Multimodalinių modelių taikymas vaizdų antraščių generavimui lietuvių kalba

Airidas Žaliauskas, Viktor Medvedev

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, LT-08412 Vilnius
airidas.zaliauskas@mif.stud.vu.lt, viktor.medvedev@mif.vu.lt

Santrauka. Straipsnyje pristatomas tyrimas, kuriame analizuojamas multimodalinių modelių taikymas automatiniam vaizdų antraščių generavimui lietuvių kalba. Kadangi išsamių lietuvių kalbai skirtų tyrimų šioje srityje iki šiol nėra atlikta, straipsnyje daugiausia dėmesio skiriama naujausiems multimodaliniems modeliams ir jų galimybėms generuoti suprantamus vaizdų aprašymus lietuvių kalba. Eksperimentai buvo atliekami naudojant „*Gemma 3*“ multimodalinį modelį, kuris buvo adaptuotas lietuvių kalbai QLORA metodu. Gauti rezultatai patvirtina metodo efektyvumą ir galimybę sėkmingai jį taikyti lietuvių kalbos turinio generavimui iš nuotraukų.

Raktiniai žodžiai: vaizdų antraščių generavimas, kompiuterinė rega, natūralios kalbos apdorojimas, dirbtiniai neuroniniai tinklai, mašininis vertimas, transformeriai, QLORA.

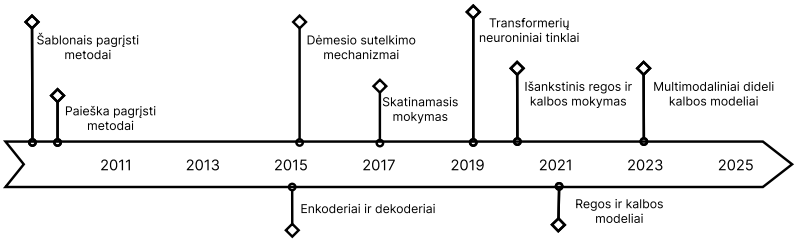
1 Įvadas

Vaizdų antraščių generavimas – tai procesas, kurio metu automatiškai generuojami vaizdų aprašymai natūralia kalba. Tai reikalauja atpažinti objektus, jų požymius ir ryšius paveikslėlyje bei išreikšti juos sakiniu ar sakinių rinkiniu, apibendrinančiu tai, kas pavaizduota. Tai paprastai sujungia kompiuterinės regos (angl. *computer vision*) ir natūralios kalbos apdorojimo metodus. Antraščių generavimas sulaukė didelio dėmesio dėl galimo pritaikymo tokiose srityse kaip prieinamumas (vaizdų aprašymas, turintiems regos sutrikimų), žmogaus ir kompiuterio sąveika [1]. Literatūros analizuojančios antraščių generavimo pritaikymą būtent lietuviškam tekstui iki šiol nebuvo, todėl šiame straipsnyje bus atliekamas tyrimas pritaikyti pažangiausius metodus vaizdų antraščių generavimui lietuvių kalba.

DOI: https://doi.org/10.15388/LMITT.2025.31

2 Antraščių generavimo metodai

Per pastaruosius du dešimtmečius vaizdų antraščių generavimas transformavosi nuo elementarių, taisyklėmis pagrįstų, sistemų iki sudėtingų multimodalinių modelių, integruojančių kompiuterinę regą ir kalbos supratimą. 1 pav. apžvelgiami didžiausią įtaką antraščių generavimo sričiai turėję akademiniai pasiekimai ir moksliniai atradimai.



1 pav. Reikšmingiausių pasiekimų, vaizdų antraščių generavimo srityje, laiko juostos diagrama.

