Usporedba algoritama umjetne inteligencije za generiranje slika

Sobotinčić, Leon

SVEUČILIŠTE U RIJECI

POMORSKI FAKULTET

LEON SOBOTINČIĆ

USPOREDBA ALGORITAMA UMJETNE INTELIGENCIJE ZA GENERIRANJE SLIKA

DIPLOMSKI RAD

Rijeka, 2024.

SVEUČILIŠTE U RIJECI

POMORSKI FAKULTET

USPOREDBA ALGORITAMA UMJETNE INTELIGENCIJE ZA GENERIRANJE SLIKA

COMPARISON OF ARTIFICIAL INTELLIGENCE IMAGE GENERATION ALGORITHMS

DIPLOMSKI RAD
MASTER THESIS

Kolegij: Umjetna inteligencija

Mentor: doc. dr. sc. Dario Ogrizović

Komentor: David Bačnar, mag. ing. el.

Student: Leon sobotinčić

Studijski program: Elektroničke i informatičke tehnologije u pomorstvu

JMBAG: 0112075035

Rijeka, rujan 2024.

Student: Leon Sobotinčić

Studijski program: Elektroničke i informatičke tehnologije u pomorstvu

JMBAG: 0112075035

IZJAVA O SAMOSTALNOJ IZRADI DIPLOMSKOG RADA

Kojom izjavljujem da sam diplomski rad s naslovom

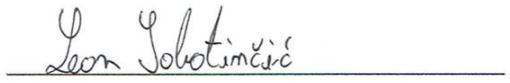
USPOREDBA ALGORITAMA UMJETNE INTELIGENCIJE

ZA GENERIRANJE SLIKA

izradio samostalno pod mentorstvom prof. dr. sc. Dario Ogrizović te komentorstvom David Bačnar, mag. el. ing.

U radu sam primijenio metodologiju izrade stručnog rada i koristio literaturu koja je navedena na kraju diplomskog rada. Tuđe spoznaje, stavove, zaključke, teorije i zakonitosti koje sam izravno ili parafrazirajući naveo u diplomskom radu na uobičajen, standardan način citirao sam i povezao s fusnotama i korištenim bibliografskim jedinicama, te nijedan dio rada ne krši bilo čija autorska prava. Rad je pisan u duhu hrvatskoga jezika.

Student



Leon Sobotinčić

Student: Leon Sobotinčić

Studijski program: Elektroničke i informatičke tehnologije u pomorstvu

JMBAG: 0112075035

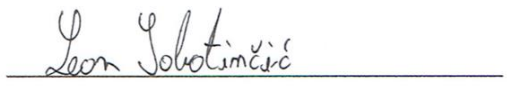
IZJAVA STUDENTA – AUTORA

O JAVNOJ OBJAVI OBRANJENOG DIPLOMSKOG RADA

Izjavljujem da kao student – autor diplomskog rada dozvoljavam Pomorskom fakultetu Sveučilišta u Rijeci da ga trajno javno objavi i besplatno učini dostupnim javnosti u cjelovitom tekstu u mrežnom digitalnom repozitoriju Pomorskog fakulteta.

U svrhu podržavanja otvorenog pristupa diplomskim radovima trajno objavljenim u javno dostupnom digitalnom repozitoriju Pomorskog fakulteta, ovom izjavom dajem neisključivo imovinsko pravo iskorištavanja bez sadržajnog, vremenskog i prostornog ograničenja mog diplomskog rada kao autorskog djela pod uvjetima Creative Commons licencije CC BY Imenovanje, prema opisu dostupnom na http://creativecommons.org/licenses/

Leon Sobotinčić – autor



SAŽETAK

U ovom diplomskom radu opisan je rad različitih modela Stable Diffusion-a kao i tehnologije koje su potakle razvoj istih modela. Opisana je kratka povijest razvoja tih modela te tehnologije zaslužne za njihov razvoj. Svaki od modela je ispitan kroz niz različitih testiranja od brzine generiranja slika do kvalitete pojedinih stilova izgleda slika te su potom međusobno uspoređeni. Testiran je rad tih modela gdje pokušavaju replicirati stil crtanja poznatih umjetnika kao i mogućnost generiranja fotorealističnih slika koje mogu zavarati ljude da su original.

Ključne riječi: Checkpoint, LoRA, Stable Diffusion, umjetna inteligencija, šum.

SUMMARY

This master thesis describes the work of different Stable Diffusion models as well as the technologies that stimulated the development of the same models. A brief history of the development of these models and the technology responsible for their development is described. Each of the models was tested through a series of different tests, from the speed of image generation to the quality of individual image appearance styles, and then they were compared to each other. The work ability of these models was tested, where they try to replicate the drawing style of famous artists, as well as the possibility of generating photorealistic images that can deceive people that they are original.

Key words: Checkpoint, LoRA, Stable Diffusion, artificial intelligence, noise.

SADRŽAJ

1. UVOD

Grana umjetne inteligencije koja se bavi obradom i generiranjem slika je u zadnjih nekoliko godina znatno napredovala. Ovaj diplomski rad se bavi poučavanjem i istraživanjem različitih novonastalih modela za generiranje slika kao i tehnologija koje su doprinijele razvoju navedenih modela. Provedeno je istraživanje u obliku ankete sa svrhom da prikaže može li čovjek raspoznati rad modela od rada samog čovjeka.

1.1. PROBLEM, PREDMET I OBJEKTI ISTRAŽIVANJA

Na osnovi relevantnih činjenica o problematici znanstvenoga istraživanja može se definirati problem istraživanja: Veliki ubrzani rast umjetne inteligencije otvorio je mnoga polja istraživanja o načinu rada i mogućnosti novih tehnologija. Nova vrsta strojnog učenja uključila je rad generiranja slika koristeći novonastale tehnologije obrade slika. Ova nova tehnologija je otvorila vrata puno mogućnosti ali u isto vrijeme i probudila pobunu autentičnosti te pitanja moraliteta. Trenutni modeli generiranja slika predstavljaju problem kod sigurnosti širenja krivih i zloćudnih informacija te se postavlja pitanje o njihovoj kvaliteti rada.

Relevantne spoznaje o problematici i problemu istraživanja predstavljaju znanstvenu podlogu za definiranje predmeta istraživanja: Istražiti način rada modela i tehnologija koje koriste. Istražiti sve mogućnosti koje imaju ulogu u generiranju slika kao i kvalitetu istih. Ispitati do koje je granice ova tehnologija napredovala te pruža li ona problem za buduće generacije.

Problem i predmet istraživanja odnose se na dva međusobno povezana objekta istraživanja, i to: Umjetna inteligencija i modeli za generiranje slika.

1.2. RADNA HIPOTEZA

Sukladno bitnim odrednicama problema, predmeta i objekta istraživanja postavljena je radna hipoteza: Rezultati ispitivanja rada modela generiranja slika s pomoću umjetne inteligencije otvaraju širu sliku u razumijevanju i istraživanju grane umjetne inteligencije te služe kao koristan alat.

1.3. SVRHA I CILJEVI ISTRAŽIVANJA

Svrha i ciljevi istraživanja u ovom diplomskom radu očituju se u sljedećemu: Upoznati se s radom modela za generiranje slika. Upoznati se s novim tehnologijama koje koriste i usporediti se one međusobno razlikuju. Ispitati kvalitetu rada svih modela i njihovu kvalitetu generiranja slika. Ispitati do koje granice je ova tehnologija napredovala na način da se pokuša generirati fotorealistična slika koja može zavarati čovjeka da se smatra da je prava, te može li model biti u mogućnosti generirati slike kao poznati umjetnici.

Ciljevi diplomskog rada će dati odgovor na sljedeća pitanja:

- Što je Stable Diffusion?
- Kako rade modeli generiranja slika?
- Koji je od modela najbolji u svom radu?
- Koje se sve nove tehnologije koriste u radu ovih modela?
- Je li moguće zavarati čovjeka sa slikom koju je generirao jedan od modela?

1.4. ZNANSTVENE METODE

Prilikom istraživanja, formuliranja i predstavljanja rezultata istraživanja korištene su u odgovarajućim kombinacijama sljedeće znanstvene metode: metoda analize i sinteze, metoda indukcije i dedukcije, metoda apstrakcije i konkretizacije, metoda specijalizacije i generalizacije, metoda dokazivanja i opovrgavanja, statistička metoda, povijesna metoda, komparativna metoda, metoda klasifikacije, metoda deskripcije, metoda kompilacije, metoda anketiranja.

1.5. STRUKTURA RADA

U prvom dijelu, "Uvod" , navedeni su problem, predmet i objekt istraživanja, radna hipoteza i pomoćne hipoteze, svrha i ciljevi istraživanja, znanstvene metode i obrazložena je struktura rada. Naslov drugog dijela rada je "Generiranje slike pomoću umjetne inteligencije". U tome dijelu rada analiziran je rad modela generiranja slika zvan Stable diffusion i opisan je njegov princip rada generiranja slika kao i kratka povijest o njegovom razvoju. "Stable Diffusion" je naslov trećeg dijela rada. U tom dijelu predočeni su rezultati istraživanja rada različitih modela Stable diffusion-a kao i usporedba njihove kvalitete rada na ančin da se ocjeni njihova brzina i kvaliteta slika generiranih u različitim stilovima. U četvrtom dijelu rada s naslovom "Testiranje granice modela za generiranje slike baziranim na umjetnoj inteligenciji" se ispituje do koje granice su modeli usavršeni provodeći testiranje gdje se ispituje ljude da otkriju koja slika je generirana pomoću modela a koja je stvarna. U posljednjem dijelu Zaključku, dana je sinteza rezultata istraživanja kojima se dokazuje postavljena radna hipoteza.

2. GENERIRANJE SLIKE POMOĆU UMJETNE INTELIGENCIJE

U ovom poglavlju će se opisati kratka povijest modela za generiranje slika i kako je ideja o njima nastala. Nakon toga će se detaljno opisati rad modela Stable diffusion koji će biti glavna tema ovog istraživanja.

2.1. POVIJEST MODELA ZA GENERIRANJE SLIKA

Najraniji pokušaji generiranja slika pomoću umjetne inteligencije su se pojavili kasnih 1960-tih godina, s prvim značajnim sustavom koji se pojavio 1973. godine pod nazivom Aaron, koji je razvio Harold Cohen. Sustav Aaron je pomoću umjetne inteligencije pomogao Cohen-u da stvori crno-bijele umjetničke crteže. Međutim, situacija se počela mijenjati s porastom dubinskog učenja i konvolucijskih neuronskih mreža, koje su zauzvrat pružile temelj za GAN (Generativne suparničke mreže) (engl. Generative Adversarial Network).

Slika 1. Jedna od slika generirana od strane sustava Aaron iz 1980. godine

Izvor: Computer history museum, https://computerhistory.org/blog/harold-cohen-and-aaron-a-40-year-collaboration/

GAN-ovi su označili značajan napredak u području generiranja slika s umjetnom inteligencijom. Ovu arhitekturu neuronske mreže su 2014. godine razvili Ian Goodfellow i njegovi kolege sa Sveučilišta u Montreal-u. GAN se sastoji od dvije neuronske mreže koju čine generator i diskriminator, koji su pritom trenirani u izmjeničnim periodima. Generatorska mreža uči generirati sintetičke podatke koji oponašaju stvarne podatke, dok diskriminatorska mreža uči razlikovati sintetičke podatke koje generira generator od stvarnih podataka.

Međutim, iako su GAN-ovi pokazali impresivne rezultate u stvaranju realističnih slika, postoji nekoliko ograničenja ove tehnologije. Jedno od tih ograničenja jest nestabilna priroda procesa treniranja. Suparnička struktura mreže može dovesti do degradacije u načinu rada, gdje generatorska mreža proizvodi samo ograničen skup slika koje ne pokrivaju cijeli raspon mogućih rezultata, što rezultira nedostatkom raznolikosti u izlaznim slikama.

Kako bi prevladali ta ograničenja, istraživači su nastavili istraživati nove tehnike i arhitekture za generiranje slika. Jedan takav primjer je DALL-E, koji je razvila tvrtka OpenAI 5. siječnja 2021. godine. Model je koristio generativne prethodno-trenirane transformatore (GPT, engl. Generative pre-trained transformer), koji se temelje na ranijem modelu transformatora. Oba su izvorno razvijena za upotrebu u obradi prirodnog jezika (engl. Natural Language).

Model GPT je arhitektura neuronske mreže koja se temelji na mehanizmu pozornosti na sebe. U tradicionalnim neuronskim mrežama svaki se ulazni element obrađuje neovisno, što može dovesti do poteškoća u modeliranju dugotrajnih ovisnosti. Mehanizam pozornosti na sebe omogućuje modelu da se selektivno fokusira na različite dijelove ulaznih podataka, omogućujući mu da uhvati složene odnose između riječi. GPT model se sastoji od kodera i dekodera, a oba su sastavljena od višestrukih slojeva neuronskih mreža pozornosti na sebe i povratnih informacija. Na temelju te arhitekture, GPT je razvila tvrtka OpenAI 2018. godine. GPT funkcionira tako da obučava veliku neutralnu mrežu na ogromnim količinama tekstualnih podataka, kao što su knjige, članci i web stranice. Model koristi proces koji se naziva nenadzirano učenje za prepoznavanje obrazaca i odnosa u podacima, bez da nam se eksplicitno kaže koji su obrasci. Nakon što se model uvježba, može se koristiti za generiranje novog teksta predviđanjem sljedeće riječi ili niza riječi na temelju konteksta prethodnih riječi u rečenici.

Izvorni GPT povećan je 2019. godine, a zatim ponovno 2020. godine, što je rezultiralo novim modelom GPT-3 koji koristi 175 milijardi parametara. Ovaj model je postao temelj za DALL-E, koji koristi svoju multimodalnu implementaciju, koristeći 12 milijardi parametara, koji zamjenjuju tekst za piksele.

