Metody uczenia głębokiego w naukach farmaceutycznych

Autoreferat

Tomasz Danel

Uniwersytet Jagielloński

Wydział Matematyki i Informatyki

Instytut Informatyki i Matematyki Komputerowej

Katedra Uczenia Maszynowego

Promotor rozprawy doktorskiej:
dr hab. Igor Podolak

Kraków 2023

Podziękowania

Pragnę serdecznie podziękować mojemu promotorowi, drowi hab. Igorowi Podolakowi, za opiekę, cierpliwość, motywację i cenne uwagi.

Dziękuję również wszystkim współpracownikom i współautorom, bez których ten cykl nie mógłby powstać. Jestem wdzięczny za możliwość pracy z wybitnymi ekspertami różnych dyscyplin nauki.

Tomasz Danel

Spis treści

Wstęp

Niniejsza praca doktorska jest zbiorem opracowanych metod uczenia głębokiego ze szczególnym ukierunkowaniem na zastosowania w problemach farmaceutycznych. Moim celem było rozwinięcie narzędzi opartych o sieci neuronowe na różnych etapach odkrywania leków, aby skrócić ten proces i zredukować jego koszty. Szczególną uwagę poświęciłem eksploracji możliwości modeli grafowych.

Dane medyczne, biologiczne i chemiczne cechują się specyficzną strukturą, którą należy wziąć pod uwagę tworząc modele uczenia maszynowego. W szczególności związki chemiczne są często reprezentowane jako grafy opatrzone dodatkowo cechami wierzchołków i krawędzi, które odpowiadają atomom i wiązaniom cząsteczki chemicznej. Do przetwarzania tego typu danych coraz częściej używa się grafowych sieci neuronowych, które zostały znacząco rozwinięte w przeciągu ostatnich lat. Modele te są obecnie używane między innymi do przewidywania własności związków chemicznych oraz proponowania nowych kandydatów na lek.

W niniejszym cyklu znalazło się sześć publikacji związanych z zagadnieniem sieci grafowych użytych do modelowania związków chemicznych [P1, P2, P3, P4, P5, P6]. Pierwsza z nich wprowadza nową architekturę sieci grafowej, uwzględniającą położenia wierzchołków w przestrzeni. Druga zajmuje się reprezentacją związków, które podawane są na wejściu do tego typu sieci grafowych. Trzecia praca jest pracą przeglądową opisującą generowanie nowych związków chemicznych, które mogą zostać rozwinięte jako potencjalne leki. W czwartej pracy pokazujemy zastosowanie sieci grafowych w studium przypadku związanym z modulacją metabolizmu. Piąta praca jest propozycją nowej interpretowalnej sieci grafowej, która tłumaczy swoje predykcje na podstawie przykładów. W końcu szósta praca pokazuje, jak sieci grafowe mogą być użyte do optymalizacji związków chemicznych.

Innym powszechnie używanym typem danych w obszarze nauk medycznych oraz farmaceutycznych są dane obrazowe. Mogą to być obrazy naturalne (zdjęcia obiektów wykonane standardowymi aparatami fotograficznymi) lub częściej obrazy pochodzące z rozmaitych instrumentów medycznych lub specjalistycznych mikroskopów, np. zdjęcia rentgenowskie lub mikroskopii fluorescencyjnej. W drugiej części cyklu znalazły się publikacje, których celem było udoskonalenie obecnych technik przetwarzania obrazów mogących znaleźć zastosowanie w obszarze chemii, medycyny i farmacji [P7, P8, P9]. Jedna z tych prac opisuje proces pozyskiwania danych i użycie sieci neuronowych na urządzeniach mobilnych do szybkiej wstępnej klasyfikacji gatunków porostów. Pozostałe dwie prace prezentują grafowe podejście do przetwarzania obrazów, co nawiązuje technicznie do pierwszej części cyklu.

Wprowadzenie. Poniżej zostały wprowadzone pojęcia istotne dla zrozumienia cyklu i powtarzające się w publikacjach. W kolejnych rozdziałach przedstawione zostały streszczenia prac P1–P9 wchodzących w skład rozprawy z cyklu publikacji. W ostatnim rozdziale został załączony mój życiorys naukowy.

Podstawowe pojęcia i notacja

W przedstawionych pracach cyklu przez **graf** będę zwykle rozumiał nieskierowany graf ocechowany $\mathcal{G} = (V, E, a_V, a_E)$, gdzie V to zbiór wierzchołków, $E \subseteq \{\{v_i, v_j\} : v_i, v_j \in V\}$ to zbiór krawędzi, natomiast $a_V : V \to \mathbb{R}^{F_V}$ to przypisanie cech do poszczególnych wierzchołków oraz $a_E : E \to \mathbb{R}^{F_E}$ to przypisanie cech krawędzi. **Graf molekularny** to szczególny przykład nieskierowanego grafu ocechowanego, gdzie wierzchołki odpowiadają atomom, a krawędzie wiązaniom związku chemicznego. Cechami wierzchołków są deskryptory atomów konieczne do identyfikacji związku, np. zakodowany symbol pierwiastka, a cechami krawędzi są deskryptory wiązań, np. krotność wiązania chemicznego.

Splotowa/konwolucyjna sieć neuronowa to rodzaj architektury sieci neuronowej przeznaczony pierwotnie do przetwarzania obrazów. Obrazy mogą być reprezentowane w komputerze jako 3-wymiarowe tensory $X \in \mathbb{R}^{W \times H \times D}$, gdzie W oraz H to szerokość i wysokość obrazu, a D odpowiada liczbie kanałów (przy obrazach zwykle $D = 3$ i odpowiada kanałom przestrzeni kolorów RGB). W tak reprezentowanych danych można zdefiniować sąsiedztwo piksela o współrzędnych (i, j) i promieniu sąsiedztwa r:

$$\mathcal{N}((i, j), r) = \{(x, y) : |x - i| \leqslant r \wedge |y - j| \leqslant r\}. \tag{1}$$

Sieci konwolucyjne używają pojęcia sąsiedztwa do filtrowania obrazów. Parametrami trenowalnymi takich sieci są wagi $F \in \mathbb{R}^{(2r+1)\times(2r+1)\times D' \times D}$ zwane też filtrami. Dla uproszczenia będę zakładał konwolucje o promieniu r, z przesunięciem filtra równym 1. W ten sposób piksel (i, j) obrazu przekształconego będzie odpowiadał temu samemu pikselowi obrazu wejściowego oraz jego r-sąsiedztwu. Każda warstwa sieci konwolucyjnej oblicza nową reprezentację obrazu $X' \in \mathbb{R}^{W \times H \times D'}$ uwzględniając lokalne cechy obrazu:

$$X'_{i,j,k} = \sum_{(x,y) \in \mathcal{N}((i,j),r)} F_{x-i+r+1, y-j+r+1, k} \cdot X_{x,y}. \tag{2}$$

Grafowe splotowe sieci neuronowe działają analogicznie do sieci splotowych na obrazach, ale pojęcie sąsiedztwa definiuje się przy pomocy sąsiedztwa wierzchołków w grafie. Dla wierzchołka v_i, jego sąsiedztwo będzie zdefiniowane następująco:

$$\mathcal{N}(i) = \{j : \{i, j\} \in E\}. \tag{3}$$

Wtedy też operację konwolucji grafowej można zdefiniować w ten sposób:

$$X'_{i,k} = \sum_{j \in \mathcal{N}(i)} F_k \cdot X_j, \tag{4}$$

gdzie $F \in \mathbb{R}^{D' \times D}$. Jest to przykład najprostszej konwolucji grafowej. Należy zauważyć, że ponieważ nie mamy możliwości rozróżnienia sąsiadów wierzchołka, filtry (wagi trenowalne) są takie same dla wszystkich sąsiadów. Dzięki temu sieci grafowe są niezmiennicze na permutację wierzchołków.

