

Karsa Róbert¹, Négyesi Imre²

Nagy nyelvi modellek és zárt információs rendszerek a védelmi szférában

Large Language Models and Closed Information Systems in the Defence Sector

Absztrakt

A mesterséges intelligencia, különösen a nagy nyelvi modellek és a gépilátás-rendszerek jelentős változásokat hoznak a mindennapi életünkben. Ez a tanulmány feltárja a mesterséges intelligencia védelmi szférában történő alkalmazásának lehetőségeit és kihívásait. Bemutatjuk a nyelvi modellek működését, képzési folyamatait, a generatív modellek képességeit és korlátait, valamint a képfeldolgozás terén elért eredményeket, beleértve a vizuális nagy nyelvi modelleket. Kiemeljük a vektorbeágyazások és vektoradatbázisok fontosságát az információkeresésben, valamint az adatlekérésre alapozott szöveggenerálás szerepét a hallucinációk csökkentésében. Vizsgáljuk a nyelvi modellek képzésének költségeit és a magyarországi lehetőségeket is. A tanulmány rámutat, hogy a nagy nyelvi modellek hadtudományban való alkalmazása jelentős potenciált rejt, de a bizalmas adatok védelme érdekében elengedhetetlen a saját, biztonságos informatikai infrastruktúra kiépítése. Egy zárt információs rendszer segítségével bemutatjuk, hogy a védelmi szektor hogyan tudja kihasználni a technológia előnyeit a biztonságos információkezelés érdekében. Következtetésként hangsúlyozzuk a hosszú távú beruházás és a folyamatos innováció szükségességét a védelmi szektor számára.

Kulcsszavak: mesterséges intelligencia, nagy nyelvi modell, hallucináció, vektoradatbázis, hadtudomány

¹ Doktori hallgató, Nemzeti Közszerológati Egyetem Katonai Műszaki Doktori Iskola; egyetemi tanársegéd, Pécsi Tudományegyetem Természettudományi Kar Matematikai és Informatikai Intézet, e-mail: karsa.robert@pte.hu

² Egyetemi docens, Nemzeti Közszerológati Egyetem Hadtudományi és Honvédtisztképző Kar, e-mail: negyesi.imre@uni-nke.hu

Abstract

Artificial intelligence, especially large language models and machine vision systems, are bringing major changes to our everyday lives. This paper explores the opportunities and challenges of applying AI in the defence domain. We present the operation and training processes of language models, the capabilities and limitations of generative models, and advances in image processing, including visual large language models. We highlight the importance of vector embeddings and vector databases in information retrieval, and the role of data query-based text generation in reducing hallucinations. We also examine the costs of training language models and the opportunities in Hungary. The paper shows that the use of large language models in military science has a significant potential, but that it is essential to build a dedicated, secure IT infrastructure to protect confidential data. By demonstrating a closed information system, we show how the defence sector can take advantage of the technology to ensure secure information management. In conclusion, we emphasise the need for long-term investment and continuous innovation in the defence sector.

Keywords: artificial intelligence, large language model, hallucination, vector database, military science

Bevezetés

A mesterséges intelligencia (MI) és ezen belül a nagy nyelvi modellek (LLM-ek) jelentős változásokat hoznak mindennapi életünkben és a hadtudomány területén is. Új készségeket, lehetőségeket kell megtanulnunk, megismernünk. A védelmi szférának ezekhez a változásokhoz alkalmazkodnia kell, sőt a gyors technológiai felzárkózással előnyt kell kovácsolnia a lehetőségekből. A kockázatkerülő háborúk korában, ahol a kibereszközök, az autonóm fegyverrendszerek és a katonai képességfejlesztés dominál, az MI központi szerepet játszik a harctéri kockázatok minimalizálásában és a hatékonyság növelésében. Az LLM-ekkel kapcsolatos kutatásunk elsősorban a katasztrófavédelmi hatósági információk feldolgozásával foglalkozik, de az elérhető eredmények túlmutatnak a katasztrófavédelmi területen, és olyan komplex védelmi rendszerek fejlesztéséhez vezetnek, amelyek segítségével új honvédelmi és rendvédelmi képességek jöhetnek létre. Az MI alkalmazása a hadtudományban több területen is megfigyelhető. A kiberhadviselés során az MI jelentősen növeli annak hatékonyságát mind a támadó, mind a védekező oldalon. Képes gyorsabban felismerni a szokatlan tevékenységeket, valamint automatizált válaszokat adni a fenyegetésekre. Az MI kulcsfontosságú az autonóm fegyverrendszerek, például a drónok működésében, lehetővé téve számukra az önálló döntéshozatalt és a célok hatékonyabb elérését. Az MI a katonák képességfejlesztésében is fontos szerepet játszhat, segítheti az agy-számítógép interfészek fejlesztését, ami javíthatja a tanulást, a kiképzést és a döntéshozatalt, továbbá az MI képes a taktikai döntések támogatására, csökkentve a bizonytalanságot a harctéren. Az MI alkalmazásakor be kell tartani az igazságos háború elméletének elveit, és az emberi kontrollt minden esetben biztosítani kell. Célunk, hogy bemutassuk azokat a módszereket és technológiákat, amelyek jelenleg a modern információs infrastruktúrák adatfeldolgozási részének belső alapköveit

képezik, valamint rámutassunk a zárt információkereső rendszerek egy lehetséges változatának védelmi célú alkalmazására. Az új eredmény ezen alkalmazások lehetséges fejlesztésének és a technológiai háttér összefüggéseinek bemutatásában rejlik.

Adatfeldolgozási infrastruktúrák

LLM-ek

Amikor számítógépek segítségével próbálunk meg leírni egy folyamatot, akkor modelleket készítünk, amelyek reményeink szerint egy elvárt viselkedést mutatnak. A nyelvi modell matematikai értelemben olyan valószínűségi eloszlás a szavak sorozatai között, ahol a modell minden egyes szóhoz valószínűségi értéket rendel egy szekvenciában, azaz a szövegben a következő szót jelzi előre az előtte lévő szavak alapján. Az LLM már olyan típusú nyelvi modell, amelynek hátterében egy neurális hálózat van. Ez a neurális hálózat architektúráját tekintve jellemzően egy transzformer, amely jelentős számítási kapacitás igénybevételével hatalmas szövegtörzseten tanul; egyik legismertebb típusa a GPT-3 vagy éppen a Llama. Ezek a neurális hálózatok az információkat a paramétereikben (számok) tárolják az általánosan megszokott dokumentumok helyett. Ezt a tudást paraméteres tudásnak is hívjuk. Az LLM-ek esetén ezeknek a paramétereknek a száma több milliárdos nagyságrendet is elérhet. A modellek nevében található számok és a B jelölés általában a paraméterszámra utalnak milliárdban kifejezve.

