Learning from Simulated and Unsupervised Images

Hongzhi Liu

June 28, 2018

Abstract

With recent progress in graphics, it has become more tractable to train models on synthetic images, potentially avoiding the need for expensive annotations. However, learning from synthetic images may not achieve the desired performance due to a gap between synthetic and real image distributions. Today, I read a thesis written by Ashish Shrivastava who is from Apple Inc. His team introduce Simulated+Unsupervised (S+U) learning where the task is to learn a model to improve the realism of a simulator's output using unlabeled real data, while preserving the annotation information from the simulator. Besides, the team develop a method for S+U learning that uses an adversarial network similar to Generative Adversarial Networks (GANs). They show that this enables generation of highly realistic images which can demonstrate both qualitatively and with a user study.

1. Overview of S+U Learning and SimGAN

The GAN framework learns two networks (a generator and a discriminator) with competing losses. The goal of the generator network is to map a random vector to a realistic image, whereas the goal of the discriminator is to distinguish the generated from the real images. The GAN framework was first introduced by Goodfellow *et al.* [2] to generate visually realistic images and since then, many improvements and interesting applications have been proposed [4].

Large labeled training datasets are becoming increasingly important with the recent rise in high capacity deep neural networks [1]. However, labeling such large datasets is expensive and timeconsuming. Thus, the idea of training on synthetic instead of real images has become appealing because the annotations are automatically available. Many efforts have explored using synthetic data for various prediction tasks. Shrivastava's work is complementary to these approaches, where he improves the realism of the simulator using unlabeled real data. This approach allows the team to generate realistic training images which can be used to train any machine learning model, potentially for multiple tasks.

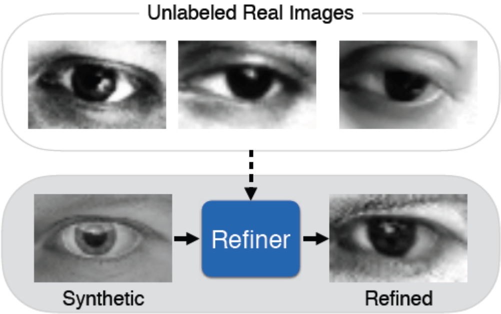

Figure 1. Simulated+Unsupervised (S+U) learning. The task is to learn a model that improves the realism of synthetic images from a simulator using unlabeled real data, while preserving the annotation information.

In this paper, Ashish Shrivastava and his team propose Simulated+Unsupervised (S+U) learning, where the goal is to improve the realism of synthetic images from a simulator using unlabeled real data [6]. The improved realism enables the training of better machine learning models on large datasets without any data collection or human annotation effort. In addition to adding realism, S+U learning should preserve annotation information for training of machine learning models, *e.g.* the gaze direction in Fig. 1 should be preserved.

Furthermore, the team develop a method for S+U learning, which they term SimGAN that refines synthetic images from a simulator using a neural network which we call the 'refiner network'. Fig. 2 gives an overview of their method: a synthetic image is generated with a black box simulator and is refined using the refiner network. To add realism, they train their refiner network using an adversarial loss, similar to Generative Adversarial Networks (GANs) [2] such that the refined images are indistinguishable from real ones using a discriminative network.

2. S+U Learning with SimGAN

The goal of Simulated+Unsupervised learning is to use a set of unlabeled real images $y_i \in \mathcal{Y}$ to learn a refiner $R_\theta(x)$

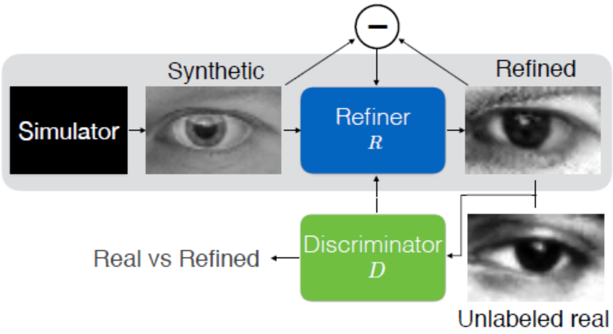

Figure 2. Overview of SimGAN. Shrivastava refines the output of the simulator with a refiner neural network, R, that minimizes the combination of a local adversarial loss and a 'selfregularization' term. The adversarial loss 'fools' a discriminator network, D, that classifies an image as real or refined. The self-regularization term minimizes the image difference between the synthetic and the refined images. The refiner network and the discriminator network are updated alternately.

that refines a synthetic image x, where θ are the function parameters. Let the refined image be denoted by $\tilde{x}$, then $\tilde{x} := R_\theta(x)$. The key requirement for S+U learning is that the refined image $\tilde{x}$ should look like a real image in appearance while preserving the annotation information from the simulator. To this end, Shrivastava proposes to learn θ by minimizing a combination of two losses as Eq. 1:

$$\mathcal{L}_R(\theta) = \sum_i l_{real}(\theta; x_i, \mathcal{Y}) + \lambda l_{reg}(\theta; x_i). \qquad (1)$$

where x_i is the i^{th} synthetic training image. The first part of the cost, l_{real}, adds realism to the synthetic images, while the second part, l_{reg}, preserves the annotation information. In the following sections, they expand this formulation and provide an algorithm to optimize for θ.