Model generatywny, w ujęciu klasycznym, uczy się (łącznego) rozkładu prawdopodobieństwa danych, co odróżnia go od modeli dyskryminacyjnych. Jego celem jest generowanie nowych przykładów danych pasujących do rozkładu danych wejściowych. W przypadku związków chemicznych najczęściej celem będzie generowanie grafów molekularnych bezpośrednio lub używając innych reprezentacji związków chemicznych dających się jednoznacznie przekształcić do poprawnych wzorów strukturalnych. Wygenerowane związki powinny być nowe i często spełniać dodatkowe założenia związane z celem projektowania tych substancji. W przypadku projektowania leków może to być, na przykład, utrzymanie odpowiednich własności fizykochemicznych czy też optymalizacja aktywności biologicznej. Matematycznie możemy zapisać, że model generatywny p przy n założeniach $\{c_i : 1 \leqslant i \leqslant n\}$ będzie generował związki $\mathcal{G} \sim p(\mathcal{M}|c_1, \ldots, c_n)$, gdzie $\mathcal{M}$ jest zbiorem poprawnych grafów molekularnych, $n \in \mathbb{N}_0$.

Cykl publikacji

[P1] **Tomasz Danel**, Przemysław Spurek, Jacek Tabor, Marek Śmieja, Łukasz Struski, Agnieszka Słowik, and Łukasz Maziarka. "Spatial graph convolutional networks". In: *International Conference on Neural Information Processing*. Springer. 2020, pp. 668–675. CORE A, MNiSW 140 pkt. *Mój wkład: optymalizacja algorytmu, opracowanie grafik i schematów, udział w napisaniu omówienia i dyskusji wyników, udział w pracach edytorskich i odpowiedzi na recenzje.*

[P2] Agnieszka Pocha, **Tomasz Danel**, Sabina Podlewska, Jacek Tabor, and Łukasz Maziarka. "Comparison of atom representations in graph neural networks for molecular property prediction". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–8. CORE A, MNiSW 140 pkt. *Mój wkład: zaprojektowanie eksperymentów, przeprowadzenie aanalizy statystycznej, opis wyników jakościowych, zaproponowanie podstawowych reprezentacji, konsultacja projektu, weryfikacja implementacji, redakcja i korekta tekstu.*

[P3] **Tomasz Danel**, Jan Łęski, Sabina Podlewska, and Igor Podolak. "Docking-based generative approaches in the service of finding new drug candidates". In: *Drug discovery today* (2023). IF 8.369, MNiSW 200 pkt. *Mój wkład: koordynowanie prac zespołu, przegląd literatury, przygotowanie schematów graficznych, opis większości metod zawartych w pracy przeglądowej, krytyczna redakcja manuskryptu, odpowiedź na recenzje.*

[P4] **Tomasz Danel**, Agnieszka Wojtuch, and Sabina Podlewska. "Generation of new inhibitors of selected cytochrome P450 subtypes– in silico study". In: *Computational and Structural Biotechnology Journal* (2022). ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2022.10.005. URL: https://www.sciencedirect.com/science/article/pii/S2001037022004536. IF 6.155, MNiSW 100 pkt. *Mój wkład: przygotowanie i analiza danych, przeprowadzenie większości eksperymentów, zaproponowanie i zaimplementowanie architektury sieci grafowej, stworzenie aplikacji do wizualizacji wyników, udział w redagowaniu manuskryptu i odpowiedzi na recenzje.*

[P5] Dawid Rymarczyk, Daniel Dobrowolski, and **Tomasz Danel**. "ProGReST: Prototypical Graph Regression Soft Trees for Molecular Property Prediction". In: *SIAM International Conference on Data Mining (SDM)*. 2023. CORE A, MNiSW 140 pkt, *Mój wkład: zaproponowanie połączenia sieci grafowych dla związków chemicznych z koncepcją prototypów, nadzorowanie prac związanych z implementacją sieci grafowych, zaproponowanie i przygotowanie zbiorów danych, udział w krytycznym redagowaniu manuskryptu i korekta tekstu.*

[P6] Łukasz Maziarka, Agnieszka Pocha, Jan Kaczmarczyk, Krzysztof Rataj, **Tomasz Danel**, and Michał Warchoł. "Mol-CycleGAN: a generative model for molecular optimization". In: *Journal of Cheminformatics* 12.1 (2020), pp. 1–18. IF 8.489, MNiSW 100 pkt. *Mój wkład: zaproponowanie i nadzór nad eksperymentami związanymi z modyfikacjami strukturalnymi związków, udział w napisaniu omówienia i dyskusji wyników, udział w procesie odpowiedzi na recenzje.*

[P7] **Tomasz Danel**, Marek Śmieja, Łukasz Struski, Przemysław Spurek, and Łukasz Maziarka. "Processing of incomplete images by (graph) convolutional neural networks". In: *International Conference on Neural Information Processing*. Springer. 2020, pp. 512–523. CORE A, MNiSW 140 pkt. *Mój wkład: zaimplementowanie modelu, przeprowadzenie większości eksperymentów, przeanalizowanie wyników i redakcja tekstu publikacji.*

[P8] Agnieszka Galanty, **Tomasz Danel**, Michał Węgrzyn, Irma Podolak, and Igor Podolak. "Deep convolutional neural network for preliminary in-field classification of lichen species". In: *biosystems engineering* 204 (2021), pp. 15–25. IF 5.002, MNiSW 100 pkt. *Mój wkład: opracowanie, wykonanie i przetestowanie modelu sieci neuronowej, opracowanie graficznych wyników, udział w napisaniu omówienia i dyskusji wyników, udział w procesie odpowiedzi na recenzje.*

[P9] Łukasz Struski, **Tomasz Danel**, Marek Śmieja, Jacek Tabor, and Bartosz Zieliński. "SONGs: Self-Organizing Neural Graphs". In: *2023 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2023. CORE A, MNiSW 140 pkt. *Mój wkład: przeprowadzenie eksperymentów na małych zbiorach danych, opracowanie schematów i wyników graficznych, udział w krytycznym redagowaniu pracy, udział w procesie odpowiedzi na recenzje.*

P1. Przestrzenne grafowe sieci splotowe

ICONIP 2020 | CORE*: A | Punkty MNiSW: 140

Wstęp. W klasycznych grafowych sieciach sąsiednie wierzchołki nie są rozróżnialne i wiadomości z nich płynące są niezależne od położenia. Rozróżnienie tych wierzchołków w trakcie treningu dałoby większą elastyczność modelowi. Proponujemy rozwiązanie tego problemu w pierwszej pracy cyklu, która wprowadza nową architekturę grafowej sieci neuronowej, przestrzenną grafową sieć splotową (ang. *spatial graph convolutional network*, SGCN). Model ten ma być bezpośrednim uogólnieniem klasycznej sieci konwolucyjnej oraz sieci grafowej.