DALL-E model je pokazao izvanredne rezultate u višestrukim zadacima generiranja slika na temelju tekstualnih unosa. Osim jednostavnog generiranja uzoraka slika raznih objekata koje je vidio tijekom treniranja, model može integrirati različite ideje i miješati nepovezane koncepte na uvjerljiv način, može čak i generirati objekte koji ne postoje u fizičkom svijetu [13].

Osim DALL-E modela, druga važna tehnologija treniranja modela koja se pojavila u isto vrijeme je CLIP (engl. Contrastive Language-Image Pretraining) [10]. CLIP je napredni model umjetne inteligencije koji su zajednički razvile tvrtka OpenAI i škola UC Berkeley. Model je sposoban razumjeti tekstualne opise i slike koristeći pristup treniranju koji naglašava povezivanje parova slika i teksta.

Prvo rješenje otvorenog koda koje je koristio tehnologiju CLIP je bio alat DeepDaze. Razvio ga je programer Phil Wang u siječnju 2021. godine gdje je kombinirao CLIP s neuronskom mrežom pod nazivom Siren [10, 14]. DeepDaze je stekao popularnost zbog svoje sposobnosti stvaranja fascinirajućih i zadivljujućih slika koje često sliče slikama iz snova ili apstraktnoj umjetnosti.

Nekoliko dana kasnije, isti programer, uz pomoć modela koje je objavio istraživač Ryan Murdock, razvio je još jedan generativni model dubokog učenja pod nazivom BigSleep. Model funkcionira kombiniranjem CLIP-a s BigGAN-om. BigGAN je vrsta generativne suparničke mreže koja koristi strojno učenje za stvaranje slika visokih rezolucija. BigGAN uključuje niz promjena i inovacija koje omogućuju bolje generiranje slika od prethodnih modela. BigSleep koristi rezultate BigGAN-a za pronalaženje slika koje se podudaraju sa slikama generiranih preko CLIP-a [10]. Model zatim postupno prilagođava unos šuma u BigGAN-ov generator sve dok proizvedene slike ne odgovaraju zadanom upitu. Ryan Murdock tvrdi da je BigSleep bio prvi model koji je mogao generirati slike visokih kvaliteta u rezoluciji od 512 x 512 piksela.

BigSleep model inspirirao je još jednu CLIP-GAN kombinaciju. Samo tri mjeseca kasnije, u travnju 2021. godine, istraživačica Katherine Crowson je razvila VQGAN-CLIP model [5]. VQGAN (engl. Vector Quantized Generative Adversarial Network) varijanta je arhitekture GAN. Sposoban je generirati slike visoke kvalitete kodiranjem slika kao detaljno opisanih unosa, što omogućuje učinkovitije treniranje i bolju kvalitetu slike u usporedbi s tradicionalnim GAN modelima.

Nastavljena je linija istraživanja koja se temelji na povezivanju CLIP-a s drugim arhitekturama. U lipnju 2021. godine, Katherine Crowson izumiteljica VQGAN-CLIP-a objavila je još jedno istraživanje, kombinirajući model prethodne obuke kontrastnog jezika i slike s difuzijskim algoritmom za stvaranje CLIP vođene difuzije. Iz ovoga se izrodio današnji model po imenu Stable Diffusion [10].

Modeli umjetne inteligencije pretvaranja teksta u sliku napredovali su daleko u samo nekoliko kratkih godina, ali do 2022. godine, ostali su prilično ograničeni. Mnogi od ranijih modela zahtijevali su mnogo računalne snage za rad, posebno tijekom faze treniranja. To je značilo da su često trenirani na manjim skupovima podataka, pa su ostali zanimljivi uglavnom istraživačima. Neke zanimljive tehnologije bile su dostupne putem Google Colab prijenosnih računala, omogućujući jednostavno izvršavanje koda na hardveru dostupnom putem usluge u oblaku. Međutim, ova metoda još uvijek zahtijeva određenu tehničku sposobnost, ograničavajući popularnost temeljne tehnologije. To se počelo mijenjati 2022. godine, kada su se počele pojavljivati nove aplikacije koje su pružale prikladna sučelja za modele koji se uglavnom temelje na nekoj implementaciji modela difuzije.

2.2. PRINCIP RADA MODELA STABLE DIFFUSION

Za razliku od običnog strojnog učenja koje koristi neuronske mreže gdje su svi neuroni međusobno povezani u slojevima, kod procesa generiranja slika se koriste dva specijalna sloja. Prvi od tih slojeva je konvolucijski sloj (engl. Convolutional layer). Razlog je taj što dolazi do sastava neuronske mreže i kako ona učitava slike s pomoću piksela. Na primjer da se običnoj umjetnoj neuronskoj mreži s međusobno povezanim neuronima da zadatak da učita sliku od 100 piksela u dužini i širini, ta mreža će prvo uzeti u obzir površinu cijele slike po broju piksela i taj broj ponovo pomnožiti samim sobom jer je svaki ulaz u isto vrijeme spojen na svaki izlaz. Osim potrebe računanja ogromnih količina umjetnih neurona potrebnih za učitavanje jedne slike, obična neuronska mreža ne uvažuje pozicije piksela na slici i samim time nije u mogućnosti prepoznavati oblike. Taj problem se rješava korištenjem konvolucijskog sloja.

Kod konvolucijskog sloja je svaki izlazni piksel određen matricom susjednih piksela koji se nalaze oko njega čija je veličina najčešće 3 za 3 ili 5 za 5. Princip je taj da se izlazni piksel računa množenjem susjednih piksela u matrici. Ovim postupkom se znatno smanjuje broj neurona potrebnih da se obradi slika time da se samo učitavaju pikseli u matrici te njihovo mapiranje i određivanje gdje koji piksel treba biti.

2.2.1 Računalni vid

Računalni vid omogućava računalima prepoznavanje slika na način sličan ljudskom. Dok ljudski vid ima prednost dugogodišnjeg iskustva u razlikovanju objekata, procjeni udaljenosti, detekciji kretanja i uočavanju nepravilnosti, računalni vid koristi kamere, podatke i algoritme kako bi postigao slične rezultate u mnogo kraćem vremenskom razdoblju. Umjesto mrežnice, optičkih živaca i dijelova mozga odgovornog za vid, računalni vid trenira strojeve za obavljanje ovih funkcija. Kada je sustav obučen za inspekciju proizvoda ili praćenje proizvodnog procesa, on može analizirati tisuće proizvoda ili procesa u minuti, otkrivajući

suptilne nedostatke ili probleme. Zbog ove sposobnosti, računalni vid može brzo nadmašiti ljudske mogućnosti.

Računali vid se sastoji od pet razina od kojih prva razina predstavlja klasifikaciju slike, odnosno gdje računalo prepoznaje što je na slici ali ne i gdje je unutar same slike. U drugoj razini računalo može prepoznati što je na slici i označava to područje u kojemu se taj objekt nalazi s okvirom u obliku pravokutnika. Treća razina omogućuje računalu prepoznavanje više različitih objekata, odnosno mogućnosti prepoznavanja i označavanja njihovih pozicija pomoću više pravokutnika. Četvrta razina koristi semantičku segmentaciju koja za svaki od različitih objekata pridodaje pikselima različitu boju. Peta razina može točno prepoznati svaki objekt u slici i izolirati ga od drugih. Svaki objekt je klasificiran te ih računalo može prepoznati. Generiranje slika s pomoću umjetne inteligencije počinje s četvrtom razinom zbog semantičke segmentacije koja svakom pikselu na slici daje klasifikaciju te time povećava preciznost rada.

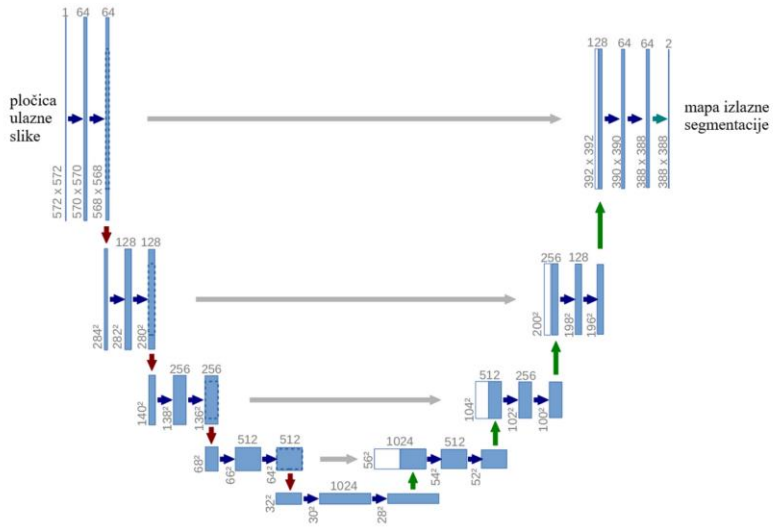
U prošlosti je segmentiranje slika bilo je veoma teško i nepouzdano, te je zahtijevalo tisuće slika za treniranje modela. Jedan od zadataka za koji se je željelo primijeniti segmentaciju slike pomoću umjetne inteligencije jest automatsko otkrivanje i označavanje tumora odnosno raka na medicinskim slikama poput rendgena, CT-a (engl. Computed Tomography) ili MRI-a (engl. Magnetic Resonance Imaging). Taj zadatak je bio gotovo nemoguć zbog malo dostupnog broja slika (između ostalog zbog privatnosti pacijenata) pa je 2015. godine grupa računalnih inženjera predložila novu konvolucijsku arhitekturu zvanu U-Net [1].

2.2.2 U-Net

U-Net je neuronska mreža koja se sastoji od više konvolucijskih slojeva i radi na način da prvo smanji rezoluciju slike te je na kraju vraća na prvobitnu veličinu. Razlog tome je u načinu obrade slike, pri ulasku slika se prvo rastavi u tri kanala boje (crvena, zelena i plava), te uz te tri boje, slika još posjeduje visinu i širinu. Ti parametri omogućuju prikaz slike u 3D tenzoru koji služi da matematički opiše fizička svojstva objekta. Pitanje glasi ako je moguće

izdvojiti više značajki te slike osim u samo tri sloja boje. Taj problem se rješava s pomoću konvolucije [1]. Konvolucija je matematička operacija koja kombinira dvije funkcije kako bi se opisalo njihovo preklapanje. Konvolucija uzima dvije funkcije i preklapa jednu od njih preko druge, te pritom množi njihove vrijednosti u svakoj točki gdje se preklapaju i zbrajaju rezultati da bi se stvorila nova funkcija.

U prvoj polovici U-Net-a, učitana slika se s tri kanala konvolucijom povećava na 64 kanala, pa potom na 128 kanala i sve tako do 1024 kanala. U konvoluciji sa 64 na 128 kanala, svaka 3D matrica ima 64 kanala dok postoji sve ukupno 128 matrica. Time U-Net ima mogućnost uzimanja sve više detaljnih značajki slike. Za razliku od običnog strojnog učenja, umjesto da se povećava veličina matrice, U-Net smanjuje veličinu slike nakon svaka dva konvolucijska sloja i time znatno smanjuje potreban broj resursa.



Slika 2.Prikaz rada U-Net-a

Izvor: U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9446143

U drugoj polovici U-Net-a, slika se povećava na isti način time da se nakon svaka 2 konvolucijska sloja rezolucija slike poveća dok se količina kanala smanjuje. Da se spriječi gubitak detalja na slici koristi se rezidualna veza koji radi tako da spaja konvolucijske slojeve gdje se rezolucija slike smanjuje ili povećava. Tim postupkom kod procesa smanjivanja rezolucije slike u prvoj polovici mreže, ta se informacija prijenosi na drugi dio mreže u isti konvolucijski sloj gdje bi se rezolucija slike natrag povećala na prvotnu veličinu i ostala u istom formatu.

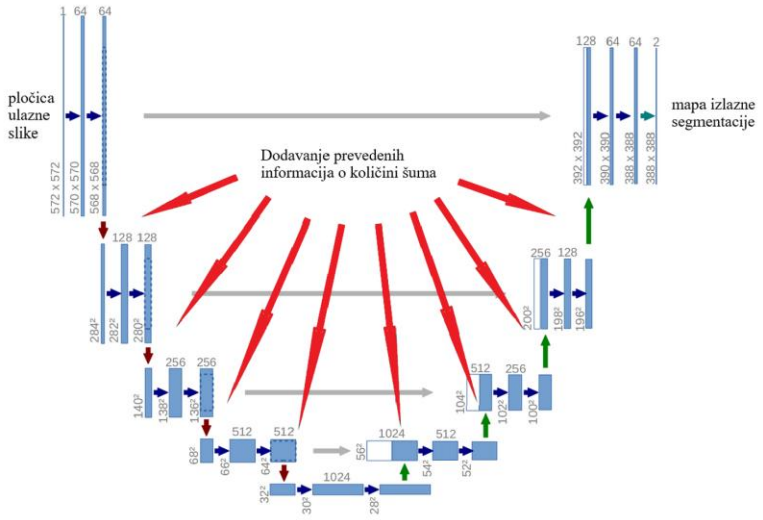
2.2.3 Uklananje šuma pomoću U-Net-a

Osim segmentacije U-Net [1] se koristi za proces koji se zove otklanjanje šuma (eng. denoising). Ako je rezultat slike kojoj je dodan šum jednak istoj koja već prethodno ima šum, onda je moguće prepoznati taj isti šum i izolirati ga iz slike sa šumom, što će rezultirati slikom bliskom originalnoj slici prije dodavanja šuma. Kod primjera koristeći U-Net [1], on bi se trenirao tako da nasumično generirani šum preko navedene slike veliki broj puta te istreniramo model da može otklanjati velike i male šumove s navedene slike.