Prieš pradedant dominuoti neuroniniams tinklams, vaizdų antraštės buvo kuriamos naudojant šablonais (angl. *template-based*) ir paieška (angl. *retrieval-based*) pagrįstus metodus. Taikant šablonais pagrįstus metodus, pirmiausia buvo nustatomi vaizdo elementai (objektai, atributai ir kt.) ir tada jais buvo užpildomi rankiniu būdu paruošti sakinių šablonai. Nors tai užtikrino, kad antraštėse būtų paminėti konkretūs aptikti objektai, sakinių struktūros buvo fiksuotos, todėl aprašymai buvo nelankstūs ir pasikartojantys, o sakinių įvairovė buvo ribota. Kita vertus, taikant paieška pagrįstus metodus, panaudojama egzistuojanti žmogaus parašytų antraščių duomenų bazė ir ieškoma tokios antraštės, kuri atitiktų naują vaizdą (dažniausiai surandant vizualiai panašų vaizdą). Tai leido sukurti gramatiškai taisyklingas antraštes, tačiau, kadangi buvo galima pasirinkti tik iš jau egzistuojančių pavyzdžių, šiam metodui trūko originalumo ar galimybės aprašyti nematytą vaizdą.

2010-ųjų viduryje įvyko reikšmingas pokytis. Sėkmingai pritaikius enkoderių (angl. *encoder*) ir dekoderių (angl. *decoder*) architektūras mašininio vertimo srityje, neuroniniai modeliai buvo pradėti taikyti vaizdų antraštėms generuoti. Proveržis įvyko 2015 metais, kai buvo išleistas modelis pavadinimu „*Show and Tell*“ [2], kuriame vaizdui koduoti buvo naudojamas konvoliucinis neuroninis tinklas (angl. *Convolutional neural network, CNN*), o tekstui nuose-

kliai dekoduoti – ilgos trumpalaikės atminties tinklas (angl. *Long short-term memory, LSTM*). Šis metodas buvo paprastas, bet ypač veiksmingas ir pranoko ankstesnius šablonais ir paieška pagrįstus metodus. Tais pačiais metais buvo išleistas „*Show, Attend and Tell*“ [3] modelis, kuriame buvo patobulinta ši architektūra panaudojant dėmesio sutelkimo mechanizmus. Šis mechanizmas leido dekoderiui dinamiškai sutelkti dėmesį į konkrečias vaizdo dalis kiekvienam sugeneruotam žodžiui, taip gerokai padidinant ir tikslumą, ir interpretaciją.

2017 metų naujovė – antraščių generavimo modelių optimizavimas, naudojant vertinimo rodiklius, tų modelių rezultatų kokybės vertinimui. Ši strategija gerokai padidino našumą, nes modeliai, adaptuoti naudojant skatinamąjį mokymą (angl. *reinforcement learning*), pasiekė ženkliai aukštesnius kokybės vertinimo rodiklius nei modeliai, apmokyti naudojant tik kryžminę entropiją [4]. Svarbus 2018 m. pasiekimas buvo „*Bottom-Up and Top-Down Attention*“ modelis [5], kuriame buvo pasiūlyta naudoti iš anksto apmokytą objektų aptikimo modelį („*Bottom-Up*“), kad būtų galima nustatyti regiono lygmens vaizdinius požymius, pavyzdžiui, stačiakampius (angl. *bounding boxes*) vaizde esantiems objektams. Tuomet, naudojamas iš viršaus į apačią (angl. *top-down*) dėmesio sutelkimo LSTM, kad dėmesys būtų sutelkiamas į šiuos nustatytus regionus, generuojant antraštes. Šis, į objektus orientuotas metodas, leido modeliams pagrįsti žodžius konkrečiais vaizdo regionais, užuot sutelkiant dėmesį į tolygiu tinkleliu padengtą vaizdą.

Dar 2017 metais išleista novatoriška transformerių neuroninių tinklų architektūra „*Attention is All You Need*“ [6] sukėlė revoliuciją natūralios kalbos apdorojime, pakeičiant rekurentinius neuroninius tinklus dėmesio sutelkimo mechanizmais (angl. *self-attention*). Bėgant laikui, tyrėjai nustatė, jog transformeriai puikiai veikia ne tik kaip kalbos dekoderiai, bet ir kaip kompiuterinės regos enkoderiai, leidžiantys modeliuoti vaizdų ir tekstų ryšius. Tuomet, antraščių generavimo modeliai pakeitė rekurentinius neuroninius tinklus, įdiegiant transformeriais grįstus modelius. Pavyzdžiui, 2019 m. buvo išleistas „*Object Relation Transformer*“ [7], kuris kodavo aptiktų objektų erdvinius ir semantinius ryšius, siekiant padidinti generuojamų antraščių sklandumą. 2020 m. antraščių generavimui buvo atsisakyta naudoti konvoliuciniais neuroniniais tinklais pagrįstus metodus, pakeičiant juos regos transformeriais (angl. *Vision transformer*, *ViT*), sujungtais su transformeriais grįstais neuroninių tinklų dekoderiais.