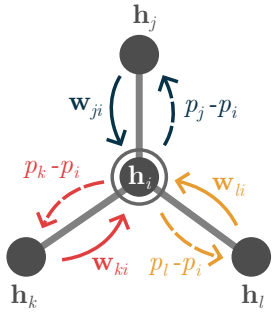
Konwolucyjne sieci grafowe posługują się pojęciem sąsiedztwa zdefiniowanym na grafie przez krawędzie między wierzchołkami. Ściślej mówiąc, wierzchołek j jest sąsiadem wierzchołka i, jeżeli istnieje w grafie krawędź $\{i, j\}$. Ponieważ nie istnieje ustalona kolejność wierzchołków w grafie (w szczególności grafy o różnej numeracji wierzchołków mogą być izomorficzne), większość grafowych sieci neuronowych przyjmuje niezmienniczość na permutację sąsiadów jako podstawowy warunek w trakcie przetwarzania grafu. Z drugiej strony, w konwolucyjnych sieciach neuronowych na obrazach jedną z podstawowych własności jest rozróżnienie między sąsiednimi pikselami, aby możliwe było nauczenie filtrów wykrywających istotne cechy w obrazie.

Metody. Aby połączyć cechy obu architektur sieci neuronowej, wprowadzamy do grafu pojęcie przestrzeni d-wymiarowej i ustalamy lokalizację poszczególnych wierzchołków w tej przestrzeni. Dla przykładu, dane obrazowe możemy przedstawić jako graf, gdzie wierzchołki odpowiadają pikselom na obrazie oraz wszystkie sąsiednie piksele są ze sobą połączone krawędzią. Na tak zdefiniowanym grafie możemy wykonać naszą operację grafowej przestrzennej konwolucji (rys. 1):

$$\mathbf{h}'_i(\mathbf{U}, \mathbf{b}) = \sum_{j \in \mathcal{N}(i)} \underbrace{\mathrm{ReLU}(\mathbf{U}^T(p_j - p_i) + \mathbf{b})}_{\mathbf{w}_{ji}} \odot \mathbf{h}_j, \tag{5}$$

gdzie $\mathbf{h}_j$ to reprezentacja j-tego wierzchołka, $p_i \in \mathbb{R}^d$ to położenie i-tego wierzchołka, a $\mathbf{U} \in \mathbb{R}^{d \times h}$ oraz $\mathbf{b} \in \mathbb{R}^h$ to parametry trenowalne. Dodanie wzajemnego położenia sprawia, że wagi sieci mogą być różne dla każdego sąsiada (w przeciwieństwie do klasycznej konwolucji grafowej przedstawionej w równaniu 4). Co więcej, dla podanego powyżej przykładu przekształcenia obrazu do postaci grafu możemy udowodnić, że operacja ta jest uogólnieniem zarówno operacji konwolucji na obrazach, jak również konwolucji grafowej. W artykule udowadniamy ten fakt formalnie.

SGCN jest architekturą, która nadaje się do analizy związków chemicznych, ponieważ atomy można łatwo umiejscowić w przestrzeni 3-wymiarowej. Związki chemiczne układają się w przestrzeni tak, by zminimalizować energię systemu. Takie chwilowe ułożenie związku nazywamy konformacją. Możliwe jest szacowanie stabilnych konformacji przy pomocy narzędzi komputerowych (metody pól siłowych) lub eksperymentalnie, np. przez krystalizację struktury związku. Jednym z problemów naszej metody w przetwarzaniu związków chemicznych jest fakt, że własności związków nie zależą od ich obrotu w przestrzeni. Aby uzyskać efekt niezmienniczości na obroty, używamy augmentacji polegającej na losowym obracaniu związku w trakcie treningu modelu.



Rysunek 1: Schemat przestrzennej konwolucji grafowej. Wagi $\mathbf{w}_{ji}$ są różne ze względu na użycie wzajemnych położeń wierzchołków.

Wyniki. Eksperymenty pokazują, że SGCN wyposażony w augmentacje daje lepsze wyniki niż inne metody grafowe na wybranych chemicznych zbiorach danych. Sugeruje to, że efektywne użycie pozycji atomów niesie ze sobą dodatkową informację. Co więcej, dodanie pozycji do cech atomowych w klasycznej sieci grafowej nie daje wyników nawet bliskich tym uzyskiwanym przez SGCN.

*Podana ranga w momencie publikacji artykułu. Obecna ranga konferencji to B.

P2. Porównanie reprezentacji atomów w grafowych sieciach neuronowych przewidujących własności związków chemicznych

IJCNN 2021 | CORE*: A | Punkty MNiSW: 140

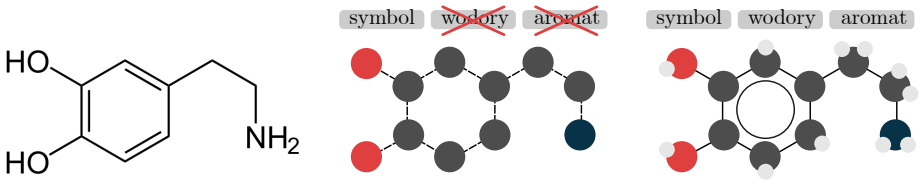
Wstęp. W kolejnej pracy cyklu skupiamy się na analizie wpływu wyboru cech atomów na wyniki przewidywań grafowych sieci neuronowych. Obecnie większość prac wprowadzających nowe architektury sieci grafowych skupia się na udoskonaleniu warstw grafowych i pomija zupełnie wpływ użytych cech atomowych. Nieraz zdarza się, nowe architektury porównywane są z poprzednimi z użyciem innego zestawu cech. W takich porównaniach nie jest jasne czy lepsze wyniki wynikają z nowych metod przetwarzania grafów czy użycia innych cech związków. W naszej pracy przeprowadzamy systematyczne porównanie różnych zestawów cech atomów na kilku chemicznych zbiorach danych. Naszym celem jest zbadanie, czy istnieje uniwersalna reprezentacja atomów dla różnych zadań chemicznych, czy raczej cechy atomów powinny stanowić jeden z istotnych hiperparametrów przy optymalizacji sieci grafowych.

Metody. Do typowych cech atomów zalicza się typ atomu (zakodowany symbol pierwiastka), liczba sąsiadów ciężkich, liczba związanych z atomem wodorów, ładunek formalny atomu, czy zawieranie się atomu w pierścieniu lub pierścieniu aromatycznym. Wśród naszych zdefiniowanych reprezentacji jest reprezentacja pusta, zawierająca tylko informację o typie atomu, oraz reprezentacje zawierające dodatkowo jedną z wymienionych wyżej cech lub reprezentacje zawierające wszystkie cechy poza jedną (przykładowe reprezentacje na rys. 2).

Reprezentacje testujemy z użyciem prostej konwolucyjnej sieci grafowej. Dla części zbiorów danych testujemy dwa podejścia do podziału danych na treningowe i testowe. Jeden podział jest losowy, drugi stara się grupować podobne strukturalnie związki w tym samym podzbiorze danych (tzw. scaffold split). Dla każdej reprezentacji szukamy optymalnej architektury spośród wylosowanych stu zestawów hiperparametrów.

Wnioski. W przeprowadzonych eksperymentach odkrywamy, że wybór odpowiedniego zestawu cech atomu może mieć równie duży wpływ na przewidywanie modelu, co dobór odpowiedniej architektury. Dodatkowo okazuje się, że wybór optymalnych cech zależny jest od wybranego problemu i sposobu podziału danych. Nie zawsze użycie wszystkich możliwych cech prowadzi do najlepszych wyników modelu. Możliwe, że ma to związek z małą ilością danych w wielu zbiorach chemicznych, przez co sieć uczy się błędnych korelacji (overfitting). Dla potwierdzenia naszych obserwacji przeprowadzamy testy statystyczne porównujące predykcje najlepszych modeli dla każdej z reprezentacji. Użyty test to test Wilcoxona z korektą Bonferroniego.

