Zero-Shot Action Recognition with Error-Correcting Output Codes

Jie Qin¹,², Li Liu³,⁴, Ling Shao⁵, Fumin Shen⁶, Bingbing Ni⁶, Jiaxin Chen¹,² and Yunhong Wang¹,²†
¹Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University
²State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
³Malong Technologies Co., Ltd., ⁴University of East Anglia
⁵University of Electronic Science and Technology of China, ⁶Shanghai Jiao Tong University
qinjiebuaa@gmail.com, li.liu@malongtech.cn, ling.shao@ieee.org
fumin.shen@gmail.com, nibingbing@sjtu.edu.cn, chenjiaxinX@gmail.com, yhwang@buaa.edu.cn

Abstract

Recently, zero-shot action recognition (ZSAR) has emerged with the explosive growth of action categories. In this paper, we explore ZSAR from a novel perspective by adopting the Error-Correcting Output Codes (dubbed ZSECOC). Our ZSECOC equips the conventional ECOC with the additional capability of ZSAR, by addressing the domain shift problem. In particular, we learn discriminative ZSECOC for seen categories from both category-level semantics and intrinsic data structures. This procedure deals with domain shift implicitly by transferring the well-established correlations among seen categories to unseen ones. Moreover, a simple semantic transfer strategy is developed for explicitly transforming the learned embeddings of seen categories to better fit the underlying structure of unseen categories. Consequently, our ZSECOC inherits the promising characteristics from ECOC as well as overcomes domain shift, making it more discriminative for ZSAR. We systematically evaluate ZSECOC on three realistic action benchmarks, i.e. Olympic Sports, HMDB51 and UCF101. The experimental results clearly show the superiority of ZSECOC over the state-of-the-art methods.

1. Introduction

During the past decade, human action recognition [1, 27, 55, 52, 54, 6, 44, 7] has been extensively explored. Robust action recognition usually relies on numerous labeled training examples. However, in many realistic scenarios, annotating sufficient examples for ever-growing new categories is exhausting and inapplicable, which inspires us to develop a system that can automatically recognize actions from novel/unseen categories.

∗ indicates equal contributions.
† indicates corresponding author.

Zero-shot learning (ZSL) [12, 61, 26, 46, 17, 5, 64] has emerged as an effective paradigm for recognizing unseen categories without any labeled examples. Usually, ZSL can be fulfilled with the help of label embeddings (or so-called intermediate representations), among which semantic attributes have been widely utilized. Nevertheless, attributes are often manually-specified and highly subjective, since they are either heuristically defined [12] or provided by domain specialists [25]. Particularly, for zero-shot action recognition (ZSAR), attribute-based methods suffer from several specific drawbacks. First, actions are usually defined by ‘verbs’, which are lack of well-defined class hierarchy relationships. Second, dynamic actions are more complex than objects, making it very difficult to specify a suitable attribute pool for different actions. The above difficulties significantly limit the capabilities of previous attribute-based ZSL approaches.

To this end, word embeddings have been preferred in recent works [23, 16, 5, 60] for addressing ZSAR. By using word vectors derived from a huge text corpus (e.g.
In this paper, we aim to enhance the conventional ECOC with the additional ability of zero-shot recognition (dubbed zero-shot ECOC, ZSECOC). Specifically, we derive the discriminative ZSECOC from category-level semantic correlations which are captured from a large-scale text corpus, i.e., Google News ($\approx 100$ billion words). The semantic correlations among categories work as tunnels to implicitly transfer crucial knowledge from seen to unseen categories, e.g., the unknown ‘triple jump’ may learn from ‘high jump’ and ‘long jump’. This kind of knowledge transfer can thus address the domain shift problem to some extent. In addition to preserving semantics, the intrinsic local structure of visual data is also considered when designing our discriminative ZSECOC. Furthermore, in contrast to transductive methods [23, 59, 60] that require the access to visual data from unseen categories, a simple semantic transfer strategy without using any unseen data is developed to generate effective ZSECOC for unseen categories. This strategy explicitly transforms the learned embeddings of seen categories to better fit the underlying semantic structure of unseen categories. In this way, we can further eliminate the influence of domain shift. As shown in Fig. 1, ZSECOC is more discriminative than attributes. Fig. 2 illustrates the whole learning process of ZSECOC for seen/unseen action categories. Our main contributions are summarized as follows:

