CS231A - Group Activity Recognition

Jayesh K. Gupta
Stanford University
jkg@cs.stanford.edu

Abstract

Group activity dynamics can be inferred from individual activity of each person in the scene. We build a deep learning based hierarchical model that learns to capture these dynamics using LSTM model. We train an LSTM model to first predict individual activity from CNN features and then pool over all individuals in a frame to train another LSTM to predict the group activity label. We evaluate this model on Collective Activity Dataset.

1. Introduction

For an autonomous agent to function properly in the real world, it needs to understand the social rules that humans implicitly follow. Given an image of a group of people, there are multiple characterisitics that can be used to describe the scene. One of them is to identify the activity that is being performed. With the current popularity of Deep-CNN architectures, one naive way to approach this problem would be to collect a large dataset of images of groups of people performing some activity; each image labeled with the group activity and train a neural network to predict this label. However, this approach fails to capture the spatio-temporal relation between the indiviuduals performing the given activity. A better approach would be to identify the individuals in the scene, analyze their appearance and reason about the relations between them. A large volume of literature explores this idea [2] [3] [14] [15]. However, most of these rely on hand-crafted features and hence are limited by their limited representation abilities. On the other hand deep-learning based representations trained on Imagenet [13], [16], [17] [8] have yielded state-of-the-art results in recognition tasks.

Taking both these views into account it is imperative that we use a hierarchical approach to identify individuals in a scene using deep conv-nets and then reason about the spatio-temporal relationship between these features. Similar approaches have also been popular for the segmentation task [18]. Ibrahim et.el. [10] proposed a hierarchical temporal algorithm in the same spirit to identify the activity being performed by the group.

2. Related Work

Human activity recognition has been an area of active research in computer vision. We discuss some related work in the field and recent work in deep learning.

Group Activity Recognition: Most works use hand-crafted features fed to structured models representing the structural information present between individuals in space and time. Lan et al. [14] represent the hierarchical features from lower level person information to higher level group interactions using an adaptive latent structure learning algorithm. Choi and Savarese [3] propose a joint framework to unify tracking multiple people using individual actions, interactions and collective activities. All these suffer from representation issues due to use of hand-crafted features and linear models.

Deep Learning: Deep convolutional networks have shown impressive performance over the last few years especially with the availability of large datasets as ImageNet [13]. These include tasks like image classification [13] and action recognition [16]. For space time inputs, recurrent neural networks, specifically LSTM [9] models have been popular. Donhaue et al. [5], stack an LSTM on top of CNN to handle sequential tasks as action recognition.

2.1. Contribution

We implement the hierarchical temporal architecture as proposed by Ibrahim et.al. [10], and replicate their results.

3. Approach

Three cues important for understanding what a group of people are doing:

- **Person level actions**: Individual actions are the building blocks towards determining the group actions.
- **Temporal dynamics of person's actions**: How individual activity changes over time is important to determine the group actions.

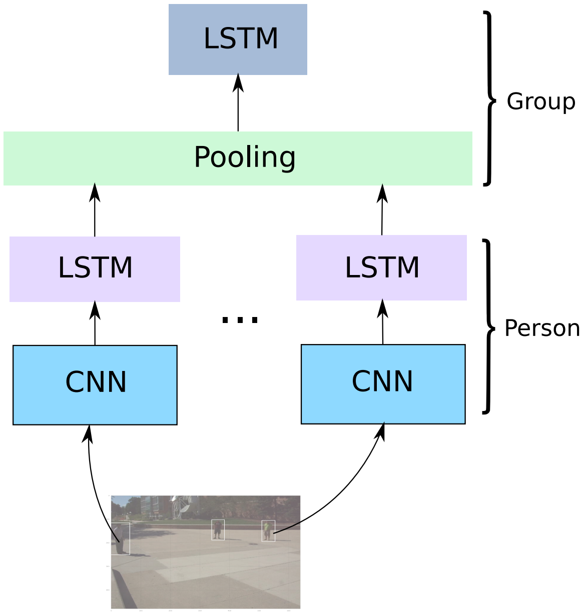

Figure 1. Proposed Solution Pipeline: Given tracklets of K-individuals, we feed each tracklet in a CNN, followed by a person LSTM layer to represent each person's action. We then pool over these temporal features in the scene. The output of pooling layer is fed to the second LSTM network to identify the whole group's activity.

