Deep Structured Models For Group Activity Recognition

Zhiwei Deng[1]
zhiweid@sfu.ca
Mengyao Zhai[1]
mzhai@sfu.ca
Lei Chen[1]
chenleic@sfu.ca
Yuhao Liu[1]
yla305@sfu.ca
Srikanth Muralidharan[1]
smuralid@sfu.ca
Mehrsan Javan Roshtkhari[2]
mehrsan@sportlogiq.com
Greg Mori[1]
mori@cs.sfu.ca

[1] School of Computing Science
Simon Fraser University
Burnaby, BC, Canada

[2] SportLogiq Inc.
Montreal, QC, Canada

Event understanding in videos is a key element of computer vision systems in the context of visual surveillance, human-computer interaction, sports interpretation, and video search and retrieval. The standard and yet state-of-the-art pipeline for activity recognition and interaction description consists of extracting hand-crafted local feature descriptors either densely or at a sparse set of interest points in the context of a Bag of Words model [4]. In recent years, it has been shown that deep learning techniques can achieve state-of-the-art results for a variety of computer vision tasks including action recognition [3]. However, modeling complex structures and high-level semantic meanings of activities in videos remains a challenge for deep learning methods. Graphical models provide a natural way to hierarchically model group activities and capture the semantic dependencies between group and individual activities [2]. A graphical model defines a joint distribution over states of a set of nodes and the relations between them. Belief propagation is often adopted as a way to infer states or probabilities of variables. However, most of the previous approaches applied a graphical model as a separate phase after using a deep learning model.

In this paper, our main goal is to address the problem of *group activity understanding* and *scene classification* in complex surveillance videos using a deep learning framework. More specifically, we are focused on learning individual activities and describing the scene simultaneously while considering the pair-wise interactions between individuals and their global relationship in the scene. This is achieved by combining a Convolutional Neural Network (CNN) with a probabilistic graphical model as additional layers in a deep neural network architecture into a unified learning framework. The structured deep neural network considers dependencies between individual actions, body poses, and group activities and mimics the message passing process for conducting the inference. Each combination of states is represented as a factor layer. Each neuron represents one type of combination between states and the parameters which are shared across the same type of neurons (weight sharing). We use a sparsely connected factor layer to mimic the message passing from variable nodes to the factor nodes. A layer with reverse connections is used to represent propagation from factor nodes to the variable nodes. The two layers together are considered as one pass iteration. Experimental results show that employing multi-step message passing procedure increases the accuracy of individual and group activity recognition. The multi-step message passing can also be considered as a label refinement procedure through information sharing. In our experimental setup, we use three components to model the interactions between people in the scene: unary terms, scene-action-pose components and global-pose components. Unary terms come from output scores of fine-tuned convolutional neural networks for individual human poses, activities, and the scene label. The scene-action-pose component captures the dependencies between states of human poses, human actions and scene for a group of people. A global-pose component models frequent patterns of different poses for all individuals in the scene. After each step of message passing, the output scores for each activity and scene labels are normalized to probability distribution by a softmax function.

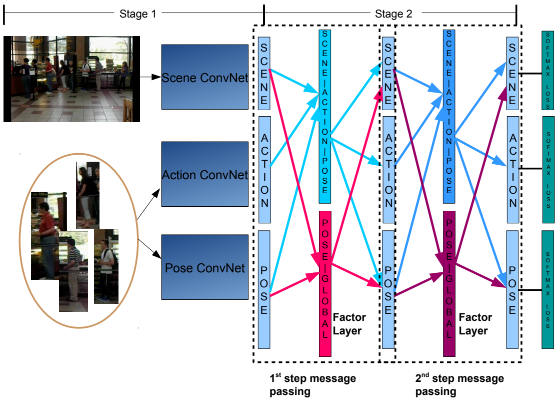

Figure 1: A schematic overview of our message passing CNN framework. Given an image and the detected bounding boxes around each person, our model predicts scores for individual actions and the group activity. The predicted labels are then refined by applying a belief propagation-like neural network. This network considers the dependencies between individual actions, body poses, and the group activity. The model learns the message passing parameters and performs inference and learning in unified framework using back-propagation.

We evaluated both the deep structured model and the features learned in the factor layers on two challenging group activity datasets: the Collective Activity [1] and the Nursing Home dataset. On both datasets, we observe a significant improvement in the scene classification accuracy by employing the message passing process. Considering the accuracy of a fine-tuned CNN classifier for the scene classification on the collective activity dataset as a baseline which is equal to 48%; the first message passing step reaches 73.6% and the second one improves the accuracy to 78.1%. In addition, by utilizing the learned high-level semantic features and a kernelized SVM as a classifier, we achieved 80.6% accuracy which is comparable to the state-of-the-art. A similar improvement in the accuracy is observed for the Nursing Home dataset, an increase from 69% to 82.5% and 84.0% by applying one and two step message passing.

[1] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *International Conference on Computer Vision Workshops on Visual Surveillance*, pages 1282–1289, 2009.

[2] Tian Lan, Wang Yang, Yang Weilong, and Greg Mori. Beyond actions: Discriminative models for contextual group activities. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[3] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 568–576. Curran Associates, Inc., 2014.

[4] Heng Wang and C. Schmid. Action recognition with improved trajectories. In *International Conference on Computer Vision (ICCV)*, pages 3551–3558, 2013.