Deep RNN Framework for Visual Sequential Applications

Abstract

The propose a new recurrent neural framework that can be staked deep effectively. Two novel designs:

- One is a new RNN module called Context Bradge Module (CBM) which splits the information flowing along the sequence and along depth.
- The other is the Overlap Coherence Training Scheme that reduces the training complexity for long visual sequential tasks.

> We are intended to build a deep RNN architecture that combines the merits of both RNN and DNN to extract more powerful temporal and representation features from visual sequential inputs.
>
> In this paper, we propose a new deep RNN architecture including two principle techniques, namely *Context Bridge Module* (CBM) and *Overlap Coherence Training Scheme*.
>
> - In CBM, we design two computing units taking charge of representation flow and temporal flow respectively, forcing these two flows relatively independent of each other with the aim of making them focus on representation and temporal information separately to ease the training process. After these two units, a merge unit is utilized to synthesize them.
> - Besides, the proposed *Overlap Coherence Training Scheme* aims at reducing the training cost of deep RNN.
>   Based on overlaps, we design *overlap coherence loss* that forces the detached clips to generate coherent results in order to strengthen

consistency of temporal information, which makes the model not a strict Markov process of order *n*, but the complexity is still reduced.

Results reveal that:
1) Deep RNN can enjoy accuracy gains from the greatly increased depth, substantially better than the shallow networks;
2) Our CBM is more suitable for stacking deep compared with other RNN structures like LSTM;
3) The overlap coherence training scheme can effectively make many computer vision problems with high-dimensional sequential inputs trainable on commonly-used computing devices.

2. Related Work

Methods for Visual Sequence Tasks
Visual sequence problems require models to extract hierarchical temporal and representation features simultaneously.
1) An inchoate approach is pooling the spatial representation features of every item in the sequence.
2) The 3D convolutional networks apprears, which treat temporal dimension equal to spatial dimension with its cubic convolutional kernel, while 3D convolutional networks need to consume large amount of computing resources.
3) RNN is designed to handle sequence problems, therefore it is a natural idea to utilize RNN to encode temporal information after obtaining spatial features.

Exploration on Deep RNN
In this paper, we focus on exploring appropriate deep structure for RNN model. There are many previous works trying to address this problem. The results show that stacked RNN has relatively better performance and more importantly, stacking method can synthesize temporal information in each layer to extract hierarchical temporal-spatial features instead of plain temporal, deep spatial features.
A new RNN structure called LSTMP is proposed to reduct the learning computational complexity.

3. Deep RNN Framework

> Context Bridge Module (CBM) designed to effectively capture temporal and representation information simultaneouly
> The Overlap Coherence Training Scheme to further simplify the training process

3.1 Context Bridge Module

- Temporal flow: temporal information from sequential inputs (e.g. a sequence of frames in a video) and oriented towards temporal depth
- Representation flow: representation information from each individual one (e.g. one frame of the sequence) and oriented towards structural depth.

Challenge A vertically stacked RNN architecture is hard to be co-adaptive to two flows simultaneously, resulting in ineffective and inefficient training. In most cases, people adopt shallow RNN which takes extracted CNN features as inputs, though it is not an end-to-end pipeline.

Our Archirecture How to capture two branches of information flows as independently as possible. **Specifically, for representation flow, we use a computing unit (e.g. CNN structure) to extract features of the individual input sample without recurrent operations, while temporal flow adopts a RNN structure.**

Temporal Dropout *Temporal Droput (TD)* scheme: forbidding back-propagation from *T* unit throught the dashed line with a certain probability. Dropout can reduce complex co-adaptations of two flows and enhance model's generalization ability.

Comparison with Convential RNN/LSTM As mentioned before, stacked RNN/LSTM is a solution for deep recurrent architecture. Actually, our proposed approach is a general version of it. This module can be considered as an extension of stacked RNN/LSTM with an extra context bridge, namely the *R* unit.

3.2 Overlap Coherence Training Scheme

Challenge A widely-used method is to sample a few items from the long sequence (successive or scattered) and learn a truncated version of the original RNN on them to solve the contradiction. Under this scheme, training on short samples instead of the long sequence greatly reduces the training complexity, which is very pratical for deep RNN.

Method In this paper, they consider shortening the long sequence to simplify the *l*-order Markov process into several *n*-order ($n < l$) ones, but we smooth the information propagation among short clips by introducing the **Overlap Coherence Training Scheme**. In training phase, they randomly sample clips that have random lengths and starting points, which will naturally generate many overlaps. The overlaps serve as the communication bridge among the disconnected clips to make information flow forward and backward smoothly throughout the whole sequence. –> **overlap coherence loss**

4. Experimental Results

4.1 Video Action Recognition and Anticipation

Implementation
Adoption Conventional Backbone-Supported Structure
Adopting Standalone RNN Structure

4.2 Polygon-RNN on Cityscapes

4.3 Video Future Prediction

5. Analysis

Analysis on Depth

6. Conclusion

> In this paper, we proposed a deep RNN framework for visual sequential applications.
> The first part of our deep RNN framework is the CBM structure designed to balance the temporal flow and representation flow. Based on the characteristics of these two flows, we proposed the Temporal Dropout to simplify the training process and enhance the generalization ability.
> The second part is the Overlap Coherence Training Scheme aiming at resolving the large resource consuming of deep RNN models, which can significantly reduce the length of sequences loaded into the model and guarantee the consistency of temporal information through overlaps simultaneously.