Rozplątana reprezentacja w modelach generatywnych z zastosowaniem w chemoinformatyce

1 Motywacja i cel badań

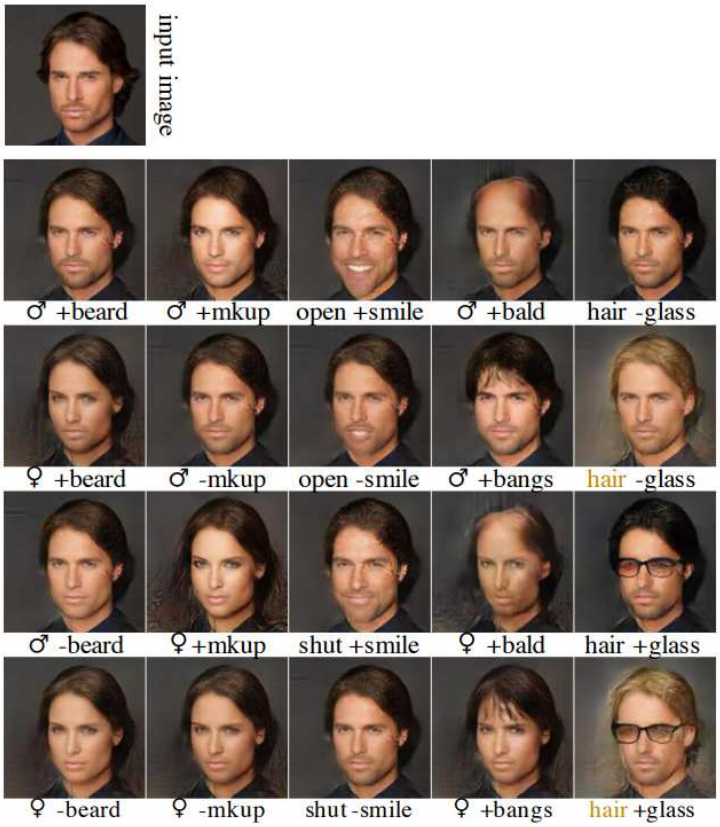
Tworzenie nowych przykładów (danych) jest jednym z podstawowych problemów uczenia maszynowego, który można rozwiązać, stosując modele generatywne (ang. generative models) [2]. W praktycznych zastosowaniach generowanie nowych przykładów, bez spełnienia dodatkowych warunków, nie jest jednak wystarczające. W szczególności przy konstrukcji związków chemicznych chcielibyśmy określić jakich właściwości oczekujemy od nowego związku (rozpuszczalność, toksyczność, oddziaływanie z różnymi białkami/enzymami/DNA). Ponadto, chcemy być w stanie zmienić wybrane cechy związku bez modyfikacji pozostałych jego właściwości. Podobna sytuacja występuje w przypadku obrazów – chcielibyśmy dokonywać modyfikacji/edycji cech twarzy bez zmiany tożsamości osoby na zdjęciu, Rysunek 1. O ile powyższe wymagania mają istotne znaczenie w praktyce, o tyle trudno je zrealizować używając standardowych modeli generatywnych. Do uzyskania takiego modelu konieczne jest stworzenie rozplątanej reprezentacji (ang. disentangled representation), w której uzyskujemy współrzędne odpowiedzialne za natężenie poszczególnych cech (właściwości).

Celem projektu jest opracowanie modelu generatywnego z rozplątaną reprezentacją, który:

- pozwala na generowanie nowych przykładów o zadanych własnościach,
- umożliwia modyfikację wybranych cech obiektu bez zmiany pozostałych charakterystyk.

Uzyskanie takiego modelu będzie miało zastosowanie w typowych dziedzinach informatyki, takich jak przetwarzanie obrazów, ale co ważniejsze przyczyni się do rozwoju innych dziedzin, takich jak chemoinformatyka. Obecnie przestrzeń związków chemicznych jest szacowana na poziomie 10^{60} związków, co oznacza że do chwili obecnej zbadano dopiero jej niewielki ułamek w kierunku leczenia różnych schorzeń. Dokładniej, baza ChEMBL[1] związków o potwierdzonej aktywności zawiera około 10^7 rekordów dla różnych strukturalnie związków chemicznych. Nawigacja w całej przestrzeni związków (nawet wirtualnie wygenerowanej) jest poza zasięgiem współczesnej chemii obliczeniowej – baza GDB[2] w chwili obecnej osiągnęła wartość około 10^{11} cząsteczek o ilości atomów do 17. Algorytmy służące do celowanej eksploracji (tworzenie związków o zadanych właściwościach) całej przestrzeni chemicznej może przyczynić się do identyfikacji strukturalnie nowych związków bioaktywnych, które mogą mieć bardzo duży potencjał komercyjny (w tworzeniu leków).

