Stvaranje slika iz prirodnog jezika koristeći modele dubokog učenja

Krajačić, Matija

SVEUČILIŠTE SJEVER
SVEUČILIŠNI CENTAR VARAŽDIN

DIPLOMSKI RAD br.

Stvaranje slika iz prirodnog jezika koristeći modele dubokog učenja

Matija Krajačić

Varaždin, rujan 2023.

SVEUČILIŠTE SJEVER

SVEUČILIŠNI CENTAR VARAŽDIN

Studij Multimedija

DIPLOMSKI RAD br.

Stvaranje slika iz prirodnog jezika koristeći modele dubokog učenja

Student:
Matija Krajačić, 0313022074

Mentor:
izv. prof. dr. sc. Emil Dumić

Varaždin, rujan 2023.

Sveučilište Sjever
Sveučilišni centar Varaždin
104. brigade 3, HR-42000 Varaždin

Prijava diplomskog rada

Definiranje teme diplomskog rada i povjerenstva

ODJEL Odjel za multimediju

STUDIJ diplomski sveučilišni studij Multimedija

PRISTUPNIK Krajačić Matija | JMBAG 0313022074

DATUM 26.06.2023. | KOLEGIJ Računalni vid

NASLOV RADA Stvaranje slika iz prirodnog jezika koristeći modele dubokog učenja

NASLOV RADA NA ENGL. JEZIKU Image creation from natural language using deep learning models

MENTOR Emil Dumić | ZVANJE izv.prof.dr.sc.

ČLANOVI POVJERENSTVA

1. doc. art. dr. sc. Mario Periša - predsjednik
2. izv. prof. dr. sc. Dean Valdec - član
3. izv. prof. dr. sc. Emil Dumić - mentor
4. doc. dr. sc. Andrija Bernik - zamjenski član
5.

Zadatak diplomskog rada

BROJ 094-MMD-2023

OPIS

U ovom radu će biti opisani i ispitani različiti modeli dubokog učenja za stvaranje realističnih i umjetničkih slika uvjetovanih prirodnim jezikom.
Pretvorba teksta u sliku označava skup modela za stvaranje slika iz ulaznog teksta. Primjena navedenih modela može biti različita, od proširivanja korištenja računalnih botova, arhitekture, modne industrije, marketinga, umjetnosti, poslovnih primjena i drugo. Specifično, bit će objašnjeni modeli StabilityAI, DALL-E, DALL-E 2 i Imagen, bazirani na metodama dubokog učenja. Također će se opisati i transformer neuronske mreže često korištene u modelima za kodiranje teksta, kao i difuzijski modeli općenito, često korišteni u koraku generiranja slike. Opisat će se i drugi mogući zadaci modela, poput nadopunjavanja nedostajećih dijelova slike (inpainting), super-rezolucija, bezuvjetno generiranje slika, klasom uvjetovano generiranje slika iz postojećih baza slika. Bit će opisani i neki noviji modeli za generiranje videozapisa iz teksta.
U praktičnom dijelu rada će se analizirati neki otvoreni kodovi od objašnjenih modela za stvaranje slika iz prirodnog jezika. Usporedit će se kvaliteta generiranih slika za modele i izvesti zaključci za ograničenja koja su još prisutna kod njih. Kvaliteta slika će se usporediti koristeći neku bazu slika (npr. MS-COCO), pomoću nekih od postojećih objektivnih mjera vezanih za umjetno generiranje slika (Inception Score, ICS mjera, za različitost slika, Frechet Inception Distance, FID mjera, za vjernost slike i Contrastive Language-Image Pre-training, CLIP mjera za povezanost slike i teksta).

ZADATAK URUČEN 03.09.2023. | POTPIS MENTORA Emil Dumić

SVEUČILIŠTE SJEVER

Sažetak

U ovom radu će biti opisani i ispitani različiti modeli dubokog učenja za stvaranje realističnih i umjetničkih slika uvjetovanih prirodnim jezikom

Pretvorba teksta u sliku označava skup modela za stvaranje slika iz ulaznog teksta. Primjena navedenih modela može biti različita, od proširivanja korištenja računalnih botova, arhitekture, modne industrije, marketinga, umjetnosti, poslovnih primjena i drugo. Specifično, bit će objašnjeni modeli StabilityAl, DALL-E, DALL-E 2 i Imagen, bazirani na metodama dubokog učenja. Također će se opisati i transformer neuronske mreže često korištene u modelima za kodiranje teksta, kao i difuzijski modeli općenito, često korišteni u koraku generiranja slike. Opisat će se i drugi mogući zadaci modela, poput nadopunjavanja nedostajećih dijelova slike (inpainting), super-rezolucija, bezuvjetno generiranje slika, klasom uvjetovano generiranje slika i uvjetovanje bez klasifikatora. Bit će opisani i neki noviji modeli za generiranje videozapisa iz teksta.

U praktičnom dijelu rada će se analizirati neki otvoreni kodovi od objašnjenih modela za stvaranje slika iz prirodnog jezika. Usporedit će se kvaliteta generiranih slika za modele i izvesti zaključci za ograničenja koja su još prisutna kod njih. Kvaliteta slika će se usporediti koristeći bazu slika MS-COCO, pomoću nekih od postojećih objektivnih mjera vezanih za umjetno generiranje slika (Inception Score, ICS mjera, za različitost slika, Frechet Inception Distance, FID mjera, za vjernost slike i Contrastive Language-lmage Pre-training, CLIP mjera za povezanost slike i teksta).

Ključne riječi: **stvaranje slika, stvaranje slika iz teksta, duboko učenje, modeli dubokog učenja, txt2img, Stable Diffusion, DALL-E, DALL-E 2**

Abstract

This paper will describe and evaluate different deep learning models for generating realistic and artistic images conditioned on natural language. Converting text to an image denotes a set of models for generating images from an input text. The use cases of such models can vary, from extending the use of computer bots, to architecture, fashion industry, marketing, art, business applications and more. The specific models that will be described are StabilityAI, DALL-E, DALL-E 2 and Imagen, based on deep learning methods. Additionally, the neural network transformer which is commonly used in natural language processing will be described, as well as the diffusion models in general, often used in the step of generating images. Some other tasks of the models will be described, such as inpainting, super-resolution, unconditional image generation, class-conditional image generation and classifier-free guidance. Some new text to video models will also be described.

In the practical section of the paper, some of the released codes from the described text to image models will be analysed. The quality of the generated images by the models will be compared, and conclusions will be made on the limitations that are still present. The quality of the images will be compared using the image database MS-COCO, through some of the existing objective measures related to text to image synthesis (Inception Score, for image diversity, Frechet Inception Distance, for image fidelity and Contrastive Language-Image Pre-training, for text and image similarity).

Keywords: **image generation, generating images from text, deep learning, deep learning models, txt2img, Stable Diffusion, DALL-E, DALL-E 2**

Popis korištenih kratica

GAN Generativna suparnička mreža (engl. *Generative Adversarial Network*)
SRCNN Super-rezolucijska konvolucijska neuronska mreža (engl. *Super-Resolution Convolutional Neural Network*)
VDSR Vrlo duboka super-rezolucija (engl. *Very deep super-resolution*)
DCGAN Duboka konvolucijska generativna suparnička mreža (engl. *Deep Convolutional GAN*)
S^2-GAN Stilska i strukturarna generativna suparnička mreža (engl. *Style and Structure Generative Adversarial Network*)
RGBD Crveni, zeleni, plavi i dubinski podaci (engl. *Red Green Blue Depth*)
WGAN Wasserstein generativna suparnička mreža (engl. *Wasserstein GAN*)
PGAN Progresivna rastuća generativna suparnička mreža (engl. *Progressive Growing GAN*)
CFG Skala usmjeravanja bez klasifikatora (engl. *Classifier-free Guidance Scale*)
SDXL Stable Diffusion XL, model za stvaranje slika iz teksta
CLIP Kontrastivno predtreniranje jezika-slike (engl. *Contrastive Language-Image Pre-training*)
GLIDE Upravljana jezično-slikovna difuzija za generiranje i uređivanje (engl. *Guided Language-to-Image Diffusion for Generation and Editing*)
ICS Inception mjera (engl. *Inception score*)
FID FID mjera (engl. *Frechet Inception Distance*)
RAM Radna memorija (engl. *Random Access Memory*)
VRAM Video radna memorija (engl. *Video RAM*)
URL Putanja do sadržaja na internetu (engl. *Uniform Resource Locator*)
TXT2IMG Stvaranje slike iz teksta (engl. *text to image*)
IMG2IMG Stvaranje slike iz slike (engl. *image to image*)
PNG Slikovni format (engl. *Portable Network Graphics*)
WSL Windows podsustav za Linux (engl. *Windows Subsystem for Linux*)
MSCOCO Microsoft podatkovni set čestih objekata u kontekstu (engl. *Microsoft Common Objects in Context*)
CIFAR-10 Podatkovni set kanadskog instituta za napredna istraživanja od 10 klasa (engl. *Canadian Institute for Advanced Research, 10 classes*)
SD Stable Diffusion, model za stvaranje slika iz teksta

Sadržaj

1. Uvod

Ovaj rad bavi se područjem stvaranja realističnih i umjetničkih slika iz prirodnog jezika koristeći modele dubokog učenja. Stvaranje slika visoke kvalitete iz tekstualnih opisa zahtjevan je zadatak koji je proteklih godina pridobio pozornost zbog potencijalnih primjena u različitim područjima kao što su umjetnost, dizajn, zabava, arhitektura, modna industrija, marketing, poslovne primjene. Primjene mogu biti i za proširivanje korištenja računalnih botova.

