Research Proposal:

Vision as Inverse Graphics

Miquel Perelló Nieto, email: miquel@perellonieto.com

March 6, 2015

Abstract – Last advances in object recognition, localization and segmentation have demonstrated the outperformance of purely supervised learning models. However, these models have shown to require large volumes of labeled samples, and the manual collection of detection and segmentation labels is becoming prohibitive. For that reason, it is really necessary to find new methods to create large labeled datasets. We propose one method to generate extensive number of synthetic 2D images where all the latent variables are known. Additionally, we propose to study and mitigate the differences between the generated images and photographs, thus generalizing our findings to real images.

Motivation and Background

COMPUTER vision is a very active and important part of machine learning. This field extends to navigation of autonomous vehicles, detection of events, visual surveillance, information retrieval, automatic inspection of industrial processes, and other tasks where the automatic processing of images can be beneficial. The majority of these examples involve the use of object recognition, detection and/or segmentation. During last decade, several datasets have been created to compare and improve computer vision models (eg. ImageNet [1], Pascal VOC [2, 3], SUN [4], Flickr8k/30k [5, 6], and more recent Microsoft COCO [7]). Thanks to these datasets, state-of-the-art models are reaching human level performance in image classification tasks [8, 9]. However, state-of-the-art models are fully supervised and have proved to require large volumes of labels, and collecting labels for localization and segmentation purposes is much more tedious and expensive than it is for classification tasks. For that reason, the number of datasets that incorporate segmentation labels is scarce. For example, the new Microsoft COCO dataset contains 91 categories and 2.500.000 segmented instances from 328.000 images. And while the classification phase took 17.751 worker hours, completing the segmentation process took above 80.000 worker hours (18k worker hours to choose the labels of every image, 8k worker hours to mark each instance, and 55k worker hours to segment each marked instance). Moreover, only 1/3 of the participants passed successfully the test to demonstrate a sufficiently good segmentation performance.

Figure 1: **Synthetic human body depth representations** (Shotton et al. [10])

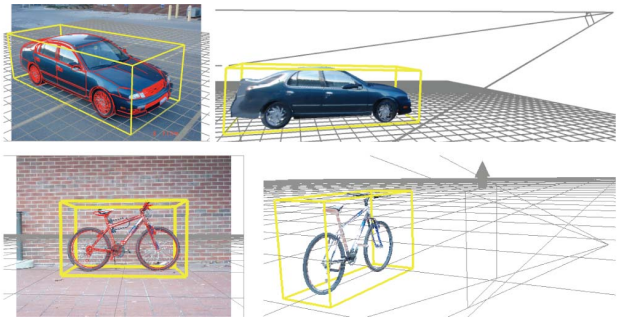

Figure 2: **(Right) 3D cad models. (Left) Detection and generalization to real images** (Zia et al.[11]).

One approach to evade this problem is to use semi-supervised learning, or weakly supervised learning. Another idea is to create synthetic data where all the possible latent variables are known. For example, Shotton et al. [10] increased the number of human body depth images modifying artificially the height, width, weight, pose, camera and location of the initial subjects (see some examples in Figure 1). With the augmented dataset, the authors achieved state-of-the-art accuracy in

human pose estimation, and their model was incorporated into the Microsoft Kinect gaming platform. Additionally, Zeeshan Zia et al. [11] used 3D CAD representations of cars and bicycles to create artificial samples (see also Stark et al. [12], and Pepik et al. [13]). The authors got state-of-the-art results in monocular 3D pose estimation for these objects. Figure 2 shows one example for car and bicycle.

Furthermore, despite learning the exact likelihood of the data is an intractable problem, Hinton and Dayan et al. [14] proposed the wake-sleep algorithm to approximate the likelihood using a generative and a recognition model. The wake-sleep algorithm consisted in two phases. During the wake phase (1), the recognition model was used to infer the latent variables from a real sample and the generative model was trained to recreate the hidden and the visible variables. During the sleep phase (2), the generative model created a sample and the recognition model was trained to recognize the hidden variables. The authors applied this technique to two directed graphical models that they called Helmholtz Machine [15]. In 2006 Hinton et al. [16] modified the wake-sleep algorithm using a contrastive divergence technique and trained a Deep Belief Network (DBN) by stacking several Restricted Boltzmann Machines (RBMs). The resulting DBN surpassed the best recognition models on classification of handwritten digits. More recent work is being developed to improve the original wake-sleep algorithm [17, 18].

Proposal

We propose to train a stochastic generative model of 3D scenes, and render the scenes to produce large amounts of image samples. At the same time, we can train a recognition model to help the learning of the generative model [15, 14, 16]. During the training, the latent variables inferred by one model are used to determine the error produced by the other. It has been demonstrated that by reducing these errors it is possible to find a good approximation of the likelihood of the real data.

