Feature voxel grid를 적용한 RGB-D 물체 인식 뉴럴 네트워크

*김가영, 윤성욱, 설상훈
고려대학교 전기전자공학부
e-mail : *gykim@mpeg.korea.ac.kr, swyoon@mpeg.kroea.ac.kr, sull@korea.ac.kr*

RGB-D Object detection neural network using feature voxel grid

*Ka-Young Kim, Seong-Wook Yoon, Sang-Hoon Sull
School of Electrical Engineering
Korea University

Abstract

Neural Networks have been widely used for object recognition. Most of neural networks are using RGB dataset, but some are using RGB-D dataset for better performances. There have been lots of methods to improve performances of neural networks that exploit RGB-D dataset by using additional hand descriptor or only deep convolutional networks. In this paper, to fully exploit 3D spatial information of RGB-D dataset, we propose a method using Depth to present spatial information of RGB features.

I. 서론

뉴럴 네트워크를 이용한 영상 기반 물체 인식에 대한 연구가 활발하게 진행되고 있다. 그 중, RGB-D 이미지의 경우, depth를 활용하여 물체의 공간 정보를 얻을 수 있다. 물체의 공간 정보를 활용하여 물체의 인식률을 향상시키기 위한 방법은 hand-designed 방법[1] 또는 뉴럴 네트워크를 사용하는 방법[2, 3]이 있다. 특히 [3]의 경우, depth 정보를 이용하여 RGB를 3차원적으로 encoding하는 방법을 사용한다.

본 논문에서는 depth 정보를 이용하여 RGB가 아닌 학습된 특징값을 3차원적으로 encoding하는 방법을 제시하고 이를 이용한 새로운 네트워크 구조를 제안한다.

II. 본론

2.1 기존의 기술

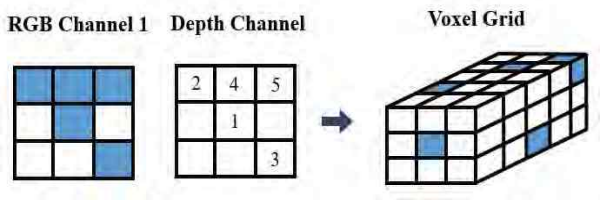

그림 1 Depth를 활용한 voxel grid encoding 방법

[3]에서, VGG3D는 그림1과 같은 방법으로 만든 RGB voxel grid를 pretrained VGG의 입력으로 사용하기 위해 만든 구조이다. 구체적으로는, 그림 1에서, quantized depth를 이용하여 RGB의 각 채널의 값을 3차원적 위치에 대입한다.

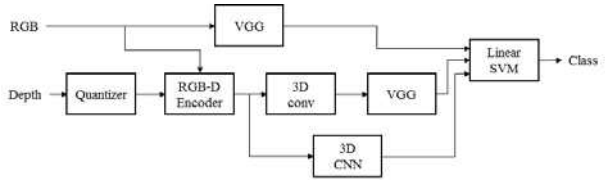

그림 2 기존방법[3]의 뉴럴 네트워크 구조

또한, 그림 2와 같이, RGB만을 사용하는 pretrained VGG, pretrained VGG 앞부분에 3D convolution을 사용한 VGG3D와 3D CNN을 각각 따로 학습한 후, linear SVM을 사용하여 3개의 특징값으로 물체 인식을 수행하였다. 이때, Quantizer는 전체 데이터셋에서의 출력들의 개수가 같도록 정해진다.

2.2 제안하는 방법

본 논문에서는, VGG의 중간 특징값을 3차원적으로 encoding하여 feature voxel grid를 생성하고자 한다. 중간 특징값의 depth를 구하기 위해, 중간 특징값을 추출하기 위해 사용된 depth의 비율을 계산하였다. 구체적으로, VGG의 2D convolution 연산에 대해서는 식 1에 의한 2D convolution 연산을 수행한다. 2D convolution에서 입력이 출력에 기여하는 바를 도함수의 절대값으로 정의하여 계산하였다.

$$r_{k,l}^{conv} = \sum_{i,j} \left| \frac{\partial y_{k,l}^{conv}}{\partial x_{k+i,l+j}} \right| r_{k,l} = \begin{cases} \sum_{i,j} \left| W_{i,j}^{conv} \right| r_{k,l}, & y_{k,l}^{conv} > 0 \\ 0, & else \end{cases} \quad \text{(식 1)}$$

VGG의 max-pooling 연산에 대해서는, pooling된 값들의 위치와 동일한 위치의 depth를 중간 특징값으로 취급하여 pooling 연산을 수행한다.

2.3 제안하는 구조

위와 같은 방법으로 생성된 feature voxel grid를 이용하여 물체 인식을 수행하는 네트워크 구조는 그림3과 같다.