1) We address ZSAR by designing discriminative ZSECOC. We equip the conventional ECOC with the capability of ZSAR by discovering the semantic correlations among seen categories, which are quantitatively measured using word vectors of well-defined class hierarchy relationships. The well-established semantic knowledge is further transferred to semantically related unseen categories. As a consequence, the proposed ZSECOC inherits the intrinsic advantages of ECOC as well as overcomes domain shift. 2) In addition to preserving category-level semantics, our ZSECOC also incorporates instance-level visual data structures. A joint optimization framework is proposed to solve the resultant challenging problem. The high-quality ZSECOC is directly learned via efficient discrete optimization without any relaxations. 3) The proposed ZSECOC is systematically evaluated on three realistic video action datasets, i.e., Olympic Sports [39], HMDB51 [24] and UCF101 [53]. The state-of-the-art performance in terms of ZSAR clearly demonstrates the superiority of our approach.

2. Related Work

1) Zero-Shot Learning. ZSL aims to recognize unseen categories without any labeled examples. As a common practice, different label embeddings have been employed, e.g., semantic attributes [12, 25, 62, 14, 26] and word vectors [2, 58, 23, 15]. A mapping from visual features to semantic embeddings is learned from seen categories and applied to unseen categories for final recognition. A majority of existing ZSL works focus on object/scene recognition [12, 14, 25, 26, 61, 62, 20, 34, 35] and there are much fewer works on ZSAR [16, 59, 23, 60] due to the challenges previously mentioned. In ZSAR, word vectors have been preferred since only category names are required for constructing the label embeddings.

In addition, previous works also attempted to address the domain shift problem [23] existing in ZSL, since it significantly deteriorates the recognition accuracy. Several domain adaptation methods [13, 14, 23, 60] have been proposed for ZSL, based on transductive learning [23] or data augmentation [60]. However, most of their models were trained using some unseen instances, which violate the fundamental assumption of the standard ZSL setting that no unseen examples could be accessed during training. In this
paper, we develop a simple semantic transfer scheme without any transductive dependency on unseen examples.

2) Error-Correcting Output Codes. ECOC [8, 3, 42, 63, 10] has been explored as an efficient and effective alternative to multi-class classification. Specifically, each class has some pre-specified codeword, e.g. ‘1010100’. ECOC methods train a set of binary classifiers in terms of each bit of the codes. The final multi-class classification is fulfilled through matching the class-level codewords with the binary code of a test point predicted by binary classifiers.

Lots of efforts have been made for simultaneously optimizing the code matrix and binary classifiers, e.g. random ECOC [3] and discriminant ECOC [42]. However, there are rarely any practices that explore ECOC for ZSL. The most related work to ours is [40], where semantic output codes (SOC) were designed for ZSL. SOC did share some similar spirits with ECOC. Nevertheless, it was directly obtained from semantic knowledge bases, thus lacked the intrinsic characteristics of ECOC (e.g. good diversity). Therefore, to the best of our knowledge, this is the first work that enhances ECOC for the purpose of ZSAR. Next, we will introduce the design of our discriminative ZSECOC in detail.

3. Discriminative ZSECOC

In ZSAR, we aim to recognize any instance $x^u$ from $C^u$ unseen action categories, given all $N$ instances $X = \{x_n\}_{n=1}^N \in \mathbb{R}^{d \times N}$ from $C$ seen categories, where $d$ is the original feature dimension. In this work, we would like to seek $m$-bit category-level ZSECOC as the label embedding, by incorporating word vectors as the side information. We denote the ZSECOC of seen and unseen categories as $B = \{b_i\}_{i=1}^B \in \{-1, 1\}^{m \times C}$ and $B^u = \{b^u_j\}_{j=1}^{m \times C^u}$, respectively. The semantic labels of seen and unseen categories are denoted as $\{y^s_i\}_{i=1}^C \in \mathcal{Y}^s$ and $\{y^u_j\}_{j=1}^{C^u} \in \mathcal{Y}^u$, respectively, where $\mathcal{Y}^s \cap \mathcal{Y}^u = \emptyset$. In the following, we will show the principles for designing ZSECOC for seen categories (i.e. $B$) from category-level semantics and visual data structures. Subsequently, a joint optimization framework based on alternating iteration is presented for solving the resultant challenging problem.

