02104*, 2016. 5, 6, 7

[35] W. Wang, Y. Pu, V. K. Verma, K. Fan, Y. Zhang, C. Chen, P. Rai, and L. Carin. Zero-shot learning via class-conditioned deep generative models. In *AAAI Conference on Artificial Intelligence (AAAI-18), Louisiana, USA.*, 2018. 6

[36] X. Xu, T. Hospedales, and S. Gong. Semantic embedding space for zero-shot action recognition. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 63–67. IEEE, 2015. 7

[37] X. Xu, T. Hospedales, and S. Gong. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, pages 1–25, 2017. 5, 7

[38] X. Xu, T. M. Hospedales, and S. Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *European Conference on Computer Vision*, pages 343–359. Springer, 2016. 5