IIT GUWAHATI

Team: Sigma

Group Member	Roll Number
AAYUSH KUMAR	200104001
DINESH PATEL	200104031
HARSH LOHIA	200104037
RAKESH KUMAR	200104087
SONU KUMAR SHAW	200104107

Zero-Shot Egocentric Action Recognition

Report(Phase I)

Date: 29.09.2023.

Zero-Shot Egocentric Action Recognition

Problem Statement:

Given a video clip, the goal is to predict a single action label for the video in a zero-shot way. However, the challenge is to do this for egocentric videos.

Understanding the problem statement:

Zero-shot learning is a machine learning paradigm where the model is trained to recognize classes or concepts it has never seen during training. So, the aim of the model is to predict the actions in an input video clip. The video clip consists of various human-object ineractions. Egocentric videos form a special class of videos as they are videos taken from a first-person point of view using wearable cameras.

Introduction:

Zero-shot action recognition is emerging as one of the most engaging and fruiful domain in the regime of deep learning and the associated fields like computer vision, natural language processing etc. Predicting actions in egocentric videos in a zero-shot manner is a complex task that requires a combination of computer vision, deep learning, and zero-shot learning techniques.

The task involves gathering dataset and annotation of actions in the preliminary stage. Feature extraction fron the egocentirc videos is done in the next step. CNNs are the most used resource in the feature extraction to analyze the frames in the video. Following this, a zero-shot learning setup is developed and on the top it action labels and the features extracted from the egocentric videos are mapped into a common embedding space. Techniques like word embeddings visual embeddings are used for this purpose. In the zero-shot learning space, a semantic space is created that represents the available action labels and their relationships. A zero-shot learning model is trained that can predict actions from the available labels while also generalizing to unseen actions. Common models for zero-shot learning include generalized zero-shot learning (GZSL) models, attribute-based models, and semantic embedding models.

Evaluation of the model is done on the testing datasets. The results of the state-of-the-arts(SOTA) of each technique is compared to judge the efficiency of a model. For any model developed, fine-tuning and iterations are performed to reach the maximum potent level of a model.

The zero-shot action recognition is in itself a novel field of active research. In addition, to perform this on egocentric videos add another complex dimension to the problem. Zero-shot action recognition is a versatile field within computer vision that enables the recognition of human actions, even in unseen scenarios, reducing the need for extensive labeled data. Its applications range from surveillance and augmented reality to healthcare and education, making

it relevant in diverse domains. By allowing models to generalize across different contexts and domains, zero-shot action recognition has the potential to significantly impact human-computer interaction and real-world applications.

Integrating Human Gaze into Attention for Egocentric Activity Recognition

Introduction:

- Attention mechanism allows the model to focus on specific parts of the input data that are deemed most relevant to the task.
- Integrating human gaze into attention for egocentric activity recognition involves using gaze data to influence where the model pays attention within the video frames.
- However, there is always uncertainty in the process of recording the gaze fixation points because of saccadic suppression and measurement errors.
- To address such problems, probabilistic modeling method can be used:
    - First, represent the locations of gaze fixation points in space and time as structured discrete latent variables to model their uncertainties.
    - Second, model the distribution of the gaze fixations using a variational method. During the training process, the distribution of gaze fixations is learned using the ground-truth annotations of gaze points.
    - The predicted gaze locations are integrated into a soft attention mechanism to make the intermediate features more attended to informative regions.

      Also perform an ablation study to verify that probabilistic modeling of gaze data is truly beneficial.

Network Architecture:

As a backbone network, the two-stream I3D (3D CNN) is used.

- To model the gaze distribution, the same convolutional blocks of the I3D is used and three convolutional layers are added on top of it.
- The two intermediate features at the end of the 4th max-pooling layer are added and the added feature map is used as an input to the network for gaze modeling.
- A sample is drawn, which is then applied with a fully connect layer and the sigmoid function to produce a soft attention map.
- The two features at the end of the 5th convolutional block are added in an elementwise way, and the soft attention map is applied to the added feature map via a residual connection.

Limitations:

- **Gaze Data Availability:** Collecting gaze data in real-world scenarios can be challenging. While datasets like EGTEA Gaze+ provide valuable gaze information, they might not cover all possible scenarios, making it difficult to generalize to other contexts.
- **Calibration and Accuracy:** Gaze tracking technology may require calibration for each user, and the accuracy of gaze tracking systems can vary. Inaccuracies in gaze data can affect the quality of attention models.