U prvom djelu rada bit će opisani zadaci modela dubokog učenja, u kontekstu generiranja slika iz teksta. Uz samo generiranje slike iz teksta, opisat će se nadopunjavanje nedostajećih dijelova slike, super-rezolucija, bezuvjetno generiranje slika i klasom uvjetovano generiranje slika. Bit će objašnjeni modeli dubokog učenja, specifično Stable Diffusion, DALL-E, DALL-E 2 i Imagen, kao i općenito difuzijski modeli i transformer neuronske mreže koji je često korišten u modelima za kodiranje teksta. Bit će opisani i modeli za generiranje videozapisa iz teksta, te metode evaluacije modela za stvaranje slika.

U praktičnom dijelu rada izvest će se objektivna evaluacija dvaju modela za stvaranje slika. Usporedit će se kvaliteta generiranih slika i izvesti zaključci za ograničenja koja su još prisutna. Objektivna evaluacija bit će izvršena pomoću nekoliko postojećih objektivnih mjera, specifično mjera za kvalitetu i različitost slika, vjernost slika, te povezanost slika i teksta. Opisat će se proces pripreme modela za generiranje slika, te njihovog pokretanja i proces pripreme modela za objektivnu evaluaciju, te pokretanja istih. Dobiveni rezultati bit će analizirani te će se iz njih izvesti zaključci.

2. Zadaci modela

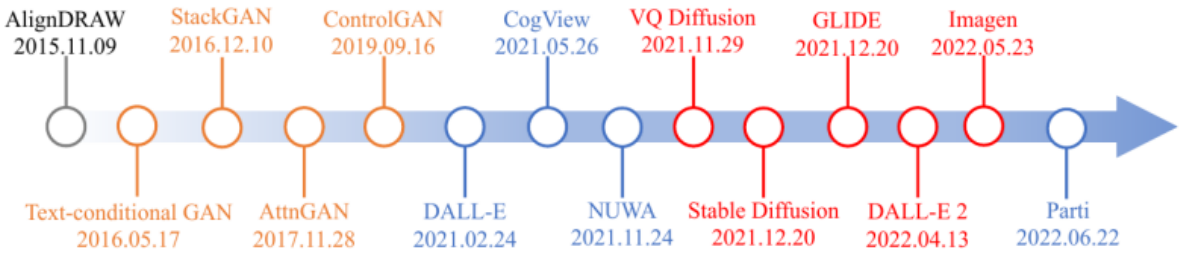
U ovom poglavlju bit će opisani neki od važnih zadataka modela dubokog učenja za generiranje slika. Svaki od tih zadataka uključuje stvaranje novih slika bazirano na dani ulaz ili grupu ograničenja. Svaki od zadataka ima jedinstven skup problema i mogućih primjena.

2.1. Generiranje slike iz teksta

Generiranje slika iz teksta (engl. *text-to-image*), je sinteza slika uvjetovana tekstom. Pionirski rad vezan uz generiranje slika iz teksta je AlignDRAW, predstavljen 2016. godine [1]. AlignDRAW iterativno stvara sitne dijelove slike, uzimajući u obzir riječi iz tekstualnog opisa. Tekstualni opisi su predstavljeni kao sekvenca uzastopnih riječi, dok su slike predstavljene kao sekvenca manjih dijelova nacrtanih na platnu kroz vrijeme. Model kombinira ponavljajući varijacijski autokoder, vrstu kodera koji pamti informacije, te model koji pomaže pri spajanju riječi (engl. *alignment model*), koji pronalazi značajne poveznice između riječi u sekvencama [2].

Kasniji modeli za generiranje slika iz teksta koriste vrstu neuronske mreže GAN (engl. *Generative Adversarial Network*), generativnu suparničku mrežu koja se sastoji od dvije neuronske mreže: generatora i diskriminatora. Korištena GAN mreža trenirala se na tekstualnim opisima. Kasniji pristupi poboljšali su kvalitetu slika koristeći višerazmjerne GAN mreže te dodavajući element pozornosti između teksta i značajka slike [3]. Za razliku od prvotnih GAN modela koji su kodirali cijele rečenice u jedan vektor kao uvjet za generiranje slike, GAN modeli sa pozornošću omogućuju izradu različitih podregija slike uvjetovanih riječima koje su najvažnije za pojedine podregije [4]. Autoregresivni GAN modeli omogućuju korištenje većih količina podataka u odnosu na prijašnje GAN modele. Ti modeli autoregresivno (uz transformer) modeliraju tekstualne i slikovne tokene kao jedan tok podataka [3]. Nedostatak takvih modela su veliki troškovi procesiranja podataka kao i sekvencijalno prikupljanje pogrešaka [1].

Trenutačno najnoviji pristup modelima za generiranje slika iz teksta su difuzijski modeli. Difuzijski modeli su klasa generativnih modela koji pretvaraju slike sa šumom u prirodne slike kroz serijsko uklanjanje šuma [5]. Primjeri difuzijskih modela su Stable Diffusion i DALL-E 2. Slika 2.1. prikazuje vremensku crtu reprezentativnih radova za generiranje slike iz teksta, gdje su žutom bojom označene metode bazirane na GAN mrežama, plavom bojom su označene autoregresivne metode, dok su crvenom bojom označeni difuzijski modeli.



Slika 2.1. Vremenska crta modela za generiranje slike iz teksta [1]

2.2. Nadopunjavanje nedostajećih dijelova slike

Nadopunjavanje nedostajećih dijelova slike (engl. *inpainting*) je zadatak generiranja realističnog sadržaja unutar dijelova koji nedostaju uz zadržavanje koherentnosti. Ima bitnu ulogu u zadacima procesiranja digitalnih slika kao što je obnova oštećenih slika, uređivanje fotografija te renderiranje slika [6]. Nadopunjavanje nedostajećih dijelova slike veoma je izazovan problem. Zadatak „inpaintinga“ zahtijeva sposobnost predviđanja onoga što nedostaje, uz određivanje pripada li dio u kontekst slike ili ne [7]. Ulazni podaci veoma su kompleksni. Različite vrste ulaza mogu zahtijevati različite strategije ili algoritme „inpaintinga“. Oštećenje na slikama može biti vrlo veliko, što često dovodi do nezadovoljavajućih rezultata za tradicionalne algoritme bazirane na nadopunjavanju, algoritme temeljene na parcijalnim diferencijalnim jednadžbama, ili interpolacijske algoritme. Dodatni problem je što rezultati „inpaintinga“ nisu jedinstveni, dok većina algoritama uzima u obzir samo jedan mogući rezultat [8].

Proteklih godina predložene su različite tehnike dubokog učenja za zadatak nadopunjavanja nedostajećih dijelova slika. Većina tehnika bazirano je na autokoderima, varijacijskim autokoderima, GAN modelima ili autoregresivnim transformerima [5]. Bazirano na vrstama problema koje rješavaju, modeli dubokog učenja za nadopunjavanje nedostajećih dijelova slike dijele se na progresivne modele, modele vođene strukturiranim informacijama, modele bazirane na pozornosti, konvolucijske modele, te pluralističke modele [8].

Progresivni modeli nadopunjuju slike tako da postupno koriste značajke neoštećenih i nedavno nadopunjenih dijelova. Ovi modeli iterativno prolaze kroz sliku te uče iz prijašnje predviđenih piksela [7]. Modeli vođeni strukturiranim informacijama oslanjaju se na strukturu poznatih regija, kao što su rubovi i segmentacija. Kada oštećene slike sadrže oštre detalje, nedostatak strukture u nedostajećim dijelovima naznaka je da nešto nedostaje. Modeli bazirani na pozornosti koriste informacije iz prostorno udaljenih lokacija da bi riješile problem konvolucijskih mreža koje nisu efektivne za posuđivanje udaljenih značajki [8]. Konvolucijski modeli koriste maske za određivanje pogođenih dijelova slike te upravljanje tokom informacija kroz brojne sekcije [9].

Pluralistički modeli iz slike sa maskom stvaraju više različitih mogućih rješenja za dovršavanje slike [10].

Difuzijski modeli sve se češće koriste za nadopunjavanje nedostajećih dijelova slike. Metode sa difuzijskim modelima mogu se podijeliti na metode sa nadzorom i metode bez nadzora. Nadzirane metode uključuju treniranje difuzijskog modela specifično za zadatak „inpaintinga“. To može stvoriti velike troškove procesiranja podataka, te su mogući lošiji rezultati sa prije neviđenim vrstama oštećenja. Metode bez nadzora koriste već trenirane difuzijske modele bez ikakvih modifikacija modela [5].

2.3. Super-rezolucija

Super-rezolucija na jednoj slici proces je stvaranja slike visoke rezolucije iz odgovarajuće slike niske rezolucije, koja je prošla kroz neku vrstu degradacije [11]. Spada pod kategoriju zadataka prevođenja iz slike u sliku. Super-rezolucija je zahtjevna jer različite izlazne slike mogu biti konzistentne sa jednom ulaznom slikom. [12]. Vrste degradacije kroz koje su prošle slike niske rezolucije mogu biti različiti oblici šuma, zamućenja ili ostalih artefakata, što dovodi do gubitka visokofrekventnih informacija. Cilj metoda super-rezolucije je oporavak što više visokofrekventih informacija.

