Alacsony paraméterszámú nyelvi modellek hatékony finomhangolása osztályozási feladatra

Efficient fine-tuning evaluation of low-parameter number language models for classification tasks

Barcza Bende[1]***, Magyar Gergely***[2]***, Szántó Mátyás***[1]

[1]Budapesti Műszaki és Gazdaságtudományi Egyetem, Villamosmérnöki Kar, Irányítástechnika és Informatika Tanszék, 1111 Budapest, Műegyetem rkp. 3.
e-mail: barczabende@edu.bme.hu
[2]Budapesti Műszaki és Gazdaságtudományi Egyetem, Gépészmérnöki Kar, Gyártástudomány és -technológia Tanszék, 1111 Budapest, Műegyetem rkp. 3.

Abstract

Language models have opened up a new era in artificial intelligence, with initial recurrent neural networks capable of generating even longer continuous texts, but the biggest breakthrough was the transformers architectures, which outperformed previous solutions. A major use of language models is in classification tasks, where the models classify a sentence or sequence into groups according to various criteria based on its content. An advantage of language models using a transformer architecture is that it is possible to activate and deactivate the weights of each layer of the network and their respective optimizations. The aim of my research is to determine to what degree the classification ability of the selected model is reduced when the different layers of the model are activated or deactivated (hereafter referred to as frozen). The analyzed dataset is classifying the given text into 6 different groups. The training is evaluated by testing the training accuracy and error, and then the trained models are evaluated by calculating the area under the Receiver Operating Curve (ROC). Using the results, we can evaluate how freezing configurations change the classification capabilities of the model, how well they meet the requirements of the intended usecase. Such classification algorithms can be used in the engineering field, for example, to aid quality control, where composition can be used to infer the quality properties of a product, or to classify materials in a recycling environment, based only on the names of the components.

Keywords: language models, sequence classification, product classification

Kivonat

A nyelvi modellek új korszakot nyitottak a mesterséges intelligenciában, a kezdetleges rekurrens neurálishálók már akár hosszabb egybefüggő szövegeket is képesek voltak generálni, azonban a legnagyobb áttörést a transformer architektúrák jelentették, amelyek felülmúlták az előző megoldásokat. A nyelvi modellek egyik jelentős felhasználási területe az osztályozási feladatok, amelyek során a modellek különböző szempontok szerint egy-egy mondatot vagy szekvenciát sorolnak csoportba azok tartalma alapján. A transformer architektúrát alkalmazó nyelvi modellek egyik előnye, hogy lehetséges a háló egyes rétegeinek súlyait, illetve azok optimalizálását aktiválni és deaktiválni. Kutatásom célja meghatározni, milyen mértékben csökken a kiválasztott modell osztályozási képessége a modell különböző rétegeinek aktiválása vagy deaktiválása (továbbiakban fagyasztása) esetén. A vizsgált adathalmaz hat különböző csoportba osztva osztályozza az adott szöveget. A tanítás kiértékelését a tanítási pontosság és hiba vizsgálatával, majd a betanított modelleket a Receiver Operating Curve (ROC) alatti terület számításával végeztük. Az eredmények felhasználásával következtetéseket tehetünk, hogy milyen fagyasztási konfigurációk hogyan változtatják a modell osztályozási képességeit, azok mennyire felelnek meg a felhasználói elvárásoknak. Az ilyen osztályozó algoritmusok a műszaki területen például minőségellenőrzést segíthetik, ahol összetétel alapján lehet következtetni a termék minőségi tulajdonságaira, vagy az újrahasznosításra szánt termékeket elnevezésük alapján képes anyagcsoportba osztani.