Szczególną uwagę należy zwrócić na wskazanie, że choć użycie pełnego zestawu cech zwykle daje zadowalające wyniki, to dla niektórych zbiorów danych usunięcie pewnych cech (np. ładunku formalnego i aromatyczności) może poprawić wyniki przewidywań. Największą poprawę wyniku daje natomiast dodanie informacji o sąsiadach ciężkich i wodorach. Stąd wnioskujemy, że odpowiedni dobór cech atomów jest istotny dla działania sieci grafowej.



(a) wzór strukturalny (b) minimalna reprezentacja (c) rozszerzona reprezentacja

Rysunek 2: Wzór strukturalny dopaminy (a), jej graf molekularny zawierający jedynie symbol pierwiastka (b) oraz graf zawierający także informację o liczbie wodorów i aromatyczności (c). Reprezentacja (b) jest niejednoznaczna, natomiast (c) zawiera cechy wzajemnie zależne od siebie.

*Podana ranga w momencie publikacji artykułu. Obecna ranga konferencji to B.

P3. Podejścia bazujące na dokowaniu w służbie odkrywania nowych kandydatów na lek

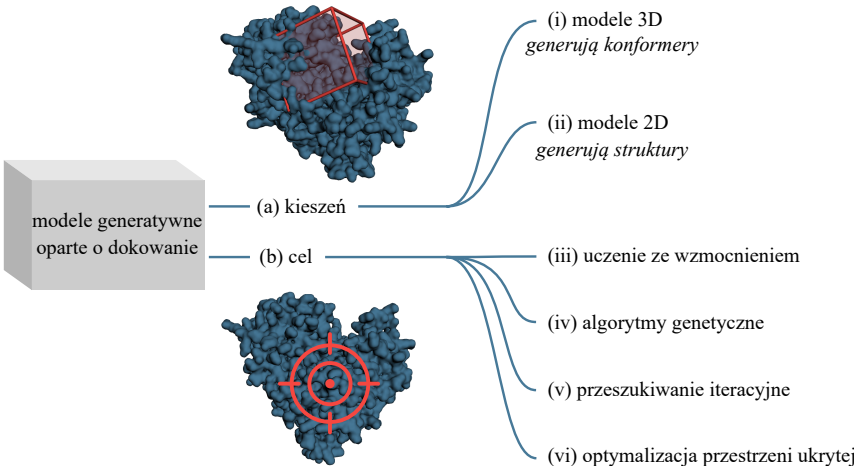
Drug Discovery Today | IF: 8.369 | Punkty MNiSW: 200

Wstęp. Dokowanie jest podstawowym narzędziem komputerowym do wstępnej oceny kandydatów na lek. Dzięki dokowaniu otrzymujemy ranking związków pod względem przewidywanego powinowactwa do swojego celu (zwykle białka) oraz ich potencjalne pozycje w kieszeni wiążącej białka. Dokowanie często wykorzystywane jest w kampaniach związanych z odkrywaniem nowych leków podczas tzw. screeningu wirtualnego, gdzie ogromne bazy związków chemicznych są dokowane, a następnie wybiera się z nich najbardziej obiecujące związki do syntezy. Wirtualny screening może być również oparty o modele uczenia maszynowego. Limitem tego podejścia jest brak możliwości znajdowania nowych związków, do czego z kolei mogą służyć modele generatywne.

Metody. W naszej pracy przeglądowej opisujemy najnowsze metody generatywne, które integrują dokowanie molekularne w celu generowania związków o polepszonym powinowactwie do swojego celu. Ponieważ dokowanie jest komputerową aproksymacją interakcji pomiędzy ligandem i białkiem, jest ono bogatym źródłem dodatkowych informacji (wartość funkcji dokowania, przewidziana pozycja liganda, interakcje chemiczne), które można użyć do prowadzenia procesu generatywnego. W tym ujęciu modele generatywne są uczone w sposób częściowo nadzorowany, ponieważ jesteśmy w stanie wygenerować etykiety przy pomocy narzędzi do dokowania, z których część jest dostępna nieodpłatnie (np. AutoDock, AutoDock Vina, smina, gnina), a inne udostępniane są na licencjach komercyjnych lub akademickich (np. Glide, GOLD, FlexX).

Taksonomia. W artykule zaproponowaliśmy nowatorską taksonomię modeli generatywnych używających dokowania (rys. 3). Metody podzieliliśmy na te używające kodowania kieszeni wiążącej oraz te, które wykorzystują jedynie dane dokowania specyficzne dla danego celu biologicznego. Te pierwsze z metod zwykle jako dodatkowe wejście otrzymują informację o kształcie kieszeni kodowaną przy pomocy reprezentacji 3D takich jak woksele lub grafy 3D. Często związek w tym wypadku jest generowany bezpośrednio w kieszeni, aby zapewnić dopasowanie.

W przypadku pozostałych metod najczęściej dokowanie jest używane jako wyrocznia. Mamy tutaj do czynienia z algorytmami genetycznymi albo uczeniem ze wzmocnieniem, w przypadku których związki są generowane tak, aby optymalizować wartość funkcji oceny pochodzącej z dokowania (docking score). Podobnie możemy wykorzystać modele generatywne z przestrzenią ukrytą, w przypadku których optymalizacja może być dokonywana właśnie na tej niskowymiarowej przestrzeni, np. przez optymalizację gradientową lub Bayesowską. Jeszcze inne metody używają dokowania do filtrowania wygenerowanych związków. Mamy nadzieję, że zaproponowany przez nas podział algorytmów posłuży naukowcom w ich pracach nad przyszłymi modelami.



Rysunek 3: Nasza taksonomia modeli generatywnych opartych o dokowanie.

P4. Generowanie nowych inhibitorów wybranych podtypów cytochromu P450 - studium *in silico*

Computational and Structural Biotechnology Journal | IF: 6.155 | Punkty MNiSW: 100

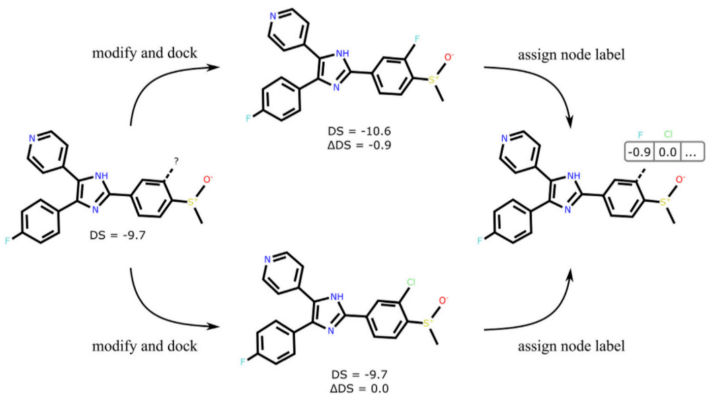
Wstęp. W tej pracy przeprowadziliśmy studium *in silico* dla wybranych podtypów cytochromu P450, odpowiedzialnego za metabolizm związków chemicznych w organizmie. Projekt miał na celu znalezienie bliskich analogów istniejących inhibitorów tych enzymów i porównanie ich powinowactwa w dokowaniu molekularnym. Z publicznej bazy związków chemicznych ChEMBL wybraliśmy te związki, które posiadają zarejestrowane eksperymentalne wartości aktywności dla wybranych enzymów. Wybraliśmy po dwa modele tych enzymów z bazy PDB, jeden zawierający ligand, drugi niezwiązany. W celu przeprowadzenia dokowania molekularnego usunęliśmy z modeli liganda i cząsteczki wody, zostawiając jedynie hem, będący kofaktorem cytochromu. Po przeprowadzeniu dokowania mogliśmy porównać wartości funkcji dokowania dla oryginalnego i zmodyfikowanego przez nas związku w obu kryształach dla każdego podtypu cytochromu P450.