Az LLM-ek képzése

A modellek képzése rendszerint két vagy három lépésből áll. Az első az előképzés, amely során a modell megszerzi az alapvető tudását és képességeit a világunkkal kapcsolatban. Ez a tanulás önfelügyelt tanulás, amelyben a modell maga generálja a tanuláshoz szükséges címkéket. Ez azt jelenti, hogy a tanulási folyamatban nem szükséges emberi közreműködés (kézi címkézés), mivel a modell saját magának hoz létre célokat a tanulás során, ami rendszerint a tanulandó szöveg következő szótörzse. A képzés során a modellnek mindig a következő szót kell kitalálnia, a képzés alatt a modellnek nincs információja a kitalálandó szó utáni szavakról. A képzés során a modell megtanulja a nyelv mintázatait, és reprezentációkat fejleszt ki a szavak és a szövegtörzset közötti kapcsolatokra. A modell megérti a kontextust és képes előrejelzéseket tenni a szövegben következő szavakra vagy mondatokra vonatkozóan.

Ez a tudás a tanítóadatokat tartalmazó akár több milliárd szavas korpusz felett értelmezhető, tehát a modell ismeretei és az ismeretek mélysége egy témában elsősorban a témához kapcsolódó tanítóadatok mennyiségén alapszik. A képzés második szakaszában felügyelt finomhangolás történik, amely már jóval kevesebb erőforrást igényel, mint az előképzés. A felhasználói szándék általában túlmutat azon, hogy egy szöveghez a következő szót előre jelezzük, vagyis egy szöveg írását folytassa. A problémának a megoldására fejlesztették ki a modell-finomhangolási technikát. Ennek során a modellt előre gondosan összeállított utasítás-válasz párokkal

tanítják tovább, amelynek eredményeképpen a modell képes lesz utasítások, kérdések megválaszolására, és annak általánosító képessége is javul. A finomhangolás során az utasítás-válasz párokból feladattól függően a néhány ezertől akár sok százezeres mennyiséget is használhatunk.

A finomhangolás azt is jelenti, hogy egy általános képességekkel rendelkező nyelvi alapmodellt továbbképzünk tartományspecifikus adatokkal, és az adott tartomány szakértőjévé válhat a nyelvi modell.

A képzés opcionális harmadik szakaszában a hangsúlyt arra helyezik, hogy a válaszok jobban kövessék az emberi preferenciákat. Ilyenkor a képzés során emberi visszajelzéseket is felhasználnak; egyik legismertebb ilyen képzési módszer az RLHF, amely során megerősítéssel tanulást alkalmaznak felhasználva az emberi visszajelzéseket. Ennek eredményeképpen a modell megtanulja az emberi preferenciákat, és válaszai sokkal „emberibbnek” tűnnek.

Képzési költségek

Az LLM-ek képzése rendkívül költségigényes feladat, ezért nagyon kevés cég képes arra, hogy saját modellt fejlesszen. A fejlesztési költségek a modell növekedésével (a paraméter számának növekedésével) skálázódnak. A BLOOM nevű, 176 milliárd paraméterrel rendelkező nyelvi modell előképzése nagyjából 118 napig tartott és 433 196 kWh áramfogyasztással járt. Az előképzést 384 NVIDIA A100 GPU-val, 48 számítási egység segítségével végezték el. Magyarországon jelenleg csak a Komondor nevű szuperszámítógép képes hasonló feladatok elvégzésére. A Komondor specifikációja szerint a GPU partíciója 232 db NVIDIA A100-as GPU-val rendelkezik, és a fenti BLOOM-modell képzése számításaink szerint több mint 190 napig tartana, amennyiben a szuperszámítógép csak az LLM képzésével foglalkozna.

Hazánkban jelenleg a legnagyobb nyilvánosan elérhető és magyar nyelvű adatokon képzett modell a Nyelvtudományi Kutatóközpont által készített PULI modellcsalád.

Generatív modellek

A legismertebb ilyen modell a GPT nevezetű LLM, amelyet az OpenAI fejlesztett ki. A GPT-modell előzetes tanítása szintén hatalmas mennyiségű szöveges adattal történt. A GPT-modell generatív jellege azt jelenti, hogy képes új szöveget generálni, amely a tanítóadatokból tanult nyelvi szerkezeteket és jellemzőket követi. A GPT modellek sokoldalúak, és széles körben alkalmazhatók szöveggenerálásra, ember és gép közötti párbeszéd megvalósítására. A GPT előnye, hogy kontextuális reprezentációkat készít az előző szövegrészeket figyelembevételével, és így képes értelmes és koherens szöveg generálására. A 2018-ban megjelent GPT-modell újabb verziókkal bővült, mint a GPT-2, a GPT-3 és a GPT-4o. Ezek egyre nagyobb méretűek és jobbak lettek. A GPT-típusú modellek ma a legjelentősebb és legfejlettebb nyelvi modellek közé tartoznak, és a természetes nyelvfeldolgozással kapcsolatos alkalmazások széles körben használják őket.