Kulcsszavak: nyelvi modellek, szöveg osztályozás, termék csoportosítás

1. BEVEZETÉS

A transformer-alapú nyelvi modellek a mesterséges intelligencia újonnan megjelent és rohamos sebességgel elterjedt eszközei. Nagy hatékonyságuknak köszönhetően jelenlegi térnyerésük széles körű felhasználási területeket hódít meg[1]. Kezdetekben olyan humán területeken jelentek meg, amelyek nagy monotonitást igényeltek és automatizálásukkal anyagi előnyökhöz lehetett jutni, így például a fordítás[2] vagy ügyfélszolgálati irányadás[3]. Az online elérhető fordítók állandó rendelkezésre állásukkal és nagy pontosságukkal számos segítséget képesek nyújtani mindennapi életünkben. A beszélgetőrobotok alkalmazása ügyfélszolgálati dolgozó mellett/helyett pedig sokat képesek gyorsítani a megfelelő ügyek intézésében[4].

Kutatásunk során vizsgált módszer elnevezése a szövegosztályozás, melynek egy, mindennapokban gyakori felhasználási területe a kártékony üzenetek kiszűrése. Amikor érkezik egy e-mail a fogadóhoz, akkor annak tartalma átvizsgálásra kerül automata rendszerek segítségével, melyek kártékony csoportosítási feladatra lettek finomhangolva. Termékek elnevezés-alapú csoportosítása szintén rendelkezik irodalommal, így Gholamian és tsai.[5] cikkükben olyan termékleírásokat csoportosító robusztus nyelvi modell finomhangolási módszert alkalmaztak, ahol jogi regulációkra és megfeleléseknek való elégtételt vizsgáltak. Üzleti indítékuk a Vásárlási Világszövetség (World Costums Organization) által közzétett adatokra alapozott, mely szerint 1,3 milliárd termék került átvizsgálásra 2022-2023-as években, ami nagy emberi és szakértői erőforrást igényelt[6]. Továbbá a termékek megfelelő vizsgálatához szakképesítés is kellett, melynek megszerzése újabb hónapokat jelentett. Egy nyelvi modell tanítása ehhez képest jelentősen kisebb erőforrást igényel, és ha nem is képes teljes mértékben lecserélni az emberi erőforrást, azt nagyban segítheti mindennapi munkájában.

A osztályozási megoldások alkalmazása a műszaki területen kevésbé gyakori, mivel használatuk nem mindig intuitív, ugyanakkor bizonyos speciális részterületeken való alkalmazásuk jelentős előnyökkel járhat. Azon termékek, melyek funkcionalitása anyagi összetételüktől függ, egy megfelelően finomhangolt nyelvi modell a rendelkezésre álló összetételi adatok alapján képes lehet eldönteni, hogy milyen minőségi osztályba tartozzon. Ehhez hasonló osztályozási adatokra finomhangolva a kiválasztott nyelvi modellt, elemezzük, hogy mely architekturális elemek a legjelentősebbek az osztályozási feladatok pontos elvégzésében. Kutatásunk célja, hogy a modell tanító paramétereinek csökkentése mellett, hogyan változnak a modell teljesítményére vonatkozó mérőszámok, melyek a tanítási pontosság és hiba, valamint az AUROC (Area Under Recieving Operator Curve - vevői működési jelleggörbe alatti terület) módszerekkel megállapítani, hogy a fagyasztásos módszerrel csökkentett tanítandó paraméterek esetében hogyan változik a finomhangolt modellek kiértékelési metrikái.

2. ADATOK ÉS MÓDSZEREK

A nyelvi modellekhez kapcsolódó áttörő újítást az úgynevezett transformer architektúra hozta el, mely a figyelemmechanizmust tartalmazó rétegek jelentős javulást tudott elérni az emberi fogalmazás gépi imitálásában[7]. Az erre építő modellek a későbbiekben számtalan optimalizáláson estek túl, így több különböző felépítésű és eltérő paraméterszámú modell elérhető. A kutatásunkhoz nyílt-forráskódú modellt használtunk, melyet a Wolf[8] által leírt weboldalon lehet elérni.