Metode bazirane na neuronskim mrežama (konvolucijske mreže i transformer) pokazale su impresivan učinak pri super-rezoluciji [11]. Većina metoda sastoji se od tri faze koje uključuju inicijalno izdvajanje plitkih dijelova, zatim računalno intenzivno usavršavanje u prostoru dubokih značajki, te završno povećanje veličine radi ostvarivanja ciljane rezolucije [13]. SRCNN prvi je primjer konvolucijske mreže za super-rezoluciju, koja se sastoji od tri sloja. Mreža direktno uči o povezanosti između slika niske i visoke rezolucije uz malo potrebnog procesiranja [14]. Uslijedio je razvoj različitih konvolucijskih mreža sa boljim performansama, poput VDSR mreže sa 20 slojeva koja koristi rezidualno učenje. Rezidualna slika je razlika između slike visoke kvalitete i slike niske kvalitete. Pošto slike visoke i niske kvalitete uveliko sadrže iste informacije, eksplicitno modeliranje rezidualne slike donosi prednosti [15]. Predlagani su različiti mehanizmi za poboljšanje kvalitete rekonstrukcije slika, poput mehanizama prostorne i kanalne pozornosti [11], koji pomažu modelu da obraća pozornost na različite dijelove i aspekte slike. Nedavno, super-rezolucijske metode bazirane na transformerima postaju sve popularnije, te ostvaruju kvalitetne rezultate [16]. Mehanizam samopažnje transformera omogućuje shvaćanje povezanosti između udaljenih informacija. Brojne metode uspješno su uključile transformer u proces super-rezolucije. Te metode ostvaruju bolje performanse te pokazuju velik potencijal za buduća istraživanja [11].

2.4. Bezuvjetno generiranje slika

Bezuvjetno generiranje slika proces je generiranja novih slika bez specifičnog ulaza. Glavna zadaća je izrada novih, originalnih slika koje nisu bazirane na postojećim slikama. Može služiti za kreiranje novih umjetničkih slika, poboljšanje algoritama za prepoznavanje slika, ili generiranje fotorealističnih slika za okruženja virtualne stvarnosti. Modeli za bezuvjetno generiranje slika sposobni su kreirati fotorealistične slike, ponekad uz dovoljnu vjernost da ih ljudi ne mogu razlikovati od pravih slika [17]. Modeli obično počinju sa sjemenkom (engl. *seed*) koja generira nasumični vektor šuma (engl. *noise vector*) [18].

Bezuvjetne suparničke mreže stvaraju slike iz nasumičnog šuma bez dodatnih uvjetnih ograničenja. Primjerice, DCGAN (duboka konvolucijska suparnička mreža) koristi duboku konvolucijsku strukturu za stvaranje slika. S^2-GAN stvara mrežu od dvije faze te dubinske mape za generiranje slika s realističnim površinskim detaljima i informacijama o dubini (RGBD slike). Wasserstein GAN (WGAN) smanjuje gubitke i poboljšava stabilnost treniranja u odnosu na prijašnje GAN-ove radi ostvarivanja boljih performansa. PGGAN (Progressive Growing GAN) mreža povećava dubinu konvolucijskih slojeva da bi proizvela prirodne slike visokih rezolucija [19].

Bezuvjetno generiranje slika ima pomalo limitiranu praktičnu svrhu. Ukoliko je potrebna neka nespecifična slika (ili druga vrsta datoteke), moguće je jednostavno nasumično preuzeti potrebnu datoteku sa jedne od brojnih medijskih baza podataka na internetu [20].

2.5. Klasom uvjetovano generiranje slika

Klasom uvjetovano generiranje slika zadatak je generiranja različitih slika koristeći informacije klasnih oznaka [21]. Kreiranje klasom uvjetovanog modela za generiranje slika zahtijeva mjerenje pripadnosti generiranih slika namijenjenoj klasi [22]. Rani pristupi ovom načinu izrade slika su uvjetni varijacijski autokoderi te uvjetne generativne suparničke mreže. Ti pristupi latentnom vektoru dodaju oznake u svrhu kontroliranja semantičkih karakteristika generirane slike. Mreže koje koriste klasne informacije pokazale su napredne performanse. Najčešći pristup kod uvjetovanih mreža je ubacivanje klasnih oznaka u generator i diskriminator [21]. Suparničke mreže postaju uvjetovane ukoliko su generator i diskriminator uvjetovani na dodatnim informacijama. Te informacije mogu biti bilo kakve pomoćne informacije, kao što su klasne oznake ili ostali oblici podataka [23].

Mreže za uvjetno generiranje slika koriste klasom uvjetovane statistike normalizacije (tretiranje podataka mijenja se zavisno o klasama), kao i diskriminatore dizajnirane da rade kao

klasifikatori. Klasne informacije su ključne za uspjeh takvih modela. Pri radu sa limitiranim brojem oznaka, pomaže stvaranje umjetnih oznaka [24].

2.6. Uvjetovanje bez klasifikatora

Uvjetovanje klasifikatorima je tehnika koja poboljšava kvalitete slika difuzijskog modela koristeći dodatni trenirani klasifikator. Uvjetovanje difuzijskih modela klasifikatorima komplicira treniranje difuzijskih modela jer zahtijeva treniranje dodatnog klasifikatora koji mora biti treniran na podacima sa šumom. Zbog toga obično nije moguća uporaba predtreniranih klasifikatora. Kod uvjetovanja bez klasifikatora, istovremeno se treniraju uvjetni i bezuvjetni difuzijski modeli, čije se rezultirajuće procjene kombiniraju te se dobiva kompromis između kvalitete slika i raznolikosti, koji je sličan kompromisu dobivenim uvjetovanjem sa klasifikatorima [25].

Uvjetovanje difuzijskih modela bez klasifikatora značajno poboljšava slike generirane uvjetnim difuzijskim modelima uz gotovo nikakav trošak. Ovaj način uvjetovanja esencijalna je komponenta DALL-E 2 i Imagen modela.

Trenira se model koji generira slike ovisno o zadanim uvjetima. Zadani uvjeti djelomični se postotak vremena (10% do 20%) uklanjaju, čime model istovremeno dobiva sposobnost generiranja slika sa prisutnošču uvjeta i generiranja slika bezuvjetno. Ovo modelu omogućuje prilagodbu generiranja slika u odnosu na prisutnost ili neprisutnost uvjeta. Model je postavljen na način da mijenja ponašanje ovisno o prisutnosti određenih uvjeta. Time se stvara mogućnost kontroliranja načina generiranja slika uzimajući u obzir različite faktore. Moguće je podešavanje koliko se model oslanja na zadane uvjete, gdje je suprotnost slobodno stvaranje slika [26]. Naziv vrijednosti koja podešava koliko model prati tekstualni opis je Classifier Free Guidance Scale ili CFG.

3. Modeli dubokog učenja

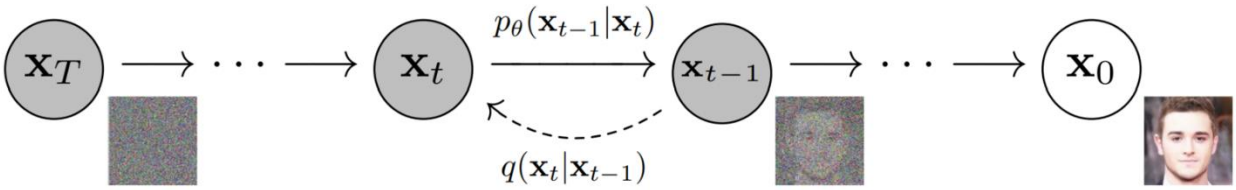
U ovom poglavlju opisat će se neki od modela dubokog učenja koji se koriste za generiranje slika iz tekstualnih opisa. Odabrani su neki od najpopularnijih modela. Razvijeni su koristeći različite arhitekture i tehnike.

3.1. Difuzijski modeli

Difuzijski modeli su vrsta generativnih modela koji pretvaraju Gaussov šum u uzorke iz naučene distribucije podataka kroz proces uklanjanja šuma. Modeli mogu biti uvjetni, primjerice prema oznakama klasa, tekstu ili slikama niske rezolucije [27].

Difuzija je proces pokretanja čestica iz mjesta visoke koncentracije prema mjestu niske koncentracije. Difuzija ujednačava koncentraciju čestica. Distribucija čestica promijenjena je u odnosu na stanje prije difuzije. U neravnotežnoj statističkoj fizici postoji ideja postepenog pretvaranja jedne distribucije u drugu. Tom idejom stvoreni su difuzijski modeli [28]. Difuzijski modeli sustavno i postepeno uništavaju strukturu unutar distribucije podataka kroz iterativni unaprijedni (engl. *forward*) difuzijski proces. Zatim se uči obrnuti (engl. *backward*) proces koji obnavlja strukturu u podacima, rezultirajući visoko fleksibilnim i prilagodljivim generativnim modelom podataka. Ovaj pristup omogućuje brzo učenje, uzorkovanje te procjenu vjerojatnosti modela dubokog učenja. Model se može opisati i kao generativan Markovljev lanac koji pretvara jednostavnu poznatu distribuciju (kao što je šum) u ciljanu (podatkovnu) distribuciju koristeći difuzijski proces [29]. Markovljev lanac znači da je stanje objekta tijekom bilo kojeg trenutka u lancu ovisno isključivo o prijašnjem stanju objekta.

Distribucija (struktura) originalne slike postepeno se (u koracima) uništava dodavajući šum. To je unaprijedni difuzijski proces na kraju kojeg ostaje izotropni Gauss, slika isključivo od šuma. Količina dodanog šuma pri svakom koraku kontrolira se pomoću planera odstupanja koji određuje koliko šuma je potrebno dodati kako bi na kraju prednjeg difuzijskog procesa ostao izotropni Gauss. Neuronska mreža koristi se za rekonstruiranje slike, tj. uklanjanje šuma iz svakog koraka. Zadaća neuronske mreže za svaki korak i sliku iz koraka predviđanje je količine dodanog šuma u sliku iz prijašnjeg koraka. Ukoliko se proces dovoljno puta ponovi sa kvalitetnim podacima, model s vremenom nauči procijeniti originalnu distribuciju slike [28]. Na slici 3.1. je prikazana ilustracija opisanog difuzijskog procesa.