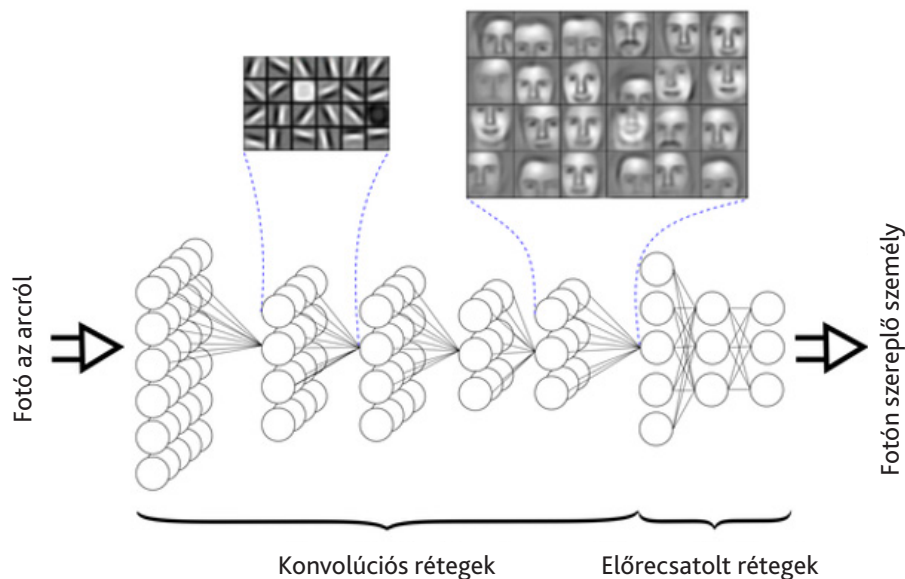
Az LLM-ek segítségével a számítógéppel emberi nyelven tudunk kommunikálni, ami olyan távlatokat nyit meg az ember-gép kommunikációban, amelyek néhány éve még teljesen elképzelhetetlennek tűntek.

Hallucináció

Az LLM-ek többnyire helyes és jó információkat adnak vissza köszönhetően a nagy paraméteres tudásbázisuknak, azonban időnként előfordul, hogy olyan tényeket állítanak, amelyek hihetőnek tűnnek, de ellentmondásosak, vagy éppen nincsenek összhangban a valós tényekkel. A hallucináció fogalma a patológia és a pszichológia területéről ered, és olyan észlelést jelent, amelyben egy entitás vagy esemény jelenik meg, miközben az valójában nem létezik. Az LLM-ek esetén a hallucináció egy olyan jelenség, amikor a generált szöveg értelmetlen, hibás vagy félrevezető. A hallucinációkat két fő csoportra bonthatjuk. Az első a ténybeli ellentmondás, amely során egy állítás valós információkra alapozható, azonban ellentmondásokat tartalmaz. Ilyen például az, amikor egy modellt az izzó feltalálójáról kérdeznek, és az helytelenül Thomas Edisont adja válaszként, aki nem a feltaláló volt, de kapcsolódott az izzóhoz, hiszen tökéletesítette az addigi tervezési eljárást. A másik hallucinációtípus a tények gyártása, amikor a modell olyan tényeket állít, amelyek nem ellenőrizhetők a meglévő tudásunk szerint. Ilyen például, amikor az „Eiffel-torony építésének főbb környezeti hatásairól” kérdezték a modellt, az helytelenül azt állította, hogy „az építkezés a párizsi tigris kihalásához vezetett”, pedig ez a faj nem létezik, így ezt az állítást semmilyen történelmi vagy biológiai feljegyzés nem támasztja alá. A hallucináció enyhítésének egyik módja a lekérdezésalapú szöveggenerálás, a RAG, amikor a nyelvi modell nemcsak a belső paraméteres tudására támaszkodik, hanem külső adatforrásokból is felhasznál információkat a válasz generálásához.

Gépi látás konvolúciós neurális hálózatok segítségével

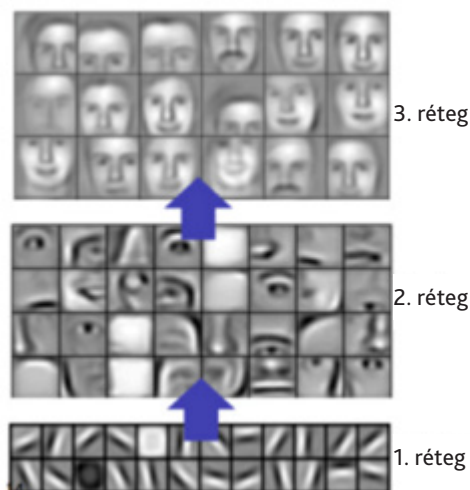
A számítógépes képfeldolgozási feladatokhoz fejlesztették ki az úgynevezett konvolúciós neurális hálózatokat (CNN), amelyek azon az ötleten alapulnak, hogy a hálózat különböző rétegei a kép egyes vonásait tanulják meg, és ezeket kombinálva annak egyre bonyolultabb részleteit képesek felismerni.



1. ábra: Konvolúciós neurális hálózat

Forrás: TÓTH 2016: 945

Ez a folyamat látható az 1. ábrán. A CNN első része a kép tulajdonságait tanulja meg, majd a CNN második része végzi az osztályozási feladatot. A CNN rétegekből áll, és az adatoknak a rétegeken kell keresztülmenni, ami egyben az adatok feldolgozását is jelenti. A CNN első rétege a bemeneti réteg, itt adjuk át az információt a hálózatnak, ami esetünkben egy kép. Az első rétegek még csak egyenes, ferde vonalakat ismernek fel, majd a következő réteg az előzőkből összeálló magasabb szintű jellemzőket képes kinyerni. A konvolúció az információk speciális kivonása a képből. Ezt a folyamatot figyelhetjük meg a 2. ábrán. Általánosságban elmondható, hogy minél több konvolúciós lépést alkalmazunk, annál összetettebb funkciókat képes megtanulni a hálózat. Az egyes rétegek által kinyert funkciókat a későbbiekben a hálózat képes kombinálni, és felismer bonyolult formákat a kép tetszőleges részén. A CNN egyes rétegei a képnek egy sajátos reprezentációját tárolják és tanulják meg.



2. ábra: Konvolúciós neurális hálózat rétegeinek vizualizációja

Forrás: ALBAWI – MOHAMMED – AL-ZAWI 2017: 1

A gépi látásban is megjelentek a transzformer architektúrák, amit a szakirodalomban *Vision Transformer*-nek (ViT) neveznek. Ezek olyan neurális hálózatok, amelyek a transzformer architektúrát alkalmazzák képek elemzésére. A ViT-ek működése és képzése hasonló az LLM-ekhez, azzal a különbséggel, hogy itt a modell bemenetét nem szavak alkotják, hanem a feldolgozandó kép kis részletei. A bemeneti képet kis, nem átfedő téglalapokra bontják. Például ha a kép 224×224 pixeles, és egy részlet mérete 16×16 , akkor a kép 196 részletre lesz felbontva. Analógiát keresve a kép- és szövegfeldolgozás között, ami a szövegfeldolgozásnál egy nagyobb szövegrészlet, az a képfeldolgozásnál a teljes kép, a szavaknak, pontosabban a tokeneknek pedig a kis képrészletek felelnek meg.

