Vaizdų aprašų generavimo modeliai

Artūr Radzivilov

Vilniaus Gedimino technikos universitetas,
Saulėtekio al. 11, LT-10223 Vilnius
artur.radzivilov@vilniustech.lt

Santrauka. Šiame straipsnyje yra nagrinėjami vaizdų aprašų generavimo modeliai, kurių pagalba galima automatizuoti teksto aprašymų kūrimą iš vaizdinės informacijos. Pateikiamos įvairios neuroninių tinklų struktūros, tokios kaip CNN ir RNN, kurios naudojamos vaizdų savybių išgavimui ir teksto generavimui, bei dėmesio mechanizmai ir „transformer" tipo tinklai, leidžiantys geriau integruoti vaizdo ir tekstinę informaciją. Analizuojami pagrindiniai duomenų rinkiniai, naudojami modelių mokymui, ir aprašymų vertinimo metodai, skirti įvertinti generuotų teksto aprašymų kokybę. Taip pat aptariamos naujausios tendencijos ir iššūkiai šioje srityje, pabrėžiant būsimų tyrimų kryptis.

Raktiniai žodžiai: vaizdų aprašų generavimas, CNN, RNN, dėmesio mechanizmai.

1 Įvadas

Šiame apžvalginiame straipsnyje nagrinėjami šiuolaikiniai vaizdų aprašymo generavimo modeliai, siekiant išanalizuoti kaip šie modeliai leidžia automatizuoti vaizdų aprašymus. Straipsnis skirtas išsamiai apžvelgti dabartinėje mokslinėje literatūroje aprašytas metodikas ir duomenų rinkinius, kuriuos naudoja vaizdų aprašymo sistemų kūrėjai. Pabrėžiama neuroninių tinklų, tokių kaip konvoliuciniai (CNN) ir rekurentiniai (RNN) neuroniniai tinklai, naudojimo svarba, taip pat aptariami dėmesio mechanizmai ir transformer tipo tinklai, kurie atlieka esminį vaidmenį siekiant efektyviai integruoti vizualinę ir tekstinę informaciją. Be to, šiame straipsnyje pateikiama vertinimo metodų analizė, leidžianti įvertinti generuotų aprašymų kokybę, ir aptariamos įvairios esamos technologijos bei jų stiprybės ir silpnybės [1].

2 Duomenų rinkiniai

Duomenų rinkiniai yra labai svarbus dalykas vaizdų aprašų generavimo sistemoje. Tam, kad sistema galėtų parodyti rezultatą, palyginamą su žmogaus vaizdo aprašymo rezultatu, reikalingi labai dideli duomenų masyvai,

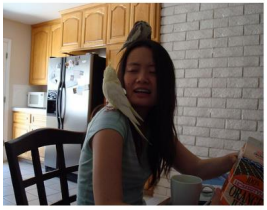
DOI: https://doi.org/10.15388/LMITT.2024.16

kuriuose privalo būti vaizdai ir nors vienas aprašas atitinkantis vaizdui. Kuo daugiau vienas vaizdas turi aprašų, tuo geriau, nes tą patį vaizdą skirtingi žmonės gali aprašyti skirtingai. Apmokant vaizdų aprašų generavimo sistemą, kuri galėtų pakeisti žmogų, reikia, kad aprašai duomenų rinkiniuose būtų sukurti žmogaus ranka.

Duomenų rinkinis „PASCAL1K“ buvo sukurtas 2010 metais. Jis turi 9000 vaizdų ir daugiau nei 40 000 aprašų šitiems vaizdams. „Amazon‘s Mechanical Turk“ (MTurk) padėjo autoriams sukurti šį duomenų rinkinį. MTurk leidžia pakankamai greitai rinkti didelį kiekį lingvistinių duomenų, nereikalaujant santykinai didelių investicijų.

