Human Activity as a Sequence to Sequences Representation

Patrick Emaase[1], Beom-Jin Lee[1], Byoung-Tak Zhang[1,2,3]
[1]School of Computer Sci. & Eng., [2]Brain Science Program, [3]Cognitive Science Program,
Seoul National University

Abstract

Sequence to sequence learning epitomizes the candid advancement in supervised learning. Currently most applications suffice single modality learning unlike to real life scenarios that integrates and combines multimodalities coherently. Sequence to sequence multimodality joint representations unearths potential to learn new facets of objectives in human activity recognition, hereby referred to as intentions. In this paper we endeavor to naturally express activities as sequences (input hierarchies) in order to learn intentions (feature map) as basis for learning personalized behavior in a robust manner. We propose a novel framework CR-CNN (Coherent Recurrent Convolutional Neural Network) to learn intention and smoothen outlier. We show empirical evidence to support our claim regarding the ability to use sequences of activities to learn, infer and predict intention thereby enabling personalization of a behavior.

1. Introduction

Transferred personalized life-log (hereby referred to as behavior) has been identified as indispensable requirement for human-like intelligent robots. Currently various non-obtrusive systems and devices collect human activity ardently on daily basis but limited techniques are able to learn, infer and predict human behavior from a set activities performed. Recently, sequence to sequence (seq2seq) learning (Sutskever, et al., 2014)

Accurate and automatic recognition of human's activities

Human action, captured through RGBD camera, can be described as a purposeful single period of unique human motion pattern that has unique predetermined intention. Hence are diverse, invariant and complex. It therefore described sequence of consecutive human body poses with specific predictive objective often referred to as intention. The sequences of moves to achieve action is occur in a specific configurable discrete manner that can provide insightful information to learn, infer and regenerate human activity map automatically. Coupled with Most existing work relies on heuristics hand-crafted features [Yang, J. B., et al., 2015] as a source of features to learning hidden states. The sequences of motions often provide input features that can be learn human activity patterns however they are heuristic and task dependent [Yang, J. B., et al., 2015].

Intention and predictions are intertwined. Often inherent intention is used to predict human behavior. People tend to desire a human-like companion capable of recognizing and responding to their respective tasks thereby support their ambient living. This will motivate and encourage them with "feel-good" phenomena while executing activities. For many applications it is important to be able to detect what a human is currently doing as well as anticipate what she is going to do next and how. There are many applications areas including monitoring and surveillance, but we need the latter for applications that require reactive responses.

In this paper, our goal is to use anticipation for predicting future activities as well as improving detection (of past activities). There has been a significant amount of work in detecting human activities from 2D RGB videos from inertial/location sensors, and more recently from RGB-D videos. The primary approach in these works is to first convert the input stream into a spatio-temporal representation, and then to infer labels over the inputs. These works use different types of information, namely skeleton data from RGB, Location-Objects-Activity, interaction with objects, object shape and appearance features. However, these methods can be used only to predict the labeling of an observed activity and cannot be used to predict intention as without modification as illustrated in figure 1.

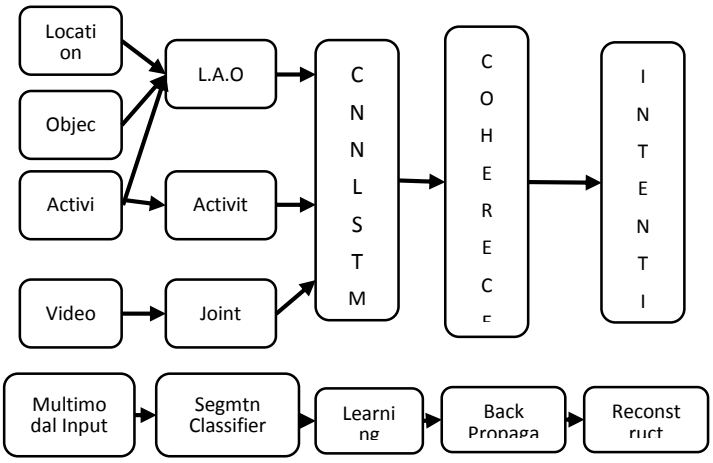

Fig. 1: Experimental set-up activities are pipelined to collect intentions. L.A.O – Locaion-Activity-Object association (input to CNN). Intention is the output of smoothed CNN + LSTM Learning through backprogropagation

Our goal is to predict the future activities as well as the details of how a human is going to perform them in short-term (e.g., 1-10 seconds). For example, if a robot has seen a child move his hand towards a book, it is possible he would move the book to a few

potential places such as his hands, to a bookshelf or just move it to a different location on the table (spatio-activity linkage). If a robot can anticipate this, then it would rather not start executing the most probable activity learnt and scheduled previously. This will avoid delays and proactively saving activity time. This scenario happens in several other settings for example in elderly care where a robot can anticipate a person's intention, plans an execution plan and executes activity in a pipelined manner.

