16 lutego 2026

Prof. dr hab. Jacek Tabor
Wydział Matematyki i Informatyki
Uniwersytet Jagielloński

Recenzja pracy doktorskiej Pana Spyridona Mouselinosa
„Visual and Language Reasoning in Deep Learning Models"

Rozprawa doktorska Pana Spyridona Mouselinosa dotyczy jednego z najbardziej fundamentalnych i jednocześnie najbardziej aktualnych problemów współczesnej sztucznej inteligencji, mianowicie pytania, czy modele głębokiego uczenia rzeczywiście rozumują, czy raczej w sposób wysoce wyrafinowany wykorzystują statystyczne regularności obecne w danych treningowych. Problem ten jest dziś absolutnie centralny w kontekście dużych modeli językowych oraz systemów multimodalnych.

Rozprawa opiera się na trzech pracach opublikowanych w bardzo dobrych, prestiżowych konferencjach międzynarodowych w tematyce sztucznej inteligencji i przetwarzania języka naturalnego. Są to miejsca należące do ścisłej czołówki w tej dziedzinie, co potwierdza wysoką jakość merytoryczną przedstawionych badań.

Przedstawione publikacje to:

- „Measuring CLEVRness: Black-box Testing of Visual Reasoning Models" (ICLR 2022),
- „A Simple, Yet Effective Approach to Finding Biases in Code Generation" (ACL 2023 Findings),
- „Beyond Lines and Circles: Unveiling the Geometric Reasoning Gap in Large Language Models" (EMNLP 2024 Findings).

Warto również zauważyć, że prace te są już cytowane w literaturze międzynarodowej. Zgodnie z danymi dostępnymi w Google Scholar, artykuł „Beyond Lines and Circles: Unveiling the Geometric Reasoning Gap in Large

Language Models" posiada obecnie 15 cytowań, „A Simple, Yet Effective Approach to Finding Biases in Code Generation" – 11 cytowań, natomiast „Measuring CLEVRness: Black-box Testing of Visual Reasoning Models" również jest cytowany w literaturze. Jak na stosunkowo niedawne publikacje, świadczy to o ich zauważalnym wpływie na rozwój badań w tej dziedzinie.

Pan Spyridon Mouselinos jest wiodącym autorem przedstawionych prac.

Przy czytaniu rozprawy bardzo pozytywne wrażenie sprawiło na mnie bardzo klarowne i czytelne wprowadzenie do rozprawy. Autor w sposób uporządkowany przedstawia historyczne tło problemu – od historycznego opisu przypadku „clever Hans" (konia który miał umieć tabliczkę mnożenia), poprzez klasyczne rozważania nad testowaniem sztucznej inteligencji, aż do współczesnych modeli głębokich i multimodalnych. Motywacja badań jest przedstawiona w sposób logiczny i przekonujący. Czytelnik od początku rozumie, dlaczego samo osiąganie wysokich wyników benchmarkowych nie może być utożsamiane z rzeczywistym rozumowaniem.

Wprowadzenie nie ma charakteru wyłącznie przeglądowego. Autor umiejętnie łączy perspektywę filozoficzną z bardzo konkretnymi problemami metodologicznymi dotyczącymi ewaluacji modeli. Dzięki temu rozprawa jest spójna koncepcyjnie, a postawione pytania badawcze są jasno sformułowane i dobrze umotywowane.

Praca „Measuring CLEVRness: Black-box Testing of Visual Reasoning Models" jest – w mojej ocenie – najbardziej nowatorską i koncepcyjnie najciekawszą częścią rozprawy. Autor podejmuje w niej problem oceny rzeczywistego rozumowania w modelach Visual Question Answering trenowanych na syntetycznym zbiorze CLEVR, który został zaprojektowany jako środowisko testujące rozumowanie relacyjne i kompozycyjne.

Punktem wyjścia pracy jest obserwacja, że wysoka skuteczność na standardowym teście nie gwarantuje, iż model rzeczywiście wykorzystuje strukturę relacyjną sceny. Autor proponuje sformalizowanie procesu ewaluacji jako gry dwuagentowej o sumie zerowej. Jeden z agentów (model VQA) odpowiada na pytania dotyczące sceny, natomiast drugi agent generuje kontrolowane modyfikacje tej sceny w taki sposób, aby – przy zachowaniu poprawnej odpowiedzi z punktu widzenia człowieka – doprowadzić do zmiany odpowiedzi modelu.

