Deep Residual Recurrent Neural Networks for Characterisation of Cardiac Cycle Phase from Echocardiograms

Fatemeh Taheri Dezaki[1(✉)], Neeraj Dhungel[1], Amir H. Abdi[1], Christina Luong[2], Teresa Tsang[2], John Jue[2], Ken Gin[2], Dale Hawley[2], Robert Rohling[1], and Purang Abolmaesumi[1(✉)]

[1] The University of British Columbia, Vancouver, BC, Canada
{fatemeht,purang}@ece.ubc.ca
[2] Vancouver General Hospital's Cardiology Laboratory, Vancouver, BC, Canada

Abstract. Characterisation of cardiac cycle phase in echocardiography data is a necessary preprocessing step for developing automated systems that measure various cardiac parameters. Accurate characterisation is challenging, due to differences in appearance of the cardiac anatomy and the variability of heart rate in individuals. Here, we present a method for automatic recognition of cardiac cycle phase from echocardiograms by using a new deep neural networks architecture. Specifically, we propose to combine deep residual neural networks (ResNets), which extract the hierarchical features from the individual echocardiogram frames, with recurrent neural networks (RNNs), which model the temporal dependencies between sequential frames. We demonstrate that such new architecture produces results that outperform baseline architecture for the automatic characterisation of cardiac cycle phase in large datasets of echocardiograms containing different levels of pathological conditions.

Keywords: Deep residual neural networks · Recurrent neural networks · Long short term memory · Echocardiograms · Frame identification

1 Introduction

According to the World Health Organization[1] millions of people worldwide suffer from the heart-related disease, a major cause of mortality. The 2-D echocardiography (echo) examination is a widely used imaging modality for early diagnosis. Echo

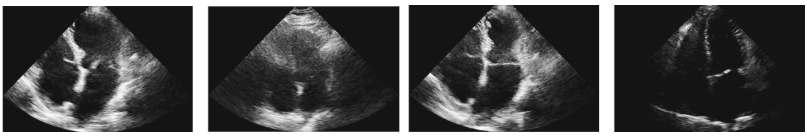
F.T. Dezaki and N. Dhungel—Contributed equally.

T. Tsang is the Director of the Vancouver General Hospital and University of British Columbia Echocardiography Laboratories, and Principal Investigator of the CIHR-NSERC grant supporting this work.

[1] Global status report on noncommunicable diseases, 2014.

M.J. Cardoso et al. (Eds.): DLMIA/ML-CDS 2017, LNCS 10553, pp. 100–108, 2017.
DOI: 10.1007/978-3-319-67558-9_12

can be used to estimate several cardiac parameters such as stroke volume, end-diastolic volume, and ejection fraction [1]. These parameters are generally measured by identifying end-systolic and end-diastolic frames from cine echos [1,2]. Current detection is either manual or semi-automatic [3,4], relying primarily on the availability of electrocardiogram (ECG) data, which can be challenging in point-of-care and non-specialized clinics. Moreover, such detection techniques add to the workload of medical personnel and are subject to inter-user variability. Automatic detection of cardiac cycle phase in echo can potentially alleviate these issues. However, such detection is challenging due to low signal-to-noise ratio in echo data, subtle differences between the consecutive frames, the variability of cardiac phases, and temporal dependencies between these frames.



(a) End-Diastolic(ED) Frames (b) End-Systolic(ES) Frames

Fig. 1. Examples of end-diastolic and end-systolic frames in 2-D echocardiograms.

There have been a number of attempts for automatic detection and labeling of frames in echo [2,4,5]. The most common approach is to use a segmentation strategy, such as level-sets or graph-cuts, for identifying the boundary of the left ventricle in an echo sequence [2,5]. Boundaries that correspond to largest and smallest ventricular areas are regarded as end-diastolic (ED) and end-systolic (ES) frames [5] (Fig. 1). A major drawback of those segmentation-based approaches is the requirement for a good initialization and localization of the left ventricle. In another approach [4], information from cine frames was projected onto a low-dimensional 2-D manifold, after which the difference of distance between points on the manifold was used to determine ED and ES frames. However, this approach does not consider the complex temporal dependencies between the frames, which may subsequently lead to sub-optimal results. Recently, methods based on convolutional neural networks (CNNs) combined with recurrent neural networks (RNNs) have produced state-of-the-art results in many computer vision problems [6–9]. A combination of CNNs and RNNs has been successfully applied in various problem domains such as detecting frames from videos, natural language processing, object detection and tracking [3,8,10,11]. However, our experience is that such network structure cannot be easily extended to the analysis of echo data, which suffer from low signal-to-noise ratio, where anatomical boundaries may not be as clearly visible compared to other imaging modalities.

