Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Visual Data Synthesis via GAN for Zero-Shot Video Classification

Chenrui Zhang and **Yuxin Peng***
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
pengyuxin@pku.edu.cn

Abstract

Zero-Shot Learning (ZSL) in video classification is a promising research direction, which aims to tackle the challenge from explosive growth of video categories. Most existing methods exploit seen-to-unseen correlation via learning a projection between visual and semantic spaces. However, such projection-based paradigms cannot fully utilize the discriminative information implied in data distribution, and commonly suffer from the information degradation issue caused by "heterogeneity gap". In this paper, we propose a visual data synthesis framework via GAN to address these problems. Specifically, both semantic knowledge and visual distribution are leveraged to synthesize video feature of unseen categories, and ZSL can be turned into typical supervised problem with the synthetic features. First, we propose *multi-level semantic inference* to boost video feature synthesis, which captures the discriminative information implied in joint visual-semantic distribution via feature-level and label-level semantic inference. Second, we propose *Matching-aware Mutual Information Correlation* to overcome information degradation issue, which captures seen-to-unseen correlation in matched and mismatched visual-semantic pairs by mutual information, providing the zero-shot synthesis procedure with robust guidance signals. Experimental results on four video datasets demonstrate that our approach can improve the zero-shot video classification performance significantly.

1 Introduction

In the past decade, supervised video classification methods have achieved significant progress due to deep learning techniques and large-scale labeled datasets. However, with the number of video categories growing rapidly, classic supervised frameworks suffer from the following challenges: (1) They rely on large-scale labeled data heavily, while collecting labels for video is laborious and costly, as well as the video instances of some categories are quite rare. (2) Particular models learned on limited categories are hard to expand automatically as video categories grow.

To overcome such restrictions, Zero-Shot Learning (ZSL) has been studied widely [Zhang and Saligrama, 2015; Changpinyo *et al.*, 2016; Xu *et al.*, 2017; Zhang *et al.*, 2017], which aims to construct classifiers dynamically for novel categories. Instead of treating each category independently, ZSL adopts semantic representation (e.g., attribute or word vector) as side information, to associate categories in source domain and target domain for semantic knowledge transfer. Existing zero-shot classification methods focus primarily on *still images*. In this paper, we focus on zero-shot classification in *videos*, which is more needed as video categories emerge on the web everyday. Furthermore, compared to *image*, zero-shot *video* classification has its own characteristics and thus is more challenging. First, video data contains more noise than image, setting a higher requirement on the robustness of zero-shot classification models. Second, video feature describes both spatial and temporal information, whose manifold is more complex. Third, video context with various poses and appearances can be changeable, leading to that video distribution is much more long-tailed than that of image.

Most existing ZSL methods learn a projection among different representation (i.e., visual, semantic or intermediate) spaces based on the data of seen domain, and the learned projection will be applied directly to unseen domain during testing. However, such *projection-based* methods have several shortcomings and thus are not robust enough for video ZSL. On the one hand, they exploit seen-to-unseen correlation only from the semantic knowledge aspect, while ignoring the discriminative information implied in visual data distribution. In fact, intrinsic visual distribution play a vital role in zero-shot video classification, since the most discriminative information is derived from visual feature space [Long *et al.*,]. On the other hand, the inconsistency between visual feature and semantic representation causes *heterogeneity gap*, and the projection between heterogeneous spaces leads to an information loss such that seen-to-unseen correlation would degrade. A specific affect caused by this gap is *hubness problem* [Radovanović *et al.*, 2010], where some irrelevant category prototypes become near neighbors of each other, as the projection is performed in high-dimensional spaces.

In this paper, we introduce Generative Adversarial Net-

*Corresponding author.

works (GANs) [Goodfellow *et al.*, 2014] to realize zero-shot video classification from the perspective of generation, which bypasses the above limitations of explicit projection learning. The key idea is to model the joint distribution over high-level video feature and semantic knowledge via adversarial learning, where discriminative information and seen-to-unseen correlation are embedded effectively for novel video feature synthesis. Once the training procedure is done, unseen video feature can be synthesized by the generator, and zero-shot classification is realized in video feature space in a supervised fashion. Namely, synthetic video features are utilized to train a conventional supervised classifier (e.g., SVM), or serve directly to the simplest nearest neighbor algorithm.

However, synthesizing discriminative video feature based on semantic knowledge is non-trivial for prevalent adversarial learning framework such as conditional GAN (cGAN) [Mirza and Osindero, 2014]. The main challenges we encountered are two-fold: (1) How to model the joint distribution over video feature and semantic knowledge robustly and ensure the discriminative characteristics of the synthetic feature? (2) How to mitigate the impact of heterogeneity gap and transfer semantic knowledge maximally?

