# Zero-Shot Egocentric Action Recognition

TeamSigma- Aayush Kumar, Dinesh Patel, Harsh Lohia, Rakesh Kumar, Sonu Kumar Shaw
*Indian Institute of Technology, Guwahati*

***Abstract*—Zero-shot learning is a very promising research topic. For instance, for a vision-based action recognition system, zero-shot learning allows us to recognize actions that are never seen during the training phase. Previous works in zero-shot action recognition have exploited the visual appearance of input videos to infer actions in several ways. Here, we propose to use external knowledge to improve the performance of purely vision-based systems. Specifically, we have explored three different sources of knowledge in the form of text corpora. Our resulting system follows the literature and segregates actions into verbs and objects. In particular, we independently train two vision-based detectors: (i) a verb detector and (ii) an active object detector. During inference, we combine the probability distributions generated from those detectors to obtain a probability distribution of actions. Finally, the vision-based estimation is combined with an action prior extracted from text corpora (external knowledge). We evaluate our approach on the EGTEA Gaze+ dataset, an Egocentric Action Recognition dataset, demonstrating that the use of external knowledge improves the recognition of actions never seen by the detectors.**

## I. Introduction

### A. Problem Statement

Given a video clip, the goal is to predict a single way label for the video in a zero-shot way. The challenge is to do it for egocentric video.

### B. Zero Shot Learning

Zero-shot machine learning is used to construct recognition models for unseen target classes that have not labelled for training. It is done in two stages. Training: Where the knowledge attribute is captured. Inference: The knowledge is then used to categorise instances among new sets of classes. ZSL recognition relies on the existence of a labelled training set of seen classes and the knowledge about how each unseen class is semantically related to the seen class.

### C. Challenges and Motivation

The regime of deep learning has been observed as the potent game changer in the field of technology. This has thus attracted multidimensional intensive researches in the stream. One of the major challenges in developing such models is of computing power requirements and training time taken by them. Thus, improvising the existing models to outperform the existing models on these parameters form the basis of ongoing quest among researchers in the field. The focus has been to reach upto to the level demanding less cost and time. This will ensure that the mankind can harness more advantages from such models.

## II. Related Works

### A. Gaze Attention[7]

To address difficulties in incorporating gaze data in attention mechanism, probabilistic modelling method is used. Specifically, the locations of gaze fixation points is represented as structured discrete latent variables and the distribution of gaze fixations is modelled using a variational method. Then the predicted gaze locations are integrated into a soft attention mechanism to make the intermediate features more attended to informative regions. Ablation study and qualitative study is performed to demonstrate effectiveness of attention mechanism

### B. Use of external knowledge[4]

Both the verb and active objects were separately inferred. In this method, the system would be capable of making predictions by combining the knowledge acquired from those two separated branches, as long as the verb and the object have been previously seen. Naively combining verbs and objects may wind up with action predictions that do not exist, for example “cut fridge” which makes no sense. Thus, we propose to add external knowledge to the system to address those problems and improve the performance.

### C. Open-world egocentric action recognition[5]

This approach integrates commonsense knowledge and symbolic reasoning with the representation learning capabilities of deep neural networks to overcome the dependence on annotated training data. This approach uses a general-purpose knowledge base that is not specifically tailored for the specific domain and has no learned correspondence in the target domain.

### D. Incorporating Visual Grounding in GCN[6]

Introduction to the problem of zero-shot learning in human object interaction actions. Proposed method to visually ground the language graph using the visual adjacency matrix. Addresses the limitation of using only the language graph in zero-shot learning, ignoring the semantics of the visual graph. The method outperforms the state-of-the-art methods on both datasets

## III. Proposed Work

### A. Using Concept Net for ZSL

ConceptNet is a freely- available semantic network, designed to help computers un- derstand the meanings of words

that people use. ConceptNet originated from the crowdsourcing project Open Mind Com- mon Sense, which was launched in 1999 at the MIT Media Lab. It has since grown to include knowledge from other crowdsourced resources, expert-created resources, and games with a purpose. Here we propose to use ConceptNet for improving current/existing ZSL models by using ConceptNet as a common sense knowledge base.

