RECOGNIZING UNSEEN ACTIONS IN A DOMAIN-ADAPTED EMBEDDING SPACE

Yikang Li[†] *Sheng-hung Hu*[†] *Baoxin Li*

Arizona State University

ABSTRACT

With the sustaining bloom of multimedia data, Zero-shot Learning (ZSL) techniques have attracted much attention in recent years for its ability to train learning models that can handle "unseen" categories. Existing ZSL algorithms mainly take advantages of attribute-based semantic space and only focus on static image data. Besides, most ZSL studies merely consider the semantic embedded labels and fail to address domain shift problem. In this paper, we purpose a deep two-output model for video ZSL and action recognition tasks by computing both spatial and temporal features from video contents through distinct Convolutional Neural Networks (CNNs) and training a Multi-layer Perceptron (MLP) upon extracted features to map videos to semantic embedding word vectors. Moreover, we introduce a domain adaptation strategy named "ConSSEV" – by combining outputs from two distinct output layers of our MLP to improve the results of zero-shot learning. Our experiments on UCF101 dataset demonstrate the purposed model has more advantages associated with more complex video embedding schemes, and outperforms the state-of-the-art zero-shot learning techniques.

Index Terms— zero-shot learning, action recognition, multi-layer perceptron, convolutional neural network.

1. INTRODUCTION

Video-based action recognition has many applications [1]. Recent rapid growth of action categories makes conducting video annotation an expensive, challenging and time consuming task. While conventional classifiers require sufficient training data to achieve acceptable results on action recognition tasks, it is difficult and costly to collect satisfactory amount of annotated spatial-temporal segments of videos. The zero-shot learning (ZSL) algorithm provide a solution to mitigate those issues by connecting human-level semantic descriptions of the action with low-level visual features and allowing different categories to share their information –thus enable new categories to be built in terms of their human descriptions rather than extending the size of the training visual-level data. Three keys are rather important in ZSL algorithm – selecting of visual descriptors, constructing human-level semantic descriptors and the mapping function to map visual to semantic space.

Most existing ZSL algorithms are realised by building human-level attribute model to bridge the visual features and their corresponding semantic space. New categories are then classified in terms of their attributes [2, 3]. However, it is rather difficult to obtain reliable attribute-based representation for objects, especially for actions [4], and this kind of semantic attribute-based ZSL classifiers suffer from lacking distributed representation of each attribute words.

An alternative approach to the attribute-based method is the Semantic Embedding Space (SES) [5, 6]. SES is trained by a skip-grim or continuous bag-of-words neural network to map the text words into a word vector space – therefore enable new categories be simply annotated by the similarity and distribution of existing text-string vectors and avoid non-scaleably growth of attribute lists as the emergence of new categories Among all SES models, *Word2Vec* model is considered to be the most efficienct model in maintaining semantic meanings while keeping low model complexity [7, 8, 9].

Although semantic embedding space has demonstrated significant advantages, most ZSL studies only focus on static images semantic embedding since it is particularly difficult to extract reliable feature descriptors which cover both seen and unseen action categories from videos to train the mapping function. Moreover, the presence of amount of neighbour vectors surrounding the mapped vectors in semantic space has been proven to be a challenge for word-based vectors [10] (i.e., "*hubness*" problem or domain shift problem).

In this paper, we design a deep two-output model for video ZSL and action recognition purposes by taking advantages of both "soft" SES labels and conventional"hard" binary labels to train a multi-layer perceptron that map CNN visual features to their corresponding semantic meanings. A new strategy called "convex combination of similar semantic embedding vectors" (ConSSEV) is also implemented to deal with the domain shift problem. The purposed model not only outperforms state-of-the-art [8] method on UCF101 [11] video action dataset on zero-shot learning but also achieve comparative high accuracy with the conventional supervised action recognition classifier on action recognition task.

[†] These two authors have equal contributions.