The main contributions of this paper are

- While most previous works consider activity detection, we consider learning, inferencing and anticipation.
- We consider rich contextual relations based on object affordances in RGB-D skeleton videos.
- We propose Deep Hierarchical Hypernetwork Activity Map, where each particle represents a CRF.
- We consider joint temporal segmentation and labeling using our approach.

2. Deep Learning

To explore the intention we propose to use restricted Boltzmann machine (RBM) which has visible input units (motion sequences) $v \epsilon\{0,1\}^D$, which is connected to other layer of hidden stochastic units $h \epsilon\{0,1\}^F$as shown in figure 2.

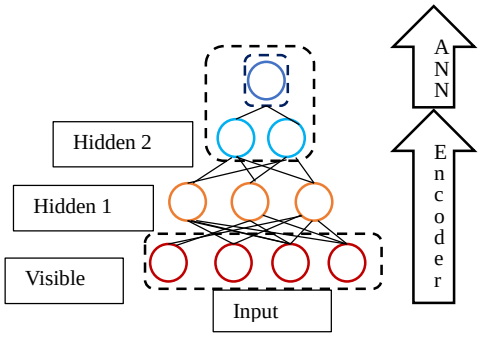

Figure 2: Restricted Boltzmann Machine

The distribution of states {v,h} of an RBM for learning intentions can be illustrated as follows:

$$E(v,h|\theta) = -\sum_{i=1}^{D}\sum_{j=1}^{F} w_{ij} v_i h_i - \sum_{i}^{D} b_i v_i - \sum_{i}^{D} a_j h_j \quad (1)$$

where **W** represents the visible-to-hidden weights matrix consisting of weights $\boldsymbol{w_{ij}}$of connections between the neurons v_i and h_j, **b** represents a visible bias vector and **a** represent a hidden bias vector. $\boldsymbol{\theta}$**= {W,a,b}** represents a set of all parameters.

The conditional distribution of the hidden vector **h** given be visible human activity sequence segment can be derived from (1) using $p(x|\theta) = \exp[-E(x|\theta)]\,/Z(\theta)$ as shown in (2)

$$p(h_i = 1|v) = g(\textstyle\sum_i w_{ij} v_i + a_i) \quad (2)$$

$$p(h_j = 1|h) = g(\textstyle\sum_j w_{ij} h_j + b_i) \quad (3)$$

Where $g(x) = 1/(1 + \exp(-x))$ is a logistic function and $Z(\theta) = \sum_i \exp[-E(x|\theta)]$ is a normalizing constant.

Contrastive Divergence Learning

The maximum likelihood estimate of parameter $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\}$ of an RBM can be obtained iteratively using a gradient-based rule with learning rate η as shown step-by-step procedure below:

$$w_{ij} \leftarrow w_{ij} + \eta[\{v_i h_j\}data - \{v_i h_j\}model] \quad (4)$$

$$b_i \leftarrow b_i + \eta[\{v_i\}data - \{v_i\}model] \quad (5)$$

$$a_j \leftarrow a_j + \eta[\{h_j\}data - \{h_j\}model] \quad (6)$$

Where {.}data is the expectation over the data distribution or positive phase distribution, $P\left(h|\{v^{(t)}\},\theta\right)$. can also be represented as $\prec.\succ$. {.}model denotes the expectation over the model or negative phase distribution, $P(v,h|\theta)$.

Learning Sequences from Trajectories

Once the location is sampled from the affordance heatmap, we generate a set of possible trajectories in which the object can be moved from its current location to the predicted target location. We parameterize the cubic equations, in particular Bézier curves, to generate human hand like motions

$$B(x) = (1-x)^3 L_o + 3(1-x)^2 x L_1 + 3(1-x)x^2 L_2 + x^3 L_3, \quad x\epsilon[0,1] \quad (8)$$

3. Sequence to Sequence Learning

Dynamic model provide appropriate solutions to sequential learning problem to learn physiological response to multimedia stimuli. The common ones are sliding window, recurrent sliding windows, input-output Markov models, conditional fields and graph transformer network.

3.1 Sliding Window Method

The sliding window algorithm converts the sequential multisensory input into the classical supervised learning problem. It uses a window classier that maps an input window of width w into an individual output value y. For our case, input we are features collected from multimedia stimuli exposure. Sliding window is "half-width" of the window. For our case the width of the window is five seconds with an overlap of 2.5 seconds. A sliding window is therefore defined by a fixed number of recently generated data elements which is the target of data mining. The window classifier is trained by converting each training examples into windows and then applying a standard supervised learning algorithm through regression models.

Recurrent Sliding Windows

This is an improvement of sliding window method where the predicted value is fed as an input to predict subsequent value. It recursively improves the predictive quality of the systems. This has been applied in various dynamic modeling environments to model dynamic responses. It has the ability to improve the quality of results.

Conditional Random Fields

They were introduced to try to overcome the label bias problem. In the CRF, the relationship among adjacent pairs is modeled as a Markov Random Field conditioned on the x inputs. In other words, the way in which the adjacent values influence each other is determined by the input features.