To tackle the first challenge, we propose *multi-level semantic inference* approach, aiming to fully utilize the distribution over video feature and semantic knowledge for discriminative information mining. It contains two opposite synthesis procedures driven by adversarial learning, where the semantic-to-visual branch synthesizes video feature given semantic knowledge, and the visual-to-semantic branch inversely infers the semantic knowledge at both feature-level and label-level. Such two-pronged semantic inference forces the generator to capture the discriminative attributes for visual-semantic alignment, and bidirectional synthesis procedures boost each other collaboratively for ensuring robustness of the synthetic video feature.

To tackle the second challenge, we propose *matching-aware mutual information correlation*, which provides the synthesis procedure with informative guidance signals for overcoming the information degradation issue. Instead of direct feature projection, the mutual information hints among matched and mismatched visual-semantic pairs are utilized for semantic knowledge transfer. Therefore, statistical dependence among heterogeneous representations can be captured, bypassing the information degradation issue in typical projection-based ZSL methods.

To verify the effectiveness of the proposal, we conduct extensive experiments on 4 widely-used video datasets, and the experimental results demonstrate that our approach improves the zero-shot video classification performance significantly.

2 Related Work

2.1 Zero-shot Learning

Zero-shot learning has been drawn a wide attention, due to its potentiality on scaling to novel categories in classification tasks. Early explorations in ZSL are mainly focus on learning probabilistic attribute classifiers (PAC), such as Direct-Attribute Prediction (DAP) [Lampert *et al.*, 2009], Indirect Attribute Prediction (IAP) [Lampert *et al.*, 2014] and CONSE [Norouzi *et al.*, 2014]. In these approaches, posterior of seen categories is predicted firstly, then the attribute classifiers are learned by the principle of maximizing posterior estimation. PAC has been proved to be poor in ZSL tasks as it ignores the relation of different attributes [Jayaraman *et al.*, 2014]. After that, research efforts shift to another direction which directly builds a projection from visual feature space to semantic (e.g. attribute or word-vector) space. In this paradigm, linear [Palatucci *et al.*, 2009], bilinear [Zhang and Saligrama, 2016] and nonlinear [Zhang *et al.*, 2017] compatibility models are explored widely by optimizing specific loss functions. Besides visual $\rightarrow$ semantic mapping, other two mapping directions are studied, namely semantic $\rightarrow$ visual mapping and embedding both visual and semantic features into another shared space [Zhang and Saligrama, 2015]. Nowadays, various hybrid models [Akata *et al.*, 2016; Changpinyo *et al.*, 2016] are also studied based on such diverse selection of embedding space.

Recently, Unseen Visual Data Synthesis (UVDS) [Long *et al.*,] views ZSL from a new perspective, i.e., it tries to synthesize visual feature of unseen instances for converting ZSL to a typical supervised problem. However, it is still an explicit projection learning framework in which visual feature is synthesized through embedding-based matrix mapping. On the one hand, such explicit projection paradigm is hard to preserve manifold information of visual feature space, or capture unseen-to-seen correlation for zero-shot generalization. On the other hand, the visual feature synthesized by UVDS suffers variance decay issue [Long *et al.*,], which is not robust enough for high-dimensional video feature synthesis. In this paper, we address these problems with deep generative models.

2.2 Generative Adversarial Networks

As one of the most potential generative models, Generative Adversarial Network (GAN) [Goodfellow *et al.*, 2014] is0 studied extensively recently. The goal of GAN is to learn a generator distribution $P_g(x)$ that matches the real data distribution $P_{real}(x)$ through a minimax game:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim P_{real}}[\log D(x)] + \mathbb{E}_{z\sim P_{noise}}[\log(1 - D(G(z)))]$$

Basic GAN is unrestricted and uncontrollable during training, as well as the priori noise z is difficult to explain. To tackle these defects, a surge of variants of GAN have emerged, such as conditional GAN (cGAN) [Mirza and Osindero, 2014], InfoGAN [Chen *et al.*, 2016] and Wasserstein GAN (W-GAN) [Arjovsky *et al.*, 2017]. cGAN tries to solve the controllability issue via providing class labels to both generator and discriminator as condition. InfoGAN aims to learn more interpretable representation by maximizing the mutual information between latent code and generator's output. W-GAN introduces earth mover distance to measure similarity of real and fake distribution, which makes training stage more stable. Our work is related to cGAN and InfoGAN closely, and details are discussed in the sequel.