3.1. Design Principles

1) Preserving Category-Level Semantics. Previous works [3, 42, 18, 8] have shown that ECOC should have good diversity. We find this property is also crucial for ZSAR, thus our code matrix $B$ is derived from the following properties:

- Column separation: $\max \sum |b_i - b_j|^2$.
- Row uncorrelation: $\frac{1}{C} \sum b_i b_j^\top = I$,
- Row-wise balancedness: $\sum b_i = 0$,

where $I$ is the identity matrix. Besides, as discussed in [45, 61], preserving semantics is crucial for discrimination. We adopt such a property with the following objective function:

$$\min b_{i} \sum s_{ij} ||b_i - b_j||_2, \text{ s.t. } b_i \in \{-1, 1\}^m,$$

where $s_{ij}$ denotes the category-level semantic affinity between the $i$-th and $j$-th categories. Specifically, we capture semantic correlations across categories based on the distributed representation of category names. In practice, we employ the skip-gram neural network model [37] trained on the Google News dataset. Each category is thus embedded by a 300-d word vector $\phi(y_i)$, where ‘$\phi(\cdot)$’ is the embedding function. We assign the cosine similarity between $\phi(y_i)$ and $\phi(y_j)$ to $s_{ij}$, i.e. $s_{ij} = \frac{\phi(y_i) \cdot \phi(y_j)}{|\phi(y_i)| |\phi(y_j)|}$, $i, j = 1, \ldots, C$, where $\cdot$ indicates the inner product operation.

The intuition behind formula (1) is that the ZSECOC of similar categories should be close to each other, while different categories should possess distinct codes. Here, we
denote this property as column ‘association’. Similar objectives are also adopted in graph-based hashing methods [57, 33], because of their capabilities of well preserving semantics and achieving high precision in the retrieval task. By combining all the above objectives, we have
\[
\min_{\beta}\;\sum s_{ij}(b_i - b_j)^2 + \lambda \sum ||b_i - b_j||^2_2,
\]
s.t. \( \frac{1}{C} \sum b_i b_j^\top = \mathbf{I}, \sum b_i = 0, b_i \in \{-1, 0, 1\}^m, \) (2)
where \( \lambda > 0 \) is the trade-off between columns ‘separation’ and ‘association’. By introducing the matrix form of (2), we have
\[
\min_{B} \mathcal{O}_{sp} := \text{trace}(BLL^\top)
\]
s.t. \( BB^\top = CI, B1 = 0, B \in \{-1, 1\}^{m \times C}, \) (3)
where the subscript ‘sp’ indicates the semantics-preserving characteristic of our ZSECOC, and \( L \) is the associated Laplacian matrix of the affinity matrix \( S' = [s'_{ij}] \in \mathbb{R}^{C \times C} \) where \( s'_{ij} = s_{ij} - \lambda, \forall i, j \). Specifically, \( L = \text{diag}(S'1) - S' \).

To tackle this challenging problem, many existing approaches [43, 50, 33, 56, 57] choose to obtain sub-optimal solutions by discarding the binary constraints. As shown in [32, 49, 51], these solutions are of low quality and will lead to less effective classification performance. In this paper, we attempt to address the problem directly without any relaxations and achieve a more accurate solution to \( B \). We will provide the details for optimization in Section 3.2.

