Laporan Synthetic Data Creation

Kevin Olind H.N./1103210140

1. Pendahuluan

Data sintetis adalah data yang dihasilkan oleh model matematis atau algoritma untuk meniru pola atau struktur dari data dunia nyata, tetapi tidak berasal dari kejadian atau entitas nyata. Data ini sangat berguna dalam melatih model kecerdasan buatan (AI), mengatasi masalah privasi, dan meningkatkan kualitas data pelatihan untuk berbagai aplikasi, terutama yang memerlukan data dengan sifat statistik tertentu.

2. Keunggulan dan Alasan Penggunaan Data Sintetis

Data sintetis banyak digunakan untuk melengkapi atau menggantikan data asli dengan alasan berikut:

- Pengelolaan Data yang Rumit: Data asli sering kali tidak terstruktur, sulit diatur, dan memerlukan upaya besar dalam pengolahan dan pemeliharaan.
- Masalah Privasi: Data sensitif, seperti data medis atau data pribadi, memiliki pembatasan dalam hal berbagi dan pemrosesannya.
- Bias Data: Data asli sering kali mengandung bias, yang dapat mempengaruhi kinerja model yang dilatih dengan data tersebut.
- Biaya Tinggi: Pengumpulan data asli dan anotasi sering memerlukan biaya yang tinggi.

Keunggulan data sintetis:

- Kemudahan Pengelolaan: Data sintetis mudah diatur dan lebih terstruktur dibandingkan data asli.
- Privasi yang Terjaga: Dapat menghasilkan data yang tidak terhubung langsung dengan entitas atau individu nyata.
- Pengurangan Bias: Dapat dilatih untuk menghasilkan data yang lebih seimbang dan adil.
- Ketersediaan Data: Dapat digunakan untuk melengkapi data asli dan meningkatkan kualitas pelatihan model.

3. Metode untuk Menghasilkan Data Sintetis

Ada beberapa pendekatan untuk menghasilkan data sintetis, yang masing-masing memiliki kelebihan dan kekurangan tergantung pada kebutuhan aplikasi:

- CAD & Blender: Menggunakan perangkat lunak seperti Blender untuk membuat model 3D yang realistis dan menghasilkan dataset gambar fotorealistik. Proses ini dapat digunakan untuk pembuatan wajah sintetis, pengawasan satwa liar, atau simulasi dunia nyata lainnya.
- Model Generatif: Seperti Generative Adversarial Networks (GAN), Transformers, dan Diffusion Models. Model ini digunakan untuk memperluas dataset, mengatasi

ketidakseimbangan data, dan menjaga privasi. Contoh penggunaan termasuk pembuatan gambar medis, pengenalan objek, dan aplikasi dalam kendaraan otonom.

Blender dan alat terkait seperti BlenderProc memberikan pipeline modular untuk pembuatan data sintetis yang realistis. BlenderProc memungkinkan pembuatan adegan 3D otomatis, simulasi fisik, dan pemrosesan paralel untuk menghasilkan data dalam skala besar.

4. Tantangan dalam Penggunaan Data Sintetis

Walaupun data sintetis menawarkan banyak keuntungan, ada beberapa tantangan yang perlu diperhatikan:

- Privasi Tidak Terjamin Secara Otomatis: Data sintetis masih rentan terhadap serangan privasi dan dapat mengungkapkan informasi dari data asli.
- Kesulitan dalam Menangkap Outlier: Kejadian langka atau data outlier sulit untuk diproduksi dengan metode data sintetis.
- Evaluasi Privasi yang Sulit: Pengukuran privasi harus diterapkan pada mekanisme pembangkitan data, bukan hanya dataset itu sendiri.
- Model Black Box yang Tidak Transparan: Model generatif yang kompleks (seperti GANs) sering kali memiliki proses yang tidak transparan, membuatnya sulit untuk mengevaluasi keakuratan atau privasi data yang dihasilkan.

