

Machine Learning Synthetic Data

Mostafa S. Ibrahim

Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher

PhD from Simon Fraser University - Canada

Bachelor / MSc from Cairo University - Egypt

Ex-(Software Engineer / ICPC World Finalist)



© 2023 All rights reserved.

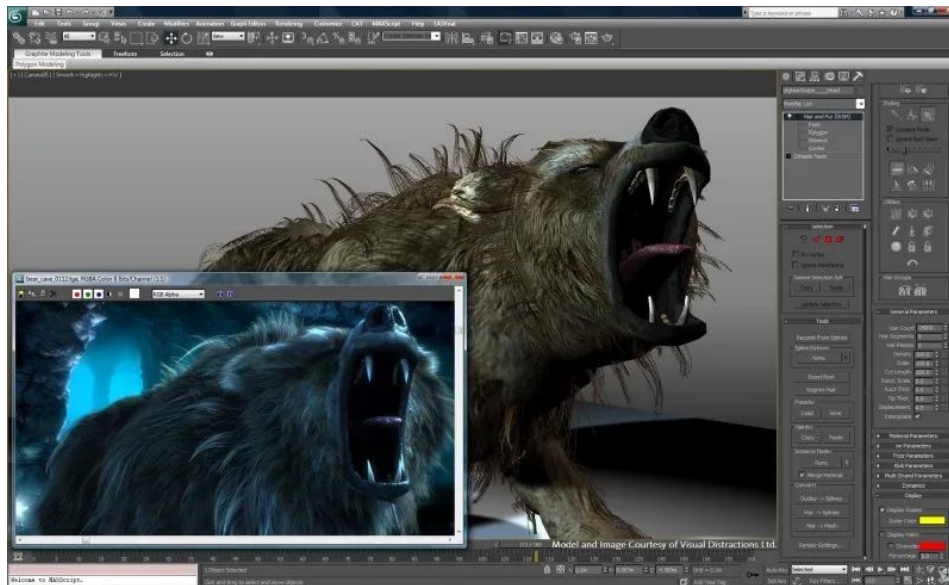
Please do not reproduce or redistribute this work without permission from the author

Synthetic data

- “Synthetic data in machine learning refers to **artificially generated data** that mimics the statistical **properties and characteristics** of real-world data.
- It is created to replicate the patterns, distributions, and relationships found in the original data.”
- The purpose of generating synthetic data is to address various challenges in machine learning, such as data **scarcity**, **privacy** concerns, or data augmentation for **model training**
- We use and evaluate carefully due missing properties or **domain gap**

Using 3D Models

- With 3D models, we can create 3D/2D videos for many things like people, cities, hand, face, etc
 - Think in cities in GTA game or car speed games
- We can control lighting conditions



Domain Gap

- Which images are **synthetic**?

A



B



C



D



- B and D are from games like GTA
- The visual gap between them is called domain gap
- **Domain adaptation techniques** aim to bridge the gap between the source domain (where the training data is available) and the target domain (where the model will be applied)