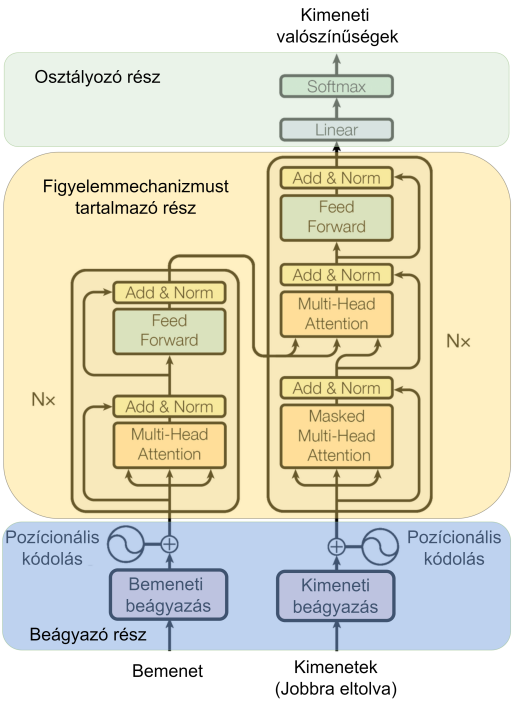
2.1. Nyelvi modell választás

Az általunk kiválasztott modell, a Muennighoff és tsai.[9] által közzétett ranglistán a 100 millió alatti paraméterszámú modellek közül a legjobb 3 közül való. A ranglistán a modelleket különböző validációs adatbázis adatain kiértékelt képességei alapján teszik sorrendbe. Az általunk választott Xiao és tsai.[10] úgynevezett BGE modell 33 millió paraméterrel rendelkezik, és Devlin és tsai.[11] által bemutatott, úgynevezett BERT típusú figyelmi mechanizmusú rétegek jellemzik. A 1 ábrán látható a transformer architektúra, mely 3 fő részre bontható:

1. beágyazó réteg, mely a tokenizált bejövő adatokon végez transzformációkat,

2. figyelmi mechanizmussal rendelkező rétegek, melyek az architektúra jellemző elemei,

3. többosztályú osztályozó réteg, mely softmax transzformációval a megfelelő mennyiségű kimenetekhez ad valószínűségeket, hogy melyik címke jellemzi a legjobban az adott bemenetet.



1. ábra. *A transformer architektúra Vaswani és tsai.[7] cikkéből fordítva, és kiemelve az általunk kiválasztott felosztást.*

2.2. Vizsgálati adathalmaz

A kiválasztott adathalmazunk He és tsai.[12] cikkben leírt internetes hirdetéseket csoportosít azok leírása alapján 6 különböző osztályba[13]. A bemenet egy, általában pár mondatból álló leírás, vagy vevő-eladó beszélgetés, ezeket a modell a következő 6 db címke egyikébe osztja: bicikli, háztartási termék, bútor, autó, elektronikai termék, telefon. A tanítás során a bélyegek kontextusa nem ismert, azok 0-5-ig tartó számokkal vannak reprezentálva. Az adatkészlet több mint 6000 db címkézett adatot tartalmaz, kiegészítve egy 800 darabos validációs adatkészlettel, melyet tanítás alatt nem lát a modell.

2.3. Vizsgálati környezet

A modellek finomhangolását az alábbi számítógépes architektúrán futtattuk:

- processzor: Intel Core i7,
- videokártya: Nvidia GeForce GTX 1650Ti, 4GB dedikált memóriával,
- számítógép RAM-jának mérete: 16GB.

2.4. Értékelési módszerek

Az AUROC elemzése eredetileg bináris osztályozás értékelésére használatos és azoknál mutat jól értelmezhető metrikákat [14]. A görbe alatti érték kiszámításával lehet következtetéseket levonni az osztályozás sikerességének számszerűsítésére. Az általunk vizsgált eset, egy többosztályos klasszifikáció, így a szokványos ROC AUC metrikák nem relevánsak, azok kiterjesztését Hand és Till[14] végezte el, melynek egy könnyedén a programba illeszthető implementációját a *Huggingface* [15] oldal tartalmazza.