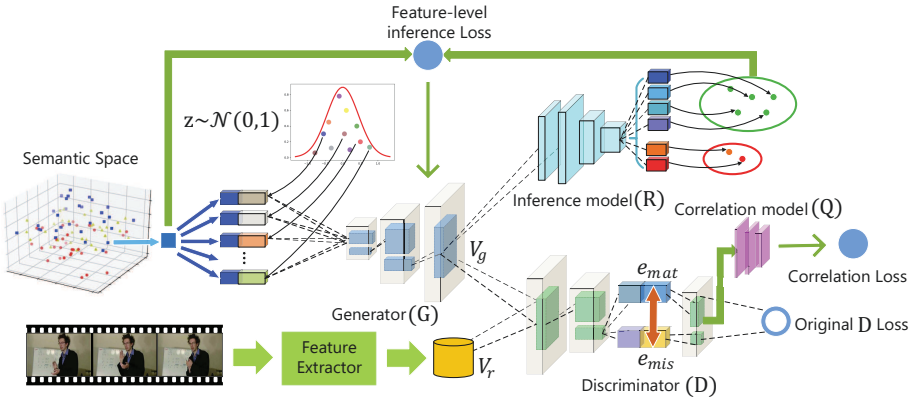

Figure 1: Architecture of the proposed framework. A group of noise is utilized by generator to synthesize video feature V_g, which is used by inference model R and discriminator D simultaneously to perform semantic inference and correlation constraint. V_r denotes the real visual feature, e_{mat} and e_{mis} respectively denote the matched and mismatched semantic embedding sets corresponding to V_r.

3 Methodology

3.1 Problem Formalization

Assume $D_{tr} = \{(V_n, y_n)\}_{n=1}^{N_s}$ denotes the training set of N_s samples and $D_{te} = \{(V_n, y_n)\}_{n=1}^{N_u}$ denotes test set of N_u samples, their corresponding label spaces are $\mathcal{S} = \{1, 2, \cdots, S\}$ and $\mathcal{U} = \{S+1, S+2, \cdots, S+U\}$ with $\mathcal{S} \cap \mathcal{U} = \varnothing$ in ZSL context. V_n and y_n respectively denote visual feature and label of the n^{st} video sample. Given a new test video feature V, the goal of ZSL is to predict its class label $y \in \mathcal{U}$. We use $g(\cdot) : Y \to E$ to denote the word embedding function that maps label to semantic embedding and $g(y_i)$ is the word-vector w.r.t. label y_i. Y and E denote label space and semantic space respectively.

3.2 Multi-level Semantic Inference

As illustrated in Figure 1, the generator G synthesizes video feature conditioned on semantic knowledge (i.e., word vector in this paper). The synthetic video feature is expected to be convincing enough to recover the distribution of real video data, as well as capture the semantic correlation and discriminative information for zero-shot classification task. Essentially, the mission of G is learning a *visual-semantic joint distribution*, rather than fitting a single distribution of only one domain. However, high-dimensional and noisy nature of video feature lead to great uncertainty for visual-semantic matching, and thus conventional GAN frameworks are hard to achieve stable synthesis. Moreover, stable visual-semantic matching cannot ensure the discriminative power of the synthetic feature, which is vital for zero-shot classification task.

To tackle the above challenges, we propose a multi-level semantic inference approach, which boosts the video feature synthesis via inversely inferring semantic information at both feature-level and label-level. Video feature synthesis and semantic inference are driven by adversarial learning, where semantic inference forces the generator to capture the seen-to-unseen correlation and discriminative information, boosting the synthesis process for robust visual-semantic alignment.

Concretely, we develop an auxiliary model R for semantic feature inference, which learns an inverse mapping from the synthetic video feature to corresponding semantic knowledge. Formally, R tries to fit conditional distribution $q(e_c|\hat{v})$, where $\hat{v}$ is the synthesized video feature and e_c is the semantic embedding. In fact, G and R model two joint distributions:

$$\begin{cases} p_\theta(v, z, e_c) = p_\theta(z, e_c)\, p_\theta(v|z, e_c) \\ q_\phi(v, z, e_c) = q_\phi(z, v)\, q_\phi(e_c|z, v) \end{cases} \tag{1}$$

where θ and ϕ are parameters of G and R, respectively. As semantic and video features are high-level representation rather than real-valued data (e.g., image pixels), typical reconstruction metrics such as ℓ_p-norm is hard to measure the semantic similarity. Hence, we adopt adversarial learning for feature-level semantic inference:

$$\begin{aligned} \mathcal{F}_{fea} &= \min_{\theta,\phi} \max_{\psi} V(\theta, \phi, \psi) \\ &= \mathbb{E}_{v\sim q_\phi(v), \hat{c}\sim q_\phi(c|\hat{v})}\big[\log D_\psi(v, \hat{c})\big] \\ &+ \mathbb{E}_{\hat{v}\sim p_\theta(v|z,c),(z,c)\sim p_\theta(z,c)}\big[\log\big(1 - D_\psi(\hat{v}, c)\big)\big] \end{aligned} \tag{2}$$

where $\mathcal{F}_{fea}$ is the adversarial object of feature inference, ψ is the parameter of the discriminator D, the meaning of θ and ϕ is same with Eq. (1).

For discriminative information mining, we enhance minimax game of Eq. (2) with category log-likelihood estimation, which can be viewed as *label-level* semantic inference:

$$\mathcal{F}_{cat} = \sum_{i=1}^{N_s} \mathbb{E}\big[\log P\big(C = C_i|v\big)\big] + \mathbb{E}\big[\log P\big(C = C_i|\hat{v}\big)\big] \tag{3}$$

where $\mathcal{F}_{cat}$ is the objective of category-aware training, C_i denotes the i-th category, N_s is the total number of seen categories and P means the category-specific distribution. The insights behind Eq. (3) is that label judgment can provide valuable information for decision boundary selection of category-level distribution, guiding G to synthesis visual features with more discriminative information.

In addition, to further stabilize such bidirectional adversarial synthesis, we design a simple yet effective outlier detection approach. As shown in Figure 1, rather than sampling only one noise z for each condition c, a group of noises are sampled randomly to generate several video features $\mathcal{V}_g$. R infers semantic features $\mathcal{R}$ with $\mathcal{V}_g$, then the similarity among elements in $\mathcal{R}$ is measured via cosine distance, outliers (denoted by $\mathcal{O}$) that exceed an adaptive threshold η will be discarded. We call $\mathcal{P} = \mathcal{R} - \mathcal{O}$ is the set of *reasonable* inferred

word vectors, and for arbitrary $r_i \in \mathcal{R}$, r_i is reasonable if and only if:

$$Cos(e_c, r_i) \geq \eta = \max_{1 \leq j \leq n} \big(Cos(e_c, r_j)\big) - \mu\Delta \quad (4)$$

where μ is trade-off parameter and Δ is the variation range of cosine similarity among elements in $\mathcal{R}$ during training:

$$\Delta = \max_{1 \leq j \leq n} Cos(e_c, r_j) - \min_{1 \leq j \leq n} Cos(e_c, r_j) \quad (5)$$

Thus, η adjusts adaptively along with the training progress to boost the robustness of semantic inference. Finally, we define the inference loss as:

$$\mathcal{L}_{re}(e_c, \mathcal{P}) = -\sum_{r_i \in \mathcal{P}} Cos(e_c, r_i) \quad (6)$$

3.3 Mutual Information Correlation Constraint

The power of GAN is attributed to that it learns a latent loss by D to classify real or fake instances in a data-driven manner, rather than relying on structured loss which is hand-engineered. This powerful latent loss guides G to fit sophisticated distribution of real data progressively. However, in ZSL video classification scenarios, D should learn to evaluate whether the output of G is aligned with conditional information, instead of only scoring its realism. Original loss of cGAN is not informative enough for G to synthesize discriminative visual feature, since D has no explicit notion of whether visual feature matches the semantic embedding. To address this problem, we propose Matching-aware Mutual Information Correlation Constraint (MMICC) to maximize semantic knowledge transfer through leveraging the Mutual Information (MI) of matched and mismatched visual-semantic pairs. Formally, MI can be defined based on entropy as:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (7)$$

where X and Y are random variables, $I(X;Y)$ denotes mutual information of X and Y, $H(X|Y)$ is conditional entropy:

$$H(X|Y) = -\sum_{X \in \mathcal{X}, Y \in \mathcal{Y}} P(X,Y) \log P(X|Y) \quad (8)$$

Eq. (7) intuitively reveals the meaning of MI, which quantitatively measures the amount of information given by one random variable about the other.