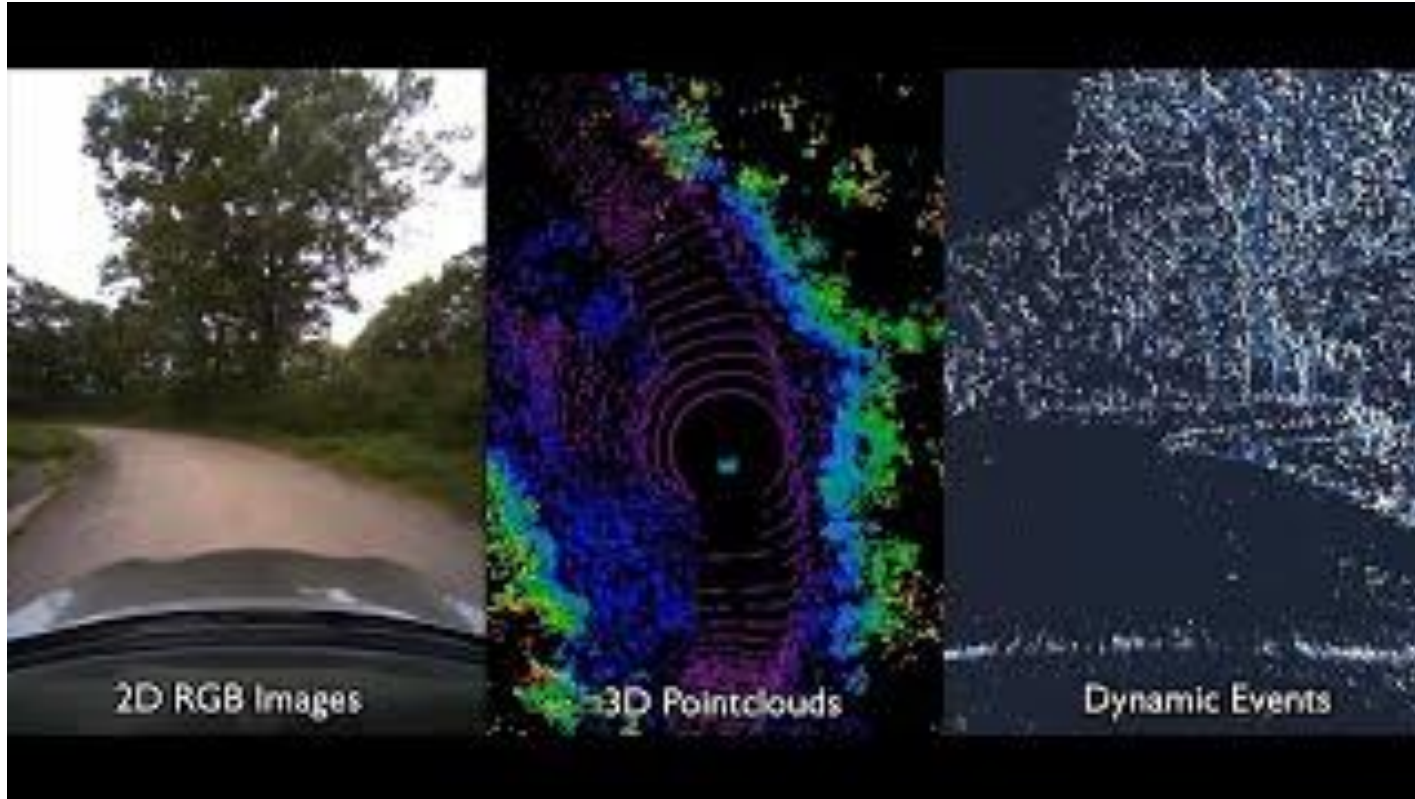
Simulators

- Simulators can **generate** realistic virtual environments with various road layouts, weather conditions, and traffic scenarios
- One common simulators nowadays are for self-driving

Neural Reconstruction Engine in NVIDIA DRIVE Sim



An open source simulator for self-driving cars



CARLA Simulator



Image Style Transfer

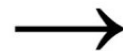
- Synthetic data can be created by applying style transfer techniques to existing images. By combining the content of one image with the style of another, new images with different artistic or visual styles can be generated

Content image



+

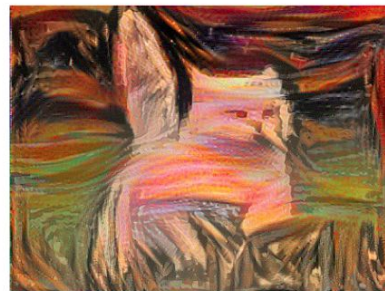
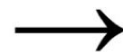
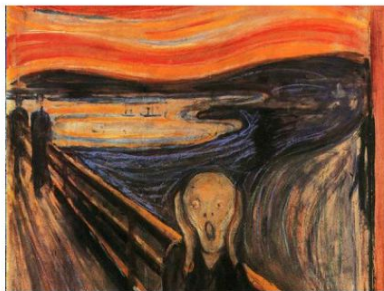
Style image



