Transformacja z przestrzeni ukrytej do przestrzeni interpretowalnej z możliwością manipulacji w modelach typu GAN [1]

Agnieszka Klimek[1], Anna Prałat[1], Daniel Zdancewicz[1]

I. Definicja problemu badawczego

Projekt skupia się na zbadaniu możliwości dokonywania transformacji przestrzeni ukrytej do przestrzeni interpretowalnych oraz umożliwiających manipulację, w celu unikania generowania pewnych cech, które mogą być problematyczne. Hipoteza badawcza zakłada, że taka transformacja jest możliwa i pozwala na skuteczne sterowanie procesem generacji danych bez wybranych cech.

Przebieg badań:

1) Przejrzenie modeli generatywnych i selekcja tych, które potencjalnie pozwalają na transformację przestrzeni ukrytej w dowolną inną.
2) Znalezienie zbioru danych, który dla każdego przykładu dostarcza wiele etykiet, co umożliwia ich eksplorację lub usunięcie.
3) Zaproponowanie architektury neuronowej do transformacji z przestrzeni ukrytej do interpretowalnej przestrzeni wraz z nauką modelu.
4) Demonstracja wizualna z zaproponowaniem i zmierzeniem metryk.

II. Przegląd literatury

Prace badawcze, przez które zapoznaliśmy się z tematyką interpretacji oraz manipulacji przestrzeniami ukrytymi to:

- Controlling generative models with continuous factors of variations[2] - Znajdowanie znaczących kierunków przestrzeni ukrytej w modelach GAN kontrolujących właściwości generowanych obrazów.
- GANSpace: Discovering Interpretable GAN Controls[3] - Identyfikowanie ważnych kierunków przestrzeni ukrytej bazujących na metodzie PCA.
- Interpreting the Latent Space of GANs for Semantic Face Editing[4] - Interpretacja ukrytej semantyki, nauczonej przez sieć GAN, pozwalająca na edycję twarzy; do separacji poszczególnych cech wykorzystano SVM.
- Unsupervised Discovery of Interpretable Directions in the GAN Latent Space[5] - Stworzenie interpretowalnej przestrzeni przy użyciu uczenia nienadzorowanego.

Również zostały przejrzane prace dotyczące interpretacji i problemu inwersji problemu:

- Interpretable latent space and inverse problem in deep generative models[6] - Wykład wprowadzeniowy do tematyki przestrzeni ukrytej
  - youtube.com/watch?v=8Hm4ad5QlUE

III. Wykorzystane technologie i narzędzia

- Python[7] - język programowania
  - python.org/
- Numpy[8] - moduł do operacji matematycznych
  - numpy.org/
- Pandas[9] - moduł do manipulacji zbiorami danych
  - pandas.pydata.org/
- PyTorch[10]
  - pytorch.org/
- TensorBoard[11]
  - tensorflow.org/tensorboard/

IV. Wymagania projektu

A. Wymagania minimalne

- Reimplementacja architektury GAN - niezbędna do podstawowych prac i tworzenia bazowego modelu.
- Pobranie i wstępna obróbka zbioru danych z wieloma etykietami - wymagane do nauki modelu oraz jako stały punkt odniesienia w badaniach.
- Utworzenie przepływów odpowiedzialnych za interpretację modeli przy wykorzystaniu TensorBoard oraz PyTorch.

Konieczne jest, aby zbiór danych zawierał wiele etykiet ze względu na potrzebę blokowania poszczególnych cech, w celu zbadania możliwości wykluczania poszczególnych elementów z generowanych obrazów.

B. Wymagania końcowe

- Utworzenie interpretacji istniejącego modelu przy wykorzystaniu utworzonych przepływów.
- Przeprowadzenie transformacji z przestrzeni ukrytej do przestrzeni interpretowalnej / kontrolowanej.
- Demonstracja zdolności do kontroli przestrzeni - Przekształcenia ukazujące możliwości kontroli przestrzeni.
- Ewaluacja i demonstracja wizualna z zastosowaniem metryk badawczych.

V. Zagrożenia projektu

- Trudności z interpretacją przestrzeni ukrytej.
- Brak możliwości manipulacji przestrzenią ukrytą.
- Ograniczenia techniczne istniejących modeli.

[1]Wydział Informatyki i Telekomunikacji, Politechnika Poznańska, Poznań {agnieszka.r.klimek, anna.k.pralat, daniel.zdancewicz}@student.put.poznan.pl

- Problemy wynikające z implementacji blokowania poszczególnych cech obrazu.
- Problematyczne pozyskanie zbioru danych o wielu etykietach.

W wypadku braku możliwości interpretacji przestrzeni ukrytej czy też manipulacji interpretowalną przestrzenią ukrytą skupimy się na elemencie dotyczącym możliwością samej manipulacji przestrzenią.

Bibliografia

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[2] A. Plumerault, H. L. Borgne, and C. Hudelot, "Controlling generative models with continuous factors of variations," 2020.

[3] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," 2020.

[4] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," 2020.

[5] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the gan latent space," 2020.

[6] P. Zatko, I. Poupyrev, R. E. Guerrab, and R. Dugan. Some cool motion sensor stuff. Youtube. [Online]. Available: https://www.youtube.com/watch?v=mpbWQbkl8˙g#t=20m15s

[7] G. van Rossum, "Python tutorial," Centrum voor Wiskunde en Informatica (CWI), Amsterdam, Tech. Rep. CS-R9526, May 1995.

[8] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[9] T. pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3509134

[10] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style"protect" @normalcr"relax-high-performance-deep-learning-library.pdf

[11] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/