[1] `https://www.ebi.ac.uk/chembl/`
[2] `https://gdb.unibe.ch/downloads/`



Rysunek 1: Przykład modyfikacji wybranych cech na obrazie [4].

2 Zespół

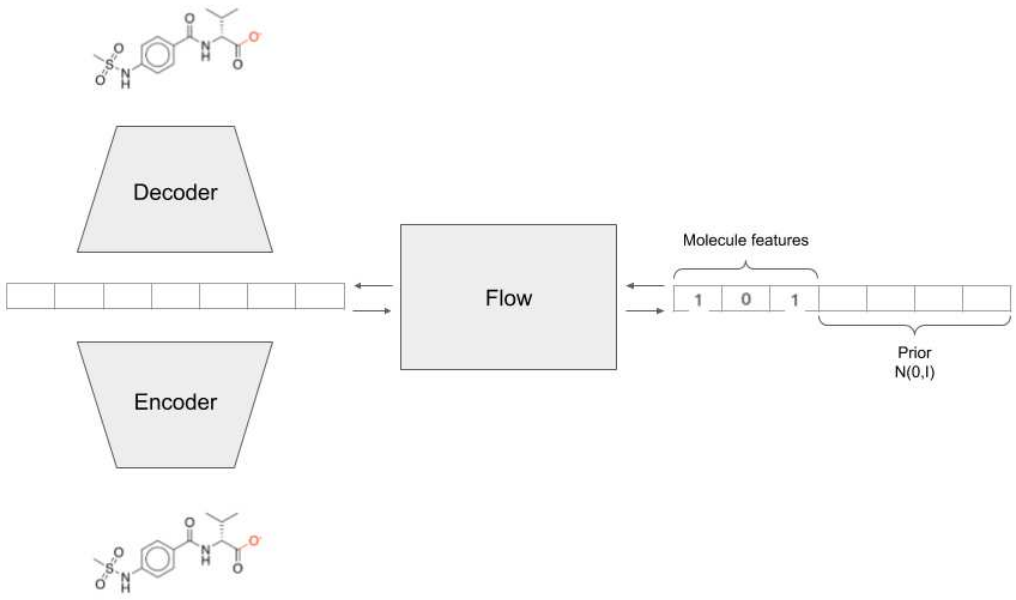
Do realizacji tego projektu konieczna jest współpraca specjalistów z zakresu głębokiego uczenia oraz chemoinformatyki. Poza kierownikiem w skład zespołu wejdzie:

- Maciej Wołczyk [doktorant WMiI], który ma doświadczenie w tworzeniu modeli generatywnych [7],
- Łukasz Maziarka [Ardigen S.A. oraz doktorant WMiI], który na co dzień stosuje modele głębokiego uczenia w zagadnieniach chemicznych [5],
- Rafał Kurczab [IF PAN], który jest specjalistą z zakresu chemoinformatyki [8].

W dalszej perspektywie chcemy wykorzystać zdobyte doświadczenie oraz nawiązaną współpracę i starać się o finansowanie ze środków zewnętrznych (w tym NCN oraz NCBiR) interdyscyplinarnych badań dotyczących rozwoju metod uczenia maszynowego w konstrukcji związków aktywnych, które mają potencjał aby stać się lekami.

3 Metodologia

Do nauczenia modelu, który potrafi modyfikować wybrane cechy obiektu, konieczne jest przekazanie informacji o tych cechach dla przykładów treningowych. Innymi słowy, musimy mieć dostęp do danych etykietowanych, dla których wiemy czy określona właściwość jest zdefiniowana. Jako że etykiety są kosztowne do pozyskania, będziemy pracować w sytuacji pół-nadzorowanej (ang. semi-supervised). Oznacza to, że dla każdego przykładu znamy tylko niewielką liczbę jego właściwości (etykiet). Takie założenie jest zwykle spełnione w praktyce. W przypadku związków chemicznych, wykorzystamy bazę ChEMBL,



Rysunek 2: Schemat proponowanego modelu. Ukryta reprezentacja koduje wybrane cechy przykładu na zmiennych kategorycznych. Zmiana wartości tych zmiennych ma spowodować zmianę tylko tych cech bez modyfikacji pozostałej części informacji.

gdzie aktywność związku jest potwierdzona dla około 2-3 białek (dla każdego związku mogą to być inne białka). Aktywność dla całego zbioru receptorów nie jest osiągalna. Kierownik projektu ma doświadczenie w konstrukcji modeli pół-nadzorowanych [6].

