Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2023, 10(3):70-74

Research Article ISSN: 2394 - 658X

Computer Vision: From Basics to Advance Applications

Kailash Alle

Sr. Software Engineer, Comscore Inc
kailashalle@gmail.com

ABSTRACT

Advanced computer vision problems include the identification and classification of moving vehicles. The study's outcome revealed a novel strategy for creating synthetic datasets with contemporary generative neural networks. To further explain this strategy, tests were conducted to create a synthetic dataset to use the most innovative diffusion process neural network models (Kandinsky 2.2) to address the fundamental computer vision problem in advanced applications. Training deep neural networks to tackle the object classification difficulty was done through research conducted on these datasets. The trained detectors' performance on a carefully chosen validation dataset was tested, and the findings proved the possibility of generating synthetic datasets with diffusion neural network models. A few challenges exist, though, including a large labor cost for fine-tuning trained diffusion and a significant discrepancy between the categories of produced and actual data when using this approach to create artificial datasets for deep neural networks of computer vision in production.

Keywords: Computer Vision, Basics, Advance, Applications

INTRODUCTION

With the advancement of deep neural networks in this century, computer vision applications have grown dramatically. Intelligent systems have been built around them to handle a wide range of activities, including industrial control, transportation control, and human video control, among many others. Many excellent data sets with metadata are needed to train deep neural networks for computer vision in these systems. There are instances in developing computer vision systems where gathering image data sets becomes incredibly challenging or impossible. Additionally, privacy concerns may make it more difficult to gather image data sets. The issues mentioned above can be solved with a synthetic data strategy. The most complex approach to producing synthetic data for computer vision issues is the multidimensional modeling technology-based method. This technique enables the use of three-dimensional modeling and scene creation tools (Blender, Unity3D) to replicate environments that closely resemble real environments in terms of factors like lighting, weather, and photorealism. From these environments, image data sets are intended to be collected to train deep neural networks in computer vision. There are some limitations to this method. When recreating a 3D simulation of a unique real environment, one must start with 3D models of individual things and work their way down to the last step of arranging the elements in the scene to match the genuine one. The team will need to put in a lot of work and time to complete this. The use of generative models is an added, equally fascinating method for producing artificial intelligence. As far as photorealism and image quality go, diffusion neural networks are currently the best models. Diffusion models' promise as sources of synthetic data has not been completely explored because they have evolved into innovative imaging solutions. Thus, one of the primaries aims of our work is to investigate this potential using the example of creating artificial picture data sets for a fundamental computer vision problem in the context of advanced applications—that is, the identification and categorization of moving automobiles.

Relevant Items

There are many works dedicated to applying the three-dimensional modeling approach to developing synthetic data. The topic of developing synthetic data for computer vision problems is currently widely developed. The part of generative models that is applied, which has not gotten as much attention as diffusion models. A method based on the generative-adversarial neural network SinGAN-Seg is proposed in the publication [1] for the creation of synthetic data in the medical domain. Medical professionals must invest a great deal of time and effort into the

laborious process of analyzing medical data, as noted by the authors. It is also seen that the medical industry faces challenges in getting trainable data because of privacy issues and a dearth of data for specific task categories. One marked frame is needed for training, which sets SinGAN-Seg apart from conventional generative-adversarial networks. The training of a U-Net++ segmentation network on a dataset created by SinGAN-Seg yields experimental results that resemble the performance level obtained from training on real data. Consequently, when there are not enough images in the training datasets, SinGAN enhances the segmentation models' performance. One generative model that was initially created for the purpose of creating synthetic data is called SinGAN-Seg, and it was designed to address computer vision issues in the highly specialized field of medicine.