Metody. Do generowania analogów wybraliśmy 15 małych ugrupowań chemicznych, które były kombinatorycznie dołączane do wszystkich możliwych atomów w pierścieniach związków oryginalnych. Zauważyliśmy, że niektóre podstawienia szczególnie przyczyniają się do poprawy lub pogorszenia wartości funkcji dokowania.

Dodatkowo zaimplementowaliśmy modeli sieci grafowej służącej do przewidywania zmiany wartości funkcji dokowania (rys. 4). Sieć otrzymuje na wejściu niezmodyfikowany związek i przewiduje dla każdego atomu w pierścieniu różnicę wartości funkcji dokowania, gdy dokonane zostanie wybrane podstawienie. Problem jest zatem sformułowany jako regresja wielowartościowa (przewidywanie wielu wartości na raz) na wierzchołkach grafu. Dla atomów poza pierścieniami (lub analogów, których nie udało się zadokować) przypisujemy puste wartości, które są ignorowane w trakcie treningu. Dla wytrenowanych modeli dokonujemy wyjaśnień predykcji przy pomocy metody map istotności (ang. *saliency maps*), aby lepiej zrozumieć działanie sieci i znaleźć ugrupowania ważne dla poprawy dokowania związku.

Implementację generatora związków, sieci grafowej i modeli wyjaśniających udostępniliśmy w publicznym repozytorium wraz z powstałą bazą danych dużej liczby zadokowanych przez nas związków chemicznych.

Wyniki. Zaproponowaliśmy nowy protokół generowania inhibitorów wybranych CYP-ów. Eksperymentalnie pokazaliśmy, że grafowa sieć neuronowa może skutecznie zaproponować podstawienia związku, które poprawiają wyniki dokowania. Umożliwić to może zawężenie przeszukiwania przestrzeni chemicznej jako alternatywa do tworzenia wszystkich możliwych kombinatorycznie pochodnych. Ponadto wizualizację wyników dokowania części naszej bazy danych można znaleźć w stworzonym przez nas narzędziu online.



Rysunek 4: Proces generowania przykładowych analogów związku i obliczania różnicy w wartości docking score. Różnice te dla poszczególnych modyfikacji służą później sieci grafowej jako etykiety.

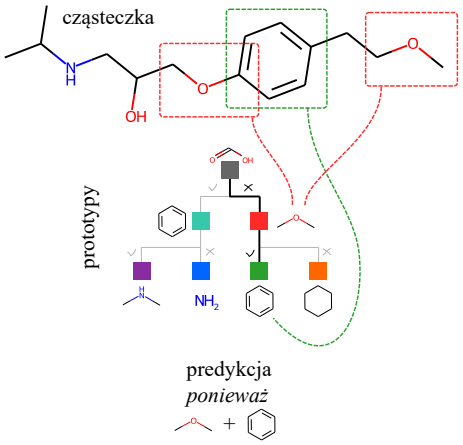
P5. ProGReST: prototypowe grafowe miękkie drzewa regresyjne do przewidywania własności cząsteczek

SDM 2023 | CORE: A | Punkty MNiSW: 140

Wstęp. Przewidywanie własności cząsteczek jest kluczowe w procesie projektowania leków. Dzięki modelom predykcyjnym naukowcy są w stanie odfiltrować związki chemiczne, które nie spełniają założonych kryteriów takich jak odpowiednia rozpuszczalność lub powinowactwo do celu biologicznego. Decyzje modeli są jednak zwykle obarczone błędami, których eliminacja jest możliwa jedynie poprzez przeprowadzenie eksperymentów w laboratorium. Ważne zatem jest zrozumienie czynników, jakie wpłynęły na decyzję modelu.

Metody. W przedstawionej pracy proponujemy nowy interpretowalny model grafowy. Jest on oparty na koncepcji prototypów, która pozwala wyjaśniać przewidywania modeli predykcyjnych przez przykłady. Prototypem jest zbiór cech lub fragmentów podobnych do posiadanych przez reprezentantów w zbiorze treningowym. W przypadku zdjęć ptaków takim fragmentem (zwanym też częścią prototypową) może być jaskrawy brzuszek lub szpiczasty dziób. W przypadku małych związków chemicznych będą to na przykład konkretne grupy funkcyjne. Nasza praca jest jedną z pierwszych wykorzystujących prototypy do wyjaśniania predykcji własności cząsteczek chemicznych i pierwszą używającą prototypy dla problemów regresyjnych.

Nasz model interpretowalny, ProGReST (rys. 5), bazuje na grafowej sieci neuronowej, która tworzy reprezentację ukrytą cząsteczki. Reprezentacja ta składa się z wektorów cech dla każdego z atomów cząsteczki. ProGReST jest decyzyjnym drzewem binarnym, który w każdym wierzchołku porównuje reprezentację atomów cząsteczki wejściowej z trenowalną częścią prototypową zawartą w danym wierzchołku. Część prototypowa odpowiada tutaj zatem charakterystycznym fragmentom związku, których opis został zagregowany w reprezentacji wierzchołka grafu odpowiadającego atomowi związku wejściowego. Jeśli część prototypowa zostanie znaleziona, to drzewo decyzyjne kieruje wybór do swojego prawego dziecka, a do lewego w przeciwnym wypadku.



Rysunek 5: Schemat pokazujący ciąg decyzji bazujących na obecności części prototypowych w cząsteczce wejściowej.

Ponieważ ProGReST jest miękkim drzewem decyzyjnym, możliwe jest kierowanie decyzji do obu dzieci jednocześnie. W tym wypadku waga decyzji prawego dziecka jest proporcjonalna do *obecności* prototypu w cząsteczce (zdefiniowanej w artykule w równaniu 3.1), a waga lewego dziecka jest dopełnieniem wagi prawego dziecka. Po dotarciu związku do liści drzewa decyzyjnego obliczana jest wartość własności związku jako kombinacja liniowa prawdopodobieństw w liściach (współczynniki są trenowalne).

Wyniki. Eksperymenty przeprowadzone na pięciu zbiorach własności chemicznych wskazują na bardzo wysoką skuteczność modelu. We wszystkich przypadkach ProGReST pokonuje model odniesienia, którym jest prosta sieć grafowa. Gdy jako sieci grafowej użyjemy RMAT-a [1] (model transformerowy dla związków chemicznych), wygrywamy z innymi modelami w aż czterech na pięć zadań, dodając jednocześnie interpretowalność przewidywań modelu. W eksperymentach jakościowych pokazujemy przykłady prototypów istotnych dla rozpatrywanych problemów chemicznych.

P6. Mol-CycleGAN: model generatywny służący do optymalizacji związków chemicznych

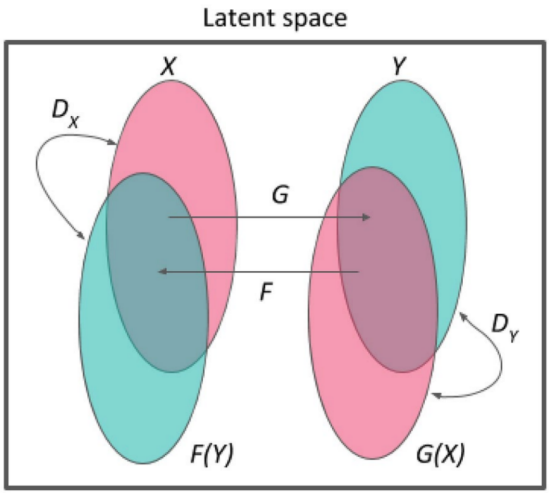
Journal of Cheminformatics | IF: 8.489 | Punkty MNiSW: 100

Wstęp. Optymalizacja molekuł jest ważnym etapem każdego projektu związanego z odkrywaniem nowych leków. Zadanie to różni się od projektowania leków *de novo* (od zera) przede wszystkim tym, że znana jest już struktura wyjściowa, która posiada pewne pożądane cechy, ale jej pozostałe właściwości nie są dopuszczalne dla dalszego postępu projektu. Przykładowo możemy znać strukturę aktywną wobec określonego celu biologicznego, ale związek ten jest zbyt szybko metabolizowany w organizmie. Modele generatywne, gdy są używane do optymalizacji związków chemicznych, starają się poprawić właściwości molekularne jedynie nieznacznie zmieniając strukturę wyjściową.