Here, we formulate a very deep CNN architecture using residual neural networks (ResNets) [7] for extracting deep hierarchical features, which are then passed to RNNs containing two blocks of long short term memory (LSTM) [12] to decode the temporal dependencies between these features. The primary motivation of using a ResNets with LSTMs is that layers in ResNets are reformulated

for learning the residual function with respect to input layers for countering the vanishing or exploding gradient problem in CNN while going deeper [7]. In addition to this, we minimize the structured loss function that mimics the temporal changes in left ventricular volume during the systolic and diastolic phase [3]. We demonstrate that the proposed method using ResNets and LSTMs with structured loss function produces state-of-the-art accuracy for detecting cardiac phase cycle in echo data without the need of segmentation.

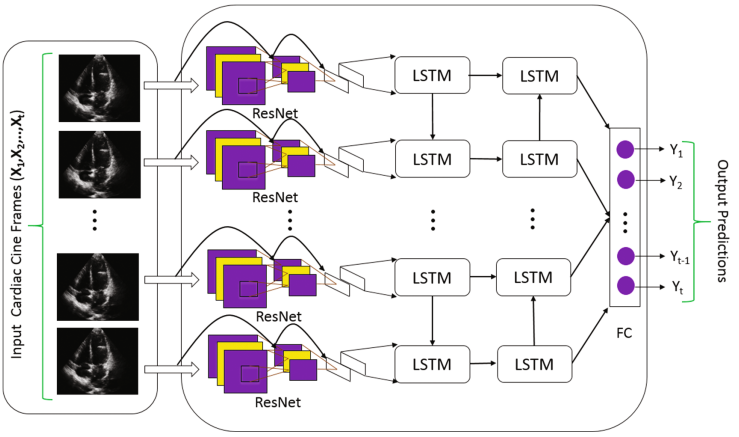


Fig. 2. The proposed method for characterisation of cardiac cycle phase from echocardiograms using deep residual recurrent neural networks (RRNs), which contains residual nets (ResNets), followed by two blocks of long term short term memory (LSTM) and a fully connected layer (FC).

2 Methods

2.1 Dataset

Let $\mathcal{D} = \{(\mathbf{c}^{(i)}, \mathbf{y}^{(i)})_j\}_{j=1}^{|\mathcal{D}|}$ represent the dataset, where $i \in \{1, 2, .., N\}$ indexes the number of cardiac cycles for an individual study j, $\mathbf{c} = \{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_t\}$ represents the collection of frames $\mathbf{x}$ for each patient such that $\mathbf{x}_t : \Omega \rightarrow \mathbb{R}$ with $\Omega \subset \mathbb{R}^2$, and $\mathbf{y} = \{y_1, y_2, .., y_t\}$ denotes the class label for individual frames computed using the following function [3]:

$$y_t = \begin{cases} \left(\frac{|t - T^i_{\mathrm{ES}}|}{|T^i_{\mathrm{ES}} - T^i_{\mathrm{ED}}|}\right)^{\tau}, & \text{if } T^i_{\mathrm{ED}} < t < T^i_{\mathrm{ES}} \\ \left(\frac{|t - T^i_{\mathrm{ES}}|}{|T^i_{\mathrm{ES}} - T^{i+1}_{\mathrm{ED}}|}\right)^{\frac{1}{\tau}}, & \text{if } T^i_{\mathrm{ES}} < t < T^{i+1}_{\mathrm{ED}} \end{cases} \tag{1}$$

where $T^{\mathrm{i}}_{\mathrm{ES}}$ and $T^{\mathrm{i}}_{\mathrm{ED}}$ are the locations of ES and ED frames in the i_{th} cardiac cycle and τ is an integer constant. The expression in Eq. 1 mimics the ventricular volume changes during the systole and diastole phases of a cardiac cycle [3].

