



dr hab. inż. Paweł Morawiecki, prof. IPI PAN

Warszawa, 15.06.2023

Recenzja rozprawy doktorskiej „Metody uczenia głębokiego w naukach farmaceutycznych” autorstwa Tomasza Danela

Tematyka rozprawy

Tematyka rozprawy dotyczy algorytmów głębokiego uczenia ze szczególnym uwzględnieniem grafowych splotowych sieci neuronowych i modeli generatywnych. Zaprojektowane i analizowane metody uczenia są osadzone w kontekście nauk farmaceutycznych w tym przede wszystkim przewidywania własności związków chemicznych i ich potencjalnego zastosowania jako leków. Tematyka podjęta przez Autora rozprawy jest bardzo obiecującym kierunkiem badań. Algorytmy uczenia maszynowego mogą pomóc w projektowaniu nowych leków poprzez przewidywanie jak różne modyfikacje strukturalne związku mogą wpłynąć na jego aktywność biologiczną. To pozwala na przyspieszenie procesu optymalizacji leków. Metody głębokiego uczenia mogą również pomóc w przewidywaniu potencjalnych efektów ubocznych leków, co jest kluczowe dla oceny ich bezpieczeństwa. Może to obejmować analizę danych z badań klinicznych lub z badań na komórkach i modelach zwierzęcych.

Charakterystyka rozprawy

Głównym celem rozprawy postawionym przez Doktoranta jest

„rozwińnięcie 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. ”

W pracy brakuje wyraźnie postawionej tezy (hipotezy) badawczej, ale do pewnego stopnia jest to zrozumiałe ze względu na formę doktoratu (cykl publikacji).

Praca składa się z cyklu 9 publikacji. Przewodnik po publikacjach zawiera krótkie omówienie podstawowych terminów i pojęć oraz charakterystykę każdego z artykułów. Pierwsze sześć prac (P1-P6) obejmuje zagadnienia dotyczące sieci grafowych użytych do modelowania związków chemicznych. Kolejne trzy prace (P7-P9) mają luźniejszy związek z deklarowanym tematem cyklu i są bardziej nastawione na udoskonalenie istniejących technik przetwarzania obrazu, co potencjalnie może być przydatne w kontekście szeroko rozumianych nauk farmaceutycznych.

Cykl publikacji

Poniżej przywołuję publikacje wchodzące do przedstawionego cyklu wraz z krótkim podsumowaniem.

[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

W tradycyjnych sieciach opartych na grafach, sąsiednie węzły nie są różnicowane, a komunikaty wysyłane z nich są niezależne od ich lokalizacji. Zdolność do rozróżniania tych węzłów podczas procesu uczenia mogłaby zapewnić modelowi większą adaptowalność. W przedstawionym artykule autorzy proponują rozwiązanie tego problemu poprzez wprowadzenie nowego modelu sieci neuronowej opartej na grafach, nazywanego przestrzenną siecią konwolucyjną na grafach (SGCN). Ten model ma stanowić bezpośrednie rozwinięcie klasycznej sieci konwolucyjnej i sieci grafowej.

[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

Aktualnie wiele prac wprowadzających nowe konstrukcje sieci grafowych koncentruje się na ulepszaniu warstw grafowych, ignorując całkowicie znaczenie wykorzystywanych cech atomowych. Często zdarza się, że nowe architektury są porównywane z poprzednimi przy użyciu odmiennego zestawu cech, co powoduje niejasności - nie wiadomo, czy lepsze wyniki są efektem nowych metod przetwarzania grafów, czy wynikają z użycia różnych cech atomów. W artykule znajdziemy metodyczne porównania różnych kombinacji cech

atomowych na kilku zestawach danych chemicznych. Wniosek płynący z badań wskazuje, że wybór odpowiedniego zestawu cech atomu może mieć równie duży wpływ na przewidywanie modelu, co dobór odpowiedniej architektury.

[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)

Dokowanie molekularne jest kluczowym narzędziem komputerowym używanym do wstępnej oceny potencjalnych kandydatów na leki. Poprzez dokowanie uzyskujemy hierarchię związków chemicznych na podstawie ich przewidywanego powinowactwa do celu (często białka) oraz potencjalne miejsca położenia w kieszeni wiążącej tego białka. W pracy Doktorant proponuje nową taksonomię modeli generatywnych używających dokowania.