Slika 3. Prikaz otklanjanja šuma u slici

Izvor: Kriomet*:* Non-local means*,* https://en.wikipedia.org/wiki/Non-local_means

Kod treniranja modela, koriste se više različitih razina šuma. Te razine su označene brojevima od kojih svaki označava koliko trenutna slika ima šuma od 0 dB do potpunog šuma. Dodjeljivanje brojeva slici vezano uz količinu šuma koje ima se zove pozicijsko kodiranje (eng. positional encoding). Pozicijsko kodiranje služi da se dane informacije poput riječi ili trenutnih razina obrade slike budu pretvorene u vektore koje neuronske mreže lakše razumiju. Svaki korak dodavanja ili otklanjanja razine šuma ima svoje mapirane brojeve u vektoru koje pomažu da svaka promjena u šumu bude zabilježena i zapamćena od strane modela. Proces otklanjanja šuma se ne može izvesti u jednom koraku jer bi rezultat bio loše kvalitete i izobličen. Prilikom treniranja modela, svaki put kada se rezolucija slike smanji ili poveća, dodaje joj se informacije o količini šuma da bi rezultat bio što točniji.



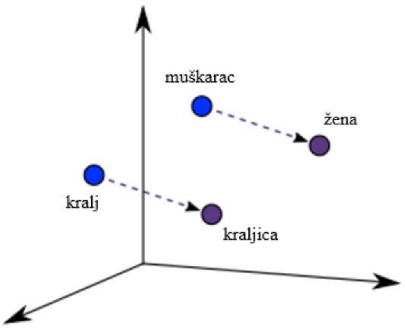
Slika 4. Otklannjajne šuma u U-Net

Izvor: U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9446143

2.2.4 Pretvorbe riječi u vektore (Word2vec)

Trenutačni modeli umjetne inteligencije generiranje slika rade na principu da korisnik opiše željenu scenu koju želi da bude prikazana na slici i model pokuša čim bolje rekreirati tu scenu. Takav način rada je moguć zahvaljujući vrsti neuronske mreže zvane autoenkoder. Autoenkoder je vrsta neuronske mreže dizajnirana za učinkovitu kompresiju ulaznih podataka i svih njihovih bitnih značajki, zatim rekonstruiranje (dekodiranje) izvornog ulaza iz ove komprimirane reprezentacije. Time je umjesto da se šum dodaje na originalnu veličinu slike i onda uklanja, u ovom slučaju se slika prvo smanji u veličini te samim time brzina njezine obrade postane puno brža.

Word2vec je tehnika kojom su sve riječi engleskog jezika mapirane vektorima tako da svaka ima svoju jedinstvenu vrijednost [2]. Svaka riječ koja ima sličnu vrijednost nekoj drugoj riječi u istoj listi znači da te dvije riječi imaju slično značenje. Shodno tome napravljene su dvije takve liste koje sadrže iste riječi ali sa različitim vrijednostima. Te dvije liste su trenirane s čitavom engleskom literaturom tako da svaka riječ iz prve liste ima sličnu vrijednost kao riječ iz druge ukoliko se te dvije riječi često koriste zajedno u rečenici. Ako bi se iz jedne od lista na primjer uzela vrijednost riječi "kralj" i oduzela s vrijednošću riječi "muškarac" te se onda tom rezultatu dodala vrijednost riječi "žena", rezultat bi točno predstavljao vrijednost riječi "kraljica". Time se dokazuje kako su sve riječi međusobno povezane u listi.



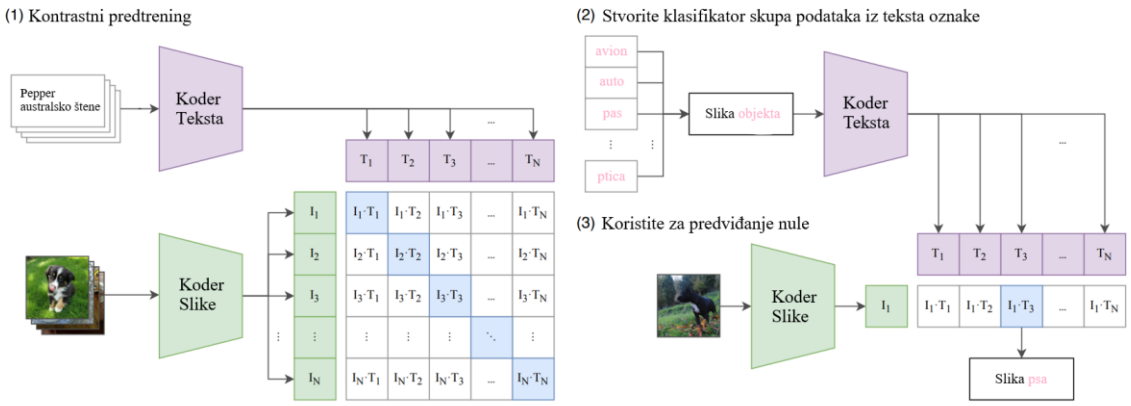
Slika 5. Prikaz relacija riječi u Word2vec

Izvor: Word2Vec Explained: https://israelg99.github.io/2017-03-23-Word2Vec-Explained/

2.2.5 Sloj samo pozornosti

Drugi od glavnih slojeva za generiranje slika se zove sloj samo pozornosti (engl. Self attention layer). Taj sloj radi sličnu zadaću kao konvolucijski sloj ali za razliku od njega koji uzima podatke iz slike u odnosu na povezanost piksela u slici, sloj pozornosti uzima podatke iz rečenica u odnosu na povezanost riječi preko njihovih vektorskih vrijednosti po Word2vec tehnici [2]. Konvolucijski sloj daje slikama vektorske vrijednosti, dok sloj samo pozornosti daje tekstu vrijednost. Organizacija OpenAI je s pomoću modela CLIP [10] (engl. Contrastive Language–Image Pre-training) trenirala model tako da slika i rečenica koja opisuje tu sliku imaju jako slične vrijednosti po oba modela.

CLIP [10] je model koji je razvila tvrtka OpenAI, koji povezuje slike i tekst u zajedničkom prostornom okruženju, omogućujući modelu da razumije i generira značajne asocijacije između vizualnog sadržaja i tekstualnih opisa. CLIP se temelji na pristupu usporednog učenja, gdje se neuronska mreža trenira kako bi razlikovala parove odgovarajućih slika i tekstova. Osnovna ideja je naučiti zajednički prostor u kojem su istim značenjem povezane slike i tekstovi, dok su nepovezane slike i tekstovi razdvojeni.



Slika 6. Prikaz rada tehnologije CLIP

Izvor: Learning Transferable Visual Models From Natural Language Supervision: https://arxiv.org/abs/2103.00020v1

Zahvaljujući tom otkriću, kod U-Net-a [1] prilikom smanjivanja ili povećavanja rezolucije slike, šum koji bi ta slika dobivala jedinstveno odgovara riječi koje je korisnik pisao. U tom procesu model umjetne inteligencije uzima slike na kojima je bio treniran, te dodaje ili izrezuje ono što mu treba ili ne treba ovisno o riječima koje je korisnik pisao. U primjeru gdje bi korisnik napisao rečenicu "Pas koji vozi biciklu." model umjetne inteligencije bi uzeo sliku psa i bicikle koju posjeduje. Te bi ih zajedno spajao i u isto vrijeme brisao sve što ne odgovara vektoru vrijednosti tih riječi i na kraju bi rezultat bio pas na bicikli. Naravno što je bolje model istreniran na određenoj stvari to bolje i detaljnije može napraviti tu sliku da odgovara riječima koje je korisnik napisao.

3. STABLE DIFFUSION

U ovom poglavlju će se ispitivati funkcionalnost različitih modela stabilne difuzije (engl. Stable Diffusion). Pregledati će se njihove opcije generiranja slika, sve postavke koje mogu uvjetovati kako će slika izgledati i njihova efikasnost generiranja slika koristeći riječi i fraze koje se normalno ne pišu zajedno u rečenici. Ovime se ispituje kvaliteta generiranja slike metodom difuzije i koliko dobro je istreniran model. Na kraju će se rezultati usporediti i procijeniti će se koji je od navedenih modela najbolji u odnosu na kvalitetu generiranih slika.

Prije početka se prvo trebaju navesti i obrazložiti dva nova pojma koja će se koristiti kroz poglavlje. Ta dva pojma su Checkpoint i LoRA [7]. Kontrolna točka (engl. Checkpoint) predstavlja stil generiranja slika umjetne inteligencije. Checkpoint je model istreniran na velikoj količini slika te mu je uloga da daje informacije drugom modelu kako generirati slike. Checkpoint je zaslužan za stil crtanja koji će slika imati nakon generiranja. Taj stil može biti u stilu crtanog filma, realističnih fotografija, renesansnih slika, itd..

LoRA (engl. Low-Rank Adaptation Model) [7] je manji model koji ne koriste svi modeli za generiranje slike, a njegova veličina je znatno manja od veličine Checkpoint-a. LoRA [7] je treniran na manjoj količini slika i koristi se da bi na neki način naredio Checkpoint-u da stvara određene objekte ili likove u slici. Ako korisnik želi generirati sliku određene zgrade npr. Eiffel-ov toranj, nakon što odabere željeni Checkpoint, umjesto da korisnik opisuje kako toranj izgleda, odabrat će LoRA-u [7] koja će to napraviti umjesto njega. Naravno, ako je Checkpoint dovoljno dobro istreniran, moći će i sam nacrtati Eiffel-ov toranj samo iz unosa koje je korisnik napisao, ali preko LoRA-e [7] će taj toranj izgledati jako slično svaki put nakon generiranja jer je bio treniran sa sličnim slikama. U slučajevima gdje korisnik želi opetovano nacrtati iste slavne osobe ili izmišljene likove, tu će taj model imati veliku prednost jer sam Checkpoint uglavnom neće svaki put znati što korisnik želi, odnosno generirane slike će imati veće varijacije.

3.1. STABLE DIFFUSION 1.0

Stable Diffusion 1.0 je prvi prototip svoje vrste napravljen od strane tvrtke Stability AI koji je trenirao model sa LAION-5B [9] bazom podataka na 256 različitih NVidia A100 grafičkih kartica od 80GB VRAM (engl. Video Random Access Memory) memorije kroz period od 150,000 sati rada. Model je bio prethodno treniran na slikama veličine 256 piksela širine i dužine, a zatim na slikama veličine 512 piksela širine i dužine. LAION-5B [9] je veliki skup podataka za istraživačke svrhe koji se sastoji od 5,85 milijardi parova slika i teksta filtriranih CLIP tehnologijom [10]. Od toga 2,3 milijardi sadrže engleski jezik, 2,2 milijardi primjera iz više od 100 drugih jezika i 1 milijarda primjera sadrže tekstove koji ne dopuštaju dodjelu određenog jezika (npr. imena).

Kasniji modeli poput Stable Diffusion 1.5 modela bili su trenirani s više skupa podataka se većim količinama podataka za bolju optimizaciju kod generiranja slika i više detalja ali još uvijek pate od istih mana koje originalno 1.0 model posjeduje. Baziran je na sličnom radu kao Google Imagen [11], ovaj model koristi smrznuti enkoder teksta CLIP ViT-L/14 za prilagođavanje modela tekstualnim upitima. Sa 860M U-Net-om [1] i 123M enkoderom teksta, model je relativno jednostavan i radi na grafičkim karticama s najmanje 10 GB VRAM-a.

3.1.1. Ispitivanje različitih stilova generiranja slika pomoću modela Stable Diffusion 1.0

S obzirom na to da je ovaj model Stable Diffusion-a prvi prototip, njegova uporaba više nije toliko popularna. Za testiranje rada ovoga modela koristiti će se online alat zvan "Stable Diffusion Online". Ovaj alat bazira rad na modelu Stable Diffusion 1.0 te isto tako može koristiti i druge modele. Stable Diffusion 1.0 ima samo teks-u-sliku opciju pa će se u ovome primjeru ispitati mogućnost modela u tome koliko dobro crta određene stilove i ako je model dovoljno

dobro istreniran da može kreirati scene koristeći riječi koje prema Word2vec [2] nemaju slične vrijednosti. Slika 7. predstavlja tri različite slike nacrtane s tri različita stila.

Na slici 7 a) vidi se stil realizma, te su riječi za generiranje slike bile: "1 person, woman, face, face closeup, green eyes, freckles, brown hair, smiling, looking at viewer, realistic". Na slici 7 b) vidi se stil crtanog filma, te su riječi za generiranje slike bile: "dog, red collar, white fur, big dog, park background, cartoon". Slika 7 c) predstavlja stil digitalne umjetnosti su riječi za generiranje ove slike bile: "asian dragon, giant dragon, mythology, white fur, wrapped around the tower, pagodas, sunset, clouds, sun rays, detailed background, old chinese city, red rooftops".

Slika 7. Prikaz generiranih slika u 3 različita stila sa lijeva na desno: a) Stil realizma, b) Stil crtanog filma i c) Stil digitalne umjetnosti

Izvor: Stable Diffusion Online, unos od strane studenta

Rezolucija svih slika je 512 piksela širine i visine. Prosječno vrijeme generiranja ovih slika je trajalo oko 9 sekundi. Svaka od slika zadovoljava svoj namijenjeni stil usprkos manjku detalja i dubine. Dok druge dvije slike zadovoljavaju svoj stil, na prvoj slici stila realizma jasno se ističe manjak fotorealizma.

3.1.2. Ispitivanje generiranja apstraktnih tema pomoću modela Stable Diffusion 1.0

Drugi test se bazira na radu programa ako može generirati slike apstraktnih tema s opisom riječi koje imaju znatno različite vrijednosti prema Word2vec [2]. Za testiranje se uzela nasumična izmišljena scena gdje su riječi korištenje za generiranje slike bile sljedeće: "Astronaut in space standing on a cake". Na slici ispod se može vidjeti da je alat u potpunosti uključio napisane riječi te generirao zadanu sliku.

Slika 8. Slika apstraktne teme

Izvor: Stable Diffusion Online, unos od strane studenta

Usprkos svojoj mogućnosti generiranja različitih slika u različitim stilovima, Stable Diffusion 1.0 još uvijek pati od raznih ograničenja poput nemogućnosti postizanja savršenog fotorealizma i mogućnosti prikazivanja čitljivog teksta. U radu se stvaraju problemi kod generiranja kompliciranih zadataka kao lica ljudi, prstiju na rukama itd.. Stable Diffusion 1.0 je treniran na ogromnom skupu podataka LAION-5B [9], koji obuhvaća različite vrste slika i tekstualnih opisa. Iako ovaj raznoliki skup podataka omogućava modelu generiranje raznolikih vizualnih stilova, istovremeno uključuje i sadržaj koji nije prikladan za sve korisnike, posebno

materijale za odrasle. Zbog toga model nije siguran za upotrebu bez dodatnih sigurnosnih mehanizama, poput filtriranja neprimjerenog sadržaja. Ovi mehanizmi filtriranja i razmatranja sigurnosti počeli su se primjenjivati tek u novijim verzijama modela, što je znatno poboljšalo njihovu sigurnost i primjenjivost u širem području aplikacija.

