dr hab. inż. Paweł Morawiecki, prof. IPI PAN Warszawa, 25.06.2025

Recenzja rozprawy doktorskiej „Adapting Deep Learning Architectures for Drug Discovery" autorstwa Łukasza Maziarki

Tematyka rozprawy

Tematyka rozprawa doktorskiej skupia się na adaptacji głębokich sieci neuronowych do kluczowych wyzwań chemoinformatyki, ze szczególnym uwzględnieniem komputerowo wspomaganego projektowania leków. Proces odkrywania nowych leków jest złożony, kosztowny i wymaga znacznych zasobów, przy czym znakomita część kandydatów na leki nie przechodzi dalszych etapów badań klinicznych. Wysoki odsetek odrzuceń na tym etapie podkreśla potrzebę ulepszonych modeli predykcyjnych oraz bardziej efektywnych metod optymalizacji związków chemicznych już we wczesnych fazach procesu opracowywania leków.

Badania przedstawione w tej pracy koncentrują się wokół dwóch głównych obszarów: predykcji właściwości molekularnych oraz generowania nowych związków chemicznych. Rozprawa wprowadza nowatorskie modele głębokiego uczenia, które wykorzystują grafowe sieci splotowe (ang. Graph Convolutional Network, GCN), mechanizmy uwagi/atencji (ang. attention) oraz architektury transformerowe w celu zwiększenia dokładności i efektywności przewidywania właściwości molekuł. Dodatkowo przedstawia innowacyjne metody generowania i optymalizacji cząsteczek, w tym zastosowanie sieci typu GAN (ang. Generative Adversarial Network) oraz modeli bazujących na przepływach (ang. flow) do warunkowego generowania molekuł.

Dysertacja obejmuje opracowanie sieci Set Aggregation Network (SAN) do ulepszonej agregacji cech w modelach grafowych, wprowadzenie Przestrzennej Grafowej Sieci Splotowej (Spatial Graph Convolutional Network, SGCN), która integruje trójwymiarową informację przestrzenną w procesie splotu, oraz stworzenie modelu MAT (Molecule Attention Transformer) i jego wariantu R-MAT, które wykorzystują mechanizmy atencji do uchwycenia złożonych interakcji molekularnych. Ponadto zaproponowano HuggingMolecules – bibliotekę z otwartym źródłem, która ułatwia dostęp do wstępnie

wytrenowanych modeli do przewidywania właściwości molekularnych, porównuje różne reprezentacje atomowe używane w GCN w celu zidentyfikowania optymalnych cech wejściowych oraz wprowadza EmBERT – adaptację architektury BERT do uczenia wielozadaniowego.

W obszarze generowania molekuł rozprawa przedstawia Mol-CycleGAN – metodę optymalizacji cząsteczek wiodących z wykorzystaniem GAN-ów do modyfikacji istniejących molekuł, oraz PluGeN – generatywną sieć typu „plugin", która przekształca przestrzenie ukryte (ang. latent) dla kontrolowanego generowania cząsteczek. Modele te pokazują potencjał głębokiego uczenia nie tylko w zakresie przewidywania właściwości związków chemicznych, ale także w generowaniu nowych molekuł o pożądanych cechach, przyspieszając tym samym proces odkrywania leków.

Cykl publikacji

Praca składa się z cyklu dziewięciu publikacji. Pierwszych siedem prac dotyczy problemu predykcji molekularnych własności związków takich jak biologiczna aktywność czy toksyczność. Z kolei prace [P8] i [P9] koncentrują się na modelach generatywnych potencjalnie pomocnych przy tworzeniu nowych związków chemicznych w tym leków. Poniżej przedstawiam krótką charakterystykę poszczególnych artykułów, główne osiągnięcia i wyniki.

[P1] Set Aggregation Network as a Trainable Pooling Layer. International Conference on Neural Information Processing (ICONIP)

Praca opisuje sieć agregacji zbiorów (Set Aggregation Network, SAN) jako rozwiązanie ograniczeń tradycyjnych metod poolingowych w zakresie agregowania informacji złożonych, strukturalnych danych, takich jak grafy molekularne. W przeciwieństwie do konwencjonalnych metod, takich jak max-pooling czy sum-pooling, SAN oferuje elastyczne, uczące się podejście, które zachowuje bogactwo danych wejściowych, co czyni ją szczególnie skuteczną w zadaniach chemoinformatycznych, takich jak predykcja właściwości molekuł. Poprzez uczenie się optymalnych projekcji do agregacji cech, SAN dostosowuje architektury głębokiego uczenia do zmienności struktur chemicznych, poprawiając uogólnianie i dokładność predykcji. W eksperymentach na zbiorze danych Tox21, SAN przewyższyła tradycyjne techniki poolingowe wykazując swoją skuteczność w dokładnym przewidywaniu właściwości toksykologicznych związków chemicznych.

