CFM AI&VR Weekly Report: Week #24 of 2019

Zhang Yifei(yidadaa@qq.com)

UESTC — June 13, 2019

1 Summary

This week, I spent two days on practicing to use pytorch framework, and wrote and trained a CNN-RNN model on UCF-101 datasets. At the same time, I took a review on the classical methods in depth estimation task, the reaction paper in this report needs your advice and guidance.

2 Code Practice: Implement LRCN using Pytorch Framework

LRCN[2] is a classical model for visual recognition and description, it is an end-to-end trainable and suitable Long-term Recurrent Convolutional Networks on action recognition tasks. To practice to use and be acquainted with pytorch framework and workflow, I spent two days implementing the model using it. I use a ResNet-152 network as CNN-Encoder, and a 3-layers GRU network as RNN-Decoder. I trained it on a subset of UCF-101 datasets, which is limited by computing resource. The model achieves a pretty good performance on the tailored UCF datasets(90% accuracy).

The code is available at `https://github.com/Yidadaa/Pytorch-Video-Classification`.

3 Reaction Paper of this Week

Predicting scene depth from input imagery is important for robot navigation, which is especially important in SLAM system. Casser. [1] trys to solve the problem by explicitly modeling 3D motions of moving objects, together with camera ego-motion. The proposed approach introduces structure in the learning proces by representing objects in 3D and modeling motion as SE3 transforms. Two models are introduced to model the ego-motion of the camera and predict the depth map of image sequences. The main idea of the approach is predicting motions of individual objects in 3D, which leads to handle the dynamic scenes well.

Intriguingly, a reconstrction loss is computed as the minimum reconstrction loss between wraping from either the previous frame or the next frame into the middle one:

$$L_{rec} = min(||\hat{I}_{1\to 2} - I_2, \hat{I}_{3\to 2} - I_2||)$$

Casser claims that the reconstruction loss is proposed by Godard[5], but a similar method could be traced back to Engel's work[3], which is also used at LSD-SLAM[4]. To refine the estimated inverse depth map from frame to frame, Engel proposes the method[3] to calclate and project the current corresponding 3D point into the new frame, and the new inverse depth d_1 could be approximated by:

$$d_1(d_0) = (d_0^{-1} - t_z)^{-1}$$

where t_z is the camera traslation along the optical axis and d_0 is the depth map of the current frame. This method is successfully applied to refine the depth map as an important part of LSD-SLAM pipeline[4].

The reconstrction loss is not the only loss to refine the depth map at [1], the total loss is applied on 3 scales with hyperparameters α_j:

$$L = \alpha_1 \sum_{i=0}^{3} L_{rec}^i + \alpha_2 L_{ssim}^i + \alpha_3 \frac{1}{2^i} L_{sm}^i$$

where the L_{ssim} denotes SSIM loss[6] and the L_{sm} denotes depth smoothness loss[5].

The approach proposed by Casser[1] is conceptually deifferent from prior works which is used optical flow for motion in 2D image space[7] in that the object motions are explicitly learned in 3D and are available at inference.

However, the approach[1] predicts the depth map from a single RGB frame and refine it with another depth map which is predicted from the next frame, it still needs further study to confirm whether it's the best way to predict depth map.

4 Plans

This week, I have taken a deep insight in reconstruction loss, which is an important part of depth estimation. There are still more questions to study, for example, the history of SSIM loss[6] and smoothness loss[5], I will try to retrospect the history of these two losses.

I will write one reaction paper per week in the next few months, please read and give your feedback to me on time.

Get more information about **Reaction Paper** at: `https://zhuanlan.zhihu.com/p/33875687`.

References

[1] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth Prediction Without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos. *arXiv:1811.06152 [cs]*, November 2018. arXiv: 1811.06152.

[2] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[3] J. Engel, J. Sturm, and D. Cremers. Semi-dense Visual Odometry for a Monocular Camera. In *2013 IEEE International Conference on Computer Vision*, pages 1449–1456, December 2013.

[4] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-Scale Direct Monocular SLAM. In *ECCV*, 2014.

[5] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611, Honolulu, HI, July 2017. IEEE.

[6] Zhou Wang, Alan Bovik, Hamid Rahim Sheikh, and Eero Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *Image Processing, IEEE Transactions on*, 13:600–612, May 2004.

[7] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, Salt Lake City, UT, June 2018. IEEE.