

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Erika Maksimova 206213IAIB

**MULTIMODAALSETE MUDELITE POTENTIAALI
UURIMINE AJALOOLISTE PILTIDE KLASSIFITSEERIMISEL**

Bakalaureusetöö

Juhendaja: Priit Järv
PhD

Tallinn 2024

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Erika Maksimova

25.05.2024

Annotatsioon

Antud lõputöö eesmärgiks on analüüsida näitetu õppe multimodaalsete mudelite võimet ajalooliste fotode klassifitseerimiseks ja pildiotsinguks. Multimodaalne mudel masinõppe kontekstis on sügava õppe mudel, mis kasutab erinevate andmemodaalsuste kombinatsioone, nagu tekst, heli või pildid. Näitetu õppe mudelid on masinõppe mudelid, mis on õpetatud objekte ära tundma ja kategoriseerima, isegi kui mudel ei ole näinud näiteid nendest kategooriatest. Selles töös analüüsidakse ja võrreldakse näitetu õppe mudelite ja traditsioonilisemate treenitud konvolutsiooniliste närvivõrkude (CNN) jõudlust ajalooliste piltide andmekogul arhiivplatvormilt *Ajapaik*[1].

Selle uurimistöö tulemusena oli analüüs erinevate asukoha ja kaamera nurga klassifitseerijate jõudluse kohta mitmekesiste piltide klassifitseerimiseks, mis jäädvustavad erinevaid ajaloolisi hooneid, portreid, maastikke ja kultuurisündmusi. Teine selle töö tulemus oli otsingumootori loomine, mis suudab leida pilte kas pildi sarnasuse või tekstiliste kirjelduste põhjal. Uuriti täiustatud pildiotsingut, ühendades fotode metaandmed fotode vektorestitusega, luues tugevama ühendatud visuaalse ja tekstilise arusaamise pildiotsingumootorile.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 40 leheküljel, 5 peatükki, 21 joonist ja 10 tabelit.

Abstract

Exploring the Potential of Multimodal Models for Historical Photos Classification

This thesis aims to analyse the capabilities of zero-shot multimodal models for historical photos classification and image search. A multimodal model, in the context of machine learning, is a type of deep learning model that uses a combination of various modalities of data, such as text, audio, or images. Zero-shot models are machine learning models which are trained to recognize and categorize objects without having seen any examples of those categories beforehand. In this work I will analyse and compare the performance of zero-shot models and more traditional trained convolutional neural networks (CNNs) on a historical images dataset from the archival platform *Ajapaik*[1].

The result of this research is an analysis of the performance of the various location and camera angle classifiers, working to classify diverse images capturing various historical buildings, portraits, landscapes and cultural events. Another output of this work was the implementation of a search engine capable of retrieving images based on either image similarity or textual descriptions. An avenue of an advanced image search has been researched, combining the photos metadata with the photo embeddings, creating a more powerful combined visual and textual understanding for the image search engine.

The thesis is written in Estonian and is 40 pages long, including 5 chapters, 21 figures and 10 tables.

Lühendite ja mõistete sõnastik

ANN	Ligikaudsed lähimad naabrid (<i>Approximate Nearest Neighbors</i>)
AP	Keskmine täpsus (<i>Average Precision</i>)
API	Rakendusliides (<i>Application Programming Interface</i>)
CLIP	Kontrastne keele-pildi eeltreenimine (<i>Contrastive Language-Image Pre-Training</i>)
CNN	Konvolutsiooniline närvivõrk (<i>Convolutional neural network</i>)
Convolutional layer	Konvolutsiooniline kiht
CV	Tehisnägemine (<i>Computer Vision</i>)
DenseNet	Tihe konvolutsiooniline närvivõrk (<i>Dense Convolutional Network</i>)
DH	Digihumanitaaria (<i>Digital Humanities</i>)
Embeddings	Vektoriesitused
Feed-forward network	Pärilevivõrk
Few-shot learning	Multinäideõpe
GPT	Generatiivne eeltreenitud transformer (<i>Generative pre-trained transformer</i>)
IR	Infootsing (<i>Information retrieval</i>)
Loss function	Kaofunktsioon
MAP	Keskliste täpsuste keskväärtus (<i>Mean Average Precision</i>)
NLP	Loomuliku keele töötlus (<i>Natural language processing</i>)
One-shot learning	Ainunäideõpe
ResNet	Residuaalne sügav närvivõrk (<i>Residual neural network</i>)
ViT	Visuaalne transformer (<i>Visual Transformer</i>)
Zero-shot learning	Näitetu õpe

Sisukord

1	Sissejuhatus	9
1.1	Taust	9
1.2	Ajalooliste fotode platvorm Ajapaik	10
1.3	Probleemi sõnastus	11
1.4	Töö struktuur	13
2	Kirjanduse ülevaade	14
2.1	Tehisnägemine ja multimodaalsus	14
2.1.1	Konvolutsioonilised närvivõrgud	14
2.1.2	Transformerid	14
2.1.3	Multimodaalne õpe	15
2.2	Näitetu õppe mudelid ja CLIP	15
2.3	Töö panus võrreldes eelneva kirjandusega	17
3	Metoodika	18
3.1	Andmete kogumine ja töötlemine	18
3.2	Pildi klassifitseerimine	18
3.2.1	Mudelite valik	19
3.2.2	CLIP	20
3.2.3	SigLIP	21
3.2.4	DenseNet	21
3.2.5	ResNet	21
3.3	Pildiotsing Annoy ja Voyager'iga	21
3.3.1	Pildiotsingu valideerimine	23
4	Tulemused	26
4.1	Klassifitseerimine	26
4.1.1	Näitetu õppe klassifitseerimine	27
4.1.2	Treenitud CNN-ga klassifitseerimine	29
4.1.3	Klassifitseerijate võrdlus	29
4.2	Pildiotsing	32
4.2.1	Lihtpildiotsing	32
4.2.2	Kombineeritud pildiotsing	33
4.2.3	Pildiotsingu valideerimine	35
5	Kokkuvõte	40

Kasutatud kirjandus	41
Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks	43
Lisa 2 – CLIP eksimismaatriksid	44
Lisa 3 – Pilditsing harvamate sõnapäringute järgi	46
Lisa 4 – Pilditsing juhuslikult valitud sõnapäringute järgi	50

Jooniste loetelu

1	Ajapaik platvormi liides. Fotol on sõjaväelane Aleksander Kärp.	11
2	Ajapaik platvormi kategooriad Aleksander Kärbi foto näitel	12
3	Vektoresituste leidmise ja sarnasuse arvutamise protsess	20
4	Ajapaik metaandmete töötlemine	23
5	Pildiotsingu tulemus, kus 2. ja 3. pildid on asjakohatud.	24
6	Pildiotsingu tulemus, kus 3. ja 4. pildid on asjakohatud.	24
7	Interior/Exterior klassifitseerijate võrdlus	30
8	Ground/Raised/Aerial klassifitseerijate võrdlus	30
9	Pildi sarnasuse alusel lihtotsing	33
10	Teksti sarnasuse alusel lihtotsing	34
11	Teksti sarnasuse alusel kombineeritud otsing	35
12	Päringu <i>Soviet</i> tulemus	37
13	CLIP Interior/Exterior eksimismaatriksid	44
14	CLIP Ground/Raised/Aerial eksimismaatriksid	45
15	Päringu <i>Aado</i> tulemus	46
16	Päringu <i>Janis</i> tulemus	47
17	Päringu <i>relations</i> tulemus	48
18	Päringu <i>yarn</i> tulemus	49
19	Päringu <i>battlefield</i> tulemus	50
20	Päringu <i>Riga</i> tulemus	51
21	Päringu <i>Fyodor</i> tulemus	52

Tabelite loetelu

1	Tulemuste osas käsitletud meetodite spikker	26
2	Interior/Exterior CLIP-i klassifitseerimisaruanded	27
3	Ground/Raised/Aerial CLIP-i klassifitseerimisaruanded	27
4	Erinevate klassinimede võrdlus CLIP ViT/B-16 klassifitseerimisel	28
5	Interior/Exterior SigLIP-i klassifitseerijad	28
6	Ground/Raised/Aerial SigLIP-i klassifitseerijad	28
7	Interior/Exterior CNN klassifitseerijad	29
8	Ground/Raised/Aerial CNN klassifitseerijad	29
9	8 sagedasemate sõnade päringute MAP	36
10	8 juhuslikult valitud sõnade päringute MAP	38

1. Sissejuhatus

Antud peatükis räägitakse fotode klassifitseerimise ja multimodaalsete mudelite lühiajaloo-
loost ning töös kasutatava ajalooliste fotode platvormist. Tutvustatakse ka töö probleemi
sõnastust ja töö struktuuri.

1.1 Taust

Ajalooliste fotode kategoriseerimine on kaalukas väljakutse visuaalsete materjalide tohutu
hulga tõttu. Traditsioonilised meetodid nõuavad sageli ulatuslikku pingutust ja ekspertiisi,
sundides teadlasi lugematuid arhiive läbi vaatama, püüdes käsitsi hinnata iga pildi konteksti,
sisu ja olulisust.

Viimase 10 aasta jooksul on tehisenägemise kiire areng võimaldanud digitaliseeritud kogude
piltide uurimist. On märgitud, et konvolutsioonilised närvivõrgud (CNN-id), st sügava
õppe mudelid saavad tuvastada arengusuundi suurtes pildikogudes[2]. Ühed enim kasu-
tatud pildiklassifikaatorite mudelid digihumanitaaria (DH) kogukonnas, CNN-id vajavad
spetsiifilise ülesande täitmiseks ulatuslikku treenimist ja suurt hulka andmeid, mis muudab
CNN-e heaks tööriistaks kitsa kategooriate komplekti kategoriseerimisele, kuid näitab oma
puudusi, kui kontekst muutub.

CNN-ide peamised probleemid on aja- ja ressursimahukas treenimine ning suured andme-
kogumid, mida tuleb käsitsi märgistada kasutamiseks treeningandmestikutena. Seetõttu
on tehisenägemise eksperdid otsinud võimalusi kasutada märgistamata andmeid tehisenä-
gemismudelite treenimiseks. Aastal 2021 arendas OpenAI välja mudeli nimega CLIP
(Contrastive Language-Image Pre-Training) – multimodaalse masinõppe mudeli, mis
suudab õppida visuaalseid mõisteid märgistamata visuaalsetest andmetest[3].

CLIP-i arendamisele järgnes teiste CLIP-i sarnaste mudelite arendamine, viies paljulubava
jõudluseni näitetu õppe (objektide äratundmine, mille instantsid võivad treeningu käigus
nägemata jääda[4]) klassifitseerimise täpsuse osas[5]. Multimodaalsete mudelite näitetu
õppe võimekus on nähtud kui ühte võimalikku lahendust digihumanitaaria töömahuka,
manuaalse andmete märgistamise probleemile. Multimodaalsetele mudelitele pole vaja
märgistatud andmed, kuna mudel on juba eelnevalt treenitud suure arvu teksti-pildi paaride
peal – CLIP’i puhul, 400 miljonit teksti-pildi paari.

Aastal 2023 avaldasid teadlased Thomas Smits ja Melvin Wevers artikli, mis tõestas, et CLIP-i saab efektiivselt kasutada et ennustada pilte, mis kujutavad endast nii üldiseid objekte nagu "loomad", kui ka konkreetsemaid objekte nagu "lõvi" või "koer", stseene ja isikuid, nagu kuninganna Victoria või Barack Obama, ja isegi abstraktseid mõisteid, nagu armastus või viha[2]. Kuid kas läänekeskne mõistmine on CLIP-i mõistmise limiit? Kui tõenäoline on, et CLIP teab, kuidas näeb välja pilt Tallinnast? Kas CLIP suudab öelda, millised pildid on vene-eesti ajakirjaniku Jüri Soonpere pere pildid?

Sarnaseid küsimusi on juba uuritud rahvusvahelises kogukonnas. OpenAI CLIP-ist inustudes, on teadlased treeninud mitmekeelseid multimodaalseid mudeleid, nt Chinese CLIP (suuremahuline mudel, eeltreenitud 200 miljoni pildi ja Hiina teksti paaride andmestikuga)[6], KELIP (Korean and English CLIP, treenitud 708 miljoni korea ja 476 miljoni inglise keele pildi-teksti paaride andmestikuga)[7] ja teised. Kevadeks 2024 pole veel saadaval ühtegi eestikeelset multimodaalset mudelit. Kuna miljonite eesti teksti-pildi paaride kogumine ja multimodaalse mudeli treenimine eesti teksti-pildi paaridel nõuaks aega väljaspool selle projekti raamistikku, ei keskendu see töö CLIP-i (või sarnase näitetu õppe multimodaalse mudeli) peenhäälestamisele ega selle arhitektuuri kohandamisele eesti keele rakenduse jaoks. Siiski julgustan tulevasi uurijaid kaaluma selle tee uurimist, kuna see võib pakkuda väärtuslikku panust tehisenägemise valdkonnas.

1.2 Ajalooliste fotode platvorm Ajapaik

Selles töös analüüsitakse, kui tõhusalt saavad näitetu õppe multimodaalsed mudelid automatiseerida ja optimeerida arhiiviuurimist Eestis. Selle projekti arhiivianalüüsi koguti *Ajapaik* platvormist[1] - avatud lähtekoodiga platvormist ajalooliste fotode rikastamiseks täiendava metaandmetega, nagu piltide kirjeldused, pealkirjad, geolokatsioon, inimeste tuvastamine ja fotokategooriad (vt Joonis 1).

