KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Generative Modelle: 08_07Diskussion_ModelCollapse

Model Collapse

Erarbeitet von
Selina Müller, M.Sc.

Lernziele

- Du kannst die Beobachtung von Qualitäts- und Diversitätsverlust bei Modellen, die mit generierten Bilddaten trainiert wurden, beschreiben

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Wie du weißt, benötigen Machine Learning Modelle Daten, um trainiert zu werden. Wenn die Datenbeschaffung sehr aufwändig, und damit kostenintensiv ist, oder wenn der Datensatz sensible Informationen beinhaltet, dann können synthetische Daten zum Einsatz kommen. Das sind künstlich erstellte Daten.

Quelle [1]

Generative Modelle benötigen sehr viele Trainingsdaten. Kann man dafür nicht auch synthetische Daten verwenden? Wie sich herausgestellt hat, könnte das zu Problemen führen. Es konnte gezeigt werden, dass Modelle, welche mittels synthetischer Daten trainiert werden, weniger diverse Outputs mit niedrigerer Qualität generieren. Dies wurde für Bild- und Textdaten beobachtet.

Quelle [2] [3] [4] [5]

Mit unterschiedlichen Modell-Architekturen und Datensätzen wurden Versuche durchgeführt, bei denen Schleifen erstellt wurden, wo der Output des einen Modells, zu den Trainingsdaten des folgenden Modells, wurde.

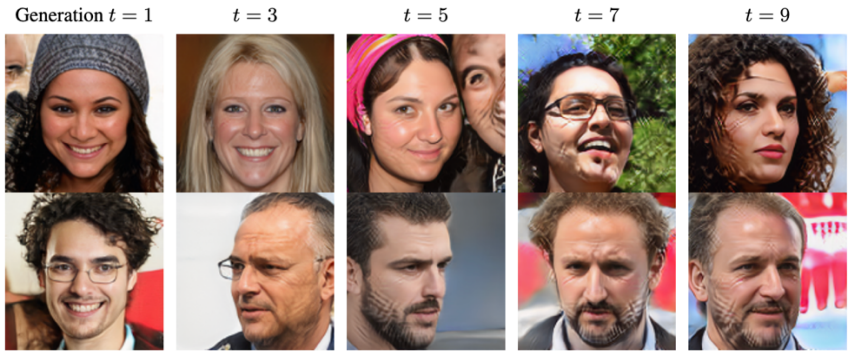

Figure 1: **Training generative artificial intelligence (AI) models on synthetic data progressively amplifies artifacts.** As synthetic data from generative models proliferates on the Internet and in standard training datasets, future models will likely be trained on some mixture of real and synthetic data, forming an *autophagous ("self-consuming") loop*. Here we highlight one potential unintended consequence of autophagous training. We trained a succession of StyleGAN-2 [1] generative models such that the training data for the model at generation $t \geq 2$ was obtained by synthesizing images from the model at generation $t - 1$. This particular setup corresponds to a fully synthetic loop in Figure 3. Note how the cross-hatched artifacts (possibly an architectural *fingerprint*) are progressively amplified in each new generation. Additional samples are provided Appendices C and D.

Generierte Bilder aus **Quelle [2]**

Quelle [2]

In diesem Beispiel seht ihr Bilder, die von Modellen der StyleGAN-2 Architektur generiert wurden. Dabei wurde jedes Modell, ab der zweiten Generation, mit generierten, also rein synthetischen Bildern des vorherigen Modells trainiert. Man sieht deutlich, wie mit

steigender Anzahl an Generationen, hier mit t bezeichnet, schraffierte Artefakte schrittweise verstärkt werden.

Quelle [6]

Dieser Qualitäts- und Diversitätsverslust ist abhängig von der Anzahl an synthetischen Daten im Trainingsdatensatz. Tatsächlich kann eine bestimmte Menge das Modell stärken. Aber wenn die Menge an synthetischen Daten eine kritische Anzahl übersteigt, leidet das Modell darunter. Diese Verschlechterung wird als „Model Collapse" – also „Modellzusammenbruch" bezeichnet.
Um die Qualität der Modelle zu erhalten sind frische, reale Daten beim Training unerlässlich. Einige Wissenschaftler*innen argumentieren, dass dies zu einem Problem werden könnte, da große Sprachmodelle beim Training auf Daten aus dem Internet zurückgreifen, wo schon jetzt generierte Bilder und Texte veröffentlicht werden. Noch ist es unklar, wie diese generierten Inhalte rückverfolgt werden können.

Quelle [2]

Quellen

Quelle [1] Nikolenko, S. I. (2021). Synthetic data for deep learning (Vol. 174). Springer Nature. [Chapter 1 Introduction: The Data Problem, p. 1-17]. https://doi.org/10.1007/978-3-030-75178-4

Quelle [2] Casco-Rodriguez, J., Alemohammad, S., Luzi, L., Ahmed, I., Babaei, H., LeJeune, D., Siahkoohi, A., & Baraniuk, R. G. (2023). Self-Consuming Generative Models go MAD. cs.LG. https://doi.org/10.52591/lxai202312101

Quelle [3] Hataya, R., Bao, H., & Arai, H. (2023). Will Large-scale Generative Models Corrupt Future Datasets? *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 20498–20508. https://doi.org/10.1109/ICCV51070.2023.01879

Quelle [4] Martínez, G. S., Watson, L., Reviriego, P., Hernández, J. A., Juárez, M., & Sarkar, R. (2024). Towards understanding the interplay of generative artificial intelligence and the internet. In Lecture notes in computer science (pp. 59–73). https://doi.org/10.1007/978-3-031-57963-9_5

Quelle [5] Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2024). The Curse of Recursion: Training on Generated Data Makes Models Forget (arXiv:2305.17493). arXiv. http://arxiv.org/abs/2305.17493

Quelle [6] Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. CVPR. https://doi.org/10.1109/cvpr.2019.00453

Disclaimer

Transkript zu dem Video „08 Generative Modelle: Model Collapse“, Selina Müller.