Az OvO (One-versus-One - egy-az-egyhez) megközelítés a többosztályos klasszifikációs problémát az egyes osztályok páronkénti másik osztállyal szembeni elemzésre osztja fel, így az egyes párok közötti klasszifikációt vizsgálja. Ezeknek az alapklasszifikálóknak a kimenetei kerülnek kombinálásra a kimenet meghatározására. Az OvR (One-versus-Rest - egy-a-többihez) vizsgáló módszer felosztja a problémát úgy, hogy minden predikcióra egy osztálynak a többi osztállyal szembeni predikcióját vizsgálja [14]. Az OvO és OvR vizsgálatok összegzésének további 2 lehetőség van: *mikro* és *makro* összegzés. A makro összegzés a külön kiszámított AUROC értékeket egymástól függetlenül összegzi az egyes kategóriákra, és ezt követően átlagolja az értékeket. Előnye, hogy kevésbé érzékeny az adatok kiegyensúlyozatlanságára a mikrohoz képest. Esetünkben egy kiegyensúlyozott adathalmazon végeztük a tanítást, és kiértékelést, így ez nem lesz mérvadó a döntésben, amelyik esetben a legnagyobb eltérések láthatóak, azt az elemzési módszert választjuk.

3. EREDMÉNYEK ÉS KIÉRTÉKELÉS

A kutatás céljaként kitűztük, hogy megvizsgáljuk, egy osztályozási feladatnál az egyes rétegek milyen hatással vannak a kimeneti eredményekre. Ha megtaláljuk azt az elemet, mely a legnagyobb hatással lesz a kimenetre, akkor jelentősen csökkenthető a tanítandó paraméterszám a finomhangolás során. Ehhez meghatároztunk különböző fagyasztási kombinációkat, melyek az 1 táblázaton láthatóak. A paraméter arány az eredeti 33 millió tanítható paraméterhez képesti százalékos arányszámot jelenti.

Konfigurációk összehasonlítása 1. táblázat

Konfiguráció	**Paraméter arány [%]**	**Rövidítés**
Teljes architektura	100,00	TOTAL
Csak osztályozó	<0,01	C
Csak beágyazó	35,72	E
Csak egy figyelemmechanizmus	5,32	A11
Teljes figyelemmechanizmus és osztályozó	64,27	AC
Beágyazó és osztályozó	35,72	EC

3.1. Tanítási hiba és pontosság kiértékelése

A 2 ábrák mutatják a finomhangolások eredményeit az egyes konfigurációk kombinációira nézve. A 2a) ábrán látható, hogy egyes görbék nem is hagyták el a kezdeti tanítási hibát, ebben az esetben a finomhangolás sem tudta pontosítani az osztályozási képességeit a C és E konfigurációnak. A 2b) ábrán látható, hogy ezek a konfigurációk jelentősen alacsony pontosságot értek el a többihez képest. Ezzel szemben, amikor minden más paramétert befagyasztottunk, csak a 12 darab figyelemmechanizmus közül az utolsót kiválasztva futtattuk a finomhangolást, már látható egy enyhe tanítási hiba csökkenés a 2a) kék görbéjén. A hozzá tartozó kimeneti pontosság már 90% feletti, így mondhatjuk, hogy csupán ennek az egy elemnek a tanítása már jelentősen megemeli az osztályozás pontosságát, pedig annak paraméterszáma a teljes modellnek csupán 5%-át teszi ki. Csupán a beágyazó és osztályozó rétegek finomhangolása esetén a tanítási hiba határozottan csökken, azonban a paraméterek száma ebben az esetben közel hétszeresére növekedett. A kisebb tanítási hiba magabiztosabb osztályozást jelent, mivel a hiba a nem megfelelő osztályokra adott predikcióból kerül számításra, így elmondható, hogyha a hiba értéke csökken, a modell magabiztosabban tudja kiválasztani a helyes osztályt.

A betanított nyelvi modelleket konfigurációkat összegyűjtve, azokon egymást követően lefuttattam az ROC görbe felrajzolásához szükséges predikciókat. A modellek a validációs adatkészletből választva kellett, hogy prediktálják a választott kimeneti címkét, az ROC alatti görbe pedig megadta, hogy milyen pontosságú volt ez az osztályozás.

