Single RGB Image Depth and Certainty Estimation via Deep Network and Dropout

Yuanfang Wang(yolandaw), Julian Gao(julianyg), Yinghao Xu(ericx)
CS229 project final report
Stanford University
yolanda.wang@stanford.edu

Abstract

Depth estimation is a useful technique for multiple applications such as obstacle detection and scene reconstruction. With the research focus on Convolutional Neural Networks (CNN), depth estimation has seen rapid development recently. Our contribution is a practical system which can inference depth from single RGB image with a measure of model uncertainty. Monte Carlo sampling with dropout at test time was used as a way of getting samples from the posterior distribution of models. By adding dropout at test time, we not only obtained the uncertainty map of our prediction, but also improved the performance of the original depth prediction task. Moreover, by testing with adding dropout after various layers, we find that adding dropout layers after every encoder convolutional layer shows potential in generating representative posterior distribution. This idea can also be applied in other neural network architecture.

1. Introduction and related work

The most commonly used devices to observe an environment are cameras. Cameras can capture 2D images from the 3D world with high density and color information. But to understand the 3D real world, we need to go from 2D image to 3D. Traditional methods for the 3D inference from 2D images are through feature matching (for example, key points, lines, super pixel) across multiple images and vanishing points.

Motivated by recent advances of deep convolutional neural network in computer vision, attempts of estimation dense per-pixel depth from a single RGB image [2, 1] using Multi-scale convolutional neural network have reached surprising success. The performances can be evaluated in two scenarios: indoor and outdoor. For the indoor scene, the ground truth depth range of pixels on an image is generally within 10 meters. Besides, the structure scenario, such as vertical walls and horizontal ceilings and ground, make the task of estimation depth from single RGB image seems not hard to solve. Moreover, the ground truth can be easily captured using Kinect (camera with depth sensor), which means an infinity amount of training data are available.

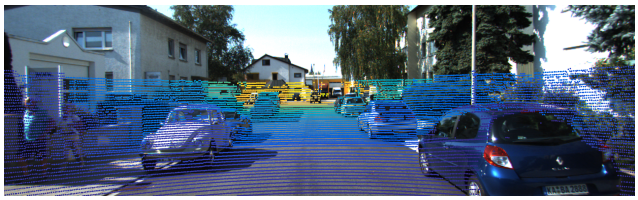


Figure 1. Sparse LiDAR points are projected onto 2D image, colored by the distance from camera.

The depth inference for outdoor scenario is not in the same situation. Depth range can be from 5 meters to more than 30 meters. The biggest problem is lack of training data. Since Kinect has limited sensing range and cannot be used outdoor, generally used depth sensors for outdoor scenario is LiDAR. Despite further sensing range, the significant difference compared with indoor Kinect sensor is that LiDAR can only get very sparse 3D point cloud. One visualization of sparse points from LiDAR is shown in Figure 1. The deficiency of ground truth makes the task of prediction depth from image using neural network even harder. Motivated by the basic application in [2], we proposed an extension version which out beat the original performance by adding dropout in the test time.

Inspired by [5], dropout cannot only be used as a regularization method during training, it provides a way of getting samples from the posterior distribution of models. More specifically, by adding dropout in teat time and run the network 30 times from the same input. The 30 samples can be seen as 30 observations of the per pixel depth distribution. The mean of distribution for a pixel can be used as the prediction of pixel depth, while the variance indicated the certainty of the prediction. [5] has used this method in the semantic segmentation problem, where the neural net-

work can be viewed as a general auto encoder and decoder. It was demonstrated that using the mean of distribution can improve the performance by 2-3% in the semantic segmentation task.

