Spatial-temporal Concept based Explanation of 3D ConvNets

Ying Ji[1], Yu Wang[2], Kensaku Mori[1], Jien Kato[3]

[1]Nagoya University [2]Hitotsubashi University [3]Ritsumeikan University

RITSUMEIKAN UNIVERSITY

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

XAI4CV @ CVPR 2022

Motivation

- Explaining 2D image recognition has achieved outstanding success.
- High-level concepts have been utilized in explaining 2D image recognition ConvNets[1].
- Due to the computation cost and complexity of video data, the explanation of 3D video recognition ConvNets is less studied.

Goal

- Extend 2D ACE[2] to 3D ACE and use spatial-temporal concepts to interpret the decision procedure of 3D video recognition ConvNets.
- Validate our method on the Kinetics dataset using popular network architectures and visualize the results.

Proposed method

- Videos are segmented into supervoxels. Similar supervoxels within each class are clustered to a set of spatial-temporal concepts.
- 3D ACE evaluates the importance score of each concept with respect to the class it belongs.
- Within the decision procedure, the network pays more attention to the concepts with high score.

Results

- Dataset: 10 classes randomly selected from Kinetics-700 dataset.
- Table 1 shows the results of adding concepts using ResNet-18. The first row is adding the highest score concepts. The second row is adding concepts randomly. The third row is adding the lowest score concepts.
- Table 2 shows the results of removing different concepts.
- Figure 2 visualize the concepts with the highest importance score and the lowest importance score.

Model	Concepts	1	2	3	4	5	baseline
	Top	11.67	23.96	32.92	39.38	**46.67**	
r3d-18	Random	10.63	21.25	32.50	37.71	41.25	75.62
	Least	9.79	16.04	26.04	33.13	41.46	

Table 1. The classification result of adding different concepts.

Model	Concepts	1	2	3	4	5	baseline
	Top	69.79	66.25	50.83	39.58	**24.38**	
r3d-18	Random	72.29	64.38	51.04	39.79	28.13	75.62
	Least	73.33	64.38	51.67	42.29	28.13	

Table 2. The classification result of removing different concepts.

