Signal Processing: Image Communication 61 (2018) 73–84

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Hierarchical Multi-scale Attention Networks for action recognition

Shiyang Yan [a,b,*], Jeremy S. Smith [a], Wenjin Lu [b], Bailing Zhang [b]

[a] Electrical Engineering and Electronic, University of Liverpool, Liverpool, United Kingdom
[b] Department of Computer Science and Software Engineering, Xi'an Jiaotong-liverpool University, SuZhou, JiangSu Province, China

ARTICLE INFO

Keywords:
Action recognition
Hierarchical multi-scale RNNs
Attention mechanism
Stochastic neurons

ABSTRACT

Recurrent Neural Networks (RNNs) have been widely used in natural language processing and computer vision. Amongst them, the Hierarchical Multi-scale RNN (HM-RNN), a recently proposed multi-scale hierarchical RNN, can automatically learn the hierarchical temporal structure from data. In this paper, we extend the work to solve the computer vision task of action recognition. However, in sequence-to-sequence models like RNN, it is normally very hard to discover the relationships between inputs and outputs given static inputs. As a solution, the attention mechanism can be applied to extract the relevant information from the inputs thus facilitating the modeling of the input–output relationships. Based on these considerations, we propose a novel attention network, namely Hierarchical Multi-scale Attention Network (HM-AN), by incorporating the attention mechanism into the HM-RNN and applying it to action recognition. A newly proposed gradient estimation method for stochastic neurons, namely Gumbel-softmax, is exploited to implement the temporal boundary detectors and the stochastic hard attention mechanism. To reduce the negative effect of the temperature sensitivity of the Gumbel-softmax, an adaptive temperature training method is applied to improve the system performance. The experimental results demonstrate the improved effect of HM-AN over LSTM with attention on the vision task. Through visualization of what has been learnt by the network, it can be observed that both the attention regions of the images and the hierarchical temporal structure can be captured by a HM-AN.

1. Introduction

Action recognition in videos is a fundamental task in computer vision. Recently, with the rapid development of deep learning, and in particular, deep convolutional neural networks (CNNs), a number of models [1–4] have been proposed for image recognition. However, for video-based action recognition, a model should accept inputs with variable length and generate the corresponding outputs. This special requirement makes the conventional CNN model that caters for a one-versus-all classification unsuitable.

For decades RNNs have been applied to sequential applications, often with good results. However, a significant limitation of the vanilla RNN models, which strictly integrate state information over time, is the vanishing gradient effect [5]: the ability to back propagate an error signal through a long-range temporal interval becomes increasingly impossible in practice. To mitigate this problem, a class of models with a long-range dependencies learning capability, called Long Short-Term Memory (LSTM), was introduced by Hochreiter and Schmidhuber [6]. Specifically, LSTM consists of memory cells, with each cell containing units to learn when to forget previous hidden states and when to update hidden states with new information.

Much sequential data often has a complex temporal structure which requires both hierarchical and multi-scale information to be modeled properly. In language modeling, a long sentence is often composed of many phrases which further can be decomposed into words. Meanwhile, in action recognition, an action category can be described by many sub-actions. For instance, 'long jump' contains 'running', 'jumping' and 'landing'. As stated in [7], a promising approach to model such hierarchical representation is the multi-scale RNN. One popular approach of implementing multi-scale RNNs is to treat the hierarchical timescales as pre-defined parameters. For example, Wang et al. [8] implemented a multi-scale architecture by building a multiple layers LSTM in which higher layers skip several time steps. In their paper, the skipped number of time steps is the parameter to be pre-defined. However, it is often impractical to pre-define such timescales without learning, which also leads to a poor generalization capability. Chung et al. [7] proposed a novel RNN structure, Hierarchical Multi-scale Recurrent

* Corresponding author at: Department of Computer Science and Software Engineering, Xi'an Jiaotong-liverpool University, SuZhou, JiangSu Province, China.
E-mail addresses: Shiyang.Yan@xjtlu.edu.cn (S. Yan), J.S.Smith@liverpool.ac.uk (J.S. Smith), Wenjin.Lu@xjtlu.edu.cn (W. Lu), Bailing.Zhang@xjtlu.edu.cn (B. Zhang).

https://doi.org/10.1016/j.image.2017.11.005
Received 12 August 2017; Received in revised form 21 November 2017; Accepted 22 November 2017
Available online 29 November 2017
0923-5965/

Neural Network (HM-RNN), to automatically learn time boundaries from data. These temporal boundaries are similar to rules described by discrete variables inside RNN cells. Normally, it is difficult to implement training algorithms for discrete variables. Popular approaches include unbiased estimator with the aid of REINFORCE [9]. In this paper, we re-implement the HM-RNN by applying the recently proposed Gumbel-sigmoid function [10,11] to realize the training of stochastic neurons due to its efficiency [12].