Metody. Nasz model, Mol-CycleGAN, jest pierwszym modelem używającym adaptacji domen do problemu optymalizacji związków chemicznych. Bazuje on na architekturze JT-VAE [2]. Jest to autoenkoder wariacyjny, który koduje molekuły w postaci drzew podstruktur chemicznych (JT, ang. *junction trees*). W naszym przypadku autoenkoder będzie służył do modyfikacji związku, więc nie może być trenowany przy użyciu standardowej funkcji kosztu rekonstrukcji obiektu wejściowego.

Problem optymalizacji cząsteczek formułujemy jako zadanie optymalizacji dziedziny danych (ang. *domain adaptation*), gdzie dziedzinami są zbiory związków o niskich (X) i wysokich (Y) wartościach interesujących nas własności chemicznych (rys. 6). Dwa autoenkodery służą nam do generowania związków z domeny Y na podstawie związków z domeny X oraz odwrotnie. Wykorzystujemy funkcję kosztu CycleGAN-a [3], która zapewnia wysokie podobieństwo wygenerowanych związków do rozpatrywanej dziedziny oraz duże podobieństwo między związkiem wejściowym i zmodyfikowanym.

Wyniki. Skuteczność naszego modelu testujemy na własności *penalized log P* będącej połączeniem przewidywanej komputerowo lipofilowości związku i łatwości jego syntezy. Eksperymenty wykazują, że jesteśmy w stanie optymalizować tę własność zachowując małe zmiany w strukturze związku. Ponadto w kolejnym eksperymencie dokonujemy optymalizacji strukturalnych, np. podmiany bioizosterycznej (wymiany podstruktur na inne o podobnych własnościach elektronowych), a także optymalizacji aktywności biologicznej. W obu przypadkach model jest w stanie nauczyć się cech strukturalnych obu domen i poprawnie optymalizuje związki.



Rysunek 6: Schemat działania Mol-CycleGAN-a. X i Y to domeny związków odpowiadające obniżonym i podwyższonym wartościom określonej własności chemicznej, F i G to generatory optymalizujące związki (modele przenoszące związki wejściowe do innej domeny), a D_X i D_Y to dyskryminatory używane w trakcie uczenia, by zapewnić podobieństwo wygenerowanych związków do domeny docelowej.

P7. Przetwarzanie niekompletnych obrazów przy pomocy (grafowych) sieci splotowych

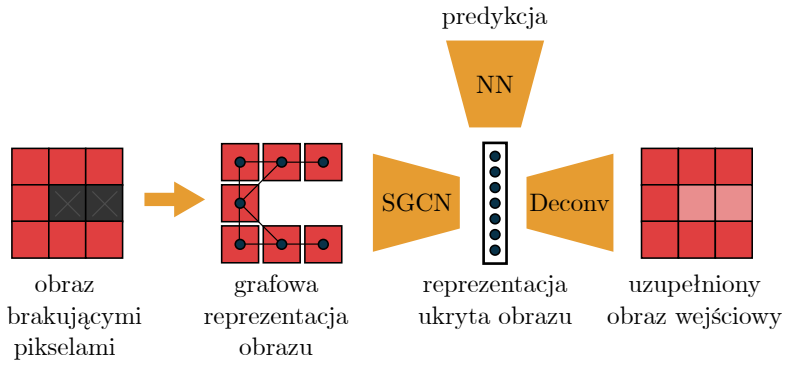
ICONIP 2020 | CORE*: A | Punkty MNiSW: 140

Wstęp. Dane obrazowe są powszechnie wykorzystywane w medycynie. Często są to obrazy specjalistyczne pochodzące z urządzeń takich jak różnego rodzaju mikroskopy i skanery. Częstymi problemami przy przetwarzaniu takich obrazów są duża rozdzielczość oraz brakujące lub niewyraźne fragmenty zdjęcia. W tej pracy zajmujemy się tym drugim problemem, skupiając się szczególnie na możliwości klasyfikacji obrazów z brakującymi regionami o dowolnym kształcie oraz uzupełnienia takich obrazów. Proponujemy wykorzystanie do tego zadania SGCN-a, czyli grafowej sieci neuronowej opisanej powyżej w sekcji P1. Do rekonstrukcji brakujących regionów używamy architektury autoenkodera, w której enkoderem jest SGCN, a dekoderem sieć dekonwolucyjna.

Metody. Główną zaletą naszego rozwiązania jest brak potrzeby uzupełniania brakujących informacji – brakujące piksele mogą być pominięte przy przetwarzaniu siecią grafową. Obraz reprezentowany jest jako graf, w którym wierzchołkami są piksele, a nieskierowane krawędzie łączą każdą parę sąsiednich pikseli (również sąsiadujące pod ukosem, patrz rys. 7). Ponieważ SGCN jest uogólnieniem sieci konwolucyjnej, obrazy mogą być przetwarzane zachowując wysoką dokładność przy jednoczesnej możliwości pominięcia części pikseli obrazu wejściowego.

W przypadku zadania polegającego na uzupełnieniu brakujących fragmentów obrazu postanowiliśmy użyć architektury autoenkodera. Aby pominąć brakujące fragmenty obrazu wejściowego, jako enkodera używamy SGCN-a jak opisano powyżej. Dekoderem jest sieć dekonwolucyjna, która stopniowo rekonstruuje pełny obraz. Model trenowany jest przy pomocy funkcji kosztu rekonstrukcji, którą jest błąd średniokwadratowy policzony pomiędzy pikselami oryginalnego a odtworzonego obrazu. Zakładamy trudniejszy scenariusz uczenia, w którym brakujące regiony nigdy nie są dostępne w treningu, a funkcja kosztu jest liczona tylko na dostępnych pikselach.

Wyniki. W eksperymentach używamy dwóch zbiorów danych, MNIST i SVHN. W przypadku zadania klasyfikacji cyfr widocznych na zdjęciu otrzymujemy lepsze wyniki niż analogiczne sieci konwolucyjne zaaplikowane na sztucznie uzupełnionych obrazach (uzupełnienie średnią wartością piksela, uzupełnienie maską składającą się z samych czarnych pikseli lub uzupełnienie metodą k-NN). W przypadku rekonstrukcji obrazów widzimy wyraźnie, że regiony uzupełnione przez nasz model najbardziej przypominają oryginalne obrazy, kształty są ostre i dobrze wpasowują się w otaczającą je ramkę dostępnych pikseli.



Rysunek 7: Schemat użycia SGCN-a do klasyfikacji lub uzupełniania niekompletnych obrazów. Obraz wejściowy jest zamieniany na graf z pominięciem brakujących pikseli. SGCN koduje obraz w niskowymiarowej przestrzeni ukrytej, która może być użyta do klasyfikacji lub rekonstrukcji obrazu wejściowego.

