Investigation of the Complex Data Distributions for their Efficient Generation

Philipp Kofman[1], Oles Dobosevych[1], and Rostyslav Hryniv[1,2]

[1] Ukrainian Catholic University, Lviv, Ukraine
[2] University of Rzeszw, Rzeszw, Poland

{kofman, dobosevych, rhryniv}@ucu.edu.ua

Abstract. Currently, the active development of image processing methods requires large amounts of correctly labeled data. The lack of quality data makes it impossible to use various machine learning methods. In case of limited possibilities for collecting real data, used methods for their synthetic generation. In practice, we can formulate the task of the high-quality generation of synthetic images as an efficient generation of complex data distributions, which is the object of study of this work. Generating high-quality synthetic data is an expensive and complicated process in terms of existing methods. We can distinguish two main approaches that are used to generate synthetic data: image generation based on rendered 3-D scenes and the use of GANs for simple images. These methods have some drawbacks, such as a narrow range of applicability and insufficient distribution complexity of the obtained data. When using GANs to generate complex distributions, in practice, we face a visible increase in the complexity of the model architecture and training procedure. A deep understanding of the real data complex distributions can be used to improve the quality of synthetic generation. Minimizing the differences in the real and synthetic data distributions can improve not only the generation process but also develop tools for solving the problem of data lack in the field of image processing.

Keywords: statistic · generative adversarial network · deep learning · synthetic data

1 Introduction

Expanding the capabilities of computer vision and deep learning opens up opportunities and approaches to solving many problems that previously remained unresolved. Many tasks that need to be solved remain beyond the reach of modern deep learning technologies - even though there is a large amount of manually annotated data.

Deep learning models do not have an understanding of the input, at least not in the human sense. People understand images based on their experience. Machine learning models do not have access to such experience, and therefore they cannot understand the input data in this way. By annotating a large number of training examples for models, we force them to learn a geometric transformation that brings data to human concepts

In: Proceedings of the 1st Masters Symposium on Advances in Data Mining, Machine Learning, and Computer Vision (MS-AMLV 2019), Lviv, Ukraine, November 15-16, 2019, pp. 21–28

for a specific set of examples, but this transformation is just a simplified outline of the original object model.

Deep learning models do not currently have a mechanism for learning abstractions through the direct definition of an object, but working with thousands, millions, or even billions of training examples solves this problem only partially [1].

Data collection for such tasks is essential, but sometimes very difficult, especially in the case of rare classes of objects. We should note that for such amount of data, manual annotation is not the best decision, since it requires a lot of resources and well-established markup strategies.

One way to solve this problem is to use artificially generated data. However, when using synthetic data, we may face the problem of a big jump in the complexity of choosing the architecture and methods for training the model. We can assume that for the model, there is a fundamental difference between real and generated data.

This study aims to compare the distributions of real and synthetic data, study the reasons for the increase of work complexity when using synthetic data and the way to eliminate it.

2 Related Work

Since 2010, research has been conducted in the field of visual domain adaptation, where the first approach to the problem was statistical methods. However, since 2014, neural network methods have gained considerable popularity. Soon the lack of data became a related problem, which led to the growth of synthetics generation methods [2-4].

The need for synthetic data often arises in many tasks. An outstanding representative of such tasks is autonomous driving. In order to make high-precision classifiers of road markings, signs, cars and other vast volumes of qualitatively marked data are needed.

To solve this problem, in 2016 was proposed the idea of generating a dataset based on the gaming world. [5] used the GTA5 game world. The purpose of this work was to get markup from screenshots of the game. The main idea was to use an existing virtual 3-D world. However, the limitations of the game did not allow obtain the complete markup necessary to solve the problem of autonomous driving.

At the same time, in 2016, using the same idea of the virtual world, the SYNTHIA dataset was generated in order to assist in semantic segmentation and the problems of understanding related scenes in the context of driving scenarios [6]. The authors a bit changed the approach and created their virtual world using the Unity development platform. They built their virtual cities based on real city prototypes [6].

