IKER GAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

VI. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2025eko maiatzaren 28, 29 eta 30a
Bilbo, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Ikusizko Hizkuntza-ereduetan Kanpo Ezagutza eta Arrazoimendu Espaziala Txertatzen

Ander Salaberria, Gorka Azkune eta *Eneko Agirre*

195-202 or.
https://dx.doi.org/10.26876/ikergazte.vi.03.24

Ikusizko Hizkuntza-ereduetan Kanpo Ezagutza eta Arrazonamendu Espaziala Txertatzen

Ander Salaberria, Gorka Azkune, Eneko Agirre

HiTZ Zentroa - Ixa, Euskal Herriko Unibertsitatea (UPV/EHU). 20018 Donostia, Euskal Herria.

ander.salaberria@ehu.eus

Laburpena

Hizkuntza naturalaren prozesamendua (NLP) eta konputagailu bidezko ikusmenaren (CV) alorrak asko hazi dira azkenaldian. NLP eta CV-ren arteko zubian aurrerapenak lortu badira ere, gaur egun testuak eta irudiak prozesatzen dituzten sistemek hainbat ahulezia erakusten dituzte. Lan honetan, doktorego-tesi baten ekarpenak aurkezten dira, non ikusizko hizkuntza-ereduen bi ahulezi aztertu diren: munduko ezagutzaren integrazioa eta arrazonamendu espaziala. Alde batetik, irudietatik goiburukoak sortu ditugu hizkuntza-ereduetan inplizituki kodetuta dagoen munduko ezagutza hobeto aprobetxatzeko. Bestetik, objektu anotazioetatik datu sintetikoak sortzen zentratu gara, bai hizkuntza-ereduetan eta baita testu bidezko irudi sortzaileetan ere arrazonamendu espaziala hobetzeko.

Hitz gakoak: Hizkuntzaren prozesamendua, Konputagailu bidezko ikusmena, Ikusizko hizkuntza-ereduak, Munduko ezagutzaren Integrazioa, Arrazonamendu espaziala

Abstract

The fields of natural language processing (NLP) and computer vision (CV) have lately emerged. Although the bridge between NLP and CV has also advanced, nowadays these systems still face weaknesses with no trivial solution. In this work, we present the findings of a PhD thesis, where we analyzed two limitations of current Vision-and-language models: world knowledge integration and spatial reasoning. On the one hand, we verbalized images to leverage better world knowledge that is implicitly encoded in language models. On the other hand, we exploited the generation of synthetic data from object annotations to aid the spatial reasoning of both language models and text-to-image generators.

Keywords: *Natural language processing, Computer vision, Vision-and-language models, World knowledge integration, Spatial reasoning*

1 Sarrera eta motibazioa

Hizkuntzaren prozesamendua (NLP) eta konputagailu bidezko ikusmena (CV) makinak giza gaitasunez hornitzea helburu duten esparruak dira. Beste era batera esanda, NLP-k hizkuntza prozesatzea eta sortzea ahalbidetzen duen bitartean, CV-k makinei ikusmena imitatu eta ingurunea bisualki ulertzeko aukera ematen die. Ikuspuntu honetatik erraz uler daiteke biek izan ditzaketen elkarrekintzak, jendeak askotan ikusten duenari buruz arrazoitu eta hitz egiten baitu. Dena den, makinen ikusmenaren eta hizkuntzaren arteko integrazioa erronka bat izan da historikoki, gaur egun datu-kopuru masiboen erabilerak eta konputazio-ahalmen handitzeak aurrerapenak sustatu baditu ere.

NLP arloak azken urteetan izan duen arrakasta handia hizkuntza-eredu handien agerpenari esker lortu da, eredu estatistiko erraldoi hauek xede orokorreko testua sortzeko ahalmena erakutsi baitute (ChatGPT[1] bezalako aplikazioek erakusten duten bezala). Gainera, testua ez ezik, beste modalitate batzuk testuarekin lotzeko ahalmena erakutsi dute. Hau da ikusizko hizkuntza-ereduen kasua, non ikusmena eta lengoaia jorratzen dituzten atazak ebazten dituzten. Adibidez, testuan baldintzatutako irudi sorkuntza edota ikusizko galdera-erantzutea.

[1] https://chatgpt.com/ estekan atzigarri.

(a) Kanpo ezagutza galdera erantzuteko beharrezkoa da.

(b) Arrazonamendu espaziala ataza ebazteko beharrezkoa da.

