Modality Illusion Learning Aids

Lina Zhu

June 12 ,2018

Abstract

Abstract We present a modality hallucination architecture for training an RGB object detection model which incorporates depth side information at training time. Our convolutional hallucination network learns a new and complementary RGB image representation which is taught to mimic convolutional mid-level features from a depth network. At test time images are processed jointly through the RGB and hallucination networks to produce improved detection performance. Thus, our method transfers information commonly extracted from depth training data to a network which can extract that information from the RGB counterpart. We present results on the standard NYUDv2 dataset and report improvement on the RGB detection task.

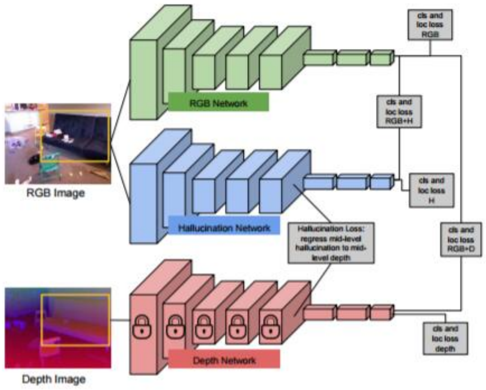

Figure 1. Training our modality hallucination architecture. We learn a multimodal Fast R-CNN convolutional network for object detection.

1. Introduction

RGB and depth images provide different and often complementary information. In fact, recent work has shown that both image formats can be used simultaneously to produce a better recognition model than either model alone[4]. The popularization of the RGB image capture device popularity depth capture device is much lower. This means that many recognition models will need to be executed And a separate RGB image as input. We introduce an algorithm that uses the available paired RGB-d training data to learn an illusionary intermediate convolution feature of an RGB image. We demonstrate through our method that we have produced a new convolutional network model that only surpasses a single RGB modal input, but the performance is better than the standard network trained only on RGB image. Therefore, our method usually passes information extracted from the deep training data to the available network to extract the information from the RGB correspondence. Convolutional networks have produced great successes in visual recognition tasks from classification [1], detection[2], semantic segmentation [3]. The standard method of training these networks is to use a large marker image corpra to initialize the network parameters and fine-tune the data source using smaller target tags. Although this strategy has proved to be very effective, it only provides a learning representation method for identifying and risking arbitrarily small nuances due to the large parameter space of the network.

We propose an additional representational learning algorithm which contains the following forms of ancillary information to generate an additional modality for the sensible test time single modal model during training. We have accomplished this by directly understanding a modal illusion network which optimizes the loss of localization of the standard class and bounding box, while being subject to additional guidance hallucinations, which reflects hallucinational features to auxiliary morphological features. Due to its practicality, we consider the production of an RGB detector that uses some pairs of RGB-D data times during training. In doing so, we produced a final model time in the test that only saw one RGB image, but was able to extract both through standard fine-tuning learning image features to monitor for loss and hallucinating features therein. Training has been received to reflect you those functions to extract if the depth image exists. We prove this our RGB and illusion detector model outperforms the most advanced RGB models on the NYUD2 dataset.

Table 1. **Detection (AP%) on NYUD2 test set:**Detection (AP%) on NYUD2 test set: We compare our performance (pool5 hallucinate) against a Fast R-CNN RGB detector trained on NYUD2 and against an ensemble of Fast R-CNN RGB detectors.

method	btub	bed	bshelf	box	chair	counter	desk	door	dresser	gbin	lamp	monitor	nstand	pillow	sink	sofa	table
RGB only	7.5	50.6	36.8	1.4	30.2	34.9	10.8	21.5	27.8	16.9	26.0	32.6	20.6	25.1	31.6	36.7	14.8
RGB ensemble	10.5	53.7	33.6	1.6	32.0	34.8	12.2	20.8	34.5	19.6	28.6	45.7	28.5	24.4	31.4	34.7	14.5
Our Net	13.9	56.1	34.4	1.9	32.9	40.5	12.9	22.6	37.4	22.0	28.9	46.2	31.9	22.9	34.2	34.2	19.4
RGB only	15.6	59.4	38.2	1.9	33.8	36.3	12.1	24.5	31.6	18.6	25.5	46.5	30.1	20.6	30.3	40.5	19.5
RGB ensemble	14.8	60.4	43.1	2.1	36.4	40.7	13.3	27.1	35.5	20.8	29.9	52.9	33.5	26.2	33.0	44.4	19.9
Our Net	16.8	62.3	41.8	2.1	37.3	43.4	15.4	24.4	39.1	22.4	30.3	46.6	30.9	27.0	42.9	46.2	22.2

2. Modal Illusion Model

We present a modal illusion architecture training that contains RGB object detection models at training time depth side information. Our hallucinatory network learns new complementary RGB image representation training to mimic deep mid-level functions. This new form of expression combined with RGB is represented by a standard fine-tuned learning image. Figure 1 shows our hallucinatory training architecture model. Our use of multi-layer convolutional networks as our basic identification framework has proven to be a very effective task for many different approvals. Previous research on RGB-D detection found that successful use of RGB and depth dual-channel model images through the final detection independent processing score is the maximum of the two predicted mean values. For our architecture, we build on this same general model. However, we tried to share information between the two both modes, especially the training time privilege depth mode, inform us of the final RGB unique detector. To do this, we have introduced a third channel that we call the illusion network. The illusion network takes RGB as the input image and a set of regions of interest and produces a score that detects each category and each region.

3. Experiments

We evaluate our model using a standard RGB-D detection dataset, NYUD2. The NYUD2 dataset consists of 1449 labeled RGB-D images. The dataset is split into train, val, and test sets. For our ablation experiments we train our model using the train set only and evaluate our model on the validation set. For our overall detection experiment which compares to prior work, we present results on the test set for our algorithm trained using the combined trainval set.

Table 1 reports the performance of our entire system, where there are two different architectures on the NYUD2 dataset. Two bases the architecture is AlexNet or VGG-1024. We trained our RGB and in-depth network using the proposed strategy, but used Fast R-CNN rather than RCNN. Then, we use slides to initialize our illusion network depth parameter values. Finally, we jointly optimized the appearance of hallucinogenic pool5 activation in the three-channel network structure. This is a specific architecture when our hallucination network is tagged. This refers to the selection of deep network and illusion network architectures and selects and indicates the RGB architecture separately.

4. Conclusion

We have introduced a novel technique for incorporating additional information, in the form of depth images, at training time to improve our test time RGB only detection models. We accomplish this through our modality hallucination architecture which combines a traditional RGB ConvNet representation with an additional and complementary RGB representation which has been trained to hallucinate depth mid-level features. Our approach outperforms the corresponding Fast R-CNN RGB detection models on the NYUD2 dataset.

References

[1] L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for heterogeneous domain adaptation. In *ICML*, 2012. 1

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[3] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[4] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*, 2014. 1