From Abstraction

1. Track People
2. Track features from RNN
3. Learn time-varying attention weights to combine these features at each time-instant
4. Attended Features are then processed using another RNN for event detection/classification

Introduction

- Attention model results in better event recognition

Related Work

- Action recognition in videos
    - RNN (state of the art)
- Multi-person video analysis
    - Some models utilize layout of participants to identify group events
    - More recently, some models use context as a cue for recognizing interaction-based group activities
- Attention models
    - Paper uses attention to identify the most relevant person during different phases of the event
- Person detection and tracking
    - Person detection, the CNN-based multi box detector from C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In NIPS, 2013.
    - For person tracking, KLT tracker from C. J. Veenman, M. J. Reinders, and E. Backer. Resolving motion correspondence for densely moving points. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23(1):54–72, 2001.

Data set:

- 257 basketball games with 14K event annotations corresponding to 11 event classes
- Videos are 1.5 hours long in length
- Amazon Mechanical Turk task for annotations
- 212 training, 12 validation, 33 test videos. 4 second clips (annotation boundaries) subsampled to 6fps
- By excluding close-up shots of players, and instant replays, it results 11436 training, 856 validation and 2256 test clips, each of which has one of 11 labels
- position of the ball is also annotated
- Also the bounding boxes of all players in a subset of 9000 frames from the training videos.

Papers Method

- Feature Extraction
    - Each video-frame is represented by a 1024 dimensional feature vector f_t
    - Features got from Inception7 network(Last fully connected layers activation)
        - Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.
        - C. Szegedy et al. Scene classification with inception-7. http://lsun.cs.princeton.edu/slides/Christian.pdf, 2015.
    - In addition, p_{ti} 2805 dimensional feature vector for each person(i'th player bounding box in frame t)
- Event Classification
    - Goal: Given f_t and p_{ti}, classify the clip into one of 11 categories.
    - First, compute a global context feature for each frame, h_t^f, derived from a bidirectional LSTM applied to the frame level feature.
        - $h_t^f = BLSTM_{frame}(h_{t-1}^f, h_{t+1}^f, f_t)$
    - Next, use unidirectional LSTM to represent the state of the event at time t:
        - $h_t^e = LSTM(h_{t-1}^e, h_t^f, a_t)$
        - where a_t is a feature vector derived from the players
    - Loss: Squared-Hinge
        - $L = \frac{1}{2}\sum_{t=1}^{T}\sum_{k=1}^{K}\max\left(0, 1 - y_k w_k^T h_t^e\right)^2$
        - where y_k is 1 if the video belongs to class k, and is -1 otherwise
- Attention Models
    - 2 issues
        - Tracking across frames
        - Player attention depends on the state of the event
    - Attention model with tracking
        - KLT tracker combined with bipartite graph matching to perform data association
        - The latent representation of player i in frame t is given by the hidden stat h_{ti}^p of the BLSTM across the player track:
        - $h_{ti}^p = BLSTM_{track}(h_{t-1,i}^p, h_{t+1,i}^p, p_{ti})$
        - $a_t^{track} = \sum_{i=1}^{N_t}\gamma_{ti}^{track} h_{ti}^p$
        - $\gamma_{ti}^{track} = softmax\left(\phi\left(h_t^f, h_{ti}^p, h_{t-1}^e\right); \tau\right)$
        - N_t is the number of detections in frame t, $\phi()$ is multi layer perceptron
    - Attention model without tracking
        - $a_t^{notrack} = \sum_{i=1}^{N_t}\gamma_{ti}^{notrack} p_{ti}$
        - $\gamma_{ti}^{notrack} = softmax\left(\phi\left(h_t^f, p_{ti}, h_{t-1}^e\right); \tau\right)$

Experimental Evaluation

- Implementation details
    - hidden state dimension of 256 for all the LSTM, BLSTM RNNs
    - ReLU non-linearity
    - 256 dimensions for embedding the player features and frame features before feeding to the RNNs
    - 32×32 bins with spatial pyramid pooling for the player location feature
    - The models were trained on a cluster of 20 GPUs for 100k iterations over one day.
- Event classification
- Event detection
    - 4 second window, with stride 2 seconds
    - All windows which do not overlap more than 1 second with any of the 11 annotated events are treated as negatives

Some Cited Papers:

- About Attention:
    - D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
    - K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044, 2015.
    - L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. stat, 1050:25, 2015.
- State of the Art Event Recognition and Caption-Generation
    - J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrentconvolutional networks for visual recognition and description. arXiv preprint arXiv:1411.4389, 2014.
    - J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. arXiv preprint arXiv:1503.08909, 2015.
    - N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. arXiv:1502.04681, 2015.
    - L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. stat, 1050:25, 2015.