On 2 kategooriat, mis üldiselt kirjeldavad kõiki fotosid *Ajapaik* platvormil: *asukoha* kategooriad ja *kaamera nurga* kategooriad. Iga foto saab olla kirjeldatud ühe asukoha kategooria ja ühe kaamera nurga kategooriaga, mida saab platvormi kasutajad käsitsi märkida või üle kirjutada. *Ajapaik* platvormi asukoha kategooriad on Interior (sisevaade) ja Exterior (välisvaade), ning kaamera nurga kategooriad on Ground (maapealne vaade), Raised (kõrgendatud vaade) ja Aerial (õhust vaade). Nagu näha Joonisel 2, on näidisfoto juba platvormi kasutajate poolt märgistatud kategooriatega *Interior* ja *Ground*.

Nagu eelnevalt mainitud, on võimalik fotosid ka märgistada, lisades käsitsi foto geolokatsiooni, pealkirja ja kirjelduse mitmes keeles, sealhulgas eesti, inglise, soome, läti, leedu, vene ja palju muud.

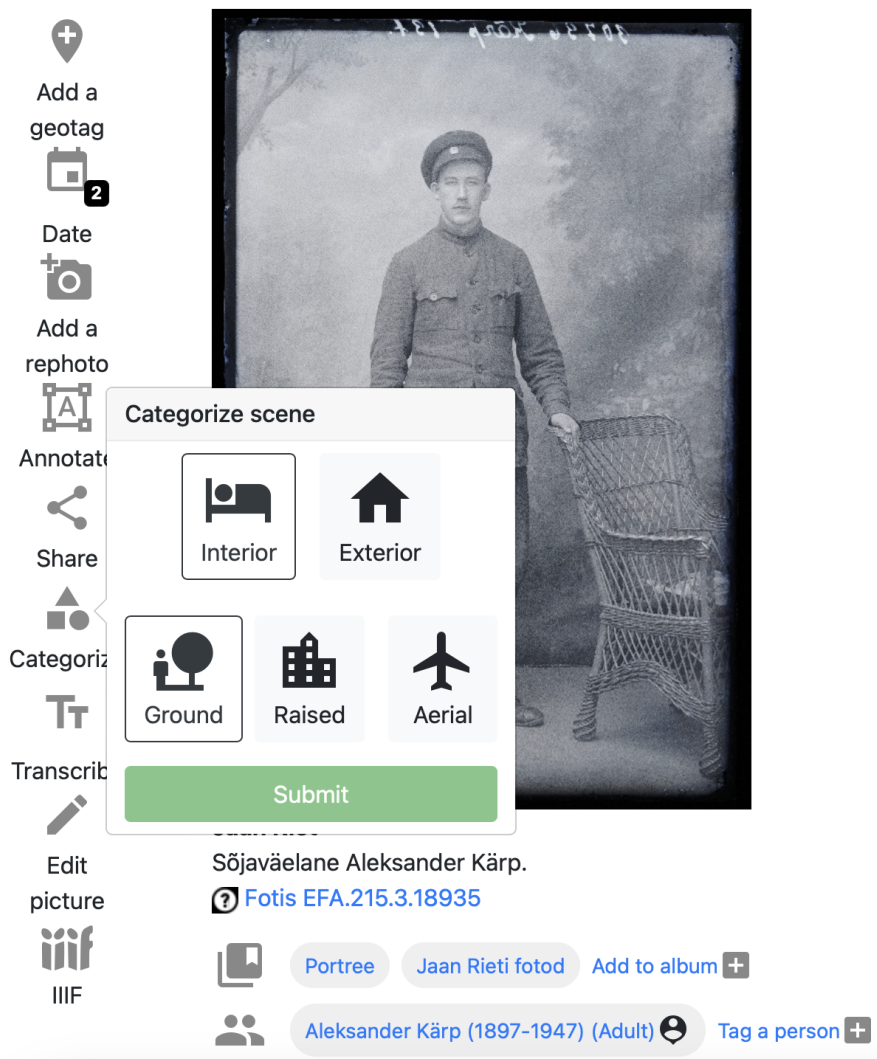


Joonis 1. Ajapaik platvormi liides. Fotol on sõjaväelane Aleksander Kärp.

1.3 Probleemi sõnastus

Ajalooliste piltide multimodaalne klassifitseerimine on väljakutset pakkuv ülesanne. Meetodikate arenguga kaasneb pidev vajadus hinnata uute sügavaõppe mudelite võimeid ja piiranguid.

Selle töö esimene eesmärk on uurida, kui tõhusalt saavad näitatu õppe multimodaalsed mudelid kategoriseerida ajaloolisi fotosid. Kasutades *Ajapaik* platvormi, saavad kasutajad käsitsi lisada foto kategooriaid kahe peamise tüübi järgi, mis on manuaalse töö puhul ajamahukas protsess. Konvolutsioonilise närvivõrgu ehitamine ja treenimine on samuti ajamahukas protsess, nõudes suure hulga korrektselt märgistatud andmete kogumist ja piisavaid ressursse mudeli treenimise ja valideerimise etappide läbiviimiseks. Selles töös uurin, kui hästi ennustab eeltreenitud näitatu õppe mudel kategooriakomplekte Interior/Exterior ja Ground/Raised/Aerial ilma mudelit peenhäälestamata, pakkudes ülevaadet mudeli



Joonis 2. Ajapaik platvormi kategooriad Aleksander Kärbi foto näitel

mõiste järeldeste võimekusest.

Teine eesmärk on võrrelda näitetu õppe mudeleid treenitud konvolutsiooniliste närvivõrkudega. Võrdlusena keskendub see töö CNN-ide, nagu ResNet ja DenseNet, ehitamisele ja treenimisele, pakkudes põhjalikku ülevaadet konkreetse klassifikaatori valimise eelistest ja puudustest arhiivsete fotode klassifitseerimise ülesande jaoks.

Kolmas eesmärk on välja selgitada, millised on multimodaalsete mudelite laiemad kasutusvõimalused arhiivuuringu valdkonnas. Selles töös uurin CLIP-i võimet abistada teadlasi infootsingu (IR) ülesannetes, peamiselt pildiotsingus. Kuna CLIP-i mudel sisaldab endas arusaamist pildi-teksti paaridest, suudab see luua pildi- ja teksti-vektoresituse samas vektorruumis. See töö keskendub ka analüüsile, kui hästi saab CLIP aidata pildiotsingus, näiteks piltidega, kus on kujutatud eesti inimesi ja kohti.

1.4 Töö struktuur

Ülejäänud töö on organiseeritud järgmiselt. Osas 2 annan näitetu õppe multimodaalsete mudelite arendamise kirjanduse ülevaate. Osas 3 tutvustan lühidalt iga kasutatud töös mudelit ning tutvustan pildiotsingu arhitektuuri ja selle hindamise vahendeid. Osas 4 tutvustan klassifitseerimistulemusi ning pildiotsingu jõudluse hindamist.

2. Kirjanduse ülevaade

Antud peatükis pakutakse ülevaadet olemasolevast kirjandusest, mis on seotud tehisnägemise, näitetu õppe multimodaalsete mudelite, pildiklassifitseerimise ja pildiotsinguga.

2.1 Tehisnägemine ja multimodaalsus

Enne kui räägime tehisnägemise rakendustest, peame tehisnägemise määratlema. Tehisnägemise peamine eesmärk on koguda kasulikku infot piltidelt[8]. Tehisintellekti valdkonnas võimaldab tehisnägemine arvutitel mõista piltide ja videote sisu ja ära tunda objekte, peamiselt sügavõppe kaudu ehk kasutades mitmekihilisi närvivõrke — sügavaid närvivõrke nagu konvolutsioonilised närvivõrgud (*CNN* ehk *Convolutional neural networks*).

2.1.1 Konvolutsioonilised närvivõrgud

CNN-id on pärilevivõrgud, mis on loodud töötleva andmeid, mis on esitatud massiividenä: 1D massiivid signaalide ja keelejadade jaoks; 2D massiivid piltide või helispektrogrammide jaoks; ning 3D massiivid video või volumeetriliste piltide jaoks [9].

CNN-id koosnevad konvolutsioonilistest kihtidest. Konvolutsiooniline kiht arvutab oma väljundi sisendi konvolutsiooniga (summeerides pildi vähemateks piksliteks[10]), lisades nihet ja viies iga tulemust läbi aktivatsioonifunktsiooni. Konvolutsioonilised kihid töötlevad iga pildipiirkonda sõltumatult, kasutades parameetreid, mis on jagatud kogu pildi ulatuses. Need kasutavad vähem parameetreid kui täielikult ühendatud kihid, ekspluateerivad ruumilisi suhteid lähedal asuvate pikslite vahel ja ei pea iga positsiooni juures pikslite tõlgendust uuesti õppima[11].

2.1.2 Transformerid

Konvolutsioonilised närvivõrgud pole aga ainsad mudelid, mis sobivad fotode klassifitseerimise ülesandeks. Aastal 2017 esitlesid Google'i teadlased uut lihtsat võrguarhitektuuri, Transformerit, mis ei kasuta rekurrentsust (*recurrence*) ja konvolutsioone ning põhineb ainult tähelepanu mehhanismidel[12].

Lühidelt öeldes, tähelepanu mehhanism on peidetud kiht, mis arvutab jaotuse, et teha valikut lähteelementide üle[13]. Tähelepanu mehhanismi kasutades saab mudel keskenduda

olulistele andmetele, andes neile rohkem kaalu kui andmetele, mida peetakse ebaolulisteks.

Transformer on tähelepanu mehhanismi arhitektuur, mis võimaldab ühe jada muundamist teiseks, kasutades kooderi ja dekooderi. Transformereid kasutatakse peamiselt loomuliku keele töötluste (NLP) mudelite jaoks, ehkki viimastel aastatel kasutatakse transformereid ikka enam arvutinägemise valdkonnas, ning on nad paljude multimodaalsete tehismärgide võrkude mootoriks.

2.1.3 Multimodaalne õpe

Kui mudel puutub kokku erinevate sisendi või väljundi modaalsustega, siis seda saab nimetada multimodaalseks. See tähendab, et piltide klassifitseerijat võib pidada multimodaalseks, kuna see võtab sisendina pildi ja väljastab tekstimärgendi[14].

Tähelepanu mehhanismi ja Transformeri arhitektuuri tutvustamine võimaldas luua mudeleid, mis suudavad suuremal määral töödelda erinevaid modaalsusi[15]. Seda võimalust kasutati algselt keelespetsiifiliste mudelite jaoks, nagu BERT[16] (mudel, mida saab häälestada tipptaseme mudelite loomiseks mitmesuguste ülesannete jaoks, näiteks küsimustele vastamise ja loomulikus keeles järeldamise jaoks) ja GPT-3[17] (tõlke-, küsimustele vastamise ja lünkade täitmise ülesanded ning ülesanded, mis nõuavad otsest järeldamist). Hiljem laiendati Transformeri arhitektuuri Visuaalsete Transformeriteni[18] – transformeerite tüüpideni, mis on mõeldud visuaalsete ülesannete, nagu pilditöötlus, jaoks.

Visuaalsete transformerite (ViT) tutvustamine on lõpuks viinud võimsate multimodaalsete mudelite arendamiseni, mis efektiivselt õpivad visuaalseid mõisteid loomuliku keele järelevalve abil. Üks selline mudel on CLIP[3], mille avaldas OpenAI 2021. aastal. Radford jt väidavad, et CLIP-i saab rakendada igale visuaalsetele klassifitseerimisülesandele, lihtsalt andes mudelile visuaalsete kategooriate nimed, sarnaselt GPT-3 "näitetu õppe" (*zero-shot*) võimalustele[19].

2.2 Näitetu õppe mudelid ja CLIP

Näitetu õppe eesmärk on tuvastada objekte, mille isendeid polnud treeningu ajal nähtud[4]. Näitetu õpe on üks kolmest järelalusviisist, millega saab mudeliga suhelda[17] ja on kontekstisõppe alamlihenemine; muud kaks kontekstisõppe tüüpi on multinäideõppe (*few-shot learning*) ja ainunäideõppe (*one-shot learning*).

Multinäideõppe eeldab, et lisatakse õppeprotsessi käigus lisaks juhistele mitmeid näiteid.

PROMPT: Palun klassifitseeri Pilt 0 kas Interior kategooriasse või Exterior kategooriasse.

Siin on mõned näited:

Pilt 1 on Interior.

Pilt 2 on Exterior.

Pilt 3 on Interior.

MUDEL: Pilt 0 on Exterior.

Ainunäideõpe eeldab, et lisatakse õppeprotsessi käigus vaid üks näide.

PROMPT: Palun klassifitseeri Pilt 0 kas Interior kategooriasse või Exterior kategooriasse.

On teatud, et Pilt 1 on Interior.

MUDEL: Pilt 0 on Exterior.

Näitetu õpe eeldab, et päringut teostatakse näiteid andmata.

PROMPT: Palun klassifitseeri Pilt 0 kas Interior kategooriasse või Exterior kategooriasse.

MUDEL: Pilt 0 on Exterior.

Põhinedes ülaltoodud näidetele, kuna CLIP-i töötamisel ei kasutata näiteid, saab seda klassifitseerida näitetu õppe mudeliks.