Vizuális nagy nyelvi modellek

A vizuális nagy nyelvi modellek (VLLM) olyan transzformeralapú neurális hálózatok, amelyek képesek a képek és szövegek közötti multimodális kapcsolatokat modellezni. Ezek a modellek nagyban hozzájárultak a gépi látás és a természetesnyelv-feldolgozás fejlődéséhez, összekapcsolásához. A VLLM-ek képesek kombinálni és integrálni képi és szöveges adatokat, így képesek az egyik modalitás információinak felhasználására a másik modalitás értelmezésében. A VLLM-ek általában külön egységeket használnak a képek és a szövegek feldolgozására, majd ezeket a reprezentációkat kombinálják. A működésük során a képeket kis részletekre bontják, amelyeket beágyaznak, a szövegeket pedig tokenekre (szótöredékekre), amelyeket szintén beágyaznak. Felhasználásuk sokrétű, alkalmasak képeket értelmezni és azokról leiratokat készíteni, kérdésekre válaszolni egy adott kép alapján, képek és szövegek kombinált felhasználásával.

osztályozási döntéseket hozni. A VLLM tulajdonképpen a hagyományos nyelvi modell kiegészítése a képi modalitással.

Beágyazások

Szöveges adatok beágyazása: a modern nyelvi modellek egyik közös tulajdonsága a szóbeágyazások használata, ahol szavakat vagy csak szórészeket sokdimenziós matematikai vektorokká alakítunk át úgy, hogy a vektortérben egymáshoz közel eső szavak jelentése hasonló legyen. A szóbeágyazásnak egyik ismert megvalósítása a Word2Vec eljárás. A Word2Vec mögött meghúzódó alapötlet az, hogy a korpuszban minden szót egyetlen vektor reprezentál, amely az összes kontextus (a közeli szavak) alapján lesz kiszámítva, amelyben a szó előfordul. Ennek a megközelítésnek az intuíciója az, hogy a hasonló jelentésű szavak hasonló kontextusban fordulnak elő. A szavakat reprezentáló beágyazott vektorok nemcsak a szavak jelentését, hanem a szót körülvevő szavak kontextusának a jelentését is megragadják. Amikor dokumentumosztályozást végzünk, tulajdonképpen ezt a beágyazási technikát használjuk fel. Mondatokat vagy akár egész dokumentumokat alakítunk át sokdimenziós vektorokká, és azokat osztályozzuk. A transzformeralapú modellek erőssége, hogy az egyes tokenek (szótöredékek) belső ábrázolása, reprezentálása függ a tokent megelőző tokenektől is. A kódolt reprezentációk így szemantikai és kontextuális információt is hordoznak.

A beágyazás segítségével hatékony keresési megoldásokat lehet készíteni, hiszen a beágyazás révén nemcsak szavak, hanem mondatok, nagyobb szövegegységek közötti jelentésbeli hasonlóságot is tudunk mérni.

Képi adatok beágyazása: a képbeágyazások tulajdonképpen a képek olyan reprezentációi, amelyeket gépi tanulási modellek, például CNN-ek segítségével generálnak. Ezek a vektorok információkat tartalmaznak a képek tartalmáról, stílusáról és struktúrájáról. A CNN-ek szűrők használatával nyerik ki a képek jellemzőit, és ezeket a jellemzőket vektorokká alakítják. A ViT-ek esetében az előbbiekhöz hasonlóan a transzformer architektúrából nyerjük ki a beágyazási vektorokat. A képi beágyazásoknál is igaz, hogy hasonló tartalmú képek a vektortérben egymáshoz közel helyezkednek el. Ez teszi lehetővé például a képekkel történő keresést.

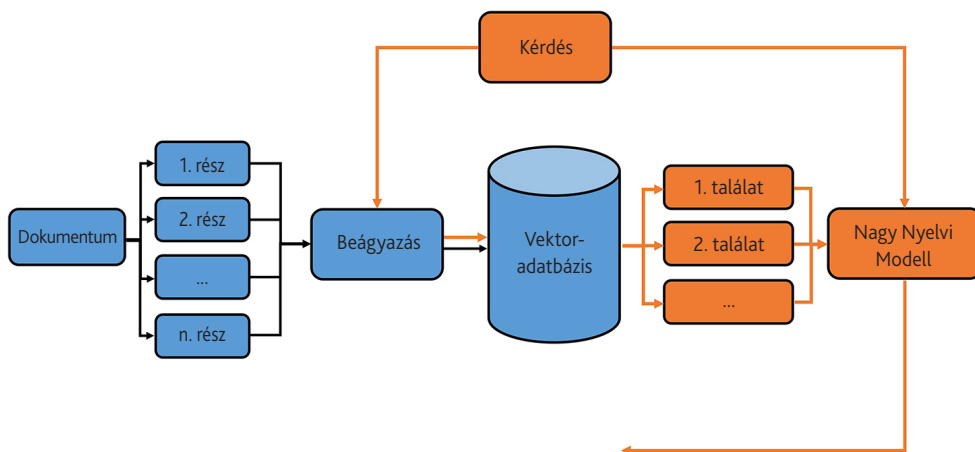
Vektoradatbázisok

A vektoradatbázisok olyan adatbázisrendszerek, amelyek képesek nagy mennyiségű vektoralakú adatot tárolni és hatékonyan lekérdezni. Ezek a rendszerek a gépi tanulás fejlődésével, a beágyazási technikák megjelenésével terjedtek el. Egyik erősségük a szemantikus keresésben rejlik, ahol a hagyományos kulcsszóalapú keresést felváltja a vektorok közötti hasonlóságon alapuló keresés. A hasonlóságot a vektorok közötti távolság (például euklideszi távolság, rebusz hasonlóság) alapján határozzuk meg. A vektorok a beágyazásuknál fogva sokkal több információt reprezentálnak, mint a kulcsszavak, ezért a keresés hatékonyabbá vált. A hagyományos keresők csak a pontos egyezéseket találják meg, míg a vektoradatbázisok a keresés során képesek

olyan elemeket is azonosítani, amelyek tartalmilag hasonlóak a keresési kifejezéshez, még akkor is, ha a pontos kulcsszavak nem szerepelnek bennük. Ez a módszertan hatalmas előrelépést jelent az olyan területeken, ahol a pontos kulcsszavak meghatározása nehézkes, például a képfelismerésben vagy a komplex szöveges adatok elemzésében. A vektoradatbázisok felhasználási területe ma már széles: képkeresés, hangalapú keresés, szemantikus keresés, témák felderítése, szövegek csoportosítása.