5. Pembuatan Data Sintetis dengan Diffusion Models

Diffusion models adalah salah satu metode yang efektif dalam menghasilkan data sintetis berkualitas tinggi. Model ini bekerja dengan menambahkan noise Gaussian ke data asli secara bertahap dan kemudian menghilangkan noise tersebut untuk menghasilkan data yang mendekati distribusi data nyata.

Contoh model berbasis diffusion yang populer adalah Stable Diffusion, yang dapat digunakan untuk menghasilkan gambar sintetis dari teks (text-to-image). Proses kerja Stable Diffusion meliputi:

1. Proses Diffusi: Menambahkan dan menghilangkan noise secara berulang untuk mempelajari distribusi data.
2. Encoder/Decoder Gambar: Mengompres gambar untuk meningkatkan efisiensi.
3. Conditional Encoder: Dapat menggunakan informasi tambahan, seperti teks, untuk mengkondisikan hasil output.

Stable Diffusion memiliki potensi besar dalam aplikasi seperti peningkatan dataset medis, pembuatan gambar senjata untuk deteksi objek, dan pelatihan model yang memerlukan data terbatas.

6. Tantangan dan Peluang dalam Menggunakan Data Sintetis

Meskipun data sintetis membawa banyak manfaat, ada beberapa tantangan yang perlu diperhatikan:

- Overfitting: Terjadi ketika model "terlalu banyak belajar" dari data pelatihan dan tidak dapat menggeneralisasi dengan baik pada data baru. Data sintetis yang terlalu sederhana atau memiliki pola yang konsisten berisiko menyebabkan overfitting.
- Bias dalam Data Sintetis: Data sintetis dapat meniru bias yang ada dalam data asli, sehingga memperburuk bias yang ada dalam model.
- Biaya Komputasi: Proses pembuatan data sintetis berkualitas tinggi bisa sangat mahal dan memerlukan sumber daya komputasi yang besar.
- Kualitas Data Sintetis: Gambar sintetis yang buruk atau tidak realistis dapat memengaruhi kinerja model dalam situasi dunia nyata. Metrik seperti Frechet Inception Distance (FID) dan Inception Score (IS) digunakan untuk menilai kualitas data sintetis dan kedekatannya dengan data dunia nyata.

7. Point Cloud dalam Pembuatan Data Sintetis

Point cloud adalah kumpulan titik dalam ruang 3D yang merepresentasikan objek atau lingkungan. Setiap titik dalam point cloud biasanya memiliki koordinat [x, y, z], serta atribut tambahan seperti warna, norma permukaan, dan albedo. Point cloud sangat penting dalam berbagai aplikasi, seperti pemetaan 3D dengan teknologi LiDAR dan perangkat Augmented Reality (AR).

Beberapa format untuk menyimpan point cloud adalah:

- PLY (Polygon File Format)
- STL (Standard Tessellation Language)
- OFF (Object File Format)
- 3DS (3D Studio)
- DAE (Digital Asset Exchange)

Point cloud digunakan untuk meningkatkan akurasi dan keberagaman dataset 3D, memungkinkan pemrosesan citra dan analisis yang lebih baik dalam aplikasi seperti kendaraan otonom, pemetaan 3D, dan AR.

8. Kesimpulan

Data sintetis merupakan alat yang sangat berguna untuk memperkaya dataset, mengurangi bias, dan menjaga privasi dalam pengembangan model AI. Namun, penggunaannya juga membawa tantangan besar seperti overfitting, biaya komputasi, dan risiko bias. Oleh karena itu, desain data sintetis harus dilakukan dengan hati-hati dan evaluasi model harus tetap dilakukan menggunakan data dunia nyata untuk memaksimalkan efektivitas dan aplikasi praktisnya.

Penggunaan metode seperti diffusion models, Blender, dan point cloud memberikan pendekatan yang kuat untuk menciptakan data sintetis berkualitas tinggi yang dapat digunakan dalam berbagai aplikasi, mulai dari pengenalan objek hingga segmentasi medis.