Standardsed tehisnägemise mudelid on head konkreetsete ülesannete lahendamisel, kuid neil võib olla raskusi uute ülesannetega kohanemisel, nõudes märkimisväärset pingutust ümberkonfigureerimiseks. CLIP-i eesmärk on lahendada probleeme praegustes tehisnägemise lähenemistes. Mõeldes mitmekülsusele, saab CLIP-i mudelit juhendada loomulikus keeles mitmesuguste klassifitseerimisülesannete lahendamiseks, ilma otsest optimeerimist individuaalse ülesande jõudluse jaoks[19].

OpenAI teadlased siiski nõustuvad, et võrreldes ülesandepõhiste mudelitega on CLIP-i näitetu õppe jõudlus endiselt üsna nõrk spetsiifiliste klassifitseerimise tüüpide puhul[3]. Kuna selle töö esimene eesmärk on kategoriseerida ajaloolisi fotosid asukoha ja kaamera-nurga järgi, mis mõlemad ei ole selgelt määratletud objektid, vaid on lihtsalt foto keskkond, pakub see töö ka täiendavat hindamist CLIP-i piiride kindlaksmääramiseks, kui mudel on teostatud mitmekesistel arhiivifotodel.

2.3 Töö panus võrreldes eelneva kirjandusega

Nagu varem mainitud, on juba tehtud uurimusi CLIP-i võimaluste multimodaalse mudelina kohta. Alec Radford jt artikkel[3] võrdleb CLIP-i jõudlust ResNet-50 mudeliga, samas kui Thomas Smitsi ja Melvin Weversi artikkel[2] uurib, kuidas multimodaalseid mudeleid saab kasutada binaarse klassifitseerimise, tekstist-pildini ja pildist-pildini pildiotsingu ülesannetes. Smits ja Wevers võrdlesid CLIP-i ka teise ResNet-mudeliga, ResNet-18.

See töö laiendab nende multimodaalsete mudelite analüüsi, võrreldes nii CLIP'i kui ka CLIP-sarnast mudelit SigLIP'i eelnevalt mainitud CNN-ide ResNet-18 ja ResNet-50 ning teistsuguse tüüpi CNN-i, DenseNet-121-ga. Sellega loodan anda piisava ülevaate näitetu õppe mudelite võimest klassifitseerida suhteliselt tundmatut ajaloolist andmestikku. Lisaks binaarsele klassifitseerimisele tegeleb see töö ka mitmeklassilise klassifitseerimisega, pakkudes ülevaadet multimodaalsete mudelite võimalustest selles valdkonnas.

See töö uurib ka, kuidas kohandada CLIP-i pildiotsingu rakendamiseks eesti ajalooliste fotode andmestikul. Nagu näidatakse hiljem, ei vaja CLIP märgistatud andmeid pildi vektorisuse väljastamiseks, kuid kui saame CLIP-ile pakkuda fotode tekstilisi kirjeldusi, on võimalik pildiotsingut täiustada. Lisaks püüab see töö kirjeldada näitetu õppe pildiotsingu valideerimise raskusi.

3. Metoodika

Antud peatükis kirjeldatakse selle töö metoodikat: andmete kogumist ja töötlemist, klassifitseerimise mudelite valikut ning pildiotsingu arhitektuuri ja valideerimist.

3.1 Andmete kogumine ja töötlemine

Klassifitseerimise ja lihtpildiotsingu ülesannete jaoks kasutati *Ajapaik* platvormi ajalooliste piltide kogumit. Pildikogum saadi *Ajapaik* platvormilt skriptide abil. Andmestik koosneb erinevatest ajaloolistest fotodest, sealhulgas portreedest, maastikest, sündmuste fotodest, postkaartide fotodest jm. Seda andmestikku nimetatakse edaspidi *ground truth* andmestiku nime all, kuna inglise keeles tähendab termin 'ground truth' tõde või tegelikkust andmetes. Andmestiku fotodel on ainult nende ID-d ja kategooriad, mis on määratud *Ajapaik* kasutajate poolt. Andmestikus pole pilte ilma kategooriateta. Kokku on *ground truth* andmestikus umbes 28450 pilte.

Ground truth andmestikku hiljem käsitsi jagati kaheks komplektiks: asukoha järgi (*Interior/Exterior*) ja kaamera nurga järgi (*Ground/Raised/Aerial*), kusjuures mõlemas hulgas olid fotod märgendatud õige kategooriaga. On võimalik, et sama foto võib esineda mõlemas hulgas, kuid erinevate kategooriatega, kuna *Ajapaik* platvormil võib foto olla kirjeldatud nii asukoha kui ka kaameranurga kategooriaga. Seejärel jagati mõlemad hulgad kolmeks alamhulgaks: treening-, test- ja valideerimisandmestikuks, järgides 70/15/15 jagunemist.

Kuna *Ajapaik* platvormil on piltide metaandmed saadaval *Ajapaik* API kaudu, täiendati hiljem osa algandmestikust, tõmmates metaandmeid (nagu fotopealkirjad, fotokirjeldused, geolokatsioon jne) *Ajapaik* API kaudu ja salvestades andmed JSON failina. Sellisel moel loodi uut andmestikku, mis nimetatakse edaspidi *täiendatud* andmestiku nime all. Kokku on täiendatud andmestikus umbes 10060 pilte. Andmestikus võivad esineda fotosid ilma määratletud kategooriateta.

3.2 Pildi klassifitseerimine

Pilte klassifitseeriti nii multimodaalsete kui CNN tüüpi mudelitega, võrreldes tulemusi kahes erinevas kategooriate komplektis.

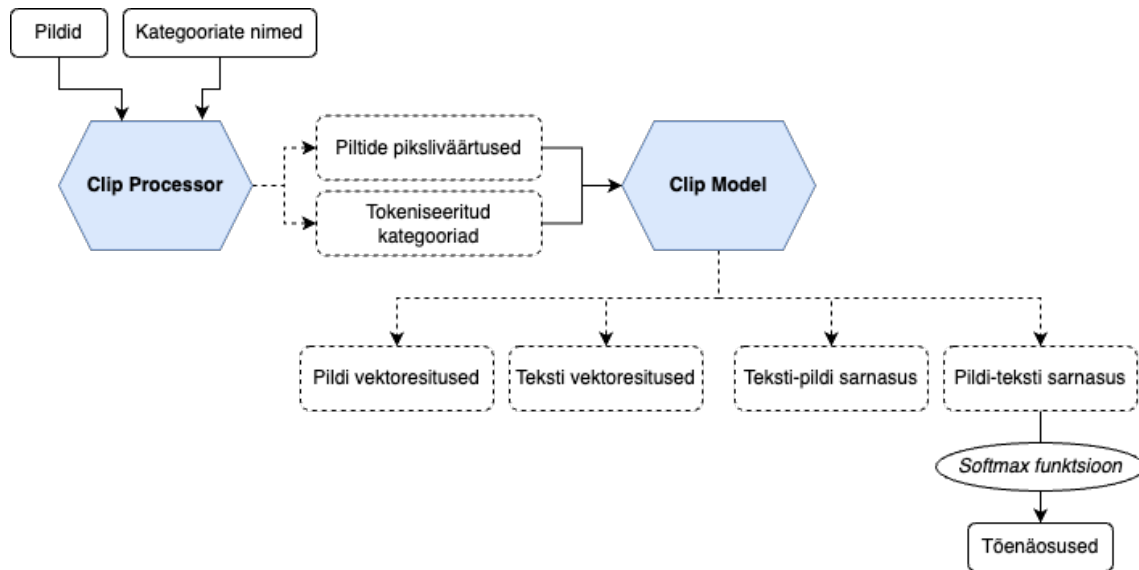
3.2.1 Mudelite valik

Klassifitseerimisülesande jaoks olid valitud järgmised mudelid:

- Näitetu õppe multimodaalsed mudelid:
 - **CLIP** – revolutsiooniline näitetu õppe arhitektuur. CLIP on olnud inspiratsiooni järgnenud multimodaalsetele mudelitele alates selle arendamisest 2021. aastal. Nagu oli selgitatud sissejuhatuses ja kirjanduse ülevaate osas, CLIP on populaarne rahvusvahelises kogukonnas, ja varasemate avaldatud artiklite põhjal saavutab CLIP tipptasemel jõudluse mitmesugustes multimodaalsetes ülesannetes.
 - **SigLIP** – arendatud 2023. aastal, SigLIP mudel asendab CLIP-s kasutatud kaotusfunktsiooni (*loss function*) lihtsa paarikaupa sigmoidkaoga (*pairwise sigmoid loss*)[5]. Artikkel väidab, et see toob kaasa parema näitetu õppe klassifitseerimise täpsuse ImageNet andmestiku[20] puhul.
- Konvolutsioonilised närvivõrgud:
 - **ResNet** – tutvustatud 2015. aastal, ResNet (*Residual Network*)-i arhitektuuri peetakse üheks populaarseimaks konvolutsiooniliseks närvivõrgu arhitektuuriks [21]. Selles töös töötan ResNet-18 ja ResNet-50-ga, sarnaselt varasemalt avaldatud CLIP-i hindamistöödega.
 - **DenseNet** – arendatud 2016. aastal, DenseNet (*Dense Convolutional Network*) on ResNet-idele konkurentsivõimeline alternatiiv, saavutades võrreldavad või paremad tulemused väiksema arvu parameetritega kui ResNetid. Selle konvolutsioonilise närvivõrgu tüübi puuduseks on kõrgem mälukasutus [22]. Selles töös kasutan DenseNet-121, valides võimalikult väikese arvu kihte, kuid siiski rohkem kihte kui ResNeti mudelite puhul, et pakkuda alternatiivi, mida varem pole võrreldud multimodaalsete mudelitega.

Mõlemad multimodaalsed mudelid olid valitud nende näitetu õppe õppimisvõime ja efektiivsuse tõttu. Rohkemate multimodaalsete mudelite testimine jäi selle projekti skoobist välja, kuid on kindlasti olemas teised CLIP-ist inspireeritud multimodaalsed mudelid, näiteks BLIP, BLIP-2, GIT, EVA-CLIP, Fuyu jm.

Alampeatükkides selgitan lühidalt iga mudeli kasutust.



Joonis 3. Vektorestituste leidmise ja sarnasuse arvutamise protsess

3.2.2 CLIP

Klassifitseerimise ülesande jaoks oli kasutatud 3 CLIP-i mudelit: ViT-B/16 Transformer mudel¹, ViT-B/32 Transformer mudel² ja ViT-L/14 Transformer mudel³.

Kuna CLIP on eelnevalt treenitud näitetu õppe mudel, siis mudelite puhul ei olnud kasutatud treenimist. Kuigi on võimalik CLIP-i uue andmetega häälestada, jääb see väljapoole probleemipüstitust, mis on CLIP-i kasutamine näitetu õppe mudelina.

CLIP HuggingFace implementatsioon on objektorienteeritud arhitektuuriga, ja kasutamiseks tuleb vastavaid klasse importida – CLIP Model (eeltreenitud mudel) ja CLIP Processor (ümbrisklass CLIP-i tunnuste eraldaja ja CLIP-i tokeniseeriija jaoks).

CLIP-i testfaasi võib ka samuti kirjeldada kui ennustusfaas. CLIP Protsessorile antakse nii testhulga andmeid (ilma määratletud kategooriateta) kui ka ennustatavaid kategoorianimesid. Pildid töödeldakse piksliväärtusteks ja kategoorianimed tokeniseeritakse ning edastatakse CLIP-i mudelile. Seejärel väljastab mudel nii pildi- kui ka kategooriate vektorestitused koos vastavate sarnasustega (vt Joonis 3). Piltide tõenäosuste saamiseks on vaja reaalarvude vektorid teisendada vektoriteks vahemikus 0 kuni 1 läbi softmax aktiveerimisfunktsiooni.

¹<https://huggingface.co/openai/clip-vit-base-patch16>

²<https://huggingface.co/openai/clip-vit-base-patch32>

³<https://huggingface.co/openai/clip-vit-large-patch14>

3.2.3 SigLIP

Klassifitseerimise ülesande jaoks oli kasutatud 3 WebLi[23] põhjal eeltreenitud SigLIP mudelit: eeltreenitud resolutsioonil 224x224 mudel⁴, eeltreenitud resolutsioonil 256x256 mudel⁵ ja eeltreenitud resolutsioonil 384x384 koos SoViT-400m arhitektuuriga mudel⁶.

Kuna SigLIP on kirjeldatud kui CLIP mudel parema kaofunktsiooniga, siis selles töös oli SigLIP kasutatud ja testitud sarnaselt CLIP-i mudelitega.

3.2.4 DenseNet

Võrdlusena oli kasutatud eeltreenitud DenseNet-121 mudel⁷. Kuna valitud mudeli versioon on treenitud ImageNet-1K[24] andmestikul, kasutati ülekandõppe meetodit mudeli treenimiseks, st uuendati ainult uusi lisatud klassifitseerimiskihtide parameetreid. Mudeli treenimiseks kasutati *ground truth* test- ja valideerimisandmestikke. Valideerimistäpsust kasutati parima mudeli salvestamiseks, ja seejärel testiti treenitud mudelit testandmestikuga.

3.2.5 ResNet

Sarnaselt DenseNet'iga, olid kasutatud ka ImageNet-1K[24] andmestikul eeltreenitud ResNet-18⁸ mudel ja ResNet-50⁹ mudel, kasutades ülekandõppe meetodit ja *ground truth* andmestikku.

