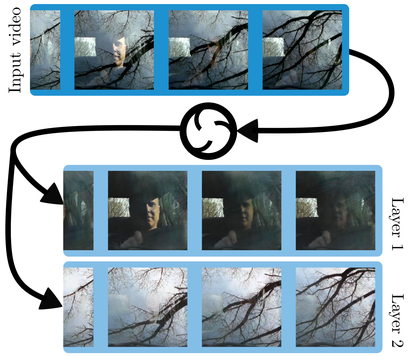
The Visual Centrifuge: Model-Free Layered Video Representations

Jean-Baptiste Alayrac*,[1], Joao Carreira*,[1] and Andrew Zisserman[1,2]

*equal contribution, [1]Deepmind, [2]University of Oxford

DeepMind

Overview



The Visual Centrifuge in a Nutshell

Goal: learn a model to decompose videos into multiple layers [Wang and Adelson].

How can we do it? by using both *semantic* and *motion* cues.

Approach: purely learning based using a simplistic *data generation scheme*, an adapted *loss* and a standard *network architecture*.

Conclusion

- The model is able to unmix artificially blended videos.
- Ablation study indicates that network uses both semantics and motion.
- Interestingly, the model can also be applied to real-world layered composition phenomenon.

References

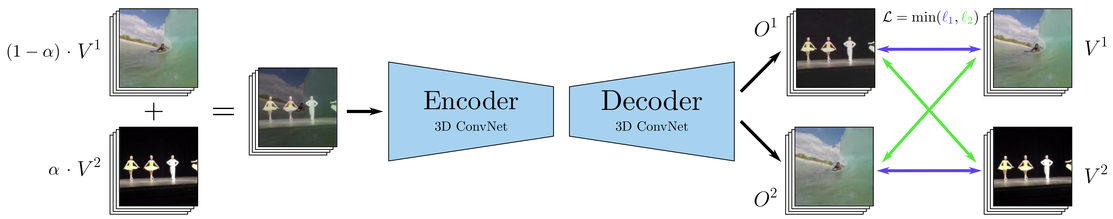
Representing moving images with layers, Wang and Adelson, *Transaction on Image Processing*, 1994.

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, Carreira and Zisserman, *CVPR*, 2017.

Deep multi-scale video prediction beyond mean square error, Mathieu et al, *ICLR*, 2016.

A computational approach for obstruction-free photography, Xue et al., *SIGGRAPH*, 2015.

The Approach



1 Data Generation

- Simply average two videos in pixel space
- Similar to reflection layer composition
- Ground Truth for free!
- Kinetics-600 train set: 100,000,000,000 unique pairs

3 Permutation Invariant Loss

Challenge: Permutation label problem

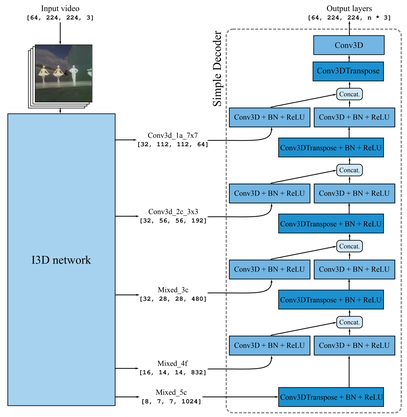
Solution: Permutation Invariant Loss:

$$\mathcal{L}\left(\{V^1, V^2\}, O\right) = \min_{(i,j)|i\neq j} \ell(V^1, O^i) + \ell(V^2, O^j)$$

with simple pixel wise reconstruction loss [Mathieu et al.]:

$$\ell(U, V) = \frac{1}{2T}\left(\sum_t \|U_t - V_t\|_1 + \|\nabla(U_t) - \nabla(V_t)\|_1\right)$$

2 The Visual Centrifuge: key ingredients

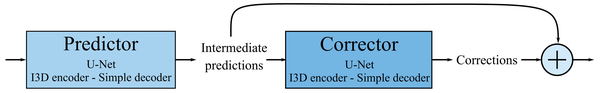


Base model (Predictor) : - I3D network encoder [Carreira and Zisserman],
- Simple decoder,
- U-Net like skip connections.

Two modifications are considered:

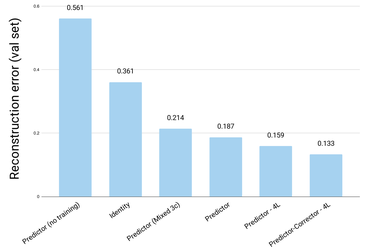
a Multiple Output Layers
- Even if videos are composed of only two layers, we allow the network to produce n layers (with n > 2).
- Why? help deal with ambiguities in the outputs.

b Predictor-Corrector



Experiments using the blending procedure

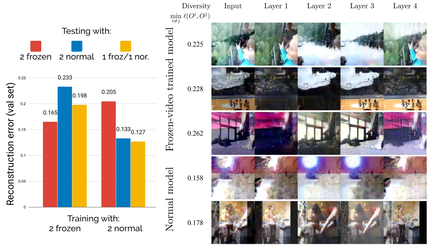
Model design choices.



Qualitative samples on Kinetics Validation set.

Conclusion: deeper model, more outputs and predictor-corrector architecture matter.

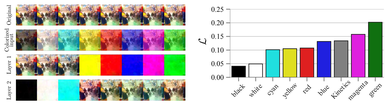
Motion vs. Static cues ablation study.



Low-level vs. high-level features:

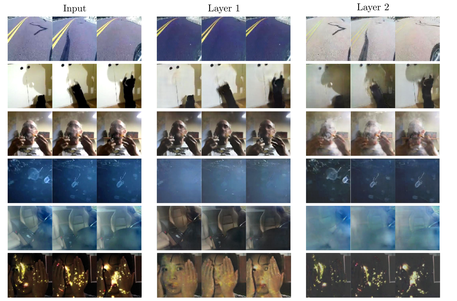
- Mixed 5c > Mixed 3c (deeper is better)
- Unmixing videos of the same class is harder (0.145 vs 0.133)
- More correlation between unmixing perf. with high level feat. than with low level feat.

Color experiment.

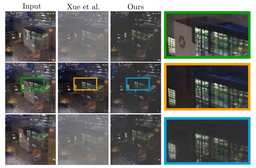


Real World Applications

Qualitative samples on real world videos.



Comparison to engineered method.



Action recognition in challenging scenario (recognizing action in blended videos)

- I3D off-the-shelf: 22% acc.
- Centrifuge + I3D off-the-shelf: 44% acc.