*Podana ranga w momencie publikacji artykułu. Obecna ranga konferencji to B.

P8. Głębokie sieci splotowe służące wstępnej terenowej klasyfikacji gatunków porostów

Biosystems Engineering | IF: 5.002 | Punkty MNiSW: 100

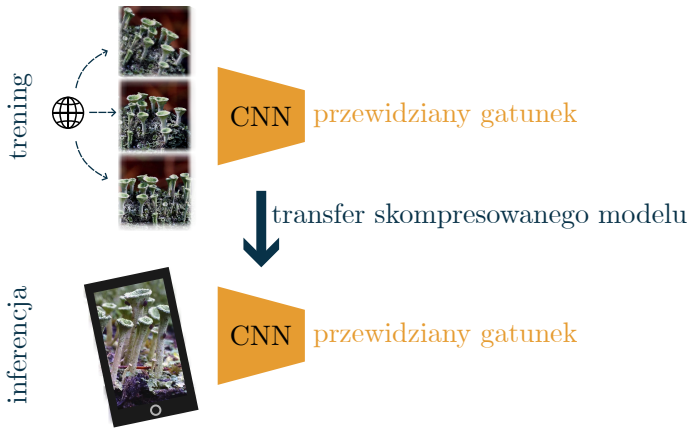
Wstęp. Organizmy takie jak rośliny, grzyby i bakterie wielokrotnie były używane jako źródła substancji aktywnych i zawarte w nich związki chemiczne były inspiracją do stworzenia leków będących obecnie w obrocie. Został nawet wyodrębniony dział farmacji, farmakognozja, który zajmuje się substancjami pochodzenia naturalnego. Poprawna klasyfikacja organizmów takich jak rośliny i grzyby jest zatem ważnym problemem farmacji. Zgodnie z naszą najlepszą wiedzą zaproponowane przez nas rozwiązanie jest pierwszym używającym zdjęć porostów do klasyfikacji ich gatunków.

Metody. W niniejszej pracy skupiamy się na porostach z rodziny chrobotkowatych (*Cladoniaceae*). Do klasyfikacji używamy sieci splotowych, które przyjmują na wejściu zdjęcia porostów i starają się zaklasyfikować obiekt na zdjęciu do jednego z 12 gatunków. Co istotne, wykorzystujemy architektury mobilne, które dostosowane są do wdrożenia na urządzenia o ograniczonej mocy obliczeniowej i dostępnej pamięci. Dzięki temu, nasze rozwiązanie może zostać zainstalowane na telefonach badaczy terenowych lub dołączone do kamer drona, który będzie automatycznie przeszukiwał teren i oznaczał gatunki na nim występujące.

W momencie przeprowadzania eksperymentów nie istniały dostępne publicznie zbiory danych zdjęć porostów, które moglibyśmy wykorzystać w naszych badaniach. Zamiast tego używamy własnego zbioru, który został automatycznie pobrany z opublikowanych w internecie zdjęć porostów. Zbiór składa się z 1164 zdjęć podzielonych losowo na 931 zdjęć treningowych i 233 testowych. Zdjęcia po pobraniu zostały ręcznie przefiltrowane w celu odrzucenia niepoprawnie oznaczonych zdjęć.

W eksperymentach testujemy dwie architektury mobilne, MobileNet v2 [4] oraz SqueezeNet [5], a także klasyczne podejście do klasyfikacji obrazów, wektory Fishera. Ponieważ zbiór danych jest bardzo mały i zawiera mniej niż 100 zdjęć na klasę, w treningu używamy augmentacji zdjęć przez losowe obroty, przybliżenia i odbicia lustrzane. Rysunek 8 przedstawia schemat uczenia i inferencji.

Wyniki. Nasz wariant modelu SqueezeNet okazuje się najlepszą architekturą w zestawieniu dla zadania klasyfikacji porostów, osiągając 61% dokładności w klasyfikacji 12-klasowej. Tak wysoki wynik klasyfikacji na małym zbiorze danych był możliwy jedynie dzięki rozbudowanej procedurze augmentacji danych.



Rysunek 8: Konwolucyjna sieć neuronowa (CNN) jest trenowana na augmentowanych obrazach z internetu, a następnie jej wagi są przenoszone do skompresowanego modelu na urządzeniu mobilnym. Urządzenie to może przewidywać gatunki porostów w czasie rzeczywistym, używając wbudowanej kamery.

P9. SONG: samoorganizujące się grafy neuronowe

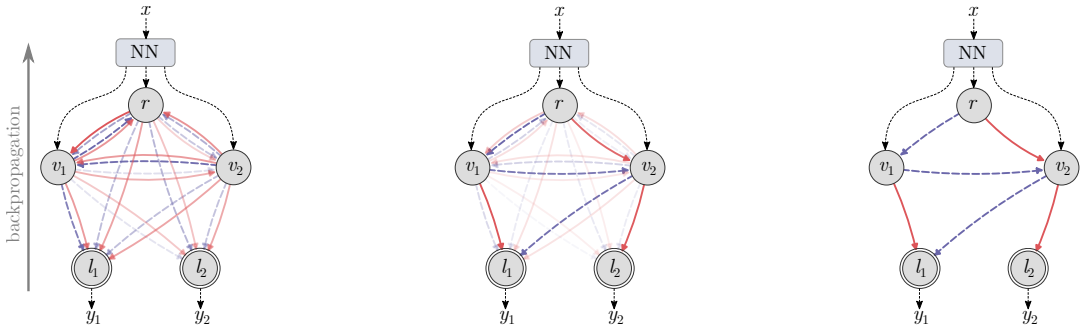
WACV 2023 | CORE: A | Punkty MNiSW: 140

Wstęp. Samoorganizujące się grafy neuronowe (SONG, ang. *self-organizing neural graphs*) to zaprojektowana przez nas nowa architektura sieci neuronowych. Motywacją tej pracy było wykorzystanie zalet drzew decyzyjnych, którymi są wysoka dokładność przewidywania i interpretowalność, zapewniając jednocześnie większą elastyczność dzięki strukturze grafu decyzyjnego. Nasz model jest w pełni różniczkowalny i może być trenowany w połączeniu z innymi architekturami sieci neuronowych.

Metody. SONG jest modelem klasyfikacyjnym, który przyjmuje wektorową reprezentację danych i dokonuje sekwencji decyzji binarnych, poruszając się po grafie decyzyjnym od korzenia do liści (rys. 9). Korzeń jest wybranym wierzchołkiem startowym, a liście to wierzchołki końcowe odpowiadające przewidywanym klasom. Połączenia w grafie zadane są przez dwie macierze przejść zdefiniowane jak w procesie decyzyjnym Markowa, czyli suma prawdopodobieństw połączeń wychodzących z każdego wierzchołka jest równa 100%. Decyzja o tym, którą macierz przejścia użyć w danym wierzchołku, jest uzależniona od danych wejściowych (rozmyta decyzja binarna). Po ustalonej liczbie kroków przykłady wejściowe trafiają do liści, tworząc rozkład prawdopodobieństwa na przewidywanych klasach. Parametrami trenowalnymi modelu są zarówno obie macierze przejść, jak i funkcje decyzyjne w wierzchołkach grafu, a sam trening może być przeprowadzony przy pomocy typowej klasyfikacyjnej funkcji kosztu, np. entropii krzyżowej.