We propose the next steps to create the stochastic generative model. First, (1) obtain 3D CAD models of the necessary objects (eg. Zia et al. [11]). Next, (2) augment the objects variability by: (a) using Principal Component Analysis to find the intraclass variability and use their largest components [11], (b) use Probabilistic Graphical models to infer the "underlying causes of the structural variability" [19], (c) use semi-automatic methods to generate the new samples using human help [20], or (d) use semi-automatic augmentation of 3D models from 2D photo examples [21]. Finally, (3) generate different scenes with the available objects and specific constraints (eg. [22, 23]). These constrains can be obtained from 3D CAD examples, or some of them can be inferred from 2D or 3D images. For example, using datasets with available contextual information it is possible to infer the common cluster of objects in different scenarios. From MS COCO there is an average of 7.7 instances per image, and other datasets like Flickr contain textual information that can be used to find these clusters.

Finally, we propose to study the photorealism of the random samples generated by the generative model. We could measure the differences and learn some type of structured noise to minimize the discrepancy between our samples and the real images (e.g. reducing the size, blurring the images, or adding random noise). The structured noise could help during the training and to generalize better.

Conclusion

Last advances in computer vision have shown that state-of-the-art models need large amounts of labeled data. However, the size of next datasets for detection and segmentation is becoming prohibitive for human labeling. We proposed a set of techniques to generate large amounts of synthetic data where all the latent variables are known. The generated latent variables can be used to train a recognition model and improve the generative model at the same time. In addition, we suggested to study the discordance of our synthetic data with respect to real photographies to reduce the difference and improve the global training.

References

[1] Olga Russakovsky, Jia Deng, and Hao Su. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv: ...*, September 2014.

[2] Mark Everingham and Luc Van Gool. The pascal visual object classes (voc) challenge. *International journal of ...*, 88(2):303–338, September 2010.

[3] Mark Everingham and SMA Eslami. The pascal visual object classes challenge: A retrospective. *International Journal of ...*, 111(1):98–136, June 2014.

[4] Jianxiong Xiao, J. Hays, and K.A. Ehinger. Sun database: Large-scale scene recognition from abbey to zoo. *Computer vision and ...*, pages 3485–3492, June 2010.

[5] Micah Hodosh, P Young, and J Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence ...*, 47:853–899, 2013.

[6] Peter Young and Alice Lai. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the ...*, 2014.

[7] TY Lin, Michael Maire, Serge Belongie, and James Hays. Microsoft COCO: Common objects in context. *Computer Vision–ECCV ...*, May 2014.

[8] Christian Szegedy, Wei Liu, Yangqing Jia, and Pierre Sermanet. Going deeper with convolutions. *arXiv preprint arXiv: ...*, September 2014.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv:1502.01852 [cs]*, February 2015.

[10] Jamie Shotton, Toby Sharp, and Alex Kipman. Real-time human pose recognition in parts from single depth images. *Communications of the ...*, 56(1):116–124, January 2013.

[11] M.Z. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3D Representations for Object Recognition and Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2608–2623, November 2013.

[12] Michael Stark, Michael Goesele, and Bernt Schiele. Back to the Future: Learning Shape Models from 3D CAD Data. *BMVC*, pages 1–11, 2010.

[13] B. Pepik and M. Stark. Teaching 3d geometry to deformable part models. *Computer Vision and ...*, pages 3362–3369, June 2012.

[14] G E Hinton, P Dayan, B J Frey, and R M Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science (New York, N.Y.)*, 268(5214):1158–61, May 1995.

[15] Peter Dayan, GE Hinton, RM Neal, and RS Zemel. The helmholtz machine. *Neural computation*, 904:889–904, 1995.

[16] GE Hinton, S Osindero, and YW Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 1554:1527–1554, 2006.

[17] Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. *arXiv preprint arXiv:1406.2751*, June 2014.

[18] DJ Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, January 2014.

[19] Evangelos Kalogerakis, Siddhartha Chaudhuri, Daphne Koller, and Vladlen Koltun. A Probabilistic Model for Component-based Shape Synthesis. *ACM Trans. Graph.*, 31(4):55:1–55:11, July 2012.

[20] Siddhartha Chaudhuri and Evangelos Kalogerakis. Probabilistic reasoning for assembly-based 3D modeling. *ACM Transactions on ...*, pages 35:1–35:10, 2011.

[21] Kai Xu, Hanlin Zheng, Hao Zhang, and Daniel Cohen-Or. Photo-inspired model-driven 3D object modeling. *ACM Transactions on ...*, pages 80:1–80:10, 2011.

[22] Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas Funkhouser, and Pat Hanrahan. Example-based Synthesis of 3D Object Arrangements. *ACM Trans. Graph.*, 31(6):135:1–135:11, November 2012.

[23] Yi-Ting Yeh, Lingfeng Yang, Matthew Watson, Noah D. Goodman, and Pat Hanrahan. Synthesizing Open Worlds with Constraints Using Locally Annealed Reversible Jump MCMC. *ACM Trans. Graph.*, 31(4):56:1–56:11, July 2012.