СЕКЦІЯ 7
СИСТЕМНИЙ АНАЛІЗ

УДК 004.8:004.932

https://doi.org/10.68134/ah.2026-11-20.023

Вознюк Вікторія Романівна
студент, Київський політехнічний інститут імені Ігоря Сікорського

АНАЛІЗ МУЛЬТИМОДАЛЬНОГО ШТУЧНОГО ІНТЕЛЕКТУ: ЕВОЛЮЦІЯ МЕХАНІЗМІВ CROSS-ATTENTION ТА ЇХ ПРАКТИЧНИЙ ПОТЕНЦІАЛ

Анотація. Незважаючи на стрімкий розвиток штучного інтелекту, він досі лишається недосконалим. Дане дослідження зумовлене зростанням використання великих візуально-мовних моделей у критичних системах прийняття рішень, де генерування ШІ об'єктних галюцинацій створює серйозні ризики. Метою роботи є системний аналіз першопричин хибних візуалізацій у моделях LLaVA, BLIP-2 та дослідження способів їх мінімізації. Також продемонстровано метрики, що дозволяють оцінити моделі в умовах випадкової, популярної та змагальної вибірки об'єктів. В результаті дослідження визначили, що галюцинації виникають переважно через домінування мовних статистичних зв'язків над візуальними доказами, отриманими через механізми перехресної уваги. У змагальних вибірках точність моделей критично знижується, що свідчить про вразливість до лінгвістичних упереджень. У висновку приходимо до необхідності перегляду підходів до візуального налаштування інструкцій, а також впровадження керування увагою для підвищення надійності систем розпізнавання у складних умовах. Практичне значення отриманих результатів полягає у можливості безпечної інтеграції мультимодальних систем у процеси дистанційного зондування та автоматизованої оцінки пошкоджень інфраструктури під час відбудови України. Подальші дослідження доцільно спрямувати на розробку алгоритмів збалансування навчальних даних і масштабування архітектур для аналізу відеоконтенту та просторових тривимірних моделей.

Ключові слова: великі візуально-мовні моделі, об'єктні галюцинації, штучний інтелект, перехресна увага, візуальне налаштування інструкцій, системний аналіз.

Abstract. Despite the rapid development of artificial intelligence, it remains imperfect. This research is driven by the increasing use of large vision-language models in critical decision-making systems, where AI-generated object hallucinations pose severe risks. The aim of this study is a systematic analysis of the root causes of erroneous visual outputs in LLaVA and BLIP-2 models, along with an exploration of methods to minimize them. Furthermore, metrics were employed to evaluate the models under conditions of random, popular, and adversarial object sampling. The research results indicate that hallucinations primarily arise due to the dominance of statistical linguistic associations over visual evidence extracted via cross-attention mechanisms. Under adversarial sampling conditions, model accuracy decreases critically, indicating a vulnerability to linguistic biases. The findings highlight the necessity of revising approaches to visual instruction tuning, as well as implementing attention management to enhance the reliability of recognition systems in complex environments. The practical significance of the obtained results lies in enabling the safe integration of multimodal systems into remote sensing and automated infrastructure damage assessment during the reconstruction of Ukraine. Future research should focus on developing training data balancing algorithms and scaling architectures for video content and 3D spatial scene analysis.

Keywords: large vision-language models, object hallucinations, artificial intelligence, cross-attention, visual instruction tuning, systems analysis.

Вступ

Останніми роками можна помітити тенденцію штучного інтелекту до переходу від ізольованих моделей обробки природної мови (NLP) та комп'ютерного зору (CV) до інтегрованих мультимодальних архітектур. Великі візуально-мовні моделі, такі як LLaVA (Large Language and Vision Assistant) та BLIP-2 (Bootstrapping Language-Image Pre-training), демонструють безпрецедентну здатність до високоякісного аналізу зображень, генерування описів та ведення діалогів [1]. Такий науковий прорив став можливим завдяки візуальному налаштуванню, яке використовує машинно-генеровані інструкції для адаптації розуміння візуальних вхідних даних.

Однак попри значні успіхи, масштабування цих моделей виявилося критично вразливим, що загрожує використанню у сферах, де помилки є недопустимими. Ця проблема відома як «об'єктні галюцинації» — феномен, при якому модель генерує відповіді, що при цьому по факту відсутні. Традиційні метрики оцінювання, такі як CHAIR (Caption Hallucination Assessment with Image Relevance), показали, що моделі з візуальним налаштуванням інструкцій часто виявляються більш схильними до галюцинацій, ніж простіші попередники [2]. Це створює серйозні проблеми для прикладного застосування штучного інтелекту, наприклад, у оцінюванні руйнувань, де хибна ідентифікація «уламків» може дезорієнтувати спецслужби. Відповідно, аналіз механізмів виникнення галюцинацій та розробка методів їх об'єктивного вимірювання є одним із найактуальніших завдань сучасної комп'ютерної інженерії.