In this project, we want to explore the method of using dropout in testing for the depth prediction task. Different from semantic segmentation, which is a classification problem with clear encoder and decoder neural network architecture, depth prediction is a regression problem. The architecture of neural network we used is described in Section 2. The dropout theoretical basement is in Section 3. Detailed experiment is discussed in 4

2. Model architecture

Our model uses the architecture proposed by Eigen *et al.* [2]. Intuitively, in order to inference depth from RGB image, the machine should have some high level understanding of the scene as well as low level processes for details. In the architecture, a coarse-scale network with 6 encoder layer and 1 decoder layers is used to predict global level depth information of the scene. Then, the fine-scale network with both the original image and the output of coarse-scale network as input finetune the prediction with better details.

The training was on KITTI dataset [4], which composed of various of outdoor scenes captured with a driving car mounted by cameras and LiDAR sensors. 56 scenes for 800 images per scene were used for training. While training, dropout is applied to the last encoder layer of the coarse-scale network. Dropout rate is 50%. The depth prediction network can process 77 frames per second.

3. Bayesian neural network

As an approximate inference in Bayesian neural network, dropout can be used as a way of getting samples from the posterior distribution of models [3, 5].

1. Find the posterior distribution $p(W|X,Y)$ over the convolutional weights W, given our observed training data X and depth Y.

2. Learn the distribution over weights, $q(W)$ by minimizing the Kullback-Leibler (KL) divergence $\mathrm{KL}(q(W)||p(W|X,Y))$ between this approximated distribution and the full posterior.

3. Approximate variational distribution $q(W_i)$ for every $K \times K$ dimensional convolutional layer i, with units j , is defined as

$$b_{i,j} \sim \text{Bernoulli}(p_i) \text{ for } j = 1, ..., K_i,$$

$$W_i = \mathrm{M}_i \mathrm{diag}(b_i)$$

4. Use dropout as approximate variational inference in Bayesian Neural Networks [3]. Since step 2 is intractable, this step is replaced by approximating the integral with Monte Carlo integration over w, and get the unbiased estimator

$$\sum_{i=1}^{N} \mathrm{E}(y_i, \hat{f}(x_i, \hat{w}_i) - \mathrm{KL}(q(w)||p(w)))$$

where $\hat{w}_i \sim q(w)$. Then the integral with Monte Carlo integration is inferred by:

$$p(y|x, X, T) \approx \int p(y|x, w)q(w)dw \approx \frac{1}{T}\sum_{t=1}^{T} p(y|x, \hat{w}_t)$$

4. Experiment

4.1. Dataset

The testing of performance is also evaluated on KITTI dataset [4]. 8 scenes with 193 images was independently selected with the training data. Ground truth of depth is obtained from the synchronized LiDAR observation by projecting the 3D point cloud of a single frame to the corresponding RGB image using provided LiDAR-camera calibration parameters.

4.2. Probabilistic variants

To analysis the distribution generated from adding dropout, we explored a number of variants that have dropout after different layers :

- **Bayesian Encoder**. In this variant we use dropout after each encoder layer in the coarse-scale network.

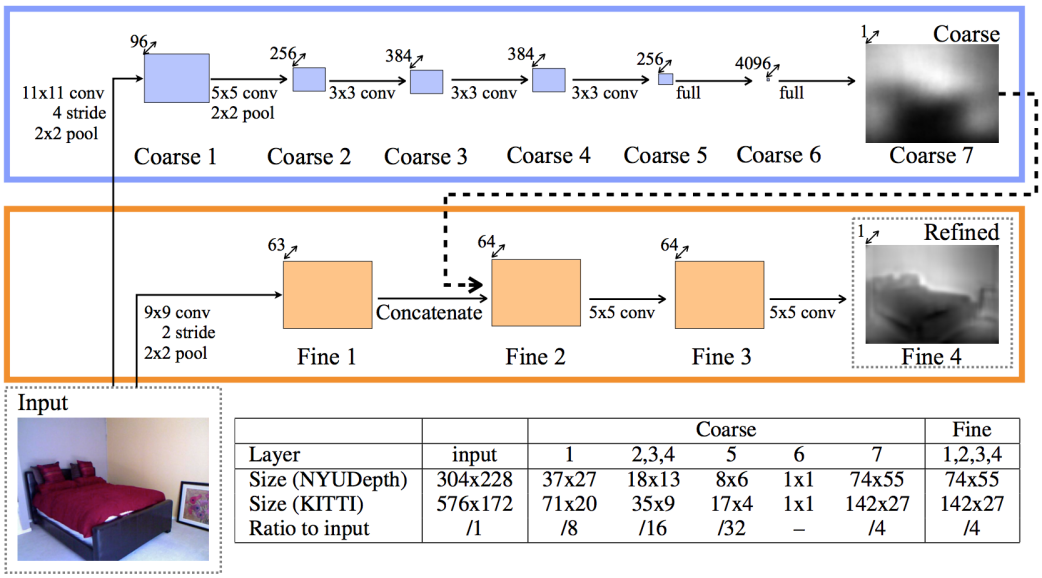
- **Bayesian Center Encoder**. In this variant we use dropout after the last encoder layer in the coarse-scale network.