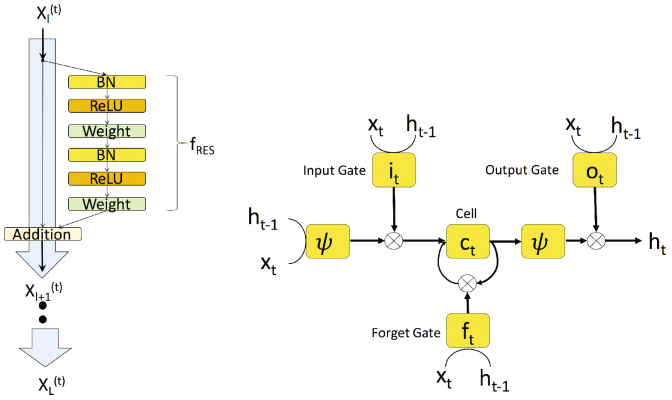
2.2 Deep Residual Recurrent Neural Networks (RRNs)

We propose deep residual recurrent neural networks (RRNs) to decode the phase information in echo data. RRNs constitute of a stack of residual units for extracting the hierarchical spatial features from echo data, RNNs that use LSTMs for decoding the temporal information, and a fully connected layer at the end. As shown in Fig. 3(a), each residual unit is made by the addition of a residual function and an identity mapping, and can be expressed in a general form by the following expression [7]:

$$\mathbf{x}_L^{(t)} = \mathbf{x}_l^{(t)} + \sum_{l=1}^{L-1} f_{\text{RES}}(\mathbf{x}_l^{(t)}; \mathcal{W}_l), \tag{2}$$

where $\mathbf{x}_l^{(t)}$ is the input feature to the $l \in \{1, ..., L\}^{th}$ residual unit (for $l = 1$, $\mathbf{x}_l = \mathbf{x}_t$), $\mathcal{W}_l = \mathbf{w}_{l,k}$ is the set of weights for the l^{th} residual unit, $k \in \{1, ..., K\}$ represents the numbers of layers in the residual unit, $f_{\text{RES}}(.)$ is called the residual function represented by a convolutional layer (weight) [6,13], batch normalization (BN) [14] and rectilinear unit (ReLU) [7,15] (Fig. 3(a)).

We use RNNs for decoding the temporal dependencies within the echo sequence. LSTM network as shown in Fig. 3(b) is a type of RNNs with the addition of memory cell that allows the network to learn to remember or forget the hidden states. In particular, LSTM updates the hidden states $\mathbf{h}_t$ and its memory units $\mathbf{c}_t$ with given sequential input $\mathbf{x}_t$ using the following expressions [8]:



(a) ResNet model (b) LSTM model

Fig. 3. Two basic building blocks of our proposed method: (a) ResNet, and (b) LSTM. A typical ResNet model contains residual units with BN, ReLU, and convolutional (weight) layers, stacked together and identity mapping (addition). An LSTM model is made of memory units (cells) which maintain a cell state using input, output, and forget gates.

$$\begin{aligned} \mathbf{i}_t &= \phi\left(\mathbf{w}_{xi}\mathbf{x}_t + \mathbf{w}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i\right); \\ \mathbf{f}_t &= \phi\left(\mathbf{w}_{xf}\mathbf{x}_t + \mathbf{w}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f\right); \\ \mathbf{o}_t &= \phi\left(\mathbf{w}_{xo}\mathbf{x}_t + \mathbf{w}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o\right); \end{aligned} \tag{3}$$

$$\begin{aligned} \mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \psi\left(\mathbf{w}_{xi}\mathbf{x}_t + \mathbf{w}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i\right), \\ \mathbf{h}_t = \mathbf{o}_t \odot \psi(\mathbf{c}_t), \end{aligned} \tag{4}$$