Мета і завдання

Метою цієї роботи є комплексне дослідження феномену об'єктних галюцинацій у сучасних великих візуально-мовних моделях, виявлення їх архітектурних першопричин та оцінювання ефективності існуючих підходів.

Методи

В основі дослідження лежить аналітичний підхід, що спирається на використання бенчмарку POPE (Polling-based Object Probing Evaluation), спеціально розробленого для стабільного та гнучкого оцінювання об'єктних галюцинацій. На відміну від метрики CHAIR, POPE використовує підхід бінарного опитування. Моделі надається зображення та запитання формату: «Чи є об'єкт на цьому зображенні?» з вимогою надати чітку відповідь «Так» або «Ні» [3].

Вибірка структурує негативні, тобто відсутні на зображенні, об'єкти за трьома маркерами:

1. Випадкова вибірка: Негативні об'єкти генеруються випадково з усього набору.
2. Популярна вибірка: Використовуються об'єкти, які найчастіше зустрічаються в усьому наборі даних, проте відсутні на поточному зображенні.
3. Змагальна вибірка: Негативні об'єкти підбираються на основі наявності спільних рис з тими, які дійсно наявні на зображенні (наприклад, питання про «стілець», якщо на фото є «стіл», але в дійсності стілець відсутній).

У межах дослідження було розглянуто особливості моделей LLaVA-1.5 та BLIP-2, які пройшли етап візуального налаштування інструкцій. Теоретичне порівняння їхньої стійкості до об'єктних галюцинацій базується на критеріях бенчмарку POPE з фокусом на ключові метрики — точність (Accuracy) та гармонійне середнє (F1-score), що відображають баланс між розпізнаванням та здатністю моделі уникати хибнопозитивних значень.

Результати та обговорення

Системний аналіз сучасних візуально-мовних моделей демонструє, що головним показником якості їхньої роботи є поєднання візуального енкодера з великою мовною моделлю. У моделі BLIP-2 для цього використовується проміжний модуль Q-Former, який за допомогою

механізму перехресної уваги витягує лише найрелевантніші візуальні ознаки та передає їх в текст [1]. Натомість LLaVA-1.5 застосовує простіший підхід, який дозволяє транслювати безпосередньо візуальні токени [1].

Попри різні методи, обидві моделі вразливі до об’єктних галюцинацій. Аналіз оцінювання за методологією POPE показує чітку закономірність: якщо на випадковій вибірці моделі зберігають високу точність, то при переході до популярної/змагальної вибірок показники Accuracy та F1-score суттєво знижуються. Першопричиною цього явища є домінування статистичних мовних зв’язків над фактичними візуальними доказами [4]. Коли модель бачить на зображенні об’єкти, що часто зустрічаються разом з іншими предметами, мовна складова по суті «додумує» відсутній об’єкт, ігноруючи реальні вимоги. У контексті відбудови України цей недолік має критичне значення. Інтеграція мультимодальних систем у процеси дистанційного зондування Землі, аналізу супутникових знімків та аерофотозйомки відкриває значний потенціал для автоматизованої оцінки пошкоджень інфраструктури. Проте в умовах складної візуальної сцени (завали, задимлення, руйнування) лінгвістичні упередження моделі можуть призводити до хибної ідентифікації критичних об’єктів або некоректної класифікації ступеня руйнувань. Мінімізація таких ризиків потребує перегляду візуального налаштування інструкцій — зокрема, збільшення частки негативних прикладів у навчальних вибірках та впровадження методів динамічного керування увагою для посилення ваги візуальних токенів під час генерації відповіді [5].

Висновки

У роботі було проведено аналіз проблеми об'єктних галюцинацій у великих візуально-мовних моделях (на прикладі LLaVA-1.5 та BLIP-2) та ролі механізмів перехресної уваги в їх виникненні. Також було встановлено, що основною причиною хибних візуалізацій є перевага накопичених лінгвістичних упереджень над візуальними даними, що найбільше проявляється в умовах змагальної вибірки бенчмарку POPE.

Підвищення надійності мультимодальних систем є необхідною умовою для їх безпечного впровадження у прикладні задачі, зокрема для автоматизованого моніторингу та оцінки пошкоджень інфраструктури. Подальші дослідження доцільно спрямувати на розроблення алгоритмів збалансування навчальних датасетів, удосконалення механізмів керування увагою, а також на масштабування мультимодальних архітектур для аналізу відеопотоків і просторових тривимірних сцен.

Список використаних джерел

1. Comparing-Image-to-Text-Models-BLIP-BLIP2-and-LLaVA. GitHub. URL: https://github.com/AthiraSreenath/Comparing-Image-to-Text-Models-BLIP-BLIP2-and-LLaVA (дата звернення: 17.11.2026)