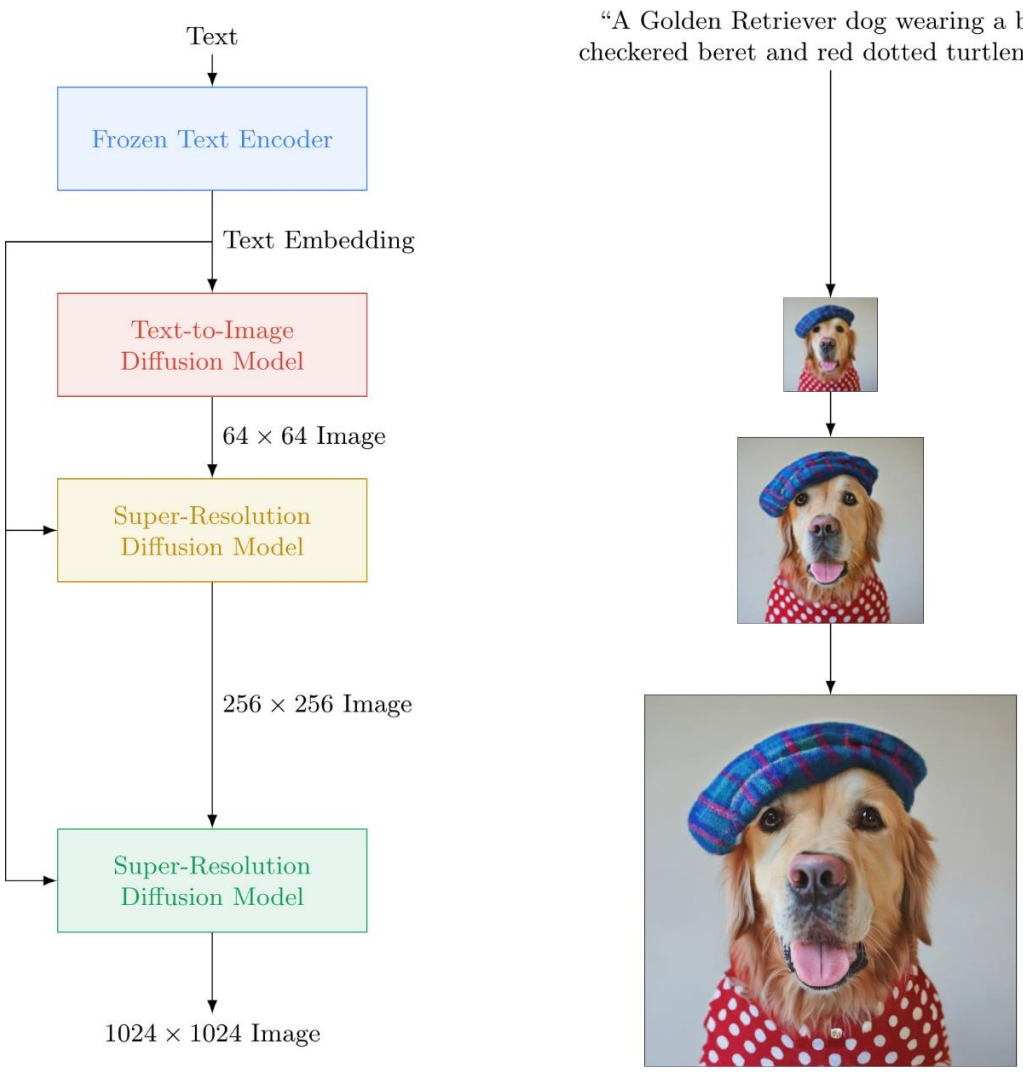
Slika 3.1. Ilustracija difuzijskih procesa [28]

3.1.1. Imagen

Google je predstavio Imagen 2022. godine. Imagen je difuzijski model za generiranje slika na temelju teksta. Model kombinira transformer sa difuzijskim modelima visoke vjernosti. Imagen uvodi novu tehniku difuzijskog uzorkovanja, nazvanu dinamični prag, koja pomaže stvaranju slika koje su fotorealističnije i detaljnije. Uvodi i novu tehniku difuzijske arhitekture, Efficient U-Net [27].

U-Net je vrsta neuronske mreže razvijena za zadatak segmentacije slika. Sastoji se od enkodera i dekodera koji izdvajaju generalne značajke slike te preskočne veze koja ponovno uvodi detaljnije značajke u dekoder [30].

Imagen se sastoji od tekstualnog kodera T5-XXL koji pretvara tekst u skup vektorskih oblika (engl. *embeddings*) te tri difuzijska modela. Prvi, bazni model stvara slike dimenzija 64x64 te koristi U-Net arhitekturu. Druga dva modela su super-rezolucijski difuzijski modeli, koji koriste modificiranu U-Net strukturu, nazvanu Efficient U-Net. Prvi super-rezolucijski model pretvara slike dimenzija 64x64 u dimenzije 256x256, dok drugi model pretvara iz slika dimenzija 256x256 u dimenzije 1024x1024. Efficient U-Net struktura pokazala se jednostavnijom, bržom i memorijski efikasnijom u odnosu na prijašnje U-Net implementacije. Promjene u strukturi su prebacivanje parametara sa blokova visoke rezolucije na blokove niske rezolucije, skaliranje preskočnih veza te preokretanje redoslijeda procesa smanjenja i povećanja razlučivosti [27]. Vizalni prikaz opisane arhitekture je na slici 3.2.



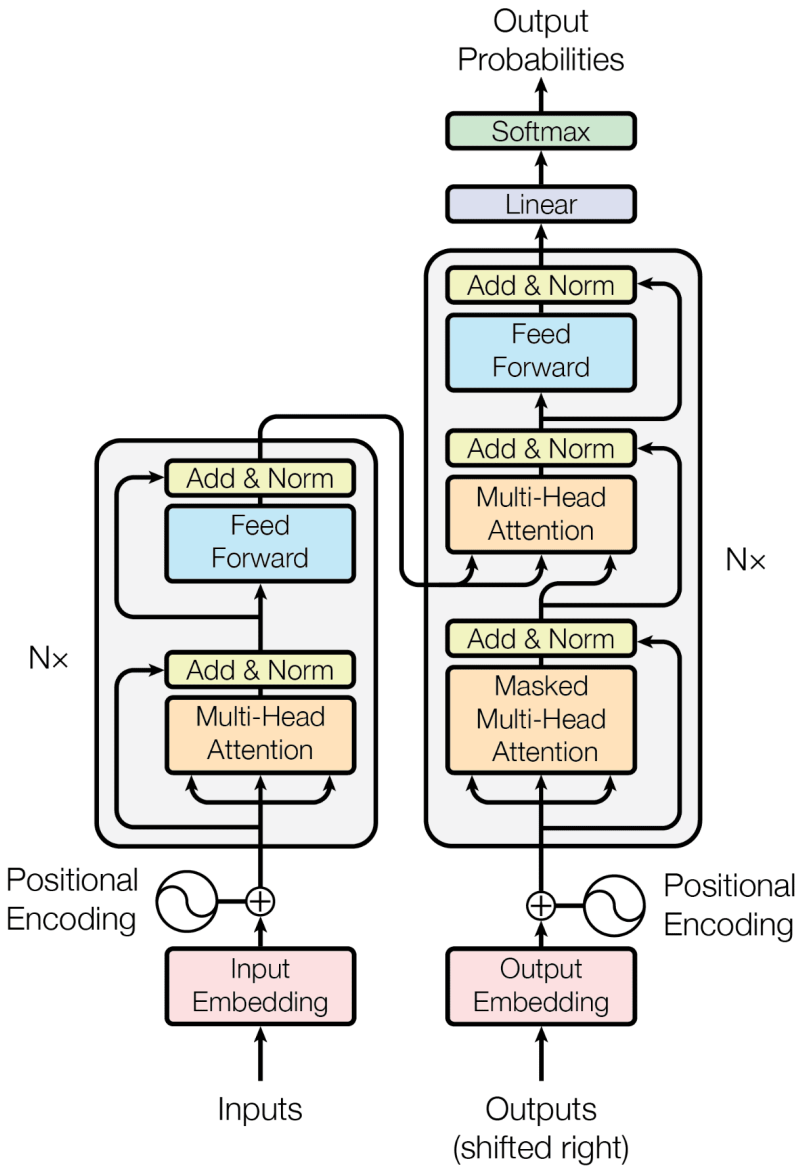
Slika 3.2. Vizualizacija arhitekture Imagen modela [27]

3.2. Transformer neuronske mreže

Transformer je neuronska mreža koja primjenjuje mehanizam pažnje. Google Brain je 2017. godine razvio transformer. Mehanizmi pozornosti modelima omogućuju fokusiranje na specifične dijelove ulaza ili izlaza, ovisno o relevantnosti svakog elementa. Omogućuju modelu da obraća pozornost na važne informacije u sekvenci, bez obzira na udaljenost između elemenata [31]. Samopažnja je mehanizam pozornosti koji uzima u obzir odnos između riječi unutar iste rečenice [32]. Arhitekture bazirane na samopažnji, posebice transformeri, postali su model izbora za obradu prirodnog jezika [33].