We realize MMICC via a mutual information regularization defined as follows:

$$\mathcal{L}_{co} = -I\big(e_{mat}; G(z, e_{mat})\big) + I\big(e_{mis}; G(z, e_{mat})\big) \quad (9)$$

where e_{mat} and e_{mis} respectively denote the matched and mismatched semantic embedding corresponding to specific visual feature. $G(z, e_{mat})/G(z, e_{mis})$ is the synthetic video feature conditioned on e_{mat}/e_{mis}. Minimizing Eq. (9) represents that MI among matched pairs is expect to be high while among mismatched pairs is expect to be low, and this smooth matching-aware regularization provides G with guidance information during training.

With conditional entropy definition in Eq. (8), Eq. (9) can be further expressed as follows (we denote e_{mat}/e_{mis} as e):

$$I\big(e; G(z,e)\big) = H(e) + P(e, G(z,e)) \log P(e|G(z,e)) \quad (11)$$

In practice, the posterior $P\big(e|G(z,e)\big)$ is hard to solve, thus $I\big(e; G(z,e)\big)$ cannot be maximized directly. Fortunately the prior work [Chen *et al.*, 2016] has solved this problem via *Variational Information Maximization* and subsequent approximation. Specifically, they introduce an auxiliary distribution $Q(c|x)$ to approximate original $P(c|x)$ and defined a variational lower bound of MI based on $Q(c|x)$. According to [Chen *et al.*, 2016], we use lower bound of MI, $LB(G, Q)$, to approximate $I\big(e; G(z,e)\big)$:

$$\begin{aligned} LB(G,Q) &= \mathbb{E}_{e \sim P(e), x \sim G(z,e)}[\log Q(e|x)] + H(e) \\ &\leq I\big(e; G(z,e)\big) \end{aligned} \quad (12)$$

where Q is the auxiliary distribution which approximates original $P\big(e|G(z,e)\big)$. $H(e)$ is irrelevant to parameters of G, thus we regard it as a constant during optimization. In our case, auxiliary distribution Q is represented by a neural network which shares all convolutional layers with D and yields conditional distribution $Q(e|x)$ by additional `fc` layers.

Algorithm 1 Training process of the proposed framework

Input: minibatch video feature v, matched semantic embedding e, the set of mismatched embedding $\hat{E}$, the number of noise samples q, training steps M, step size m.

for $i = 1$ **to** M **do**
  $Score_r \leftarrow D(v, e)$ {$\langle$real v, matched $e\rangle$}
  $Score_w \leftarrow 0, \mathcal{R} \leftarrow \varnothing$ {Initialization}
  for each $\hat{e}_i \in \hat{E}$ **do**
    $Score_w = Score_w + D(v, \hat{e}_i)$ {$\langle$real v,mismatched $e\rangle$}
  end for
  $Score_w = Score_w / \#\hat{E}$ {Average the $Score_w$}
  for $i = 1$ **to** q **do**
    $z_i \sim \mathcal{N}(0,1)^Z$ {Draw noise samples randomly}
    $\hat{v_i} \leftarrow G(z_i, e)$ {Forward through generator}
    $Score_f = Score_f + D(\hat{v_i}, e)$ {$\langle$fake v, matched $e\rangle$}
    $r_i \leftarrow R(\hat{v_i}), \mathcal{F}_{cat} \leftarrow D(v_i, \hat{v_i})$ {Semantic inference}
    $\mathcal{R} = \mathcal{R} \cup \{r_i\}$
  end for
  $Score_f = Score_f / q$ {Average the $Score_f$}
  $\mathcal{L}_{in} \leftarrow$ (6) {Get semantic inference loss via Eq. (6)}
  $\mathcal{L}_{co} \leftarrow$ (9) {Get correlation loss via Eq. (9)}
  $\mathcal{L}_{\mathcal{D}} \leftarrow log(Score_r) + [log(1 - Score_w) + log(1 - Score_f)]/2 + \lambda_2 \mathcal{L}_{co}$
  $D \leftarrow D - m \dfrac{\partial \mathcal{L}_{\mathcal{D}}}{\partial D}$ {Update discriminator}
  $\mathcal{L}_{\mathcal{G}} \leftarrow log(Score_f) + \lambda_1 \mathcal{L}_{in}$
  $G \leftarrow G - m \dfrac{\partial \mathcal{L}_{\mathcal{G}}}{\partial G}$ {Update generator}
end for

3.4 Objective and Optimization

Comprehensively, objective of the proposed framework is summarized as Eq. (10), where v is the video feature, λ_1 and λ_2 are hyper-parameters which balance semantic inference loss $\mathcal{L}_{in}$ and correlation loss $\mathcal{L}_{co}$. We summarize the training procedure of our framework in Algorithm 1. Similar to [Reed *et al.*, 2016], input of D is three-fold: 1) $\langle v_{real}, e_{mat}\rangle$ pair, 2) $\langle v_{real}, e_{mis}\rangle$ pair and 3) $\langle v_{fake}, e_{mat}\rangle$ pair. We use $Score_r$, $Score_w$ and $Score_f$ to denote the confidence scores given by D w.r.t. three kind of pairs, respectively.