1. Irudia: Ikusizko galdera-erantzutearen bi adibide, non helburua irudi bati buruzko galdera bat erantzutea den.

Har ditzagun 1. Irudiko bi adibideak ikusizko galdera-erantzutea ondo ebazteko behar diren gaitasun batzuk erakusteko. 1a. Irudiko adibidean karnabaren batezbesteko bizialdia zein den erantzun behar da. Horretarako, lehenengo irudiko txoria karnaba bat dela identifikatu eta, ondoren, karnabaren bizialdia lortu behar da. Informazio hori ez dago irudian eta datu hori lortzeko iruditik kanpo dagoen ezagutza[2] txertatu behar da ereduan. Bigarrenean, berriz, arrazonamendu espaziala ezinbestekoa da kolibriaren mokoa lore barruan ez dagoela prozesatzeko.

Gaur egungo ikusizko hizkuntza-ereduek (VLM) kanpo ezagutza txertatu edota arrazonamendu espaziala burutzeko arazoak izaten jarraitzen dituzte (Marino *et al.*, 2019). Hortaz, lan honen helburua VLM hauen mugei aurre egiteko alternatiba berriak proposatu eta haien gaineko analisiak burutzea izan da.

2 Arloko egoera eta ikerketaren helburuak

Atal honetan ikerketa honi dagokion artearen egoera lantzen dugu, erabilitako eredu eta teknika ezberdinak arloko egoeran kokatuz eta, horren harira, gure ikerketaren helburuak laburtuz.

2.1 Arloko egoera

Ikusizko hizkuntza-ereduak (VLM) terminoa oso erabilia izan da literaturan ikusizko eta testuzko datuak prozesatzen dituzten ereduak izendatzeko. Orokorrean, eredu hauek transformer arkitekturan oinarritzen dira (Vaswani *et al.*, 2017), irudi sortzaileak izan ezik, artearen egoera konboluzio sareetan oinarritutako difusio ereduek zehazten baitute (Rombach *et al.*, 2022). Nahiz eta VLM batzuk bi modalitateak prozesatzeko zerotik entrenatuak izan, VLM gehienak sistema unimodaletan oinarritzen dira (Li *et al.*, 2019), hizkuntza-eredu eta ikusizko kodetzailetan zehazki. Kasu hauetan, ikusizko kodetzaileek irudien adierazpenak kodetzen dituzte, ondoren hizkuntza-ereduek adierazpen horiek testuarekin batera inferentziak egiteko.

Modalitate anitzeko datu-multzoak. Hizkuntza-ereduak ikusizko datuetara adaptatzeko ezinbestekoa da bi modalitateak lerrokatuta dituzten datuak erabiltzea. Ebatzi behar den atazaren arabera, hainbat datu-multzo aurki daitezke. Adibidez, irudi goiburukoen sorkuntzarako, COCO datu-multzoa oso erabilia da (Lin *et al.*, 2014), non irudi bakoitzeko 5 goiburuko sortu diren giza-anotazioen bitartez. Datu-multzo honetan erabilitako 200k irudiak beste ataza batzuk sortzeko berrerabili dira, ikusizko galdera-erantzuterako datu-multzoetan, besteak beste (Antol *et al.*, 2015).

Kanpo ezagutzaren integrazioa. VLM-etan integrazio hau aztertzeko ohiko ataza ikusizko galdera-erantzutea (VQA) da. VQAv2 datu-multzoa oso erabilia bada ere (Goyal *et al.*, 2017), datu-multzo hau ez da egokia kanpo ezagutzaren integrazioa neurtzeko, galdera gehienak irudiaren edukiaren bitartez erantzun daitezke eta. Gure lanean, OK-VQA (Marino *et al.*, 2019) erabili dugu, eskuragarri dauden datu-multzoetatik handiena eta orokorrena baita. Ataza hau ebazteko VLM-ei kanpo ezagutza esplizituki txertatzen zaie, ezagutza-grafo edota ezagutza-iturri ezberdinak gehituz.[3]

[2]Lan honetan zehar, kanpo ezagutza edota munduko ezagutza bezala izendatu dugu.

[3]Hau zen orain dela 2-3 urteko artearen egoera, lan hau burutu zenekoa. Gaur egun, hizkuntza-eredu handien eragina dela eta artearen egoera zeharo aldatu da, ezagutza-grafoen erabilera guztiz ezabatuz literaturatik, OK-VQA atazarako behintzat.

Arrazonamendu espaziala. VLM-ek arrazonamendu espazial mugatua dutela zehazten duten lan asko publikatu dira azken urteetan, bai hizkuntza-ereduen kasuan eta baita irudi sortzaileen kasuan ere (Gokhale *et al.*, 2023). Honen azalpena entrenamendu datuetan erlazio espazialen agerpen kopurua urria dela izan daiteke. Horregatik, espazialki aberatsak diren datu sintetikoen sorrera burutu dugu, gabezi honi aurre eginik arrazonamendu espaziala hobetuko den hipotesia frogatu nahian.