MTurk taip pat turi ir minusų, vienas iš jų tai ribota galimybė kontroliuoti tai, kas iš personų gali dalyvauti tam tikroje užduotyje. Atliekant užduotis, reikalaujančias teksto įvedimo laisva forma, tai tas minusas, dėl kurio gali iškilti problemos. Tokio tipo užduotys skiriasi nuo užduočių su keliais galimais atsakymo variantais, kurie dažniausiai būna sukurti testine forma. Tai reiškia, kad tokio tipo užduočių rezultatą negalima bus patikrinti naudojant testinę užduotį, atsakymas į kurią yra žinomas. Dar vienas minusas susijęs su tuo, kad MTurk neturi įrankio, kuris garantuotų, kad visos personos, kurios spręs užduotis, gerai žinos anglų kalbą.

Kvalifikaciniai testai yra procedūros, kurias atlieka asmenys, norintys dalyvauti duomenų rinkimo procese, siekiant įsitikinti, kad jie atitinka tam tikrus kvalifikacijos standartus, pavyzdžiui, kalbos mokėjimą arba gebėjimą teisingai atlikti užduotis, prieš jiems leidžiant dalyvauti tekstų generavimo ir kitose panašiose užduotyse.



Without qualification test
(1) lady with birds
(2) Some parrots are have speaking skill.
(3) A lady in their dining table with birds on her shoulder and head.
(4) Asian woman with two cockatiels, on shoulder
head, room with oak cabinets.,
(5) The lady loves the parrot
With qualification test
(1) A woman has a bird on her shoulder, and another bird on her head
(2) A woman with a bird on her head and a bird on her shoulder.
(3) A women sitting at a dining table with two small birds sitting on her.
(4) A young Asian woman sitting at a kitchen
table with a bird on her head and another on her shoulder.
(5) Two birds are perched on a woman sitting in a kitchen.

1 pav. Aprašymų su ir be kvalifikacinio testo pavyzdys „PASCAL1K“ duomenų rinkinyje [2].

Kaip galima pamatyti 1 pavyzdyje su aprašais, rezultatai po kvalifikacinių testų ir be jų labai skiriasi. Kvalifikaciniai testai padėjo išvengti punktuacijos, rašybos ir kitų klaidų. Pasirinkti konkretesni aprašymai be tariamų santykių tarp objektų, kaip, pavyzdžiui, 5 punkte: „The lady loves the parrot“ (Ponia myli papūgą).

„Flickr8K“ ir „Flickr30K“ duomenų rinkiniai, sukurti nuotraukų aprašymų generavimo modelių mokymui, sudaryti iš tūkstančių įvairių tematikų nuotraukų, kurioms būdingi detalūs aprašymai, padedantys mokymosi procese [3]. „Flickr30K Entities“ papildo šiuos rinkinius, įtraukiant objektų lokalizaciją ir žodžių susiejimą su vaizdais, taip pagerinant modelių sugebėjimus generuoti tikslų ir kontekstinį turinį. Klaidos rinkinyje rodo, jog yra tobulėjimo galimybių, ypač kalbant apie rėmelių tikslumą ir objektų identifikaciją. „Microsoft COCO Captions“ duomenų rinkinys, praturtintas gausiais vaizdais ir aprašymais iš įvairių kategorijų, teikia dar didesnę įvairovę modelių mokymui, orientuotą į objektų atpažinimą, lokalizavimą ir aprašymų generavimą. Šie duomenų rinkiniai atlieka svarbų vaidmenį tobulinant ir vertinant vaizdų aprašymų generavimo technologijas, suteikdami tyrėjams reikiamus įrankius aiškesnėms ir tikslesnėms vizualizacijoms kurti.

3 Sugeneruotų aprašų vertinimas

Automatinių vaizdų aprašymų generavimo sistemos vertinimas yra sudėtingas procesas, kuris reikalauja objektyvių metrikų, kad būtų galima išmatuoti sugeneruotų aprašymų kokybę. Šiam tikslui mokslininkai ir inžinieriai naudoja įvairias vertinimo metrikas, tokias kaip BLEU, ROUGE, METEOR, CIDEr ir SPICE, kurios leidžia vertinti aprašymus remiantis skirtingais aspektais, įskaitant gramatiką, turinio tikslumą, sinonimų naudojimą ir semantinę prasmę. Kiekviena iš šių metrikų turi savo stiprybes ir silpnąsias puses, todėl jų kombinavimas suteikia išsamesnį ir objektyvesnį modelių vertinimą [4].