Punktem wyjścia dla nas jest praca [4], która przedstawia model auto-enkoderowy z rozplątaną reprezentacją ukrytą. Wspomniany model musi jednak być uczony w klasycznej sytuacji nadzorowanej, w której konieczna jest znajomość wszystkich etykiet dla każdego przykładu. Jest to bardzo restrykcyjne założenie, które nie jest zwykle spełnione w rzeczywistych danych. Ponadto, ten model nie jest generatywny, czyli pozwala jedynie na modyfikację cech danego obiektu, ale nie pozwala na generowanie nowych danych. Pierwszym krokiem będzie rozwinięcie tego modelu, zapewniając spełnienie dodatkowych wymagań. W tym celu zastosujemy modele odwracalne (ang. invertible normalized flows), tak aby móc mapować wyjściową reprezentację autoenkodera do reprezentacji rozplątanej, a także generować dane o określonych cechach (kierunek przeciwny). Schemat planowanego podejścia jest przedstawiony na Rysunku 2. Będziemy rozważać także podejście adwersarialne [3]. Modele adwersarialne są trudniejsze w uczeniu, ale mają znaczenie większe spektrum zastosowań, gdyż nie wymagają definiowania precyzyjnego kryterium uczenia.

4 Plan badań

Badania rozpoczniemy od ewaluacji istniejących rozwiązań do tworzenia modeli generatywnych i reprezentacji rozplątanej w przypadku danych obrazowych. Z naszych dotychczasowych badań wynika, że jednym z najbardziej obiecujących podejść do tworzenia rozplątanej reprezentacji jest model MSP [4], który jednak nie pozwala na generowanie nowych danych oraz wymaga danych w pełni etykietowanych. Zbadamy także podejścia adwarsarialne [3].

Na podstawie uzyskanych wyników będziemy budować własny model, który spełni wszystkie wymagane warunki w projekcie: (1) generatywność (2) rozplątana reprezentacja (3) trenowanie z użyciem niewielkiej liczby etykiet. Naszym punktem wyjścia będzie model

MSP. Planujemy dodać model odwracalnej sieci neuronowej z odpowiednim rozkładem w przestrzeni ukrytej aby zapewnić zarówno generatywność jak i rozplątaną reprezentację. Korzystając z doświadczenia kierownika w tworzeniu modeli częściowo-nadzorowanych, dostosujemy model do treningu z użyciem niewielkiej liczby etykiet.

Kolejnym krokiem będzie adaptacja zbudowanego modelu do danych chemicznych. W tym celu zastosujemy architekturę opartą o konwolucyjne sieci grafowe, która jest w stanie przetwarzać związki chemiczne. W szczególności wykorzystamy opracowaną przez nas przestrzenną sieć grafową [1], która dobrze nadaje się do przetwarzania grafów cząsteczek chemicznych. Tutaj kluczowe będzie również dostosowanie modelu do niewielkiej liczby danych treningowych oraz etykiet, gdyż takie dane występują w rzeczywistości.

Literatura

[1] Tomasz Danel, Przemysław Spurek, Jacek Tabor, Marek Śmieja, Łukasz Struski, Agnieszka Słowik, and Łukasz Maziarka. Spatial graph convolutional networks. *arXiv preprint arXiv:1909.05310*, 2019.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in neural information processing systems*, pages 5967–5976, 2017.

[4] Xiao Li, Chenghua Lin, Chaozheng Wang, and Frank Guerin. Latent space factorisation and manipulation via matrix subspace projection. *In Proceedings of the 37th International Conference on Machine Learning*, 2020.

[5] Łukasz Maziarka, Agnieszka Pocha, Jan Kaczmarczyk, Krzysztof Rataj, Tomasz Danel, and Michał Warchoł. Mol-cyclegan: a generative model for molecular optimization. *Journal of Cheminformatics*, 12(1):1–18, 2020.

[6] Marek Śmieja, Łukasz Struski, and Mário AT Figueiredo. A classification-based approach to semi-supervised clustering with pairwise constraints. *Neural Networks*, 2020.

[7] Marek Śmieja, Maciej Wołczyk, Jacek Tabor, and Bernhard C Geiger. Segma: Semi-supervised gaussian mixture autoencoder. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[8] Sabina Smusz, Rafał Kurczab, Grzegorz Satała, and Andrzej J Bojarski. Fingerprint-based consensus virtual screening towards structurally new 5-ht6r ligands. *Bioorganic & medicinal chemistry letters*, 25(9):1827–1830, 2015.