3.3 Pildiotsing Annoy ja Voyager'iga

Annoy (*Approximate Nearest Neighbors Oh Yeah*) on Spotify poolt välja töötatud teek, mis otsib ruumis lähedasi punkte antud päringupunktile[25]. Voyager on veel üks Spotify poolt välja töötatud teek, mida kirjeldatakse kui järgmise põlvkonna Annoy-d, kuid palju suurema saagisega[26] – st Voyager leiab vasteid otsingule usaldusväärsemalt. Mõlemad teegid on implementeeritud objektorienteeritud arhitektuuriga.

⁴<https://huggingface.co/google/siglip-base-patch16-224>

⁵<https://huggingface.co/google/siglip-large-patch16-256>

⁶<https://huggingface.co/google/siglip-so400m-patch14-384>

⁷<https://pytorch.org/vision/main/models/generated/torchvision.models.densenet121.html>

⁸<https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html>

⁹<https://pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html>

Pildiotsingu ülesande jaoks kasutati Annoy ja Voyager-i teeki, et luua kolm lähimate naabrite otsimise indeksit. Peale piltide töötlemist vektoresitusteks CLIP-i abil, vektoresitused on lisatud tühjadesse indeksklassidesse, luues *forest of trees* indekseid. Iga loodud indeks sisaldab vektorandmeid ja nendega seotud ID-sid. Indeksit saab pärida, et saada antud vektorite lähimad N naabrit.

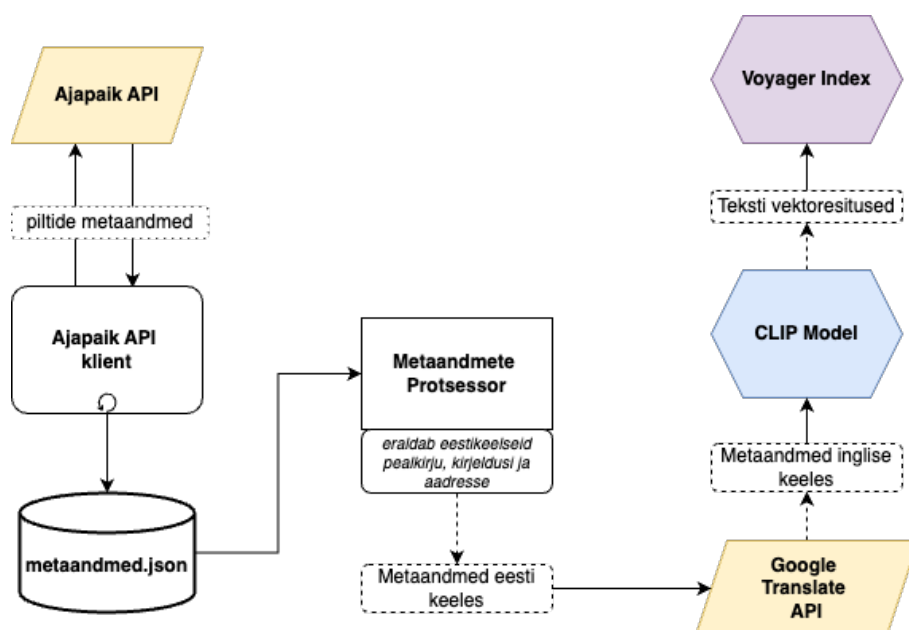
Esimese katsena arendati *lihtpildiotsingutööriista*, kasutades Annoy indeksit. Indeksile olid lisatud CLIP-i abiga saadud piltide vektoresitused, et otsida läbi esitatud pilte ja leida N arv sarnaseid pilte antud tekstpäringule või antud pildile. Kui otsingut tehakse pildi sarnasuse järgi (indeksis eksisteeriva pildi järgi), tagastatakse N lähimat naabrit pildi vektoresitustele. Kui otsingut tehakse teksti sarnasuse järgi (kasutaja poolest esitatud ingliskeelse päringu järgi), siis teksti töödeldakse kasutades CLIP-i, et saada teksti vektoresitused, mis on projitseeritud koos pildi vektoresitustega samasse vektorruumi. Otsingut teostatakse seejärel vektori sarnasuse otsingu abil. Päring peab olema teostatud inglise keeles, kuna CLIP oli eeltreenitud inglise keeles, ja mudel ei saa luua mõistetavaid vektoresitusi eestikeelsetest sõnadest.

Lihtpildiotsingu tööriista otsest järeltulijatena oli loodud *kombineeritud pildiotsingutööriist*. Kombineeritud pildiotsingut loodi kasutades kahte Voyageri indeksit: CLIP-i piltide vektoresituste indeksit ja *Ajapaik* fotode kirjelduste vektoresituste indeksit. Voyager-i teeki valiti varasemalt kasutatud Annoy teegi asemel, et uurida, kas see võiks pakkuda paremat jõudlust.

Kahe indeksi kasutamise peamine põhjus oli hüpoteesi testimine, et visuaalse otsingu laiendamine piltide kirjeldustega annaks paremaid ja täpsemaid tulemusi.

Fotode kirjelduste indeks oli rakendatud nagu kirjeldatud Joonisel 4. Esiteks koguti fotode metaandmed Ajapaik API kaudu ja salvestati JSON-failis. Seejärel töödeldi metaandmeid, eraldades eestikeelseid pealkirju, kirjeldusi ja aadresse. Metaandmeid tõlgiti Google Translate API abil inglise keelde. Viimaks loodi Voyager-i indeksit, kasutades metaandmete vektoresitused, mis oli saadud CLIP mudeli abil.

Pildiotsingul kasutati kirjelduste indeksit piltide vektoresituste indeksi kõrval. Nimetasin selle tööriista kombineeritud pildiotsinguks, kuna see võtab arvesse fotode vektoresitusi, ja rikastab otsingut, kasutades fotodega seotud metaandmeid.



Joonis 4. Ajapaik metaandmete töötlemine

3.3.1 Pildiotsingu valideerimine

Infootsingusüsteemide tõhususe mõõtmine pole lihtne. Tavaliselt nõuaks infootsingusüsteemi valideerimine testandmestikku, milles on selgelt märgitud päringule vastavad kirjed. Sellise testandmestikku saamiseks *Ajapaik* andmestikust oleks vaja lisatööd, kus tuleks pilte läbi vaadata ja neid käsitsi märkida vastavalt pildil olevatele teemadele. Kuna selline ülesanne oleks selle projekti skoobist väljaspool, olen loonud hindamissüsteemi, mis sõltub pildi metaandmetest, et pakkuda konteksti tagastatud fotodele.

Ajapaik täiendatud andmestikku kasutamiseks valideerimisandmestikuna peaks kõigepealt saama fotokirjeldused, mis on seotud fotodega andmestikus, läbitud töötamise metaandmete protsessoris ja on tõlgitud inglise keelde. Pärast seda saame kirjeldused analüüsida, jagades need sõnakottidesse – loenditesse, mis põhinevad sõnade sagedusel tekstis, arvestamata sõnade järjekorda ega konteksti. Kirjelduste segmenteerimiseks sõnakottideks kasutati spaCy teeki[27]. Valideerimisskripti jaoks valiti 8 kõige sagedamini esinevat sõna, mis hiljem olid kasutatud päringutena. Samuti märgiti ära, millistel fotodel on kirjelduses, nimes või asukohas mainitud konkreetne sõna.

Pildiotsingu hindamiseks kasutati ühte tuntumatest infootsingu tulemuste hindamise mõõdikutest *Mean Average Precision* (MAP – keskmiste täpsuste keskväärtsus), mis pakub kvaliteedimõõdet kõigi saagiste tasemete ulatuses[28].

Ühe päringu puhul arvutatakse $AP(n)$ (*Average Precision*(n) – keskmine täpsus n kohta),

mis saadakse n kogumi jaoks.

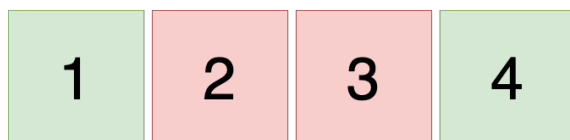
$$AP(n) = \frac{1}{R} \sum_k^n \frac{TP(k)}{k} \quad (3.1)$$

kus: $AP(n)$ = keskmine täpsus esimeste n fotode tagastatud järjekorras puhul

R = asjakohaste tulemuste koguarv

$TP(k)$ = asjakohaste tulemuste arv k -tulemuse puhul

Näidis: Pildiotsingu tööriist tagastas neli fotot, kus teine ja kolmas tagastatud foto tegelikult ei ole asjakohased (vt Joonis 5).

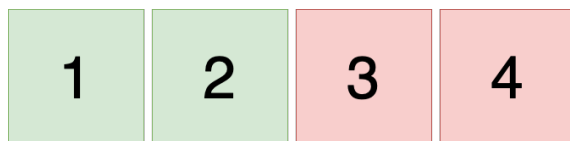


Joonis 5. Pildiotsingu tulemus, kus 2. ja 3. pildid on asjakohatud.

Keskmise täpsuse leidmiseks tagastatud nelja foto puhul tuleb teha järgmine arvutus:

$$AP(4) = \frac{1}{2} \left(\frac{1}{1} + \frac{0}{2} + \frac{0}{3} + \frac{2}{4} \right) = 0.75$$

Kui pildid oleksid tagastatud järjekorras, kus asjakohased pildid oleksid järjekorras eespool ja asjakohatud pildid oleksid järjekorras lõpus (vt Joonis 6), siis $AP(4)$ ja $AP(2)$ oleksid järgmised:



Joonis 6. Pildiotsingu tulemus, kus 3. ja 4. pildid on asjakohatud.

$$AP(4) = \frac{1}{2} \left(\frac{1}{1} + \frac{2}{2} + \frac{0}{3} + \frac{0}{4} \right) = 1$$

$$AP(2) = \frac{1}{2} \left(\frac{1}{1} + \frac{2}{2} \right) = 1$$

Saab näha, et keskmine täpsus (AP) otseselt autasustab asjakohaste piltide paigutamist järjekorra algusesse, ning ei karista asjakohatud pilte järjekorra lõpus.

Mean Average Precision (MAP) arvutatakse, tehes N päringut ja arvutades kõigi päringute kohta nende *Average Precision* (AP) keskmist:

$$MAP = \frac{1}{N} \sum_{i=1}^N AP_i$$

4. Tulemused

Tabelis 1 on toodud meetodid, mille tulemusi arutatakse alampeatükkides.

Tabel 1. Tulemuste osas käsitletud meetodite spikker

Meetod	Liik	Andmestik	Sisend	Väljund	Viide
Klassifitseerimine	Näitetu õppe mudelid	Ground truth	Pildid + kategooriate nimed	Tõenäosused	4.1.1
	CNN mudelid	Ground truth	Pildid määratud kategooriatega + kategooriate nimed	Tõenäosused	4.1.2
Lihtpilditsing	Pildi sarnasuse alusel	Ground truth	Pildi vektorsitused	N sarnaseid pilte	4.2.1
	Teksti sarnasuse alusel	Ground truth	Teksti vektorsitused	N sarnaseid pilte	4.2.1
Kombineeritud pilditsing	Teksti sarnasuse alusel	Täiendatud	Teksti vektorsitused	N sarnaseid pilte	4.2.2

4.1 Klassifitseerimine

Klassifitseerimise ülesande jaoks käivitati iga mudel asukoha (Interior/Exterior) testhulga ja kaamera nurga (Ground/Raised/Aerial) testhulgaga, saades tulemusena iga mudeli jaoks kaks klassifitseerijat. Klassifitseerimisülesande puuduste hindamiseks, iga klassifitseerijat hinnatakse täpsuse, saagise ja F-1 skooriga, mainides ka testhulga tuge.

- Täpsus - tõeliste klassi ennustuste osakaal kogu klassi ennustuste hulgas. Kui palju klassi ennustusi olid tegelikult tõesed?
- Saagis - tõeliste klassi ennustuste osakaal kogu tegelike klassi hulgas. Mitu protsenti klassihulgast suudeti õigesti ennustada?
- F-1 skoor - täpsuse ja saagise harmooniline keskmine.
- Tugi - klassi tegelike esinemiste arv andmestikus.

Iga klassifitseerija jaoks loodi ka eksimismaatriksi. Eksimismaatriks on tööriist, mis visualiseerib kui palju pilte ennustati õigesti ning kui palju pilte ennustati ekslikult teise klassi alla.

4.1.1 Näitetu õppe klassifitseerimine

CLIP

Klassifitseerijate loomiseks kasutati 3 CLIP'i mudelit: ViT-B/16, ViT-B/32 ja ViT-L/14. Pärast klassifitseerijate käitamist fototesthulkadel, hinnati klassifitseerijaid luues klassifitseerimisaruandeid (tabelid 2, 3) ja eksimismaatriseid (lisa 2).