Metrikos, kaip BLEU ir ROUGE, dažnai naudojamos vertinant tekstų sutapimą ir aprašymų išsamumą [5], o METEOR prideda papildomą sluoksnį, atsižvelgdamas į gramatiką ir sinonimus [6]. Tuo tarpu CIDEr ir SPICE yra orientuoti į semantinį turinį ir aprašymų atitikimą žmogaus sukuriam konsensusui [7]. Visos šios metrikos padeda identifikuoti, kiek gerai automatinių sistemų generuoti aprašymai atitinka realius vaizdus ir jų kontekstą, tačiau taip pat pabrėžia, kad nėra vienos universalios vertinimo sistemos.

Būsimų tyrimų kryptys apima tiek esamų metrikų tobulinimą, tiek naujų metodų kūrimą, siekiant dar tiksliau įvertinti aprašymų kokybę ir atspindėti sudėtingesnius modelių generavimo aspektus. Svarbu atsižvelgti į modelių sudėtingumą, apmokymo laiką ir efektyvumą, kad būtų galima išsamiai įvertinti jų privalumus ir trūkumus. Tokie vertinimai yra būtini ne tik mokslinės pažangos požiūriu, bet ir praktine prasme, siekiant sukurti efektyvesnes ir energiją taupančias modelių architektūras, kurios atveria naujas taikymo sritis ir pagerina technologijų naudojimą kasdieniame gyvenime.

4 Egzistuojančios vaizdų aprašų generavimo sistemos

Vaizdų aprašymų generavimo sistemos yra svarbus dirbtinio intelekto ir kompiuterinės regos tyrimų segmentas, kuris siekia sukurti modelius, galinčius analizuoti vaizdus ir generuoti apie juos natūralios kalbos aprašymus. Šios sistemos remiasi pažangiomis dirbtinio intelekto technologijomis, tokiomis kaip konvoliuciniai neuroniniai tinklai (CNN), vaizdų savybių išgavimui, ir rekurentiniai neuroniniai tinklai (RNN), tekstui generuoti, taip pat dėmesio mechanizmais ir „transformer“ tipo tinklais, siekiant efektyviai apdoroti ir integruoti vaizdo bei tekstinius duomenis.

NIC yra vienas iš pirmųjų metodų šioje srityje, kuris efektyviai derina CNN, vaizdų savybių išgavimui, ir RNN, aprašymų generavimui, demonstruodamas kaip šie du komponentai gali bendradarbiauti generuojant aprašymus. NIC modelis, kuriame CNN naudojamas kaip koduotojas ir RNN kaip dekoduotojas, yra fundamentali koncepcija, kuri buvo pritaikyta ir tobulinta įvairiose vėlesnėse sistemose [8].

„Show, Attend and Tell“ metodas įvedė dėmesio mechanizmą į vaizdų aprašymų generavimo procesą, leidžiant sistemai ne tik identifikuoti vaizde esančius objektus, bet ir nustatyti kuriuos objektus ir kada reikėtų akcentuoti generuojant kiekvieną aprašymo žodį. Tai suteikė modeliams galimybę generuoti žymiai tikslesnius ir konteksto atžvilgiu prasmingesnius aprašymus [9].

„Transformer“ tipo tinklai, su jų naujovišku dėmesio mechanizmu, leidžia efektyviai tvarkyti ir analizuoti didelius duomenų kiekius, taip pat pritaikyti modelius įvairioms NLP užduotims, įskaitant vaizdų aprašymų generavimą. Jų gebėjimas apdoroti visą įvestį vienu metu, o ne paeiliui reiškia didžiulį žingsnį efektyvumo ir veikimo supratimo požiūriu [10].

„Image Transformer“ pritaiko „transformer“ tipo tinklų architektūrą tiesiogiai vaizdams, naudodamas lokalų dėmesį ir kvadratinį kontekstą vaizdo

pikselių analizei. Ši metodika leidžia modeliui efektyviai ir tikslingai generuoti vaizdo aprašymus, koncentruojantis į svarbiausius vaizdo elementus [11].