2021 metais vaizdų antraščių generavimo srityje buvo pradėti naudoti regos ir kalbos modeliai (angl. *Vision-Language Model, VLM*), gebantys susieti vaizdo ir teksto įterpinius bendroje įterpinių reprezentacijų erdvėje (angl.

embedding space), padidindami generuojamų antraščių tikslumą ir rišlumą. 2025 m. kovo 12 d. „*Google DeepMind*“ pristatė „*Gemma 3*“ [8] – multimodalinį modelį, pagrįstą regos ir kalbos modelių architektūra, galintį vienu metu apdoroti ir vaizdą, ir tekstą. „*Gemma 3*“ 27B parametrų modelis, atsižvelgiant į jo dydį, pasižymi geresniais rezultatais nei Llama3-405B, DeepSeek-V3 ir o3-mini modeliai, remiantis išankstiniais „*LMArena*“ žmonių vertinimų reitingais[1].

2023 metais antraščių generavimo srityje buvo pradėti naudoti multimodaliniai dideli kalbos modeliai, pavyzdžiui, GPT-4, galintis sklandžiai derinti vaizdo, teksto ir kalbos modalumus. Šie multimodaliniai modeliai geba interpretuoti sudėtingas vaizdų kompozicijas ar šnekamąją kalbą. Tai leidžia išplėsti jų pritaikymą sudėtingesnėms užduotims spręsti, pavyzdžiui, vaizdų apibūdinimui šnekamąja kalba. Konkrečiai vaizdų antraščių generavimo kontekste multimodaliniai modeliai pasinaudoja turtingu kalbos supratimu. Jie geba kurti tikslias ir išsamias, žmogaus rašyseną primenančias antraštes. Pavyzdžiui, 2024 metais „*Neurotechnology*“ pristatė didelį kalbos modelį lietuvių kalbai [9], kuris ne tik patobulino teksto generavimą lietuvių kalba, bet ir parodė, kaip tokie modeliai gali būti pritaikomi konkrečioms lingvistinėms savybėms. Ateityje, integruojant tokias technologijas į sistemas, gebančias apdoroti ir tekstą, ir vaizdus, bus galima generuoti kokybiškesnes bei tikslesnes antraštes skirtingoms kalboms, įskaitant mažiau paplitusias, kaip lietuvių.

3 Duomenų rinkiniai

„*Flickr*“ platforma, įkurta 2004 metais, yra internetinė nuotraukų talpinimo ir dalinimosi bendruomenė. Ji suteikė pradžią dviem populiariems duomenų rinkiniams – Flickr8k (8 000 nuotraukų, 40 000 antraščių) [10] ir išplėstam jo variantui Flickr30k (31 783 nuotraukos, 158 915 antraščių) [11]. MS COCO (2014 m.) [12] duomenų rinkinį sudaro 123 287 vaizdai, apimantys 80 objektų kategorijų, anotuoti taip, kad atspindėtų realaus pasaulio scenas natūraliame kontekste. Šio duomenų rinkinio antraštėse (vidutiniškai 52,49 simbolių) pirmenybė teikiama glaustam informacijos perteikimui. Tuo tarpu DOCCI duomenų rinkinys [13] skiriasi struktūriškai – jame 14 847 nuotraukų su pavienėmis žmogaus parašytomis antraštėmis vidutiniškai sudarytomis iš 640,79 simbolių arba maždaug 136 žodžių, t. y. daugiau nei 12 kartų ilgesnėmis nei MS COCO. DOCCI duomenų rinkinio privalumas yra tas, kad

[1] Šis teiginys paremtas „*Google DeepMind*“ tinklaraščio straipsniu https://blog.google/technology/developers/gemma-3/

jo struktūra leidžia kiekvieną vaizdą aiškiai atskirti nuo panašių vaizdų, remiantis jų aprašymu [13].