A 2c) ábrán látható, hogy az egyes konfigurációk milyen magabiztossággal tudták csoportosítani a validációs adatkészletet. A figyelemmechanizmus nélküli esetek láthatóan alulmaradtak a többihez képest, így kijelenthetjük, hogy ez az eset is alátámasztja a figyelemmechanizmus robusztusságát.

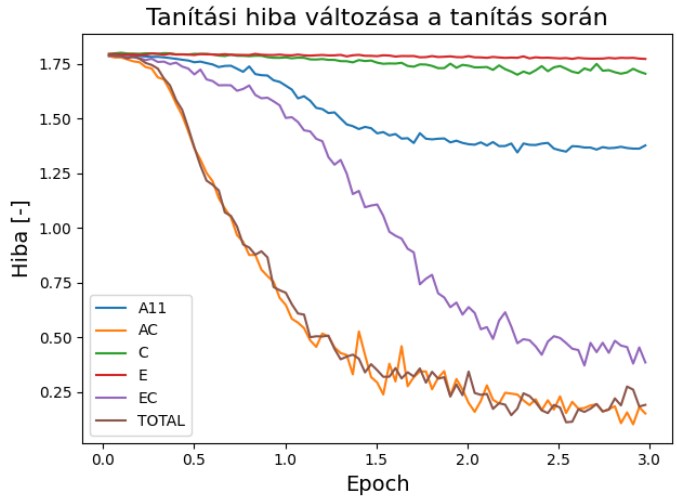
4. ÖSSZEFOGLALÁS

A fenti elemzésekből következtethetünk, hogy osztályozási feladatokra használt alacsonyabb paraméterszámú nyelvi modelleknek melyek azok a rétegei, amelyek a legnagyobb hatással vannak a nyelvi szekvenciák címkézési feladatainak ellátására. Az általunk használt modell jelentősen kevesebb paraméterszám esetén is hasonló magabiztossággal volt képes számára új, validációs adatkészletet csoportosítani. A fenti eredmények ismeretében kijelenthető, hogy a nyelvi modellek alkalmazását osztályozási feladatra hatékonyabb finomhangolási módszerrel tudjuk beépíteni egy új alkalmazási környezetbe. A redukált paraméterszámú finomhangolásnak köszönhetően nagyobb modellek is elérhetőbbé vállnak, mivel kisebb számítási kapacitás is elegendő a kisebb paraméterszám finomhangolásához. Nagyobb modellek alkalmazásával pedig szélesebb körű feladatok robusztusabb elvégzése lehetséges.

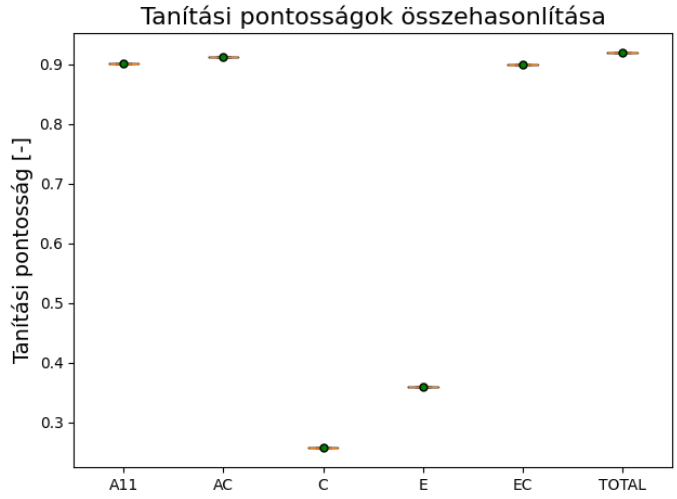
5. KÖSZÖNETNYILVÁNÍTÁS

Ez a kutatás a Kulturális és Innovációs Minisztérium által a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott Egyetemi Kutatói Ösztöndíj Program EKÖP-24-3-BME-383 számú Ösztöndíj támogatásával készült, továbbá a projekt a Doktoranduszi Kiválósági Ösztöndíj Program (DKÖP) által támogatott a Kulturális és Innovációs Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott, valamint a Budapesti Műszaki és Gazdaságtudományi Egyetem közös támogatásával, a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatallal kötött támogatásiszerződés alapján valósult meg.