Tabel 2. Interior/Exterior CLIP-i klassifitseerimisaruanded

	Interior			Exterior			
Mudel	Täpsus	Saagis	F1-skoor	Täpsus	Saagis	F1-skoor	Kesk. F1
ViT-B/16	0.75	0.85	0.80	0.94	0.90	0.92	0.86
ViT-B/32	0.88	0.77	0.82	0.92	0.96	0.94	0.88
ViT-L/14	0.67	0.93	0.78	0.97	0.84	0.90	0.84

Tabel 3. Ground/Raised/Aerial CLIP-i klassifitseerimisaruanded

	Ground			Raised			Aerial			
Mudel	Täpsus	Saagis	F1-skoor	Täpsus	Saagis	F1-skoor	Täpsus	Saagis	F1-skoor	Kesk. F1
ViT-B/16	0.85	0.74	0.79	0.01	0.00	0.00	0.46	0.99	0.63	0.47
ViT-B/32	0.87	0.78	0.82	0.12	0.03	0.04	0.45	0.98	0.62	0.49
ViT-L/14	0.82	0.73	0.77	0.11	0.08	0.09	0.58	0.97	0.73	0.53

Tulemusi analüüsidest saab näha, et CLIP annab väga häid tulemusi asukoha klassifitseerijana, kus parimaid tulemusi saavutab CLIP ViT-B/32 mudel. ViT-B/32 mudeli eksimismatriksi analüüsidest on selge, et mudel mõistab, millal on tegemist õues tehtud fotodega, kuid eksib siseruumides tehtud fotode ennustamisel. See võib olla tingitud sellest, et inglise keeles võib termin 'interior' olla veidi mitmetähenduslik, tähendades ka "riigi või piirkonna sisemust". Märgatavalt, CLIP'il oli ka raskusi kaamera nurga ennustamisega, kus ühtegi mudel ei suutnud ära tunda *Raised* kategooriat, st ei saanud õigesti ennustada rohkem kui 8% *Raised* fotodest. Eksimismatriksite vaadates saab öelda, et CLIP-mudelil ei ole selgelt piisavalt infot, et luua seost fotode tunnuste ja sõna *Raised* vahel, kuigi mudel suudab selgelt tuvastada *Ground* ja *Aerial* fotosid.

Kuna CLIP-i arendamise käigus oli seda treenitud mitteavalikkude internetist kogutud pildi-teksti paaridega, on võimatu ette teada, kas ja kui hästi suudab CLIP mõista mudelile antud klassinimesid. Selle leevendamiseks on kaks võimalust: häälestada CLIP-i uue pildihulgaga, et valmistada mudel ette teatud ülesande jaoks (seega töötab CLIP paremini kitsal ülesannete komplektil, kuid mudelil ilmnevad raskusi laiemate ülesannete peal), või pakkuda alternatiivseid nimesid fotokategooriatele, testides sõnakombinatsioone, mis mõistlikult sobivad CLIP-i tekstilise arusaamise konteksti. Selle hüpoteesi testimiseks selle töö raamis valiti CLIP ViT-B-16 klassifitseerijat, kuna see mudel suutis õigesti ennustada

ainult ühe *Raised* fotot (vt lisa 2). Mudel käivitati, et ennustada sama kaamera nurga testhulka, kuid nüüd klassinimedega asendatud alternatiivsete nimedega (vt Tabel 4).

Tabel 4. Erinevate klassinimede võrdlus CLIP ViT/B-16 klassifitseerimisel

Klass 1	F1-skoor	Klass 2	F1-skoor	Klass 3	F1-skoor	Kesk. F1
Ground	0.79	Raised	0.00	Aerial	0.63	0.47
Street Level	0.33	Building Level	0.28	Sky Level	0.37	0.33
Ground View	0.49	Raised View	0.31	Aerial View	0.92	0.57
Ground View	0.76	Elevated View	0.36	Bird's-eye View	0.76	0.63
Ground	0.70	Elevated View	0.45	Aerial	0.81	0.65
Ground Level	0.76	Elevated View	0.44	Aerial	0.82	0.67
Ground View	0.76	Elevated View	0.38	Aerial View	0.91	0.68

Näeme, et CLIP-i tulemused muutuvad üsna äärmuslikult, kui me defineerime klassi nime teisiti. Kategooria nimede lõppu sõna 'View' lisamine tõi *Raised* kategooria F1-skoori 0%-st 31%-le. *Raised View*'d *Elevated View*-ga asendades ennustas mudel paremini, jõudes mudeli keskmisele F1-skoorile 68%. Tulemused näitavad õigesti valitud klassinimede olulisust näitetu õppe mudelitega töötamisel. Parimate tulemuste saamiseks peab uurija panustama aega klassifitseerimiskriteeriumide täpsustamiseks.

SigLIP

Vastavalt tabelile 5 saavutasid parimad tulemused asukohakategooriates SigLip mudelid suurustega 224x224 ja 256x256, mõlemad saades üle 95% kõigist testitud *Interior* piltidest ja üle 80% kõigist testitud *Exterior* piltidest. Kuna SigLIP on CLIP-ilaadne mudel, suutis mudel teha ennustusi testandmetel ilma treenimiseta. Siiski, nagu CLIP, esines SigLIP kaamera nurga ülesandega (Tabel 6) sarnaselt halvasti *Raised* kategooria kontekstist madala arusaamise tõttu.

Tabel 5. Interior/Exterior SigLIP-i klassifitseerijad

Mudel	Interior			Exterior			Kesk. F1
	Täpsus	Saagis	F1-skoor	Täpsus	Saagis	F1-skoor	
224x224	0.64	0.96	0.77	0.98	0.81	0.89	0.83
256x256	0.66	0.95	0.78	0.98	0.83	0.90	0.84
384x384	0.76	0.61	0.68	0.87	0.93	0.90	0.79

Tabel 6. Ground/Raised/Aerial SigLIP-i klassifitseerijad

Mudel	Ground			Raised			Aerial			Kesk. F1
	Täpsus	Saagis	F1-skoor	Täpsus	Saagis	F1-skoor	Täpsus	Saagis	F1-skoor	
224x224	0.74	0.78	0.76	0.22	0.21	0.21	0.85	0.72	0.78	0.58
256x256	0.70	0.75	0.73	0.14	0.13	0.14	0.85	0.64	0.73	0.53
384x384	0.58	0.47	0.52	0.17	0.45	0.24	0.50	0.00	0.00	0.25

Tabel 7. Interior/Exterior CNN klassifitseerijad

	Interior			Exterior			
<i>Mudel</i>	<i>Täpsus</i>	<i>Saagis</i>	<i>F1-skoor</i>	<i>Täpsus</i>	<i>Saagis</i>	<i>F1-skoor</i>	<i>Kesk. F1</i>
ResNet-18	0.92	0.78	0.85	0.93	0.97	0.95	0.90
ResNet-50	0.94	0.79	0.86	0.93	0.98	0.96	0.91
DenseNet-121	0.93	0.85	0.89	0.95	0.98	0.96	0.93

Tabel 8. Ground/Raised/Aerial CNN klassifitseerijad

	Ground			Raised			Aerial			
<i>Mudel</i>	<i>Täpsus</i>	<i>Saagis</i>	<i>F1-skoor</i>	<i>Täpsus</i>	<i>Saagis</i>	<i>F1-skoor</i>	<i>Täpsus</i>	<i>Saagis</i>	<i>F1-skoor</i>	<i>Kesk. F1</i>
ResNet-18	0.85	0.96	0.90	0.72	0.40	0.51	0.86	0.85	0.85	0.75
ResNet-50	0.84	0.98	0.91	0.82	0.40	0.53	0.87	0.85	0.86	0.77
DenseNet-121	0.92	0.97	0.94	0.84	0.67	0.75	0.92	0.93	0.92	0.87

4.1.2 Treenitud CNN-ga klassifitseerimine

Asukoha ja kaamera nurga foto kategooriate klassifitseerimistulemused on esitatud tabelites 7 ja 8.

Nagu oligi oodata, konvolutsioonilised närvivõrgud toovad suurepäraseid tulemusi, aga ei võta tabel arvesse mudelite treenimiseks kulutatud aega ja ressursse. Erinevalt varasemalt analüüsitud multimodaalsetest mudelitest, mis tegutsesid otsestest ennustustel 2.5 tuhande foto testikomplektides, pidi CNN mudelid esmalt olema treenitud ülekandeõppe abil treening- ja valideerimisandmestikutel (10 tuhat fotot asukoha klassifikaatori jaoks ja 14 tuhat fotot kaamera nurga klassifikaatori jaoks). Suurepäraseid tulemusi saavutati tänu õigesti märgistatud treeningandmestikule ja pikaajalisele treenimisprotsessile, mis kumbki ei olnud vajalikud CLIP-i ja SigLIP-i kasutamiseks samale testandmestikule ennustamiseks.

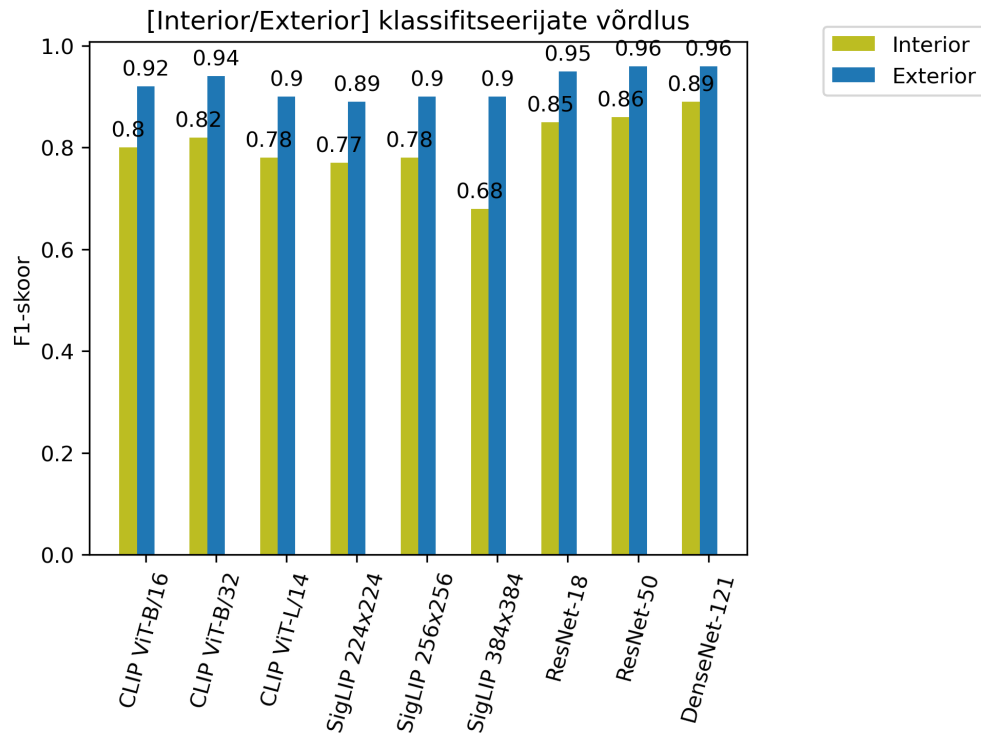
Märkimisväärne on ka see, et isegi vaadates DenseNet-121-i (ehk klassifitseerijat parimate tulemustega), suutis see ennustada vaid 67% kõigist *Raised* fotodest, andes kinnitust andmekogumi keerukusele ja fotode, mis on märgistatud kui *Raised*, mitmekesisusele.

4.1.3 Klassifitseerijate võrdlus

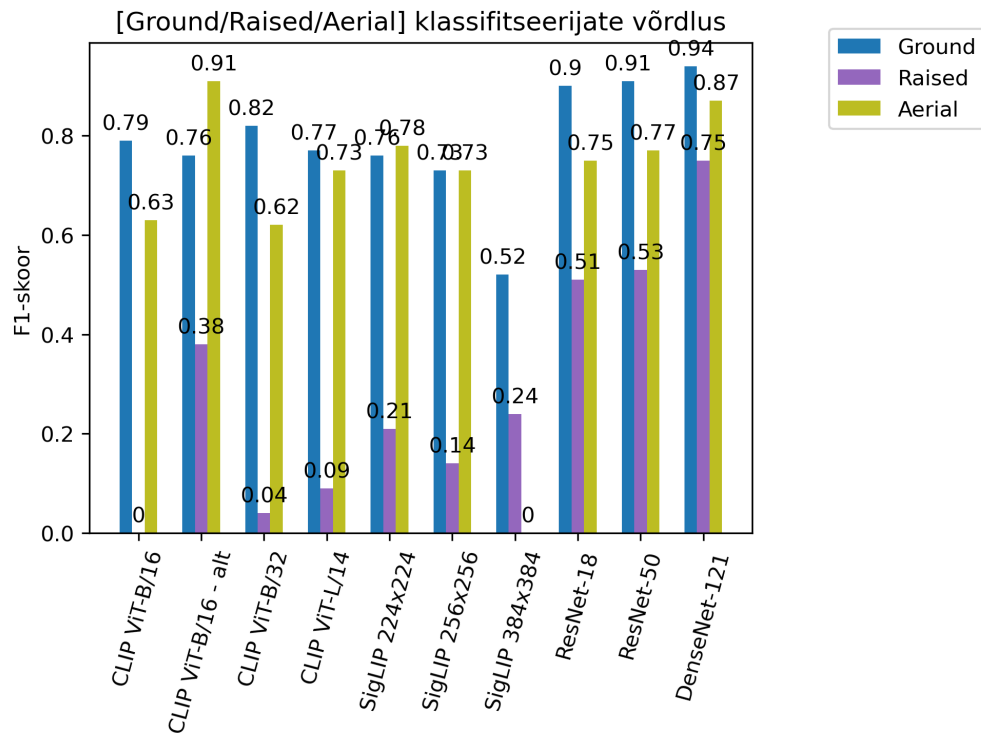
Visualiseerimise eesmärgiks on varasemalt tutvustatud klassifitseerijad koondatud tulp-diagrammidesse, mis põhinevad mudelite kategooriate F1-skooridel.

Joonisel 7 saab näha, et asukoha klassifitseerimise ülesande parim näitetu õppe multimodaalne mudel oli CLIP ViT-B/32 mudel, parim CNN mudel aga DenseNet-121. CNN

model esines napilt paremini kui näitetu õppe model, olles *Interior* F1-skoori osas 7% võrra parem.