Output image



+



+



Synthetic data in NLP

- “Synthetic data can be generated by language models to create new text samples that resemble the language patterns and context found in real-world data.
- Techniques like GPT-3 or Transformer models can generate coherent and contextually relevant text for tasks such as language modeling, dialogue generation, or content creation”

Write a possible dialog between 2 people living in the seventh century about the possibility of humans reaching the moon



Person A: Have you ever wondered what lies beyond the vast expanse of the night sky?



Person B: Indeed, the stars and the moon have always fascinated me. But do you think it is possible for humans to ever reach the moon?

Person A: It is an intriguing thought, isn't it? Our ancestors believed the celestial bodies were deities or unreachable realms. But as time passes, we uncover new knowledge and technologies. Who knows what the future holds?

Person B: True, advancements in navigation and astronomy have allowed us to understand the movements of the heavenly bodies. But the moon, so distant and ethereal, seems like an insurmountable challenge.

Person A: Consider this: If we can build mighty ships to navigate treacherous seas and explore distant lands, is it not within the realm of possibility to build machines that can take us beyond the skies?

Person B: But the moon? It seems unimaginable to traverse such a vast distance. How could we ever achieve that?

Person A: Perhaps our understanding of physics will evolve, and we will discover new principles to propel us into the heavens. Just as the ingenious inventors of our time have crafted remarkable mechanisms, so too might future generations devise means of reaching celestial bodies.

Person B: It is difficult to conceive of such possibilities. Our world is so vast, and the moon appears unattainable, almost divine. Would it not be a violation of the natural order for humans to set foot on its surface?

Person A: Ah, but is it not human nature to explore and push the boundaries of what is known? Our insatiable curiosity has led us to new lands, new cultures, and new understanding. The moon, as distant as it seems, may hold secrets waiting to be unraveled.

Person B: You raise a compelling point. Perhaps it is not for us to decide what is possible or impossible. Our descendants may yet embark on a remarkable journey to the moon, unraveling its mysteries and expanding our understanding of the universe.

Person A: Precisely! Let us not limit our imagination or dismiss the potential of future generations. For even the wildest dreams of today may become the reality of tomorrow. Who knows? One day, humans might indeed set foot on the moon and leave an indelible mark on history.

Person B: How awe-inspiring it would be to witness such an extraordinary feat. Let us keep our minds open to the wonders that lie ahead and embrace the possibilities that await us, both on Earth and among the stars.

Synthetic data vs Augmented Data

- Synthetic data

- It is created **from scratch** using algorithms and **models** in a way that **approximate** the reality, with possible domain gap (Or based on seed but with complex generation **model**)
- Synthetic data is generated entirely from algorithms that does not rely on existing real-world data
- Purpose: critical if we can't collect real data (now) for scarcity, cost of privacy. Also can be used to enhance the training process

- Augmented Data

- it is created from **existing** dataset already in a way that preserves the characteristics of the original data while introducing **controlled modifications**
- Purpose: It can help address **training issues** like overfitting, data imbalance, or lack of diversity in the original dataset

Data Tracking

- It is very very very important to be able to track your data and metadata
- Say you collected some image dataset of people
 - Record metadata for all relevant information
 - For example, camera details, location (indoor, outdoor), lighting conditions, etc
 - Date of data collection
 - ID and hashcode for users (to split train/val if needed)
 - Give the dataset some name
- Whenever you create data from data (e.g. oversampling), you must be able to track that
- **Never lose raw data**
 - You crop the image and remove origin. What is something is wrong?
 - You replace numbers with their mean and variance. What if we need other stats?
- What is proper data storage format? DB? Json?

“Acquire knowledge and impart it to the people.”

“Seek knowledge from the Cradle to the Grave.”