$$\min_{G,Q}\max_{D} V(G,D,Q) = \mathbb{E}_{v,e\sim P(v,e)}\big[\log D(v|e_{mat}) + \log(1-D(v|e_{mis}))\big] + \mathbb{E}_{z\sim P_{noise}}\big[\log\big(1-D(G(z|e_{mat}))\big)\big] + \lambda_1\mathcal{L}_{in}(e_{mat}, G(v|e_{mat})) - \lambda_2\mathcal{L}_{co}(G,Q) \tag{10}$$

3.5 Zero-shot Classification

At the test stage, we simply use the nearest neighbor (NN) search and SVM as classifiers for evaluating the discriminative capability of synthetic video feature. For NN search, label in test set is predicted by:

$$\hat{y} = arg\min_{y\in y_{te}, v^*\in V_{te}} ||G\big(z, g(y)\big) - v^*|| \tag{13}$$

where the $G\big(z, g(y)\big)$ is the synthetic video feature of label y, v^* is the original video feature in test set V_{te} and $\hat{y}$ is the predicted label.

For experiments with SVM, we synthesize video feature of same amount with unseen categories in test set. Then the synthetic feature of unseen categories and original visual feature of seen categories are merged to train a SVM with 3^{rd}-degree polynomial kernel, whose slack parameter is set to 5.

4 Experiments

4.1 Datasets and Settings

Datasets: Experiments are conducted on four popular video classification datasets, including HMDB51 [Kuehne *et al.*, 2013], UCF101 [Soomro *et al.*, 2012], Olympic Sports [Niebles *et al.*, 2010] and Columbia Consumer Video (CCV) [Jiang *et al.*, 2011], which respectively contain 6.7k, 13k, 783 and 9.3k videos with 51, 101, 16 and 20 categories.

Zero-Shot Settings: There are two ZSL settings, namely strict setting and generalized setting [Xian *et al.*, 2017]. The former assumes the absence of seen classes at test state while the latter takes both seen and unseen data for testing. In this paper, we adopt strict setting in all experiments.

Data Split: There are quite few zero-shot learning evaluation protocols for video classification in the community. Xu et al. [Xu *et al.*, 2017] establish a baseline of this filed. In order to compare our framework with the state-of-the-art, we follow the data splits proposed by [Xu *et al.*, 2017]: 50/50 for every dataset, i.e., video feature of 50% categories are used for model training and the other 50% categories are held unseen until test time. We take the average accuracy and standard deviation as evaluation metrics and report the results over 50 independent splits generated randomly.

4.2 Implementation Details

Our model is implemented with PyTorch[1]. Both traditional feature and deep feature are investigated in experiments. For the former, similar to [Xu *et al.*, 2017], we extract improved trajectory feature (ITF) with three descriptors (HOG, HOF and MBH) for each video, then encode them by Fisher Vectors (FV) and we get combined video feature with 50688 dimension. For the latter, we use the two-stream [Simonyan and Zisserman, 2014] framework based on VGG-19 to extract spatial-temporal feature of videos, and frame and optical flow feature from last pooling layer are concatenated to form final visual embedding. We adopt GloVe [Pennington *et al.*, 2014] trained on Wikipedia with more than 2.2 million unique vocabularies to obtain semantic embedding and its dimension is 300. The dimension of Gaussian noise is 100 and the cardinality of noise set is set to 30. We train our framework for 300 epochs using Adam optimizer with momentum 0.9. We initialize the learning rate to 0.01 and decay it every 50 epochs by a factor of 0.5. Both λ_1 and λ_2 are set to 1.

[1] http://pytorch.org/

4.3 Compared Methods

We compare our framework with several ZSL methods in video: (1) Convex Combination of Semantic Embeddings (CONSE) [Norouzi *et al.*, 2014]. CONSE is a posterior based model which trains classifiers on seen categories, then the prediction models are built on linear combination of existing posterior. (2) Structured Joint Embedding (SJE) [Akata *et al.*, 2015]. SJE optimizes the structural SVM loss to learn the bilinear compatibility. This model utilizes bilinear ranking to maximize the score among relevant labels and minimize the score among irrelevant labels. (3) Manifold regularized ridge regression (MR) [Xu *et al.*, 2017]. MR enhances the conventional projection pipeline by manifold regularization, self-training and data augmentation in transductive manner.