Adatlekérésre alapozott szövegenerálás

A hallucinációk csökkentésének egyik lehetséges megoldása olyan eljárás, amikor a keresett információt nem a nyelvi modell tanult paraméteres tudásbázisából nyerjük ki, hanem más módszerrel a prompthoz, azaz a nyelvi modell részére átadott szöveghez csatolunk információkat egy külső tudásbázisból. Ezt az eljárást nevezzük adatlekérésre alapozott szövegenerálásnak, az angol terminológiában RAG-nak. A továbbiakban mi is ezt a RAG elnevezést fogjuk használni.



3. ábra: A RAG folyamata

Forrás: a szerzők szerkesztése

A 3. ábrán láthatjuk a RAG-rendszer működési folyamatait. A feldolgozandó dokumentumokat először részekre tördeljük, egy rész jellemzően néhány száz szót tartalmaz. Az így keletkezett szövegrészeket beágyazással átalakítjuk sokdimenziós vektorokká, és eltároljuk azokat egy vektoradatbázisban. Amikor kérdést teszünk fel a RAG-rendszernek, akkor a kérdést átalakítjuk a beágyazás segítségével egy nagydimenziós vektor formájába. Ezt követően a kérdésvektorhoz hasonló vektorokat keresünk a vektoradatbázisban. Az így kapott hasonló vektorokat, amelyek egy-egy szövegrészt

reprezentálnak, visszaalakítjuk szöveggé, ezek lesznek a találatok. Amennyiben a kereső jól működik, a visszakapott találatok ténylegesen kapcsolódnak a kérdéshez. Ezeket a találatokat pedig átadjuk az eredeti kérdéssel együtt a nyelvi modellnek, amely választ generál a kérdésünkre. Egy RAG segítségével felépített rendszer esetén az LLM-nek azt a tudását használjuk elsősorban, amely a promptból képes információkat feldolgozni. Erre lehet egy példa, amikor a Wikipédiát mint külső tudásbázist használjuk például Zrínyi Miklóssal kapcsolatos kérdés megválaszolására. Első lépésben a Wikipédiát le kell tölteni, fel kell darabolni, és a szövegrészeket be kell ágyazni, majd eltárolni egy vektoradatbázisban. A RAG-rendszer a kérdésből („Mikor született Zrínyi Miklós?”) megkeresi a számára elérhető adatokat Zrínyi Miklósról a vektoradatbázisból, majd ezeket az információkat becsomagolja a promptba. A nyelvi modellnek pedig a kérdésből és a Zrínyi Miklósról átadott információkból kell választ generálnia.³

Egy gyakorlati alkalmazás bemutatása hadtudományi témában

LLM-ek a védelmi szférában

A műveletek támogatása kapcsán paradigmaváltás zajlik, hiszen a mesterséges intelligencia olyan adatokon alapuló döntés-előkészítési lehetőségekkel bír, amelyek korábban elképzelhetetlenek voltak. A védelmi szférában nagyon sok írott szöveges formában elérhető dokumentum keletkezik. Ezek a dokumentumok rendkívül jelentős tudást és rengeteg értékes információt reprezentálnak. A dokumentumok jelentős része a védelmi és biztonsági megfontolások alapján nyilvánosan nem elérhető. Ezeknek az adatoknak a feldolgozása, tudásreprezentációk kialakítása kizárólag teljesen ellenőrzött körülmények között lehetséges, azaz nem engedélyezhető, hogy az interneten elérhető – például Chat GPT, Gemini vagy más hasonló – rendszerek részére ezekből az adatokból bármi is átadható legyen. Amennyiben ezt a megszorítást elfogadjuk, úgy kénytelenek vagyunk saját, a védelmi szféra egyes résztvevői számára külön-külön megoldásokat létrehozni.

Gyakorlati alkalmazás

Képzeljünk el egy olyan szituációt, ahol nagy mennyiségű szöveges, képi és akár hang-információt, például rádióüzeneteket, műholdfelvételeket, drónfelvételeket és jelentéseket kell feldolgozni. Ezek az adatok különböző forrásokból származnak, különböző formátumban érkeznek, és gyors feldolgozásuk kritikus a döntéshozatali folyamatban.

A feldolgozás lépései: Adatfeldolgozás során a különböző formátumú adatokat (szöveg, kép) egységes formátumba kell konvertálni. Az adatokból vektoros reprezentációkat kell létrehozni és tárolni egy nagy teljesítményű vektoradatbázisban. Ez lehetővé teszi a gyors és hatékony keresést a hasonlóságon alapuló lekérdezések segítségével.

³ RAM 2023.

RAG-alkalmazási példa lehet, ahol az értékelő-elemző egy komplex kérdést tesz fel: „Milyen katonai tevékenységet észleltek az elmúlt 24 órában X település környékén, figyelembe véve a rendelkezésre álló műhold- és drónfelvételeket, valamint a lehallgatott rádióüzeneteket?” A RAG-rendszer a kérdést vektorizálja, és a vektoradatbázisban megkeresi a legrelevánsabb szöveges és képi információkat. Ezeket az információkat összegyűjti, és egy LLM-nek adja át, amely egy koherens és informatív jelentést generál az elemző számára. A RAG-rendszer minimalizálja a hallucináció kockázatát, mivel a válasz közvetlenül az adatokon alapul.