2) Capturing Visual Data Structures. According to [63, 65], visual data structures should also be considered for discriminative ECOC. For instance, [63] utilized spectral analysis and [65] employed sum match kernel to acquire useful information from data. We adopt the similar spirit but learn our ZSECOC from data in a different way by using latent factor decomposition (LFD). Specifically, we formulate the problem of learning data-driven ZSECOC as
\[
\min_{R, V} \mathcal{O}_{dd} := ||X - DV||_F^2 + ||V||_F^2 + \alpha ||BP - RV||_F^2,
\]
s.t. \( R^\top R = I, V \in \{-1, 1\}^{m \times N}, \) (4)
where ‘dd’ denotes the data-driven characteristic of ZSECOC, \( || \cdot ||_F^2 \) denotes the Frobenius norm, \( D \in \mathbb{R}^{d \times m} \) is the pre-computed dictionary (or so-called bases) usually obtained by applying k-means or Gaussian mixture models on seen data \( X, V \in \mathbb{R}^{m \times N} \) is the latent factor matrix, \( P \in \{0, 1\}^{C \times N} \) is the category-instance indicator matrix, \( R \in \mathbb{R}^{m \times m} \) is the orthogonal transformation matrix, \( \alpha > 0 \) is the penalty parameter, and \( \gamma > 0 \) is the regularization parameter w.r.t. \( V \). In particular, each entry \( p_{ij} \) of \( P \) is defined as follows:
\[
p_{ij} = \begin{cases} 1, & \text{if } x_j \text{ belongs to the } i\text{-th category,} \\ 0, & \text{otherwise.} \end{cases}
\] (5)
By multiplying the category-level \( B \) with \( P \), we can reconstruct the codes for all seen instances.

The first two terms in (4) correspond to the latent factor decomposition problem. With the dictionary \( D \), a data point \( x_n \) can be reconstructed as \( DV_n \) by using its latent factor \( v_n \). To approximate the final ZSECOC, we derive the latent factors in terms of the same dimension as with the length of the codes (i.e. \( m \)). Furthermore, to fit the codes and the decomposed latent factors, we introduce a penalty term, i.e. the last term in (4). Theoretically, with a sufficiently large \( \alpha \), the resulting ZSECOC can well preserve the intrinsic structure of the visual data. We additionally impose an orthogonal rotation on the factors because such rotation will reduce the quantization loss effectively [19].

Overall Objective Function. By coupling the above two problems, we can learn discriminative ZSECOC from both category-level semantics and visual data structures. The overall objective function is
\[
\min_{R, V, B} \mathcal{O}(R, V, B) := \mathcal{O}_{dd} + \beta \mathcal{O}_{sp}
\]
s.t. \( R^\top R = I, BB^\top = CI, B1 = 0, B \in \{-1, 1\}^{m \times C}, \) (6)
where \( \beta > 0 \) weights the importance between the two characteristics, i.e. semantics-preserving and data-driven.

3.2. Alternating Optimization

The above joint problem (6) is generally NP-hard and non-convex due to the discrete constraint on \( B \). Here, we attempt to tackle it by iteratively computing each of the three variables, i.e. \( R, V \) and \( B \). In other words, we find the solution to one variable while fixing the other two. Similar techniques are adopted in [30, 48, 49].

R-Step: With fixed \( B \) and \( V \), the subproblem w.r.t. \( R \) is
\[
\min_{R} ||BP - RV||_F^2, \text{ s.t. } R^\top R = I. \quad (7)
\]
This objective function is equivalent to the classic Orthogonal Procrustes Problem (OPP) [47]. OPP tries to find a rotation to align one point set (i.e. \( V \)) with another (i.e. \( BP \)). Specifically, the solution to \( R \) is obtained as follows:
\[
U\Sigma\hat{U}^\top = \text{svd}(BPV^\top), \quad R = \hat{U}\hat{U}^\top, \quad (8)
\]
where ‘svd(·)’ denotes the singular value decomposition.

V-Step: The subproblem by fixing \( B \) and \( P \) becomes:
\[
\min_{V} ||X - DV||_F^2 + \alpha ||BP - RV||_F^2 + \gamma ||V||_F^2 \Leftrightarrow \min_{V} ||X_{BP} - D_{RF}V||_F^2 + \gamma ||V||_F^2, \quad (9)
\]
where
\[
X_{BP} = \frac{X}{\sqrt{\alpha BP}} \quad \text{and} \quad D_{RF} = \frac{D}{\sqrt{\alpha R}}.
\]
Particularly, we obtain the solution to OC can be obtained directly without any relaxations. Particularly, we obtain the solution to OC can be obtained directly without any relaxations. Par-
In [23, 60, 21], several techniques were proposed to solve domain shift in a transductive way, by incorporating some examples from unseen categories to refine the visual-semantic mapping. However, as we claimed, this is opposed to our standard ZSL setting. Here, we propose a simple semantic knowledge transfer strategy to acquire $B^u$ without employing any instances from unseen categories. An existing work [61] attempted to utilize a matrix indicating the similarities between seen and unseen categories to fulfill this task. However, the matrix was provided by some volunteers, thus it was highly subjective and unreliable. As we aim to learn ZSECOC automatically without manual intervention, we instead employ a matrix that captures semantic correlations between seen and unseen categories. Particularly, we construct the similarity matrix $S^u = \{s^u_{ij}\} \in \mathbb{R}^{C \times C^u}$ based on the cosine distances between word vectors of seen and unseen categories:

$$s^u_{ij} = \frac{\langle \phi(y_i), \phi(y^u_j) \rangle}{||\phi(y_i)|| \cdot ||\phi(y^u_j)||}, \quad i = 1, \ldots, C, \quad \text{and} \quad j = 1, \ldots, C^u.$$  

Subsequently, we generate $B^u = \text{sign}(BS^u)$. In this way, $S^u$ can well transfer the semantic knowledge from correlated seen categories to unseen ones. More importantly, our $S^u$ is obtained without utilizing any unseen instances.

**Zero-Shot Recognition.** Based on $B$, i.e. the category-level ZSECOC of seen categories, we can learn the visual-semantic mapping via a set of independent binary classifiers (e.g. linear SVMs). Specifically, we regard each row of $B$ as the binary labels for seen categories and one binary classifier is trained based on all the seen data and the associated labels. This will result in $m$ independent binary classifiers: $\{f_i\}_{i=1}^m$. Subsequently, for any unseen data point $x^u$, we can acquire its code through the outputs of these classifiers, i.e. $F(x^u) = [f_1(x^u), \ldots, f_m(x^u)]^\top$. Finally, we formulate zero-shot recognition as the Hamming decoding process [3] of ECOC. We assign the unseen category label $y_j$ to a test data point $x^u$ as follows:

$$j^* = \arg\min_j d_H(F(x^u), b^u_j),$$

where $d_H$ denotes the Hamming distance and $b^u_j$ is the prototype code of the $j$-th unseen category.

## 5. Experiments

### 5.1. Experimental Setup

**Datasets and Settings.** We conduct our experiments on three realistic video action datasets, i.e. Olympic Sports [39], HMDB51 [24] and UCF101 [53], in which there are totally 783, 6766 and 13320 action videos from 16, 51 and 101 categories, respectively. For action representation, we adopt the 50688-d features kindly provided by [60], which are improved dense trajectory (IDT) [55] features encoded by Fisher Vectors (FV) [41]. We adopt the skip-gram neural network model [37] trained on the Google News dataset ($\approx 100$ billion words) and represent each category name by an L2-normalized 300-d word vector. For any multi-word category name (e.g. ‘ride horse’), we generate its vector by accumulating the word vectors of each unique word [38]. For the visual-semantic mapping, we adopt a set of independent linear SVMs [11], as used by the conventional ECOC methods [3]. The lengths of ZSECOC are empirically set to $m = 10 \log_2(C + C^u)$ as suggested in [3], i.e. 40, 70 and 100 w.r.t. Olympic Sports, HMDB51 and UCF101, respectively. We use cross-validations on seen categories to determine the hyper-parameters for our model.

**Evaluation Metric.** Following [60], we adopt the class-wise data splits by evenly dividing each dataset into seen/unseen categories, i.e. 8/8, 27/26 and 51/50 splits with regard to Olympic Sports, HMDB51 and UCF101, respectively. We randomly generate 10 splits for each dataset, and the average recognition accuracies and standard deviations are reported. Due to the randomness of initializing $B$, we report the average results for each split of each dataset based on 5 trials. We conduct the experiments on a PC with an Intel quad-core 3.4GHz CPU and 32GB memory.

In the following, we systematically evaluate our ZSECOC in different aspects. Firstly, ZSECOC is compared to conventional label embeddings. Subsequently, we compare ZSECOC with state-of-the-art ECOC and ZSL methods. Finally, we visualize various qualitative results and present some further analyses as well.

