Action Recognition by Hierarchical Sequence Summarization

Yufeng Jiang

Abstract

Recent progress has shown that learning from hierarchical feature representations leads to improvements in various computer vision tasks. Motivated by the observation that human activity data contains information at various temporal resoluetions, authors present a hierarchical sequence summarization approach for action recognition that learns multiple layers of discriminative feature representations at different temporal granularities. They build up a hierarchy dynamically and recursively by alternating sequence learning and sequence summarization. For sequence learning they use CRFs with latent variables to learn hidden spatio-temporal dynamics; for seqnence summarization they group observations that have similar semantic meaning in the latent space. For each layer they learn an abstract feature representation through non-linear gate funcitons.

1. Introduction

Recent progress has shown that learning from hierarchical feature representations leads to significant improvements in various computer vision tasks, including spatial pyramids of image pathches in object detection [2], higher order potentials in object segmentation [1], and the deep learning with multiple hidden layers [3, 5]. Although there is much differnece in algorithmic details, these approaches share the common goal of learning from hierarchical feature representations in order to capture high-level concepts that are otherwise difficult to express with a single reperesentation approach.

Action recognition is one particular area that can benefit from such representations, because human activity data contains information at varitous spatio-temporal resolutions. People may perform gestures slowly to enphasize a point, but more rapidly on unintentional movements or meaningless geatures. The resulting data stream will have many similar obsevations with occasional and irregualr changes. As a result, capturing discriminative information from a single temporal representation may prove to be difficult. Section 2 details their Hierarchical Sequence Summarization (HSS) model.

2. Hierarchical Sequence Summarization

Authors propose to capture complex spatio-temporal dynamics in human activity data by learning from a hierarchical sequence summary representation. Intuituvely, each layer in the hiercrchy is a temporally coarser-grained summary of the sequnce from the preceding layer.

Their approach builds the hierarchy by alternating sequence learning and sequence summarization. They define their notation in Section 2.1, describe sequence learning in Section 2.2.

2.1. Notation

Input to their model is a time-ordered sequence $\mathbf{x} = [\mathbf{x}_1; \ldots; \mathbf{x}_T]$ of length T; each per-frame observation $\mathbf{x}_t \in \mathbb{R}^D$ is of dimension D and can be any type of action feature. Each sequence is labeled y from a finite alphabet set, $y \in \mathcal{Y}$.

They denote a sequence summary at the l-th layer in the hierarchy by $\mathbf{x}^l = [\mathbf{x}_1^l]; \ldots; \mathbf{x}_Y^l$. A super observation $\mathbf{x}_t^l$ is a group of observations from the preceding layer, and they define $c(\mathbf{x}_t^l)$ as a reference operator of $\mathbf{x}_t^l$ that returns the group of observations; for $l = 1$ they set $c(\mathbf{x}_t^l) = \mathbf{x}_t$.

2.2. Sequence Learning

Following [4], they use CRFs with latent variables to capture hidden dynamics in each layer in the hierarchy. Using a set of latent variables $\mathbf{h} \in \mathcal{H}$, the conditional probability distribution is defined as [6]

$$p(y|\mathbf{x}; \mathbf{w}) = \frac{1}{Z(\mathbf{x}; \mathbf{w})} \sum_h expF(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) \qquad (1)$$

where $\mathbf{w}$ is a model parameter vector, $F(\cdot)$ is a generic feature function, and $Z(\mathbf{x}; \mathbf{w}) = \sum_{y', \mathbf{h}, \mathbf{x}; \mathbf{w}}$ is a normalization term.

Feature Function: They define the featuer function as [6]

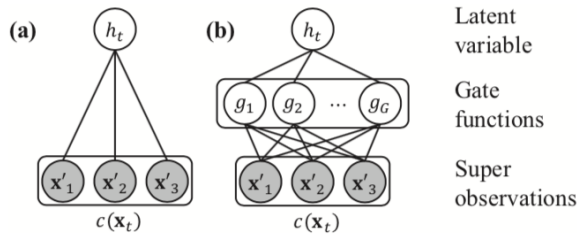

Figure 1. **Illustration of our super observation feature function.** (a) Observation feature function similar to Quattoni *et al.* [4], (b) our approach uses an additional set of gate functions to learn an abstract feature representation of super observations.

$$F(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) = \sum_t f^1(\mathbf{h}, \mathbf{x}, t; \mathbf{w}) + \sum_t f^2(y, \mathbf{h}, t; \mathbf{w}) + \sum_t f^3(y, \mathbf{h}, t, t+1; \mathbf{w}) \quad (2)$$

Their definition of feature function is different from that of [4] to accommondate the hierarchical nature of their approach. Specifically, they define the super observation feature function that is different from [4].

Let $\mathbb{1}[\cdot]$ be an indicator function, and $y' \in \mathcal{Y}$ and $(h', h'') \in \mathcal{H}$ be the assignments to the label and latent variables. The second and the third terms in Equation 2 are the same as those defined in [4], *i.e.* the label feature function $f^2(\cdot) = w_{y,h}\mathbb{1}[y = y']\mathbb{1}[h_t = h']$ and the transition feature function $f^3(\cdot) = w_{y,h,h}\mathbb{1}[y = y']\mathbb{1}[h_t = h']\mathbb{1}[h_{t+1} = h'']$.

Their super observation feature function incorporates a set of non-linear gate functions G, as used in neural networks, to learn an abstract feature representation of super observations (see Figure 1). Let $\psi_g(\mathbf{x}, t; \mathbf{w})$ be a function that computes, using a gate function $g(\cdot)$, an average of gated output values from each observation contained in a super observation $\mathbf{x}' \in c(\mathbf{x}_t)$ [6],

$$\psi_g(\mathbf{x}, t; \mathbf{w}) = \frac{1}{|c(\mathbf{x}_t)|} \sum_{\mathbf{x}' \in c(\mathbf{x}_t)} g(\sum_g w_{g,d} x'_d) \quad (3)$$

They adopt the popular logistic function as their gate function. $g(z) = 1/(1 + exp(-z))$, which has been shown to perform well in various tasks [1]. They define their super observation feature function as [6]

$$f^1(\mathbf{h}, \mathbf{x}, t; \mathbf{w}) = \sum_{g \in G} w_{g,h}\mathbb{1}[h_t = h']\psi_g(\mathbf{x}, t; \mathbf{w}). \quad (4)$$

where each $g \in G$ has the same form. The set of gate functions G creates an additional layer between latent variables and observations, and has a similar effect to that of the neural network. This feature function automatically learns an abstract representation of super observations and provides more discriminative information for capturing complex spatio-temporal patterns in human activity data.

References

[1] Y. Bengio. Learning deep architectures for AI. *FTML*, 2(1):1–127, 2009. 1, 2

[2] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 1

[3] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, 2011. 1

[4] A. Quattoni, S. B. Wang, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE TPAMI*, 29(10):1848–1853, 2007. 1, 2

[5] M. Ranzato, J. Susskind, V. Mnih, and G. E. Hinton. On deep generative models with applications to recognition. In *CVPR*, 2011. 1

[6] Y. Song and R. Davis. Action recognition by hierarchical sequence summarization. In *CVPR*, 2013. 1, 2