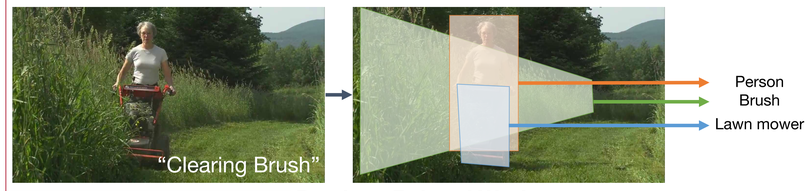
Attentional Pooling for Action Recognition

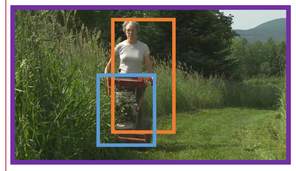
Rohit Girdhar and Deva Ramanan

Carnegie Mellon University — The Robotics Institute

Code & Models — NIPS 2017

Recognizing fine-grained actions





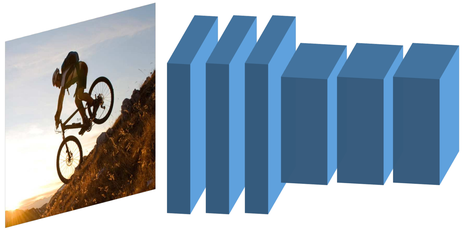
Previous Related Work

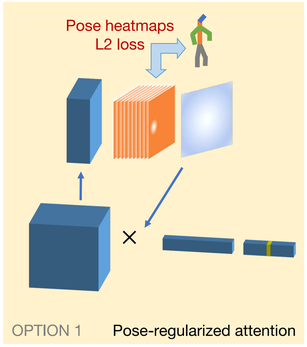
- Sharma *et al.*, ICLR-W'15 uses recurrent attention model, shows minor improvements
- R*CNN (Gkioxari *et al.*, ICCV'15) focus on features from the person and context boxes
- Mallya and Lazebnik, ECCV'16 simplify it to only use full image as the context

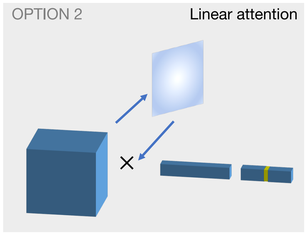
Recent methods hard-code attention on the human

Our Model

Jointly predict an attention map to modulate the final convolutional feature before spatial average pool. The attention map can be unconstrained, or regularized using human pose keypoints.





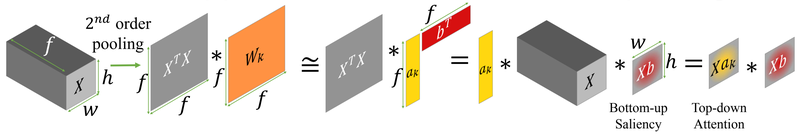


Improved action recognition performance!
(MPII human-pose/action dataset)

Model (ResNet-101)	Val mAP
No attention	26.2%
Linear attention	30.3%
Pose-regularized attention	**30.6%**

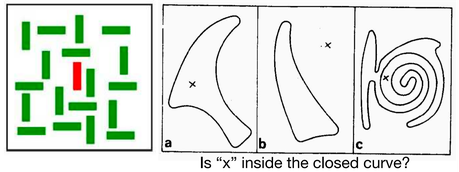
What makes the model work?

Rank-1 approximation of second order pooling



Inspiration from Human Vision

- Bottom-up saliency: Certain image regions seem to pop-out
- Top-down task-guided attention: Similar to Ullman's *Visual Routines* (Cognition, 1984), extract desired information from base representation modulated by saliency



Is "x" inside the closed curve?

Why not only top-down attention?
Standard average pool models already have top-down attention and adding bottom-up saliency improves performance significantly

Method (HICO dataset)	Val mAP
ResNet-101 (in-built top-down attention)	30.2%
Ours, adding bottom-up saliency to above	**35.0%**

Comparison with the state of the art

MPII pose dataset
Classify into one of 393 action classes

Method	Test mAP
R*CNN (ICCV'15)	26.7%
Mallya and Lazebnik (ECCV'16)	31.9%
Ours (Linear attention)	36.0%
Ours (Pose regularized attention)	**36.1%**

HICO dataset
Detect 600 human-object interactions

Method	mAP
R*CNN	28.5%
Mallya & Lazebnik (w/o wtd loss)	33.8%
Ours (Linear attention)	**35.0%**

HMDB-51 dataset →
Classify short-video clips into 51 action classes. We train a frame-level model (RGB stream from two-stream architectures) with attentional pooling. At test time, we average predictions from uniformly sampled frames

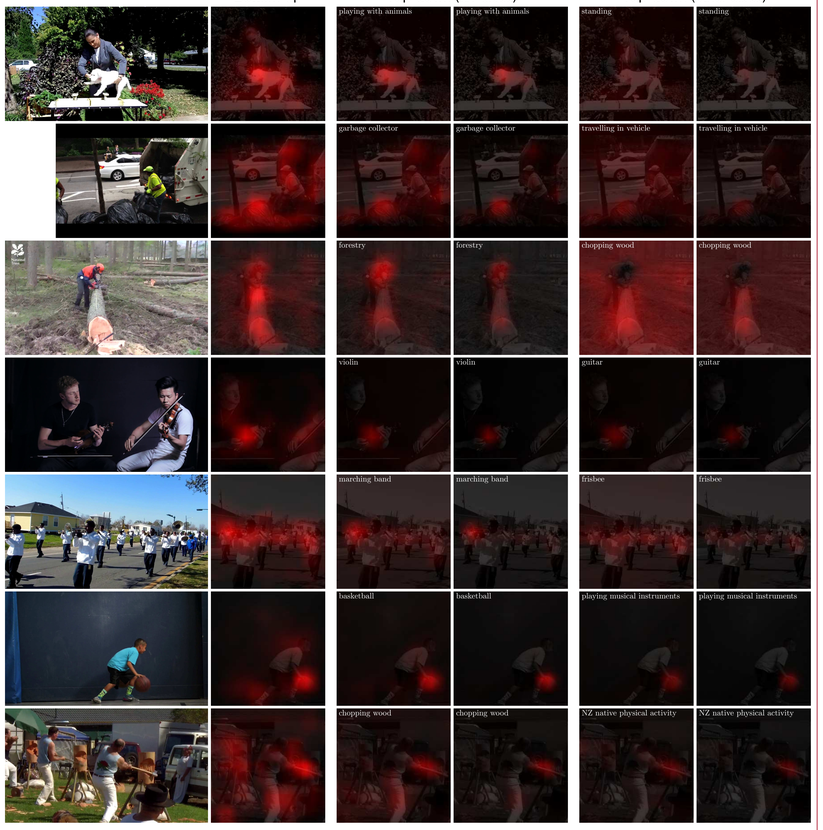
Method (RGB stream only)	Accuracy
TSN (ECCV'16)	51.0%
ResNet-152 (NIPS'16)	46.7%
TSN, ResNet-101 (ours)	47.1%
Ours (Linear attention)	50.8%
Ours (Pose reg. attention)	**52.2%**

Conclusion

- Consider replacing pooling with attentional pooling. Lightweight yet powerful!
- Connection to bilinear pooling suggests action recognition ~ fine-grained task!
- Self attention out-performs sequential attention. Our approach is an efficient implementation for self-attention.

Where does the model look?

Validation images in MPII that obtain strongest improvement in performance



The model learns to look for objects to recognize interactions