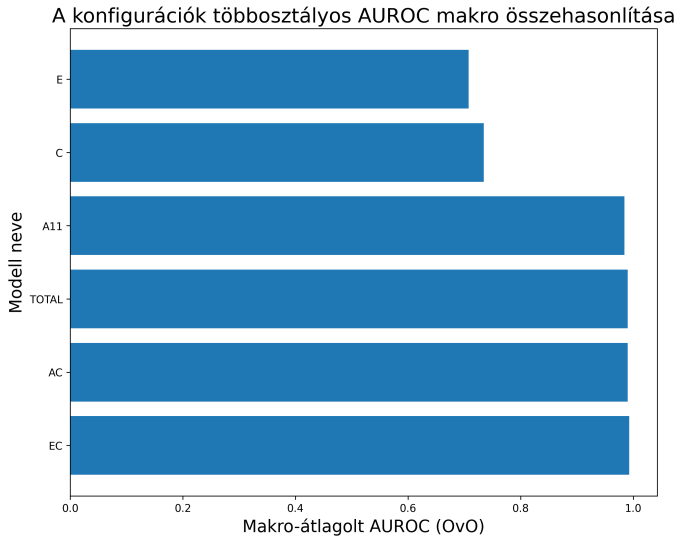
a)



b)



c)



2. ábra. *A tanítási hiba alakulása az epoch-okon keresztül az a) ábrán, míg a b) ábra a finomhangolásokat követő kimeneti pontosság alakulását mutatja. A c) ábra a konfigurációk többosztályos AUROC makro összehasonlítása.*

IRODALOMJEGYZÉK

[1] Weixin Liang és tsai. „The Widespread Adoption of Large Language Model-Assisted Writing Across Society”. *arXiv preprint arXiv:2502.09747* (2025).

[2] Álvaro López Caro. „Machine translation evaluation metrics benchmarking: from traditional MT to LLMs”. (2023).

[3] Farooq Shareef. „Enhancing Conversational AI with LLMs for Customer Support Automation”. *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*. IEEE. 2024, 239–244. old.

[4] Jingzhe Shi és tsai. „Chops: Chat with customer profile systems for customer service with llms”. *arXiv preprint arXiv:2404.01343* (2024).

[5] Sina Gholamian és tsai. „LLM-Based Robust Product Classification in Commerce and Compliance”. *arXiv preprint arXiv:2408.05874* (2024).

[6] World Customs Organization. *World Customs Organization*. Accessed on March 2025. 2025. URL: https://www.wcoomd.org/.

[7] Ashish Vaswani és tsai. „Attention is all you need”. *Advances in neural information processing systems* 30 (2017).

[8] T Wolf. „Huggingface's transformers: State-of-the-art natural language processing”. *arXiv preprint arXiv:1910.03771* (2019).

[9] Niklas Muennighoff és tsai. „MTEB: Massive text embedding benchmark”. *arXiv preprint arXiv:2210.07316* (2022).

[10] Shitao Xiao és tsai. *C-Pack: Packaged Resources To Advance General Chinese Embedding*. 2023. arXiv: 2309.07597 [cs.CL].

[11] Jacob Devlin és tsai. „Bert: Pre-training of deep bidirectional transformers for language understanding”. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, 4171–4186. old.

[12] He He és tsai. *Decoupling Strategy and Generation in Negotiation Dialogues*. 2018. arXiv: 1808.09637 [cs.CL].

[13] aladar. *Craigslist Bargains*. Hugging Face Datasets. Utolsó elérés dátuma 2025 március 8. 2023. URL: https://huggingface.co/datasets/aladar/craigslist_bargains.

[14] David J. Hand és Robert J. Till. „A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems”. *Mach. Learn.* 45.2 (2001. okt.), 171–186. old. ISSN: 0885-6125. URL: https://doi.org/10.1023/A:1010920819831.

[15] Hugging Face. *ROC AUC Metric on Hugging Face*. Accessed: 2024-11-03. URL: https://huggingface.co/spaces/evaluate-metric/roc_auc.