Istotnym elementem jest to, że modyfikacje nie są przypadkowymi perturbacjami pikselowymi, lecz semantycznie uzasadnionymi transformacjami

w przestrzeni sceny: zmiany pozycji obiektów, zamiany atrybutów, dodawanie lub usuwanie elementów, przy zachowaniu spójności logicznej pytania. Wykorzystana zostaje strukturalna reprezentacja sceny dostępna w CLEVR, co pozwala operować na poziomie obiektów i relacji, a nie surowych obrazów.

Co szczególnie ważne, całe podejście ma charakter black-box – nie zakłada dostępu do wag ani gradientów modelu. Atakujący agent korzysta wyłącznie z odpowiedzi modelu i na tej podstawie adaptacyjnie generuje kolejne modyfikacje. W pracy analizowane są różne strategie generowania perturbacji, w tym procedury systematycznie eksplorujące przestrzeń relacji przestrzennych oraz kombinacji atrybutów.

Autor wprowadza również ilościową miarę „CLEVRness", która pozwala ocenić odporność modelu na tego typu semantyczne modyfikacje. Dzięki temu możliwe jest porównywanie modeli nie tylko pod względem surowej dokładności, ale również pod względem stabilności odpowiedzi w obliczu logicznie neutralnych zmian sceny.

Wyniki eksperymentalne są bardzo przekonujące. Modele osiągające niemal perfekcyjne wyniki na standardowym benchmarku wykazują znaczące spadki skuteczności po wprowadzeniu niewielkich, semantycznie uzasadnionych zmian. Autor pokazuje, że modele często opierają się na skrótach heurystycznych, takich jak lokalne korelacje między atrybutami obiektów, zamiast budować pełną reprezentację relacyjną sceny. Oznacza to, że problem ma charakter strukturalny, a nie jedynie implementacyjny.

W pracy przeprowadzono również analizy porównawcze różnych architektur, co pozwala wskazać, które typy modeli są bardziej odporne na tego rodzaju testy. Co istotne, nawet modele zaprojektowane specjalnie z myślą o rozumowaniu relacyjnym nie są całkowicie odporne na proponowane procedury testowe.

Podsumowując, praca ta wprowadza nową, ogólną metodologię testowania modeli rozumowania wizualnego, która może być adaptowana do innych zbiorów danych i innych modalności. Praca ta nie tylko diagnozuje ograniczenia współczesnych modeli, ale redefiniuje sposób myślenia o ich ewaluacji. W mojej opinii stanowi ona istotny wkład w dyskusję nad tym, czym jest rzeczywiste rozumowanie w systemach głębokiego uczenia i może mieć trwały wpływ na rozwój metod oceny modeli multimodalnych.

Druga praca, „A Simple, Yet Effective Approach to Finding Biases in Code Generation", dotyczy analizy uprzedzeń oraz skrótów heurystycznych w modelach generujących kod. Punktem wyjścia jest obserwacja, że wysoka

skuteczność modeli na benchmarkach programistycznych (np. zadaniach typu HumanEval) nie musi oznaczać rzeczywistego rozumienia specyfikacji problemu.

Autor proponuje elegancką i jednocześnie bardzo przejrzystą metodologię nazwaną „Blocks of Influence". Prompt zadania programistycznego zostaje rozłożony na trzy zasadnicze komponenty: (1) nazwę funkcji (Name Block), (2) opis zadania w języku naturalnym (Description Block) oraz (3) przykłady wejścia–wyjścia (Example Block). Następnie, w kontrolowany sposób, usuwa lub modyfikuje poszczególne bloki, zachowując semantyczną równoważność zadania.

Kluczowe jest to, że modyfikacje nie zmieniają logicznej treści problemu, lecz eliminują powierzchowne wskazówki, takie jak sugestywne nazwy funkcji czy charakterystyczne frazy. Autor analizuje, w jakim stopniu skuteczność modelu zależy od każdego z tych bloków oraz czy model potrafi poprawnie wygenerować rozwiązanie w sytuacji, gdy dostępne są wyłącznie formalne przykłady, bez opisu słownego, bądź odwrotnie.

Przeprowadzone eksperymenty obejmują różne modele generujące kod i różne konfiguracje promptów. Wyniki pokazują wyraźnie, że modele w dużym stopniu polegają na tokenach-kluczach oraz powierzchownych korelacjach między nazwą funkcji a typowym wzorcem rozwiązania. W wielu przypadkach usunięcie nazwy funkcji powoduje istotny spadek skuteczności, mimo że opis problemu pozostaje niezmieniony. Wskazuje to na obecność efektów memorystycznych oraz silne uzależnienie od statystycznych regularności w danych treningowych.