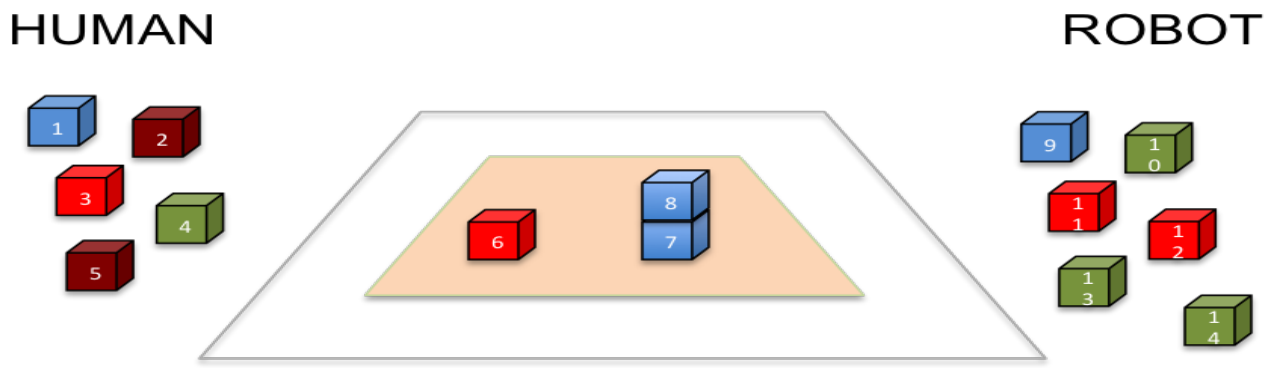

Fig. 1. The framework of our deep two-output model for video ZSL and action recognition purposes. Notice that errors from both output layers are back-propagated.

2. METHOD

The overall framework of the proposed model is illustrated in Fig.1. We first extract frames and optical flow [12] from video contents and then pass them into two different pre-trained CNN models [13]. The appearance features and optical flow features are collected from the second last fully connected layer (i.e., "fc7") from each CNN. Next we aggregate and concatenate both features to represent one action by using a sliding window strategy. Then, a two output layer multi-layer perceptron is trained by backpropagating errors from semantic labels and fine-tuned on softmax hard binary labels to serve as a mapping function from visual to semantic space. Finally, zero-shot learning (ZSL) is performed by mapping visual features to semantic space vectors through our model and Convex combination of Similar Semantic Embedding Vectors (ConSSEV) is implemented as a domain shift method.

2.1. Visual Feature Extraction

Considering the success of Pose-based Convolution Neural Network [13] on recognizing human-posed and action, we extracted features in a similar way. Videos are first sampled to RGB frames to represent appearances and optical flows are computed to represent motions [12]. To describe both appearance features f^t_{app} and motion features f^t_{of}, we use two different CNNs on RGB and optical flow frames respectively and the output of the second last fully-connected layer [14] with dimension $k = 4096$ is served as our descriptors. For RGB frames the publicly available "VGG-f" pre-trained network [14] is used while the pre-trained motion network [15] is applied on optical flows.

A Sliding Window Strategy: We applied a sliding window with size T and step size S on both f^t_{app} and f^t_{of}. Features within the same window are avg aggregated to obtain fixed-length video descriptors d^i over T frames

$$d^i_{app} = \frac{1}{T}\sum_{t=t_0}^{t_o+T-1} f^t_{app}(i) \quad (1)$$

$$d^i_{of} = \frac{1}{T}\sum_{t=t_0}^{t_o+T-1} f^t_{of}(i) \quad (2)$$

$$d^i = [d^i_{app}, d^i_{of}]. \quad (3)$$

Finally d^i is the extracted visual features for the i^{st} class with dimension $k = 8192$.

2.2. Semantic Embedding Space

In this paper, we construct the semantic embedding space with the help of the *word2vec* neural network [5, 6] because of its reliable mapping between word corpses and mathematical meanings. Through this networks, semantic labels $\{y^i\}_{i=1...n}$ are assigned to 500-D vectors $Z^i = g(y^i)$ and are divided by the amount of unique words in $Z^i = \frac{1}{N}\sum_{j=1}^{N} g(y^i_j)$ for normalization purpose.