### *B. Methodology*

In the context of egocentric action recognition, a ZSL approach aims to create a model which is capable of recognising actions that have never been seen during the training phase. As the existing literature reviews suggested, an action has been separated into a verb and an active object. As per this an action performed in an input video is identified using the combinations of a verb and an object. Following that idea, two identical neural networks have been developed to detect the verb and the active object. The problem is thus simplified as a classification problem, where given a video, i.e., a sequence of frames, the detectors estimate the probability distribution over the set of verbs and active objects. Consequently, both probability distributions are combined to infer the action such that a = maxipv(ai), where a is the action label and pv(ai) denotes the probability for the ith action estimated by the vision-based system. This probability is calculated as pv (ai ) = p(vai ) × p(oai ), where p(vai ) and p(oai ) denote the probability of the verb and object disentangled from ai. Those probabilities are estimated by the neural networks DV and DO. On the top of it, a concept net has been deployed to compute a probability distribution of all the combinations of verbs and objects. Specifically, using the concept net the probability of whether a particular combination of a verb and an active object makes sense or not is calculated. The output of the concept net is combined with the probabilities of the actions obtained from the combination of the verb and object detectors. The final action prediction is the one with the highest probability. More formally, given pv(ai) and pt(ai) (the output from the concept net for ith action), the inferred action a is calculated as a = maxipv(ai) × pt(ai). An overview of the system is shown in Figure 1.

### *C. Network Architecture*

A video X = F1, F2, ..., Fn (an ordered list of frames of the video, where Fi  R224x224x3), is fed to both the networks DV and DO, the verb and object detectors, as an input. As the videos have a varying length, 25 frames are uniformly sampled from each one. The network architecture is based on the work of Sudhakaran et al., being composed of a Convolutional Neural Network (CNN) with a Recurrent Neural Network (RNN) on top, as the feature extraction part, and a single Fully-Connected (FC) layer as the classifier. To be more specific, a ResNet50(He,K.,Zhang,X.,Ren,S.,Sun,J et al.) architecture as the CNN (with a 1x1 convolution of 256 filters on top to reduce the dimensionality) and a Convolutional Long Short-Term Memory (ConvLSTM) (Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C. et al.) as the RNN has been used. The architecture of the ResNet-50 has been shown in the Fig. 2 and that of ConvLSTM in Fig. 3. The detector outputs a probability distribution p = p1 , p2 , ..., pn , where pi  [0, 1] is the probability of the class i for a given video. Thus, Object and verb detection is done using a combination of ResNet-50 and LSTM. It's worth noting that as CNNs tend to struggle to capture the temporal dynamics of the video recurrent neural networks (RNNs), in the form of ConvLSTMs have been used. ResNet-50 is also being used for Feature Extraction. In the DV and DO networks, the orange blocks have been frozen and the yellow ones have been trained(ref. fig.1). The first layer of the ResNet-50 is made up of Conv block followed by Batch Norm, followed by ReLU, followed by max pool. The remaining four blocks are made up of Conv Block and ID Block. ConvLSTM has been used for temporal modeling.Temporal modeling refers to the process of capturing and understanding the temporal (time-related) aspects of data, particularly in the context of time series analysis, sequential data, or any data that varies with time. Thus, in the model used, ResNet-50 is combined with the ConvLSTM and classification is being done by Fully Connected Layer. Both the ConvLSTM and the Fully Connected Layer have been trained. The major benefits leveraged from this combination are:

1. Leveraging Spatial Feature Extraction: ResNet-50 provides strong spatial feature extraction capabilities, allowing it to capture the details of objects within each frame.

2. Capturing Temporal Dynamics: ConvLSTM effectively models the temporal relationships between frames, enabling it to learn how objects move and interact over time.

3. Improved Classification Accuracy: By combining the strengths of both architectures, the model can potentially achieve better classification performance compared to using ResNet-50 or ConvLSTM alone.

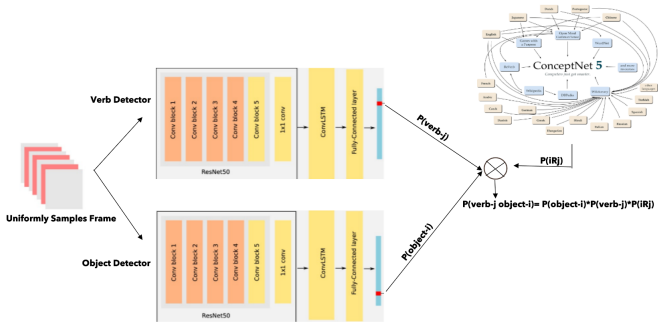


Fig. 1: Architecture overview

**Two neural networks composed of a ResNet50 and a Con- vLSTM take as input a video (uniformly sampled frames) and output two probability distributions (verbs and objects). The resulting probability distributions are combined with a probability value from concept net infer the most probable action. The layers or blocks of layers in orange are frozen while the yellow ones are trained.**

### *D. Using concept net*

We use concept net to provide a sense of common sense knowledge to the model. Concept net is basically a relational

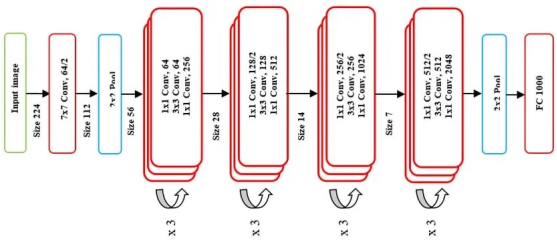


Fig. 2: ResNet-50 Architecture

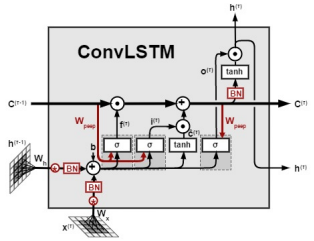


Fig. 3: ConvLSTM Architecture

knowledge graph. We input object and verb into the concept net probability calculation code and obtain a probability indicating, how much sense the particular action make.