2.1. Adversarial Loss with Self-Regularization

To add realism to the synthetic image, the team need to bridge the gap between the distributions of synthetic and real images. An ideal refiner will make it impossible to classify a given image as real or refined with high confidence. This need motivates the use of an adversarial discriminator network, D_ϕ, that is trained to classify images as real vs refined, where ϕ are the parameters of the discriminator network. The adversarial loss used in training the refiner network, R, is responsible for 'fooling' the network D into classifying the refined images as real. Following the GAN approach [2], they model this as a two-player minimax game, and update the refiner network, R_ϕ, and the discriminator network, D_ϕ, alternately. Next, they describe this intuition more precisely. The discriminator network updates its parameters by minimizing the following loss as Eq. 2:

$$\mathcal{L}_D(\phi) = -\sum_i \log(D_\phi(\tilde{x}_i)) - \sum_j \log(1 - D_\phi(y_j)). \quad (2)$$

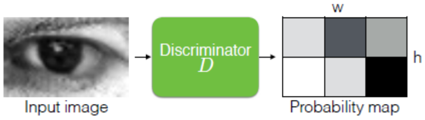

Figure 3. Illustration of local adversarial loss. The discriminator network outputs a $w \times h$ probability map. The adversarial loss function is the sum of the cross-entropy losses over the local patches.

This is equivalent to cross-entropy error for a two class classification problem where $D_\phi(.)$ is the probability of the input being a synthetic image, and $1{-}D_\phi(.)$ that of a real one. They implement D_ϕ as a ConvNet whose last layer outputs the probability of the sample being a refined image.

2.2. Local Adversarial Loss

Another key requirement for the refiner network is that it should learn to model the real image characteristics without introducing any artifacts. When Shrivastava and his team train a single strong discriminator network, the refiner network tends to over-emphasize certain image features to fool the current discriminator network, leading to drifting and producing artifacts. A key observation is that any local patch sampled from the refined image should have similar statistics to a real image patch. Therefore, rather than defining a global discriminator network, they can define a discriminator network that classifies all local image patches separately. This division not only limits the receptive field, and hence the capacity of the discriminator network, but also provides many samples per image for learning the discriminator network. The refiner network is also improved by having multiple 'realism loss' values per image.

In their implementation, they design the discriminator D to be a fully convolutional network that outputs $w \times h$ dimensional probability map of patches belonging to the fake class, where $w \times h$ are the number of local patches in the image. While training the refiner network, they sum the cross-entropy loss values over $w \times h$ local patches, as illustrated in Fig. 3.

3. Evaluation of the Method

Shrivastava and his team evaluate their method for appearance-based gaze estimation in the wild on the MPIIGaze dataset [8] and hand pose estimation on the NYU hand pose dataset of depth images. They use a fully convolutional refiner network with ResNet blocks for all of their experiments.

Table 1. Comparison of SimGAN to the state-of-the-art on the MPIIGaze dataset of real eyes. The second column indicates whether the methods are trained on Real/Synthetic data. The error the is mean eye gaze estimation error in degrees. Training on refined images results in a 2.1 degree improvement, a relative 21% improvement compared to the state-of-the-art.

Method	R/S	Error
Support Vector Regression (SVR) [5]	R	16.5
Adaptive Linear Regression ALR) [3]	R	16.4
Random Forest (RF) [7]	R	15.4
kNN with UT Multiview [9]	R	16.2
CNN with UT Multiview [9]	R	13.9
kNN with UnityEyes [8]	S	9.9
CNN with UnityEyes Synthetic Images	S	11.2
CNN with UnityEyes Refined Images	S	7.8

Tab. 1 shows a comparison to the state-of-the-art. Training the CNN on the refined images outperforms the state-of-the-art on the MPIIGaze dataset, with a relative improvement of 21%. This large improvement shows the practical value of our method in many HCI tasks.

References

[1] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2

[3] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE TPAMI*, 2014. 3

[4] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. *arXiv preprint arXiv:1606.03498*, 2016. 1

[5] T. Schneider, B. Schauerte, and R. Stiefelhagen. Manifold alignment for person independent appearance-based gaze estimation. In *ICPR*, 2014. 3

[6] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017. 1

[7] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3D gaze estimation. In *CVPR*, 2014. 3

[8] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *ACM ETRA*, 2016. 2, 3

[9] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, 2015. 3