A strategy to fine-tuning large-scale diffusion models for the text2image problem is presented in the study [2], which aims to enhance metrics performance in benchmark tasks that have been professionally researched, such as image classification on ImageNet datasets. The authors correctly note that deep generative models are developing and are now able to produce highly correct, photorealistic images. Denoising diffusion probabilistic models, for instance, can produce images that are on par with those produced by generative-adversarial networks in terms of quality. Therefore, the authors inquire as to whether the diffusion models in use today are strong enough to produce images that are proper for complicated issues and of a high enough quality. To generate face images that may be used to address face recognition problems, the publication [3] introduces the DGFace diffusion model. The authors point out that the process of creating synthetic datasets for face recognition is difficult because it requires not only producing photorealistic images but also multiple images of the same subject with different lighting conditions, expressions on their faces, and other variations that mimic the distribution of real data. The creation of face images of the same subject in many styles and with exact control is made possible by the DGFace diffusion model. Higher verification accuracy than earlier works is achieved by face recognition models trained on synthetic images from the proposed DCFace in 4 out of 5 test datasets. [—] AgeDB, LFW, CFPFP, and CPLFW along with CALFW. To solve challenging computer vision tasks like face recognition, this work shows that diffusion models have a high loss of usage as an alternative instrument for creating synthetic datasets.

EXPERIMENTS

The following procedure was established in order to test the hypothesis regarding the applicability of synthetic images produced by contemporary diffusion models for training neural networks to solve fundamental computer vision problems using the example of vehicle identification and classification: 1) choosing a diffusion model to create synthetic images; 2) creating a synthetic dataset to be used in the task of identifying and classifying vehicles in urban environments; 3) annotating the synthetic dataset for this purpose; 4) choosing neural networks to be used in the task of identifying and classifying vehicles; 5) training the neural networks for this purpose on the synthetic dataset; 6) testing the neural networks that were trained in the preceding step on a validation dataset of real images.

Choice of Diffusion Model

The process of creating the images is not considered; that is, potential models may produce images from text searches or from previously created images. The later models were chosen as potential candidates: DeepFloyd IF, Kandinsky 2.2, DALL-E 2, Midjourney v5, and Stable Diffusion XL 1.0.

Presented in spring 2022, OpenAI's DALL-E 2 is a second-generation generative model for the text-based image generation task. Two stages of the diffusion visual model DALL-E, the contrastive model CLIP, and a diffusion decoder are crossed in DALL-E 2. In tasks like developing robust representations of images with semantics and style, contrastive models like CLIP perform well, as the authors point out. (7) DALL-E 2 is not an open model, but it does enable the creation of text-based graphics with a high degree of photorealism. However, OpenAI services enable access to it.

Table 1. Comparative table of the diffusion models

Model name	Availability of the model in the open source	Ease of use for experiments	The ability to generate photorealistic images
Stable Diffusion XL 1.0	+	+	+
Kandinsky 2.2	+	+	+
DALL-E 2	-	-	+
Midjourney v5	-	+	+
Deep Floyd IF	+	-	+

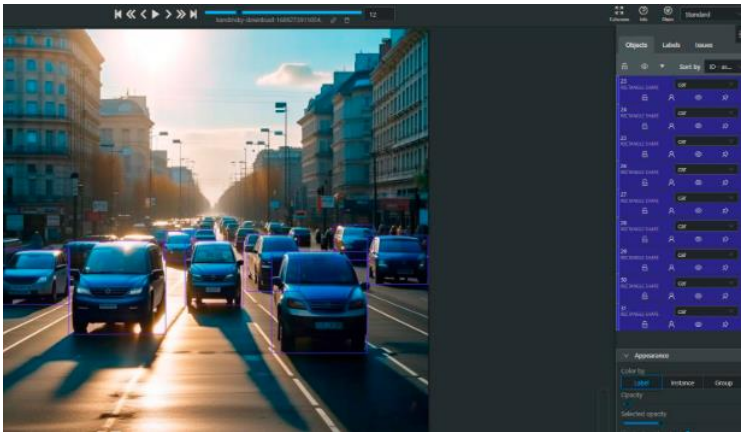
Creation of a fictitious collection

It was decided to create a synthetic dataset using the FusionBrain service from Sber, which offers free image production with various aspect ratios and stylistic options. In addition, any image can be downloaded to a local computer. It is vital to accurately provide textual prompts with the highest level of precision for the generation results to be as relevant to the vehicle detection and classification task at hand as possible. There is a list of keyword categories that need to be in the text prompt because the job is to find the vehicles in the flow. The following are included in these categories: 1) The frequency or rarity of traffic; 2) The kind of vehicle in the traffic (truck or car);

