ACM multimedia
Chengdu, China OCT 20-24 2021

Counterfactual Debiasing Inference for Compositional Action Recognition

Pengzhan Sun*, Bo Wu*, Xunsong Li, Wen Li, Lixin Duan, Chuang Gan

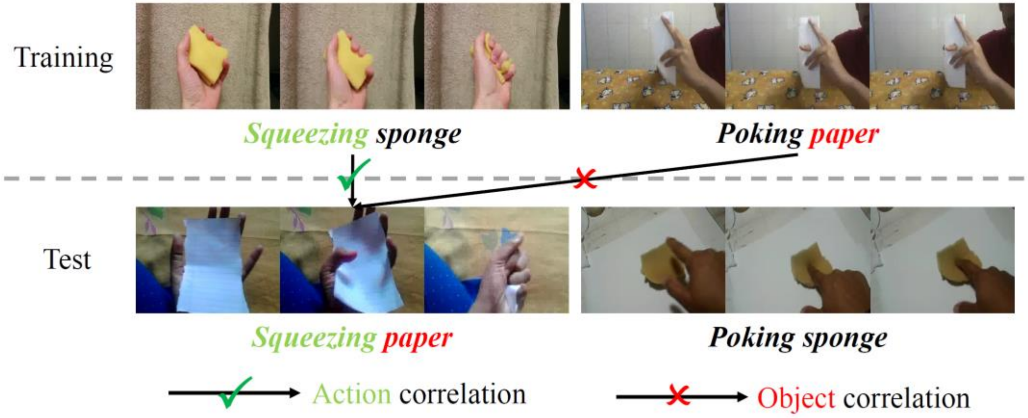
University of Electronic Science and Technology of China
&
MIT-IBM Watson AI Lab

Introduction

Compositional action recognition[1]:

- ➢ Motivated by the appearance bias
- ➢ Learn real action knowledge

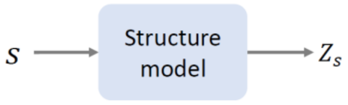
- **How to split**
  - Actions -> Group 1 Group 2
  - Object -> Group A Group B
  - Training set {*1+A, 2+B*}
    validation set {*1+B, 2+A*}
- **Target**： Recognize action with unseen object appearance



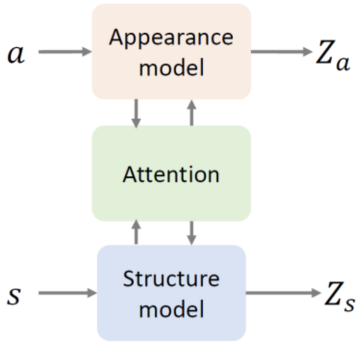
[1] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. 2020. Something-Else: Compositional Action Recognition With Spatial-Temporal Interaction Networks. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1046–1056.

Related Work

- Previous Work
  - Decreasing the dependency on instance appearance: STIN[1]
  - Fusing visual information with structure information: SAFCAR[2] and Interactive Fusion[3]

