Summary of CVPR 2017

Jun Gao
Peking University

July 28 2017

1 Introduction

I'm very lucky to have an opportunity to attend the conference of Computer Vision and Pattern Recognition (CVPR) this year. Hawaii is a great place, both for traveller and scholar. I would like to write down my own experiences in this conference and share them with others.

To beginning, I want to claim that all of the following is my own opinion, there exists bias because of my own research interests. I will not cover too much about some computer vision application tasks, such as Optical Flow, Person Re-Identification, Face Recognition, Computational Photography, Motion Tracking etc.

In Summary, although not every paper is great enough, CVPR this year gives us many directions of current Computer Vision research community. What impressed me most are the works about self-supervised / unsupervised / physical constraints. Also the Multi-modal, 3D deep learning, reinforcement learning and the combination between computer graphics and computer vision are very popular.

2 Model-Based Constraints

I came to know this idea in workshop of the Deep Learning for Robotics Vision. Prof. Honglak Lee and Prof. Dirter Fox introduced this idea on their works. From a high-level view, it wants to find the constraints in the real prediction process, then utilize these constraints as loss function or network architecture. E.g. In one of Prof. Honglak Lee's paper [20], we have a prior that 2D images are from the projection of 3D volume, therefore, we could project one 3D Volume and get multi-view 2D images, these images should be the same as input images (and ground truth). Then the L2 Loss could be applied to the projection result and input images. [17] is an improvement work by Tinghui Zhou, it used ray tracing instead of projection to reconstruct the image.

Also, the Prof. Dieter Fox introduced [3], where the network only predicts the mask and transformation, then we could calculate the position of each object and reconstruct the original image.

Other paper relates to this topic is [16]. You could read these papers if you are interested in :)

3 Multi-Modal

Whether learning multiple tasks simultaneously could improve performance of every task? How to combine multiple modals such that we could gain improvement from the combination? Prof. Gupta introduced this in the workshop of the Deep Learning for Robotics Vision, too. He found robots those trained from multi-tasks could perform better than those trained from single one.

[11] uses knowledge graph to improve the classification performance, it propagate nets over the knowledge graph to get the final output of the knowledge, then add into the classification.

[10] is an adaptive method to do multi-task training, the network learns to decompose tasks, mainly by the task similarity and task grouping. The way it defines the similarity is based on the easy and hard data at different task.

[12] finds a way to composite different tasks (in this paper, it composite attribute and objects). Basically, it pretrained SVM for each attribute and object, and then train a neural network to predict the combined SVM weights for each kind of combination.

[8] is a network that could do almost every general tasks relates to vision.

4 Reinforcement Learning

Applying reinforcement Learning to many vision tasks: Image captioning, Object relationship. Basically, we need to formulate the problem and define reward function.

[14] used Reinforcement Learning to do image captioning. Actor-Critic Method has a benefit that avoid beam search, we could use critic to get the expected value (reward) if generate this word currently.

[9] used reinforcement learning to iteratively predict the relationship between subject and object, next object and attribute of the subject. Q-Learning and LSTM

5 Computer Graphics

I tried my best to understand these papers, but unfortunately :(......

Inspired from the section 2, if we want to find the constraints, we have to consider the graphics, e.g. rendering procedure for 2D image [17].

6 3D Deep Learning

3D data has multiple representation format: volume, point cloud, mesh etc. For volume data, it's conventional to use 3D CNN....

...

I find many papers relates to 3D, but due to the reason that Leo's lab already has many work on this field[13][6], I did not care too much about these papers.

7 Iterative Optimization

[21] iteratively get the classification result by treating the current feature map as the input the network.

[15] iteratively find the ground truth of unlabelled data.

8 Others

I'm interested in the ideas in these papers.

[4] defined Part Affinity Fields to find the correspondence among the keypoints in one person.

[19] plugs in a subnetwork that generates hard example to fool original network (Fast RCNN) by masking feature maps. This could be regarded as a more intelligent dropout.

[2] finds a way to quantify the interpretability of network. Interestingly, some neurons could be regarded as the feature detector (just like Canny Edge Detector). Neurons may have a response to certain object or certain part of a object.

[1] redefines the pipeline to do instance segmentation, it defines an energy map, which have zero value in the edge and background, and energy grows as deeper into an object. Then the network only need to predict this energy map.

[18] utilized the attention module for classification, it predicted the residual attention for the feature map. More interestingly, they do not apply any supervision for attention, but the network could learn the attention in a reasonable manner. One more things, we could not apply semantic segmentation supervision here, because semantic segmentation could not be used to attention on the feature map directly.

[5] uses an polygan to surround one instance, then the network only need to predict this polygan.

[7] learns to do Non-Maximum Suppression.

References

[1] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. *arXiv preprint arXiv:1611.08303*, 2016.

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. *arXiv preprint arXiv:1704.05796*, 2017.

[3] Arunkumar Byravan and Dieter Fox. Se3-nets: Learning rigid body motion using deep neural networks. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 173–180. IEEE, 2017.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016.

[5] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. *arXiv preprint arXiv:1704.05548*, 2017.

[6] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image. *arXiv preprint arXiv:1612.00603*, 2016.

[7] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. *arXiv preprint arXiv:1705.02950*, 2017.

[8] Iasonas Kokkinos. Ubernet: Training auniversal'convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. *arXiv preprint arXiv:1609.02132*, 2016.

[9] Xiaodan Liang, Lisa Lee, and Eric P Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. *arXiv preprint arXiv:1703.03054*, 2017.

[10] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. *arXiv preprint arXiv:1611.05377*, 2016.

[11] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844*, 2016.

[12] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. 2017.

[13] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016.

[14] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. *arXiv preprint arXiv:1704.03899*, 2017.

[15] Changshui Zhang Renping Cui, Hu Liu. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization.

[16] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. *arXiv preprint arXiv:1704.04131*, 2017.

[17] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. *arXiv preprint arXiv:1704.06254*, 2017.

[18] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. *arXiv preprint arXiv:1704.06904*, 2017.

[19] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. *arXiv preprint arXiv:1704.03414*, 2017.

[20] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 1696–1704, 2016.

[21] Amir Roshan Zamir, Te-Lin Wu, Lin Sun, William Shen, Jitendra Malik, and Silvio Savarese. Feedback networks. *CoRR*, abs/1612.09508, 2016.