1 lentelė. Duomenų rinkinių palyginimas.

Pavadinimas	Nuotraukų (vaizdų) skaičius	Antraščių skaičius	Vidutinis antraščių ilgis (simboliais)
Flickr-8k	8 000	40 000	55,13
Flickr-30k	31 783	158 915	64,04
MS COCO (2014)	123 287	616 435	52,49
DOCCI	14 847	14 847	640,79

Pagrindinis iššūkis, pritaikant regos ir kalbos modelius lietuviškam turiniui yra tai, kad šiuose duomenų rinkiniuose pateiktos antraštės yra anglų kalba. Siekiant pašalinti šią kliūtį buvo atliktas keturių mašininio vertimo modelių vertinimas: „*Helsinki Opus*“, „*Madlad400*“ (3 mlrd. ir 7 mlrd. parametrų versijos), „*Seamless*“ ir „*NLLB-200*“ Vertimo į lietuvių kalbą kokybė buvo vertinama naudojant Flores-200 duomenų rinkinį pagal tris pagrindinius rodiklius, leidžiančius įvertinti vertimo tikslumą, rišlumą ir prasmę: BLEU, METEOR ir ROUGE-L. Flores-200 duomenų rinkinį sudaro 204 kalbos, įskaitant lietuvių. Angliški sakiniai buvo išversti į lietuvių kalbą naudojant visus anksčiau minėtus vertimo modelius, o rodikliai apskaičiuoti palyginant modelio sugeneruotą tekstą su žmonių pateiktais vertimais į lietuvių kalbą. BLEU – tai populiariausias mašininio vertinimo kokybės vertinimo rodiklis, kuris vertina, kiek išverstas tekstas sutampa su pateiktais pavyzdžiais pagal tikslias žodžių kombinacijas (n-gramas). METEOR – papildo BLEU tuo, kad vertina ne tik tiesioginį žodžių sutapimą, bet taip pat atsižvelgia į sinonimus ir žodžių tvarką. ROUGE-L – vertina vertimo kokybę pagal tai, kaip gerai išverstas tekstas išlaiko originalaus teksto prasmę, vertindamas ilgiausią bendrą žodžių seką [1].

2 lentelė. Mašininio vertimo modelių vertinimo rodiklių rezultatai Flores-200 duomenų rinkiniui.

Pavadinimas	BLEU	METEOR	ROUGE-L
Helsinki Opus	0,2303	0,4595	0,4830
Madlad400 (3 mlrd.)	0,2236	0,4660	0,4830
Madlad400 (7 mlrd.)	0,2250	0,4639	0,4833
Seamless	0,1953	0,4243	0,4541
NLLB-200	0,1825	0,4037	0,4360

Kaip matyti 2 lentelėje, „*Helsinki Opus*“ modelis verčiant iš anglų kalbos į lietuvių kalbą pasiekė geriausius rezultatus, kadangi BLEU balas buvo 0,2303, o ROUGE-L – 0,483. Tai rodo, kad jis efektyviai atlieka anglų-lietuvių kalbų vertimo užduotis. „*Madlad400*“ (3 mlrd. parametrų) pasiekė konkurencingus rezultatus, ypač pagal METEOR rodiklį (0,466), o tai rodo, kad jis geriau susidoroja su perfrazavimu, kas yra svarbu generuojamų antraščių natūralumui. Tačiau parametrų skaičiaus didinimas iki 7 mlrd. reikšmingo kokybės pagerėjimo nesuteikė (+0,0015 BLEU, -0,002 METEOR), tad daugiau nei 2 kartus daugiau parametrų turinčio modelio rezultatai nepateisina papildomų resursų reikalavimų. Sekantis, „*Seamless*“ modelis, pasiekė prastesnius rezultatus (BLEU – 0,195), o „*NLLB-200*“ surinko žemiausią įvertinimą (BLEU – 0,183), jo plati 200 kalbų aprėptis galimai sumenkino pritaikymą vertimui būtent į lietuvių kalbą. Atsižvelgiant į šiuos rezultatus, specializuotas „*Helsinki Opus*“ modelis yra pranašesnis anglų-lietuvių vertime, nei bendrosios paskirties sistemos.