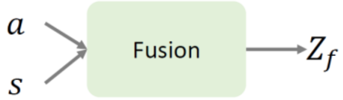


STIN



SAFCAR

a: appearance information
s: structure information



Interactive Fusion

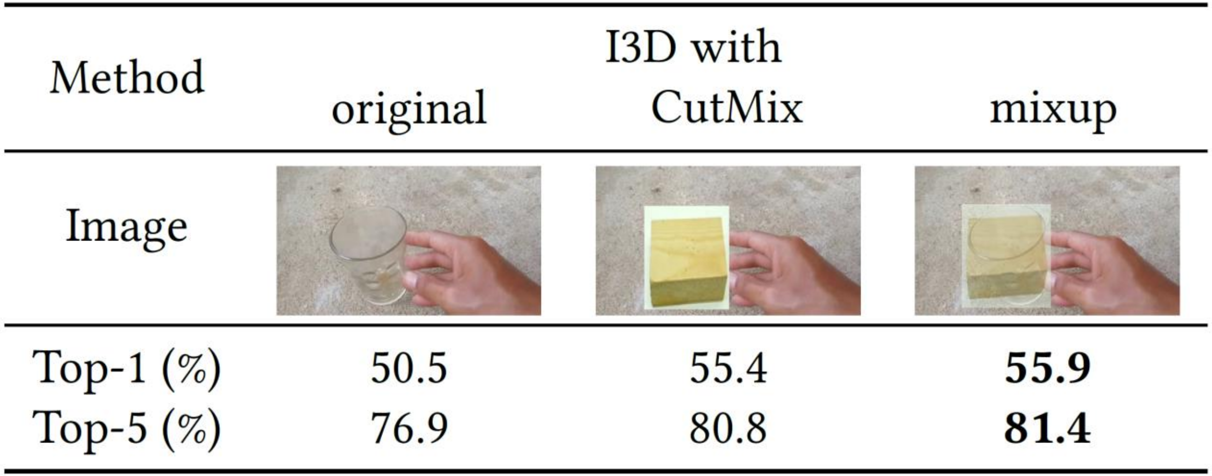
[1] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. 2020. Something-Else: Compositional Action Recognition With Spatial-Temporal Interaction Networks. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1046–1056.

[2] Tae Soo Kim and Gregory D. Hager. 2020. SAFCAR: Structured Attention Fusion for Compositional Action Recognition. abs/2012.02109 (2020). arXiv:2012.02109

[3] Rui Yan, Lingxi Xie, Xiangbo Shu, and Jinhui Tang. 2020. Interactive Fusion of Multi-level Features for Compositional Activity Recognition. arXiv:2012.05689

Novelty

- Shortcoming of Existing Methods
  - Ignore the negative effect introduced by instance appearance bias
  - Obvious improvement achieved after breaking visual correlation by CutMix[4] and mixup[5]



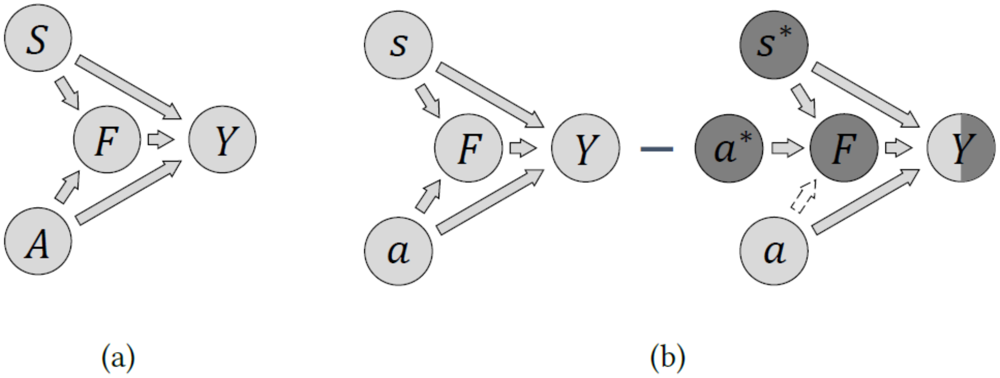
Method		I3D with	
	original	CutMix	mixup
Image			
Top-1 (%)	50.5	55.4	**55.9**
Top-5 (%)	76.9	80.8	**81.4**

[4] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[5] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017).