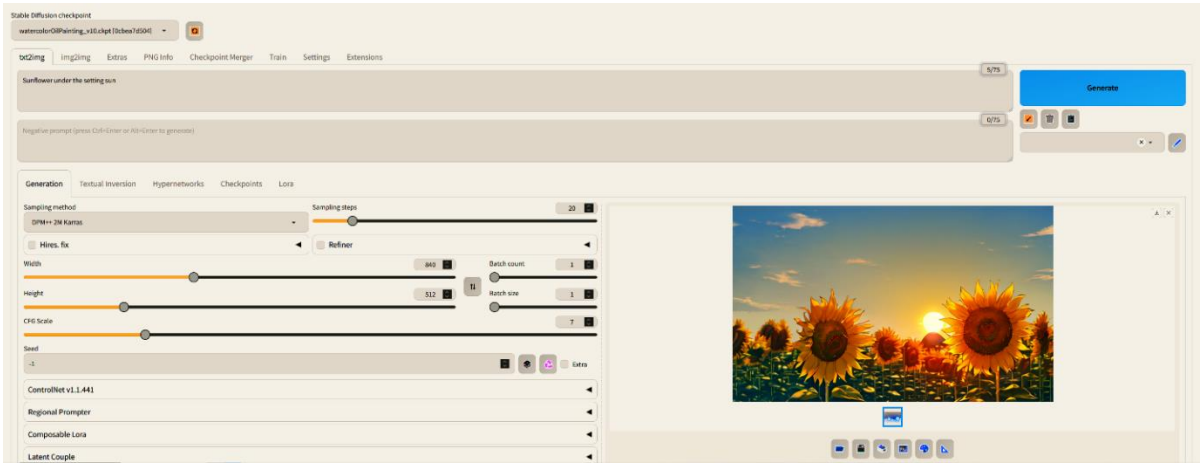
3.2. STABLE DIFFUSION 2.0

Stable Diffusion 2.0 je nova i bolja verzija 1.0 tekst-u-sliku modela napravljena od strane tvrtki Stability AI i LAION. Stable Diffusion 2.0 uključuje nove poboljšanije modele pretvaranja teksta u sliku trenirane s pomoću potpuno novog enkodera teksta OpenCLIP [12], koji je razvio LAION uz podršku Stability AI i znatno poboljšava kvalitetu generiranih slika u usporedbi s prijašnjim 1.0 modelom. Modeli pretvaranja teksta u sliku u ovoj verziji mogu generirati slike s rezolucijom od 512 piksela širine, odnosno dužine i 768 piksela širine i dužine. Kao i 1.0 model, Stable Diffusion 2.0 je isto treniran na LAION-5B [9] bazi podataka koja je dodatno filtrirana da makne bilo kakve nepoželjne sadržaje za odrasle. Naspram originalnom načinu treniranja CLIP [10] tehnologije, OpenCLIP je treniran na većim paketima informacija za bolju efikasnost tako da se maksimizira količina informacija koje grafičke kartice obrađuju [12]. Isto kao i u Stable Diffusion 1.0 modelu, ovaj model je također optimiziran da može raditi na samo jednoj grafičkoj kartici tako da je dostupan za koristiti gotovo svakome.

Osim generiranja slika, Stable Diffusion 2.0 ima više mogućnosti obrade slika poput upscaler-a koji povećava rezoluciju slike za do četiri puta svoje originalne rezolucije i inpaint opcije koja omogućuje da program nacrta novu sliku preko određenog označenog dijela unutar slike. Ove opcije daju ovome modelu mnogo veću slobodu u generiranju slika kao i uređivanja i dodavanja detalja time dajući veću dubinu slikama.

3.2.1. Ispitivanje različitih stilova generiranja slika pomoću modela Stable Diffusion 2.0

Za ispitivanje modela Stable Diffusion 2.0 će se za početak kao i u prijašnjem ispitivanju 1.0 modela ispitati različiti stilovi crtanja, te će se zatim ispitati nove opcije povećavanja rezolucije slike s pomoću upscaler-a kao, mogućnost uređivanja slika i dodavanja dubine slici. Ovo ispitivanje će se izvoditi pomoću alata "Stable Diffusion web UI" koji pruža puno mogućnosti u pomaganju procesa generiranja slika, te je jedan od najpopularnijih besplatnih alata za izradu slika. Njegov rad koristi resurse grafičke kartice korisnika. Model grafičke kartice u ovom ispitivanju je NVidia GeForce RTX 3080 SUPRIM X 10G koji sadrži 10 GB VRAM memorije, brzinu memorije od 19 Gbps i 8704 CUDA (engl. Compute Unified Device Architecture) jezgri s brzinom rada do 1920 MHz. Na slici dolje se može vidjeti sučelje programa Stable Diffusion web UI.



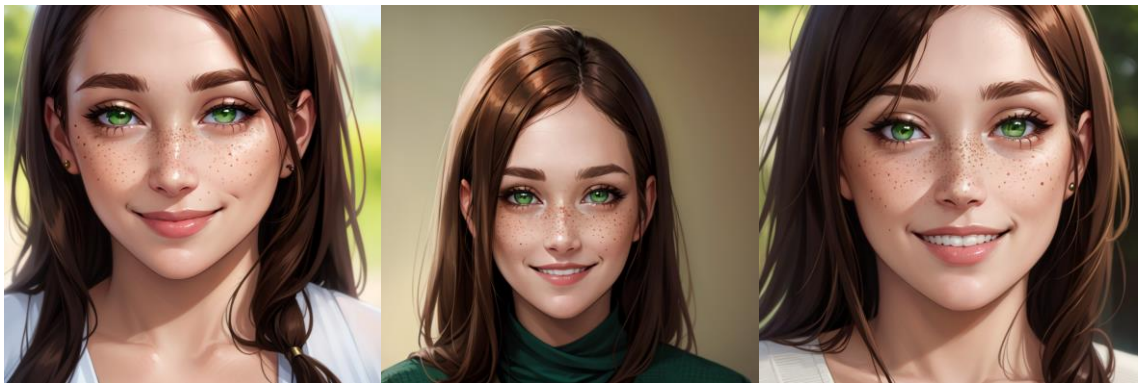
Slika 9. Sučelje programa Stable Diffusion web UI

Izvor: Stable Diffusion web UI

Prije generiranja slika, podesit će se sljedeći parametri. Odabrani Checkpoint će u ovome slučaju biti "darksun_v7b". U tekstualni prozor negativa bit će ubačena dva LoRA [7] modela, "BadDream" i "badhandv4". Ta dva modela imaju ulogu da ublaže distorziju slike prilikom

generiranja i potencijalno zaustave da generirana slika ima loše nacrtane ruke ili oblike lica. Pod stavku "Sampling method" ostat će opcija "DPM++ 2M Karras". Ova stavka određuje način na koji se otklanja šum pri generiranju slike te zadana opcija iz prijašnjih testiranja pruža najbolje rezultate. Širina i visina slike će biti postavljeni na 768 piksela. Stavka "Sampling steps" koja određuje u koliko koraka će se šum uklanjati te time kolika će kvaliteta slike biti postavljena na 35. "CFG Scale" (engl. Classifier-Free Guidance) skala koja određuje koliko strogo program mora slijediti Checkpoint pri generiranju slika je postavljena na vrijednost 5. Stavka "Batch count" će ostati na 1, a stavka "Batch size" će se dignuti na 3. To znači da će se pri pokretanju rada alata generirati 3 slike odjednom ali te slike neće biti slične već će nakon svakog generiranja alat generirati sliku na način da nova slika ne bude vezana uz prethodnu. Stavka sjeme (eng. seed) će ostati na vrijednosti -1 što znači da će generirane slike biti nasumično nacrtane bez utjecaja od prijašnjih slika. Kod generiranja sljedećih slika, prikazat će se sve 3 slike. Ispod se može vidjeti 3 različita stila generiranja slika putem modela 2.0.

Slika 10 predstavlja stil realizma. Za generiranje ove slike koristile su se riječi: "1 person, woman, face, face closeup, green eyes, freckles, brown hair, smiling, looking at viewer, realistic". Na slici se može vidjeti da je model generirao tri jako slične slike te se zaključuje da je Checkpoint treniran specifično da generira ljude.



Slika 10. Stil realizma

Izvor: Stable Diffusion web UI, unos od strane studenta

Slika 11 predstavlja stil crtanog filma. Riječi generiranja ove slike bile us: "dog, red collar, white fur, big dog, park background, cartoon". Po slikama se može vidjeti da za razliku od prethodne slike, ovaj Checkpoint se više trenirao na slikama lica pa su slike o ovome primjeru rezultirale s tri potpuno drukčije slike istog psa.

Slika 11. Stil crtanog filma

Izvor: Stable Diffusion web UI, unos od strane studenta

Slika 12 predstavlja stil digitalne umjetnosti. Riječi generiranja u ovom primjeru su bile sljedeće: "asian dragon, giant dragon, mythology, white fur, wrapped around the tower, pagodas, sunset, clouds, sun rays, detailed background, old chinese city, red rooftops". Checkpoint "darksun_v7b" ne može sam nacrtati sliku zmaja, stoga se u ovom primjeru koristila LoRA [7] "ChineseDragon_v2" koja je specifično trenirana na generiranju azijskih zmajeva da pomogne Checkpoint-u nacrtati željenu sliku. Kao što je prije naglašeno, ovaj checkpoint je uglavnom treniran da generira slike ljudi pa mu je potrebna pomoć od drugih modela. Svaka od slika je zadovoljila svoje upite te je prosječno vrijeme generiranja iznosilo 4 sekunde.

Slika 12. Stil digitalne umjetnosti

Izvor: Stable Diffusion web UI, unos od strane studenta

3.2.2. Ispitivanje generiranja apstraktnih tema pomoću modela Stable Diffusion 2.0

Sljedeće ispitivanje će biti apstraktnih scena s opisom riječi koje imaju znatno različite vrijednosti prema Word2vec [2]. Za primjer će se uzeti ista rečenica kao u testiranju prijašnjeg modela: "Astronaut in space standing on a cake". Korišteni Checkpoint ostaje "darksun_v7b", te su kod prozora negativa postavljene LoRA-e [7] "BadDream" i "badhandv4" koje imaju ulogu da pomažu u procesu generiranja slike time da izoliraju način otklanjanja šuma koji bi rezultirao generiranom slikom osobe s dodatnim prstima ili neproporcionalnim tijelom.

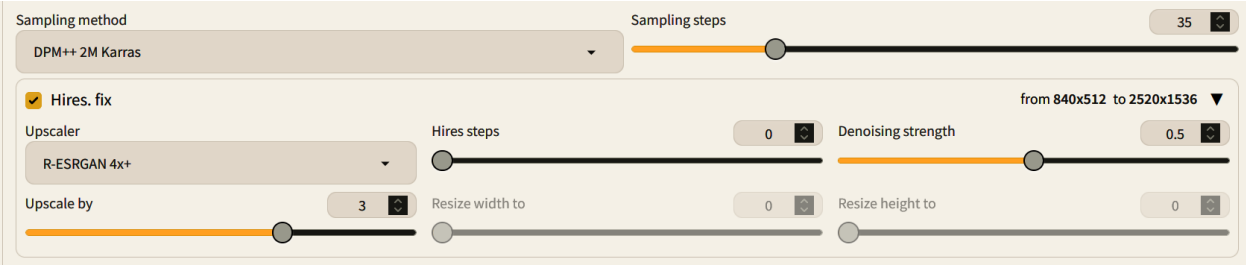
Slika 13. Slika apstraktnih tema

Izvor: Stable Diffusion web UI, unos od strane studenta

Na slici iznad se može vidjeti da alat nije mogao napraviti sliku opisanu u rečenici. Odabrani Checkpoint nije dovoljno treniran na ovoj tematici te zato je potrebna LoRA [7] u ovim slučajevima da pomogne u generiranju slike. Iz slika se isto može vidjeti da je u svakoj od njih nacrtan ženski astronaut što još više potvrđuje da je Checkpoint najviše treniran da generira ženske likove. Ovaj problem se srećom može lako zaobići korištenjem drugog Checkpoint-a ili LoRA-e koji su trenirani na generiranju slika astronauta. Zahvaljujući svojoj popularnosti i velikoj zajednici, Stable Diffusion web UI ima tisuće LoRA i Checkpoint-a na raspolaganju za odabir [7].

3.2.3. Ispitivanje povećanja rezolucije slike pomoću modela Stable Diffusion 2.0

Sljedeće će se ispitivati opcija povećanja rezolucije slike, gdje će se uzeti generirana slika i usporediti s originalnom nakon povećavanja rezolucije. Na slici ispod se može vidjeti ta opcija kao i podešeni parametri koji će se koristiti. Za generiranje slike će se koristiti riječi: "bar, wooden walls, drinks on the wall, table, cocktail, glass, texture, martini glass, condensed water on the glass".



Slika 14. Postavke za povećanje rezolucije slike

Izvor: Stable Diffusion web UI

Opcija "Denoising strength" je postavljena na vrijednost 0,5. Ona određuje balans između uklanjanja šuma na slici. Ovu opciju je važno podesiti kod povećanja rezolucija slika jer ako je njezina vrijednost iznad ili ispod 0,5 (osnovne razine) to će rezultirati u slici koja izgleda drukčije od slike prije povećanja rezolucije. Na slici ispod se može vidjeti slika prije i poslije povećanja rezolucije.

