Semantic Video CNNs Through Representation Warping

Raghudeep Gadde¹, Varun Jampani² and Peter V. Gehler³
¹Max Planck ETH Center for Learning Systems ²NVIDIA ³Amazon

The “NetWar” module turns image CNNs into video CNNs with tiny computational overhead.

1 Semantic Video CNNs

- Significant improvements in semantic image segmentation with CNNs.
- But data comes in sequence of frames, videos.

How do we extend existing image CNNs to videos?

Easy Solution: Frame wise CNN + Optical Flow

Motivation

- Clockwork Convnets: Intermediate CNN representations change only slowly over adjacent frames, especially for deeper CNN layers [1].
- Bilateral Inceptions: Averaging intermediate CNN representations for locations across the image that are similar close improves performance [2].
- Does combining temporally close representations also lead to stable and consistent semantic predictions?

2 From Image CNNs to video CNNs with ‘NetWar’

Main Idea: Use optical flow of adjacent frames to warp intermediate CNN representations

- Three stages in a NetWar module:
  1. Transform input optical flow using a small CNN.
  2. Warp previous frame representation.
  3. (Linearly) Combine the warped representations to those of the present frame.
- Multiple NetWar modules at different depths in a CNN.

Advantages:

- Fast
- End-to-end trainable
- Online

3 Experiments: CamVid Dataset

4 Experiments: Cityscapes Dataset

5 Effect of Optical Flow

- Two recent optical flow techniques.
  - DIS-flow (fast and less accurate)
  - FlowFields (slow and accurate)
- Almost identical IOU performance.
- Traditional optical flow might not be the best suited transformation.

- Radical change in transformed optical flow.
- Scene structure is much more pronounced in the transformed flow.

6 Conclusion

NetWar is a fast, efficient and easy way to turn an image CNNs into a video CNN.

References


Code: http://segmentation.tuebingen.mpg.de