and 3) The location on a small scale 4) a generalized place (urban, rural, or highway), 5) the climate weather (rainy or sunny), 6) the time of day (day or night), 7) the position of the camera (building wall or lamppost)
{traffic density}, {kind of traffic in the traffic}, {location on a wide scale}, {narrow scale location}, {weather conditions}, {time of day}, {camera location} are the text prompt templates for creating photos for the detection and classification task. An apt illustration would be the little car traffic on the city's main thoroughfare, the sunny weather, the daytime hours, and the camera mounted atop a lamppost. It is possible to build 192 different text prompt variations using the Kandinsky 2.2 text prompt technique described above. Using the FusionBrain service, 1000 photos of automobile traffic with varying combinations of weather and time of day were generated based on some variations of the prompts created using the above template.

An explanation of the fictitious statistics

For this paper, a publicly available tool for annotation of synthetic images was chosen: the Computer Vision Annotation Toolkit (CVAT). The following computer vision tasks can be annotated with this tool: classification of images; detection of objects.
Furthermore, annotation can be readily organized using CVAT. Additionally, datasets can be imported and exported into a variety of formats using CVAT, including MS COCO, PASCAL VOC, YOLO, and CVAT for pictures. An SDK from CVAT enables the development of Python applications for managing the status of jobs and tasks in addition to automatically creating tasks using neural network models for preannotation. For annotation, the publicly accessible CVAT picture at cvat.ai was utilized.



Choosing neural networks for the identification and categorization of vehicles

Numerous neural network architectural variations have arisen since the peak of deep learning techniques in computer vision to address fundamental computer vision issues, such as the object detection job. The YOLO family of neural network detectors is one of the most advanced at addressing the real-time object detection challenge. The following YOLO models are still relevant today: YOLOv4, YOLOv5, YOLOv6, YOLOv7, and YOLOv8.
PyTorch is used in the implementation of YOLOv8 in the Ultralitycs repository, opening the door for a wide range of community aid. The YOLOv5 family of detectors was selected as the preferred object detectors for this paper because of its widespread community support and ease of use. However, based on the information provided, it can be concluded that all detectors from YOLOv4 through YOLOv8 are equally suitable in terms of performance and accuracy in the context of the experiments within this paper.

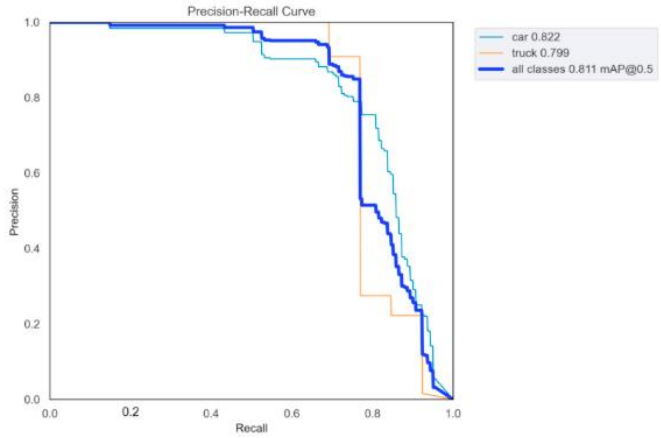
Neural network training for vehicle recognition and classification using a synthetic dataset

On MS COCO, training iterations were started from nothing without using previously learned weight values. The repository's standard Yolov5 model setups were applied. For every training iteration, a single YAML file with hyperparameters indicating modest augmentation was used. Following the training iterations from YOLOv5n to

YOLOv5x, a test sample of the synthetic dataset was used to decide which checkpoints performed the best in terms of accuracy. The table below provides an overview of the accuracy measures.