LLM-ek a hadtudományban

A hadtudományban nagyon jelentős szerepe van az írásbeliségnek, hiszen a hadtudományhoz szorosan köthető a dokumentálás mint az ismeretek felhalmozásának egyik formája. Ez a jelentős írásbeli tudásanyag, amennyiben digitalizált formában is elérhető, kiváló alapot adhat egy nyelvi modell finomhangolásához egy adott szakterületen.

Az MI katonai alkalmazása számos etikai kérdést is felvet. Kiemelt figyelmet kell fordítanunk az emberi kontroll biztosítására, a dehumanizáció elkerülésére és az igazságos háború elmélete elveinek betartására. Az MI alkalmazásának átláthatónak és indokoltnak kell lennie, és az MI-t irányító embereknek kell felelősséget vállalniuk az MI által végrehajtott műveletekért. A kockázatkerülő háború fogalmi rendszere is összefügg az MI-vel, hiszen a technológia egyik célja a katonákra leselkedő kockázatok csökkentése, a veszteségek minimalizálása.⁴

A rendelkezésünkre álló hadtudománnyal kapcsolatos szövegek segítségével a nyelvi modell szert tehet olyan ismeretekre, amelyek erre a tudásbázisra épülnek. Egy kellően nagy nyelvi modellben⁵ már kialakultak bizonyos érvelési képességek, amelyek a tanult tudásbázis alapján képesek lehetnek koherens és következetes válaszokat adni bizonyos kérdésekre. Az ilyen modellek nagyon gyorsan képesek releváns információk összegyűjtésére és azokból stratégiai, taktikai döntéseket előkészítő anyagok létrehozására.

A védelmi szektorban gyakran előfordulnak olyan „Szigorúan titkos”, „Titkos”, „Bizalmas” vagy „Korlátozott terjesztésű” dokumentumok, amelyeknek a tudásanyagára támaszkodnunk kell, azonban nem lehetséges azokat vagy azoknak valamely részhalmazát egy külső fél tudomására hozni. Gondolhatunk stratégiai tervdokumentációkra, védelmi folyamatleírásokra stb. Ezekben az esetekben teljesen zárt és hozzáférés-kontrollált rendszerekben kell gondolkodnunk.

A fentiek alapján az MI és az LLM-ek nem csupán a hadtudomány elméleti megközelítését változtatják meg, hanem a gyakorlati alkalmazásokban is forradalmi újításokat hoznak.

⁴ BODA 2024.

⁵ WEI 2022.

Zárt információkereső rendszer

A védelmi szférában is működőképes rendszer kialakítása megköveteli, hogy akár hálózati kapcsolat nélkül is működő okos információkereső rendszert alakítsunk ki. Az információk visszakeresése túlmutat egy egyszerű keresőmotoron, amely a kérdésre csak találatokat ad. A mai technológiai lehetőségek lehetővé teszik, hogy a kérdésekre válaszokat kapjunk, ami felgyorsítja a munkafolyamatainkat azáltal, hogy a keresést és az eredmények szintetizálását egy program elvégzi helyettünk.

Egy zárt információs rendszert készítettünk el teljesen nyílt forráskódú szoftverekkel, és a *Hadtudomány* folyóirat 29–34. évfolyamok közötti számaiból 234 cikket használtunk fel a vektoros adatbázis készítéséhez. A cél az volt, hogy modellezzük nagy mennyiségű tudásanyag feldolgozását és tárolását egy zárt környezetben, majd abból kérdések alapján releváns információkat nyerjünk ki RAG segítségével, és válaszokat készítsünk nyelvi modellek felhasználásával.

A rendszert a GitHub⁶ kódtárban is elérhető Open-WebUI⁷ keretrendszerre építettük fel. A szövegek beágyazására használt modult magyar nyelvű szövegeken finomhangoltuk,⁸ generálásukhoz pedig a Llama 3.1 8B⁹ utasításkövető modellt használtuk. A teljes rendszert Ubuntu Linux¹⁰ operációs rendszerre telepítettük. A megfelelő sebességű működéshez szükséges egy GPU¹¹ is, Nvidia RTX 3090-est használtunk a projekthez.

Az így létrejött rendszer az internettől teljesen függetlenül is képes működni és releváns válaszokat adni teljesen specifikus kérdésekben is, amelyekre a válaszadáshoz szükséges tudásanyag csak helyi gépen található meg. A válaszadáshoz semmilyen információ nem kerülhet ki, hiszen a választ generáló nyelvi modell is a helyi gépen fut.

A 4. ábrán láthatjuk a keresőrendszer felületét, ahol a keresőbe beírt kérdést, utasítást válaszolja meg, hajtja végre a rendszer. A választ a RAG rendszerbe feltöltött dokumentumok alapján készíti el, megadva a válaszhoz felhasznált dokumentumokat.

A kérdés a következő volt: „Milyen fejlesztési lehetőségek vannak a virtuális valóság katonai rendszerekben történő alkalmazására?”

Válasz: „Az alkalmazott virtuális valóság (Virtual Reality, VR) jelentősen fejleszthető és alkalmazható a katonai rendszereken belül. Azokban az eszközökben, amelyekben a jelenlegi technológia már meglehetősen fejlettségre van, a VR-technológia továbbfejleszthető és kiegészítheti a rendszereket.

⁶ Lásd: <https://github.com>

⁷ Lásd: <https://github.com/open-webui/open-webui>

⁸ Lásd: <https://huggingface.co/karsar/paraphrase-multilingual-MiniLM-L12-hu-v2>

⁹ Lásd: <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

¹⁰ Lásd: <https://ubuntu.com/>

¹¹ Graphics Processing Unit.



4. ábra: Zárt információkereső rendszer

Forrás: a szerzők felvétele

A virtuális valóság alkalmazása a katonai rendszerekben számos előnnyel jár. Ezek közé tartozik, hogy:

- Az ejtőernyős kiképzésben használt VR-technológia jelentősen csökkenti a tényleges ejtőernyős ugrás kockázatait.
- A szimuláció javítja az ejtőerősök képzettségi szintjét, és platformot biztosít számukra az új ejtőerő típusok, ismeretlen környezetek és új képzési tárgyak megtapasztalásához.
- A VR-technológia lehetővé teszi a csapatok számára, hogy észleljék a különböző légi vészhelyzeteket, ami csökkenti a tényleges ejtőernyős ugrás kockázatait.
- A VR-technológia lehetővé teszi a szimulációkat, amikor a katonai egységek a valóságos körülmények között gyakorolnak.