Model originalnog transformera je stog od 6 slojeva. Slojevi su koder stoga, umetanje ulaza, pozicijsko kodiranje, pozornost s više glava, mreža za prosljeđivanje i dekoder stog. Transformer koristi prethodne izlazne sekvence kao dodatni ulaz. Buduće riječi su skrivene od transformera,

što ga prisiljava da nauči kako predviđati [34]. Slika 3.3. prikazuje arhitekturu transformera neuronske mreže.



Slika 3.3. Arhitektura transformera neuronske mreže [31]

Direktna primjena samopažnje na slike zahtijevala bi da svaki piksel uzima u obzir sve ostale piksele. Zbog velikog broja piksela u slikama, to nebi bilo praktično za slike realističnih dimenzija. Isprobani su različiti načini primjena transformera za slike, poput primjene samopažnje samo unutar specifičnih sekcija umjesto cijele slike, ili uz korištenje manje računalno intenzivnih tehnika. Neke metode selektivno primjenjuju pažnju na blokove različitih dimenzija ili samo na određenu os. Vision Transformer model primjenjuje transformer na slike uz što je manje modifikacija originalnog modela. Slike se dijele na manje sekcije koje se u brojčanom obliku šalju

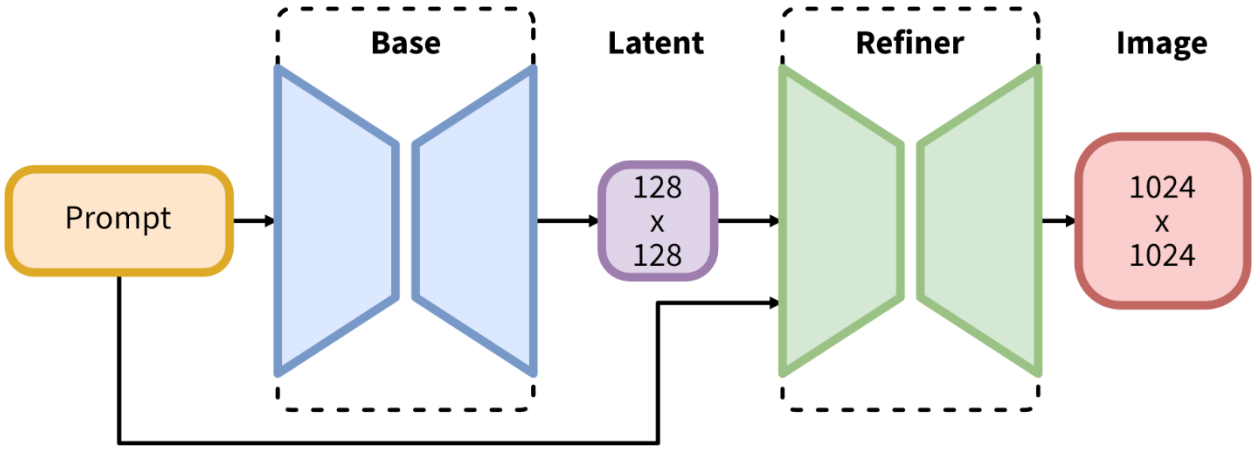
kao ulaz u transformer. Sekcije slike tretiraju se na isti način kao i riječi (tokeni) pri obradi prirodnog jezika [33].

3.3. StabilityAI

Stability AI i Runway razvili su stabilnu difuziju, latentni difuzijski model iz teksta u sliku [35]. Latentni prostor označava prikaz komprimiranih podataka na način da su slični podaci međusobno blizu [36]. Latentni prostor obuhvaća bitne informacije istovremeno smanjujući veličinu podataka. Smanjenje računalne intenzivnosti difuzijskih modela bez oštećenja svojstava ključno je za poboljšanje pristupačnosti difuzijskih modela. Stable Diffusion bazira se na latentnim difuzijskim modelima. Pristup uključuje treniranje autokodera koji uči iz pojednostavljenih podataka koji su ekvivalentni originalnim podacima, ali efikasniji. Difuzijski modeli treniraju se u naučenom latentnom prostoru, što je bolje za skaliranje. Jednostavnija kompleksnost omogućuje efikasno generiranje slika iz latentnog prostora uz jedan prolaz kroz mrežu. Treniranje autokodera moguće je odraditi samo jednom te ponovo iskoristiti za treniranje različitih difuzijskih modela ili ostalih zadataka.

Za razliku od ostalih pristupa, latentni difuzijski modeli bolji su u skaliranju na visoko-dimenzijskim podacima te ostvaruju impresivne rezultate uz bitno smanjenje troškova procesiranja. Modeli koriste mehanizam unakrsne pažnje, koji omogućuje treniranje sa više oblika podataka (engl. *multi-modal training*) [37]. Mehanizam unakrsne pažnje mehanizam je arhitekture transformera koji kombinira dvije različite sekvence vektorskih oblika. Sekvence moraju biti istih dimenzija, te mogu biti različitih modaliteta (tekst, slika, zvuk itd.) [38].

U vrijeme izrade rada, najnovija verzija Stable Diffusion modela je SDXL. Studije pokazuju da SDXL konzistetntno nadmašuje prijašnje verzije Stable Diffusion modela. Glavni razlozi poboljšanja performansi su tri puta veća U-Net mreža u odnosu na prijašnje modele, dvije jednostavne i efektivne tehnike uvjetovanja te poseban difuzijski model za usavršavanje, koji odrađuje difuzijski proces na baznim SDXL slikama radi poboljšanja vizualne kvalitete slika [39]. Bazni SDXL model stvara pojednostavljene slike dimenzija 128x128 ispunjene šumom, koje se zatim dalje procesuiraju modelom za usavršavanje specijaliziranim za uklanjanje šuma, te se iz njih stvaraju slike dimenzija 1024x1024 [40]. Slika 3.4. prikazuje SDXL arhitekturu s modelom za usavršavanje.



Slika 3.4. SDXL arhitektura [40]

Alternativni pristup SDXL arhitekturi je korištenje baznog modela za stvaranje pojednostavljenih slika željenih dimenzija, te korištenje tehnike SDEdit za generiranje i uređivanje slika. Tehnika SDEdit koristi se u Stable Diffusion modelima pri funkciji „img2img“ [41], koja omogućuje stvaranje slika iz već postojećih slika.

3.4. DALL-E

DALL-E je model dubokog učenja kojeg je razvio OpenAI. Prvi puta je predstavljen u siječnju 2021. godine [42]. DALL-E je autoregresivni transformer sa 12 bilijuna parametara, treniran na 250 milijuna parova tekst-slika prikupljenih s interneta. Autoregresivno označava arhitekturu koja iz prijašnjih vrijednosti predviđa buduće vrijednosti [43]. Rezultat treniranja je fleksibilan generativni model slika visoke vjernosti, koji se može kontrolirati kroz prirodan jezik [3].

DALL-E kombinira dva modela dubokog učenja. Prvi je transformer (samo dekoder dio [3]) koji pretvara tekst u latentni slikovni prostor, te varijabilni enkoder/dekoder model koji pretvara latentni slikovni prostor u izlazne slike. Varijabilni enkoder/dekoder model inicijalno je treniran na slikama iz skupa podataka. Takav predtrenirani model zatim je korišten kao dio transformera kako bi transformer naučio parametre za pretvorbu teksta u slike [44].

Model stvara visoko kvalitetne slike bez korištenja ikakvih oznaka treniranja, tj. „zero-shot“ [3]. „Zero-shot“ označava koncept generiranja slike iz tekstualnog opisa na način koji sliku čini konzistentnom sa tekstom. Naziv „zero-shot“ dolazi iz činjenice da model nije specifično treniran sa fiskiranim setom tekstualnih ulaza, što znači da je u principu sposoban generiranja iz bilo kojeg tekstualnog unosa (uz različite razine uspješnosti) [42].

Ovim modelom pokazano je da skaliranje skupa podataka može dovesti do poboljšane sposobnosti „zero-shot“ generiranja slika, kao i sposobnosti modela da do određene razine

odrađuje zadatke za koje nije specifično kreiran, poput pretvorbe iz slike u sliku kontrolirane prirodnim jezikom ili transformacije slike [3].

3.5. DALL-E 2

Godinu dana nakon predstavljanja DALL-E modela, OpenAI predstavlja DALL-E 2, model koji generira realističnije i preciznije slike sa četiri puta većom rezolucijom u odnosu na DALL-E [45].

CLIP model pokazao se uspješnim u stvaranju opisa za slike koji obuhvaćaju semantiku i stil. CLIP model sastoji se od dva enkodera koji pretvaraju tekst i slike u vektorske oblike [43]. Model je multimodalan, što znači da predstavlja više oblika informacija, koji su u slučaju CLIP modela tekst i slike. DALL-E 2 kombinacija je CLIP modela i difuzijskih modela [46]. Drugi naziv za DALL-E 2 je unCLIP.

Model radi na način da CLIP tekstualni enkoder uzima tekstualni opis te pretvara tekst u CLIP vektorski oblik. Tekst u CLIP vektorskom obliku služi kao ulaz za model koji stvara CLIP vektorski oblik slika. Taj model naziva se „prior“. Na kraju iz stvorenih CLIP vektorskih oblika slikovni dekoder generira završnu sliku [47].

Isprobana su dva oblika „prior“ modela, autoregresivni i difuzijski. Odabran je difuzijski model koji je procesivno efikasniji te stvara kvalitetnije rezultate [46]. Slikovni dekoder koji generira završnu sliku je modificirani difuzijski model GLIDE. GLIDE model u proces treniranja difuzijskog modela dodaje tekstualne informacije, te u verziji korištenoj u DALL-E 2 modelu, CLIP vektorske oblike [47].