where $\odot$ is element-wise product, ψ is the hyperbolic tangent function, ϕ is the sigmoid function, and $\mathbf{w}_{xi}$, $\mathbf{w}_{hi}$, and $\mathbf{b}_i$ are the weights and biases between input $\mathbf{i}_t$ and hidden state $\mathbf{h}_{t-1}$; $\mathbf{w}_{xf}$, $\mathbf{w}_{hf}$, and $\mathbf{b}_f$ are the weights and biases between forget state $\mathbf{f}_t$ and hidden state $\mathbf{h}_{t-1}$; and $\mathbf{w}_{xo}$, $\mathbf{w}_{ho}$, and $\mathbf{b}_o$ are the weights and biases between output $\mathbf{o}_t$ and hidden state $\mathbf{h}_{t-1}$. Our proposed RRNs (Fig. 2) can be expressed as the following function:

$$\{\tilde{y}_1, \tilde{y}_2, .., \tilde{y}_t\} = f_{\text{RRNs}}\left(\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_t; \mathcal{W}_{\text{RRNs}}\right), \tag{5}$$

where f_{RRNs} takes the sequential echo images as input to RRNs and outputs the predicted values $\tilde{y}_t$ with parameters of RRNs as $\mathcal{W}_{\text{RRNs}} = [\mathcal{W}_{\text{ResNets}}, \mathcal{W}_{\text{LSTM}}, \mathbf{w}_{\text{fc}}]$. $\mathcal{W}_{\text{ResNets}}$ represents the parameters of ResNets, $\mathcal{W}_{\text{LSTM}}$ represents the parameters of LSTM, and $\mathbf{w}_{\text{fc}}$ is associated with weights of the final fully connected layer. We train the proposed RRNs in an end-to-end fashion with stochastic gradient decent to minimise the following loss function [3] containing a first term as L^2 norm and a second term as structured loss (for notation simplicity, we dropped index i denoting the echo sequences):

$$\begin{aligned} \ell(\mathcal{W}_{\text{RRNs}}) = \sum_{j=1}^{|\mathcal{D}|} \Bigg[\frac{\alpha}{T} \sum_{t=1}^{T} ||y_{(j,t)} - \tilde{y}_{(j,t)}||^2 + \frac{2\beta}{T} \sum_{t=2}^{T} \big(\mathbb{I}(y_{(j,t)} > y_{(j,t-1)}) \\ \max(0, \tilde{y}_{(j,t-1)} - \tilde{y}_{(j,t)}) + \mathbb{I}(y_{(j,t)} < y_{(j,t-1)}) \max(0, \tilde{y}_{(j,t)} - \tilde{y}_{(j,t-1)})\big) \Bigg], \end{aligned} \tag{6}$$

where $\mathbb{I}$ denotes the indicator function, T is the maximum frame length, and α, β are user defined variables which are cross validated during the experiment. The structured loss term takes into account monotonically decreasing nature of the cardiac volume changes during the systole phase and monotonically increasing nature of cardiac volume changes in the diastolic phase. Finally, inference in the purposed method is done in a feed-forward way using the learned model.

3 Experiments

We carried out the experiments on a set of 2-D apical 4 chambers (AP4) cine echoes obtained at Vancouver General Hospital (VGH) with ethics approval from the Clinical Medical Research Ethics Board of the Vancouver Coastal Health (VCH) (H13-02370). The dataset used in this project consists of 1,868 individual patient studies containing a range of pathological conditions. The data were obtained using different types of ultrasound probes from various manufacturers,

which consists of both normal and abnormal cases. Experiments were run by randomly dividing these cases into mutually exclusive subsets, such that 60% of the cases were available for training, 20% for validation, and 20% for the test. The average length of echo sequence in each study was about 30 frames. The electrocardiogram (ECG) signals, synchronized with cine clips, were also available. We considered the ECG signal to generate the ground-truth labels for ED and ES frames by identifying R and end of T points in that signal. Subsequently, the intermediate labels for all frames in each echo sequence were derived using Eq. (1), where the value of τ is selected to be 3 and labels are normalized in the range of [0–1] (Note that 0 represents ES and 1 represents ED). The experiment was conducted on the image resolution of 120×120, where we sub-sample the original image resolution using bi-cubic interpolation.