Joonis 7. Interior/Exterior klassifitseerijate võrdlus



Joonis 8. Ground/Raised/Aerial klassifitseerijate võrdlus

Kaamera nurga klassifitseerijate tulpdiagrammile (Joonis 8) oli ka võrdluseks lisatud parim

alternatiivsete kategooriate nimedega CLIP mudel (Tabel 4) nime all "CLIP ViT-B/16 - alt". Valiti mudel, kus kategoorianimed olid asendatud sellisel viisil: Ground → Ground View, Raised → Elevated View, ja Aerial → Aerial View. Taas ilmnes, et DenseNet-121 oli täpsem klassifitseerija, kusjuures alternatiivne CLIP mudel oli kõige täpsem näitatu õppe mudelite hulgast. Kaamera nurga klassifitseerimise ülesanne osutus kõigile mudelitele väljakutseks tänu kategooria *Raised* fotode mitmetähenduslikkusele. Siiski näitab CLIP endiselt konkurentsivõimelist jõudlust *Ground* ja *Aerial* klassifitseerimise ülesannetes.

4.2 Pildiotsing

Pildiotsinguks kasutatakse näitetu õppe mudelite võimet erineva modaalsusega sisendeid projitseerida samasse vektorruumi.[14]

Osades 4.2.1 (Lihtpildiotsing) ja 4.2.2 (Kombineeritud pildiotsing) käsitletakse kvalitatiivseid katseid, osa 4.2.3 (Pildiotsingu valideerimine) tegeleb kvantitatiivsete katsetega.

4.2.1 Lihtpildiotsing

Lihtpildiotsingu tööriista rakendamisega näitan CLIP-i mõistmise sügavust ja kuidas seda saab kasutada sarnaste piltide leidmise kontekstis. Kuna tööriist töötab kohalike failidega, vajab see otsingu tegemiseks pildiandmestikku.

Kõik lihtsa pildiotsingu ülesanded teostati Ajapaiga *ground truth* pildiandmestikuga (umbes 28,5 tuhat fotot).

Pildiotsing pildi sarnasuse alusel

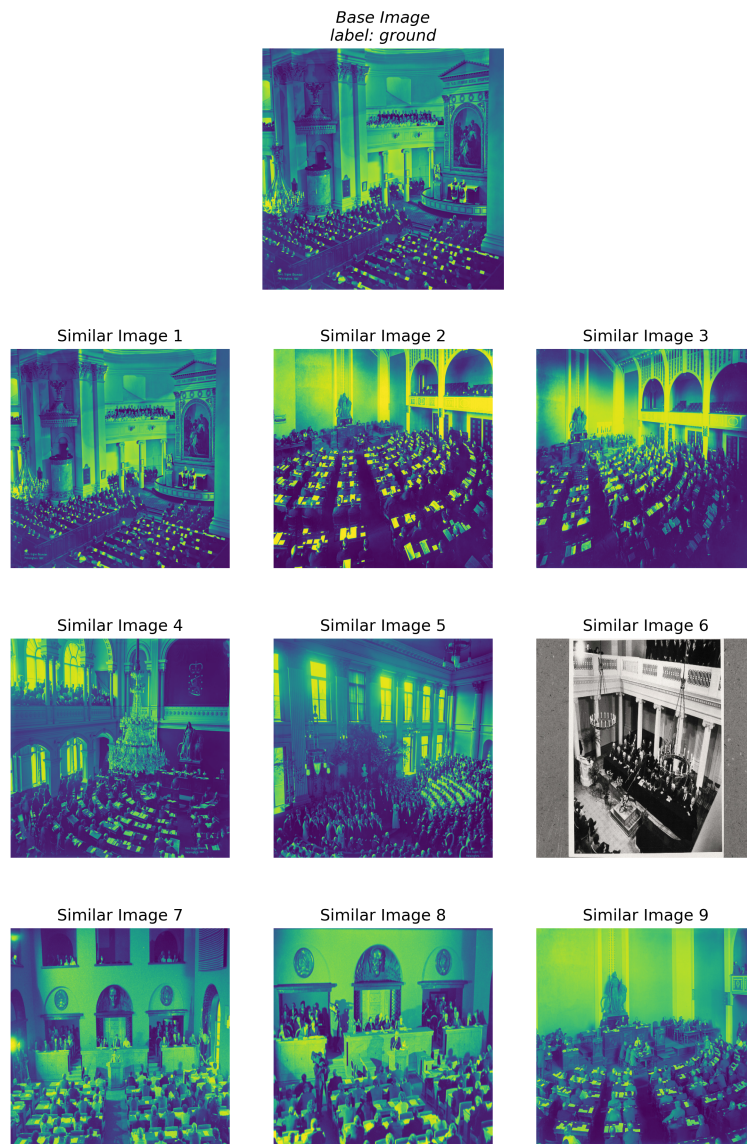
Lihtpildiotsingu tööriist oli loodud kasutades CLIP ViT-B/32-st saadud pildi vektoreksitused. Pärast multimodaalse mudeli seadmist nõuab tööriist baaspildi määratlemist pildi ID-na ning samuti kuvatavate kõige sarnasemate fotode arvu määramist.

Testpildi jaoks valiti pilt suurest inimeste kogunemisest ilusas kahekorruselises hoones. Joonisel 9 saab näha, et pildiotsingu tööriist tõi välja sarnaseid pilte suurtest inimeste rahvamassidest erinevates hoonetes. Paljud neist sisaldavad ka baaspildiga sarnaseid rõdusid.

Pildiotsing teksti sarnasuse alusel

Lihtpildiotsingu tööriist teksti sarnasuse alusel on seadistatud samamoodi nagu pildiotsing pildi sarnasuse alusel. Ainuke erinevus seisneb selles, et sisendina saab kasutaja esitada sõna või lause, ning tööriist tagastab N fotot, mis on kõige paremini kirjeldatud esitatud pealkirjaga.

Kui paluti tagastada 9 pilti, mis on paremini kirjeldatud inglise sõnaga *gathering*, tööriist tagastas pilte, mis on näha joonisel 10. Kuna *gathering* võib tõlgendada nii kui 'koosolek või kohtumine' kui ka 'millegi korjamine/kogumine', on huvitav, et seostab CLIP sõna *gathering* fotodega, kus on palju inimesi.



Joonis 9. Pildi sarnasuse alusel lihtotsing

4.2.2 Kombineeritud pildiotsing

Selle tööriista rakendamise eesmärk oli uurida, kas CLIP-i visuaalse konteksti mõistmise kombineerimine Ajapaik platvormilt võetud fotokirjeldustega võiks laiendada otsinguvaldkonda, võimaldades otsida teatud kohti ja inimesi lisaks teistele visuaalsetele kontekstidele, mida saab mõista CLIP-i pildi-teksti paaride kaudu.

Kõik kombineeritud pildiotsingu ülesanded viidi läbi Ajapaik platvormilt pärit testhulgal (umbes 10 tuhat fotot), millele olid lisatud tõlgitud inglise keelde pildikirjeldused või pealkirjad.

Kombineeritud otsingut testiti ainult teksti sarnasuse alusel, kuna kirjeldustel ei oleks pildi

Query text: gathering



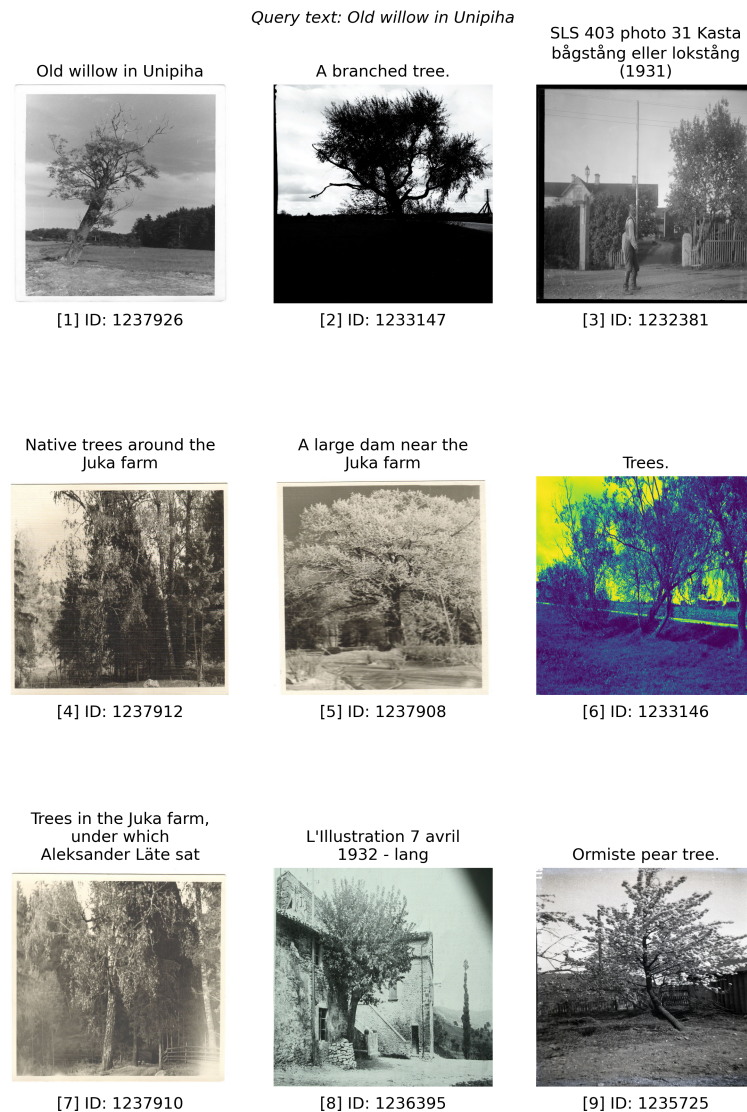
Joonis 10. Teksti sarnasuse alusel lihtotsing

sarnasuse alusel pildiotsingu suhtes mingit mõju.

Pildiotsing teksti sarnasuse alusel

Katsepildina valiti pilti vanast pajupuust kirjeldusega 'Vana paju Unipihas' ning kasutati kombineeritud otsingut täpse sama fraasi otsimiseks. Testhulga läbivaatus kinnitas, et ükski teine pildi kirjeldus ei maini puid Unipihal. Kui otsing töötab õigesti, see ei pea tagastama midagi muud kui puid.

Kombineeritud otsingu kasutamise järel saame Joonis 11. Otsing tagastas esimese pildina kasutatud kirjeldusega pilti. Ülejäänud 8 pilti ei kujuta pajusid ega Unipihat, kuid nad kujutavad suuri, okstega puid, visuaalselt väga sarnased esimese pildiga.



Joonis 11. Teksti sarnasuse alusel kombineeritud otsing

Tulemus on oodatud, kuna teatakse, et andmestikus on ainult üks foto, mis vastab päringule. Päringuga '*vana paju Unipihas*' (kus päringu fookus on puul) peaksid fotodel ideaalis olema pajud või sarnase välimusega puud. Sellega näitab otsingutööriist häid tulemusi.

4.2.3 Pildiotsingu valideerimine

Pildiotsingu valideerimiseks testiti tööriista kolmes erinevas olukorras, kasutades sõnu erineva esinemissagedusega. Seda tehes kontrolliti, kas otsingutööriist leiab kõige sagedasemaid fotosid (fotosid sagedasemate sõnadega kirjeldustes) lihtsamini.

Pildiotsingu valideerimine sagedasemate sõnapäringute järgi

Pärast Ajapaik andmestiku fotokirjelduste uurimist märgiti spaCy abil kõige sagedasemateks järgmised sõnad: *rowing* (sõudmine), *competition* (võistlus), *Tallinn*, *Soonpere* (perekonnanimi), *Iisaku*, *Tartu*, *portrait* (portree) ja *Soviet* (Nõukogude). Pärast fotode märkimist kas *asjakohasteks* või *asjakohatuteks*, teostati kombineeritud otsingut, kasutades kõige sagedasemaid sõnu päringutena. AP-d (keskmist täpsust) arvutati iga päringu jaoks. Pärast kõigi päringute lõpuleviimist arvutati MAP-d (keskmiste täpsuste keskväärtust).

Tabel 9. 8 sagedasemate sõnade päringute MAP

Päring i.k.	Päring e.k.	Tagastatud piltide arv	Asjakohaste piltide arv	AP
<i>rowing</i>	sõudmine	9	361	0.82
<i>competition</i>	võistlus	9	883	0.17
<i>Tallinn</i>	Tallinn	9	575	1.0
<i>Soonpere</i>	Soonpere	9	535	0.93
<i>Iisaku</i>	Iisaku	9	290	0.17
<i>Tartu</i>	Tartu	9	301	0.45
<i>portrait</i>	portree	9	301	1.0
<i>Soviet</i>	Nõukogude	9	561	0.0
8 päringute MAP:				0.57

Tabel 9 näitab iga päringu AP-d ja kaheksa päringute MAP-i, mis on 0.57. Alguses tundub, et see on tõend selle kohta, et päringute keskmine täpsus on kesine, ning et kombineeritud otsing väljastab asjakohatuid fotosid. Näiteks, päringu *Soviet* tulemusel polnud ühtegi asjakohast pilti – vähemalt nii näib, kui vaadata valideerimise tulemusi. Tegelike piltide uurimine (Joonis 12) annab aga teise pildi.