In the general RNN framework for sequence-to-sequence problems, the input information is treated uniformly without discrimination on the different parts. This will result in the fixed length of intermediate features and hence subsequent sub-optimal system performance. The practice is in sharp contrast to the way humans accomplish sequence processing tasks. Humans tend to selectively concentrate on a part of information and at the same time ignores other perceivable information. The mechanism of selectively focusing on relevant contents in the representation is called attention. The attention based RNN model in machine learning was successfully applied in natural language processing (NLP), and more specifically, in neural translation [13]. For many visual recognition tasks, different portions of an image or segments of a video have unequal importance, which should be selectively weighted with attention. Xu et al. [14] systematically analyzed stochastic hard attention and deterministic soft attention models and applied them in image captioning tasks, with improved results compared with other RNN-like algorithms. The hard attention mechanism requires a stochastic neuron which is hard to train using the conventional back propagation algorithm. They applied REINFORCE [9] as an estimator to implement hard attention for image captioning.

The REINFORCE is an unbiased gradient estimator for stochastic units, however, it is very complex to implement and often has high gradient variance during training [12]. In this paper, we study the applicability of Gumbel-softmax [10,11] in hard attention because Gumbel-softmax is an efficient way to estimate discrete units during the training of neural networks. To mitigate the problem of temperature sensitivity in Gumbel-softmax, we apply an adaptive temperature scheme [12] in which the temperature value is also learnt from the data. The experimental results verify that the adaptive temperature is a convenient way to avoid manual searching for the parameter. Additionally, we also test the deterministic soft attention [14,15] and stochastic hard attention implemented by REINFORCE-like algorithms [16,17,14] in action recognition. Combined with HM-RNN and the two types of attention models, we systematically evaluate the proposed Hierarchical Multi-scale Attention Networks (HM-AN) for action recognition in videos, with improved results.

Our main contributions can be summarized as follows:

- We propose a Hierarchical Multi-scale Attention Network (HM-AN) by implementing HM-RNN with Gumbel-sigmoid to realize the discrete boundary detectors.
- We also propose four methods of realizing an attention mechanism for action recognition in videos, with improved results over many baselines.
- By incorporating Gumbel-softmax and Gumbel-sigmoid into HM-RNN, we make the stochastic neurons in the networks end-to-end trainable by error back propagation.
- For the hard attention model based on Gumbel-softmax, we propose to use an adaptive temperature for the Gumbel-softmax, which generates much improved results over a constant temperature value.
- Through visualization of the learnt attention regions, the boundary detectors of HM-AN and the adaptive temperature values, we provide insights for further research.

2. Related works

2.1. Hierarchical RNNs

The modeling of hierarchical temporal information has long been an important topic in many research areas. The most notable model is LSTM proposed by Hochreiter and Schmidhuber [6]. LSTM employs the multi-scale updating concept, where the hidden units' update can be controlled by gating such as input gates or forget gates. This mechanism enables the LSTM to deal with long term dependencies in the temporal domain. Despite this advantage, the maximum time steps are limited to within a few hundred because of the leaky integration which makes the memory for long-term gradually diluted [7]. Actually, the maximum time steps in video processing is several dozen frames which makes the application of LSTM in video recognition very challenging.

To alleviate this problem, many researchers tried to build a hierarchical structure explicitly, for instance, Hierarchical Attention Networks (HAN) proposed in [8], which is implemented by skipping several time steps in the higher layers of the stacked multi-layer LSTMs. However, the number of time steps to be skipped is a pre-defined parameter. How to choose these parameters and why to choose a certain number are unclear.

More recent models like clockwork RNN [18] partitioned the hidden states of a RNN into several modules with different timescales assigned to them. The clockwork RNN is more computationally efficient than the standard RNN as the hidden states are updated only at the assigned time steps. However, finding the suitable timescales is challenging which makes the model less applicable.

To mitigate the problem, Chung et al. [7] proposed the Hierarchical Multiscale Recurrent Neural Network (HM-RNN). The HM-RNN is able to learn the temporal boundaries from data, which allows the RNN model to build a hierarchical structure and enables long-term dependencies automatically. However, the temporal boundaries are stochastic discrete variables which are very hard to train using the standard back propagation algorithm.

A popular approach to train the discrete neurons is the REINFORCE-like [19] algorithms. This is an unbiased estimator but often with high gradient variance [7]. The original HM-RNN applied a straight-through estimator [9] because of its efficiency and simplicity in implementation. Instead, in this paper, we applied the more recent Gumbel-sigmoid [10,11] to estimate the stochastic neurons. This is much more efficient than other approaches and achieved state-of-the-art performance among many other gradient estimators [10].

2.2. Attention mechanism

One important property of human perception is that we do not tend to process a whole scene, in its entirety, at once. Instead humans pay attention selectively on parts of the visual scene to acquire information where it is needed [16]. Different attention models have been proposed and applied in object recognition and machine translation. Mnih et al. [16] proposed an attention mechanism to represent static images, videos or as an agent that interacts with a dynamic visual environment. Also, Ba et al. [17] presented an attention-based model to recognize multiple objects in images. These two models are all with the aid of REINFORCE-like algorithms.

The soft attention model was proposed for the machine translation problem in NLP [13], and Xu et al. [14] extended it to image caption generation as the task is analogous to 'translating' an image into a sentence. Specifically, they built a stochastic hard attention model with the aid of REINFORCE and a deterministic soft attention model. The two attention mechanisms were applied to the image captioning task, with good results. Subsequently, Sharma et al. [15] built a similar model with soft attention applied to action recognition from videos.

There are a number of subsequent works on the attention mechanism. For instance, in [20], the attention model is utilized for video