2.2 Ikerketa helburuak

Lehen esan bezala, tesi honen helburu nagusia egungo VLM-en mugak arakatzea eta horiei aurre egiteko soluzio berriak garatzea izan da. Bi mugatan zentratu gara: i) *munduko ezagutzaren integrazioa*, non hizkuntza ereduetan aurkitutako ezagutza inplizitua hobeto aprobetxatzeko modu desberdinak aztertu ditugun, eta ii) *arrazonamendu espaziala*, non VLM-en entrenamenduko datuak sintekikoki sortu ditugun, irudi eta testu pareetan agertzen diren erlazio espazial kopurua artifizialki handituz. Arrazonamendu espazialari dagokionez, datu sintetikoen sorrera bai testua eta baita irudiak sortzen dituzten VLM-etara aplikatu dugu.

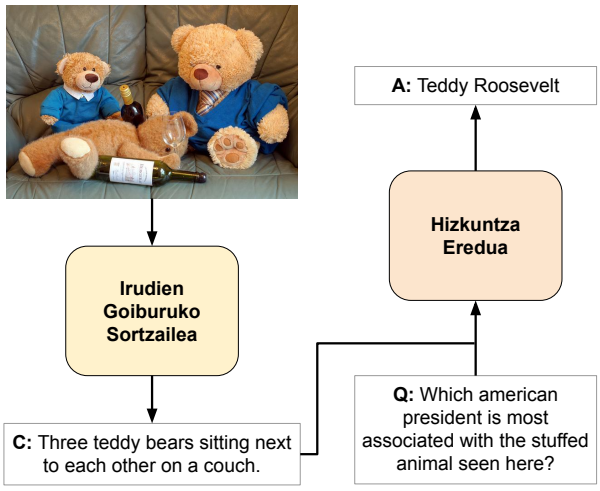
3 Ikerketaren muina

Atal hau hiru zatitan banatzen da. Alde batetik, 3.1. Atalean kanpo ezagutzaren inguruan arituko gara. Bestetik, arrazonamendu espazialaren ingurukoak 3.2. eta 3.3. Ataletan eztabaidatuko dira, hizkuntza-eredu eta irudi sortzaileen arrazonamenduan hain zuzen ere, hurrenez hurren.

3.1 Ezagutza Inplizituaren erabilera VQA sistemetan

Ekarpen honetan kanpo ezagutzan oinarritutako VQA ataza batean murgildu gara, OK-VQA atazan hain zuzen ere (Marino *et al.*, 2019). Bertan, galderak ondo erantzuteko irudiaren edukia izatea ez da nahikoa. Irudi eta testu pareekin ebatz daitezkeen VQA atazak ez bezala, ataza honek kanpo ezagutza txertatu, prozesatu eta ezagutza horretatik erantzunak inferitzeko ahalmena duten ereduen beharra dauka.

Guk aurrentrenatutako hizkuntza-ereduak ezagutza inplizitu iturri gisa erabili ditugu, hizkuntza-ereduen pisuetan kodetzen den ezagutza hain zuzen ere. OK-VQA atazan hizkuntza-ereduen erabilera ohikoa bada ere, eredu hauek ikusizko eta testuzko sarrera datuez elikatu behar dira ataza behar bezala ebazteko. Hizkuntza-ereduak testua bakarrik prozesatzeko diseinatuta daude, testu corpus erraldoietan estentsiboki entrenaturik. Hori dela eta, testuan soilik oinarritutako sistema batek ezagutza inplizitu hau hobeto aprobetxatuko duenaren hipotesia jarri dugu mahai gainean.



2. Irudia: Galdera eta irudi bat emanda, irudiaren edukia hitzez adierazten dugu goiburuko baten bidez, inferentzia unean aurrentrenatutako hizkuntza-eredu bat erabiliz.

3.1.1 Metodologia

OK-VQA ikusmen-testu ataza denez, irudiak automatikoki berbalizatzea proposatzen dugu irudiari buruzko informazioa hitzez adierazteko modu gisa. Gure kasuan, OSCAR eredua erabili dugu honetarako (Li *et al.*, 2020). Behin goiburuko hauek sortzean, planteatzen dugun metodoak testuzko datuak bakarrik erabiltzen ditu. Irudia berbalizatzeko garaian, informazio galera bat dagoela badakigu eta testua bakarrik erabiltzeak hasierako galera hori konpentsa dezakeen egiaztatu nahi dugu. 2. Irudian goiburukoetan oinarritutako eredu hau (*Caption based Model* edo CBM) antzeman daiteke.

CBM ereduaren bi aldaera definitu ditugu erabilitako hizkuntza-ereduaren arabera:

- CBM_{BERT}: BERT hizkuntza-ereduan oinarrituta dago (Devlin *et al.*, 2019), modalitate anitzeko MM_{BERT} ereduarekin konparaketa zuzenak egiteko (Marino *et al.*, 2021). MM_{BERT} ereduak objektu detektore batetik inferitutako adierazpenak jasotzen ditu goiburukoa beharrean.