Table 2. Metrics of YOLOv5 networks on test samples of synthetic dataset

Network name	Precision	Recall	mAP@0.5	mAP@0.5-0.05
YOLOv5n 1024x1024	0.886	0.725	0.815	0.501
YOLOv5s 1024x1024	0.891	0.633	0.779	0.494
YOLOv5m 1024x1024	0.75	0.673	0.716	0.485
YOLOv5l 1024x1024	0.788	0.664	0.74	0.473
YOLOv5x 1024x1024	0.876	0.573	0.74	0.504



CONCLUSION

In this paper, we assessed the potential for producing high-quality, photorealistic data to address the fundamental computer vision problem in the context of urban applications, which is the identification and categorization of automobiles in traffic using innovative diffusion models to address the issue of producing images from text. The studies conducted for this paper's framework on neural network detector training on a dataset created with images of car traffic presented in different external conditions using Kandinsky 2.2 have demonstrated the great potential that modern diffusion models have to become fully functional and potent tools for producing data sets of potentially infinite volume when there aren't enough real data sets available to solve computer vision tasks. But it's likely that contemporary diffusion models need added adjustment. But it's likely that further fine-tuning is needed with current diffusion models to produce images that are more photorealistic and in closer domain to the real ones.

REFERENCES

[1]. Thambawita V., Salehi P., Sheshkal S.A., Hicks S.A., Hammer H.L., Parasa S. SinGAN-Seg: Synthetic training data generation for medical image segmentation, PLoS ONE 17(5): e0267976, 2022: https://doi.org/10.1371/journal.pone.0267976

[2]. Tripathi S., Chandra S., Agrawal A., Tyagi A., Rehg J. M., Chari V. Learning to Generate Synthetic Data via Compositing. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): https://ieeexplore.ieee.org/document/8953554?denied=

[3]. Voetman R., Aghaei M., Dijkstra K. The Big Data Myth: Using Diffusion Models for Dataset Generation to Train Deep Detection Models, 2023: https://www.researchgate.net/publication/371684920_The_Big_Data_Myth_Using_Diffusion_Models_for_Dataset_ Generation_to_Train_Deep_Detection_Models

[4]. Azizi S., Kornblith S., Saharia C., Norouzi M., Fleet D. J. Synthetic Data from Diffusion Models Improves ImageNet Classification, 2023: https://openreview.net/pdf/1ebc4a57598471f9a31f2adb9a161b3f7e241f9c.pdf

[5]. Kim M., Liu F., Jain A., Liu X. DCFace: Synthetic Face Generation with Dual Condition Diffusion Model, 2023: http://cvlab.cse.msu.edu/pdfs/kim_liu_jain_liu_cvpr2023.pdf

[6]. He R., Sun S., Yu X., Xue C., Zhang W., Torr P., Tori P., Bai S., Qi X. Is synthetic data from generative models ready for image recognition? 2023 International Conference on Learning Representations: https://openreview.net/pdf?id=nUmCcZ5RKF

[7]. Ramesh A., Dhariwal P., Nichol A., Chu C., Chen M. Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022: https://www.semanticscholar.org/paper/Hierarchical-Text-Conditional-Image-Generation-withRamesh-Dhariwal/c57293882b2561e1ba03017902df9fc2f289dea2

[8]. Dmitrov D. Kandinsky 2.2 — новый шаг в направлении фотореализма, 2023: https://habr.com/ru/companies/sberbank/articles/747446/
[9]. Podell D., English Z., Lacey K., Blattmann A., Dockhorn T., Müller J., Penna J., Rombach R. SDXL: Improving Latent Diffusion Models forHigh-Resolution Image Synthesis, 2023: https://www.researchgate.net/publication/372136709_SDXL_Improving_Latent_Diffusion_Models_for_High Resolution_Image_Synthesis
[10]. Bochkovskiy A., Wang C.-Y., Mark Liao H.-Y. YOLOv4: Optimal Speed and Accuracy of Object Detection, 2020: https://arxiv.org/abs/2004.10934
[11]. Spodarets D. A Guide to the YOLO Family of Computer Vision Models, 2023: https://dataphoenix.info/a-guideto-the-yolo-family-of-computer-vision-models/
[12]. Li C. et. al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications, 2022: https://arxiv.org/abs/2209.02976
[13]. Wang C.-Y., Bochkovskiy A., Mark Liao H.-Y. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022: https://arxiv.org/abs/2207.02696