4. Modeli generiranja videozapisa iz teksta

Kreiranje jasnih i realističnih videozapisa važna je prekretnica u istraživanju generativnih modela. Difuzijski modeli su počeli stvarati visoko kvalitetne fotografije i audio zapise, te postoji velik interes za korištenje difuzijskih modela kod novih vrsta podataka. Videozapisi visoke kvalitete mogu se generirati koristeći standardnu formulaciju Gaussovog difuzijskog modela, uz male izmjene strukture modela za rukovanje videopodacima unutar memorijskih ograničenja akceleratora u dubokom učenju [48].

Teškoće u stvaranju modela za generiranje videozapisa prouzrokovane su kompleksnom strukturom podataka kod videozapisa, kao i računalno intenzivnim video reprezentacijama od strane modernih generatora. Generatori tretiraju videozapise kao odvojene skupove slika, što je jako intenzivno za predstavljanje dugih videozapisa visoke rezolucije [49].

Napredak u generiranju videozapisa iz teksta zaostaje u odnosu na generiranje slika iz teksta, najviše zbog nedostatka velikih skupova podataka sa visokokvalitetnim parovima teksta i videozapisa, te zbog kompleksnosti modeliranja videopodataka visokih dimenzija [50].

Meta je 2022. godine predstavila novi sustav Make-A-Video. To je model koji korisnicima omogućuje upis opisa scene, iz koje će se generirati kratak video koji odgovara opisu teksta. Videozapisi koje model kreira su vidljivo umjetni, uz zamagljene subjekte te iskrivljenu animaciju, no model i dalje predstavlja važan korak u razvoju sustava za generiranje sadržaja umjetnom inteligencijom [51]. 2022. godine Google je predstavio Google Imagen Video – svoju verziju modela za generiranje videozapisa iz teksta [52].

4.1. Difuzijski videomodeli

Za generiranje videozapisa koristeći difuzijske modele moguće je korištenje standardnih difuzijskih modela uz strukturu neuronske mreže prikladnu za videopodatke. U-Net neuronsku mrežu može se proširiti za videopodatke, koristeći posebnu vrstu 3D U-Net mreže koja je faktorizirana kroz prostor i vrijeme. Faktorizirana metoda prostora i vremena dobra je opcija za videotransformere zbog efikasnosti računalnih resursa [48].

U-Net mreža može se koristiti na sekvencama različitih duljina te se može zajednički trenirati na slikovnim i videozadacima, što je važno za kvalitetne rezultate. Pri takvom treniranju pojavljuje se potreba za balansiranjem između pristranosti i varijabilnosti (engl. *Bias-Variance tradeoff*) [53].

4.2. Make-A-Video

Make-A-Video koristi modele za generiranje slike iz teksta za učenje povezanosti teksta sa vizualima, te koristi učenje bez nadzora na neoznačenim videozapisima kako bi naučio realistično kretanje. Tim pristupom Make-A-Video stvara videozapise iz teksta bez potrebe za pariranim podacima teksta i videozapisa [50].

Kao model za generiranje slike iz teksta korišten je Make-A-Scene, model tvrtke Meta. Da bi model počeo uzimati vrijeme u obzir, potrebno je dodati prostorno vremenski tok za procesiranje videozapisa. Model već treniran na slikovnim podacima adaptira se na informacije relevantne za videozapise. Za stvaranje videozapisa visoke kvalitete, koristi se mreža za interpolaciju sličica koja povećava sličice te nadopunjuje praznine, što rezultira većom kvalitetom [54].

Make-A-Video stvara videozapise visoke kvalitete tako da dodaje privremene informacije u predtrenirane modele za generiranje slike iz teksta. Za poboljšanje vizualne kvalitete treniraju se prostorni super-rezolucijski modeli te mreže za interpolaciju sličica [1].

5. Evaluacija modela

Kvaliteta slike te povezanost slike s tekstom dva su glavna kriterija za procjenu modela za generiranje slika iz teksta. Kvaliteta slike ukazuje na fotorealizam ili vjernost generiranih slika [1], te sličnost generiranih slika sa slikama iz podatkovnog seta za treniranje [55]. Povezanost slike s tekstom pokazuje odgovara li sadržaj generirane slike značenju tekstualnog opisa. Za ocjenu kvalitete percepcije slike, općenito trebamo koristiti modele bez referentne slike za umjetno generirane slike. Modeli se mogu podijeliti u sljedeće skupine, od kojih su prve 3 skupine za općenitu percepcijsku kvalitetu slike:

- Modeli temeljeni na unaprijed definiranim značajkama slika: Ova skupina uključuje mjere percepcijske kvalitete slike poput CEIQ (Contrast Enhancement Image Quality), CPBD (Cumulative Probability of Blur Detection), NIQE (Natural Image Quality Evaluator), itd... Ovi modeli koriste unaprijed definirane značajke na temelju prethodnog znanja o kvaliteti slike.
- Modeli temeljeni na regresiji potpornih vektora (SVR. Support Vector Regression): Ova skupina uključuje BMPRI (Blind Image Quality Estimation via Distortion Aggravation), GMLF (Gradient Magnitude and Laplacian Features), HIGRADE (HDR Image GRADient based Evaluator) itd… Ovi modeli kombiniraju unaprijed definirane značajke pomoću regresije SVR-a za predstavljanje perceptivne kvalitete.
- Modeli temeljeni na dubokom učenju: Ova skupina uključuje najnovije mjere bazirane na dubokom učenju, poput DBCNN (Deep Bilinear Convolutional Neural Network), CLIPIQA (CLIP Image Quality Assessment), CNNIQA (Convolutional Neural Networks for no-reference Image Quality Assessment), HyperNet itd... Ovi modeli karakteriziraju informacije o kvaliteti treningom neuronskih mreža iz označenih podataka o subjektivnoj kvaliteti.
- Drugi modeli za specifične namjene kod umjetno generiranih slika - modeli bazirani na funkciji gubitka (engl. loss function): ova skupina uključuje modele koji se obično koriste u umjetno generiranim slikama, naime FID (Frechet Inception Distance), ICS (InCeption Score) i KID (Kernel Inception Distance) mjere. FID i KID računaju udaljenost između umjetno generiranih slika i MS-COCO baze slika, dok ICS mjeri raznolikost generiranih slika [56].

Za mjerenje povezanosti slike s tekstom, mogu se koristiti najpopularniji model CLIP mjera (Contrastive Language-Image Pretraining score) [1], ImageReward, HPS (Human Preference Score), PickScore i drugi.

U ovom radu za mjerenje povezanosti slike s tekstom, koristit će se CLIP mjera [1]. Za mjerenje raznolikosti i sličnosti generiranih slika s drugom grupom slika, koristit će se ICS i FID mjere [55].

Spomenute mjere rezultat računaju automatski, uz pokretanje svojevrsnog modela. Postoje i mjere koje uključuju ljudsku procjenu. Primjerice, za bolju procjenu vjernosti generiranih slika te povezanosti sa tekstom, DrawBench, PartiPropts i UniBench uključuju ljudske ispitivače koji uspoređuju generirane slike različitih modela. Mjera PaintSKills uz kvalitetu slike i povezanost slike s tekstom procjenjuje vještine vizualnog zaključivanja te društvene pristranosti [1].

5.1. Objektivne metode evaluacije

Cilj generativnog učenja je da model stvara podatke koji odgovaraju promatranim podacima. Samim time, razlika između vjerojatnosti za promatrane podatke iz stvarnosti te vjerojatnosti za generativni model može služiti kao način procjene performansa generativnih modela. Međutim, definiranje prikladnih načina procjena za generativne modele je komplicirano [57]. Postojeće automatske mjere za evaluaciju imaju svoje limitacije, primjerice FID nije uvijek ujednačen s perceptualnom kvalitetom [1].

5.1.1. ICS mjera (*Inception Score*)

ICS mjera automatska je metoda za evaluaciju uzoraka [58]. Mjera je pokazala dobru korelaciju s ljudskim ocjenjivanjem generiranih slika. Metoda koristi Inception V3 mrežu treniranu na ImageNet podatkovnom setu. Inception V3 mreža dizajnirana je za zadatatke klasifikacije. Izlaz mreže je vektor vjerojatnosti koji pokazuje vjerojatnost da slika pripada određenim klasnim oznakama. Pri računanju ICS mjere, računa se statistika izlaza mreže primjenjene na generiranim slikama [59]. Inception score procjenjuje kvalitetu generiranih slika bazirano na uspješnosti Inception V3 modela u klasifikaciji slika na jedan od 1000 poznatih objekata [60]. Viši Inception score označava da je model sposoban stvarati mnogo raznolikih slika. Kao uvjeti za ostvarenje visokog Inception scorea, slike moraju jasno nalikovati nekom objektu, te moraju biti raznolike [61]. Generirane slike trebaju sadržavati jasne objekte. Mreža mora biti uvjerena da se na slici nalazi jedan objekt. Model za generiranje slika treba stvoriti raznolike slike iz različitih klasa koje se nalaze u ImageNet podatkovnom setu [59].

ICS mjera je limitirana na ono što Inception mreža može detektirati, što je povezano sa podatkovnim setom na kojem je mreža trenirana. To znači da se za klase koje nisu prisutne u podatkovnom setu uvijek ostvaruje niži IS unatoč kvaliteti slika, jer slika nije pravilno

klasificirana. Ako Inception mreža ne može detektirati značajke slike bitne za kvalitetu generirane slike, tada slike lošije kvalitete mogu ostvariti visoke rezultate. ICS mjera loša je za procjenu različitosti unutar istih klasa, mogući su visoki rezultati ako su generirane iste slike istih klasnih oznaka više puta [61].