2.3. Mapping Function

Given visual features d^i and semantic embedded space labels Z^i, our goal now is to build a projection model: $Z^* = M(d^i)$ that can best map each video to a vector in the corresponding semantic embedding space. Inspired by the idea of [16], A two-output multi-layer perceptron (MLP) is trained to achieve this goal.

Both the semantic space soft labels Z^i and hard binary labels y^i are applied on training the MLP. Two distinct loss functions are calculated and both errors are backpropagated. For semantic soft label loss, we use hinge rank loss function (similar to [7]) to measure the similarity of the semantic output Z^* and semantic labels Z. The hinge rank loss function is defined as:

$$L_{se} = -\sum_{j\neq i} \max[0, m - Z^i \cdot Z^* + Z^j \cdot Z^*] \quad (4)$$

where m represents the margin value.

For hard binary label loss, the softmax loss function is selected because of its robustness for multi-class classification. The probability and loss for the softmax layer input Z^* be classified to the j^{th} class (i.e., $\bar{y} = j|Z^*$) is defined as:

$$P(\bar{y} = j|Z^*) = \frac{e^{Z^{*T} w_j}}{\sum_{k=1}^{K} e^{Z^{*T} w_k}} \quad (5)$$

$$loss_{sm} = -\sum_{j} 1\{\bar{y} = j\} \log(P(\bar{y} = j|Z^*)). \quad (6)$$

Here $1\{\bar{y} = j\}$ equals to 1 when predicted label $\bar{y}$ equals to target label, otherwise it equals to 0.

Both softmax and semantic output leverage visual and semantic similarity to train MLP since two outputs share information in input and hidden layer. Moreover, the summation of softmax and semantic loss backpropagate to learn weights in each layer, so that each loss can be serve as a compensation and fine-tuning method for the other.

2.4. Zero-Shot Learning and ConSSEV strategy

Zero-Shot learning is then performed on a completely disjoint dataset from training set. Utilizing our deep-multi output model and $word2vec$, we are able to project both "unseen" visual contents d^i_{test} and their corresponding word labels y^i_{test} to semantic embedded space vectors (Z^*_{test} and Z^i_{test})). Since both vectors are normalized, a simple cosine similarity is performed to match projected visual contents and labels

$$\bar{y} = \arg\max cos(Z^*_{test}, Z^i_{test}). \quad (7)$$

ConSSEV strategy: Due to the disjointness of train and test set in ZSL, the trained mapping function M may not be the best fit in the case of mapping test set and thus biases the similarity measurement [7, 8]. In this paper, we introduce a self-adaptive domain shift method by utilizing both semantic and softmax outputs in our model to adjust both semantic output Z^*_{test} and label prototype Z^i_{test}.

For semantic output, we first find the top K training labels that have the highest similarities with test semantic label vector $\{Z^*_k\}_{k=1...K}$ as that in Eq. (7). Then we form a new semantic vector $\bar{Z}^*$ by weighted sum of all k vectors with their corresponding softmax output $P(\bar{y} = k|Z^*_k)$. The adaptive semantic output vector performs better since it penalized "ambiguous feature results" by weighting smaller softmax probabilities ("confidences").

$$\bar{Z}^* = \frac{1}{K} \sum_{k=1}^{K} P(\bar{y} = k|Z^*_k) \cdot Z^*_k \quad (8)$$

For label prototype, we perform the same $self - training$ techniques as that in [8] on the adaptive data vector obtained by Eq. (8). The new convex combination of semantic output $\bar{Z}^*$ and test label prototype $\bar{Z}^i_{test}$ are more directly comparable by using Eq. (7).

2.5. Conventional Video Recognition

Our deep two-output model can also serve as a conventional video recognition classifier by supervised training the mapping function M with all categories. Then we map each test visual feature d^i_{test} to a vector Z^* in semantic space through $Z^* = M(d^i_{test})$ and matching them with all projected labels Z^i through Eq. (7).