Novelty

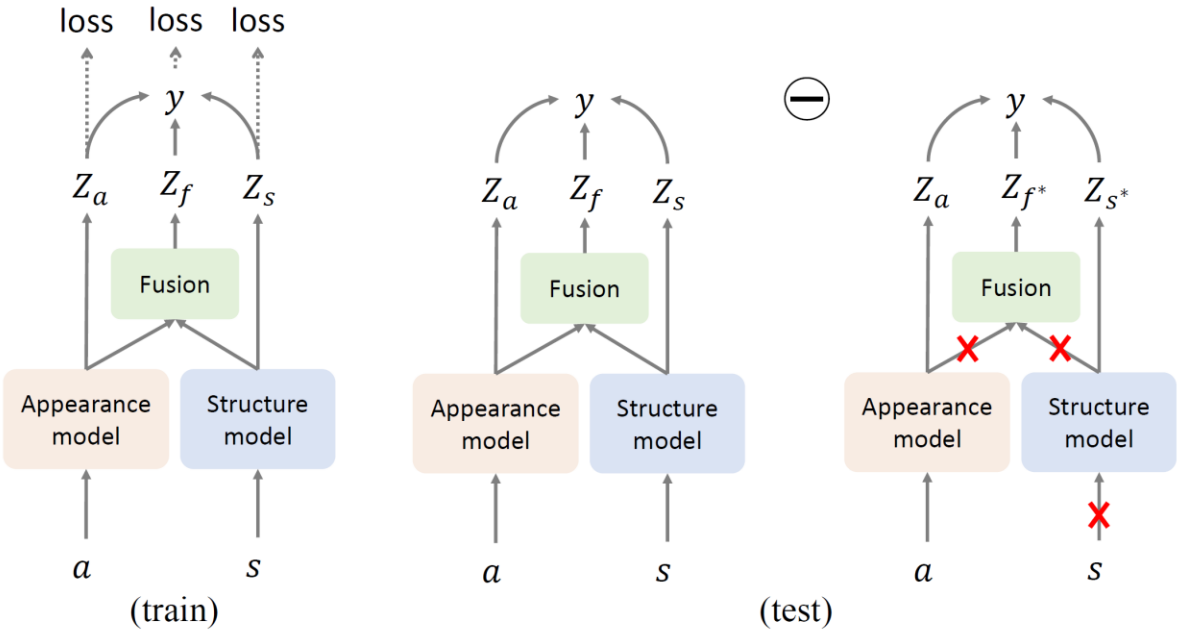
- Causal graph for compositional action recognition
- Counterfactual Debiasing Network (CDN)



S: structure information
A: appearance information
F : fusion information
Y: prediction scores
Light node: real value input
dark node: dummy value input