W trakcie uczenia zdecydowaliśmy się użyć dodatkowych składników regularyzacyjnych. Wprowadzamy regularyzację wierzchołków, wymuszającą równy podział danych w każdym wierzchołku grafu, oraz regularyzację na liściach, pomagającą w doprowadzeniu całości danych do liści w ustalonej liczbie kroków. Przy uczeniu grafu decyzyjnego metodą gradientową zauważyliśmy szybkie nasycanie się ścieżek. Zjawisko to polega na utrwalaniu przejść w grafie tak szybko, jak tylko jakiś przykład treningowy trafi do poprawnego liścia i nastąpi propagacja gradientu wzdłuż tej ścieżki. Aby zapobiec wzmacnianiu nieoptymalnych połączeń, zachęcamy model do eksploracji innych ścieżek przez zastosowanie stochastycznej operacji Gumbel softmax [6] zamiast zwykłego softmaxu.

Wyniki. Stosując odpowiednie sieci uczące się reprezentacji złożonych danych, jesteśmy w stanie zastosować SONG do różnych problemów klasyfikacyjnych. Na przykład w przypadku przewidywania własności związków chemicznych możemy użyć reprezentacji z sieci grafowej, natomiast dla obrazów używamy warstwy reprezentacji pochodzącej z sieci konwolucyjnej (ResNet [7] lub kilkuwarstwowe ekstraktory cech). Nasze eksperymenty pokazują, że SONG potrafi osiągnąć dokładność równą innym metodom wykorzystującym drzewa decyzyjne przy mniejszej liczbie użytych wierzchołków decyzyjnych. Ponadto dowodzimy teoretycznie i empirycznie, że w pełni nauczony SONG zbiega do rzadkiego acyklicznego grafu binarnego.



(a) losowa inicjalizacja (b) faza pośrednia (c) nauczony SONG

Rysunek 9: Trening SONG-a. (a) Na początku wagi połączeń są losowo inicjalizowane. (b) W trakcie treningu poprawne ścieżki są wzmacniane, a niepoprawne osłabiane. (c) SONG staje się rzadkim acyklicznym grafem prowadzącym od korzenia r do liści l_i odpowiadającym klasom predykcji.

Życiorys naukowy

W roku 2018 ukończyłem z wyróżnieniem studia magisterskie na Uniwersytecie Jagiellońskim na kierunku informatyka o specjalizacji informatyka stosowana. Tego samego roku rozpocząłem studia doktoranckie w Katedrze Uczenia Maszynowego na Wydziale Matematyki i Informatyki Uniwersytetu Jagiellońskiego. Dołączyłem jednocześnie do grupy metod uczenia maszynowego GMUM.

Do moich zainteresowań badawczych zaliczyć można: uczenie głębokie, komputerowo wspomagane projektowanie leków, chemoinformatykę oraz wizję komputerową. W swoich badaniach dokonywałem prób syntezy tych dziedzin poprzez rozwój algorytmów uczenia maszynowego ukierunkowanych na rozwiązywanie problemów związanych z farmaceutyką.

Również zawodowo pracuję z podobnymi modelami komputerowymi. W latach 2019-2021 pracowałem jako Data Scientist w firmie Ardigen, gdzie zajmowałem się analizami danych biologicznych oraz automatycznym przetwarzaniem danych obrazowych. W 2021 roku zacząłem pracę na stanowisku Machine Learning Lead Engineer w Insitro, gdzie pracuję nad rozwojem metod uczenia maszynowego w projektowaniu leków.

Publikacje (poza ujętymi w cyklu)

1. Gaiński, P., Maziarka, Ł., **Danel, T.**, & Jastrzebski, S. (2022, June). HuggingMolecules: An Open-Source Library for Transformer-Based Molecular Property Prediction (Student Abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 11, pp. 12949-12950).

   Mój wkład polegał na weryfikacji implementacji i przygotowaniu oprawy graficznej.

2. Maziarka, Ł., Majchrowski, D., **Danel, T.**, Gaiński, P., Tabor, J., Podolak, I., Morkisz, P., & Jastrzębski, S. (2021). Relative molecule self-attention transformer. In *Machine Learning for Molecules Workshop at NeurIPS* (Vol. 2021).

   Mój wkład polegał na konsultacji projektu, przygotowaniu schematów graficznych, udziale w krytycznym redagowaniu artykułu.

3. Maziarka, Ł., & **Danel, T.** (2021, July). Multitask Learning Using BERT with Task-Embedded Attention. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-6). IEEE. 2021.

   Mój wkład polegał na konsultacji projektu, przygotowaniu tekstu i schematów graficznych.

4. **Danel, T.**, Szymczak, M., Maziarka, Ł., Podolak, I., Tabor, J., & Jastrzębski, S. (2020). De Novo Drug Design with a Docking Score Proxy. In *Machine Learning for Molecules Workshop at NeurIPS* (Vol. 2020).

   Mój wkład polegał na nadzorowaniu projektu, implementacji modelu generatywnego, przeprowadzeniu większości eksperymentów, przygotowaniu tekstu artykułu.

5. Maziarka, Ł., **Danel, T.**, Mucha, S., Rataj, K., Tabor, J., & Jastrzębski, S. (2019). Molecule-augmented attention transformer. In *Workshop on Graph Representation Learning, Neural Information Processing Systems.*

   Mój wkład polegał na konsultacji projektu, przeprowadzeniu eksperymentów z udziałem modeli podstawowych i weryfikacji modelu na zadaniach symulowanych.

6. Struski, Ł., Sadowski, M., **Danel, T.**, Tabor, J., & Podolak, I. (2023) Feature-based interpolation and geodesics in the latent spaces of generative models. **Under review** in *IEEE Transactions on Neural Networks and Learning Systems.*

   Mój wkład polegał na przeprowadzeniu eksperymentów na danych chemicznych, korekcie tekstu i udziale w procesie odpowiadania na recenzje.

Granty

PRELUDIUM
Narodowe Centrum Nauki
2021-2023
Stanowisko: Kierownik
Tytuł: „Połączenie symulacji molekularnej i uczenia głębokiego w projektowaniu leków de novo"
Numer: 2020/37/N/ST6/02728

OPUS
Narodowe Centrum Nauki
2022-2025
Stanowisko: Stypendysta
Tytuł: „Głębokie samoorganizujące się grafy neuronowe"
Numer: 2021/41/B/ST6/01370

Minigrant POB DigiWorld
Uniwersytet Jagielloński
2021-2022
Stanowisko: Kierownik
Tytuł: „Grywalizacja procesu projektowania leków"

Minigrant POB DigiWorld
Uniwersytet Jagielloński
2021
Stanowisko: Wykonawca
Tytuł: „Stabilność reprezentacji ukrytej molekuł w autoenkoderach"

Dydaktyka

Współorganizacja nowego kursu na kierunku informatyka: **Uczenie maszynowe w projektowaniu leków**. Kurs po raz pierwszy odbył się w roku akademickim 2021/2022 i był koordynowany przez dr Sabinę Podlewską. Do kursu zostały przygotowane materiały dydaktyczne dostępne pod adresem `github.com/gmum/umwpl2021`.

Bibliografia

[1] Łukasz Maziarka, Dawid Majchrowski, Tomasz Danel, Piotr Gaiński, Jacek Tabor, Igor Podolak, Paweł Morkisz, and Stanisław Jastrzębski. "Relative molecule self-attention transformer". In: *arXiv preprint arXiv:2110.05841* (2021).

[2] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. "Junction tree variational autoencoder for molecular graph generation". In: *International conference on machine learning*. PMLR. 2018, pp. 2323–2332.

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[4] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[5] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size". In: *arXiv preprint arXiv:1602.07360* (2016).

[6] Eric Jang, Shixiang Gu, and Ben Poole. "Categorical reparameterization with gumbel-softmax". In: *arXiv preprint arXiv:1611.01144* (2016).

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.