6N-03

Visually Explaining Videos using Recurrent Neural Networks with Gradient-Based Localization

Naoya Yamashita† Naoto Iwahashi† Seiya Nakano‡ Takamitsu Sakai‡ Mitsunori Hamano‡

Okayama Prefectural Univ.† AISIN AW Co. Ltd‡

1. Introduction

Artificial intelligence systems are outperforming humans in an increasing number of tasks. The effective collaboration between artificial intelligence systems and humans has been increasingly pursued, for which the mechanisms that enable human understanding and trust in such systems are important. However, as the systems are constructed through machine learning in a highly abstracted manner, it is difficult for human to understand why the systems make decisions under individual situations. The understanding, visualization, and interpretation of artificial intelligence systems have recently been attracting attention [1].

In this context, Samek et al. [2] proposed the visual explanation technique called gradient-weighted class activation mapping (Grad-CAM), which enables the systems to present humans with the spatial portions important for decision making by using convolutional neural network (CNN). However, as the Grad-CAM technique has previously been applied to a feedforward CNN, it is difficult for the technique to treat tasks with time-series input.

In this paper, we propose a recurrent Grad-CAM, which is the extension of the conventional Grad-CAM. The proposed technique applies the gradient-based localization principle to a recurrent CNN to solve the aforementioned difficulties.

2. Recurrent grad-CAM

Figs. 1 and 2 present the overviews of the conventional Grad-CAM and our proposed recurrent Grad-CAM, which applies the gradient-based localization principle to a long-term recurrent convolutional network (LRCN) [3].

The main difference between the Grad-CAM and recurrent Grad-CAM is that the gradient of score $y_{t'}^{c}$ of class c at output time t' with respect to feature map A_{tij}^{k} at input time t of the convolutional layer, i.e., $\partial y_{t'}^{c}/\partial A_{tij}^{k}$, is calculated between different

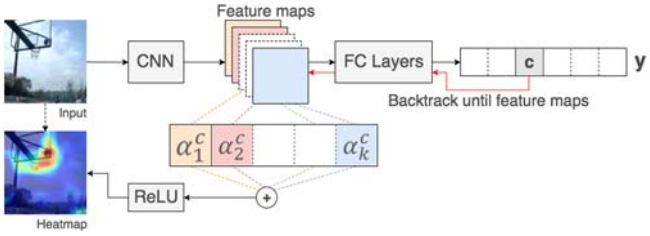

Fig. 1: Overview of Grad-CAM

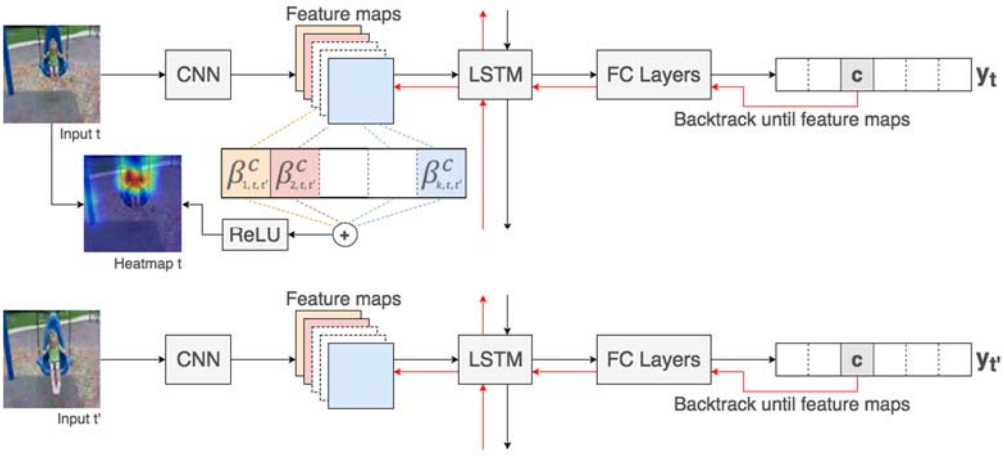

Fig. 2: Overview of our proposed recurrent Grad-CAM

time indexes t and t' . When flowing back, these gradients are global average pooled to obtain the importance weights as

$$\beta_{ktt'}^{c} = \frac{1}{Z}\sum_{i}\sum_{j}\frac{\partial y_{t'}^{c}}{\partial A_{tij}^{k}} \quad (1)$$

By using these weights, the weighted combination of forward activation maps is conducted, and is followed by a ReLU to obtain a heatmap as

$$L_{t}^{c} = ReLU\left(\sum_{k}\beta_{ktT}^{c}A_{t}^{k}\right) \quad (2)$$

where T denotes the final time index of a video. Note that this is only an example calculation for obtaining a heatmap by using the importance weights represented by Eq. 1; many variations are possible.