[P2] Spatial Graph Convolutional Networks. International Conference on Neural Information Processing (ICONIP)

Publikacja wprowadza sieć konwolucyjną grafów przestrzennych (Spatial Graph Convolutional Network, SGCN) – nową architekturę głębokiego uczenia, która adaptuje grafowe sieci konwolucyjne (GCN) poprzez integrację trójwymiarowych informacji

przestrzennych grafów molekularnych. Podejście to zwiększa zdolność modeli głębokiego uczenia do uchwycenia zarówno topologicznych, jak i przestrzennych relacji wewnątrz cząsteczek, co prowadzi do dokładniejszych predykcji właściwości chemicznych. Zdolność SGCN do uwzględniania cech przestrzennych oraz przeprowadzania augmentacji danych poprzez transformacje geometryczne zwiększa odporność modelu, szczególnie w zadaniach chemoinformatycznych. Analiza teoretyczna potwierdza, że SGCN stanowi właściwe uogólnienie konwolucyjnych sieci neuronowych (CNN), a wyniki eksperymentów wykazują przewagę zaproponowanej metody w klasyfikacji obrazów i predykcji właściwości chemicznych.

[P3] Molecule-augmented attention transformer. Workshop on Graph Representation Learning, Neural Information Processing Systems (NeurIPS Workshop)

Praca wprowadza model Molecule-augmented Attention Transformer (MAT), który był jednym z pierwszych modeli wykorzystujących mechanizmy self-attention do przewidywania właściwości molekuł. MAT integruje tradycyjną konwolucję grafową opartą na macierzy sąsiedztwa z mechanizmem atencji by działać jako ucząca się macierz odległości. Pozwala to modelowi MAT skutecznie uchwycić zarówno strukturę grafową, jak i trójwymiarowe relacje przestrzenne wewnątrz cząsteczek. Dodatkowo, strategia wstępnego uczenia (pre-training) modelu MAT poprawia jego zdolność do uogólniania oraz wydajność na zróżnicowanych zbiorach danych, ograniczając potrzebę intensywnego dostrajania hiperparametrów. MAT osiągnął bardzo dobre wyniki na benchmarkach przewidywania właściwości molekuł.

[P4] Relative Molecule Self-Attention Transformer. Journal of Cheminformatics, 16, 3, 2024.

Praca stanowi rozwinięcie badań z [P3], umożliwiono skuteczniejsze modelowanie złożonych relacji między atomami w cząsteczce, uwzględniając kluczowe cechy, takie jak trójwymiarowe odległości atomowe, informacje o wiązaniach chemicznych oraz relacje sąsiedztwa w grafie. To usprawnienie pozwala modelowi lepiej uchwycić złożone zależności międzyatomowe, niezbędne do dokładnego przewidywania właściwości molekularnych. Wprowadzono również ulepszoną procedurę wstępnego uczenia, obejmującą zarówno lokalne, jak i globalne zadania predykcyjne na poziomie molekuły, co dodatkowo zwiększa zdolność modelu do uogólniania na różnorodnych zbiorach danych molekularnych.

[P5] An Open-Source Library for Transformer-Based Molecular Property Prediction (Student Abstract). AAAI Conference on Artificial Intelligence (AAAI)

Publikacja opisuje bibliotekę w języku Python, która zapewnia łatwy w użyciu interfejs do stosowania modeli opartych na transformerach w zadaniach przewidywania właściwości molekuł. Poprzez ujednolicenie implementacji różnych modeli oraz ułatwienie ich

zastosowania dzięki spójnemu interfejsowi API, HuggingMolecules obniża próg wejścia dla badaczy chcących korzystać z tych zaawansowanych technik, czyniąc je bardziej dostępnymi dla szerszej społeczności naukowej. HuggingMolecules oferuje solidny i elastyczny zestaw narzędzi, który może przyspieszyć badania i innowacje w dziedzinach chemii i odkrywania leków, umożliwiając efektywniejsze dostosowanie modeli głębokiego uczenia do unikalnych wyzwań danych molekularnych.

[P6] Comparison of Atom Representations in Graph Neural Networks for Molecular Property Prediction. International Joint Conference on Neural Networks (IJCNN)

Doktorant przeprowadza kompleksową ocenę reprezentacji atomowych w sieciach grafowych (GCN) stosowanych do przewidywania właściwości molekuł. Wyniki wskazują, że wybór cech atomowych może znacząco wpływać na wydajność modelu – niektóre cechy, takie jak liczba ciężkich sąsiadów czy obecność atomów wodoru, przynoszą największe korzyści. Z kolei pominięcie niektórych cech, takich jak ładunek formalny czy aromatyczność, może w niektórych przypadkach poprawić skuteczność modelu. Przedstawione analizy jakościowe i ilościowe przyczyniają się do lepszego zrozumienia roli, jaką odgrywają reprezentacje atomowe w przewidywaniu właściwości molekuł z wykorzystaniem sieci GCN.