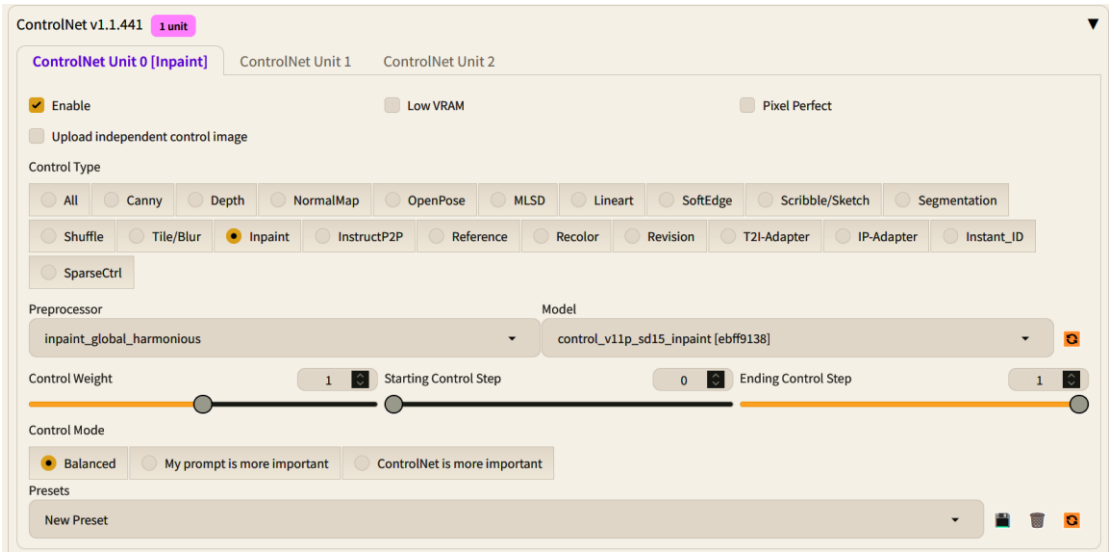
Slika 15. Prikaz povećanja rezolucije slike sa 128 x 128 (lijevo) na 2304 x 2304 piksela (desno)

Izvor: Stable Diffusion web UI

Povećanje rezolucije slike je ujedno izoštrilo kontrast slike i time pojačalo dubinu kao i detalje. Spektar svijetla izgleda puno prirodnije kao i sjene koje prostor stvara. Opcija "Denoising strength" u ovom slučaju određuje koliko će se slika mijenjati od originala dok se povećava rezolucija. Radi limitacije VRAM memorije grafičke kartice, slika se mogla samo tri puta povećati umjesto maksimalno četiri puta. Cijeli proces je trajao oko 160 sekundi i zauzimao je 90 % VRAM memorije.

3.2.4. Ispitivanje opcije inpaint pomoću modela Stable Diffusion 2.0

Sljedeća opcija uređivanja slike koju Stable diffusion 2.0 nudi se zove inpaint. Ova opcija omogućuje korisniku da doda ili nacrta novu sliku prijeko dijela postojeće i time joj da više detalja ili izbriše neželjene objekte u slici. Ova opcija radi neovisno o Checkpoint-u ili LoRA-i [7] koje je korisnik odabrao. Nakon što korisnik generira sliku putem alata, odabrat će opciju inpaint ispod generirane slike koja će ga odvesti u drugi prozor na slici ispod.



Slika 16. Prikaz postavki opcije inpaint

Izvor: Stable Diffusion web UI

Opcije na slici 16 podešene su da služe opciji inpaint. U ovome slučaju prvo je opcija inpaint upaljena i pod opciju "Control Type" odabrana je opcija "Inpaint". Ovaj odabir daje alatu doznanja da korisnik želi crtati nove oblike po postojećoj slici. Opcije "Preprocessor" i "Model" su odabrane po svojoj efektivnosti rada. Prvi proces opcije inpaint je da se generira slika. Slika je generirana prema sljedećim riječima: "clouds, sky, sunset, town, sea, port, day, sunrays, ships" te je rezolucija povećana sa 768x768 na 1536x1536 piksela. Slika je pritom prebačena u prozor alata inpaint gdje se označio mali dio slike koji se želi izmijeniti. Nakon što je dio slike označen,

s podešenim postavkama napisala se nova rečenica u tekstualni prozor: "plane, flying, red". Time alat zanemaruje prethodno generiranu sliku, već generira novu zadanu po njoj. Alat pritom pokušava generirati napisane riječi u označenom prozoru te proces kao i gotov rezultat se može vidjeti na slici ispod.

Slika 17. Prikaz rada opcije inpaint

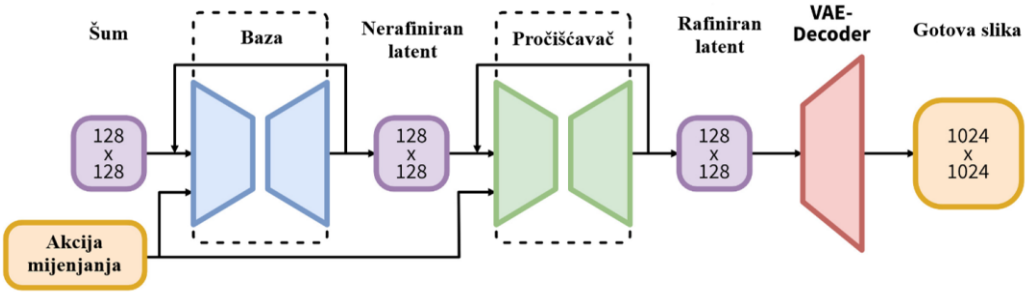
Izvor: Stable Diffusion web UI, unos od strane studenta

Opcija inpaint nije savršena i treba nekoliko pokušaja da se dobije zadovoljavajući rezultat. Sam po sebi Stable diffusion 2.0 je znatno bolji model od prethodnog 1.0 i pruža puno mogućnosti kod uređivanja slika. Njegova prednost u boljoj optimizaciji resursa pri generiranju slika daje mu puno veću popularnost i lakše korištenje. Dodatne opcije koje pruža ovaj model nisu savršene ali uz vrijeme, mogu se unaprijediti na bolje uz bolju optimizaciju i kvalitetnije treniranje modela.

3.3. SDXL

SDXL (engl. Stable Diffusion Extra Large) je bolja i više optimizirana verzija modela Stable Diffusion modela 1.5 i 2.1. Njegova prednost dolazi od tri puta većeg U-Net-a [1] koji kombinira drugi tekstualni enkoder da mu omogući korištenje velikog broja parametara. Njegov rad se zasniva na načinu rada dvije cijevi. Prvo se osnovni model koristi za generiranje slike željene izlazne veličine. U drugom koraku se koristi specijalizirani model visoke razlučivosti i primjenjuje se tehnika pod nazivom SDEdit [4] također poznatu kao (engl. "img2img") na sliku generiranu u prvom koraku, koristeći isti upit. SDEdit omogućuje sintezu slika na temelju poteza (engl. stroke-based), uređivanje slika na temelju poteza i sastavljanje slika bez optimizacije za specifičan zadatak. SDEdit se može izravno uključiti u gotove difuzijske modele [3].

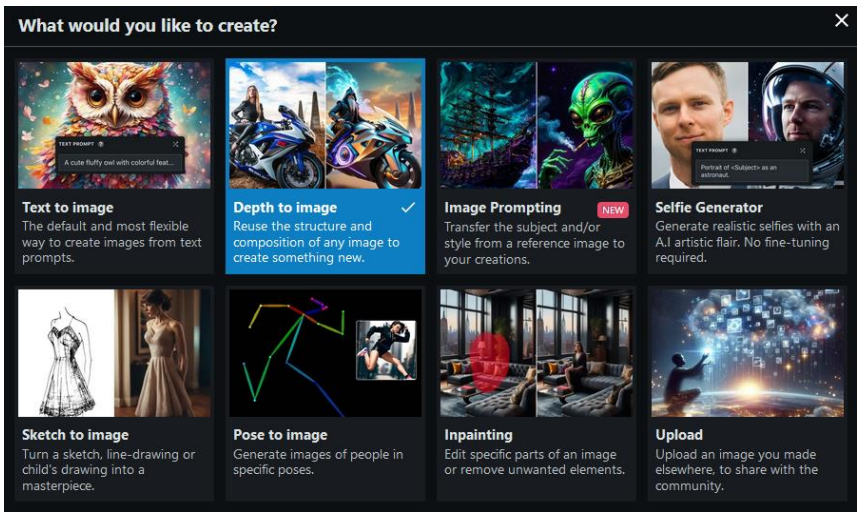
SDEdit prvo dodaje šum na ulaz, zatim naknadno uklanja šum rezultirajuće slike kroz SDE prije povećanja njezinog realizma. SDEdit ne zahtijeva obuku specifičnu za zadatak ili inverzije i može prirodno postići ravnotežu između realizma i vjernosti. SDEdit značajno nadmašuje najsuvremenije metode temeljene na GAN-u (engl. Generative Adversarial Network) do 98,09% na realističnosti i 91,72% na ukupnim rezultatima zadovoljstva, prema testiranju ljudske percepcije, na više zadataka, uključujući sintezu slike temeljenu na crtama i uređivanje također kao sastavljanje slike. Mana ove tehnike je ta što ima usporen rad jer zahtijeva više funkcija [4].



Slika 18. Prikaz rada modela SDEdit

Izvor: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis: https://arxiv.org/pdf/2307.01952

Za prikaz rada ovog modela koristi se online alat NightCafe. NightCafe je popularan online alat za generiranje slika pomoću modela SDXL. Alat je napravljen je 2019. godine zahvaljujući programeru Angus Russell-u. NightCafe koristi tekst-u-sliku metodu VQGAN+CLIP [5] koja pruža više upita slici za više detalja, ali time detalje u generiranim slikama odvaja na sekcije koji mogu rezultirati nesmislenim slikama. Iako je izgled slika lošiji od običnih prethodno treniranih modela, zahvaljujući ovoj metodi generiranja slika, korisnik ima na raspolaganju puno više opcija i tema koje alat može nacrtati. Ono što će se testirati s ovim alatom su različiti stilovi crtanja, nekoliko drugih ponuđenih opcija te koliko dobro model crta slike kada mu se daju riječi različitih vektorskih vrijednosti [2]. NightCafe naime posjeduje nekoliko dodatnih zanimljivih opcija kod generiranja slika, stoga će se još ispitati opcija generiranja slika preko već prethodno nacrtanih poza.



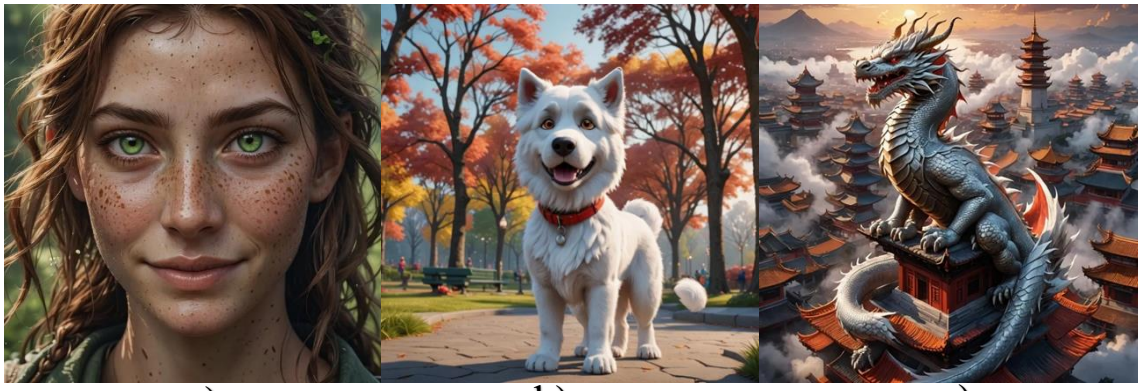
Slika 19. Različite mogućnosti koje nudi program Nightcafe

Izvor: NightCafe

3.3.1. Ispitivanje različitih stilova generiranja slika pomoću modela SDXL

Prvo će se testirati nekoliko različitih stilova crtanja koji će se kasnije uspoređivati s drugim modelima kao i njihova brzina, odnosno vrijeme generiranja. Na slikama ispod se mogu vidjeti tri različita stila crtanja.

Riječi napisane za generiranje svake slike su sljedeće: Na slici 20 a) predstavlja stil realizma, te su riječi za generiranje slike bile: "1 person, woman, face, face closeup, green eyes, freckles, brown hair, smiling, looking at viewer, realistic". Slika 20 b) predstavlja stil crtanog filma, te su riječi za generiranje slike bile: "dog, red collar, white fur, big dog, park background, cartoon". Slika 20 c) predstavlja stil digitalne umjetnosti su riječi za generiranje ove slike bile : "asian dragon, giant dragon, mythology, white fur, wrapped around the tower, pagodas, sunset, clouds, sun rays, detailed background, old chinese city, red rooftops".



Slika 20. Prikaz generiranih slika u 3 različita stila sa lijeva na desno: a) Stil realizma, b) Stil crtanog filma i c) Stil digitalne umjetnosti

Izvor: NightCafe, unos od strane studenta

Vidljivo je da je model u mogućnosti generirati sliku prema ponuđenim riječima te isto tako dodavati detalje u sliku ovisno o tome što korisnik želi da ta slika prikazuje. U usporedbi s prijašnjim modelima Stable Diffusion, SDXL nema problema kod generiranja slika različitih tema i ne zahtijeva pomoć LoRA-e [7]. Zbog tih opcija NightCafe je jednostavniji za korištenje

od modela Stable Diffusion 1.0. Brzina generiranja je veoma dobra s prosjekom generiranja od 1,5 sekunde za sliku veličine 768 piksela širine i dužine. Što se tiče kvalitete slika, kako što je prije navedeno, zbog uporabe metode VQGAN+CLIP [5], NightCafe pruža malo opcija kod odabira stila pa kvaliteta slika ostaje limitirana u tom aspektu.