- **Bayesian Center-2 Encoder**. In this variant we use dropout after the last 2 encoder layer in the coarse-scale network.

- **Bayesian Center**. In this variant we use dropout after the last encoder layer and the first decoder layer in the coarse-scale network.

- **Bayesian Center-4**. In this variant we use dropout after the last 2 encoder layer and the first decoder in the coarse-scale network, and the first convolutional layer in the fine-scale layer after concatenating of original image and coarse prediction.

We also applied four kinds of dropout rate for these variants: $\{5\%, 10\%, 20\%, 30\%\}$.



		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

Figure 2. Model architecture.

For each image in the testing data, we run each variant of network for 30 times to get 30 samples for the depth distribution of every pixel. Mean of the distribution served as our depth prediction and variance served as the uncertainty of prediction.

4.3. Qualitative analysis

Figure 3 shows the frames randomly chosen among the testing results of Bayesian Encoder with 30% dropout rate: predicted depth image (top), input image (middle) and uncertainty image (down). The darker pixels in the depth image represents closer points and lighter pixels represents farther points. In the uncertainty map, darker points refers to higher variance of the pixel depth distribution, i.e. the network is more uncertain with the prediction.

From the images we can find that the depth images can generally predict the distances of objects, for instance cars and road has darker color indicating that they are closer to the camera. Even pixels of the traffic sign in the Figure 3(c) can be reasonably inferred with smaller depth. The visualization of prediction certainty is also plausible: higher certainty for closer objects and lower for farther objects. Moreover, pixels referring to vegetation and shadows in all the frames has higher uncertainty, indicating that the complicated visual features of vegetation and light change of shadows hindered the prediction of depth.

Another thing wroth to mention is the holes in depth image predicted from Bayesian Center and Bayesian Center Encoder. Shown in Figure 4, image (a) is the network input. (b) and (c) are the output of Bayesian Center and Center Encoder respectively with dropout rate 5%. The upper images in (b) and (c) are single samples obtained from running the neural network once. The lower images in (b) and (c) are the means of 30 samples. The only difference between the variant used for (b), Bayesian Center, and the variant for (c), Bayesian Center Encoder, is an extra dropout layer with 5% dropout rate was added after the first decoder layer in the coarse-scale network. It is obvious that dropout can lead to holes in layer output inside the network. But the results show that holes in encoder layer output can be recovered at the end of the pipeline, while the holes in decoder layer output becomes larger as passing through the network, even with the concatenated information from original input image in the fine-scale layer.

4.4. Quantitative analysis

The prediction of our network can only provide the relevant pixel depth for the input image. As a result, for quantitative analysis, we need to first retrieve the scale between our estimation and the ground truth by finding the global scale of the scene with minimizing the L2 error between ground truth and rescaled prediction. We used the scale invariant method proposed in [2] for our quantitative evaluation.