4 Multimodalinių (regos ir kalbos) modelių pritaikymas

Pirmiausia, antraščių generavimui buvo išbandytas „*Neurotechnology*“ pristatytas didelis kalbos modelis, pritaikytas lietuvių kalbai. Jis sukurtas papildomai apmokant transformeriais grindžiamus „*Llama2*“ architektūros modelius lietuviškais duomenų rinkiniais [9]. Kadangi „*Llama2*“ yra tekstinio pobūdžio modelis, ši sistema gali apdoroti tik tekstinę informaciją. Šiai kliūčiai išspręsti buvo pritaikytas „*BLIP-2*“ [14] metodas, kuris naudoja tarpinį komponentą, vadinamą „*Q-Former*“. Šis komponentas leidžia sujungti užšaldyto vaizdo enkoderio arba regos transformerio išgaunamus vaizdo požymius su tekstu, generuojamu didelio kalbos modelio. Tokio būdo privalumas – efektyvus adaptavimas, nes papildomai apmokomas tik „*Q-Former*“ komponentas, o pagrindiniai modeliai išlieka užšaldyti, taip sumažinant skaičiavimo išteklius [14]. 2 pav. pateikiama adaptuoto „*Neurotechnology*“ modelio sugeneruotos antraštės palyginimas su originalia, į lietuvių kalbą išversta, antrašte. Modelis sugebėjo sugeneruoti įskaitomą, tačiau su vaizdu nesusijusį tekstą.

„*Gemma 3*“ modelis [8], kuris buvo pristatytas 2025 m. kovo 12 d., buvo adaptuotas „*Helsinki Opus*“ vertimo modeliu išverstu DOCCI duomenų rinkiniu, panaudojant QLORA (angl. *Quantized Low-Rank Adaptation*). QLORA – efektyvus derinimo metodas dideliems kalbos modeliams, leidžiantis sumažinti skaičiavimo išteklių poreikį, tačiau išlaikant aukštą našumą [15]. Pagal šią

Originalus aprašymas, išverstas į lietuvių kalbą: Aukšto kampo vaizdas iš seno išblukusio gatvės kampo. Vaizdas yra oranžinės spalvos purškimo dažytas žodisROW", virš kurio yra horizontali raidė". Dešinėje vaizdo pusėje yra iš dalies matoma ir išblukusi raudona linija gatvės kampe su žodžiaisFIRE LANE", labai išblukęs baltais dažais.

Sugeneruota antraštė: įvarių spalvų kombinacija, kurią sudaro įvairių atspalvių mėlyna, žalia, geletona, raudona, oražinė, rožinė, oranžinė, rožinė, oranžinė, rožinė, oranžinė, rožinė, oranžinė, rožinė, oranžinė, rož

2 pav. Originalios antraštės (išverstos į lietuvių kalbą) ir „*Neurotechnology*“ pritaikyto modelio sugeneruotos antraštės palyginimas.

metodiką, iš anksto apmokyti „*Gemma*“ svoriai kvantuojami 4 bitų tikslumu (naudojant NF4 kvantavimą kartu su dvigubu kvantavimu), išlaikant įšaldytus bazinius parametrus ir įvedant apmokomus LoRA adapterius (angl. *Low Rank Adapters*) visuose linijiniuose sluoksniuose. Mokymo procese naudojamas „*Hugging Face*“ SFTTrainer, kuris duomenis suformatuoja į pokalbių užklausas (sistemos instrukcija, naudotojo užklausa su paveikslėliu ir pavyzdine antrašte) ir mokymosi metu maskuoja nereikšmingas žymas. Atnaujinant tik nedidelę dalį modelio parametrų, naudojant LoRA adapterius, šis metodas išsaugo pradinį „*Gemma 3*“ gebėjimą analizuoti vaizdus ir kartu patobulina teksto generavimą lietuvių kalbai. Po apmokymo adapterius pasirinktinai galima sujungti su baziniu modeliu, taip sukuriant vientisą lietuviškų vaizdo antraščių generavimo sistemą, nereikalaujančią brangios techninės įrangos. 3 pav. ir 4 pav. palyginama originali antraštė išversta į lietuvių kalbą, naudojant „*Helsinki Opus",* ir modelio sugeneruota antraštė prieš ir po adaptavimo.