- CBM_{T5}: T5 familiako ereduetan oinarrituta dago (Raffel *et al.*, 2020). Handituz doazen hizkuntza-ereduen gaitasuna konparatzeko erabili dugu.

Eredu hauek OK-VQA atazako irudi eta galdera pareetan ez ezik, VQAv2-ko instantzietan entrenatu ditugu (Goyal *et al.*, 2017). Entrenamendurako OK-VQA datu-multzoak 9k instantzia dituenez, ereduak irudien modalitatera moldatzeko datu gehiago erabiltzeak dakarzkien onurak aztertu nahi izan ditugu.

3.1.2 Emaitzak

Atal honetan gure esperimentazioaren emaitza nagusiak azaltzen dira. Esperimentu guztiak 3 aldiz errepikatu dira eta lortutako batezbesteko asmatze-tasa eta desbiderapen estandarrak ipini ditugu.

Eredua	Asmatze-tasa	+ VQAv2	Param.
Q_{BERT}	21,2 ±0,2	23,0 ±0,2	112M
MM_{BERT}	29,6 ±0,6	35,7 ±0,3	114M
CBM_{BERT} (ours)	**32,5** ±0,4	**36,0** ±0,4	112M

1. Taula: BERT ereduan oinarritutako ereduen errendimendua OK-VQA atazan. Zutabeetan, VQAv2 atazan doitu gabe eta doituta azaltzen dira emaitzak, hurrenez hurren, ereduen parametro kopuruekin batera.

Eredua	Asmatze-tasa	Param.
$CBM_{T5\text{-}Small}$	29,2 ±0,2	60M
$CBM_{T5\text{-}Base}$	36,1 ±0,5	220M
$CBM_{T5\text{-}Large}$	40,8 ±0,4	770M
$CBM_{T5\text{-}3B}$	44,0 ±0,7	3B
$CBM_{T5\text{-}11B}$	**47,9** ±0,2	11B

2. Taula: CBM_{T5} eredu sortzaileen errendimendua OK-VQA atazan, ereduen parametro kopuruarekin batera.

1. Taulan, MM_{BERT} eta CBM_{BERT} ereduak konparatu ditugu. Irudiaren informaziorik jasotzen ez duen Q_{BERT} eredua ere probatu dugu, bi aldaeren ekarpenak neurtzeko erreferentzia gisa. Emaitzetatik goiburukoak irudiak baino eraginkorragoak direla ondorioztatu daiteke OK-VQA atazan, tamaina berdineko ereduak erabiltzen direnean behintzat. Gainera, pareko emaitzak lortzen dituzte beste VQA osagarriekin doitzen badira, irudi eta testua lerrokatzen ikasteko datu kopuru handiagoa erabiltzea lagungarria dela erakutsiz.

2. Taulan, CBM_{T5} aldaerarekin lortutako emaitzak erakusten ditugu. Bertan, hizkuntza-ereduen tamaina eta kapazitatea handitzeak dituen onurak antzeman daitezke. Gainera, hobekuntza hau oraindik egonkortu ez dela antzeman dugu. Emaitzak artearen egoerarekin konparatu ditugu eta ezagutza-iturri ezberdinak erabiltzen dituzten VLM-ak hein handi batean gainditu ditugu (42,0 puntutik behera lortzen dituzte).

3.2 Arrazonamendu espaziala ikasten hizkuntza-ereduetan

Ezker eta *azpian* bezalako erlazio espazialen gaineko arrazonamendua burutzea ez dago ebatzita VLM-etan. Gainera, egoera are okerragoa da testua soilik prozesatzen duten hizkuntza-ereduentzat. Atal honetan, VLM-ak alde batera utzi ditugu eta testu hutsezko hizkuntza-ereduen oinarritze espazialean zentratu gara. Ikusizko informazioa testura itzultzeko joera jarraituz, testu tokenak era berri batean erabiltzea proposatzen dugu irudiko informazio espaziala kodetzeko. Xehetasunetan sartuz, kokapen tokenak erabiltzea proposatzen dugu eszena bateko objektuen posizioak eta tamainak zehazteko.

Gure hipotesia kokapen token horiek hizkuntza-ereduen arrazonamendu espaziala ikasteko modu eraginkor bat ahalbidetzen dutela da. Hau balioztatzeko, *Visual Spatial Reasoning* (VSR) datu-multzoaren bertsio berbalizatu batean egin ditugu esperimentuak (Liu *et al.*, 2023). Datu multzoak irudi eta goiburuko pareak ditu, non goiburukoak irudian agertzen diren bi objekturen arteko erlazio espazial bat zehazten duen. Horiekin batera, etiketa boolear bat dator, goiburukoa irudian betetzen den ala ez zehazten duena.