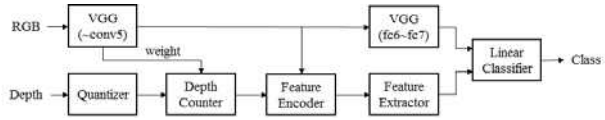

그림 3 제안하는 구조

RGB의 경우, VGG의 conv5까지의 레이어들을 거쳐 이미지의 중간 특징값을 생성한다. Depth의 값들은 Quantizer를 통해 양자화되는데, 양자화 간격은 총 6개의 일정한 간격으로 이루어져 있다. 양자화된 depth는 VGG의 위에서 설명한 연산들로 이루어진 Depth Counter를 거쳐 depth의 비율로 계산된다. 이를 이용하여, Feature Encoder에서는 feature voxel grid를 얻을 수 있다.

RGB의 중간 특징값은 앞서 계산되지 않은 fully connected 레이어들을 거치고 feature voxel grid는 다른 구조의 Feature Extractor를 거친다. 이후, 두 개의 출력을 합쳐 Linear Classifier로 class를 얻는다.

기존 방법과 달리 3차원적 encoding을 네트워크의 중간에서 수행함으로써, 제안하는 구조는 크게 3가지의 차이점을 가진다. 첫째, 2D convolution을 사용한다. 둘째, 동시에 학습될 수 있다. 셋째, 전체 데이터셋을 고려하지 않고 일정한 간격을 사용한다.

Ⅲ. 결과

제안된 구조를 구현하기 위해, TensorFlow[4]를 이용하였고 환경 구성은 GTX 1080Ti를 사용하였다. 또한, 데이터셋은 기존 기술 [3]에서 사용한 washington RGB-D Object Dataset [5]을 사용하였다.

기존 방법과 제안하는 방법의 성능을 비교하기 위해, 각기 다른 baseline으로부터의 성능 증가량을 비교했다. 특히, Linear SVM으로 얻는 이득을 제외하여, 3차원 encoding 방법의 차이만을 고려하고자 했다. 그럼에도 불구하고, 기존 방법 [3]에서, VGGnet을 fine-tune하지 않았다는 점과 VGG3D의 효과보다 더 큰 3D CNN의 효과가 포함되었다는 점 때문에 공정한 비교가 되기 힘들다.

제안하는 방법의 baseline은 pretrained VGG fc8만 fine-tune한 결과이다.

	VGG (baseline)	Ours	Increment
trial 1	66.97	76.0131	9.0431
trial 2	74.2112	81.6295	7.4183
trial 3	71.8219	76.8726	5.0507
trial 4	70.5493	79.5634	9.0141
trial 5	73.8038	77.9009	4.0971
trial 6	76.4037	78.2847	1.881
trial 7	78.663	80.0461	1.3831
trial 8	75.3481	79.4828	4.1347
trial 9	73.3568	81.886	8.5292
trial 10	78.5562	84.4277	5.8715
Mean	**73.9684**	**79.61068**	**5.64228**

표 1 trial별 각 방법들의 정확도 및 증가량

Method	RGB (baseline)	RGB-D	Increment
[3]	88.96±2.1	91.84±0.89	2.88

표 2 기존방법[3]의 방법의 정확도

본 논문에서 제안하는 방법의 경우 RGB-D 데이터의 정확도가 RGB 데이터를 사용하였을 때 보다 약 5.6%의 증가율을 보인다. 이는 기존 방법[3]보다 높은 증가량임을 볼 수 있다.

Ⅳ. 결론 및 향후 연구 방향

본 논문에서는, RGB-D 이미지를 사용하기 위한 새로운 방법으로써, RGB에 대한 중간 특징값의 depth를 추정하고 이를 이용한 3차원 encoding 방법을 제시했다. RGB-D 데이터를 사용한 제시한 방법은 RGB 데이터를 사용한 VGG의 성능보다 높은 정확도를 보여주었다.

본 논문에서는 VGG와 Linear Classifier를 사용했지만, VGG가 아닌 네트워크를 사용하거나 Linear SVM을 이용하여 다양한 구조에 대한 성능 증가를 확인해 볼 수 있다.

Acknowledgement

"본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음" (IITP-2018-2016-0-00464)

참고문헌

[1] Gupta, S., Girshick, R., Arbeláez, P., &Malik, J. (2014, September). Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision* (pp. 345-360). Springer, Cham.

[2] D. Maturana and S. Scherer. Voxnet: A 3D convolutional neural network for real-time object recognition. In Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, pages 922 - 928. IEEE, 2015.

[3] S. Zia, B. Yüksel, D. Yüret and Y. Yemez, "RGB-D Object Recognition Using Deep Convolutional Neural Networks," 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, 2017, pp. 887-894. doi: 10.1109/ICCVW.2017.109

[4] https://www.tensorflow.org/

[5] https://rgbd-dataset.cs.washington.edu/