### E. Cost Functions

In the context of the zero-shot egocentric learning model developed, two tasks: verb detection and object detection have been separately dealt with. Each task can have its own loss function. In our model, since ResNet-50 and ConvLSTM has been employed for the object classification, cross-entropy loss has been calculated. The loss function used for training neural network for that matter, depends on the specific task for which the network is being used. Since, the task is primarily a classification-based one, the cross-entropy cost function or the categorical cross-entropy loss has been referred to. The categorical cross-entropy loss, also known as log loss, is used for multi-class classification problems, where each input is assigned to one of several possible classes. Given a ground truth label (one-hot encoded) for each input and the predicted class probabilities produced by the network, the loss measures the dissimilarity between the predicted probabilities and the true label.The formula used is:

$$L = -\sum ytrue * log(ypred)$$

Where L represents the loss

ytrue is the true class distribution for the video sequence.

ypred is the predicted class distribution for the video sequence.

The model is equipped with the loss function by implementing these steps: A loss function, CrossEntropyLoss(), which is used for image classification tasks. is defined. It combines a softmax activation function and the categorical cross-entropy loss. Stochastic Gradient Descent, or SGD is used as an optimizer to update the model's weights based on the computed gradients. In the training loop, the dataset is iterated, the loss is calculated, backpropagation is performed, and the model's weights are updated. The network is trained to minimize this categorical cross-entropy loss. The goal is to have the predicted class probabilities be as close as possible to the true labels for the training data.

## IV. Experimental Details

### A. Datasets and Training Details

We chose the EGTEA Gaze+10 dataset for our experiments. Launched in 2017, this dataset contains 28 hours of egocentric videos with 32 subjects performing cooking related actions. It is composed of 10,325 action segments, with 19 verbs, 53 nouns and 106 actions.

Official splits in EGTEA are not suitable for ZSL, since the actions in the test set are also represented in the train set. Therefore, in our experiments, we employ new splits.

First, we removed action videos containing verbs and objects that only appear once, as they are not appropriate for the zero-shot task, as formulated in given papers. This left us with 9 verbs and 29 objects.

Second, to generate the test set, we randomly took 20 persent of the action classes under the condition that any verb and object contained in that test set must appear in the training set (in any action). That is, all the verbs and objects must appear in the training set.

The validation set is created taking a stratified subset from the resulting training set, using the 10 persent of the videos in train. Note that the validation set is important not for the ZSL task itself, but to train and tune the detectors.

A video input is segmented into 25 frames uniformly.

## V. Conclusion

The model proposed here aims to increase the accuracy of the SOTA model in the case of ZSL for egocentric action without affecting other parameters, such as training time and loss functions. The model employs ResNet 50 and convLSTM for object detection and verb detection and further combines the obtained probability distribution with a relational probability obtained from the concept net. Using the concept net as a knowledge base, it aims to reduce or avoid blunders. It could be limited by the existing depth of knowledge graph(concept net) and may increase run time since concept net is not available locally on the system.

Code- https://github.com/shawsk04/CS590-Course-Project

## VI. References

1) He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770778 (2016)
2) Sudhakaran, S., Lanz, O.: Attention is all we need: nailing down object-centric attention for egocentric activity recognition. arXiv preprint arXiv:1807.11794 (2018)
3) Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems. pp. 802810 (2015)

4) Using External Knowledge to Improve Zero-Shot Action Recognition in Egocentric Videos Adrián Núñez-Marcos, Gorka Azkune, Eneko Agirre, Diego López-de-Ipiña  Ignacio Arganda-Carreras
5) Knowledge guided learning: Open world egocentric action recognition with zero supervision Author links open overlay panelSathyanarayanan N. Aakur, Sanjoy Kundu, Nikhil Gunti
6) Incorporating Visual Grounding in GCN for Zero-Shot Learning of Human Object Interaction Actions, Chinmaya Devaraj, Cornelia Fermuller, and Yiannis Aloimonos,  University of Maryland
7) Integrating Human Gaze into Attention for Egocentric Activity Recognition by Kyle Min, Jason J Corso.