Ezek mellett a virtuális valóság alkalmazása a katonai rendszerekben továbbfejleszthető és kiegészítheti a rendszereket, például:

- A digitális katona program, amely a honvédek egyéni felszerelésének átfogó modernizálását célozza.
- A robototechnológia, amely a robotoknak olyan képességeket ad, hogy meg tudják oldani bizonyos feladatokat, például a szimulációban.
- Az autonóm csapásmérő eszközök, amelyek olyan képességekkel rendelkeznek, hogy megszerezhetik céljaikat a valóságos körülmények között.

A virtuális valóság alkalmazása a katonai rendszerekben számos lehetőség és előnyt tartalmaz. Azonban, ahhoz, hogy ezt a fejlesztését megvalósítsuk, szüksége van olyan erőforrásokra, amelyekkel a rendszereket kiegészítik és fejlesztik tovább."

A válasz elkészítéséhez a következő forrásanyagokat használta fel a rendszer:

1. forrás: 12266-Cikk szövege-52251-1-10-20230804.pdf

Relevancia: 66.18%(0.6618)

Tartalom: „HADMŰVÉSZET Négyesi Imre/c71 A virtuális valóság technológia alkalmazásának lehetőségei a katonai készségfejlesztésben* DOI 10.17047/HADTUD.2023.33.2.3 ..."

2. forrás: 9734-Cikk szövege-44567-1-10-20230207.pdf(oldal 6)

Relevancia: 56.52%(0.5652)

Tartalom: „–robothadviselés és robottechnológia; –digitális katona program; –autonóm csapásmérő eszközök; –kvantum számítástechnika; –szimulációs, virtuális és kiterjesztett valóság. ..."

3. forrás: 6780-Cikk szövege-28285-1-10-20210713.pdf(oldal 16)

Relevancia: 56.12%(0.5612)

Tartalom: „A Digitális Katona Program a honvédek egyéni felszerelésének átfogó modernizálását célozza. Kiemelkedően fontos eleme a programnak, hogy ahol kivitelezhető, ott a rendszerelemek kutatás -fejlesztése és gyártása a hazai védelmi ipar bevonás ával valósuljon meg. ..."

A források tartalmi elemei nem teljes körűek a fenti idézetekben.

A kérdésre adott válasz nem tökéletes, csak egy résszel foglalkozik, illetve számos helyesírási hibát is tartalmaz. A válasz nagyban függ a rendszer beállításaitól, hogy hány forráscikket használjon, milyen hosszú legyen a válasz, illetve fontos paraméterek a vektros adatbázis létrehozásánál a kontextusméret, átfedés. A bemutatott kérdés-válasz párhoz tartozó válasz további beállításokkal finomítható, változtatható.

Összefoglalás, következtetések

Összefoglalás

Ez a tanulmány a mesterséges intelligencia, különösen az LLM-ek és a gépilátás-rendszerek alkalmazásának lehetőségeit és kihívásait vizsgálta a védelmi szférában. Részletesen bemutattuk a nyelvi modellek működését, képzési folyamatait, valamint a generatív modellek képességeit és korlátait, mint például a hallucináció jelensége. A képfeldolgozás terén a legkorszerűbb modellek, valamint a VLLM-ek lehetőségeit elemeztük. Kitértünk a vektorbeágyazások és vektoradatbázisok jelentőségére a hatékony információkeresés és feldolgozás szempontjából, valamint az adatlekérésre alapozott szöveggenerálás technikájának szerepére a hallucinációk csökkentésében. Tanulmányunk rámutatott az LLM-ek képzésének jelentős költségigényére és a jelenlegi magyarországi

lehetőségekre. Az LLM-ek alkalmazása nemcsak az általános szöveges feladatokban, hanem a speciális területeken, például a hadtudományban is komoly potenciált rejt. Konklúzióként kijelenthetjük, hogy az LLM-ek és a gépilátás-rendszerek alkalmazása paradigmaváltást hozhat a katonai műveletek támogatásában, azonban a bizalmas adatok védelme érdekében elengedhetetlen egy saját, biztonságos, zárt informatikai infrastruktúra kiépítése, a megfelelő szakértelem fejlesztése és a folyamatos innováció. A védelmi szektor számára tehát egy hosszú távú, jelentős beruházást igénylő projekt megvalósítása válik szükségessé a biztonság érdekében.

Következtetések

A műveletek támogatása kapcsán paradigmaváltás zajlik, hiszen a mesterséges intelligencia olyan adatokon alapuló döntés-előkészítési lehetőségekkel bír, amelyek korábban elképzelhetetlenek voltak. A védelmi szférában nagyon sok írott szöveges és képi formában elérhető dokumentum keletkezik. Ezek a dokumentumok jelentős tudást és rengeteg értékes információt reprezentálnak. A dokumentumok jelentős része a védelmi és biztonsági megfontolások alapján nyilvánosan nem elérhető. Ezeknek az adatoknak a feldolgozása, tudásreprezentációk kialakítása kizárólag teljesen ellenőrzött körülmények között lehetséges, azaz nem engedélyezhető, hogy az interneten elérhető rendszerek (például Chat GPT, Gemini stb.) ezekből az adatokból bármihez is hozzáférjenek. Amennyiben ezt a megszorítást elfogadjuk, úgy kénytelenek vagyunk saját, a védelmi szféra egyes résztvevői számára külön-külön megoldásokat létrehozni.