Nagu fototulemustest näeme, pildiotsingutööriist tagastas tegelikult asjakohaseid Nõukogude pilte, nt 1980. aasta suveolümpiamängud Moskvast (tol ajal NSV Liit), Zaporozhets auto Nõukogude Ukrainast, Oktoobrirevolutsiooni tähistamine, jne. Inimese silmale paistavad pildid asjakohastena, aga kui valideerimisprogramm saab tegelikult tegutseda ainult piltide kirjelduste põhjal, mis selles konkreetses juhtumis ei näita kusagil selgelt sõna *Soviet* – siis märgitakse pilte asjakohatuteks.

Pildiotsingu valideerimine harvemate sõnapäringute järgi

Teine valideerimiskatse viidi läbi kasutades kõige harvemaid sõnu täiendatud andmestiku kirjeldustest. SpaCy teeki kasutades, valiti 8 järgmiseid sõnu: *Rainis* (pseudonüüm) ja *Janis* (mehe eesnimi), *shipping* (laevandus), *countries* (riigid), *relations* (suhted/sugulused), *choirmaster* (koorijuht), *yarn* (lõng) ja *Aado* (mehe eesnimi). Iga sõna esineb kogu 10 tuhande pildi andmestikus vaid üks kord.



Joonis 12. Päringu *Soviet* tulemus

Pärast katse läbiviimist oli iga pildi jaoks saadud AP 0.0 ning selle tulemusel oli MAP-i väärtus loomulikult samuti 0. Iga pildi jaoks tagastas valideerimistööriist ainult ühte asjakohast pilti kogu andmestikust.

Tulemuste analüüsidest on selge, et mudelil oli suuri raskusi nimede nagu Rainis, Janis ja Aadoga. *Rainis* ja *Aado* päringud tagastasid asjakohastuid fotosid inimestest ja linnadest. Päring *Janis* tagastas naiste fotosid, tõenäoliselt seetõttu, et Janis võib-olla interpreteeritud kui inglise nimi Janice, mis on naisenimi. Päring *relations* tagastas laste ja pere pilte, visuaalselt päringuga seotud. Päring *lõng* tõi esile pilte lõnganäitustest Eesti Kunstimuuseumis

(1920. aastatel Tallinna Eesti Muuseum) ning mõningaid pilte kootud rõivastest.

Tulemused päringutele *Aado*, *Janis*, *relations* ja *yarn* on esitatud Lisas 3.

Pildiotsingu valideerimine juhuslikult valitud sõnapäringute järgi

Kolmas valideerimiskatse viidi läbi kasutades juhuslikult valitud (SpaCy teegi abil) sõnapäringuid: *battlefield* (lahinguväli), *teahouse* (teemaja), *Riga* (Riia), *school* (kool), *Sigulda* (linn Lätis), *Bauskas* (linn Lätis), *Fyodor* (vene mehenimi) ja *revolution* (revolutsioon). Tabel 10 näitab iga päringu AP-d ja kaheksa päringute MAP-i, mis on 0.15.

Tabel 10. 8 juhuslikult valitud sõnade päringute MAP

Päring i.k.	Päring e.k.	Tagastatud piltide arv	Asjakohaste piltide arv	AP
<i>battlefield</i>	lahinguväli	9	1	0.0
<i>teahouse</i>	teemaja	9	1	0.0
<i>Riga</i>	Riia	9	114	0.0
<i>school</i>	kool	9	232	0.0
<i>Sigulda</i>	Sigulda	9	86	1.0
<i>Bauskas</i>	Bauskas	9	7	0.0
<i>Fyodor</i>	Fjodor	9	1	0.0
<i>revolution</i>	revolutsioon	9	22	0.17
8 päringute MAP: 0.15				

Katset analüüsid näeme, et ainult kahel päringul on positiivne keskmine täpsus: 'Sigulda' ja 'revolutsioon'.

'Sigulda' päringust tagastatud pildid osutusid Sigulda ordulinnu ja Sigulda uue lossi piltideks ning revolutsiooni piltidel olid paraadid ja mälestusmärgid. Tulemuste vaatamisel pole selget seost andmestikus olevate asjakohaste piltide hulga ja õigesti tagastatud piltide hulga vahel. Seetõttu uurime suvaliselt valitud tulemusi, mis tagastasid 0.0 keskmist täpsust. Lisasse 4 on lisatud tulemused kolmale päringule - *battlefield* (1 asjakohane foto kogu andmestikus), *Riga* (114 asjakohast fotot) ja *Fyodor* (1 asjakohane foto).

Vaadates tulemusi päringule *battlefield* (Joonis 19), näeme, et esimesed kolm ja viimane tagastatud fotod kujutavad sõdureid. Ülejäänud fotod kujutavad viljatuid, tühje rohumaid. Kuigi fotod näivad visuaalselt asjakohastena, ei maini ükski fotode kirjeldus lahinguvälja, seega on AP 0.0.

Joonisel 20 (päring *Riga*) on näha CLIP-i pildiotsingu piiranguid. 9 tagastatud fotost 8 on tegelikult Tallinnast. Selgelt mõistis mootor, et Riia on linna nimi, kuid tal oli raskusi täpse asukoha väljaselgitamisega. Huvitava detailina on ainus mitte-Tallinna foto mehest nimega Mart Latvia. Eestikeelse kirjelduse põhjal on mehe tegelik nimi Mart Lätt, mida valesti

tõlgiti tõlgiteegi abil Mart *Latviaks* (Mart *Lätiks*). On ilmne, et CLIP-il on arusaam, et Riia on seotud Lätiga, vastasel juhul ei oleks 'Mart Latvia' sarnasuse indeksis nii kõrge.

Joonisel 21 (päring *Fyodor*) näeme, et CLIP-i pildiotsing ei tagastanud ühtegi pilti inimestest nimega Fjodor – tagastas see aga ainult fotosid meestest, tõestades, et CLIP-il on arusaam, et Fjodor on mehe nimi.

Pildiotsingu valideerimise järeldus

Pildiotsingu tulemusi on raske valideerida kvantitatiivselt. Fotode kirjeldused, mis on määratletud Ajapaik platvormi kasutajate abil, on mõnikord vaid üks sõna, mis kirjeldab terve pilti. Nagu oli näidatud 4.2.3 osas, on sageli pildiotsinguga saadud pilt visuaalselt asjakohane, kuid pildi kirjelduses endas otsinguõna ei mainita. Piltide märgistamine asjakohasteks või asjakohatuteks ilma saadud piltide käsitsi läbivaatamiseta on keeruline.

MAP-mõõt kipub alahindama mudeli jõudlust, kuna enamasti loetakse testides asjakohaseid pilte mitteasjakohasteks, aga mitte vastupidi. Sellega andis MAP sagedasemate sõnadega positiivseid tulemusi, ja ei saanud korralikult hinnata harvemate sõnadega otsingut. Arvestades mõõdu puudustega, MAP-i saab kasutada ligilähedase täpsuse hindamiseks, ja ei soovitata kasutada jõudluse täiusliku hindamismõõduna.

Tulemusi analüüsid, on selge, et isegi juurdepääsuga fotode kirjeldustele pole vektorite sarnasuse alusel pildiotsing täiuslik. Pildiotsingutööriistaga oli raske saada asjakohaseid tulemusi linnade ja inimeste nimede päringutega. Üllatavalt näitas tööriist arusaamist sellest, millised nimed loeti naise- või mehenimedeks.

On oluline meeles pidada, et otsingutööriist põhineb ligikaudsete naabrite otsimisel, mis ei taga, et kõige asjakohasemad fotod oleksid järjekorra alguses.

5. Kokkuvõte

See töö demonstreerib näitetu õppe multimodaalsete mudelite potentsiaali ajalooliste fotode klassifitseerimise ja pildiotsingu ülesannetes. Arendasin ja võrdlesin näitetu õppe klassifitseerijaid tavapärase CNN-klassifitseerijate kõrval, tõestades, et näitetu õppe mudelid pakuvad konkurentsivõimelist jõudlust. Olen samuti näidanud näitetu õppe multimodaalsete mudelite puudusi, nagu madalam täpsus mitmeklassilise klassifitseerimise korral. Näitetu õppe mudelid võivad toimida halvemini ebamääraselt määratletud ülesannetega, seetõttu on oluline nimetada kategooriad õigesti. Mudelid tõestasid end konkurentsivõimelistena selgelt määratletud, binaarse klassifitseerimise korral.

Lisaks olen loonud pildiotsingu tööriista, kasutades näitetu õppe mudeli CLIP-i võimet välja tõmbama vektoretsusi antud piltidest. CLIP-i kontekstuaalse mõistmise rakendamine võimaldab efektiivset pildiotsingut ilma täiendavate metaandmeteta. Kui aga lisatakse ka fotode metaandmeid vektorruumi, parandab see otsingut, tagades, et kõige asjakohasemad fotod jäävad N kõige sarnasemate fotode järjekorra algusse. Sellega olen näidanud, et CLIP ei vaja peenhäälestamist selle võimete laiendamiseks pildiotsingu ülesande jaoks. Siiski on pildiotsingul ka puudusi: isegi juurdepääsuga fotokirjeldustele ei saa pildiotsinguga saada asjakohaseid tulemusi inimeste nimede ja geograafiliste asukohade päringutest.

Kuna CLIP mudel on arendatud ja eeltreenitud inglise keeles, pidin selles töös kõik eestikeelsed tekstid inglise keelde tõlkima, et CLIP saaks luua seoseid tekstide ja piltide vahel. Sellele lahenduseks võiks olla treenida eesti keele CLIP-i rakendus suuremahulise eestikeelsete pildi-teksti paaride andmestikuga, tagades, et mudel sisaldaks eesti keele kultuurilisi semantilisi omadusi.

Kasutatud kirjandus

- [1] Ajapaik. <https://ajapaik.ee/>. Kasutatud: 2024.04.11.
- [2] Thomas Smits ja Melvin Wevers. „A multimodal turn in Digital Humanities. Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections“. *Digital Scholarship in the Humanities* 38.3 (märts 2023), lk. 1267–1280. ISSN: 2055-7671. DOI: 10.1093/llc/fqad008. eprint: <https://academic.oup.com/dsh/article-pdf/38/3/1267/51309490/fqad008.pdf>. URL: <https://doi.org/10.1093/llc/fqad008>.
- [3] Alec Radford *et al.* *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].
- [4] Yongqin Xian *et al.* *Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly*. 2020. arXiv: 1707.00600 [cs.CV].
- [5] Xiaohua Zhai *et al.* *Sigmoid Loss for Language Image Pre-Training*. 2023. arXiv: 2303.15343 [cs.CV].
- [6] An Yang *et al.* *Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese*. 2023. arXiv: 2211.01335 [cs.CV].
- [7] Byungsoo Ko ja Geonmo Gu. *Large-scale Bilingual Language-Image Contrastive Learning*. 2022. arXiv: 2203.14463 [cs.CV].
- [8] S.J.D. Prince. *Computer Vision: Models Learning and Inference*. Cambridge University Press, 2012.
- [9] Yann LeCun, Y. Bengio ja Geoffrey Hinton. „Deep Learning“. *Nature* 521 (mai 2015), lk. 436–44. DOI: 10.1038/nature14539.
- [10] Pemi Nguyen. *Introduction to Machine Learning (Non-Majors)*. 2022. URL: https://courses.cs.washington.edu/courses/cse416/24sp/resources/CSE_416_Notes.pdf.
- [11] Simon J.D. Prince. *Understanding Deep Learning*. The MIT Press, 2023. URL: <http://udlbook.com>.
- [12] Ashish Vaswani *et al.* *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].
- [13] Yoon Kim *et al.* *Structured Attention Networks*. 2017. arXiv: 1702.00887 [cs.CL].