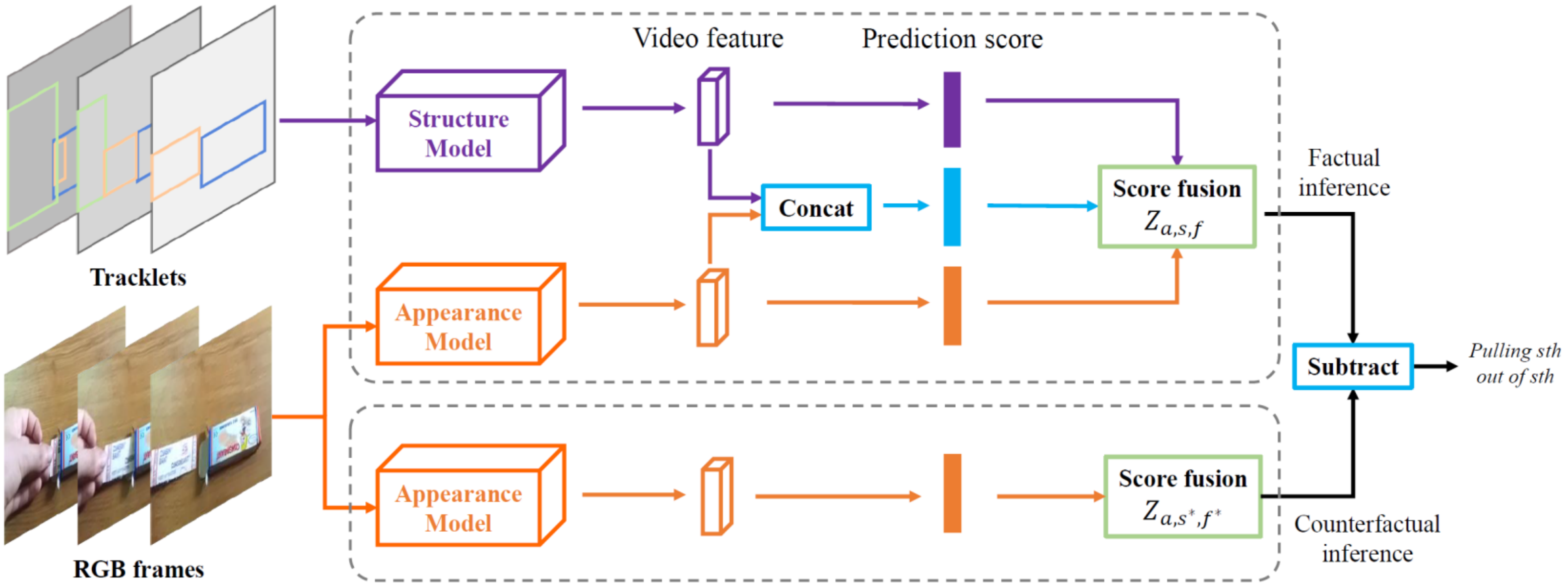
Proposed model

- Design



Proposed model

- Architecture
  - No strict requirements in the specific implementation
  - The factual outcome: fusion of three branch activation
  - The counterfactual outcome: fusion of one branch activation and two zero values



Proposed model

- Implementation
  - Appearance model output: Z_a
  - trajectory model output: Z_s
  - fusion model output: Z_f
  - The fusion function's output: $Z_{a,s,f} = h(Z_a, Z_s, Z_f)$
  - The fusion function: $h(\cdot) = \log(\sigma(sum(\cdot)))$

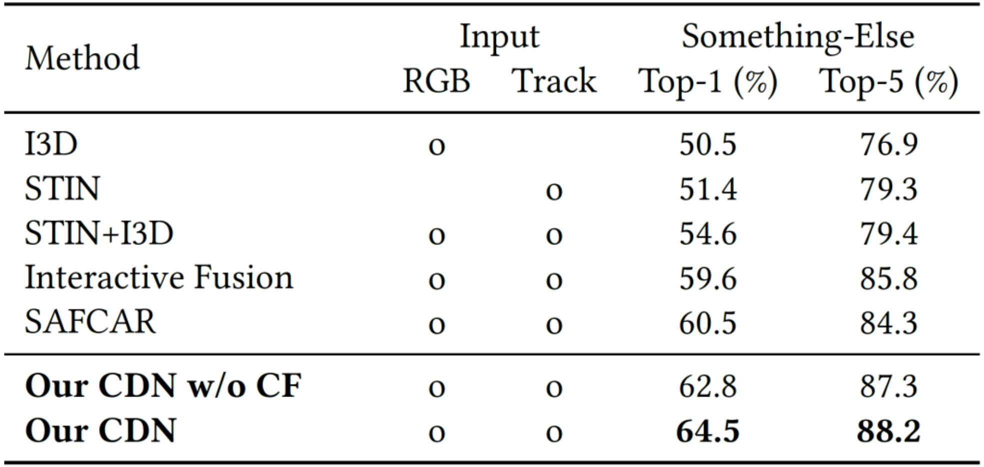
  - Choose total indirect effect as our criterion:

$$TIE = Z_{a,s,f} - Z_{a,s^*,f^*} \approx Z_{a,s,f} - \alpha \cdot Z_{a,s^*,f^*}$$