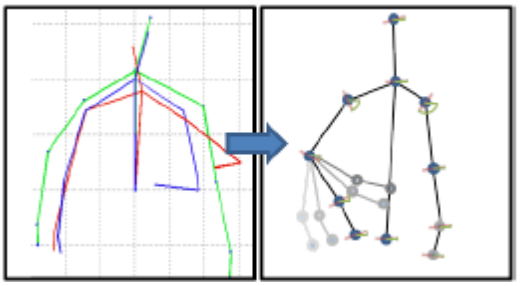

Figure 3: Joint Augmented Context

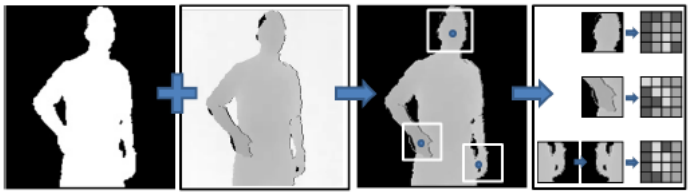

Figure 4: Extracted Depth motion Histograms for learning

4. Experimental Results

Seventeen neurologically healthy participants (8 males and 7 females), aged between 20 and 29 (mean 24.04, standard deviation 2.29), undergraduate or graduate students, participated in the experiment. All participants could clearly understand the stimuli mode of communication. This was supported by their respective response to stimuli.

Experimental Setup

We collect data to learn this model in a pipelined manner (Fig. 1). To achieve this, the mobile robot is follows a 'child' recognizing and tracking the activity. It also records related objects in the room. For example, vision data will be collected with a camera and other child's status will be collected as a ground truth by human annotator. The input and output schedule data is handmade script, which describes the robot's entire behaviors.

Discussion and recommendation

Results from figures 3 and 4 show that prediction is possible and feasible. This is confirmed by the respective values of CC, MAE and RMSE. There is an increasing body of evidence supporting the application of modeling of intention and responses since the use of wearable and indeed ubiquitous devices. The results of the experiments presented above indicate that intentions are predictive hence can be generalized and learned autonomously. Additional empirical work is needed, in order to establish dynamic modeling of cognitive response to. Further studies to elaborate dynamic physiological responses in a dynamic environment need to be explored. Wearables are on the rise, hence there is a need to understand spatio-temporal responses through further research.

Overall, we found that there are different intentions have unique learnable patterns. Machine learning provides robust method model intentions automatically.

Acknowledgement

This work was partly supported by the Institute for Information & Communications Technology Promotion (R0126-16-1072-SW.StarLab), Korea Evaluation Institute of Industrial Technology (10044009-HRI.MESSI, 10060086-RISF), and Agency for Defense Development (UD130070ID-BMRR) grant funded by the Korea government (MSIP, DAPA).

References

CogDIEM: 인지컴퓨팅 연구를 위한 멀티미디어 시청자의 암묵적 감정 반응 데이터베이스, 온경운, 김병희, 김경민, 곽동현, 박태서, 장병탁, *한국정보과학회 동계학술발표회 논문집*, pp. 571-573, 2014.

Yang, J.B.. Nguyen, M. N., San P. P., Li, X.L and Krishnaswamy (2015). Deep Convolutional Neural Networks On Multichannel Time Series For Human Activity Recognition, IJCAI 2015 pp. 3995 - 4001

Jiang, X,G., Xu, Baohan & Xue Xiangyang (2014). Predicting Emotions in User-Generated Videos. Association for the Advancement of Artificial Intelligence. Pp. 1-13

McDuff, D., Kaliouby, R., Cohn, J. and Picard R.W. (2014) Predicting Ad Liking and Purchase Intent: Large-scale Analysis of Facial Responses to Ads, IEEE Transactions on Affective Computing.

On, K., Kim, B.-H., Kim, K., Kwak, D., Park, T.-S., and Zhang B.-T. (2014) CogDIEM: Database of Implicit Emotional Responses to Multimedia for Cognitive Computing, *KIISE Winter Conference*, pp. 456-458.

Park, T.-S. Kim, B.-H. and Zhang B.-T. (2014) A viewer preference model based on physiological feedback, *Journal of the Korean Institute of Intelligent Systems*, Vol 24. No. 3, pp. 316-322.

Aipperspach, R., Cohen, E., and Canny, J., "Modeling human behavior from simple sensors in the home," Pervasive Computing, pp. 337–348, 2006.

Le Roux N. and Bengio, Y., "Deep belief networks are compact universal approximators," Neural Computation, vol. 22, no. 8, pp. 2192–2207, 2010.

Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. Sequence to sequence learning with neural networks. In Proc. NIPS, 2014

Bengio, Y., Lamblin, P., D. Popovici, and Larochelle H., "Greedy layerwise training of deep networks," in Advances in NIPS, vol. 19. MIT; 1998, 2007, pp. 153–160.