3. Experiments

We conducted experiments for visual explanation in a video classification task, and used CNN and LRCN as baseline grad-CAM and our proposed recurrent Grad-CAM, respectively. A total of 1,600 videos (1,200 for training and 400 for testing) from UCF11 [4] were used, each of which falls into one of 11 motion classes. The image time sequences were extracted from the original videos at an interval of 7.5 fps. The CNN and LRCN were trained frame-by-frame and according to the overall sequence, respectively. The accuracies of the training data were 100% in both CNN and LRCN and those of the test data were 94% and 98%, respectively. The resultant heatmap sequence examples are shown in Fig. 3. We confirmed that compared to the conventional Grad-CAM, the recurrent Grad-CAM successfully produced more suitable heatmap sequences that explained temporal-spatial portions important to the classification of decision making.

4. Conclusion

In this paper, we proposed a recurrent Grad-CAM for visual explanation of video tasks based on deep neural networks, and evaluated its validity. It can be applied to a wide range of video tasks, such as captioning, question-answering, and more general human-machine multimodal interactions.

Reference

[1] R. Selvaraju, et al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv preprint arXiv:1610.02391v3,* 2017.

[2] W. Samek, et al.: Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv preprintarXiv:1708.08296,* 2017.

[3] J. Donahue, et al.: Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *CVPR*, 2015.

[4] J. Liu, J. Luo and M. Shah, Recognizing Realistic Actions from Videos “in the Wild”. *CVPR*, 2009.

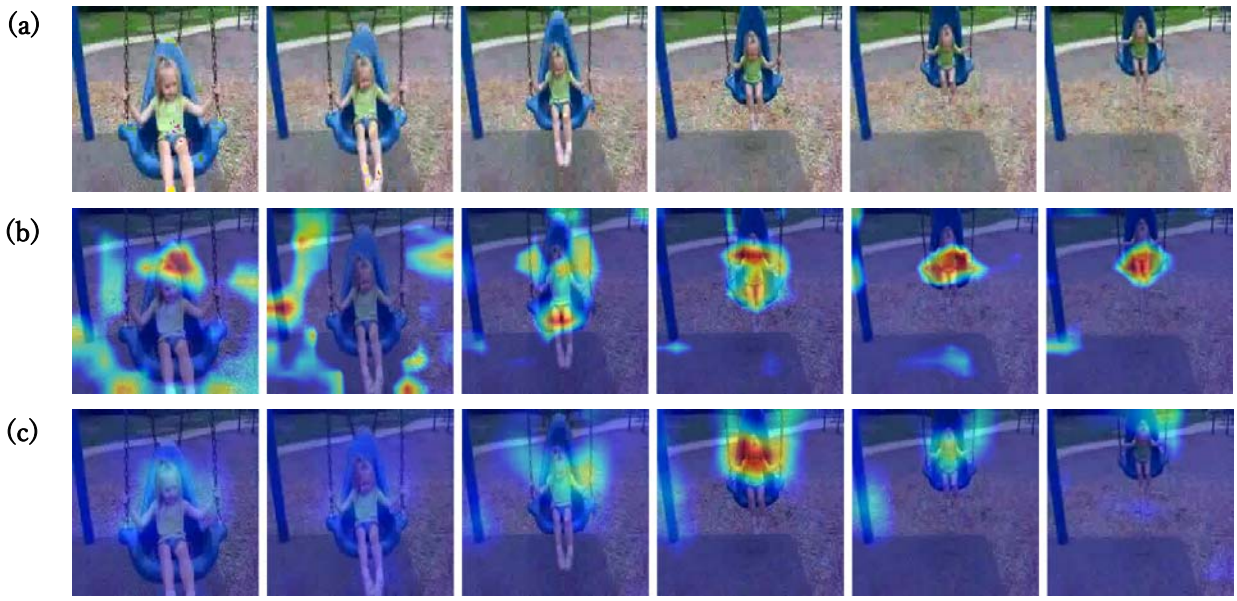

Fig. 3: Resultant heatmap examples: (a) original video, (b) Grad-CAM, and (c) recurrent Grad-CAM