3. EXPERIMENTS

3.1. Dataset

We train and test our model on one of the most challenging video action dataset – UCF101 [11] which contains 13320 videos from 101 cation categories (e.g."Apply Eye Make","Basketball Dunk",and "Brest Stroke"). Videos in each action category are grouped into 25 groups where each group shares some common features, such as background, viewpoints, objects,...etc.

For evaluating ZSL, we use the same evaluation protocol as in [8] – 30 independent splits for UCF101 dataset with each split contains a completely disjoint 51 categories for training purpose and 50 for testing purpose. For conventional action recognition, on the other hand, we use the standard splits ("Three Train/Test Splits") for UCF101 dataset.

3.2. Experiment Setting

Semantic Embedding Space: *Word2Vec* [5, 6] method is used to embed the text labels. We trained a skip-gram text model on a corpus containing 5.4 billion words extracted from wikipedia.org. Dimension of word vector is set to 500-D to trade-off training complexity and maintaining semantic meanings [7, 8].

Visual Feature Extraction: To further decrease model complexity, only one frame is sampled for each three consecutive frames from video and optical flows [12] are computed upon them. Two distinct convolutional neural networks are applied to extract features – both contain 5 convolutional and 3 fully-connected layers. The appearance and optical flow features are extracted from the second last fully-connected layer (i.e.,"fc7") which dimension for each feature is 4096. Then we use the sliding window strategy to aggregate and concatenate the appearance and optical flow feature vectors into one 8192-D vector.

Mapping function training: A Multilayer Peceptron (MLP) is trained with those aggregated visual features for each video clip. Target labels for softmax output are so-called "hard binary labels" (i.e., "0" and "1"), and target labels for semantic embedding output are the 500-D semantic space word vectors. The number of hidden nodes is set to be 1000, learning rate to be 0.001, the momentum to be 0.9 and margin value for rank loss function to be 0.9 based on the result of cross-validation. Moreover, for each splits, we train five iterations

of all training features with random training order.

Results Comparison: For zero-shot learning, we compared the following methods on the same split data of UCF101: (1) Random Guess: the method randomly guesses one label from the unseen labels. (2) Attribute Based-Indirect Attribute Prediction (IAP) [2]: the method selects the unseen label by the video representation attributes. (3) Convex Combination of Semantic Embeddings (ConSE) [9]: the method use the conventional neural network classifier output (i.e. softmax output) to weight the training labels and combine the top K embedded labels to denote a new semantic embedding word vectors. (4) Dense Trajectories Based Regression Model with Nearest Neighbour (DTRM+NN) [8]: the model is trained a SVM classifier with RBF kernel on the dense trajectory descriptors [17]. This method is treated as our baseline. (5) Our deep two-output model with Nearest Neighbour: finding the nearest neighbour in terms of maximal cosine similarity. (6) DTRM+NN+ST: Apply Self-training domain shift method with DTRM. (6) Our deep two-output model with ConSSEV approach: Apply ConSSEV domain shift strategy on our model.

For video recognition, our model is validated with the following: (1) Dense Trajectories [17]. (2) Binary SVM classifier with RBF kernel (DTRM) [8]. (3) Our model Semantic output. (4) Our model Softmax output.

3.3. Evaluations

3.3.1. Performance of Zero-shot Learning

The results of zero-shot learning are presented in Tab.1. All listed methods are significantly better than random guess which shows successful ZSL. Without applying any kinds of domain shifting techniques and only consider the Nearest Neighbour (NN), our deep two-output model achieves a slightly better performance than existing state-of-the-art semantic-based ZSL (DTRM) – suggesting visual contents are effectively mapped to semantic space vectors that are near to its human-level semantic meanings. Although our model fails to demonstrate better performance than attribute-based model [2] when evaluating by NN, it does not suffer from lack of attributes and costly attribute annotation. By applying our ConSSEV domain shift strategy, our model significantly outperforms other domain shift counterpart (DTRM+NN+ST).

Overall, our zero-shot learning technique based on MLP has a great performance among the existing state-of-the-art ZSL methods and the ConSSEV domain shift strategy between test and train categories proves to be a significant performance boost on ZSL technique.