[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)

Praca przekonująco pokazuje jak metody maszynowego uczenia mogą wspomagać studium *in silico*. Konkretnie przeprowadzono analizę dla wybranych podtypów cytochromu P450, odpowiedzialnego za metabolizm związków chemicznych w organizmie. Celem pracy było znalezienie bliskich analogów istniejących inhibitorów tych enzymów i porównanie ich powinowactwa w dokowaniu molekularnym.

[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.

Autorzy proponują nowy model ProGReST. Bazuje on na grafowej sieci neuronowej, która generuje reprezentację cząsteczki. Ta reprezentacja zawiera wektory cech dla wszystkich atomów danej cząsteczki. ProGReST funkcjonuje jako binarne drzewo decyzyjne, gdzie na każdym węźle porównuje reprezentację atomów z cząsteczki wejściowej do treningowej reprezentacji prototypowej zapisanej w tym węźle. Uzyskane wyniki są obiecujące (wygrywa z konkurencyjnym rozwiązaniem RMat w 4 z 5 zadań).

[P6] Łukasz Maziarka, Agnieszka Pocha, Jan Kaczmarczyk, Krzysztof Rataj, Tomasz Danel, and Michał Warchoń. “Mol-CycleGAN: a generative model for molecular optimization”. In: Journal of Cheminformatics 12.1 (2020)

W pracy zaproponowany został autoenkoder wariacyjny służący do modyfikacji związku. W każdym projekcie związanym z poszukiwaniem nowych leków, istotnym elementem jest optymalizacja molekuł. To zadanie jest odmienne od projektowania leków de novo (od podstaw) głównie dlatego, że mamy już znane punkty wyjściowe - struktury z pewnymi pożądanymi charakterystykami, jednakże inne ich właściwości stanowią przeszkodę dla dalszego postępu projektu.

[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

Artykuł dotyczy problemu uzupełniania brakujących lub niewyraźnych fragmentów zdjęć. Zaproponowana metoda wykorzystuje grafowe sieci neuronowe. Do rekonstrukcji brakujących regionów używamy architektury autoenkodera, w której enkoderem jest SGCN, a dekoderelem sieć dekonwolucyjna. Ewaluacja przeprowadzona została na zbiorach MNIST i SVHN i uzyskane rezultaty są zadowalające.

[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)

Charakter pracy jest typowo aplikacyjny. Głębokie sieci konwolucyjne zostały wykorzystane do klasyfikacji gatunków porostów.

[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

Interesująca praca próbująca połączyć zalety drzew decyzyjnych (łatwa interpretowalność) z większą elastycznością sieci neuronowych. Ogólny pomysł wzbogacony jest dodatkowymi technikami regularyzacyjnymi i prowadzi do konkurencyjnych wyników.

Wyżej przytoczone prace zostały przedstawione na dobrych i bardzo dobrych konferencjach (ranga B lub A) i opublikowane w wysoko-punktowanych czasopismach.

Uwagi

Lektura rozprawy była dla mnie trudna ze względu na interdyscyplinarny charakter badań. Szczególnie prace [P3] i [P4] zawierają sporo żargonu z chemioinformatyki i nauk farmaceutycznych.

Uważam, że przedstawiony cykl prac mógłby się ograniczyć do pozycji [P1-P6]. Ostatnie trzy artykuły mają bardzo luźny związek z tematyką cyklu (na siłę każdą pracę z głębokich sieci konwolucyjnych czy grafowych można by próbować włączyć).

Zauważyłem kilka niezręczności językowych, głównie wynikających z trudności tłumaczenia anglojęzycznych terminów na język polski. Przykładowo „model dyskryminacyjny” (raczej dyskryminatywny), „augmentacje”. Doktorant tłumaczy na polski tytuły artykułów w „przewodniku”, co wydaje mi się sztuczne i niepotrzebne.

Konkluzja

Przedłożony cykl prac ma wysoką wartość merytoryczną, prace napisane są starannie, a uzyskane wyniki przekonywujące. Uważam, że złożona rozprawa mgr Tomasza Danela spełnia wymagania ustawowe i zwyczajowe stawiane pracom doktorskim i może stanowić podstawę nadania stopnia doktora. Co więcej oceniam, że cykl publikacji wykracza pozytywnie poza zwykłe wymagania stawiane rozprawom doktorskim w informatyce technicznej. Podjęta tematyka jest poznawczo bardzo ciekawa i społecznie ważna. Dlatego też wnioskuję o wyróżnienie rozprawy.

Piotr Kowalecki