One of the main advantages was the ability to add natural events, such as time of day, rain, snow, fog and other. These methods are automatic from generating datasets point of view but challenging at the design stage of the virtual world.

Another remarkable example of using synthetic data is the task of a person's gaze direction recognition. In 2016, [7] presented a method for generating eyes, taking into account the eyeball biological features, as well as the skin around it. This work paid great attention to the light characteristics of the eye surface. It used the same idea of generating 3-D models of objects, but took into account their physical characteristics for higher realism.

Often there are also problems in which the original dataset contains data of a different nature. From here, naturally arise the tasks of domain adaptation [19] and style transfer [20]. In 2018, [20] discussed the methods of transferring style from one image to another using neural network methods. The ideas were based on the principle that neural networks highlight the features of style. The first articles on this topic used features obtained by neural networks VGG [13], as well as the principles of auto-encoders [17]. Style transfers were carried out due to tricks with intermediate outputs of neural networks, as well as in various ways of constructing a loss function. In 2017, Judy Hoffman introduced a domain adaptation method called CYCADA [21]. Its essence was the use of a complex architecture consisting of two generators, two discriminators, and four auxiliary decision networks. The method showed good results; however, for training, it was necessary to have labelled semantic segmentation of data [22, 23].

Recently, a large number of approaches, methods, and architectures have been developed to solve this and similar problems. However, analyzing the work in this area, we can say that insufficient attention was paid to the problem consideration of generating synthetic data precisely from the statistical methods point of view.

3 Research Hypothesis and Problem

The main problem considered in this paper is the difficulty of generating highquality synthetic data for their further use in deep learning models for image processing. So, the central objective is to identify the hidden differences between real and synthetic data for their high-quality generation. We highlight related objectives:

- Hypothesis confirmation of the presence of a statistically significant difference in the distributions of real and synthetic data
- Building a pipeline for image conversion
- Quality criterion selection for assessing the generated data

The objects of the study are four primary datasets: real photos collected from autorecorders [8], generated pictures ”SYNTHIA” transport routes [9], real photos of dogs [10] and generated images of dogs using GANs [11, 12].

We assume that the identification of distinctive features in the distributions of real and synthetic data will help to avoid the difficulty of transferring the machine learning model between them.

The formal statement of the problem:

1. Conversion of images and their transformation into vector space using neural network methods
2. Construction of space and two presentations: from images to hidden space and vice versa
3. Analysis of distributions in a new hidden space and their investigation using statistical methods

4. Conducting transformations on data in hidden space to minimize differences
5. Display modified synthetic data into the image space
6. Selection of a formal criterion for assessing the quality of artificially generated data so that machine learning models in the field of computer vision containing synthetics in the training dataset show high quality working with test and validation samples of real data.

4 Envisioned Approach

4.1 Dataset Collecting

For the experiments, we were selecting data according to the two criteria: the relevance of the task for which these could be used; and the simplicity of objects for human perception. The principal requirement was the existence of a pair (real data, synthetic data) since the generation of large amounts of synthetic data from scratch is a costly and time-consuming process.

By the first criterion, we selected the SYNTHIA dataset[1]. SYNTHIA is a dataset that has been generated to aid semantic segmentation and related scene understanding problems in the context of driving scenarios. SYNTHIA consists of a collection of photo-realistic frames rendered from a virtual city and comes with precise pixel-level semantic annotations for 13 classes: miscellaneous, sky, building, road, sidewalk, fence, vegetation, pole, car, sign, pedestrian, cyclist, lane marking [9].

The self-driving task requires maximum accuracy in its solution, and therefore large high-quality datasets, which is consistent with the relevance of our work. In this case, we chose the Berkeley DeepDrive dataset[2] as real data on which three complex tasks for the CVPR 2018 Autonomous Driving Workshop were conducted: detection of road objects, segmentation of the driving region, and adaptation of semantic segmentation domains [8].