Drugi test se bazira na radu alata ako je u mogućnosti generirati slike apstraktnih tema s opisom riječi koje imaju znatno različite vrijednosti prema Word2vec [2]. Za testiranje se uzela nasumična izmišljena rečenica kao i u prethodnom ispitivanju: "Astronaut in space standing on a cake". Program je bez problema uspio spojiti sve riječi u tekstualnom prozoru i napraviti izmišljenu scenu. Detalji svemira na slici kao i torta se savršeno uklapaju u scenu bez preklapanja ili ikakve distorzije. Model SDXL za razliku od prethodnih modela nema nikakvih problema u generiranju ovakvih slika, već je napravljen upravo za tu svrhu.

Slika 21. Slika apstraktne teme

Izvor: NightCafe, unos od strane studenta

3.3.2. Ispitivanje opcije generiranja slika preko gotovih poza pomoću modela SDXL

U području vizualnih umjetnosti i komunikacije, ljudske poze imaju duboku važnost. Način na koji se ljudi tijelom izražavaju nešto je što su umjetnici koristili tisućljećima, od pećinskih slika do portreta iz 19. stoljeća ili ekspresivnih pokreta u modernom plesu. Način na koji osoba stoji, sjedi i pozira mogu otkriti njezine osjećaje i emocije, ispričati priče i prenijeti namjere bez riječi. Za kraj će se ispitati opcija generiranja slika preko pred-odabranih poza. Ova opcija je specijalizirana za prevođenje položaja ljudskih figura s referentne slike na generiranu sliku, gdje otkriva položaj nacrtanih dvodimenzionalnih figura i nastoji preslikati njihov položaj tijela. Ova opcija je jako korisna za dizajn i pozicioniranje likova, važno je uzeti u obzir da poze gdje su ruke ili noge jako blizu jedna drugoj mogu predstavljati izazov za točnost replikacije.

NightCafe pruža 46 različitih poza koje korisnik može odabrati od skakanja do ležanja. Jedino što korisnik mora napraviti je odabrati pozu Checkpoint kojim želi da se slika nacrta preko odabrane poze. U ovome primjeru će se uzeti sljedeća poza. Po slikama se može vidjeti da ova opcija ne radi svaki put ali je definitivno u mogućnosti rekreirati pokrete i napraviti sliku preko njih u par pokušaja. Na žalost NightCafe nema opciju da korisnik sam napravi pozu koju želi.

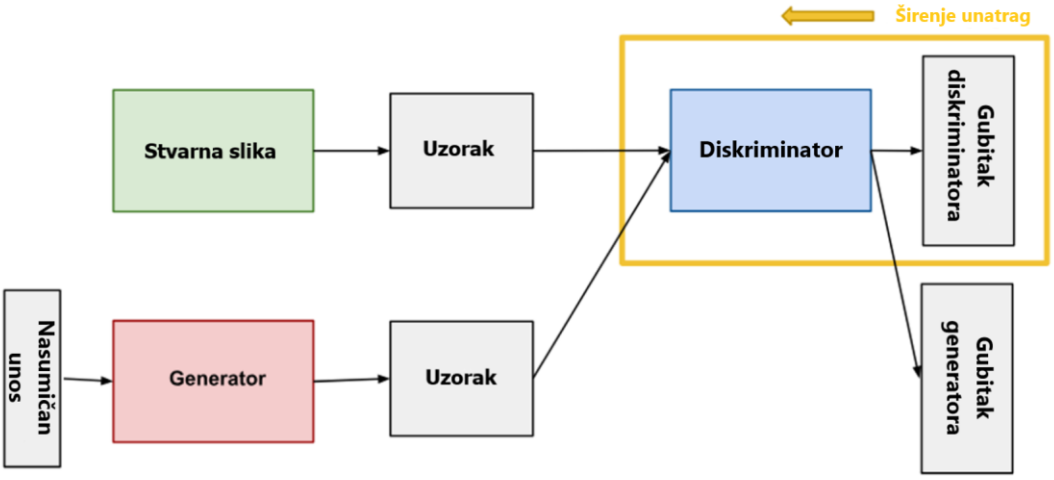
Slika 22. Slike generirane prema pozi iznad na slici lijevo

Izvor: NightCafe, unos od strane studenta

3.4. SDXL TURBO

SDXL turbo je modificirana verzija SDXL 1.0 koja je istrenirana da radi izrazito brzo na način da u trenutku kada korisnik napiše riječ, slika će biti generirana u realnom vremenu. Ta opcija je moguća zahvaljujući novom načinu treniranja modela zvanog ADD (Suprotstavljena difuzijska destilacija, engl. Adversarial Diffusion Distillation) koji omogućuje procesiranje slika velikih rezolucija u jedan do četiri koraka [6].

ADD [6] bazira svoj rad na optimizaciji diskriminirajuće mreže (engl. discriminator network) čija je uloga da razlikuje stvarne podatke iz pravih slika i generirane podatke napravljene od strane računala. Stvarne podatke poput slika ljudi diskriminator klasificira kao pozitivne podatke dok generirane i umjetne podatke klasificira kao negativne. Rad diskriminatora se može vidjeti na slici dolje.



Slika 23. Prikaz treniranja diskriminatora

Izvor: Google, Machine Learning advanced courses: The Discriminator: https://developers.google.com/machine-learning/gan/discriminator

Ovaj način rada je bio prethodno predstavljen kao GAN (Generativna kontradiktorna mreža, engl. Generative Adversarial Network) koji bi trebao sam prepoznavati često postavljene upite i time pokušati predvidjeti što korisnik želi napisati, "ali je njegova funkcija zakazala zbog nestabilnosti".

SDXL Turbo je nova verzija te ideje koja koristi ADD [6]. On koristi novu vrstu diskriminatora zvan DINOv2 [8] za ubrzavanje treniranja modela. Ovaj način rada dolazi sa svojim ograničenjima gdje je model prisiljen raditi sa svim pikselima zasebno i samim time mu je vrijeme procesiranja znatno povećano kao i korištenje resursa memorije odnosno vrijeme treniranja. Zbog tih limitacija, SDXL turbo može raditi samo sa slikama malih rezolucija te mu je korištenje LoRA [7] znatno ograničeno.

Kao i u prijašnjim ispitivanjima, testirat će se funkcionalnost ovoga modela kroz nekoliko različitih testova. Prvi od njih će biti generiranje tri slike sa tri različita stila gdje će se usporediti kvaliteta slika kao i brzina generiranja. Na kraju će se ponovo ispitati koliko dobro model raspoznaje riječi različitih vektorskih veličina [2].

Za ispitivanje rada SDXL Turbo, koristit će se alat ComfyUI. ComfyUI nije online alat, već sučelje na kojemu se mogu različiti prozori grafički spojiti s različitim vezama i time rekreirati rad bilo kojeg Stable Diffusion modela. Na slici ispod se može vidjeti struktura koja prikazuje jedan SDXL turbo model.

Slika 24. Prikaz sučelja SDXL turbo preko ComfyUI

Izvor: ComfyUI alat

Prije početka ispitivanja rada, prije će se pojasniti rad ovoga sklopa. Prvi prozor sa slike pod imenom "Load Checkpoint" ima ulogu da željenoj slici odredi stil crtanja. Taj Checkpoint određuje kako će slika izgledati i što sve alat uopće može nacrtati. Za ovaj primjer je odabran Checkpoint "moxieDiffusionXL_v12". Ovaj prozor se pritom grana u sljedeća tri prozora, "SDTurboSchelduer" i dva prozora tekstualnog enkodera pod imenom "CLIP Text Encode (prompt)". "Load Checkpoint" i "SDTurboSchelduer" su povezani sa vezom "MODEL". Taj spoj daje doznanja prozoru "SDTurboSchelduer" o kojem se modelu radi. Uloga te veze je važna jer je svaki Checkpoint napravljen da radi na različitom modelu Stable Diffusion-a. Prvi prozor tekstualnog enkodera je spojen s Checkpoint prozorom preko veza "MODEL", "CLIP" i "VAE". Veza "MODEL", kao što je i prije navedeno, služi da daje prozoru tekstualnog enkodera doznanja o kojem modelu Checkpoint-a se radi. "CLIP" veza služi da šalje određene informacije tekstualnom enkoderu kako bi on pritom mogao, preko napisanih riječi od strane korisnika, generirati željene slike. "VAE" (engl. Variational Autoenkoder) spoj daje bolju kvalitetu slici koje se rade putem tekst-u-sliku prozora. Drugi prozor tekstualnog enkodera ima ulogu negativa. Uloga negativa je ta da onemogućuje određenim riječima napisanih u njegov tekstualni enkoder

da se pojave u gotovoj slici. Njegova uloga daje puno mogućnosti pri generiranju slika time da minimizira neželjene rezultate. Uloga prozora "SDTurboSchelduer" je da korisnik odabere koliko šuma želi da mu se na slici generira te u koliko koraka želi da se taj šum ukloni. Podešavanje ovoga parametra ovisi najviše o vrsti Checkpoint-a.

Druga dva prozora u istom stupcu pod nazivom "Empty Latent Image" i "KSamplerSelect" imaju ulogu oblikovanja slike. Prozor "Empty Latent Image" služi da korisnik sam namjesti željenu veličinu generirane slike te koliko slika želi da mu program generira odjednom za naveden unos. "KSamplerSelect" određuje način na koji se otklanja šum i time daje drugačiji prikaz slike. U ovome ispitivanju koristio sampler "euler_ancestral".

Svi ti prozori se zajedno povezuju u prozor "SamplerCustom". Taj prozor ima nekoliko važnih opcija podešavanja poput "noise_seed" i "cfg". "noise_seed" parametar na početku ima vrijednost nula. Ta vrijednost se mijenja kada se slika generira. Svaka slika ima svoju vrijednosti te se ona može koristiti ako korisnik želi da mu se slika ponovo generira ali da izgleda jako slično prethodnoj. CFG (engl. Classifier-Free Guidance) skala određuje koliko strogo program mora slijediti Checkpoint pri generiranju slika. Ova opcija je veoma korisna ako korisnik želi koristiti neki LoRA [7] model jer kod korištenja takvih modela, može doći do izobličenja slika. Ovime korisnik podešava vrijednost na CFG skali dok ne dobi savršenu ravnotežu LoRA [7] i Checkpoint-a. Zadnji prozor "VAE Decode" dekodira vektorske informacije natrag u sliku s pikselima i iz njega proizlazi rezultat generirane slike.

3.4.1. Ispitivanje različitih stilova generiranja slika pomoću modela SDXL Turbo

Na slikama ispod se prikazuje tri različita stila crtanja kao i proces pri pisanju cijelog upita u tekstualni prozor. Za prikaz ovih slika, podesili su se parametri kod odabira količine koraka kod otklanjanja šuma s 1 koraka na 10, te se tip samplera promijenio s "euler_ancestral" na "euler" radi bolje kvalitete slika.

Slika 25 prikazuje kako su se generirane slike mijenjale dok se nije cijela rečenica napisala. Riječi generiranja su iste kao i u prijašnjim primjerima: "1 person, woman, face, face closeup, green eyes, freckles, brown hair, smiling, looking at viewer, realistic". Može se točno vidjeti prijelaz gdje se su se unijele riječi "green eyes" i "freckles" kao i na zadnjoj slici kada je unesena riječ "smiling".

Slika 25. Stil realizma

Izvor: ComfyUI tool, unos od strane studenta

Slika 26 prikazuje različite generirane slike istog psa dok se cijela rečenica pisala. Riječi generiranja slike glase: "dog, red collar, white fur, big dog, park, background, cartoon". Prve tri slike sa lijeva na desno izgledaju više manje isto, dok je uočljivo da se na zadnjoj slici stil potpuno promijenio kada se napisala zadnja riječ "cartoon".

Slika 26. Stil crtanog filma

Izvor: ComfyUI tool, unos od strane studenta

Na slikama ispod se može vidjeti da je samo prva slika u nizu različita od drugih te se nakon prvih par napisanih riječi, model pogubio u prepoznavanju riječi te je ostatak slika ostao gotovo identičan. Razlog tome je u svojstvu odabranog Checkpoint-a koji nije dovoljno dobro istreniran da generira ovakvu tematiku. Riječi generiranja su iste kao i u prethodnom ispitivanju.

Slika 27. Stil digitalne umjetnosti

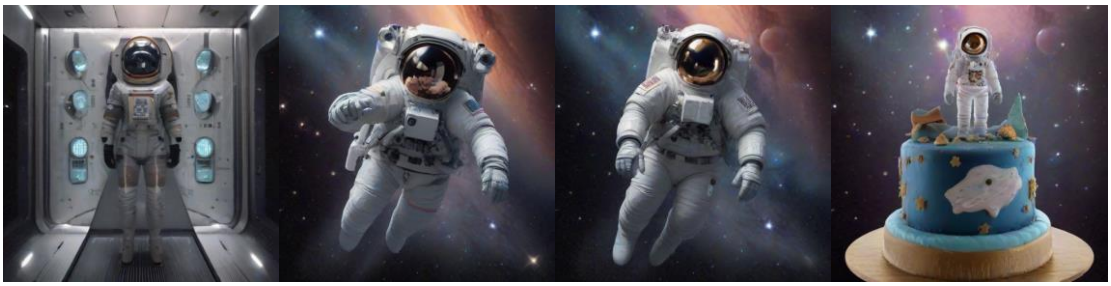
Izvor: ComfyUI tool, unos od strane studenta

Slike gore prikazuju proces kako su se mijenjale prilikom pisanja cijeloga teksta opisa. Iz njih se može vidjeti da je Checkpoint jako dobro istreniran u crtanju realističnih slika ljudi i realizma, ali zato pati kod crtanja digitalne umjetnosti i nerealističnih scena. Što se tiče brzine generiranja, svaka slika se uspjela generirati u prosijeku od pola sekunde. Kvaliteta slika i rezolucija u usporedbi s drugim modelima nije najbolja, ali to je za očekivati uzimajući u obzir da se ovaj model bazira na brzini u zamjenu za lošiju kvalitetu i manje detalja.