3.2.1 Metodologia

Testuzko adierazpen espazialak. Aipatutako kokapen token hauek dagoeneko hizkuntza-ereduaren hiztegian dauden zenbakien tokenak erabiliz definitzen dira, hots, tokenizatzailean zehaztutako hiztegian agertzen diren tokenak. Horrela, objektu baten posizio eta tamaina zehazteko lau kokapen token eta objektuaren izena erabiltzen ditugu. Irudiaren testuzko adierazpen honi esker, hizkuntza-ereduek *ezker* bezalako erlazio espazialak dagozkien kokapen token sekuentziekin erlaziona ditzakete, erlazio horiek oinarritzeko modua eskainiz.

Testuzko VSR ataza. Esan bezala, VSR ataza eraldatu dugu testuzko adierazpen espazialekin lan egin ahal izateko: (i) objektu detektore bat erabiliz eszenaren informazioa jasotzen dugu, (ii) deskribapen horietan kokapen tokenak gehitzen ditugu, *testuzko adierazpen espazialak* lortuz, (iii) goiburukoa eta eszenaren testuzko adierazpen

espaziala kateatzen ditugu eta lortutako token sekuentzia hizkuntza-ereduari elikatzen diogu, (iv) hizkuntza-eredua sailkapen bitarra burutzeko doitzen dugu.

Gainera, aldez aurretik hizkuntza-eredua adierazpen espazial hauetan trebatzeko aukera eskaintzen dugu guk garatutako datu-multzo espazial sintetikoan (*Synthetic Spatial Training Dataset* edo SSTD).

SSTD datu-multzoa. Datu-multzo hau hizkuntza-ereduei kokapen token eta erlazio espazialen arteko loturak erakusteko dago pentsatuta. COCO-ko irudi bat emanik (Lin *et al.*, 2014), objektu detektore bat erabiltzen dugu testuzko adierazpen espazialak sortzeko. Deskribapeneko bi objektu eta hauen kaxa inguratzaileak kontuan hartuta, arau eta txantiloi sinple batzuk erabili ditugu bi objektuen arteko erlazio espazialei buruzko galdera bitarrak sortzeko, bai positiboak eta baita negatiboak ere.

Bai SSTD eta baita VSR-en ere planteatutako atazetan sailkapen bitarrak egin behar dira, biak ebazteko arkitektura berdinak erabili daitezkeelarik. 3.1. Atalean bezala, BERT eta T5 ereduetan oinarritu gara sailkatzaile hauek eraikitzeko. Hortaz, kokapen tokenak eta SSTD entrenamenduaren bitartez hizkuntza-ereduek espazialki hobeto arrazona dezaketen aztertuko dugu.

3.2.2 Emaitzak

Eredua	Kokapen Tokenak	Asmatze-tasa
BERT	Ez	$62{,}11_{\pm 0{,}88}$
	Bai	$61{,}60_{\pm 0{,}92}$
$\text{BERT}_{\text{SSTD}}$	Ez	$61{,}83_{\pm 0{,}28}$
	Bai	$\mathbf{73{,}69_{\pm 0{,}88}}$

3. Taula: BERT-base eredua oinarritzat harturik, kokapen token edota ikasketa espaziala (SSTD datu-multzoan) erabiliz lortutako emaitzak VSR-ko ebaluazio azpimultzoan.

Eredua	Kokapen Tokenak	Asmatze Tasa
$\text{BERT}_{\text{SSTD}}$	Ez	76,96
	Bai	94,49

4. Taula: SSTD-ko garapen azpimultzoko asmatze-tasak kokapen tokenen erabileraren arabera.

3. Taulan VSR-ko ebaluazioan lortzen ditugun emaitzak azaltzen dira, ikasketa espaziala edota kokapen tokenak erabiltzearen arabera. Taulako lehenengo blokeak VSR-en doitutako BERT ereduaren errendimendua erakusten du. Bertan kokapen tokenen erabilerak ezberdintasunik ez dakarrela ikus daiteke. Hala ere, bigarren blokean antzemangarria da ezberdintasuna. Bloke hauetako ereduak SSTD datu-multzoan doitu dira lehenago, baina kokapen tokenak erabiltzen dituen eredua da hobekuntzak lortzen dituen bakarra. Hobekuntza hau nabarmena da beste ereduekiko, $\sim$12 puntu absolutuko aldea erakutsiz.

4. Taulan SSTD datu-multzoko garapenean lortutako emaitzak ikus ditzakegu. Ausaz erantzuten duen eredu batek $50{,}0$ puntuko asmatze-tasa lortuko balu ere, interesgarria da kokapen tokenik gabeko ereduak $76{,}96$ puntu lortzea. Honek kokapen tokenik gabe datuetan dauden hainbat alborapen ikasi ditzakeela erakusten digu, baina ataza ondo ebazten ikasteko nahikoa ez dela antzematen da ere, kokapen tokenak gehitzean $94{,}49$ko asmatze-tasa lortu baita.