VLP (Vision and Language Pre-training) metodai, tokie kaip „ViLBERT“ ir „VisualBERT“, rodo, kad išankstinis mokymas, naudojant tiek vaizdo, tiek teksto duomenis, gali suteikti modeliams reikiamą lankstumą ir pritaikymą įvairioms vaizdo ir kalbos sąveikos užduotims atlikti. Šie metodai, derinantys galingas „transformer“ tipo tinklų architektūras su išankstiniu mokymu, atveria naujas galimybes modelių veikimo gerinimui vaizdų aprašymų generavimo srityje.

Šių sistemų plėtra ir tobulinimas leidžia ne tik generuoti tikslius ir prasmingus vaizdų aprašymus, bet ir gilinti mūsų supratimą apie vaizdo ir teksto sąveiką. Toliau vykstantys tyrimai ir technologinė pažanga tikrai atneš dar daugiau inovacijų ir tobulinimų, leidžiančių dar labiau pagerinti šių sistemų efektyvumą ir universalumą.

5 Neuroninių tinklų struktūros

Neuroninių tinklų struktūros yra esminė dirbtinio intelekto technologijų dalis, lemianti jų efektyvumą ir taikymą plačiame spektre užduočių. Šios struktūros atlieka svarbų vaidmenį analizuojant vaizdus, generuojant tekstus, atpažįstant kalbą ir sprendžiant kitus uždavinius.

CNN yra pagrindiniai vaizdo apdorojimo ir analizės darbuose, įskaitant objektų atpažinimą, vaizdų klasifikavimą ir segmentavimą. Jų struktūra, sudaryta iš konvoliucinių, aktyvavimo ir slenkstinių sluoksnių, leidžia efektyviai išmokti vaizdų savybes nuo paprastų iki sudėtingų. CNN naudojimas yra itin veiksmingas dėl gebėjimo atpažinti ir išgauti svarbius vaizdo bruožus, pritaikant mažiau parametrų ir išteklių nei kitos architektūros [12].

RNN yra skirti dirbti su sekos duomenimis, tokiomis kaip tekstas ar garso įrašai. Jų pagrindinė savybė – gebėjimas išsaugoti informaciją per ilgesnį laikotarpį. RNN architektūra leidžia informacijai keliauti nuo vieno apdorojimo žingsnio prie kito. Šis „žingsnis“ reiškia vieną iteraciją per duomenų seką, kurioje tinklas apdoroja vieną elementą (pvz., žodį ar garsą) ir perduoda savo būseną į kitą iteraciją. Tokiu būdu modelis įgauna gebėjimą atskleisti ir išmokti priklausomybes tarp iš eilės einančių sekos elementų, kas yra ypač svarbu užduotims kuriose reikia suprasti ir prognozuoti sekos tęstinumą.

Pagrindinės RNN sudedamosios dalys yra vidinė būsena, kuri atnaujinama kiekviename žingsnyje perduodant informaciją, ir grįžtamoji ryšio

struktūra, leidžianti informacijai cirkuliuoti per tinklą [13]. Tačiau RNN dažnai susiduria su gradientų nykimo problema, kuri apsunkina ilgalaikių priklausomybių mokymąsi [14].

Dėmesio mechanizmai padidina neuroninių tinklų efektyvumą, leisdami modeliams sutelkti dėmesį į svarbiausius duomenų elementus konkrečiu metu [9]. Tai yra ypač naudinga ilgų sekų apdorojimui, pavyzdžiui, teksto vertimui ar vaizdų analizei, nes dėmesio mechanizmai dinamiškai supranta, kiekvieno įvesties ar sekos elemento svarbą [15].

Architektūra „transformer" tipo tinklų yra pažangi, remiasi dėmesio mechanizmais ir atsisako tradicinių CNN ar RNN elementų. Ši architektūra yra labai veiksminga NLP užduotims ir vaizdų analizei dėl savo gebėjimo efektyviai apdoroti ilgas sekas, nenaudojant brangių rekursijos ar konvoliucijos operacijų. „Transformer" tipo tinklai yra pagrindas šiuolaikiniams AI modeliams, tokiems kaip GPT ir BERT, suteikiantys jiems gebėjimą suprasti ir generuoti kalbą bei analizuoti vaizdus nepaprastai aukštu lygiu.