3.4.2. Ispitivanje generiranja apstraktnih tema pomoću modela SDXL Turbo

Zadnji test kao i kod prijašnjih ispitivanja rada modela ostaje na provjeri koliko se dobro model slaže s izrazima apstraktnih tema s opisom riječi koje imaju znatno različite vrijednosti prema Word2vec [2]. Kao i u prijašnjem testiranju, riječi generiranja su ostale iste: "Astronaut in space standing on a cake". Na slici ispod se može vidjeti kako su se generirane slike mijenjale dok se rečenica pisala. Samo prva i zadnja slika su se vidno promijenile dok su druga i treća ostale gotovo identične. Iz slika dolje se može vidjeti da ovaj model nema problema kod

generiranja slika apstraktnih tema. U suštini, SDXL turbo obavlja svoj posao kako je i namijenjen. Uspijeva bez problema prevesti sve napisane riječi u generirane slike s velikim brzinama. Njegova najveća trenutna mana je mala rezolucija i manjak dubine slike. Ovisno o svojoj upotrebi ovaj model može biti jako koristan alat.

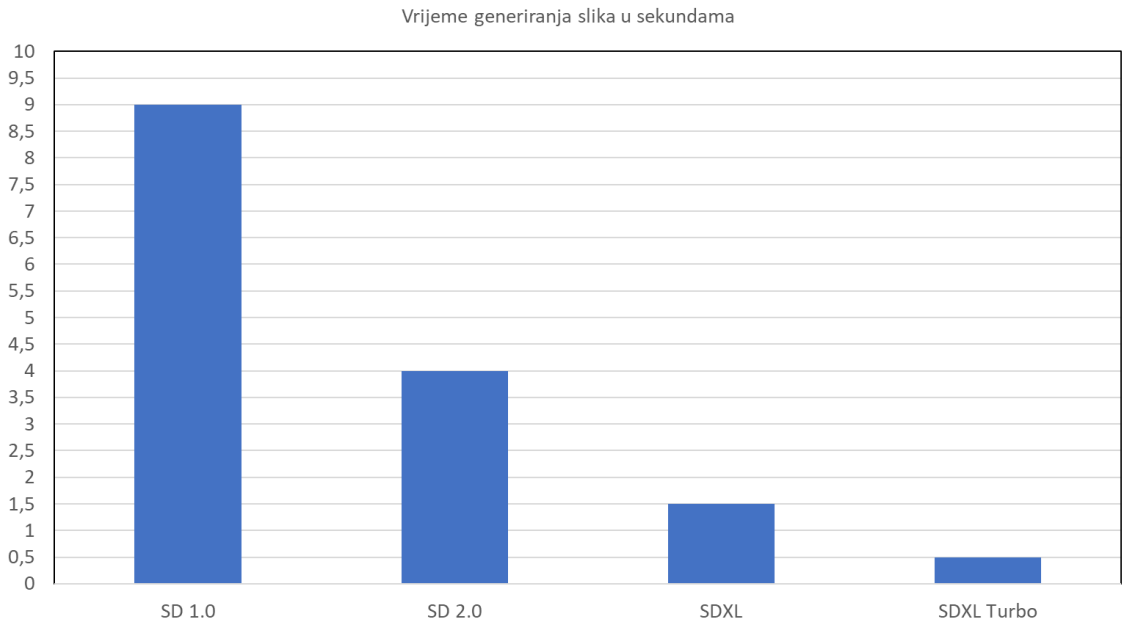


Slika 28. Slika apstraktnih tema

Izvor: ComfyUI tool, unos od strane studenta

3.5. USPOREDBA RAZLIČITIH STABLE DIFFUSION MODELA

Svaki od modela ima svoje prednosti i mane. Sada kada je opisan i testiran rad svih navedenih modela, vrijeme je da se usporedi njihova brzina rada kao i kvaliteta generiranja slika. Za početak će se usporediti prosječno vrijeme potrebno za generiranje jedne slike za svaki model. Brzina generiranja ovisi o nekoliko faktora, uključujući složenost modela, optimizaciju algoritama, te dostupnost hardverskih resursa poput GPU-a. U opisanom eksperimentu, izmjereno je vrijeme potrebno za generiranje slike zadane rezolucije koristeći identične ulazne parametre za sve generirane slike. Rezultati prikazani na grafu ispod jasno pokazuju kako se pojedini modeli razlikuju u brzini obrade, s nekima koji su optimizirani za brže generiranje, dok drugi posjeduju veću preciznost i detalje uz manjak brzine.



Slika 29. Grafikon koji prikazuje brzine generiranja slike za svaki model
Izvor: Napravljeno od strane studenta

Error! Reference source not found.. pokazuje da je SDXL Turbo znatno superioran po t om pitanju zahvaljujući njegovoj tehnologiji rada DINOv2 [8] diskriminatora. Ostali modeli zaostaju za njime te se može vidjeti kako sa svakom novom verzijom Stable Diffusion-a, brzina generiranja raste.

Nakon analize brzine, ključno je razmotriti i kvalitetu generiranih slika. Kvaliteta se ocjenjuje prema nekoliko kriterija, uključujući oštrinu, realističnost slike, razinu detalja, te sposobnost modela da rekreira zadane prizore. Sljedeće će se usporediti kvaliteta generiranja slika svakoga modela iz prijašnjih ispitivanja subjektivnim ocjenama loš, dobar i odličan. Ove ocjene se mogu mijenjati ovisno o Checkpoint-u koje model koristiti kao i maksimalnoj mogućoj rezoluciji generirane slike, stoga će se uzeti prosječna ocjena kvalitete slike koji pojedini modeli mogu generirati.

Tablica 1. Subjektivna usporedba generiranih slika za svaki stil svakog modela

Model	Stil realizma	Stil crtanog filma	Stil digitalne umjetnosti	Apstraktne teme
SD 1.0	loš	dobar	dobar	dobar
SD 2.0	dobar	odličan	odličan	dobar
SDXL	odličan	dobar	dobar	odličan
SDXL Turbo	odličan	dobar	loš	loš

Izvor: Napravljeno od strane studenta

Svaki od modela ima svojih prednosti i mana. Neku su dobri u brzini uz smanjenu kvalitetu slika dok su neki namijenjeni da rade duboke i detaljne slike za koje je potrebno više resursa i samim time veće vrijeme generiranja. Neki od modela su dobri za generiranje jednog stila dok imaju problema kod generiranja drugog. Na kraju korisnik je onaj koji odlučuje što mu najviše odgovara u ovisnosti o zadaći koju je potrebno ispuniti.

4. TESTIRANJE GRANICE MODELA ZA GENERIRANJE SLIKE BAZIRANIM NA UMJETNOJ INTELIGENCIJI

U ovom poglavlju ispitat će se do koje granice modeli umjetne inteligencije rade dobro za generiranje zadovoljavajućih slika. Ta ispitivanja će se izvoditi u obliku testa sličnog Alan Turing-ovom testu. No umjesto da robot pokušava zavarati čovjeka da je on sam čovjek, ovime ispitivanjem će se ispitati ako su trenutni modeli dovoljno dobri da naprave fotorealističnu sliku koju prosječni čovjek ne može razlikovati od prave slike neke osobe. Drugo ispitivanje će obuhvatiti temu umjetnosti gdje će se pomoću jednog od modela pokušati replicirati stil nekog poznatog umjetnika i vidjeti ako je prosječna osoba u mogućnosti prepoznati surogat od originala.

4.1. GENERIRANJE FOTOREALISTIČNE SLIKE

Fotorealizam u likovnoj umjetnosti predstavlja smjer crtanja i prikazivanja stvarnosti na papiru. Ova grana umjetnosti se bavi slikama koje se čine toliko stvarnim, da mogu zavarati ljude da su fotografije. S obzirom na to da modeli generiranja slika uče od postojećih slika, koristit će se modeli i Checkpoint-i koji su već prethodno trenirani na sličnim slikama.

Za generiranje testne slike koristit će se jedan od prijašnje ispitanih alata pod nazivom Stable Diffusion web UI. Ovaj alat je odabran zbog svojih raznih opcija uređivanja slika kao i toga što se za taj model rade veliki broj Checkpoint-a i LoRa [7]. Dobar Checkpoint će biti ključan element u procesu generiranja testne slike.

Za početak će se odabrati Checkpoint koji je treniran na fotorealističnim slikama. Odabrani Checkpoint je "Realistic Vision V6.0 B1". Ovaj Checkpoint je odabran zbog svoje iznimno dobre kvalitete u generiranju realističnih slika ljudskih lica. Kod postavke "Sampling method" odabrana je opcija "DPM SDE++ Karras". Postavka "CFG Scale" je spuštena na 2, te

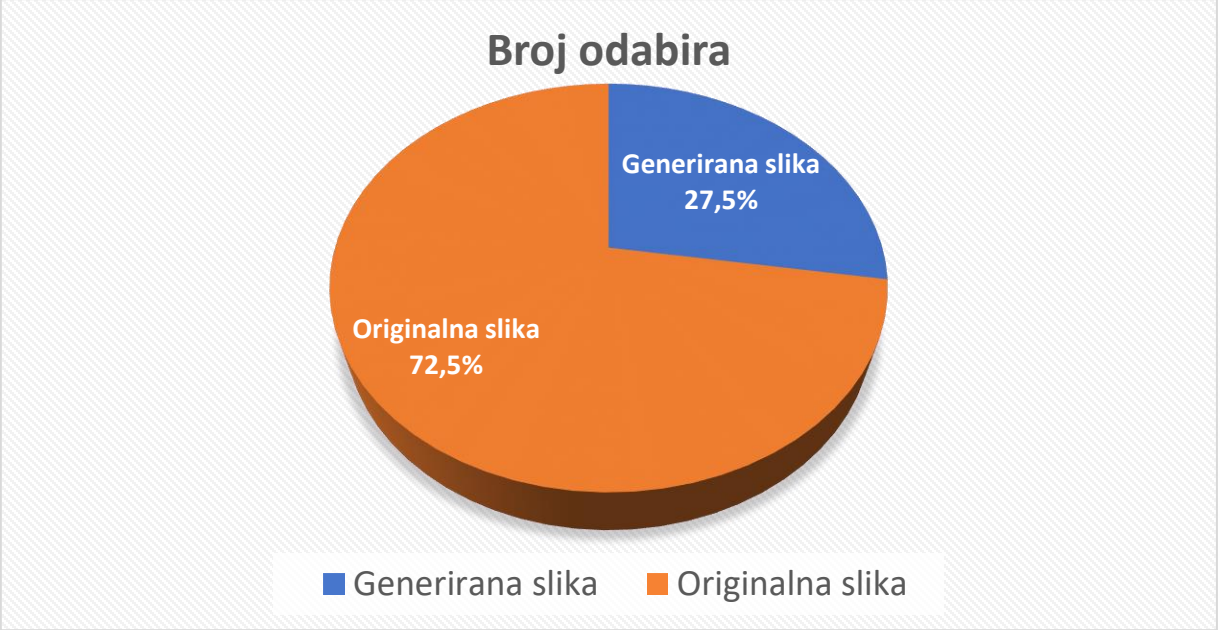
je rezolucija slike podešena na 768x768 piksela. Ove postavke su odabrane po preporuci developera ovog Checkpoint-a. U prozor negativa odabrane su LoRa-e [7] "BadDream" i "badhandv4" koje, kao što je prije objašnjeno, služe da smanje količinu izobličenja u slici. Uz te dvije korištene LoRa-e [7] dodalo se još nekoliko riječi u negativ koje su sljedeće: "(deformed iris, deformed pupils, semi-realistic, cgi, 3d, render, sketch, cartoon, drawing, anime, mutated hands and fingers:1.4), (deformed, distorted, disfigured:1.3), poorly drawn, bad anatomy, wrong anatomy, extra limb, missing limb, floating limbs, disconnected limbs, mutation, mutated, ugly, disgusting, amputation". Riječi koje su se koristile za generiranje slike su sljedeće: "young woman, smiling, dynamic pose, makeup, white shirt, beach background, people in background, European seaside city, beach, sea, trees, blonde hair, Realistic, best quality, sharp focus, DSLR, Studio lighting, DoF, instagram photo, front shot, photo, beautiful face, cinematic shot, realistic skin, high quality fabric, 4k, 8k".

Generiranoj slici se povećala rezolucija za dva puta te prije rada tog procesa, podesile su se postavke koje određuju kvalitetu slike pri povećanju rezolucije. Na Postavci "Upscaler" je odabrana opcija "R-ESRGAN 4x+" koja obavlja dobar posao kod povećanja rezolucije realističnih slika, te je opcija "Denoising strength" podešena na 0,5 da se slika sa povećanom rezolucijom ne mijenja od prvobitne. Generirana slika se može vidjeti ispod.

Slika 30. Primjer generirana fotorealistične slike

Izvor: Stable Diffusion web UI, unos od strane studenta

Generirana slika izgleda vrlo realno i detaljno napravljeno. Zahvaljujući Checkpoint-u "Realistic Vision V6.0 B1", svi detalji lica poput bora, tamnih mrlja u koži ili ispucanih vrhova vlasi kose se mogu detaljno uočiti. Za ispitivanje koliko je ova slika zapravo realistična, svrstana je s tri druge prave fotografije ljudi koje su onda date nekolicini ispitanika da pokušaju prepoznati generiranu sliku znajući na umu samo da je jedna od tih četiri slike generirana pomoću modela umjetne inteligencije. Ispit je napravljen preko alata za provođenje anketa Google forms. Ispitanicima je bilo postavljeno pitanje da od četiri ponuđene slike odaberu onu za koju misle da je napravljena koristeći umjetnu inteligenciju. Na grafu ispod se mogu vidjeti rezultati glasanja.



Slika 31. Grafikon koji prikazuje rezultate ispitivanja fotoreasilstičnih slika
Izvor: Napravljeno od strane studenta

Kao što se može vidjeti na rezultatima grafikona, samo je 27,5 % (11 glasova od ukupnih 40) ispitanika moglo raspoznati sliku generiranu pomoću umjetne inteligencije od originalne. Prema različitim komentarima ispitanika, troje je naglasilo da nisu uopće mogli prepoznati generiranu sliku dok je samo 5% ispitanika prepoznalo sliku zbog toga što oni sami koriste takve alate. Ovo ispitivanje je dokazalo da novi modeli za generiranje slika predstavljaju problem kod ljudi koji nisu upoznati s tom tehnologijom dovoljno dobro da znaju kako radi te bi kao

posljedice toga mogli biti zavarani da je slika koja je generirana s pomoću modela umjetne inteligencije stvarna.