Autor przeprowadza również analizy porównawcze pomiędzy modelami o różnej wielkości oraz architekturze, pokazując, że problem nie zanika wraz ze wzrostem liczby parametrów. Co istotne, metodologia ma charakter black-box i może być stosowana bez dostępu do wnętrza modelu, co czyni ją praktycznym narzędziem ewaluacyjnym.

Praca ta ma duże znaczenie praktyczne. W kontekście rosnącego wykorzystania modeli generujących kod w zastosowaniach przemysłowych oraz edukacyjnych, zrozumienie ich ograniczeń jest kluczowe. Zaproponowana metodologia może stanowić element bardziej rygorystycznej procedury testowania systemów programistycznych opartych na LLM.

Trzecia praca, „Beyond Lines and Circles: Unveiling the Geometric Reasoning Gap in Large Language Models", dotyczy rozumowania geometrycznego jako szczególnie wymagającego przypadku rozumowania symboliczno-

przestrzennego. Autor wychodzi z założenia, że geometryczne zadania konstrukcyjne stanowią istotny test zdolności do planowania sekwencji operacji oraz operowania relacjami przestrzennymi.

W pracy wykorzystano środowisko zadań konstrukcyjnych, w których należy – przy użyciu operacji typu rysowanie linii, okręgów czy punktów przecięcia – zrealizować określoną konstrukcję geometryczną. Zadania te wymagają nie tylko wiedzy deklaratywnej, lecz również planowania wieloetapowego i utrzymywania spójnej reprezentacji przestrzennej.

Autor pokazuje, że duże modele językowe, które osiągają bardzo dobre wyniki w zadaniach algebraicznych i arytmetycznych, mają wyraźne trudności w zadaniach geometrycznych wymagających konstrukcyjnego myślenia przestrzennego. Błędy mają często charakter strukturalny: modele generują sekwencje operacji, które są lokalnie sensowne, lecz globalnie niespójne lub nie prowadzą do realizacji celu.

Szczególnie interesującym wkładem pracy jest zaproponowanie architektury wieloagentowej. W zaprojektowanym systemie różne moduły odpowiadają za odmienne role: planowanie strategii rozwiązania, generowanie konkretnych operacji konstrukcyjnych oraz weryfikację poprawności uzyskanej konfiguracji. Takie rozdzielenie odpowiedzialności pozwala ograniczyć typowe błędy, takie jak halucynacje narzędzi czy generowanie operacji niespójnych z dotychczasową konstrukcją.

Autor analizuje również wpływ technik takich jak few-shot prompting, adaptacyjne dobieranie przykładów czy translacja reprezentacji wizualnej do tekstowej. Wyniki pokazują, że choć możliwe jest częściowe zmniejszenie luki w rozumowaniu geometrycznym, problem ten pozostaje istotnym wyzwaniem dla obecnych architektur.

Praca ta stanowi ważny krok w kierunku bardziej precyzyjnej diagnozy ograniczeń dużych modeli językowych. Wskazuje ona wyraźnie, że zdolności algebraiczne i językowe nie przekładają się automatycznie na kompetencje przestrzenne i konstrukcyjne, co ma istotne znaczenie dla dalszego rozwoju modeli multimodalnych i systemów wspomagających rozwiązywanie problemów matematycznych.

Podsumowując, rozprawa doktorska Pana Spyridona Mouselinosa stanowi bardzo wartościowy i dojrzały wkład w badania nad rozumowaniem w sztucznej inteligencji. Szczególnie wysoko oceniam pierwszą pracę, która w mojej opinii jest najbardziej nowatorska i metodologicznie przełomowa.

Rozprawa spełnia zarówno zwyczajowe, jak i ustawowe wymagania sta-

wiane pracom doktorskim w dyscyplinie informatyka. W konsekwencji, wnoszę o dopuszczenie Pana Spyridona Mouselinosa do dalszych etapów przewodu doktorskiego oraz o przyznanie stopnia doktora nauk ścisłych i przyrodniczych w dyscyplinie informatyka. Finalnie, w związku z wysoką jakością prezentowanych wyników wnoszę o wyróżnienie rozprawy.

Jacek Tabor

Podpisany elektronicznie przez
Jacek Tabor
19.02.2026
11:05:52 +01'00'