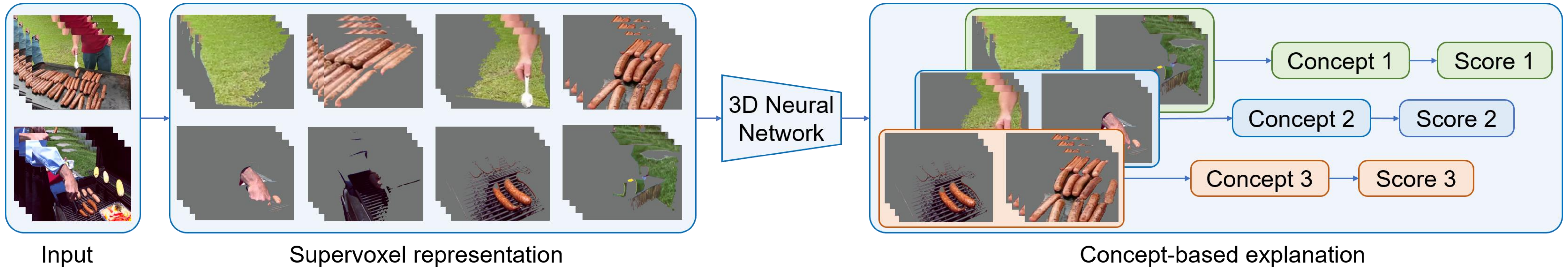

Fig. 1: pipeline of proposed method

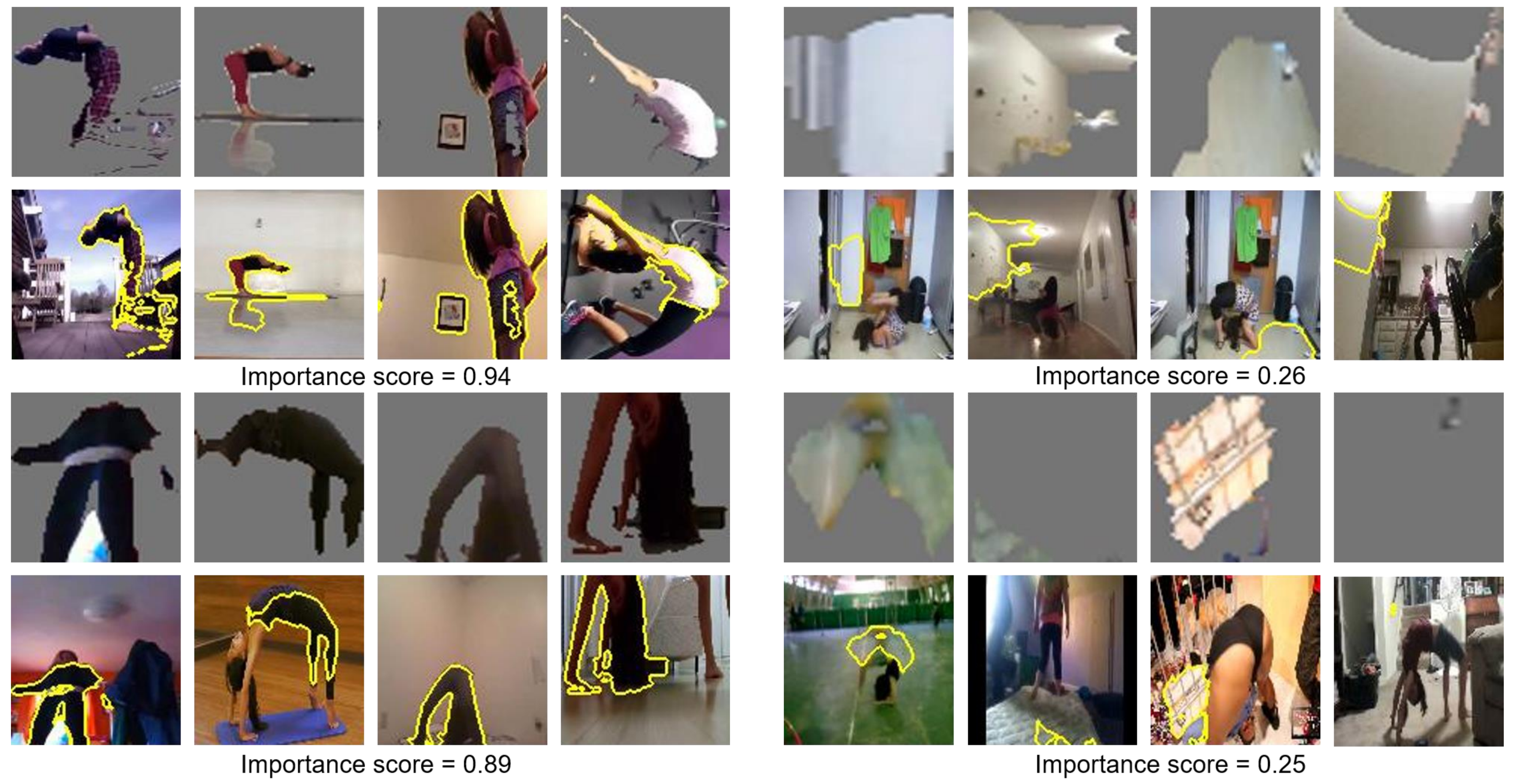

Fig. 2: visualization of different concepts

Conclusion

- We proposed a spatial-temporal concept-based explanation framework for 3D ConvNets.
- Different from the previous low-level pixel-based method, our research provides a human-understandable high-level explanation.
- Experiments show the effectiveness of our proposed method.

Reference

[1] Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.

[2] Ghorbani, Amirata, et al. "Towards automatic concept-based explanations." Advances in Neural Information Processing Systems 32 (2019).

Contact information

- E-mail: jiying@nagoya-u.jp
- GitHub: a TensorFlow implementation is available on: https://github.com/OrangeeJi/3D-ACE