- **Temporal evolution of group activities**: How the groups's movements are changing over time helps determine the group activity.

Most classic approaches model this problem as a structured prediction problem based on hand crafted features [14], [3]. This work instead takes inspiration from the success of hierarchical deep learning models to model the dynamics in a unified end-to-end framework.

Since group activities are inherently sequential in nature, Recurrent Neural Network (RNN) architectures are important to model them. We specifically use LSTM layers which are a type of RNN structure. The proposed model takes as input a sequence of images and first extracts individual person features from the scene using a CNN. These features are then transferred to an LSTM which predicts the individual activity being performed. Since we have a sequence of images, the LSTM learns to map this sequence to individual activity label. We next pool together the hidden states from these individual LSTMs and feed it to another LSTM that models the *temporal* dynamics of the group activity ans uses a softmax layer to predict the group activity label. See pipeline in Fig.1.

3.1. Implementation

We use the Tensorflow [1] library to implement our model. We initialize the CNN using a pre-trained AlexNet [13] [1]. This allows us to take a transfer learning approach [6] and train the first LSTM layer with the CNN in an end to end fashion. The hidden state of LSTM keeps track of an individual's behavior over the short period of time. For the first state, the output of the LSTM layer is passed to a softmax classification layer to predict the individual action class label.

For the next stage, we concatenate the fc7 layer features from the AlexNet with the LSTM hidden layer features for each person and feed it to a pooling layer. This pooling operation's output provides a representation for the action being performed by all people in the scene. These features will then be fed to the second LSTM layer followed by a softmax layer to predict the group activity label. We postulate that to perform this task, the second LSTM learns to directly model the **temporal dynamics of a group activity**.

Mathematically, the pooling layer can be expressed as [10]:

$$P_{tk} = x_{tk} \oplus h_{tk} \tag{1}$$

$$Z_t = P_{t1} \diamond P_{t2} \diamond \ldots \diamond P_{tk} \tag{2}$$

In this equation, h_{tk} is the first LSTM's output and x_{tk} represent the AlexNet fc7 features for the k-th person at time t. $\oplus$ concatenates these two features which are then max-pooled ($\diamond$) over to give us the frame's features at time t, Z_t. Finally we feed these frame level features to the second LSTM stage similar to the first one and learn the group level dynamics. Z_t, passed through a fully connected layer is input to the LSTM layer. The hidden state of LSTM layer carries the temporal information for the whole group dynamics which is fed to a softmax layer for classification. The network consists of 3000-node fully connected layer followed by a 9-timestep 500-node LSTM layer without the softmax layer.

3.2. Dataset

- **Collective Activity Dataset** [3]: This dataset contains 33 videos which results in 21470 frames. It contains six group activities: *Gathering*, *Talking*, *Dismissal*, *Walking together*, *Chasing* and *Queueing*. It also has three individual activity labels: *Standing*, *Walking* and *Running*. Although the dataset has some labels for pose and interaction features we do not use them. We do a 80-20 split for all frames.

3.3. Differences from the Original Paper

- We use Adam [12], to perform stochastic optimization using a fixed learning rate of 0.01, instead of the moment based SGD used in the paper.

[1] We use Caffe-Tensorflow to convert the Caffe [11] model to its Tensorflow equivalent.