Originalus aprašymas, išverstas į lietuvių kalbą: Aukšto kampo vaizdas iš seno išblukusio gatvės kampo. Vaizdas yra oranžinės spalvos purškimo dažytas žodisROW", virš kurio yra horizontali raidė". Dešinėje vaizdo pusėje yra iš dalies matoma ir išblukusi raudona linija gatvės kampe su žodžiaisFIRE LANE", labai išblukęs baltais dažais.

Prieš adaptavimą: Paveikslelis matomas su plytelėmis ir šviesiai žalia linija. Linija brūkšneliu. Paveikslas yra išdėstytas iš apačios į viršų. Pagrįstas kaimo namo architektūra. Gali būti senoviškas. Paveikslas yra išdėstytas iš apačios į viršų. Pagrįstas kaimo namo architektūra. Gali būti senoviškas. Užrašyti "ROW" ir "FIRE LANE". Paveikslas yra išdėstytas iš apačios į viršų. Pagrįstas kaimo namo architektūra. Gali būti senoviškas.
Po adaptavimo: Vidaus rėmo, arti, akių lygis vaizdas iš keliu su medinio denio, einančiu per jį. Vaizdas yra nukreiptas į denio galą. Lentų jungtys klojimo iš viršaus į apačią dešinėje. Vaizdo apačioje kairėje yra rausvos dažytos juostos su tekstu ant jo, einančiu vertikaliai. Jis skaito "GAISRO PRADŽIA" su baltu fonu. Dešinėje pusėje denio yra oranžinė žodžio "Row" su smailia rodykle nukreipta į dešinę.

Oranžinis spalvos žodis "ROw" yra ant betono dangos, jo žodis yra parašytas didelėmis raidėmis, o žodis yra šiek tiek kreivas. Žodis yra įstrižai, žodis yra šiek tiek ištrauktas į dešinę ir į viršų, jo žodis yra ant betono dangos. Po žodžio rodomas mažas kvadratinis drenažo dangtelis, po drenažo dangtelio yra kvadratinė plytelių forma ir įstrižainė žemyn iš trapecijos formos, po plytelių yra betono dangos.

3 pav. Originalios antraštės (išverstos į lietuvių kalbą) ir „*Gemma 3*“ modelio sugeneruotų antraščių, šaligatvio žymėjimų nuotraukai, palyginimas prieš ir po adaptavimo.

4 pav. Lėktuvo nuotraukos antraščių palyginimas: originalios antraštės (išverstos į lietuvių kalbą) ir „*Gemma 3*“ modelio sugeneruotų antraščių palyginimas prieš ir po adaptavimo.

Prieš adaptavimą modelis dažnai „atsisakydavo“ generuoti antraštes lietuvių kalba, kaip matoma 4 pav., arba pateikdavo paviršutiniškus aprašymus anglų kalba. Po adaptavimo modelis geba pateikti detalius, kontekstą atitinkančius vaizdų aprašymus lietuvių kalba.

