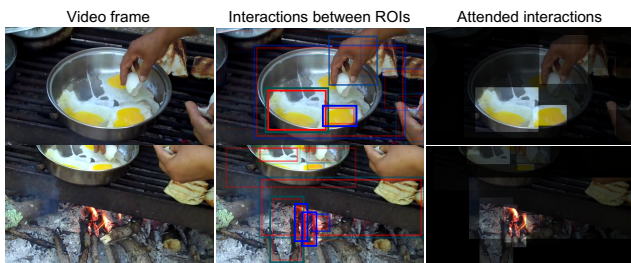


Attend and Interact: Higher-Order Object Interactions for Video Understanding (extended abstract)

Chih-Yao Ma^{*1}, Asim Kadav², Iain Melvin², Zsolt Kira³, Ghassan AlRegib¹, and Hans Peter Graf²

¹Georgia Institute of Technology, ²NEC Laboratories America, ³Georgia Tech Research Institute



Action prediction: *cooking on campfire*, *cooking egg*, ...

Figure 1. *Higher-order object interactions* are progressively detected based on selected inter-relationships. ROIs with the same color (weighted **r**, **g**, **b**) indicating there exist inter-object relationships. Groups of inter-relationships then jointly model higher-order object interaction of the scene (interaction between different colors). *Right*: ROIs are highlighted with their attention weights for higher-order interactions. The model further reasons about the interactions through time.

Abstract

In this paper, we propose to efficiently learn higher-order interactions between arbitrary subgroups of objects for fine-grained video understanding. We demonstrate that modeling object interactions significantly improves accuracy for both action recognition and video captioning. The proposed method achieve state-of-the-art performances on the Kinetics and ActivityNet Captions datasets even though the videos are sampled at a maximum of 1 FPS. To the best of our knowledge, this is the first work modeling object interactions on open domain large-scale video datasets.

1. Introduction

Video understanding tasks such as activity recognition and caption generation are crucial for various applications in surveillance, video retrieval, human behavior understanding, etc. Recently, datasets for video understanding such

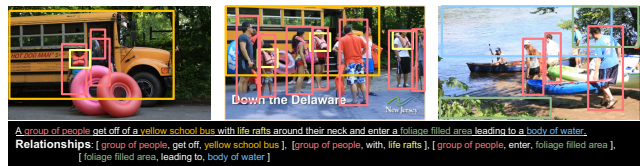


Figure 2. Video captions are composed of multiple visual relationships and interactions. We detect higher-order object interactions and use them as basis for video captioning.

as Charades [3], Kinetics [1], and ActivityNet Captions [2] contain diverse real-world examples and represent complex human and object interactions that can be difficult to model with state-of-the-art video understanding methods [3]. Consider the example in Figure 1. To accurately predict *cooking on campfire* and *cooking egg* among other similar action classes requires understanding of fine-grained object relationships and interactions. For example, a hand breaks an egg, eggs are in a bowl, the bowl is on top of the campfire, campfire is a fire built with wood at a camp, etc. Existing approaches for action recognition often focus on representing the overall visual scene (coarse-grained) as sequence of inputs that are combined with temporal pooling methods. These approaches ignore the fine-grained details of the scene and do not infer interactions between various objects in the video. On the other hand, in video captioning tasks, although prior approaches use spatial or temporal attention to selectively attend to fine-grained visual content in both space and time, they too do not model object interactions. These methods do not ground their predictions on object relationships and interactions. However, modeling visual relationships and object interactions in a scene is a crucial form of video understanding as shown in Figure 2

Prior work in understanding visual relationships in the image domain has recently emerged as a prominent research problem. However, it is unclear how these techniques can be adapted to open-domain video tasks, given that the video is intrinsically more complicated in terms of temporal reasoning and computational demands. Toward this end, we

^{*}Work performed as a NEC Labs intern

Figure 3. **Tobogganing**: Identifying *Tobogganing* essentially need three elements: toboggan, snow scene, and a human sitting on top. The three key elements are accurately identified and their interaction are highlighted as we can see from $t = 1$ to $t = 3$. The model also tracks the person and toboggan throughout the while video and ignore the irrelevant background snow scene.

present a generic recurrent module for fine-grained video understanding, which dynamically discovers higher-order object interactions via an efficient dot-product attention mechanism combined with temporal reasoning. Our work is applicable to various open domain video understanding problems. In this paper, we validate our method on two video understanding tasks with new challenging datasets: action recognition on Kinetics [1] and video captioning on ActivityNet Captions [2] (with ground truth temporal proposals). By combining both coarse- and fine-grained information, our **SINet** (Spatiotemporal Interaction Network) for action recognition and **SINet-Caption** for video captioning achieve state-of-the-art performance on both tasks while using RGB video frames sampled only at maximum 1 FPS.

2. Qualitative Results

2.1. Qualitative analysis on Kinetics

To validate the proposed method, we qualitatively show how the SINet selectively attends to various regions with relationships and interactions across time. In this extended abstract, we show one example in Figure 3. In each of the figure, the top row of each video frame has generally multiple ROIs with three colors: red, green, and blue. ROIs with the same color indicates that there exist inter-relationships. We then model the interaction between groups of ROIs across different colors. The color of each bounding box is weighted by the attention generated by the proposed method. Thus, if some ROIs are not important, they will have smaller weights and will not be shown on the image. The same weights are then used to set the transparent ratio for each ROI. The brighter the region is, the more important the ROI is.

2.2. Qualitative analysis on ActivityNet Captions

A common problem with the state-of-the-art captioning models is that they often lack the understanding of the relationships and interactions between objects, and this is often the result of dataset bias. For instance, when the model de-

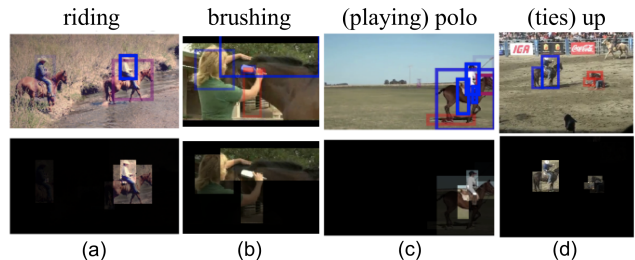


Figure 4. What interactions (verb) learned for video captioning. We verify how the SINet-Caption distinguishes various type of interactions with a common object - *horse*. (a) People are riding horses. (b) A woman is brushing a horse. (c) People are playing polo on a field. (d) The man ties up the calf.

fects both a person and a horse. The caption predictions are very likely to be: A man is riding on a horse, regardless whether if this person has different types of interactions with the horse.

We are thus interested in finding out whether if the proposed method has the ability to distinguish different types of interactions when common objects are presented in the scene. In Figure 4, each video shares a common object in the scene - *horse*. We show the verb (interaction) extracted from a complete sentence as captured by our proposed method. While all videos involve horses in the scene, our method successfully distinguishes the interactions of the human and the horse.

References

- [1] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2
- [2] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 2
- [3] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 1