Azkenik, 5. Taulan gure emaitzak VLM-ekin konparatzen dugu. VLM onenak, LXMERT-ek, $70{,}1$eko asmatze-tasa lortzen du, eta gure eredu guztiek nabarmenki gainditzen dute. Hala ere, guk informazio garrantzitsua galtzen dugu irudiaren deskribapena erabiltzean, objektuen izenak eta hauen kaxa inguratzaileak bakarrik erabiltzen ditugu eta.

Eredua	Asmatze-tasa	Param.
ViLT	$69{,}3_{\pm 0{,}9}$	87M
LXMERT	$70{,}1_{\pm 0{,}9}$	240M
BERT-base	$73{,}69_{\pm 0{,}88}$	110M
T5-Base	$73{,}09_{\pm 0{,}59}$	220M
T5-Large	$\mathbf{74{,}49_{\pm 0{,}36}}$	770M
T5-3B	$\mathbf{74{,}52_{\pm 0{,}25}}$	3B

5. Taula: VSR-ko ebaluazio azpimultzoan lortutako emaitzak. Lehenengo blokean VLMak azaltzen dira eta bigarren blokean, berriz, guk doitutako hizkuntza-ereduak aurkitzen dira.

3.3 Erlazio espazialek baldintzatutako irudien sorrera

Stable Diffusion (Rombach *et al.*, 2022) bezalako testu bidezko irudi sortzaileak arreta handia jaso dute azken aldian, beraien errendimendua asko hobetu baita. Hala ere, sistema hauek erlazio espazialak ez dituztela ondo adierazten antzeman da (Gokhale *et al.*, 2023).

Erlazio espazialak dituzten goiburukoen faltagatik motibatuta, entrenamendu datuak sortzean zentratu gara irudi sortzaileen artearen egoera hobetzeko. Zehazki, erlazio espazialak dituzten goiburuko sintetikoak automatikoki sortu ditugu, irudi errealekin lerrokatuz. COCO datu multzoko (Lin *et al.*, 2014) objektu anotazioak erabiltzen ditugu bi kaxa inguratzaileren arteko erlazio espazialak inferitzeko. Horietatik, goiburuko sintetikoak sortzen ditugu, irudi errealekin lerrokatuta daudenak, eta pare horiekin datu multzo bat osatzen dugu, *Spatial Relations for Generation* edo SR4G deritzoguna.

SR4G datu multzoa Stable Diffusion (SD) familiako ereduak doitzeko erabili dugu, erlazio espazial esplizituak dituzten irudi eta goiburuko pareekin ikasteak ereduen gaitasunak hobetuko dituela aurreikusiz. Doitutako ereduak ebaluatzeko eta jatorrizko SD ereduekin konparatzeko proposatu berri den VISOR metrika erabili dugu (Gokhale *et al.*, 2023).

3.3.1 Metodologia

SR4G datu-multzoa. Lan honetan testu bidezko irudi sortzaileentzat entrenamendu, garapen eta ebaluaziorako datu multzo bat sortu dugu. Irudi sortzaileak doitzeko irudi eta goiburuko pareak behar ditugu, baina ereduen ebaluaziorako, berriz, goiburukoekin soilik nahikoa dugu. Izan ere, aurreko lanetan egindakoa jarraituz (Gokhale *et al.*, 2023), irudi sortzailearen irteerak ez dira irudi errealekin konparatzen ebaluazioan.

- Ebaluaziorako goiburukoak: Goiburukoen sorkuntza burutzeko, $\langle$*subject, relation, object*$\rangle$ hirukote espazialen zerrenda bat definitu dugu. Hirukote espazial batek erlazio espazial bat (*relation*) zehazten du, goiburuko batean azalduko den izena (*subject*) eta objektuaren (*object*) arteko ezaugarri espazial bat zehaztuz. Gure hasierako hirukote espazialen zerrenda osatzeko COCO (Lin *et al.*, 2014) datu multzoan definitutako 80 objektuak erabili ditugu, hauekin eraikitako pareak 14 erlazio espazialekin konbinatuz: *left of* eta *inside*, besteak beste. Hirukoteak goiburuko bihurtzeko eskuz definitutako txantiloiak erabili ditugu. Goiburuko hauek bi objekturen arteko erlazio espaziala zehazten dute soilik, beste xehetasunik gehitzea ekiditen dugularik. Adibidez, $\langle$*dog, left of, cat*$\rangle \longrightarrow$ "A dog left of a cat.".
- Entrenamendurako irudi-goiburuko pareak: COCO 2017 bertsioko entramendu azpimultzoa erabili dugu irudi errealak eta hauen objektu anotazioak lortzeko. Horietatik goiburuko sintetikoak lortzeko metodologia bat definitzen dugu bi pausotan zatitzen dena: i) objektu anotazioetatik hirukote espazialak sortzen ditugu eta ii) hirukote hauetatik goiburukoak eraikitzen ditugu (ebaluazioko goiburukoetan bezala).
- SR4G Bertsioak: SR4G-ren bi bertsio eraiki ditugu: *Main* eta *Unseen*. *Main* bertsioan hirukote bera entrenamendu (*train*), garapen (*val*) edota ebaluazioan (*test*) zehar ager daitekeela esan nahi du. *Unseen* bertsioan, berriz, COCO datu multzoko 80 objektuak hiru azpimultzotan banatu ditugu honako banaketarekin: $|O_{\text{train}}| = 45$, $|O_{\text{val}}| = 5$ and $|O_{\text{test}}| = 30$. Horrela, erlazio espazialak entrenamenduan ikusi ez diren objektuetara orokortzeko ahalmena neurtu dezaekgu.