5 Išvados

Tyrime analizuojami ir vertinami šiuolaikiniai multimodaliniai modeliai, kurie buvo taikomi generuojant vaizdų antraštes lietuvių kalba. Tai vienas pirmųjų tokio pobūdžio tyrimų, nes iki šiol dauguma panašių darbų buvo orientuoti į antraščių kūrimą anglų kalba. Vertinant keturių skirtingų mašininio vertimo modelių kokybę pagal BLEU, METEOR ir ROUGE-L metrikas, „*Helsinki Opus*“ modelis parodė geriausius rezultatus verčiant tekstą iš anglų į lietuvių kalbą. Šis modelis pasižymi dideliu vertimo tikslumu, ženkliai pranokdamas universalius daugiakalbius modelius („*Madlad400*“, „*Seamless*“, „*NLLB-200*“), todėl buvo panaudotas išverčiant DOCCI duomenų rinkinį į lietuvių kalbą. Eksperimentų metu efektyviai adaptuotas „*Gemma 3*“ multimodalinis modelis, naudojant QLORA metodą, buvo pritaikytas antraščių generavimui lietuvių kalba. Pastarasis metodas pasirodė itin efektyvus, nes leidžia reikšmingai sumažinti skaičiavimo išteklių poreikį išlaikant pakankamai aukštą antraščių generavimo kokybę. Prieš adaptaciją „*Gemma 3*“ modelis sunkiai generavo antraštes lietuviškai arba pateikdavo paviršutiniškas anglų kalbos antraštes. Tačiau po adaptavimo pastebėtas reikšmingas patobulėjimas – generuojamos antraštės tapo nuoseklios, prasmingos, detalesnės ir tiksliai susietos su vaizdų turiniu. Empirinių tyrimų rezultatai rodo, kad

pritaikytas „*Gemma 3*“ modelis gali būti efektyviai naudojamas automatinėse vaizdų antraščių generavimo sistemose lietuvių kalba bei prieinamumo didinimui žmonėms su negalia.

Literatūra

[1] D. Sharma, C. Dhiman, and D. Kumar, “Evolution of visual data captioning Methods, Datasets, and evaluation Metrics: A comprehensive survey,” *EXPERT SYSTEMS WITH APPLICATIONS*, vol. 221, p. 119773, Jul. 2023, doi: 10.1016/j.eswa.2023.119773.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and Tell: A Neural Image Caption Generator,” in *2015 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR)*, in IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2015, pp. 3156–3164. doi: 10.1109/cvpr.2015.7298935.

[3] K. Xu *et al.*, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” Apr. 19, 2016, *arXiv*: arXiv:1502.03044. doi: 10.48550/arXiv.1502.03044.

[4] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-Critical Sequence Training for Image Captioning,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1179–1195. doi: 10.1109/CVPR.2017.131.

[5] P. Anderson *et al.*, “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 6077–6086. doi: 10.1109/CVPR.2018.00636.

[6] A. Vaswani *et al.*, “Attention Is All You Need,” in *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (NIPS 2017)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., in Advances in Neural Information Processing Systems, vol. 30. La Jolla: Neural Information Processing Systems (nips), 2017.

[7] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, “Image Captioning: Transforming Objects into Words,” Jan. 11, 2020, *arXiv*: arXiv:1906.05963. doi: 10.48550/arXiv.1906.05963.

[8] G. Team *et al.*, “Gemma 3 Technical Report,” Mar. 25, 2025, *arXiv*: arXiv:2503.19786. doi: 10.48550/arXiv.2503.19786.

[9] A. Nakvosas, P. Daniušis, and V. Mulevičius, “Open Llama2 Model for the Lithuanian Language,” Aug. 23, 2024, *arXiv*: arXiv:2408.12963. doi: 10.48550/arXiv.2408.12963.

[10] “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics | Journal of Artificial Intelligence Research.” Accessed: Mar. 26, 2025. [Online]. Available: https://www.jair.org/index.php/jair/article/view/10833

[11] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models,” in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, in ICCV ’15. USA: IEEE Computer Society, Dec. 2015, pp. 2641–2649. doi: 10.1109/ICCV.2015.303.

[12] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[13] Y. Onoe *et al.*, “DOCCI: Descriptions of Connected and Contrasting Images,” Apr. 30, 2024, *arXiv*: arXiv:2404.19753. doi: 10.48550/arXiv.2404.19753.

[14] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models,” Jun. 15, 2023, *arXiv*: arXiv:2301.12597. doi: 10.48550/arXiv.2301.12597.
[15] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLORA: efficient finetuning of quantized LLMs,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, in NIPS ’23. Red Hook, NY, USA: Curran Associates Inc., Dec. 2023, pp. 10088–10115.