Véleményünk szerint el kell kezdenünk kiépíteni azokat a saját védelmi informatikai infrastruktúrákat, amelyek képesek feldolgozni és kezelni a védelmi szektorban található adatokat. Ez a kiépítés több szempontból is komplex feladat. Első lépésben egy olyan biztonságos, zárt környezetet kell létrehoznunk, amely megfelel a legmagasabb szintű adatvédelmi és biztonsági előírásoknak. Ez magában foglalja a fizikai biztonságot, a hálózati biztonságot, valamint az adat titkosítását és hozzáférés-szabályozását. A rendszernek skálázhatónak kell lennie, hogy képes legyen kezelni a folyamatosan növekvő adatmennyiséget, és rugalmasnak, hogy alkalmazkodni tudjon a változó igényekhez. A technológiai infrastruktúra mellett elengedhetetlen a megfelelő szakértelem biztosítása is. Képzett adattudósokra, biztonsági szakértőkre és rendszergazdákra van szükség, akik képesek a rendszer üzemeltetésére, karbantartására és folyamatos fejlesztésére. Ezenfelül a különböző védelmi szervek közötti együttműködés és adatmegosztás is kritikus fontosságú, ami standardizált adatformátumok és protokollok bevezetését követeli meg. A különböző rendszerek integrációja is jelentős kihívás lehet, és gondos tervezést igényel. Hosszú távon pedig a folyamatos fejlesztés és innováció biztosítása a legfontosabb. A mesterséges intelligencia gyors fejlődése miatt folyamatosan figyelemmel kell kísérni az új technológiákat, és be kell építeni azokat a rendszerbe, amennyiben növelik a hatékonyságot és a biztonságot. Ez a folyamatos fejlesztés azonban csak megfelelő finanszírozás és szakmai támogatás mellett valósulhat meg. A védelmi szféra számára tehát egy hosszú távú, jelentős beruházást igénylő projekt elindítása válik szükségessé, amelynek sikeressége a nemzeti biztonság szempontjából is rendkívül fontos.

Felhasznált irodalom

- ALBAWI, Saad – MOHAMMED, Tareq Abed – AL-ZAWI, Saad (2017): Understanding of a Convolutional Neural Network. *2017 International Conference on Engineering and Technology (ICET)*, Antalya, Turkey, 1–6. Online: <https://doi.org/10.1109/ICEngTechnol.2017.8308186>
- BIRHANE, Abeba et al. (2021): Science in the Age of Large Language Models. *Nature Reviews Physics*, 5, 277–280. Online: <https://doi.org/10.1038/s42254-023-00581-4>
- BODA Mihály (2024): A kockázatkerülő háború és a bátorság a 20–21. század fordulóján. *Honvédségi Szemle*, 152(3), 113–125. Online: <https://doi.org/10.35926/HSZ.2024.3.9>
- BROWN, Tom B. et al. (2020): *Language Models are Few-Shot Learners*. Online: <https://doi.org/10.48550/arXiv.2005.14165>
- DEVLIN, Jacob et al. (2018): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Online: <https://arxiv.org/abs/1810.04805>
- FILIPPOVA, Katja (2020): Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data. In COHN, Trevor – HE, Yulan – LIU, Yang (szerk.): *Findings of the Association for Computational Linguistics EMNLP 2020*. [H. n.]: Association for Computational Linguistics, 864–870. Online: <https://doi.org/10.18653/v1/2020.findings-emnlp.76>
- HUANG, Lei et al. (2025): A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Office Information Systems*, 43(2), 1–55. Online: <https://doi.org/10.1145/3703155>
- JURAFSKY, Daniel – MARTIN, James H. (2023): N-gram Language Models. In JURAFSKY, Daniel – MARTIN, James H.: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Third Edition draft. [H. n.]: [k. n.], 31–57. Online: <https://web.stanford.edu/~jurfafsky/slp3/3.pdf>
- KUKREJA, Sanjaj et al. (2023): Vector Databases and Vector Embeddings-Review. In *2023 International Workshop on Artificial Intelligence and Image Processing (IWAIP)*. Online: <https://doi.org/10.1109/IWAIP58158.2023.10462847>
- LEWIS, Patrick (2020): Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle H. et al. (szerk): *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. Online: <https://arxiv.org/pdf/2005.11401>
- LUCCIONI, Aleksandra Sasha – VIGUIER, Sylvain – LIGOZAT, Anne-Laure (2023): Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *Journal of Machine Learning Research*, 24, 1–15. Online: <https://doi.org/10.48550/arXiv.2211.02001>
- MIKOLOV, Thomas et al. (2013): *Efficient Estimation of Word Representations in Vector Space*. Online: <https://arxiv.org/pdf/1301.3781>
- PAN, James Jie et al. (2024): Vector Database Management Techniques and Systems. In *SIGMOD '24: Companion of the 2024 International Conference on Management of Data*. New York: Association for Computing Machinery, 597–604. Online: <https://doi.org/10.1145/3626246.3654691>

- RAM, Ori et al. (2023): *In-Context Retrieval-Augmented Language Models*. Online: https://doi.org/10.1162/tacl_a_00605
- STIENNON, Nisan et al. (2020): *Learning to Summarize From Human Feedback*. Online: <https://arxiv.org/pdf/2009.01325>
- TÓTH Bálint Pál (2016): Beszélő számítógépek mély gondolatokkal. *Neurális hálózatok. Élet és Tudomány*, 71(30), 944–946.
- TOUVRON, Hugo et al. (2023): *LLaMA: Open and Efficient Foundation Language Models*. Online: <https://doi.org/10.48550/arXiv.2302.13971>
- VASWANI, Ashish et al. (2017): Attention Is All You Need. *Advances in Neural Information Processing Systems*, 5998–6008. Online: <https://arxiv.org/pdf/1706.03762.pdf>
- WEI, Jason et al. (2022): *Emergent Abilities of Large Language Models*. Online: <https://doi.org/10.48550/arxiv.2206.07682>
- YANG, Min et al. (2021): DOLG: Single-Stage Image Retrieval With Deep Orthogonal Fusion of Local and Global Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11772–11781. Online: <https://doi.org/10.1109/ICCV48922.2021.01156>
- YANG ZIJIAN Győző et al. (2023): Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia*, 247–262. Online: https://acta.bibl.u-szeged.hu/78417/1/msznykonf_019_247-262..pdf
- ZHANG, Shengyu et al. (2023): *Instruction Tuning for Large Language Models: A Survey*. Online: <https://doi.org/10.48550/arXiv.2308.10792>
- ZHOU, Chunting et al. (2023): *LIMA: Less Is More for Alignment*. Online: <https://doi.org/10.48550/arXiv.2305.11206>