5.1.2. FID mjera (*Frechet Inception Distance*)

Frechet Inception Distance (ili skraćeno FID), mjera je kojom se računa razlika između vektora značajki izračunatih za stvarne slike i generirane slike, koristeći Inception V3 mrežu treniranu na ImageNet podatkovnom setu [55]. FID rezultat prikazuje koliko su te dvije grupe statistički slične. Niži rezultati ukazuju da su dvije grupe slika međusobno sličnije, sa više sličnih statistika, dok savršen 0.0 rezultat ukazuje da su dvije grupe slika jednake. Niži rezultati pokazali su korelaciju sa slikama više kvalitete.

Cilj razvoja FID mjere je evaluacija generiranih slika bazirano na statistikama kolekcije generiranih slika, usporedno sa statistikama kolekcije stvarnih slika iz ciljane domene [60]. Glavna razlika između FID i IS jedinica je usporedna uporaba sa stvarnim slikama, što za FID omogućuje analiziranje stvarnih slika sa generiranim slikama, čime se bolje simulira ljudska percepcija [62].

FID je definiran kao Frechet razlika distribucija Inception značajki između dvaju seta slika. Inception značajke posebne su vrste dubokih značajki, tj. aktivacija specifičnog međusloja trenirane neuronske mreže. Duboke značajke efektivne su za snažne semantičke opise [63].

Pri procesu predprocesiranja kod računanja FID mjere, radnje kao što su mijenjanje veličine slika i kompresiranje mogu dovesti do velikih varijacija te neočekivanih posljedica. Primjerice, za istu sliku, različite implementacije za procesiranje slika proizvode različite rezultate, koje uzrokuju značajne varijacije pri korištenju evaluacijskih protokola [64].

5.1.3. CLIP mjera

CLIP je multimodalni model treniran na 400 milijuna parova slika i opisa skupljenih sa interneta [43]. 500 tisuća upita za pretraživanje poslana su kroz pretraživač. Za svaki upit prikupljeno je do 20 tisuća parova slika i opisa. CLIP model moguće je koristiti za robusnu automatsku evaluaciju titlovanja slika. Specifični CLIP model korišten za računanje CLIP mjere je ViT-B/32 verzija, koja predstavlja slike kroz Vizualni Transformer.

Za procjenu kvalitete opisa slika, opis i slika prolaze kroz transformere koji služe kao ekstraktori. Nakon toga računa se sličnost kosinusa rezultirajućih vektorskih oblika. Sličnost kosinusa je mjera koja određuje kosinus kuta između dva vektora u višedimenzijskom prostoru.

Skala sličnosti kosinusa je od -1 do 1, gdje +1 pokazuje da su vektori jednaki, 0 označava da su vektori ortogonalni, dok -1 označava da su vektori suprotni [65]. Zatim se vrši operacija ponovnog skaliranja te se izračunava CLIP mjera prema formuli [66]. CLIP rezultat prikazuje sličnost dane slike sa zadanim opisom [67]. Viši rezultat ukazuje da su slika i tekst više semantički povezani, dok niži rezultat ukazuje da nisu [65].

Kada su male razlike između točnih testnih opisa i netočnih testnih opisa, CLIP ne uspijeva odrediti točan opis. CLIP pokazuje neosjetljivost prema strukturi rečenica u opisu. Broj i veličina objekata relevantnih za sliku koji su spomenuti u opisu utječu na rezultat. Neuvjerljivi opisi slabo utječu na rezultate, dok nedostatak vizualne poveznice ima snažan utjecaj [67].

6. Praktični dio rada

Praktični dio rada objektivna je evaluacija modela za stvaranje slika. Modeli koji će biti evaluirani su Stable Diffusion i DALL-E 2. Specifično, evaluirat će se Stable Diffusion verzija 1.5 te SDXL. Objektivna evaluacija izvršit će se pomoću objektivnih mjera za kvalitetu i različitost slika, vjernost slika te povezanost slika i teksta. Specifično, mjere koje će se koristiti su ICS, FID te CLIP. Opisat će se proces pripreme modela Stable Diffusion i DALL-E 2 za generiranje slika te pokretanja istih. Bit će opisan proces pripreme modela za objektivnu evaluaciju, te njihovog pokretanja. Dobiveni rezultati će se analizirati te će se iz njih izvesti zaključci.

6.1. Priprema modela za stvaranje slika

Ovisno o tome jesu li korišteni modeli otvorenog ili zatvorenog koda, moguća je instalacija modela za stvaranje slika na uređaj te pokretanje modela koristeći hardverske sposobnosti računala na koji je model instaliran. Moguće je i korištenje modela preko web servisa koji omogućuju korištenje grafičkih kartica u oblaku, kao što su Google Collab i HuggingFace. Za potrebe ovog rada, Stable Diffusion model verzije 1.5 instaliran je lokalno na računalo. Stable Diffusion SDXL koristit će se preko web servisa Replicate. DALL-E 2 model je zatvorenog koda, što znači da model nije moguće instalirati lokalno na računalo. Način na koji će se koristiti DALL-E 2 model je preko web sustava tvrtke OpenAI.

6.1.1. Instalacija Stable Diffusion

Performanse modela instaliranog lokalno na računalo ovisit će o hardverskim specifikacijama računala. Autor Stable Diffusion predlaže korištenje NVIDIA grafičke kartice koja sadrži minimalno 6GB VRAM memorije. Predlaže se minimalno 8GB RAM memorije na računalu. Za instalaciju modela potrebno je minimalno 10GB prostora na disku [68].

Kao preduvjet za instalaciju Stable Diffusion, potrebno je instalirati Python programski jezik te sustav za upravljanje izvornim kodom git [69]. Python je pri instalaciji potrebno dodati u varijable okoline na računalu. Idući korak je instalacija grafičkog sučelja za Stable Diffusion. Neka od postojećih grafičkih sučelja su stable-diffusion-webui autora AUTOMATIC1111 [69], InvokeAI [70] te diffuzers autora abhishekkrthakur [71]. U ovom radu koristit će se sučelje stable-diffusion-webui autora AUTOMATIC1111. Za instalaciju sučelja potrebno je pokrenuti naredbu sa slike 6.1., unutar željene mape za Stable Diffusion. Naredba koristi prijašnje instaliran sustav za upravljanje izvornim kodom git za preuzimanje datoteka grafičkog sustava.

```
C:\Users\MK\sd>git clone https://github.com/AUTOMATIC1111/stable-diffusion-webui.git
Cloning into 'stable-diffusion-webui'...
remote: Enumerating objects: 26032, done.
remote: Counting objects: 100% (419/419), done.
remote: Compressing objects: 100% (187/187), done.
remote: Total 26032 (delta 271), reused 348 (delta 228), pack-reused 25613
Receiving objects: 100% (26032/26032), 31.83 MiB | 2.88 MiB/s, done.
Resolving deltas: 100% (18229/18229), done.
```

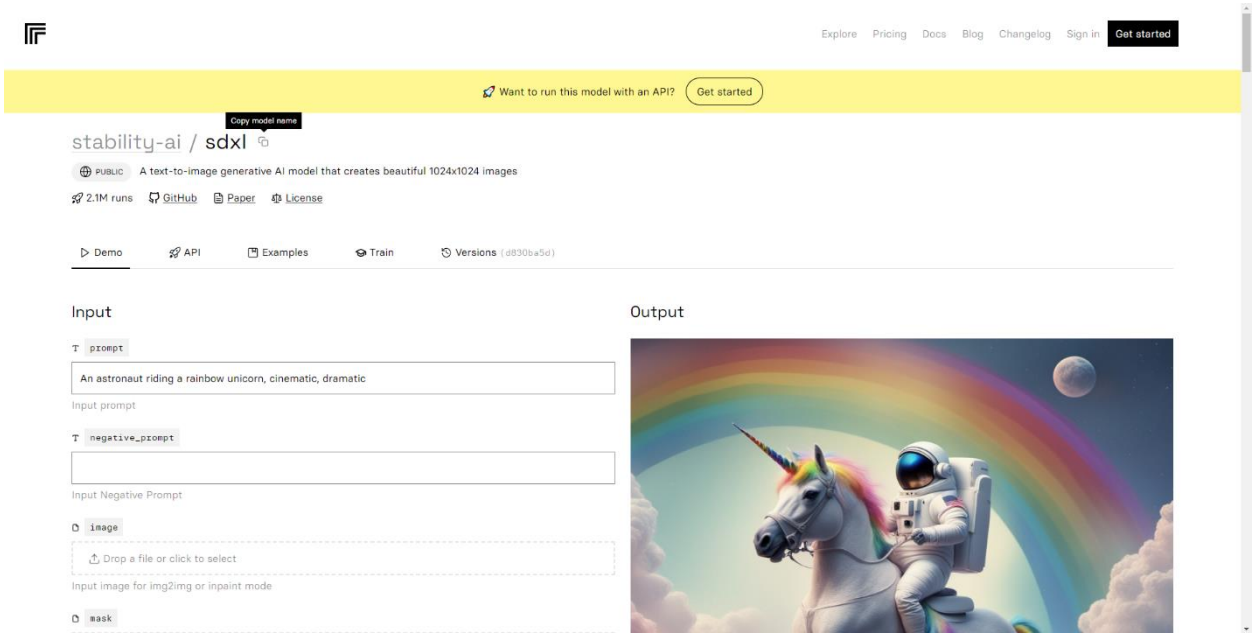
Slika 6.1. Naredba za instalaciju grafičkog sučelja za Stable Diffusion te izlaz nakon uspješnog pokretanja naredbe
Izvor: Autor

Nakon instalacije sučelja, potrebno je preuzeti predtrenirani model za Stable Diffusion. Postoje različite vrste modela koji su optimizirani za generiranje različitih vrsta sadržaja. Stable-diffusion-v1-5 predtrenirani je model izbačen od strane autora Stable Diffusion [72] koji će se koristiti za potrebe ovog rada. Potrebno se preseliti u mapu čija je putanja prikazana na slici 6.2. te u nju spremiti preuzeti predtrenirani model.