Ebaluazioa. Irudi sorkuntzan erlazio espazialak ebaluatzeko (Gokhale *et al.*, 2023) laneko metrikak erabili ditugu, ondoren laburtzen ditugunak.

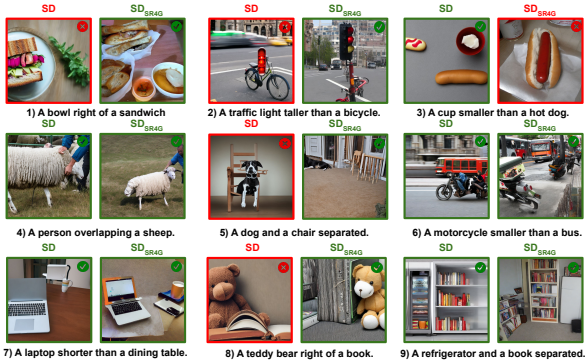
- Objektu Asmatze-tasa (OA): Metrika honek goiburukoan aipatutako bi objektuak sortutako irudian azaltzen diren ala ez neurtzen du.
- VISOR: bi objektuak sortu ote diren eta bien arteko erlazio espaziala zuzena ote den neurtzen du, OwL-VIT-ek aurkitutako objektuen arteko erlazio espazialak goiburukoan definitu direnak direla beramtuz.
- $\text{VISOR}_{\text{Cond}}$: VISOR-en parekoa, baina bi objektuak sortu diren kasuak bakarrik hartzen ditu kontuan. Sistema ezberdinek objektuak irudikatzeko ahalmen ezberdinak izan ditzaketenez, gure esperimentuetarako informazio baliagarriena $\text{VISOR}_{\text{Cond}}$ metrikak ematen digu, erlazio espazialen ulermena isolatzen baitu.

3.3.2 Esperimentuak

SD v2.1 doikuntza espazialik gabeko eredua eta SD_{SR4G} v2.1 eredu doitua izanik, 6. Taulan SR4G-ko datuekin doitzeak erlazio espazialen ulermena hobetzen duela antzeman daiteke. Hobekuntza hau entrenamenduan

Eredua	$\text{VISOR}_{\text{Cond}}$ ↑	VISOR ↑	OA ↑	Param.
Main split				
LayoutGPT	64,7	24,7	38,1	8.1B
VPGen	67,7	**34,5**	**51,0**	14.1B
SD v2.1	64,0	27,4	42,8	1.3B
SD_{SR4G} v2.1	**69,5**	31,7	45,6	1.3B
Unseen split				
LayoutGPT	64,7	24,7	38,1	8.1B
VPGen	68,4	**37,0**	**54,1**	14.1B
SD v2.1	64,0	28,4	44,4	1.3B
SD_{SR4G} v2.1	**69,4**	29,4	42,4	1.3B

6. Taula: Artearen egoerarekin konparaketa bi SR4G bertsioetan, ereduen parametro kopuruak zehaztuz.



3. Irudia: SD v2.1 eta *main* bertsioan doitutako SD_{SR4G} v2.1 ereduekin sortutako hainbat irudi.

zehar ikusi ez dituen objektuetan antzeman da baita ere (*Unseen* bertsioan), eredu doituaren orokortze ahalmenak erakutsiz. $\text{VISOR}_{\text{Cond}}$ metrikan erreparatu ezkero, gure emaitzak irudi sortzaileen ulermen espazialaren alorrean artearen egoera gainditzen dutela ikus daiteke. Gainera, VPGen (Cho *et al.*, 2023) ereduak duen arkitekura konplexu edota hizkuntza-eredu handien erabilera ekidin dugu. Hala ere, eredu hauek objektuak sortzeko erraztasun handiagoak jarraitzen izaten dituzte (OA eta VISOR metrikan antzeman daitekeena). Bukatzeko, 3. Irudian eredu hauen hainbat sorkuntza ikus daitzeke.