4.2. USPOREDBA GENERIRANJA STILA CRTANJA

Ono što čini dobrog umjetnika je njegov ili njezin prepoznatljiv stil kojim crta slike. Jedne od najpoznatijih umjetnina danas koje su pomogle obilježiti novi slikarski pokret upravo su nacrtane jedinstvenim stilom. Danas je jako popularna digitalna umjetnost koja gubi svoju prepoznatljivost u načinu crtanja i dodavanja boja. Takve digitalne slike su crtane preko računala i zato osoba koja ih poučava ne može vidjeti način kojim je umjetnik prolazio s kistom po platnu ili koliko slojeva boje sama slika ima. Takve slike se danas sve lakše uspijevaju krivotvoriti tako da se model trenira slikama tog istog umjetnika. Pitanje dolazi ako je s modelom za generiranje slika moguće replicirati stil crtanja poznatog umjetnika koji crta fizičke umjetnine sa svojim jedinstvenim stilom. Stil koji će se probati replicirati u ovom ispitivanju će biti kubizam.

Kubizam je umjetnički stil koji je nastao početkom 20. stoljeća u Francuskoj i smatra se kao jedan od najutjecajnijih umjetničkih pokreta tog stoljeća. Njegovi osnivači su Georges Braque i Pablo Picasso zaslužni su u otkrivanju novog jedinstvenog stila crtanja. Pablo Picasso je jedan od najpoznatijih svjetskih umjetnika, zaslužan za dodavanje geometrije u svoje slike. Zbog toga, njegova kasnija djela su jako prepoznatljiva po svojim geometrijskim obličjima i jednostavnim bojama.

U ovom pokusu, probat će se generirati slika s pomoću modela generiranja slika i provesti ispitivanje ako je prosječna osoba u mogućnosti raspoznati pravu sliku Picasso-a od surogata. Za generiranje ovog stila koristit će se jedan od prijašnje ispitanih alata pod nazivom Stable Diffusion web UI. Ovaj alat je odabran zbog svojih raznih opcija uređivanja slika kao i toga što se za taj model radi veliki broj Checkpoint-a i LoRa [7]. U procesu generiranja ove slike bit će potreban dobar Checkpoint i dobra LoRa [7]. Uloga Checkpoint-a će biti ta da

generira slike na način da izgledaju kao da su crtane uljanim bojama, dok će uloga LoRa-e [7] biti ta da replicira stil crtanja Picasso-a.

Prvo je odabran Checkpoint koji može replicirati stil crtanja uljanim bojama. Odabrani Checkpoint je "watercolorOilPainting_v10". Nakon toga, u tekstualni prozor je dodana LoRa [7] koja je trenirana sa Picasso-tovim slikama pod nazivom "bijiasuov0". Kada su Checkpoint i LoRa [7] odabrani, vrijeme je za podešavanje drugih postavki. Na postavki "Sampling method" je postavljena opcija "DPM++ 3M SDE" koja u ovom primjeru daje zadovoljavajuće rezultate. Postavka "CFG Scale" je podignuta na vrijednost osam tako da alat pri generiraju što bolje slijedi napisane riječi u tekstualnom prozoru. Rezolucija slike stavljena na 512x712 piksela kako bi izgledala kao portret. U prozor negativa su opet odabrane LoRa-e [7] "BadDream" i "badhandv4" koje, kao što je prije objašnjeno, služe da smanje količinu mutacija u slici. Iako se ovo čini nepotrebnim u slici koja bi sama po sebi trebala izgledati apstraktno, još uvijek Checkpoint može dodati elemente u sliku koji će odvući izgled slike od željenog stila.

Riječi koje su se koristile za generiranje slike su sljedeće: " woman, colors, Picasso style, sitting, black outlines, vibrant, style, sharp, 1 person, masterpiece, detailed".

Slika 32. Primjer generirana slike stila Pablo Picasso-a

Izvor: Stable Diffusion web UI, unos od strane studenta

Slika 32 izgleda dosta dobro s obzirom na to da je generirana s pomoću modela umjetne inteligencije. Na njoj se može dobro vide tragovi bojanja uljenim bojama koje je Checkpoint stavio te joj daju izgled autentičnosti. Sada je na redu ispitanicima da prokušaju raspoznati surogat od originala. Za onoga tko se bavi umjetnošću ili je barem donekle proučavao slike od Picasso-a, bit će im veoma jasno da ovoj slici nedostaje kontrast jednostavnih i oštrih boja kao i crnih linija koje odvajaju svaki geometrijski lik. Sada pitanje ostaje ako je slika dovoljno dobra da zavara prosječnu osobu. Kao i u prijašnjem ispitivanju, odabrane su tri prave Picasso-tove slike i ispitanik mora odabrati jednu od njih za koju smatra da je surogat. Ovaj test je isto napravljen pomoću alata za anketiranje Google Forms. Ispitanicima je bilo postavljeno pitanje da od četiri ponuđene slike odaberu onu za koju misle da je napravljena koristeći umjetnu inteligenciju. Na grafu ispod se mogu vidjeti rezultati glasanja.

Slika 33. Grafikon koji prikazuje rezultate ispitivanja slike stila Pablo Picasso-a
Izvor: Napravljeno od strane studenta

Za razliku od prijašnjeg testiranja gdje većina ispitanika nije mogla raspoznati pravu osobu od one generirane pomoću umjetne inteligencije, u ovom slučaju, 55% ispitanika je uspjelo uočiti surogat. U ovom slučaju ispitanici su komentirali da su prepoznali vidljivu razliku

u oštrini boje i kontrastu slike bez ikakvog problema dok je jedan od ispitanika bio upoznat sa svakom slikom Picasso-a pa su odmah znali koja slika nije originalna.

Generiranje slika pomoću umjetne inteligencije je dosta napredovalo u posljednjih osam godina, ali dok su se neke stvari gotovo usavršile kao generiranje ljudskog lica, ostale još imaju dug put dok se usavrše. Naravno ovaj primjer je bio napravljen koristeći prava originalna umjetnička djela, dok današnje slike digitalne umjetnosti bi mogle biti puno lakše repliciranje. Sve ove slike mogu predstavljati veliki problem onima koji nisu upoznati tehnologijom generiranja slika ili onima koji nisu vidjeli dovoljno generiranih slika da mogu uočiti pravu sliku od lažne.

5. ZAKLJUČAK

Generiranje slika s pomoću umjetne inteligencije prošlo je kroz značajan razvoj od svojih ranih dana do danas. Počevši s sustavom Aaron iz 1970-ih godina, evolucija ove tehnologije ubrzana je s pojavom naprednih tehnologija poput GAN-a i modela poput DALL-E, koji su omogućili integraciju složenih tekstualnih i vizualnih elemenata. Danas, zahvaljujući novim tehnologijama poput CLIP-a, stvoren je model Stable Diffusion koji je promijenio način na koji se gleda rad umjetne inteligencije [10].

U ovom radu su se istraživale nove tehnologije i tehnike generiranja slika s pomoću umjetne inteligencije. Istraživao se i ispitivao rad modela za generiranje slika s pomoću umjetne inteligencije te su se uspoređivali različiti modeli koji su produkt tih novih tehnologija. Istražila se povijest kako je ova tehnologija napredovala iz generiranja jednostavnih crno-bijelih slika do generiranja fotorealističnih slika koje već danas počinju zavarati ljude da su autentične.

Od kad je pojam umjetne inteligencije stvoren, ljudi su se bojali budućnosti gdje umjetna inteligencija upravlja njihovim životima i gdje pitanje morala više ne vrijedi. Umjetna inteligencija nije nikad bila napravljena s idejom da se koristi kao surogat za čovjeka, niti su modeli generiranja slika stvoreni da se ljudska umjetnost zamijeni s onom generiranom preko modela, već da služi kao alat u istraživanju novih tehnologija i olakšavanja posla čovjeka.

Iz različitih ispitivanja u ovome radu može se zaključiti da usprkos njihovoj mogućnosti generiranja slika iz mašte i gotovo savršenih lica, modelima generiranja slika još uvijek nedostaje originalnost i jedinstvo koje samo čovjek posjeduje. Zbog ovoga, ti modeli su trenutno limitirani trenutnim tehnološkim napretkom i za sljedeći korak u evoluciji umjetne inteligencije će biti potrebno jako puno godina.

Usprkos svojim korisnim funkcijama i dobroj naravi, još uvijek postoje problemi koje je ova tehnologija izazvala. Jedan od najvećih problema je kršenje autorskih prava. Modeli se često treniraju na velikim skupovima podataka koji sadrže slike, uključujući one zaštićene autorskim pravima. Ovo može dovesti do stvaranja slika koje su previše slične originalima,

kršeći prava umjetnika. To se najviše može vidjeti kod digitalne umjetnosti gdje se je tehnika crtanja umjetnika kistovima zamijenjena s dvodimenzionalnim kistovima i bojama koje su zahvaljujući svojem digitalnom obliku, znatno lakše uzeti i iskoristiti za treniranje novih modela.

Modeli se mogu koristiti za generiranje lažnih slika koje mogu zavarati ljude. Ovo je posebno opasno u kontekstu politike ili pokreta gdje jedna kriva slika može pokrenuti zastrašujuće pojave. Iako modeli generiranja slika mogu pomoći u stvaranju umjetnosti, postoji problem da bi mogli zamijeniti ljudsku kreativnost i rad umjetnika, smanjujući vrijednost ručno izrađene umjetnosti i originalnih djela. Modeli generiranja slika mogu isto tako generirati neprikladan ili uvredljiv sadržaj, bilo slučajno ili namjerno, što može dovesti do moralnih i društvenih problema. Kvaliteta generiranih slika u velikoj mjeri ovisi o podacima na kojima je model treniran.

Kada se oslanjamo na modele za generiranje slika, postoji rizik da izgubimo osobni izraz i jedinstvenost koju ljudi unose u svoja umjetnička djela. Iako je ovo samo alat, treba uzeti u obzir da alat služi da se koristi za olakšavanje zadatka, a ne da zamijeni onoga tko ga koristi.

LITERATURA

[1] N. Siddique, S. Paheding, C. P. Elkin and V. Devabhaktuni, 'U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications, ' in IEEE Access, vol. 9, pp. 82031-82057, 2021, doi: 10.1109/ACCESS.2021.3086020. (9.7.2024)

[2] CHURCH KW. 'Word2Vec. *Natural Language Engineering*'. 2017;23(1):155-162. doi:10.1017/S1351324916000334. (9.7.2024.)

[3] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, Robin Rombach: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, https://arxiv.org/pdf/2307.01952 (9.7.2024.)

[4] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations, https://arXiv:2108.01073 (9.7.2024.)

[5] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, Edward Raff: VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance, https://arxiv.org/pdf/2204.08583 (9.7.2024.)

[6] Shanchuan Lin, Anran Wang, Xiao Yang: SDXL-Lightning: Progressive Adversarial Diffusion Distillation, https://arxiv.org/pdf/2402.13929 (9.7.2024.)

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen: LoRA: Low-Rank Adaptation of Large Language Models, https://arxiv.org/abs/2106.09685 (20.7.2024.)

[8] Maxime Oquab, Timoth´ee Darcet, Th´eo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Je- gou, Julien Mairal, Patrick Labatut, Armand Joulin, and Pi- otr Bojanowski: DINOv2: Learning robust visual features without supervision, https://arxiv.org/pdf/2304.07193 (20.7.2024.)

[9] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, Jenia Jitsev: LAION-5B: An open large-scale dataset for training next generation image-text models, https://arxiv.org/abs/2210.08402 (20.7.2024.)

[10] Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D. and Houlsby, N., 2021. 'Scaling vision with sparse mixture of experts'. *Advances in Neural Information Processing Systems*, *34*, pp.8583-8595. (20.7.2024.)

[11] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T. and Ho, J., 2022. 'Photorealistic text-to-image diffusion models with deep language understanding'. *Advances in neural information processing systems*, *35*, pp.36479-36494. (26.8.2024.)

[12] Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L. and Jitsev, J., 2023. 'Reproducible scaling laws for contrastive language-image learning'. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2818-2829). (26.8.2024.)

[13] Giannis Daras, Alexandros G. Dimakis: Discovering the Hidden Vocabulary of DALLE-2, https://arxiv.org/pdf/2206.00169 (26.8.2024.)

[14] Sitzmann, V., Martel, J., Bergman, A., Lindell, D. and Wetzstein, G., 2020. 'Implicit neural representations with periodic activation functions'. *Advances in neural information processing systems*, *33*, pp.7462-7473. (26.8.2024.)

POPIS KRATICA

Kratica	Puni naziv na engleskom jeziku	Tumačenje na hrvatskom jeziku
ADD	Adversarial Diffusion Distillation	Način treniranja Diffusion modela
CFG	Classifier-Free Guidance	Parametar koji kontrolira koliko proces generiranja slike slijedi tekstualni upit
CLIP	Contrastive Language-Image Pretraining	Neuronska mreža trenirana na različitim parovima slika i teksta
CUDA	Compute Unified Device Architecture	Paralelna računalna platforma koja softveru omogućuje korištenje određenih vrsta grafičkih procesorskih jedinica
GAN	Generative Adversarial Network	Arhitektura dubokog učenja
GPT	Generative Pre-trained Transformer	Modeli neuronske mreže obučeni na velikim skupovima podataka

LoRA	Low-Rank Adaptation Model	Manji model treniran za specifičnu funkciju
SDXL	Stable Diffusion Extra Large	Model Stable Diffusion-a
VRAM	Video Random Access Memory	Vrsta RAM memorije koja služi za pohranjivanje slikovnih podataka

POPIS SLIKA

POPIS TABLICA