[P7] Multitask learning using BERT with task-embedded attention. International Joint Conference on Neural Networks (IJCNN)

Publikacja dotyczy nowej architektury EmBERT do uczenia wielozadaniowego, która rozszerza model BERT poprzez włączenie embeddingów do mechanizmu atencji, co umożliwia trenowanie modelu w trybie wielozadaniowym. EmBERT wprowadza minimalną liczbę dodatkowych parametrów, jednocześnie skutecznie rozpraszając informacje specyficzne dla zadań w całym modelu.

Wyniki na benchmarku GLUE pokazują, że EmBERT osiąga state-of-the-art w kilku zadaniach i przewyższa inne podejścia do uczenia wielozadaniowego, szczególnie w scenariuszach z ograniczoną ilością danych treningowych. Zaproponowany model można zastosować w odkrywaniu leków, gdzie przewidywanie wielu właściwości cząsteczek może skorzystać ze wspólnego uczenia się między zadaniami, co wpisuje się w szerszy cel adaptacji modeli głębokiego uczenia w celu zwiększenia efektywności odkrywania leków.

[P8] Mol-CycleGAN: a generative model for molecular optimization. Journal of Cheminformatics

Praca przedstawia Mol-CycleGAN – model generatywny zaprojektowany na potrzeby odkrywania nowych leków, w szczególności do optymalizacji właściwości cząsteczek przy jednoczesnym zachowaniu ich strukturalnego podobieństwa. Dzięki połączeniu architektur CycleGAN i JT-VAE, Mol-CycleGAN rozwiązuje kluczowe wyzwania w komputerowo wspomaganym projektowaniu leków (CADD), takie jak generowanie poprawnych cząsteczek oraz jednoczesne ulepszanie wielu właściwości. Wykorzystując

reprezentację cząsteczek opartą na grafach i zapewniając strukturalne podobieństwo dzięki funkcji straty tożsamości (identity loss), Mol-CycleGAN stanowi solidne rozwiązanie w zakresie optymalizacji struktur w projektowaniu leków.

Wyniki pokazują, że Mol-CycleGAN skutecznie przeprowadza transformacje strukturalne i optymalizuje właściwości cząsteczek, co czyni go wartościowym narzędziem we wczesnych etapach opracowywania leków.

[P9] PluGeN: Multi-Label Conditional Generation from Pre-trained Models. AAAI Conference on Artificial Intelligence (AAAI)

Artykuł prezentuje rozwiązanie PluGeN – elastyczny moduł zaprojektowany w celu rozszerzenia możliwości wcześniej wytrenowanych modeli generatywnych poprzez umożliwienie warunkowego generowania z uwzględnieniem wielu etykiet. Dzięki dekompozycji przestrzeni ukrytej modeli generatywnych na interpretowalne komponenty, PluGeN umożliwia niezależne sterowanie wieloma właściwościami cząsteczek, takimi jak rozpuszczalność, dostępność syntetyczna czy bioaktywność. Taka architektura ułatwia generowanie nowych związków o określonych właściwościach, a także optymalizację już istniejących cząsteczek. PluGeN stanowi przykład na to, jak architektury głębokiego uczenia mogą zostać zaadaptowane do przyspieszenia procesu generowania i optymalizacji kandydatów na leki o pożądanych, wieloparametrowych cechach.

Uwagi i pytania

Prace wchodzące w skład cyklu zostały opublikowane w materiałach konferencyjnych najlepszych konferencji i wysoko-punktowanych czasopismach. Zarówno prace jak i przewodnik napisane są bardzo starannie i czytelnie. W mojej ocenie przewodnik mógłby być krótszy i niektóre podrozdziały swobodnie można scalić w jeden bardziej zwięzły (np. opisujące prace P3 i P4).

Cały cykl zawiera aż 9 prac, uważam, że można było zrezygnować z tych mniej znaczących, szczególnie, że przedstawiony dorobek jest bogaty i opublikowany w bardzo dobrych miejscach.

Część prac zostało spisanych i opublikowanych ponad 5 lat temu. Ciekawi mnie na ile przeszły one „próbę czasu”? Na ile zmieniły się metody, podejście do problemu?

Praca P7 jest w napisana w kontekście modeli językowych. Doktorant stwierdza, że zaproponowany model można zastosować w odkrywaniu leków, gdzie przewidywanie wielu właściwości cząsteczek. Czy udało się to sprawdzić?

Konkluzja

Przedłożony cykl prac oceniam wysoko, prace napisane są starannie, a uzyskane wyniki przekonywujące. Uważam, że złożona rozprawa mgr Łukasza Maziarki spełnia wymagania ustawowe i zwyczajowe stawiane pracom doktorskim i może stanowić podstawę nadania stopnia doktora.

Biorąc pod uwagę uzyskane wyniki, społecznie ważką tematykę i realny wpływ pracy, wnioskuję o wyróżnienie rozprawy.