Our proposed RRNs, shown in Fig. 2, takes as input a cardiac sequence containing 30 frames irrespective of its location in the cardiac phase. Each frame is passed through the convolutional layer (weights) plus a ReLU, where the convolutional layer contains eight filters of size 3×3 followed by nine subsequent residual units. Each residual unit is made up of batch normalization (BN) plus ReLU plus weights. Each convolutional layer in the first three residual units contained the same eight filters (size 3×3), the fourth, fifth and sixth residual units contained 16 filters of size 3×3, and the seventh, eighth and ninth units had 32 filters of size 3×3. In the second to the last layer, we concatenated the 32 output features from each ResNet to form 192 features (32×6), followed by two blocks of LSTM layer and a final fully connected layer containing 30 neurons, to generate a label for each frame of the cardiac cycle. For comparison, we also used the shallow CNN model, Zeiler-Fergus (ZF), [3] in combination with different loss functions. The performance of our approach is calculated based on three metrics, the coefficient of determination (R^2 score), average absolute frame detection error, and computation time of each sample. R^2 score is defined by $R^2 = 1 - \frac{\sum(y_t - \tilde{y}_t)^2}{\sum(y_t - \bar{y})^2}$, where $\bar{y}_t$ is the mean of true labels. Average absolute frame detection error is calculated as $Err_S = \frac{1}{|\mathcal{D}|} \sum |T_S - \widetilde{T}_S|$, where S is either the ED or the ES frame. All experiments were performed on a system with an Intel(R) Core(TM) i7-2600k 3.40 GHz×8 CPU with 8 GB RAM equipped with the NVIDIA GeForce GTX 980Ti graphics card.

4 Results and Discussion

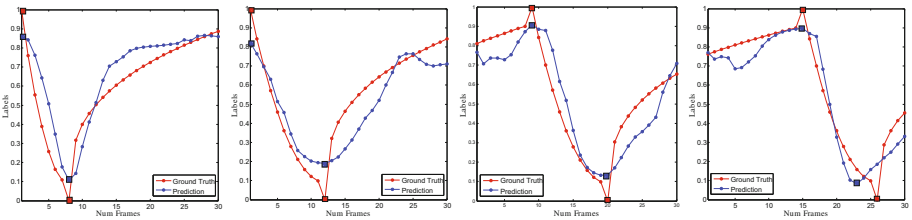
Table 1 shows the result of the proposed approach using two different loss functions: a) L_2 loss, and b) L_2 loss + structured loss [3]. The R^2 values for these two settings are 0.36 and 0.66, respectively. Similarly, (Err_{ED}, Err_{ES}) for these two settings are (4.4, 4.7) and (3.7, 4.1), respectively. The baseline method [3] using CNN with the same setting produces the R^2 score of 0.13, whereas the average absolute frame detection error for both ED and ES error is 6.3 and 7.3 in setting (a), and 6.4 and 7.3 in setting (b).

Results in Table 1 show that the proposed RRN method outperforms the baseline method in both settings with a large margin. The advantage of the

proposed method lies in its ability to go deeper with 126 layers in ResNets compared to the shallow architecture in the baseline model (CNN). Furthermore, we note that performance of the proposed RRNs model improves with the use of structured loss with higher R^2 score and lower frame detection error compared to the use of L_2 loss only. This is due to fact that structured loss introduces a structured constraint in the loss function which helps with smoothing the predicted labels, thus increasing the overall accuracy. It is also important to note that R^2 score of our proposed model is far better than that of baseline model (0.66 vs. 0.13), indicating that our approach is a better function approximation in the global context (with respect to assigning the correct label to individual frames). The method takes only 80 ms to characterise each cardiac cycle showing its efficiency in terms of computation cost. Figure. 4 shows some visual results of frame characterisation in a sequence of 30 frames as a function of labels represented in Eq. (1). In particular, Fig. 4(a and b) shows cases where the first frame in the sequence starts as an ED frame, and Fig. 4(c and d) shows cases where there is a shift in the location of the ED frame. Similarly, in Fig. 5, we also show one fairly accurate visual example of detection of ED and ES frames using the proposed method in the test set.

Table 1. Performance of the proposed and state-of-the-art methods on the test set.