LITERATURE REVIEW: Exploring Ways To Improve ZSL For Egocentric Videos

Title: Using External Knowledge to Improve Zero-shot Action Recognition in Egocentric Videos

Authors: Adri´an N´u˜nez-Marcos, Gorka Azkune, Diego L´opez-de-Ipi˜na, Ignacio Arganda-Carreras

Introduction:

- To improve Egocentric Action Recognition, the paper suggested using external knowledge from text corpora to create action priors.
- The paper's contributions include a novel method for enhancing EAR using external knowledge and an analysis of the effects of various knowledge sources.

Methodology:

- Actions were divided into verbs and active objects, each detected using separate neural networks.
- This was treated as a classification problem where the networks estimate probability distributions over verbs and active objects based on video input.
- Action recognition was achieved by combining the probability distributions, selecting the action with the highest probability.
- The probability of an action was calculated as the product of verb and object probabilities, learned by individual neural networks for object and verb detection.
- External knowledge in the form of text corpora was used to create an "action prior" by identifying co-occurrences of verbs and objects in N-grams from the text.
- The action prior was combined with probabilities from the neural networks to make the final action prediction.

Action priors:

- The action prior represents a probability distribution over actions resulting from the combination of verbs and objects in a given dataset.
- The action prior aims to estimate the likelihood of specific verb-object combinations based on external knowledge sources independent of action recognition videos.
- The paper utilizes three external knowledge sources to estimate action priors:
    - Cookbook wiki: Extracts a corpus from the Cookbook wiki containing recipes and cooking-related actions, focusing on specialized knowledge.
    - Google searcher API: Uses this API to search for actions and determine their frequency, providing a more general prior estimation not limited to a specific domain.
    - Phrasefinder searcher API: Similar to Google API but searches through Google Books' N-grams, offering a controlled alternative to the Google API for prior estimation.
- To create the action prior from the Cookbook source:
    - The Wi-kicook corpus is scraped and cleaned, including removing non-ASCII characters, lowercase text, eliminating stop words, and applying WordNet lemmatization.
    - N-grams of size 4 are extracted.
    - Actions are identified in N-grams based on the presence of both verb and object, not necessarily in adjacent positions.
    - Synonyms for verbs and objects are considered, and the action is considered present if at least one synonym of the verb and one synonym of the object appear in the N-gram.
    - The final prior for an action is the number of N-grams containing the action divided by the total number of N-grams.
- For the Google and Phrasefinder sources:
    - The API returns the number of results for a query constructed as "verb * object" (Google API) or "verb ? object" (Phrasefinder API), using wildcards as placeholders.
    - Synonyms for verbs and objects were considered, and the mean of all non-zero results is used as the final frequency of the action.
    - The frequency is normalized by the sum of frequencies for all actions to obtain the action prior.

Architectural Review:

- The verb and object detectors, processed video inputs composed of ordered frames, with each frame represented as $Fi \in R224x224x3$.
- To handle varying video lengths, the network uniformly sampled 25 frames from each video.
- The network architecture combined a Convolutional Neural Network (CNN) based on ResNet50 with a Convolutional Long Short-Term Memory (ConvLSTM) for feature extraction and temporal context modeling.
- Classification was performed using a single Fully-Connected (FC) layer.

- Both detectors outputted probability distributions (p) over classes, with pi representing the probability of class i, ensuring that the sum of all probabilities equaled 1 (Σ pi = 1).
- Depending on the task, the network defined p as p(v) for verb prediction (output of DV) or p(o) for active object prediction (output of DO).

Limitations:

- Limited datasets hinder EAR system scalability.
- The choice of the knowledge source is critical.

Open world egocentric action recognition with zero supervision:

#Title: Knowledge guided learning: Open world egocentric action recognition with zero supervision.

#Authors: *Sathyanarayanan N. Aakur, Sanjoy Kundu, Nikhil Gunti Department of Computer Science, Oklahoma State University, Stillwater, OK 74078, USA*

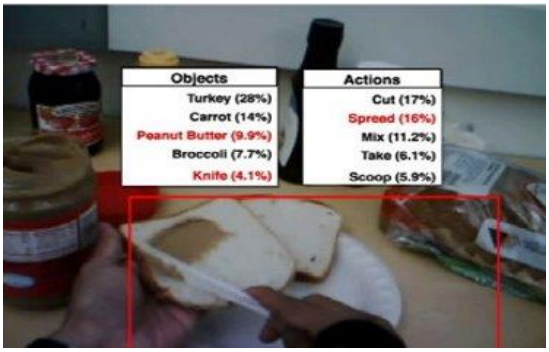
#Introduction:

- This paper presents one of the first works on open-world action recognition in egocentric video.
- This paper formulates a novel approach that integrates commonsense knowledge and symbolic reasoning with the representation learning capabilities of deep neural networks to overcome the dependence on annotated training data.