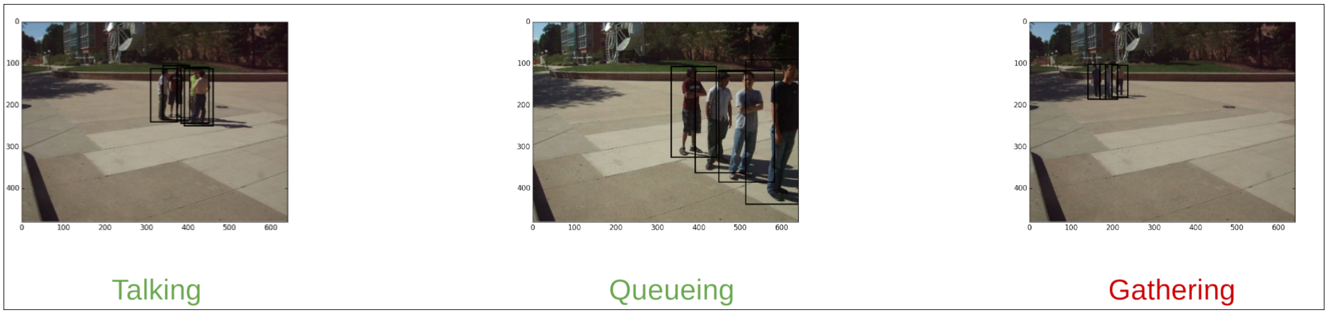

Figure 2. Predictions by the two staged hierarchical network. In the last image network predicts the *walking together* activity incorrectly as *gathering*.

4. Experiments

4.1. Image Baseline

We fine tune the AlexNet model to directly predict group activity label for each frame.

4.2. Person Baseline

We pool over fc7 features of each person and train a softmax classifier to predict the group activity label.

4.3. Fine-tuned Person Baseline

We fine tune AlexNet model for each person to predict individual activity and then pool over these fc7 features to predict the group activity label.

4.4. Temporal with Person Features

We extend the second baseline with LSTM. fc7 features are pooled over all individuals in a frame and fed to an LSTM model to predict group activity label.

4.5. Results

Table 1. Comparison with Baseline experiments

Experiment	Accuracy
Image classification	61.5
Person classification	60.2
Fine-tuned Person classification	64.6
Temporal with Person features	61.6
Two staged hierarchical	**78.6**

It is clear that our two staged model has improved the performance when compared to our baselines and temporal information helps in the task. Also identifying the relevant parts of the frame (people) helps over frame level features. Some example prediction in Fig.2.

Comparing our method with the state of the art we see that we show comparable performance to them. The given approach would be even more useful if we had a larger dataset.

Table 2. Comparison with State of Art

Method	Accuracy
Two staged hierarchical (this)	**78.6**
Contextual [14]	79.1
Deep structured [4]	80.6
Two staged hierarchical (original) [10]	81.5
Cardinality kernel [7]	83.4

5. Conclusion

We presented a deep structured architecture to solve the group activity recognition problem. This model consists of two stages: we first learn the individual person level action representation and combine these to recognize the group activity. We evaluated the model on Collective Activity Dataset. Results show that using the temporal information improves upon the baselines lacking the hierarchical structure. The source code is available at `https://gitlab.com/rejuvyesh/cs231a-project`.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015. Software available from tensorflow.org.

[2] M. R. Amer, P. Lei, and S. Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *Computer Vision–ECCV 2014*, pages 572–585. Springer, 2014.

[3] W. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1282–1289. IEEE, 2009.

[4] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. J. Roshtkhari, and G. Mori. Deep structured models for group activity recognition. *CoRR*, abs/1506.04191, 2015.

[5] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, abs/1310.1531, 2013.

[7] H. Hajimirsadeghi, A. Vahdat, W. Yan, and G. Mori. Visual recognition by counting instances: A multi-instance cardinality potential kernel. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] M. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. *CoRR*, abs/1511.06040, 2015.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] T. Lan, L. Sigal, and G. Mori. Social roles in hierarchical models for human activity recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1354–1361, June 2012.

[15] V. Ramanathan, B. Yao, and L. Fei-Fei. Social role discovery in human events. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9. IEEE, 2015.

[18] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision (ICCV)*, 2015.