According to the second criterion, we took dogs images dataset because they are easy for human perception, but challenging to formalize for a computer. It follows that the distribution is complex, and this is a vital aspect to consider in our study. As a real dataset, we selected Stanford Dogs Dataset[3] [10], which contains images of 120 dog breeds from around the world. This dataset was created using images and annotations

[1] The SYNTHIA dataset (https://synthia-dataset.net/) is provided by the Computer Vision Center, Barcelona, and may be used for non-commercial purposes only, subject to the CC BY-NC-SA 3.0 license (http://creativecommons.org/licenses/by-nc-sa/3.0/legalcode)/

[2] Berkeley DeepDrive dataset is freely available for download and use at https://bdd-data.berkeley.edu/

[3] Open source Stanford Dogs Dataset is freely available for download and use at http://vision.stanford.edu/aditya86/StanfordDogs/

from ImageNet[4] for the task of detailed categorization of images. As its synthetic analogue, we chose images of dogs generated using GAN [12] method from the Kaggle Generative Dog Images[5] [11] competition. It contains 10 000 examples of synthetically generated dogs without markup.

4.2 Problem Solution

Before the experiment starts, we converted our data to a single image of 224x224 size and three-color channel format [13]

Our hypothesis assumes that the distributions of real and synthetic data have statistically significant differences. For humans, the difference between synthetic images and real data is intuitive, but like many similar processes, hard to formalize. Based on this statement, our approach attempts to formalize these differences.

The approach we chose for the first iteration of the experiment involves using trained neural networks to extract image information in vector form.

We will use the VGG16 network trained on ImageNet [14] with batch normalization as feature extractor [15]. Using statistical tests such as Student's T-test [16], Kullback-Leibler divergence (relative entropy) [24], we can confirm our assumption about the distinguishability of synthetics and real data with a certain level of confidence. For evaluating how synthetic data is suitable for modeling, we will use the approach proposed in [25], which is grounded in a particular application of synthetic data generation.

The next step is to train the variational auto-encoder [17] on real and synthetic data, thereby constructing a hidden space and two pretensions: from pictures to hidden space and vice versa. We will analyze and compare the basic statistical characteristics of real and synthetic data in a hidden space. We will use simple mathematical operations in order to approximate the statistical characteristics of synthetic data to real ones.

Then we pass the converted hidden representations of the synthetic data through the decoder. At the output, we expect to get images close to real.

4.3 Hypothesis Verification

Two experiments can serve as verification of our hypothesis.

First, we can re-pass the generated data through the trained VGG16 [14] with batch normalization. Then, a measure of quality will be a statistically insignificant difference in data distributions.

As a second experiment, let us pass the transformed synthetic and initial real data through a simple neural network, which will solve the binary classification problem, i.e., determine the nature of the image.

[4] ImageNet data are freely available for non-commercial research and/or educational use at http://image-net.org/download-images

[5] Open source Kaggle Generative Dog Images dataset is freely available for download and use at https://www.kaggle.com/c/generative-dog-images/data

After that, we will use the validation dataset to predict the binary classification label. If a neural network cannot accurately predict the correct label, then the conversion quality of synthetic data can be considered high. We assume that a neural network cannot distinguish class labels if the ROC-AUC [18] value is about 0.5 on the validation dataset.

5 Research Methodology and Plan

5.1 Dataset Preparation

Data volume is a critical parameter for statistical methods. In our case, we operate on four central datasets. However, for further work, it may be necessary to extend them to obtain greater representativeness. Therefore, it is planned to finish the preparation of data for experiments in the middle of October.

5.2 Hypothesis Confirmation

It is necessary to allocate one month of work to test the central hypothesis about the statistical difference in the distributions of synthetic and real data. Since trained, machine learning models can highlight non-representative features. Statistical tests in the first approximations can give mixed results. As a result, it may be necessary to adjust the design of the experiment or move to more stringent statistical tests. Since this is a fundamental hypothesis, we plan to end this stage in the middle of November.