- [14] Dan Bogdanov *et al.* *Tehisintellekti ja masinõppe tehnoloogia riskide ja nende leevendamise võimaluste uuring*. Tehniline raport D-16-377. Mäealuse 2/1, 12618 Tallinn, Estonia: Cybernetica AS, veebruar 2024, lk. 111. URL: <https://ria.ee/sites/default/files/documents/2024-03/Tehisintellekti-masinoppe-tehnoloogia-riskide-uuring-2024.pdf>.
- [15] Davide Caffagni *et al.* *The (R)Evolution of Multimodal Large Language Models: A Survey*. 2024. arXiv: 2402.12451 [cs.CV].
- [16] Jacob Devlin *et al.* *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].
- [17] Tom B. Brown *et al.* *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].
- [18] Alexey Dosovitskiy *et al.* *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].
- [19] Alec Radford *et al.* *CLIP: Connecting text and images*. Kasutatud: 2024.05.03. Jaanuar 2021. URL: <https://openai.com/index/clip>.
- [20] Jia Deng *et al.* „ImageNet: A large-scale hierarchical image database“. Teoses: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, lk. 248–255. DOI: 10.1109/CVPR.2009.5206848.
- [21] Kaiming He *et al.* *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].
- [22] Gao Huang *et al.* *Densely Connected Convolutional Networks*. 2018. arXiv: 1608.06993 [cs.CV].
- [23] Xi Chen *et al.* *PaLI: A Jointly-Scaled Multilingual Language-Image Model*. 2023. arXiv: 2209.06794 [cs.CV].
- [24] Olga Russakovsky *et al.* „ImageNet Large Scale Visual Recognition Challenge“. *International Journal of Computer Vision (IJCV)* 115.3 (2015), lk. 211–252. DOI: 10.1007/s11263-015-0816-y.
- [25] Erik Bernhardsson. *Annoy*. <https://github.com/spotify/annoy>. Kasutatud: 2024.04.25.
- [26] *Voyager*. <https://github.com/spotify/voyager>. Kasutatud: 2024.04.25.
- [27] *spaCy – Industrial-Strength Natural Language Processing in Python*. <https://spacy.io>. Kasutatud: 2024.05.08.
- [28] Christopher D. Manning, Prabhakar Raghavan ja Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks¹

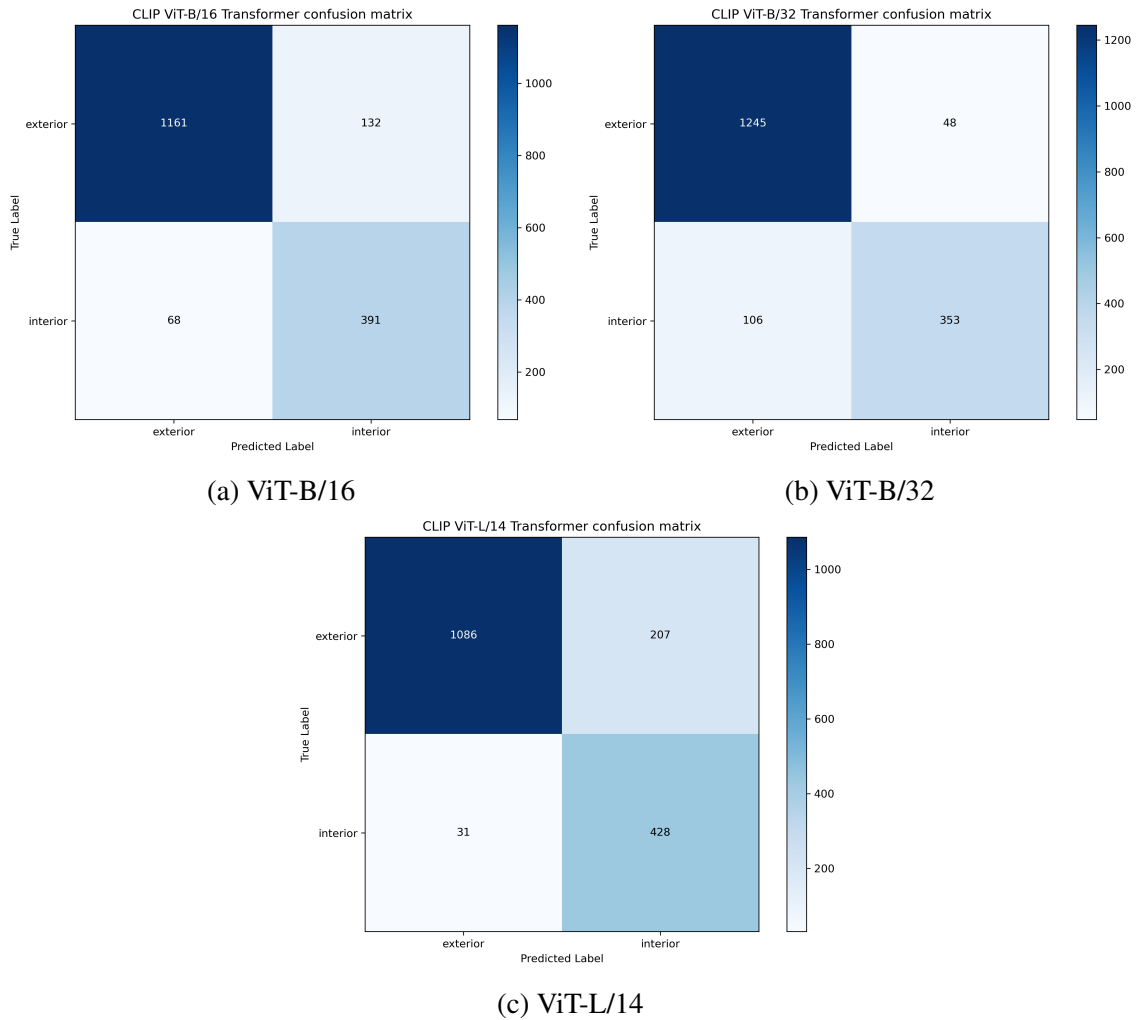
Mina, Erika Maksimova

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose “Multimodaalsete mudelite potentsiaali uurimine ajalooliste piltide klassifitseerimisel”, mille juhendaja on Priit Järv
 - 1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
 - 1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

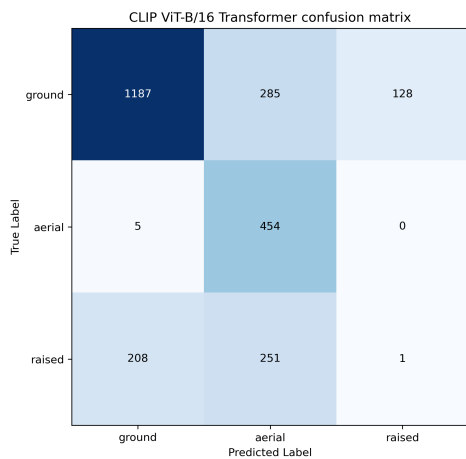
25.05.2024

¹Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingu tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtjaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.

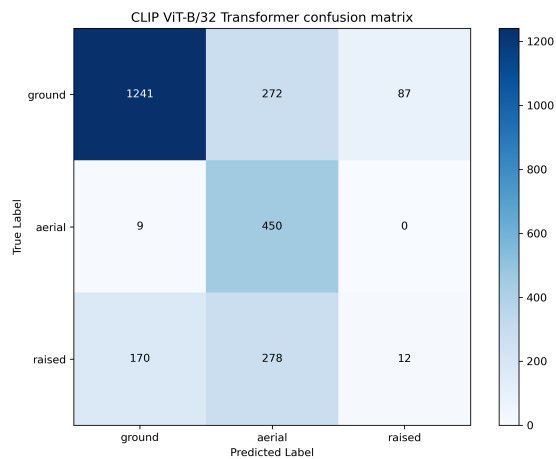
Lisa 2 – CLIP-i eksimismatriksid



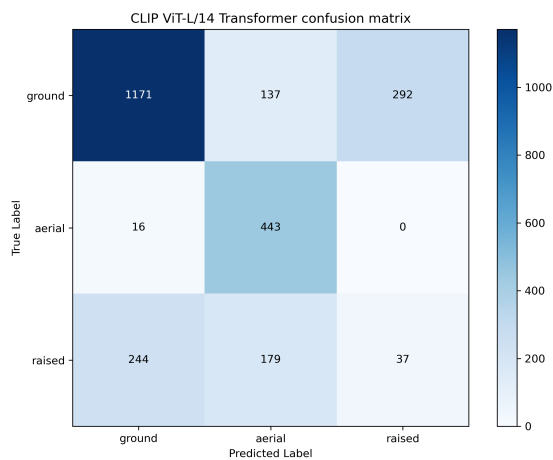
Joonis 13. CLIP Interior/Exterior eksimismatriksid



(a) ViT-B/16



(b) ViT-B/32



(c) ViT-L/14

Joonis 14. CLIP Ground/Raised/Aerial eksimismaatrisid

Lisa 3 – Pildiotsing harvamate sõnapäringute järgi

Query text: *Aado*

L. Soonpere's daughter
Reet.



[1] ID: 1234475

Villa Johannesberg,
Helsinki Old Town



[2] ID: 1230975

SLS 407 b photo 29
Slunga (1932)



[3] ID: 1232427

L. Soonpere's daughter
Reet in the Stroom
forest.



[4] ID: 1234519

Blank frame



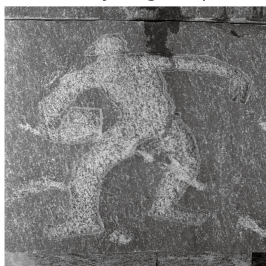
[5] ID: 1238056

Kesklinn, 51004 Tartu,
Estonia



[6] ID: 1231932

Monument to Kristjan
Raua by Gagarin pst.



[7] ID: 1237004

Kassar's chapel.



[8] ID: 1235495

Laupa manor house.



[9] ID: 1234582

Joonis 15. Päringu *Aado* tulemus

Query text: *Janis*

A woman.



[1] ID: 1234014

A girl.



[2] ID: 1234452

A girl.



[3] ID: 1234081

A girl.



[4] ID: 1234279

A girl.



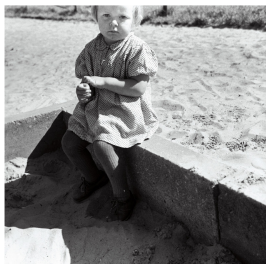
[5] ID: 1234280

A girl.



[6] ID: 1234334

A girl.



[7] ID: 1234335

A girl.



[8] ID: 1234336

TSK control exercise at the node.



[9] ID: 1235391

Joonis 16. Päringu *Janis* tulemus

Query text: *relations*

Children.



[1] ID: 1234248

Children.



[2] ID: 1234269

Children.



[3] ID: 1234377

Juidik Soonpere's
daughter Reeda, a
family friend of L.
Soonpere.



[4] ID: 1234123

Customer reference.



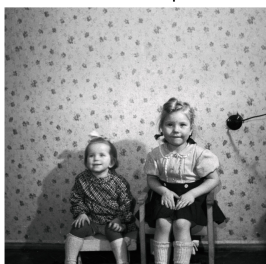
[5] ID: 1233832

SLS 397 b photo 69 Dra
tumsticka (-kil) (1929)



[6] ID: 1232681

Children sitting, on the
left Reet, daughter of
communications worker
Lembit Soonpere.



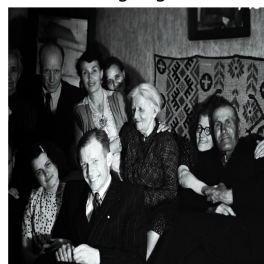
[7] ID: 1234101

Children.



[8] ID: 1234853

Getting together.



[9] ID: 1233983

Joonis 17. Päringu *relations* tulemus

Query text: yarn

Exhibition of the Tallinn
Estonian Museum



[1] ID: 1237599

Virk.



[2] ID: 1233804

TEM exhibition in
Raikküla



[3] ID: 1237558

TEM exhibition in
Raikküla



[4] ID: 1237562

TEM exhibition in
Raikküla



[5] ID: 1237559

A child.



[6] ID: 1234453

Kreenholm History
Museum. Exhibition
Headwear



[7] ID: 1236194

L. Soonpere's wife Esta.



[8] ID: 1234868

photos of Jüri Kurul's
family and friends.



[9] ID: 1240612

Joonis 18. Päringu *yarn* tulemus

Lisa 4 – Pildiotsing juhuslikult valitud sõnapäringute järgi

Query text: *battlefield*

Soldiers.



[1] ID: 1233544

Soldiers.



[2] ID: 1233545

Soldiers.



[3] ID: 1233546

Side Republican Cross
Country Competitions.



[4] ID: 1235379

Haystacks in Rae
municipality Lehmja.



[5] ID: 1233202

Nõmmerga lauter
place, view from the
land to the sea. A boat
cart in the foreground,
a boat on a cart in the
back left.



[6] ID: 1238402

On the castle ruins.



[7] ID: 1234755

Mountain.



[8] ID: 1235698

A soldier.



[9] ID: 1233542

Joonis 19. Päringu *battlefield* tulemus

View of the Old Town
and Victory Square
from the upper
windows of Kiek in de
Kök.



[1] ID: 1237028

Query text: Riga

Tallinn



[2] ID: 1237032

Tallinn



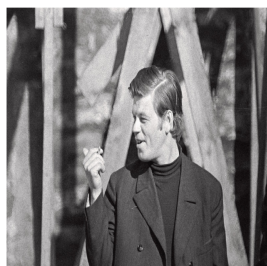
[3] ID: 1237033

Tallinn



[4] ID: 1237034

Tallinn



[5] ID: 1237035

Tallinn



[6] ID: 1237036

Tallinn



[7] ID: 1237037

Mart Latvia



[8] ID: 1232996

View of the Old Town
and Victory Square
from the upper
windows of Kiek in de
Kök.



[9] ID: 1237029

Joonis 20. Päringu *Riga* tulemus

Query text: *Fyodor*

(Military) Nemegalov.



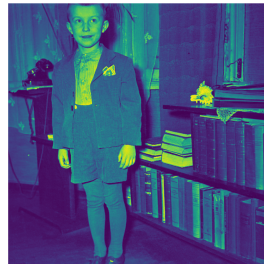
[1] ID: 1233344

(Military) Poblichenko.



[2] ID: 1233346

Boy.



[3] ID: 1234517

Max Läte, assistant
headmaster of
industrial high school



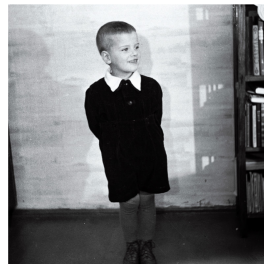
[4] ID: 1232995

Boy.



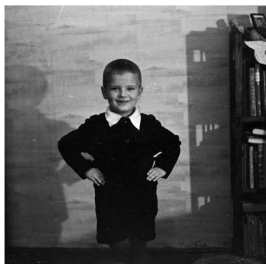
[5] ID: 1234244

Boy.



[6] ID: 1234245

Boy.



[7] ID: 1234246

Boy.



[8] ID: 1234247

Boy.



[9] ID: 1234271

Joonis 21. Päringu *Fyodor* tulemus