3.3.2. Performance of Action Recognition

The performance of our model conducting conventional action recognition task is listed in Tab.2. The final results reveal our approach performs comparatively with the state-of-the-art method including Dense Trajectories [17] and Binary SVM classifier with RBF kernel [8]. Thus demonstrating the ability of our deep two-output model on addressing conventional action recognition tasks. We can also observe that Dense Trajectory [17] performs slightly better than our approach. This may due to the sliding window strategy that is used to extract video features. Even though motion information of optical flow is utilized in our model, the averaged features within a sliding window will lose some temporal information of video sequences compared with HOF and MBH features in Dense Trajectories [17].

Table 1. Zero-shot learning performance

Method	**Accuracy $\pm$ Variation**
Random Guess	2.0
IAP [2]	12.8 ± 2.0
ConSE [9]	10.5 ± 2.0
DTRM + NN [8]	10.9 ± 1.5
Ours + NN	11.3 ± 2.1
DTRM + NN + ST [8]	15.8 ± 2.3
Ours + NN + ConSSEV	26.8 ± 4.4

Table 2. Conventional action recognition performance

Method	**Accuracy**
Dense Trajectories [17]	75.1
DTRM [8]	73.7
Ours(Semantic)	74.1
Ours(Softmax)	72.7

4. CONCLUSION

In this paper we train a deep two-output model to realise the zero-shot learning paradigm on video recognition. A domain shift technique, convex combination of similar semantic embedding vectors (ConSSEV), which proves to provide a significant improvement in terms of zero-shot learning accuracy by utilizing the known semantic space to express the unknown semantic space, is purposed. We show that our zero-shot learning model with ConSSEV strategy greatly outperforms state-of-the-art zero-shot video action recognition techniques.

Acknowledgments: The work was supported in part by ONR grant N000141512344 and ARO grant W911NF1410371. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of ONR or ARO.

5. REFERENCES

[1] Zheshen Wang and Baoxin Li, "Human activity encoding and recognition using low-level visual features.," in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2009, pp. 1876–1883.

[2] C.H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 3, pp. 453–465, March 2014.

[3] Heng-Tze Cheng, Martin Griss, Paul Davis, Jianguo Li, and Di You, "Towards zero-shot learning for human activity recognition using semantic attribute sequence model," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, New York, NY, USA, 2013, UbiComp '13, pp. 355–358, ACM.

[4] Chuang Gan, Ming Lin, Yi Yang, Yueting Zhuang, and Alexander G.Hauptmann, "Exploring semantic interclass relationships (sir) for zero-shot action recognition," 2015.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, Eds., pp. 3111–3119. Curran Associates, Inc., 2013.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[7] Andrea Frome, Greg Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances In Neural Information Processing Systems, NIPS*, 2013.

[8] Xun Xu, T. Hospedales, and Shaogang Gong, "Semantic embedding space for zero-shot action recognition," in *Image Processing (ICIP), 2015 IEEE International Conference on*, Sept 2015, pp. 63–67.

[9] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean, "Zero-shot learning by convex combination of semantic embeddings," *CoRR*, vol. abs/1312.5650, 2013.

[10] Georgiana Dinu and Marco Baroni, "Improving zero-shot learning by mitigating the hubness problem," *CoRR*, vol. abs/1412.6568, 2014.

[11] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

[12] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European Conference on Computer Vision (ECCV)*. May 2004, vol. 3024 of *Lecture Notes in Computer Science*, pp. 25–36, Springer.

[13] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid, "P-CNN: Pose-based CNN Features for Action Recognition," in *ICCV 2015 - IEEE International Conference on Computer Vision*, Santiago, Chile, Dec. 2015.

[14] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *CoRR*, vol. abs/1405.3531, 2014.

[15] G. Gkioxari and J. Malik, "Finding action tubes," in *CVPR*, 2015.

[16] Ragav Venkatesan and Baoxin Li, "Diving deeper into mentee networks," *arXiv preprint arXiv:1604.08220*, 2016.

[17] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013.