5.3 Building a Pipeline for Image Conversion

Learning deep neural networks is a labor-intensive process. As a result, we laid the time until the middle of December for the second stage of the experiment. The complexity of this stage includes the fact that difficulties may arise in processing hidden data representations.

5.4 Result Evaluation

The mechanisms for conducting validation will be partially implemented in the first part of the experiment. We allocate two weeks for training simple models required for validation. Therefore, we plan to produce conclusions on the described experiments by the end of December.

6 Conclusion

Lack of data is the cornerstone of a large number of computer vision tasks. Synthetic data can be the solution to this problem. The use of classical methods of statistical analysis in conjunction with new ways of neural networks can give a much deeper understanding of the data and lead to the emergence of plans for the efficient generation of synthetic data.

References

1. Marcus, G.: Deep learning: a critical appraisal. arXiv preprint, arXiv:1801.00631 (2018)
2. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 2314, pp. 213–226. Springer, Berlin, Heidelberg (2010). doi: 10.1007/978-3-642-15561-1_16
3. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1521–1528. IEEE Press, New York (2011). doi: 10.1109/CVPR.2011.5995347
4. Tzeng, E., Darrell, N.: Deep domain confusion: maximizing for domain invariance. arXiv preprint, arXiv:1412.3474 (2014)
5. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: ground truth from computer games. In: Leibe B., Matas J., Sebe N., Welling M. (eds.) ECCV 2016. ECCV 2016. LNCS, vol. 9906, pp. 102–118. Springer, Cham (2016). doi: 10.1007/978-3-319-46475-6_7
6. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.: The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3234–3243. IEEE Press, New York (2016). doi: 10.1109/CVPR.2016.352
7. Wood, E., Baltrusaitis, T., Morency, L.-P., Robinson, P., Bulling, A.: Learning an appearance-based gaze estimator from one million synthesised images, In: 9th Biennial ACM Symposium on Eye Tracking Research & Applications, pp. 131–138. ACM (2016). doi: 10.1145/2857491.2857492
8. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: BDD100K: a diverse driving video database with scalable annotation tooling. arXiv preprint, arXiv:1805.04687 (2018)
9. SYNTHIA Home page. http://synthia-dataset.net/
10. Robots pets page. https://www.robots.ox.ac.uk/~vgg/data/pets/
11. Kaggle Biggan competition. https://www.kaggle.com/dvorobiev/doggies-biggan-sub-final
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014)
13. VGG16 Homepage. https://keras.io/applications/#vgg16
14. Simonyan, K., Zisserman, A.,: Very deep convolutional networks for largescale image recognition. arXiv preprint, arXiv:1409.1556 (2015)
15. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint, arXiv:1502.03167 (2015)
16. William H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C: the Art of Scientific Computing. Cambridge University Press, Cambridge (1992)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint, arXiv:1312.6114 (2014)
18. Understanding AUC-ROC Curve. https://towardsdatascience.com/ understanding-auc-roc-curve-68b2303cc9c5
19. Su, J.-C., Tsai, Y.-H., Sohn, K., Liu, B., Maji, S., Chandraker, M.: Active adversarial domain adaptation. arXiv preprint, arXiv:1904.07848 (2019)

20. Li, H.: A literature review of neural style transfer. https://www.cs. princeton.edu/LiteratureReview/COSBspr/NealStyleTransfer.pdf
21. Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y.: CYCADA: cycle consistent adversarial domain adaptation. arXiv preprint, arXiv:1711.03213 (2017)
22. Karacan, L., Akata, Z., Erdem, A., Erdem, E.: Learning to generate images of outdoor scenes from attributes and semantic layouts. arXiv preprint, arXiv:1612.00215 (2016)
23. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(4), 640–651 (2015). doi: 10.1109/TPAMI.2016.2572683
24. Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Statist. 22(1), 79–86 (1951). doi:10.1214/aoms/1177729694
25. Jordon, J., Yoon, J., van der Schaar, M.: Measuring the quality of synthetic data for use in competitions. arXiv preprint, arXiv:1806.11345 (2018)