Because the holes of decoder layer output influenced badly for the prediction, we only evaluated the variants without dropout after decoder layer, i.e. Bayesian Encoder, Bayesian Center Encoder, Bayesian Center-2 Encoder. The scores are in Table 1. Less score indicates better performance. It shows that adding dropout only in the last or last two encoder layers tends to act as noises impeding the depth prediction. While adding dropout after all encoder layers

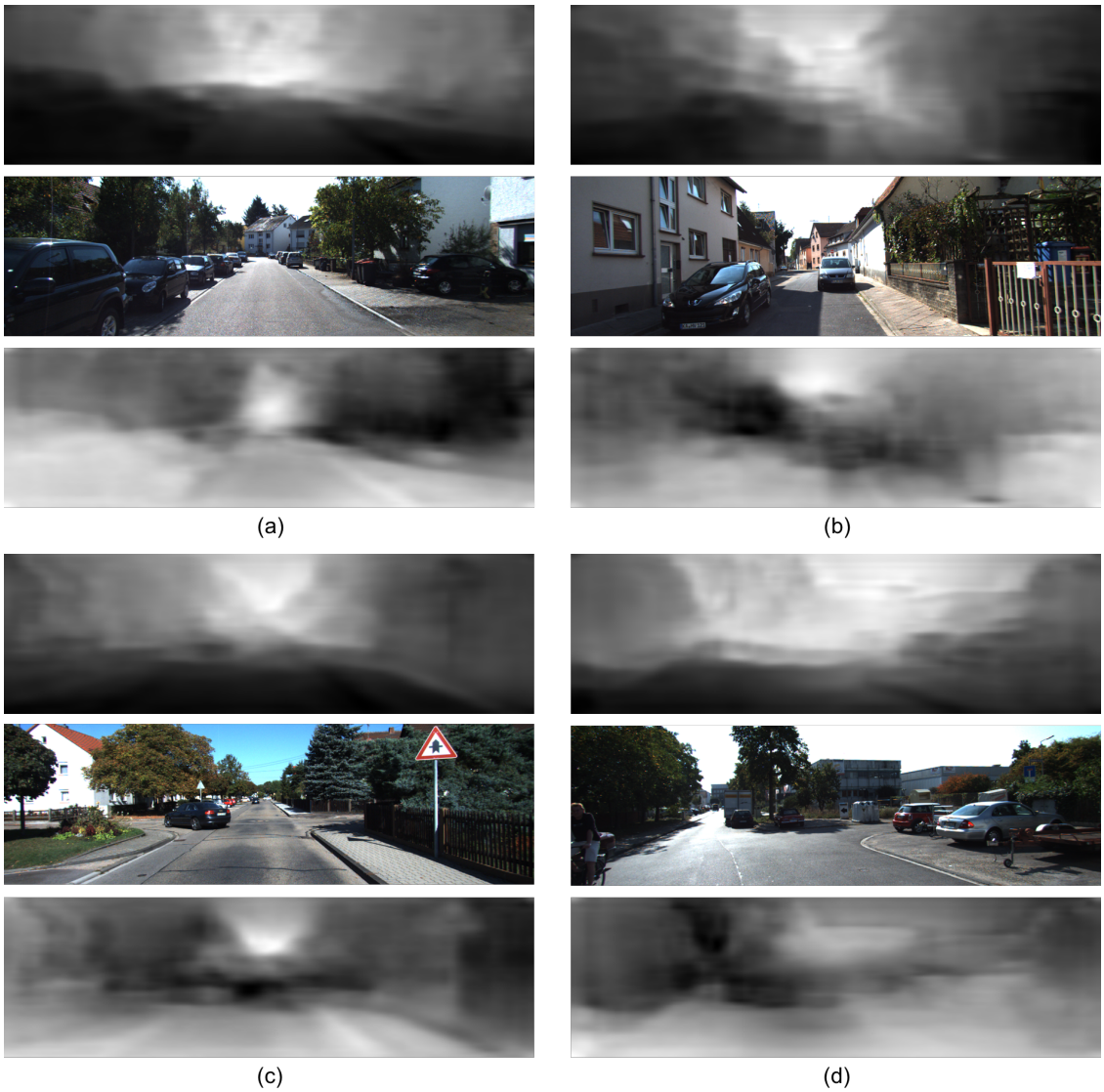


Figure 3. Visualization of predicted depth image (top), input image (middle) and uncertainty image (down). The darker pixels in the depth image represents closer points and lighter pixels represents farther points. In the uncertainty map, darker points refers to higher variance of the pixel depth distribution, i.e. the network is more uncertain with the prediction.

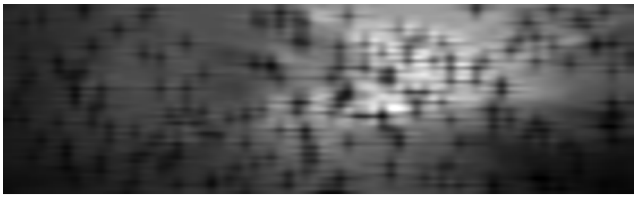
Outdoor scenes depth prediction				
Probabilistic Variants	5% dropout	10% dropout	20% dropout	30% dropout
Bayesian Encoder	1.186	1.158	1.084	**1.023**
Bayesian Center Encoder	1.225	1.225	1.227	1.226
Bayesian Center-2 Encoder	1.225	1.230	1.229	1.233
No dropout	**1.214**			

Table 1. Quantitative results

tends out to be a good random way for generating the posterior distribution, and the depth prediction significantly out preformed the original method with no dropout in testing time.

Besides, shown in Figure 5, we find an interesting correspondence between the prediction depth(mean) and prediction uncertainty(variance). It is intuitive that for farther points, our prediction might be with worse accuracy and higher variance, but why there exists a variance drop down for extremely far points (value over 200) needs further dis-