### 5.2. Experimental Results

**Evaluation of Label Embeddings.** We first evaluate different strategies for the label embedding, including semantic attributes, word vectors and our ZSECOC. For the real-valued word vectors, we employ linear support vector regression (SVR) instead of SVMs for learning the visual-semantic mapping. In terms of semantic attributes, [29] and [22] provided the 40 and 115 category-level attributes for Olympic Sports and UCF101, respectively. As no semantic
attributes are available for HMDB51, we omit the attribute-based results on HMDB51. Table 1 shows the ZSAR accuracies on the three datasets. We can have the following observations: (1) Semantic attributes based embeddings can achieve better accuracies than word vectors. (2) Our ZSECOC is the best choice for the label embedding because of its superior characteristics. (3) Shorter codes are required by ZSECOC compared with word vectors and attributes, leading to lower memory load.

We further visually depict the similarity matrices (see Fig. 4) among categories for the three embeddings on Olympic Sports as in [64]. For the binary attributes and ZSECOC, we create their matrices based on the Hamming distance, and the cosine distance is adopted for word vectors. As our ZSECOC is data-driven, we also show the similarity matrix of visual data using the Euclidean distance. We can observe that, in most cases, the colors of the blocks shown in Table 3. Generally, data-dependent ECOC methods perform better than data-independent ones. This implies the necessity of incorporating visual data structures when learning ECOC. Particularly, our ZSECOC consistently achieves the best accuracies on all the three datasets. The performance gains are especially obvious on large-scale datasets, i.e. HMDB51 and UCF101. This mainly owes to the employment of both category-level semantic correlations and instance-level visual data structures when designing our ZSECOC.

Comparison with State-of-the-Art ZSL Methods. We compare ZSECOC with various contemporary ZSL methods: (1) Direct/Indirect Attribute Prediction (DAP/IAP) [26]: the classic attribute-based ZSL strategy; (2) Human Actions by Attributes (HAA) [28]: we adopt the simplified version provided in [60]; (3) The Self-Training model (ST) [59]: domain shift is solved by a transductive self-training procedure; (4) Embarrassingly Simple Zero-Shot Learning (ESZSL) [46]: the mean square loss is used instead of the regression loss w.r.t. the objective function; (5) Structured Joint Embedding (SJE) [2]: a triplet hinge loss is employed to ensure more related labels correspond to higher mapping values from visual features; (6) Unsupervised Domain Adaptation (UDA) [23]: a target domain specific dictionary is learned by using some unseen data; (7) Multi-Task Embedding (MTE) [60]: multi-task regression is developed to learn the visual-semantic mapping, together with a data augmentation strategy.

We notice that some compared methods (e.g. UDA and ST) require the access to unseen instances and we still compare with their results under this transductive setting. However, [59, 60] developed some data augmentation techniques by using examples from some auxiliary categories. In this setting, categories used for training may be re-used during testing, which seriously violates the fundamental assumption.
Figure 6. t-SNE [36] visualization between different embeddings of unseen categories w.r.t. two representative splits on Olympic Sports.

Figure 5. Top-5 returned video examples for unseen categories on Olympic Sports (left) and HMDB51 (right).

Figure 7. Recognition accuracies with increasing code lengths. All the compared methods adopt the fixed-length 300-d word vectors.

6. Conclusion

In this paper, we formulated zero-shot learning as designing error-correcting output codes (ECOC). Discriminative ZSECOC was learned in terms of preserving category-level semantics as well as maintaining intrinsic visual data structures. A joint optimization scheme was proposed to iteratively learn the optimal ZSECOC for seen categories. An intuitive semantic transfer strategy was developed to obtain the ZSECOC of unseen categories without any transductive dependency on test data. The extensive experiments in terms of zero-shot action recognition on three public video action datasets demonstrated the state-of-the-art performance of the proposed ZSECOC.

Acknowledgement

This work was partly supported by the National Natural Science Foundation of China (No. 61573045), the Foundation for Innovative Research Groups through the National Natural Science Foundation of China (No. 61421003), the National Natural Science Foundation of China (No. 61502301), and China’s Thousand Youth Talents Plan.
References