„Encoder-decoder" architektūra yra būdinga užduotims, kuriose koduotojas apdoroja įvestį, o dekoduotojas generuoja išvestį. Ši schema naudojama įvairiose užduotyse, pavyzdžiui, mašininiame vertime, teksto generavime ir vaizdų aprašymų kūrime [16]. „Encoder-decoder" modeliai gali būti pagrįsti RNN, CNN ar „transformer" tipo tinklų architektūra ir dažnai įtraukia dėmesio mechanizmus, kad padidintų jų veiksmingumą ir gebėjimą išmokti sudėtingas priklausomybes tarp įvesties ir išvesties duomenų [10].

Šios pagrindinės neuroninių tinklų architektūros atlieka lemiamą vaidmenį formuojant šiuolaikinio dirbtinio intelekto technologijas, suteikdamos mokslininkams ir inžinieriams įrankius sudėtingų problemų sprendimui ir naujų, pažangių sistemų kūrimui.

6 Išvados

Šiame straipsnyje apžvelgti vaizdų aprašymų generavimo modeliai, remiantis naujausiomis dirbtinio intelekto technologijomis, tokiomis kaip CNN, RNN, dėmesio mechanizmai ir „transformer" tipo tinklai, parodo, kaip šios technologijos gali efektyviai transformuoti vizualinę informaciją į tekstinius aprašymus. Sistemos naudoja įvairius duomenų rinkinius modelių mokymui ir tobulinimui, o sugeneruotų aprašymų vertinimas atliekamas naudojant išsamią metrikų analizę, tokią kaip BLEU, ROUGE ir kitos, leidžiančią išsamiai įvertinti aprašymų kokybę ir nustatyti tobulinimo kryptis.

Tęsiant mokslinius tyrimus ir technologijų plėtrą, galime tikėtis, kad vaizdų aprašymų generavimo technologijos toliau evoliucionuos, tobulindamos jų sugebėjimą suprasti ir interpretuoti vizualinius bei tekstinius duomenis. Ateityje šios technologijos gali transformuoti tai, kaip mes suvokiame ir naudojame vaizdinę informaciją, integruodamos jas į įvairias pramonės šakas ir kasdienę veiklą, tokiu būdu suteikdamos naujas galimybes vartotojų sąveikai su technologijomis.

Literatūra

[1] Koršunova KP. Tasks and methods of automatic image description. Systems of Control, Communication and Security 2018;1:30–77.

[2] Rashtchian C, Young P, Hodosh M, Hockenmaier J. Collecting Image Annotations Using Amazon's Mechanical Turk 2010:139–47.

[3] Hodosh M, Young P, Hockenmaier J. Framing Image Description as a Ranking Task Data, Models and Evaluation Metrics Extended Abstract. IJCAI International Joint Conference on Artificial Intelligence 2015:4188–92.

[4] Kuznetsova P, Ordonez V, Berg AC, Berg TL, Choi Y. Collective Generation of Natural Image Descriptions 2012:8–14.

[5] Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02 2002:311–8. https://doi.org/10.3115/1073083.1073135.

[6] Denkowski M, Lavie A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. Proceedings of the Ninth Workshop on Statistical Machine Translation, Stroudsburg, PA, USA: Association for Computational Linguistics; 2014, p. 376–80. https://doi.org/10.3115/v1/W14-3348.

[7] Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE; 2015, p. 4566–75. https://doi.org/10.1109/CVPR.2015.7299087.

[8] Vinyals O, Toshev A, Bengio S, Erhan D. Show and Tell: A Neural Image Caption Generator. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2014;07-12-June-2015:3156–64. https://doi.org/10.1109/CVPR.2015.7298935.

[9] Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhutdinov R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. 32nd International Conference on Machine Learning, ICML 2015 2015;3:2048–57.

[10] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. Adv Neural Inf Process Syst 2017;2017-December:5999–6009.

[11] Parmar N, Vaswani A, Uszkoreit J, Kaiser L, Shazeer N, Ku A, et al. Image Transformer. 35th International Conference on Machine Learning, ICML 2018 2018;9:6453–62.

[12] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition 2015.

[13] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks. Adv Neural Inf Process Syst 2014;4:3104–12.

[14] Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation 2014.
[15] Luong M-T, Pham H, Manning CD. Effective Approaches to Attention-based Neural Machine Translation 2015.
[16] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate 2014.