Method	R^2 Score	Err_{ED}	Err_{ES}
ResNet+LSTM+L_2 loss (Proposed)	0.36	4.4	4.7
ResNet+LSTM+L_2 loss+struct loss (Proposed)	**0.66**	**3.7**	**4.1**
CNN+LSTM+L_2 loss [3]	0.13	6.3	7.3
CNN+LSTM+L_2 loss+struct loss [3]	0.13	6.4	7.3



(a) (b) (c) (d)

Fig. 4. Label approximation of the proposed method (blue) plotted against the ground-truth labels (red) on some sample cases, where the ES and ED frames (denoted by rectangular boxes) are correctly identified within an error margin of 3 to 4 frames. (Color figure online)

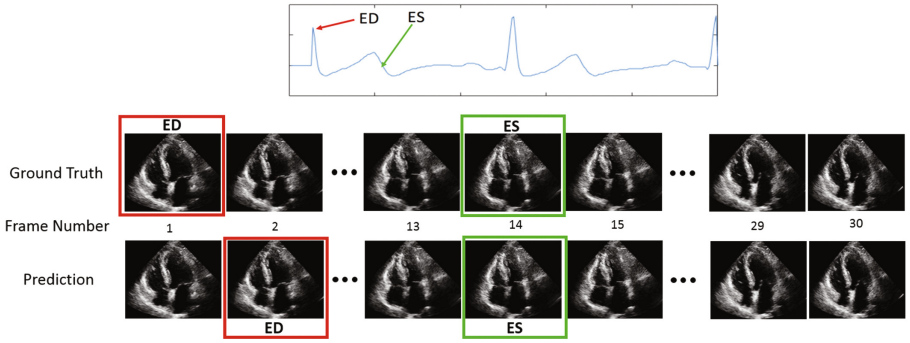


Fig. 5. A sample echo cine, consisting of 30 frames, alongside its ECG signal. The results of the proposed method in terms of ED and ES frame detection are demonstrated and compared with their ground-truth labels.

5 Conclusion and Future Works

In this paper, we proposed a deep residual recurrent neural net (RRN) for automated identification of cardiac cycle phases (ED and ES) from echocardiograms. We also showed that the proposed method produces results that outperform a baseline method using CNN with a large margin in detecting ED and ES frames. We achieved the R^2 score of 0.66 and the average absolute frame detection error of 3.7 and 4.1 for ED and ES, respectively. Our results suggest that the method has the potential to be used in clinical setting and is robust to all sorts of the pathological condition of the patient. In future, we plan to use the method as a pre-processing step for assessing several cardiac function parameters, including the ejection fraction.

References

1. Barcaro, U., Moroni, D., Salvetti, O.: Automatic computation of left ventricle ejection fraction from dynamic ultrasound images. Pattern Recogn. Image Anal. **18**(2), 351 (2008)
2. Abboud, A.A., Rahmat, R.W., et al.: Automatic detection of the end-diastolic and end-systolic from 4D echocardiographic images. JCS **11**(1), 230–240 (2015)
3. Kong, B., Zhan, Y., Shin, M., Denny, T., Zhang, S.: Recognizing end-diastole and end-systole frames via deep temporal regression network. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 264–272. Springer, Cham (2016). doi:10.1007/978-3-319-46726-9_31
4. Gifani, P., Behnam, H., Shalbaf, A., Sani, Z.A.: Automatic detection of end-diastole and end-systole from echocardiography images using manifold learning. Physiol. Meas. **31**(9), 1091 (2010)
5. Darvishi, S., Behnam, H., Pouladian, M., Samiei, N.: Measuring left ventricular volumes in two-dimensional echocardiography image sequence using level-set method for automatic detection of end-diastole and end-systole frames. Res. Cardiovasc. Med. **2**(1), 39 (2013)

6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR, pp. 770–778 (2016)
8. Donahue, J., Anne Hendricks, L., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: IEEE CVPR, pp. 2625–2634 (2015)
9. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: CNN-RNN: a unified framework for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2016)
10. Sundermeyer, M., Schlüter, R., Ney, H.: LSTM neural networks for language modeling. In: Interspeech, pp. 194–197 (2012)
11. Milan, A., Rezatofighi, S.H., Dick, A., Reid, I., Schindler, K.: Online multi-target tracking using recurrent neural networks, arXiv preprint arXiv:1604.03635 (2016)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
13. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. In: Arbib, M.A. (ed.) Handbook of Brain Theory and Neural Networks, vol. 3361. MIT Press (1995)
14. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167 (2015)
15. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML, pp. 807–814 (2010)