(a)



(b)

(c)

Figure 4. Image (a) is the network input. (b) and (c) are the output of Bayesian Center and Center Encoder respectively with dropout rate 5%. The upper images in (b) and (c) are single samples obtained from running the neural network once. The lower images in (b) and (c) are the means of 30 samples.

covery.

5. Conclusions

In this project we explored the architecture of multi-scale deep convolutional neural network for the task of predicting depth image from single RGB image. Meanwhile, by adding dropout at test time, we not only obtained the uncertainty map of our prediction, but also improved the performance of the original depth prediction task. Moreover, by testing with adding dropout after various layers, we find that adding dropout layers after every encoder convolutional

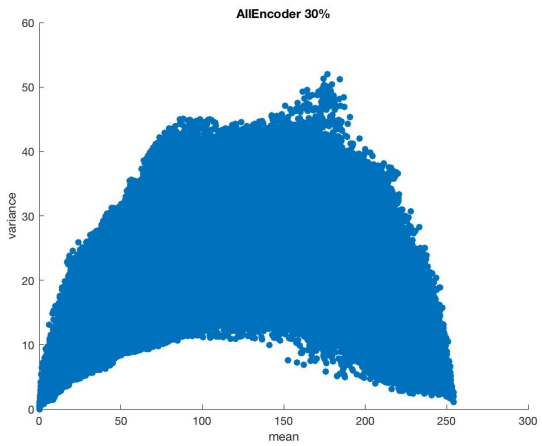


Figure 5. The relevance of prediction depth(mean) and prediction uncertainty(variance) for Bayesian Encoder with 30% dropout

layer shows potential in generating representative posterior distribution. This idea can also be applied in other neural network architecture.

References

[1] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *arXiv:1411.4734*, 2015. 1

[2] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. 2014. 1, 2, 3

[3] Y. Gal and Z. Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015. 2

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2

[5] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015. 1, 2