4.4 Experimental Results

The experimental results are shown in Table 1. Note that FV denotes the Fisher Vectors encoded dense trajectory feature and DF denotes the deep video feature extracted by two-stream neural networks. From the results we draw several conclusions: (1) All methods are far beyond the random guess bound, demonstrating the success of ZSL in video classification. (2) Our approach beats the most existing methods in terms of average accuracy/mAP, while the standard deviation is slightly higher. This fact illustrates that generative framework is not as stable as projection-based methods, such as SJE and MR. (3) Compared to SVM, NN search suffers from more hubness problem and thus achieves suboptimal results. However, NN that based on our approach performs better than CONSE and SJE without extra payment, indicating the discriminative power of our synthetic video feature. (4) Our approach can improve the zero-shot video classification performance significantly both on traditional visual feature and deep video feature, indicating that it is robust enough to capture both discriminative information and seen-to-unseen correlation for zero-shot video classification purpose.

4.5 Ablation Studies

We conduct ablation studies to further evaluate the effect of two components of our proposal and the results are exhibited in Table 2. For the sake of simplicity, we only report the results on UCF101 and CCV, since similar results are yield on the other two datasets. From Table 2, we can draw the

Methods	Feature	HMDB51	UCF101	Olympic Sports	CCV
Random Guess	–	4.0	2.0	12.5	10.0
CONSE [Norouzi *et al.*, 2014]	FV	15.0 ± 2.7	11.6 ± 2.1	36.6 ± 9.0	20.7 ± 3.1
SJE [Akata *et al.*, 2015]	FV	12.0 ± 2.6	9.3 ± 1.7	34.6 ± 7.6	16.3 ± 3.1
MR [Xu *et al.*, 2017]	FV	24.1 ± 3.8	22.1 ± 2.5	43.2 ± 8.3	33.0 ± 4.8
Ours (NN)	FV	22.8 ± 4.0	23.7 ± 4.5	39.5 ± 9.2	28.3 ± 5.7
Ours (SVM)	FV	**25.3 ± 4.5**	**25.4 ± 3.1**	**43.9 ± 7.9**	**33.1 ± 5.8**
CONSE [Norouzi *et al.*, 2014]	DF	12.9 ± 3.1	8.3 ± 3.0	21.1 ± 8.4	17.2 ± 5.3
SJE [Akata *et al.*, 2015]	DF	10.7 ± 3.5	9.9 ± 2.6	25.8 ± 8.3	14.5 ± 4.0
MR [Xu *et al.*, 2017]	DF	19.6 ± 4.9	24.1 ± 4.2	33.5 ± 9.2	22.6 ± 6.4
Ours (NN)	DF	18.5 ± 5.3	27.3 ± 5.9	30.1 ± 9.5	22.3 ± 7.6
Ours (SVM)	DF	**21.6 ± 5.5**	**30.1 ± 5.7**	**35.5 ± 8.9**	**26.1 ± 8.3**

Table 1: Experimental results of our approach and comparison to state-of-the-art for ZSL video classification on four datasets. Average % accuracy ± standard deviation for HMDB51 and UCF101, mean average precision ± standard deviation for Olympic Sports and CCV.

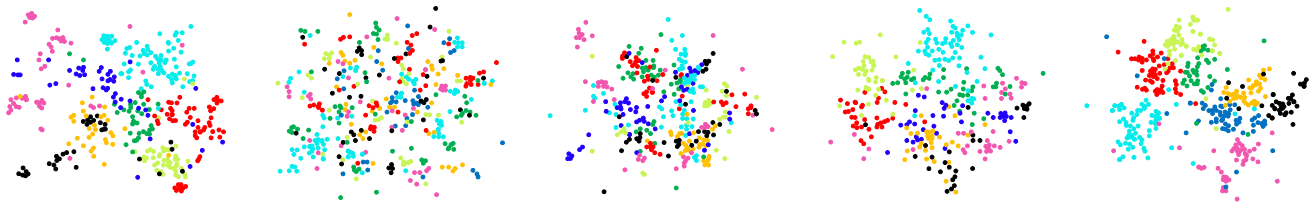

Figure 2: Visualization of the traditional visual feature distribution of 8 random unseen classes on Olympic Sports dataset. Different classes are shown in different colors.