#Approach:

- The common approaches are to either use an attribute space or embedding space that captures the semantics of a scene and helps extend beyond the training label by exploiting the semantic correspondences across classes. However, the success of such models relies on the presence of "seen" training classes that allow it to establish semantic correspondences to recognise a finite set of unseen actions.
- This constraint does not restrict this concept since it exploits an object's compositionality and functionality to move beyond a fixed vocabulary.
- This model employs the concept Net as the source of graph-based commonsense Knowledge.
- This approach uses a general-purpose knowledge base that is not specifically tailored for the kitchen domain and has no learned correspondence in the target domain.
- This approach generalizes well across domains without any supervision, including access to target domain data. This is a key difference between this approach and other models (including ZSL approaches), which have access to the target domain data for other known

classes and are only expected to learn correspondences for a small number of "unseen" classes.



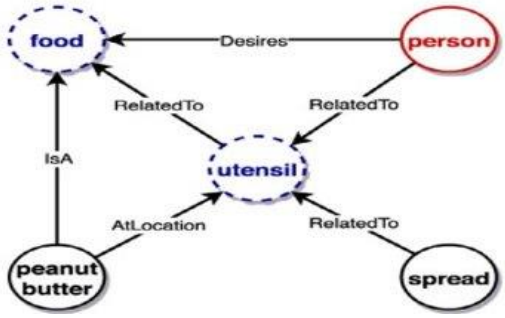


Fig2:

The model was able to arrive at the correct interpretation even when the target noun and verb were not the top-1 prediction.

The label peanut butter was not in the top 5 labels for the noun initially. Still, the inference process considered it as a possible object based on the presence of the verb spread and helped arrive at the final interpretation, which captures the semantics of the scene beyond semantic correspondences between nouns.

#Data Sets:

It uses the GTEA Gaze and the GTEA Gaze+ as the test environment for open-world object, action, and activity recognition in egocentric videos. The two datasets consist of several video sequences on meal preparation tasks by different subjects and ground-truth annotations of their gaze positions. The activity annotations include an action (verb) and the corresponding object (noun). GTEA Gaze contains 10 different verbs and 38 different nouns, while GTEA Gaze+ contains 15 verbs and 27 nouns. We also test our approach's generalisation capability to scenes beyond egocentric videos for generalised object detection. We use a subset of Open Images with 10 classes called the Open Images OW10 dataset with 3095 images and 5686 bounding box annotations. Each of these 10 classes can be found in the GTEA Gaze dataset. It allows us to evaluate our model beyond egocentric videos where the gaze positions isolate the object of interest. The goal is to expand the vocabulary beyond MS COCO without any supervision.

#Limitation:

- The limitation of this model is its dependency on external knowledge source, which is not available locally, hence increasing the processing time and lack of full control of the model.

Incorporating Visual Grounding in GCN for Zero-Shot Learning of Human Object Interaction Actions

#Title: Incorporating Visual Grounding In GCN For Zero-shot Learning Of Human Object Interaction Actions

#Author: *Chinmaya Devaraj, Cornelia Fermuller, and Yiannis Aloimonos University of Maryland*

Introduction:

- The literature proposes a GCN(Graph Convolution Network) based zero shot learning approach.
- Here, visual grounding is incorporated in GCN for zero-shot learning of human object interaction actions.
- A novel method that enhances the performance of GCNs in recognizing human manipulation actions by visually grounding the external knowledge graph, has been put forward by the authors.