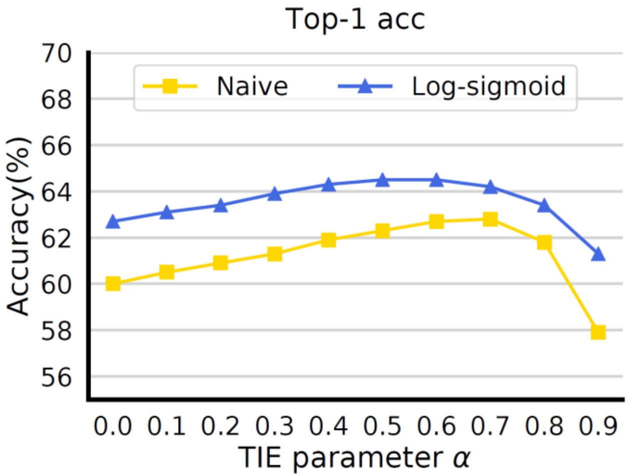
The subscribe * represents the null input

The parameter α represents the effect weight we remove.

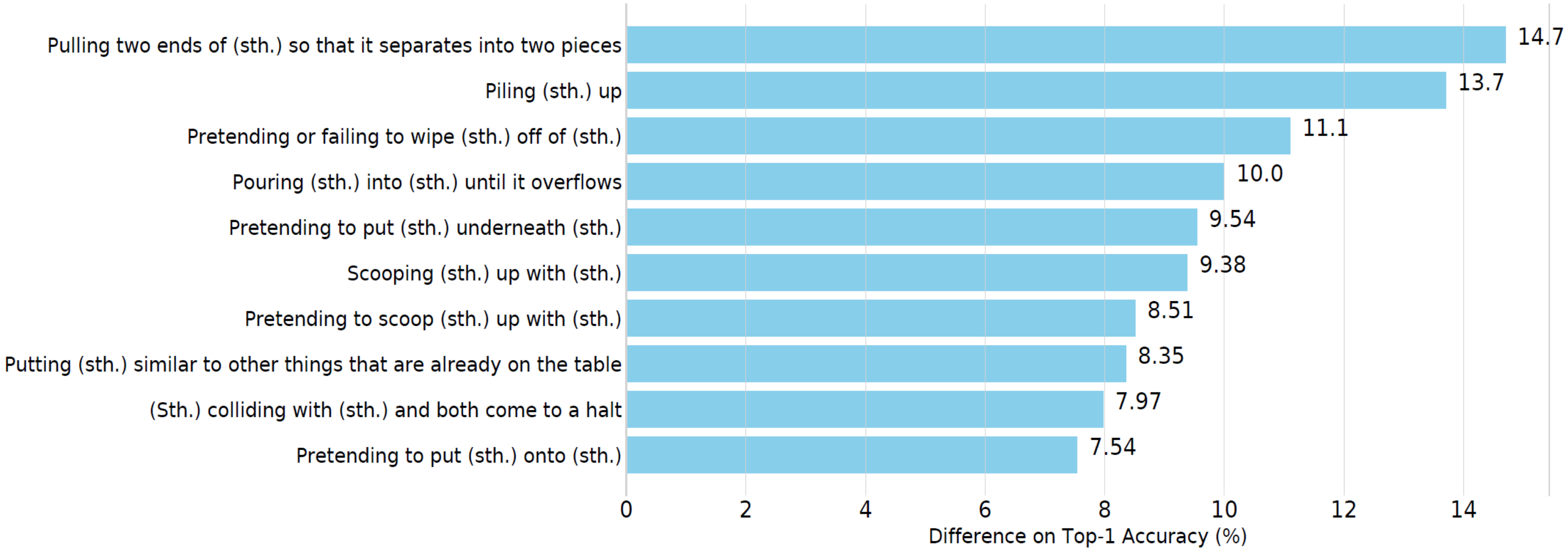
Result & Ablation



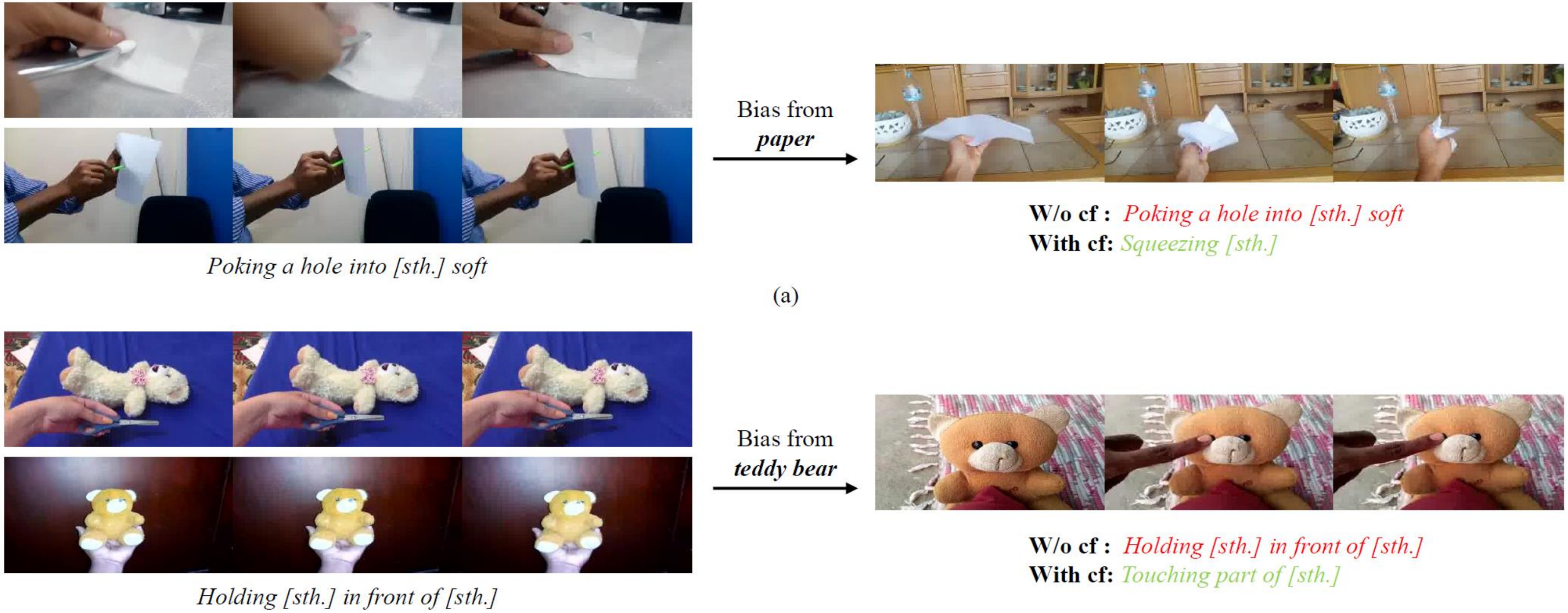
Method	Input		Something-Else	
	RGB	Track	Top-1 (%)	Top-5 (%)
I3D	o		50.5	76.9
STIN		o	51.4	79.3
STIN+I3D	o	o	54.6	79.4
Interactive Fusion	o	o	59.6	85.8
SAFCAR	o	o	60.5	84.3
Our CDN w/o CF	o	o	62.8	87.3
Our CDN	o	o	**64.5**	**88.2**



Result & Ablation



Case study



Poking a hole into [sth.] soft

Bias from ***paper***

W/o cf : *Poking a hole into [sth.] soft*
With cf: *Squeezing [sth.]*

(a)

Holding [sth.] in front of [sth.]

Bias from ***teddy bear***

W/o cf : *Holding [sth.] in front of [sth.]*
With cf: *Touching part of [sth.]*

(b)

Summary

- We observe that prior knowledge learned from appearance information is mixed with the spurious correlation between action and instance appearance, which badly inhibits the model's ability of action learning.

- We remove the pure appearance effect from total effect by counterfactual debiasing inference on our novel framework CDN proposed for compositional action recognition.

- We achieve state-of-the-art performance for compositional action recognition on the Something-Else dataset.

Thanks for Your Attention!