Dataset	Method	NN		SVM	
		FV	DF	FV	DF
UCF101	Real feature	5.1	5.4	8.1	8.0
	cGAN-baseline	5.5	5.9	9.0	9.4
	cGAN+Inf	13.8	15.2	14.9	18.7
	cGAN+Co	20.6	23.0	21.2	25.1
	cGAN+Inf+Co	**23.7**	**27.3**	**25.4**	**30.1**
CCV	Real feature	11.9	11.6	14.1	13.8
	cGAN-baseline	12.5	13.9	14.5	15.1
	cGAN+Inf	18.2	16.0	21.9	19.5
	cGAN+Co	23.8	19.7	29.0	24.4
	cGAN+Inf+Co	**27.2**	**22.3**	**33.1**	**26.1**

Table 2: Baseline experiments on performance of semantic inference and MMICC, which are denoted by Inf and Co respectively.

following observations: (1) With the aid of semantic condition, synthetic video feature exceeds the real feature, suggesting semantic knowledge transfer is meaningful for ZSL problems. (2) Both semantic inference and MMICC can improve the performance of vanilla cGAN by a large margin, suggesting both intrinsic structure information of video feature and semantic correlation are vital for ZSL. (3) MMICC plays a major role in improving classification performance, revealing mutual information correlation is effective and robust for countering heterogeneity between visual and semantic representations. (4) Our proposal incorporates semantic inference and MMICC into an unified framework and achieves a significant improvement when compared to the single component, demonstrating they complement each other for learning a robust visual-semantic alignment for zero-shot generalization.

Moreover, we investigate the performance of our model on alleviating hubness problem by qualitative illustration. We randomly sample 8 unseen categories from Olympic Sports and visualize both original and synthetic FV feature yield by different methods with t-SNE [Maaten and Hinton, 2008]. As shown in Figure 2, (a) illustrates the distribution characteristics of original FV feature. In (b), vanilla cGAN conditioned on word-vectors is hard to synthesize FV feature with high discriminative power, where video feature of different classes mix together, resulting in a severe hubness problem. In (c), when cGAN is equipped with semantic inference, the synthetic feature starts to show clustering properties, but instances of different classes still overlap with each other. In (d), MMICC can improve discriminative power of the synthetic feature to a large extent, which proves the effectiveness of mutual information constraint. Finally, in (e), both semantic inference and MMICC are adopted to achieve a best performance for mitigating the hubness problem. Compared to original feature, synthetic feature yield by our framework suffers less hubness problem intuitively.

5 Conclusion

In this paper, we have adopted deep generative model GAN to overcome the limitations of explicit projection function learning in zero-shot video classification. The distributions of video feature and semantic knowledge are fully utilized to facilitate visual feature synthesis. We have proposed semantic inference and mutual information correlation to endow conventional GAN architecture with zero-shot generalization ability. The synthetic feature owns high discriminative power and suffer much less information degradation issue than pre-

vious methods. State-of-the-art zero-shot video classification performance is achieved on four video datasets.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 61771025 and Grant 61532005.

References

[Akata *et al.*, 2015] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015.

[Akata *et al.*, 2016] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. Multi-cue zero-shot learning with strong supervision. In *CVPR*, pages 59–68, 2016.

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[Changpinyo *et al.*, 2016] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016.

[Chen *et al.*, 2016] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, pages 2172–2180, 2016.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[Jayaraman *et al.*, 2014] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, pages 1629–1636, 2014.

[Jiang *et al.*, 2011] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ICMR*, page 29, 2011.

[Kuehne *et al.*, 2013] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In *HPCSE*, pages 571–582. 2013.

[Lampert *et al.*, 2009] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

[Lampert *et al.*, 2014] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *PAMI*, 36(3):453–465, 2014.

[Long *et al.*,] 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*, pages 1627–1636.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008.

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

[Niebles *et al.*, 2010] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, pages 392–405, 2010.

[Norouzi *et al.*, 2014] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

[Palatucci *et al.*, 2009] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, pages 1410–1418, 2009.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[Radovanović *et al.*, 2010] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *JMLR*, 11(Sep):2487–2531, 2010.

[Reed *et al.*, 2016] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[Soomro *et al.*, 2012] 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.

[Xian *et al.*, 2017] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, pages 1980–1990, 2017.

[Xu *et al.*, 2017] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *IJCV*, 123(3):309–333, 2017.

[Zhang and Saligrama, 2015] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, pages 4166–4174, 2015.

[Zhang and Saligrama, 2016] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint semantic similarity embedding. In *CVPR*, pages 4166–4174, 2016.

[Zhang *et al.*, 2017] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. *CVPR*, pages 2021–2030, 2017.