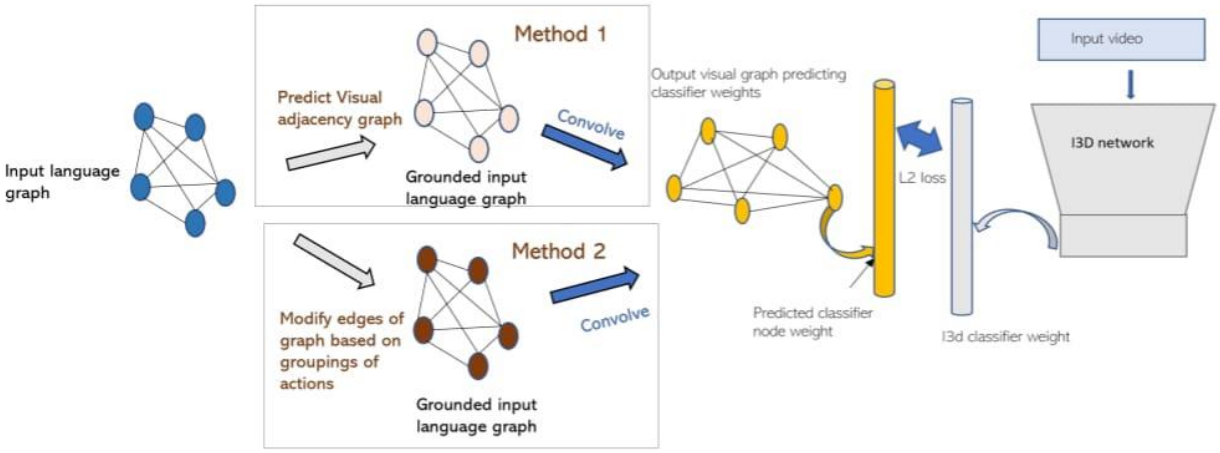
Datasets:

- The authors of this PDF file used two datasets for their experiments: the EPIC-Kitchens-55 dataset and the Charades dataset.
- The EPIC-Kitchens-55 dataset is a highly unbalanced dataset when it comes to the number of videos per class, so the authors sampled the dataset such that approximately 50 samples per class were present in the training set.
- The Charades dataset is a dataset of crowdsourced videos of activities in people's homes. It has 157 action classes, and for the experiments in this PDF file, the authors used splits of 79 training and 78 test classes.
- The authors formed groups of actions based on the object involved, and since the object manipulated in each video is provided, there was no need for further annotation.

Architectural review:

- The model uses a graph convolutional network (GCN) to learn the representations of the actions and objects in the videos.
- The GCN takes as input a knowledge graph that encodes the relationships between the actions and objects.
- The authors used two different concepts to define shared concepts to group actions: inhibitory and excitatory feedback for message passing in graph convolutions.

- The model uses a visual graph that encodes the visual features of the objects in the videos.
- The authors proposed two methods to visually ground the language graph: modifying the input language graph by changing the weights of the edges to reflect the visual semantics of the visual graph, and integrating the visual graph in the GCN architecture. The model uses a zero-shot learning setting, where the test classes are not seen during training.
- The authors formed groups of actions based on the object involved, and since the object manipulated in each video is provided, there was no need for further annotation.



Results:

- The model was evaluated on two datasets: the EPIC-Kitchens-55 dataset and the Charades dataset.
- The proposed method outperforms the state-of-the-art methods on both datasets.
- Specifically, on the Charades dataset, the proposed method achieves a mean class accuracy of 38.3%, which is a significant improvement over the previous state-of-the-art method that achieved a mean class accuracy of 28.9%.
- On the EPIC-Kitchens-55 dataset, the proposed method achieves a mean class accuracy of 23.5%, which is also a significant improvement over the previous state-of-the-art method that achieved a mean class accuracy of 18.9%.
- The authors also conducted ablation studies to show the effectiveness of the proposed methods for visually grounding the language graph.
- The proposed method can be used for other tasks such as action localization and action segmentation.
- The proposed method can be extended to other modalities such as audio and text.

- The proposed method can be used for other datasets and tasks beyond human manipulation actions.

#Limitations:

- The proposed method relies on the availability of a knowledge graph that encodes the relationships between the actions and objects, which may not be available for all datasets.
- The proposed method may not generalize well to other modalities such as audio and text, and may require additional modifications to be effective in those modalities.
- Training the transformer function F is hard as limited data points are present for training and it is a complex function to learn. Learning a better F depends on accuracy of ALanguagetrain(adjacency matrix of language graphs) and Avisualtrain(adjacency matrix of visual graphs). If both of these are unreliable, then the overall accuracy of the system will be poor.
- Finally, the proposed method may not be suitable for real-time applications due to its computational complexity.

#Noble Recommendation:

- The intuition is to make ZSL more knowledge driven by improving and expanding the knowledge base.
- We could use a common sense knowledge graph for knowledge base giving the model a sense of reasoning. But the hurdle is to create a good enough locally available knowledge graph, and adding more layers to process the result from the graph.
- We could use more than one external knowledge base to increase the knowledge of our model but the problem will be processing these external knowledge.