4 Ondorioak

Tesi honetan, gaur egungo VLM-en bi muga aztertu ditugu: munduko ezagutzaren integrazioa eta arrazonamendu espaziala. Bertan hainbat alternatiba proposatu dira muga hauei aurre egin ahal izateko. Alde batetik, hizkuntza-ereduetan aurki daitekeen ezagutza inplizitua eraginkorki nola erabili aztertu da, ikusizko galdera-erantzute sistemak erabiliz azterketa hau burutzeko. Irudia adierazteko testuzko modalitatea erabiltzea lagungarria dela ikustea ez ezik, eredu hauen tamaina kanpo ezagutza behar den atazetan lortzen duten errendimenduarekin korrelatuta dagoela ikusi dugu. Hala ere, joera hau kanpo ezagutza behar ez den atazetan ez dela betetzen antzeman dugu, atazaren arabera irudia adierazteko modua aukeratzea kontuan hartzekoa dela erakutsiz. Beste aldetik, VLM-en arrazonamendu espaziala hobetzeko datuak sintetikoki sortzea lagungarria dela frogatu dugu bi ataza ezberdinetan. Objektu anotazioak erabiliz, irudi eta testu adierazpenen arteko erlazio espazialen lerrokatzea hobetu dugu VSR atazan. Gainera, difusio ereduek burututako irudi sorkuntzan erlazio espazialen sorrera hobetu dugu hauen konplexutasuna eta arkitektura ukitu gabe. Azkeneko honek, gure hobekuntzak arkitektura aldaketekin osagarriak direla erakusten du, gure metodoa aplikagarria bihurtuz arkitektura ezberdineko edozein ereduetara.

5 Etorkizunerako planteatzen den norabidea

Tesi honetan jorratu diren lerroen bi mugei dagokienez, munduko ezagutzaren integrazioa arrazonamendu espaziala baino askoz gehiago ikertu da azken urteotan. Munduko ezagutza txertatzeari begira, ezagutza inplizitua bakarrik ez da gomendagarria. Azken finean, hizkuntza-ereduek (eta baita VLM-ek ere) errealak ez diren testuak sortzeko ohitura dute askotan. Hau ekiditeko, RAG sistemak proposatu dira (Lewis *et al.*, 2020), inferentzia unean behar den ezagutza eskuratzeko ahalmena ematen diotenak hizkuntza-ereduari. Ezagutza erauzketa pausoa irudi eta testuekin baldintzatzea ikerketa-lerro interesgarria dirudi OK-VQA bezalako atazetan. Bestalde, arrazonamendu espazialak ikasteko ereduen arkitekturan eta entrenamenduko galera funtzioetan aldaketak egiteko ikerketa-lerroak ireki nahi ditugu, sortutako datu sintetikoekin batera ereduen errendimendua hobetzeko.

Erreferentziak

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, eta Devi Parikh. 2015. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2425–2433.

Cho, Jaemin, Abhay Zala, eta Mohit Bansal. 2023. Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328* .

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, eta Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gokhale, Tejas, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, eta Yezhou Yang, 2023. Benchmarking spatial relationships in text-to-image generation.

Goyal, Yash, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, eta Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 6325–6334. IEEE Computer Society.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, eta others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33.9459–9474.

Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, eta Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* .

Li, Xiujun, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, eta others. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 121–137. Springer.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, eta C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, Fangyu, Guy Emerson, eta Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics* 11.635–651.

Marino, Kenneth, Xinlei Chen, Devi Parikh, Abhinav Gupta, eta Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14111–14121.

——, Mohammad Rastegari, Ali Farhadi, eta Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3190–3199. IEEE.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, eta Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21.1–67.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, eta Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Salaberria, Ander, 2024. *Integrating Outside Knowledge and Spatial Reasoning in Vision-and-language Models*. University of the Basque Country tesia. Available at `https://www.ixa.eus/node/14167?language=eu`.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanÑ. Gomez, Lukasz Kaiser, eta Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V.Ñ. Vishwanathan, eta Roman Garnett, 5998–6008.

6 Eskerrak eta oharrak

Artikulu hau lehenengo autorearen doktorego tesiaren laburpena da (Salaberria, 2024), gainontzeko autoreak tesi honen zuzendariak izanik. **Esker instituzionalak:** Eusko Jaurlaritzako Hezkuntza Sailari, ikerketa-lan hau egiteko emandako ikertzaileak prestatzeko bekarengatik, baita egonaldi internazionala burutzeko diru-laguntzarengatik ere. Tesiaren azkeneko urtea LUMINOUS proiektupean burutu da HE-CL4-DIGITAL23/02 (101135724), ikerketa batzorde Europarrak esleituta.