Full-GRU Natural Language Video Description for Service Robotics Applications

Introduction

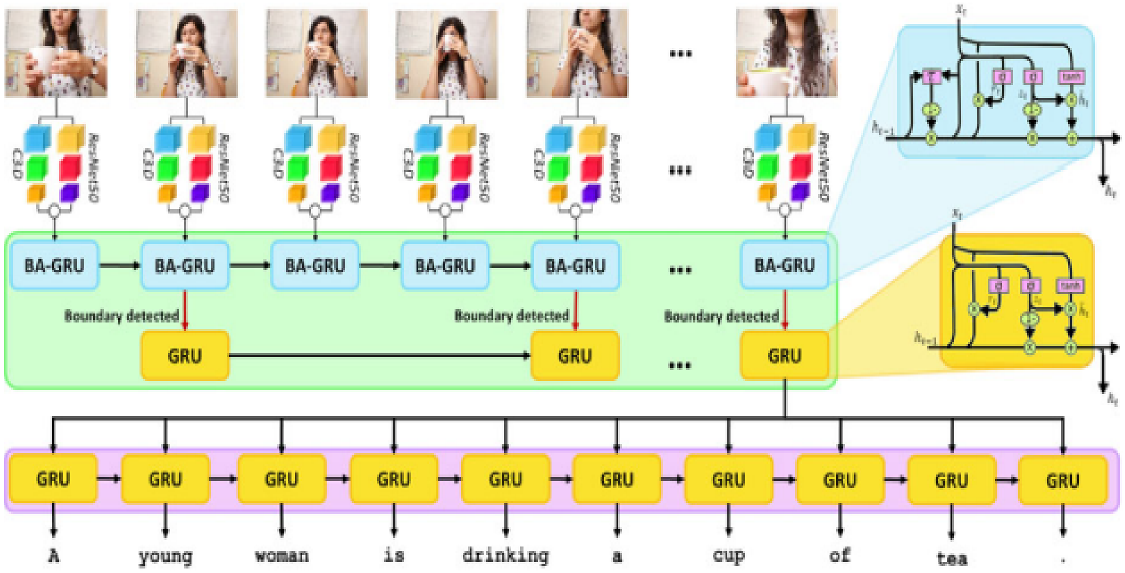
The ability to provide a description of the scene in a form that every user can easily understand is keystone for the success of effective and user-friendly service robotics products. In fact, a natural language description offers an interpretable manifestation of the robot's inner representation of the scene and is also a good basis for natural language question answering about what is happening in the environment. Hence, this functionality would provide a friendly interface also for non-expert people who would then be able to easily interact with their home robot in the near future.

Enabling effective human–robot interaction is crucial for any service robotics application. In this context, a fundamental aspect is the development of a user-friendly human–robot interface, such as a natural language interface. In this paper, the robot side of the interface, in particular the ability to generate natural language descriptions for the scene it observes is investigated. This capability is achieved via a deep recurrent neural network architecture completely based on the gated recurrent unit paradigm. The robot is able to generate complete sentences describing the scene, dealing with the hierarchical nature of the temporal information contained in image sequences. The proposed approach has fewer parameters than previous state-of-the-art architectures, thus it is faster to train and smaller in memory occupancy. These benefits do not affect the prediction performance. In fact, this method outperforms or is comparable to previous approaches in terms of quantitative metrics and qualitative evaluation when tested on benchmark publicly available datasets and on a new dataset introduced in this paper.

In this paper, a full-GRU NLVD system is proposed, that is able to deal with the hierarchical nature of the temporal information typical of natural and generic video sequences and obtains comparable performance with respect to more complex State-of-the-Art (SotA) systems. The proposed system features a GRU cell modified in order to automatically change its temporal connection if a boundary, i.e., a significant modification in the scene, is detected. This is the first full-GRU encoder-decoder architecture applied to the problem of NLVD.

Encoder-Decoder Full-Gru Architecture

The video frames are described via the ResNet50 and the C3D ConvNets. The obtained feature vectors are then fed, one at each time-step, in the first layer of the encoder. The proposed BA-GRU recurrent block, that encodes the video frames until a boundary is detected. Afterwards, the first-layer encoding is fed to the second layer of the encoder, which consists of a classical GRU block. The output of the encoding phase is a vector representing the entire video sequence. Finally, the GRU decoder produces the description emitting the most probable word at each time-step, conditioned to the video vector representation and the previous emitted words. The captioning process ends when a tag (i.e., the full-stop) is emitted. A pictorial representation of the system is shown in following figure.



. Architecture of the proposed system. Recurrent layers are depicted as unfolded graphs for explanatory purpose.

Video Frames and Caption Words Preprocessing

The video frames are preprocessed as follows. The output of the last fully connected layer of the ResNet50 ConvNet is computed every five video frames, to capture the appearance of the scene. To the same video frames is associated also the output of the C3D ConvNet to capture the movement in the scene, based on partially overlapped sliding windows of frames. The output of the two ConvNets are concatenated (forming a 2048+4096-dimensional vector) and mapped in a learned 512-dimensional linear embedding. The entire video is then represented by a sequence of features vectors (x1 , x2 , ..., xn), where the x· vectors are the feature vectors extracted from the frames of the video.

The captions are preprocessed as follows. First, the words are converted to lower-case and the punctuation characters are removed. Then, begin-of-sentence () and end-of-sentence () tags are added before and behind the sentence, respectively. Finally, the sentences are tokenized. From the tokenized sentences, we build a vocabulary (D). To prevent the formation of a large vocabulary containing many rare words, we retain only those tokens that appear at least five times in the caption corpus. To each token is associated an index in the vocabulary, based on its frequency in the vocabulary. A caption is then represented by a list of one-hot vectors (y1 , y2 , ..., yL), each of them corresponding to the representation of its words in the vocabulary. Similarly to what is done for the frames features, the captions are mapped in a learned 512-dimensional linear embedding.

Video Encoder

In this work, the boundary-aware LSTM (BALSTM) cell is built and devise a boundary-aware GRU (BA-GRU) cell. This cell is the first layer of a two-layers encoder. The second layer of the encoder is a simple GRU cell. The BA-GRU is a modification of the classical GRU cell (see Fig. 2, top right). The GRU is a recurrent neural networks with gating strategies to model wider temporal dependencies in the input sequence. The GRU is characterized by an update gate z_t and a reset gate r_t. At each timestep, a candidate activation h_t is computed based on the current input x_t, the previous inner state h_{t-1} and the values of the gates. In particular, the zt gate controls how much the inner state h_t has to be updated, the r_t gate controls how much the previous inner state h_{t-1} influences the candidate inner state value h_t. In this work, the GRU is modified by adding a boundary aware gate st, that modifies the inner connectivity of the unit based on the input and the inner state. In particular, when a substantial change in input sequence occurs, a boundary is estimated by a learnable function. Consequently, the inner state ht−1 is emitted as output and then re-initialized to zero.

Conclusions And Future Developments

This letter focuses on the NLVD task and presents a fullGRU encoder-decoder architecture to address it. We show that the proposed approach is faster to train and less memory consuming that other State-of-the-Art algorithms. Our method is also competitive in terms of performance on the public datasets which were partially used also for training. The experimental results on the devised dataset show that all methods have serious overfitting, making the generalization capabilities of new algorithm one of the most important questions to solve in future work. Other future work is the ability to better cope with videos of variable lengths. This issue could be tackled by cutting the continuous video sequence in shorter chunks and describing each chunk using our proposed method as it is. However, being able to deal with much longer videos is surely of great interest and the development of effective solutions to this problem will be the subject of future work.