Slika 6.2. Mapa u koju je potrebno spremiti predtrenirane modele. U ovom slučaju, „sd“ je naziv željene mape za Stable Diffusion
Izvor: Autor

6.1.2. SDXL

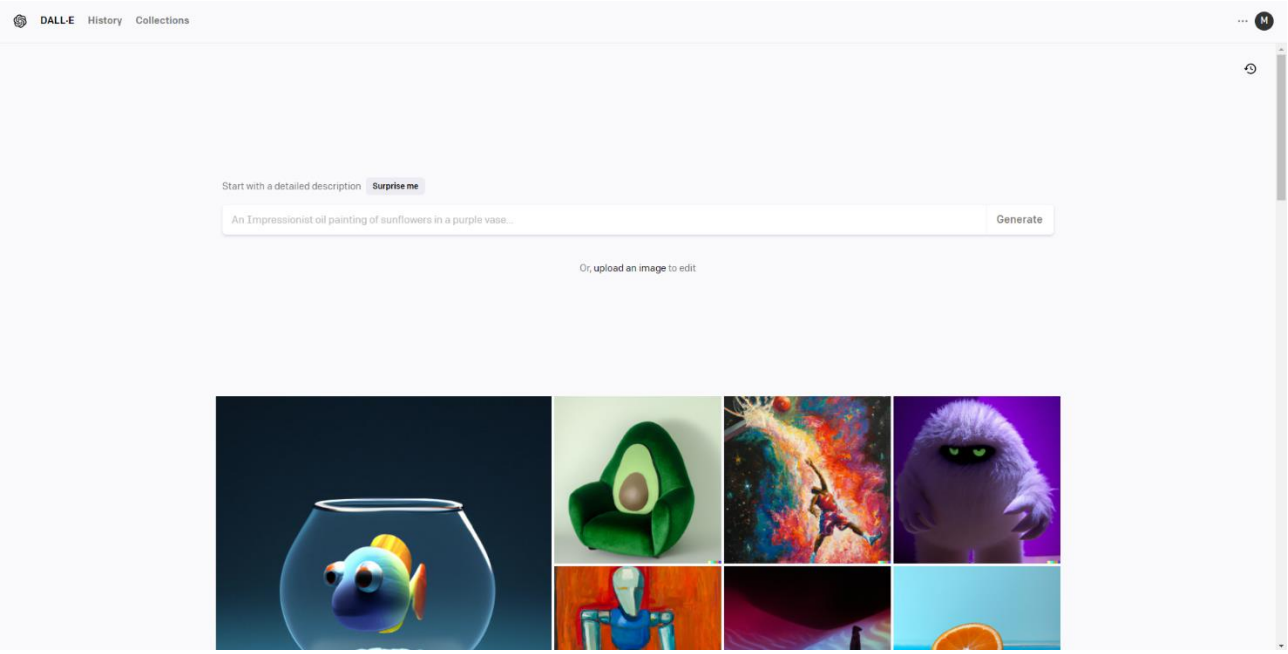
Na web sustavu Replicate moguće je koristiti Stable Diffusion SDXL model. Korištenje modela je besplatno, bez kreiranja korisničkog računa, do određene mjere. Trajanje besplatnog korištenja razlikuje se zavisno o korištenom modelu. Troškovi korištenja ovise o računalnom vremenu potrošenom na pokretanje vremena [73]. Slika 6.3. prikazuje web sustav Replicate, preko kojeg se može koristiti SDXL model.



Slika 6.3. Web sustav Replicate
Izvor: Autor

6.1.3. DALL-E 2

DALL-E 2 model moguće je koristiti kroz web sustav tvrtke OpenAI. Potrebno je izraditi korisnički račun za web sustav. Korištenje DALL-E 2 modela limitirano je kreditima (engl. *credits*). Korisnici koji su otvorili račun prije 6.4.2023 imaju određen broj besplatnih kredita koji se ponovo nadopunjuju svaki mjesec. Korisnici koji su otvorili račun nakon spomenutog datuma moraju kupiti kredite kako bi počeli koristiti DALL-E 2 model. Jedan kredit može se iskoristiti za generiranje 4 slika iz opisa, te za slanje zahtijeva za uređivanje slike ili varijaciju [74]. Na slici 6.4. je prikazan izgled web sustava OpenAI preko kojeg je moguće koristiti DALL-E 2 model.



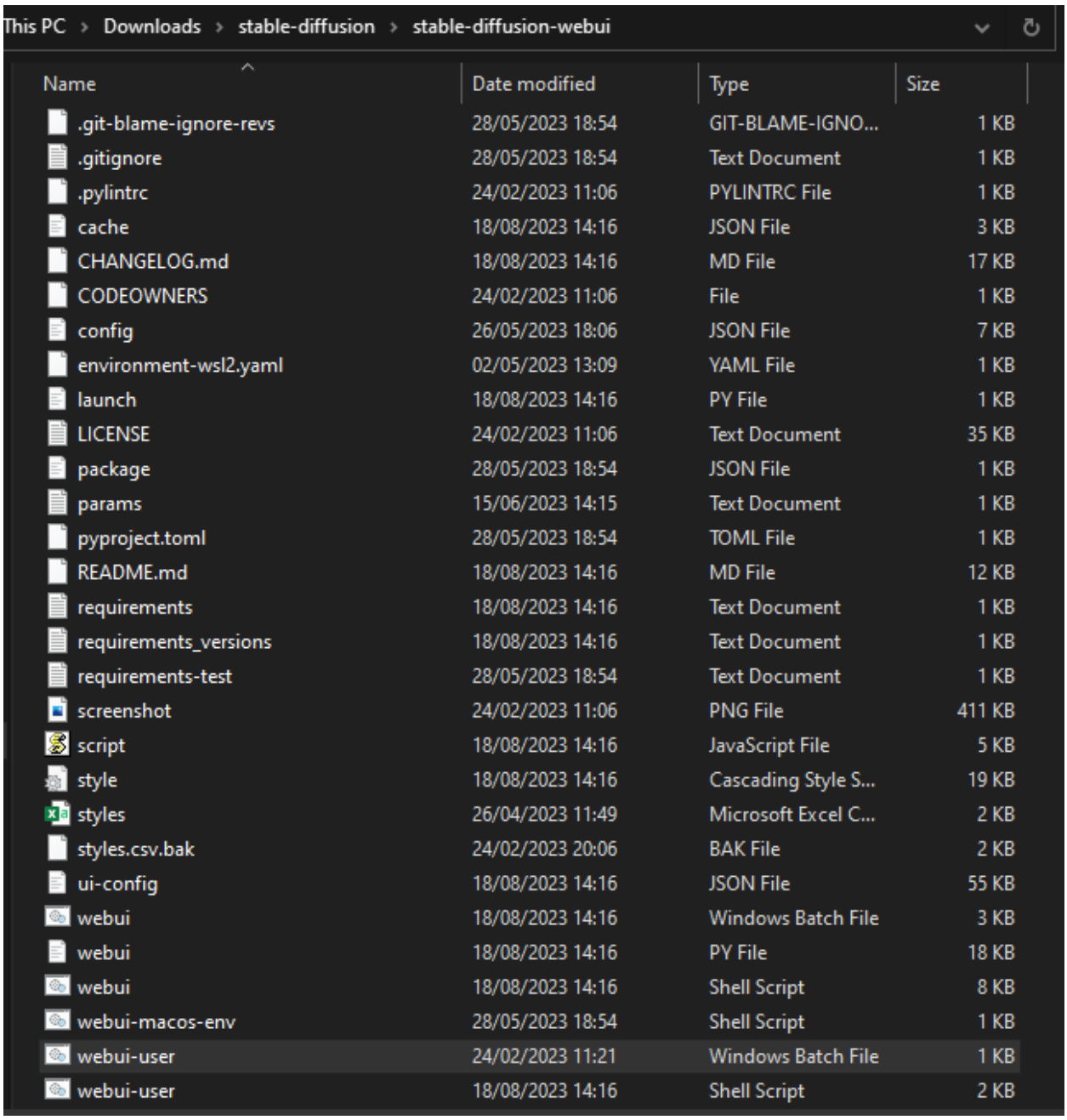
Slika 6.4. OpenAI web sustav za korištenje DALL-E 2 modela
Izvor: Autor

6.2. Pokretanje modela za stvaranje slika

Pri pokretanju modela instaliranog na računalo, vrijeme potrebno za generiranje slika, moguće rezolucije generiranja, broj slika koje se mogu generirati istovremeno, faktori su koji su uvjetovani hardverskim specifikacijama računala. Pri korištenju modela preko web sustava, spomenuti faktori ovise o pružatelju sustava. Lokalno instalirani modeli nude više mogućnosti prilagodbe modela [75], koje često nisu moguće preko web sustava. Primjerice, podešavanje modela sa svojim podacima, korištenje različitih vrsta modela ili spajanje više različitih modela.

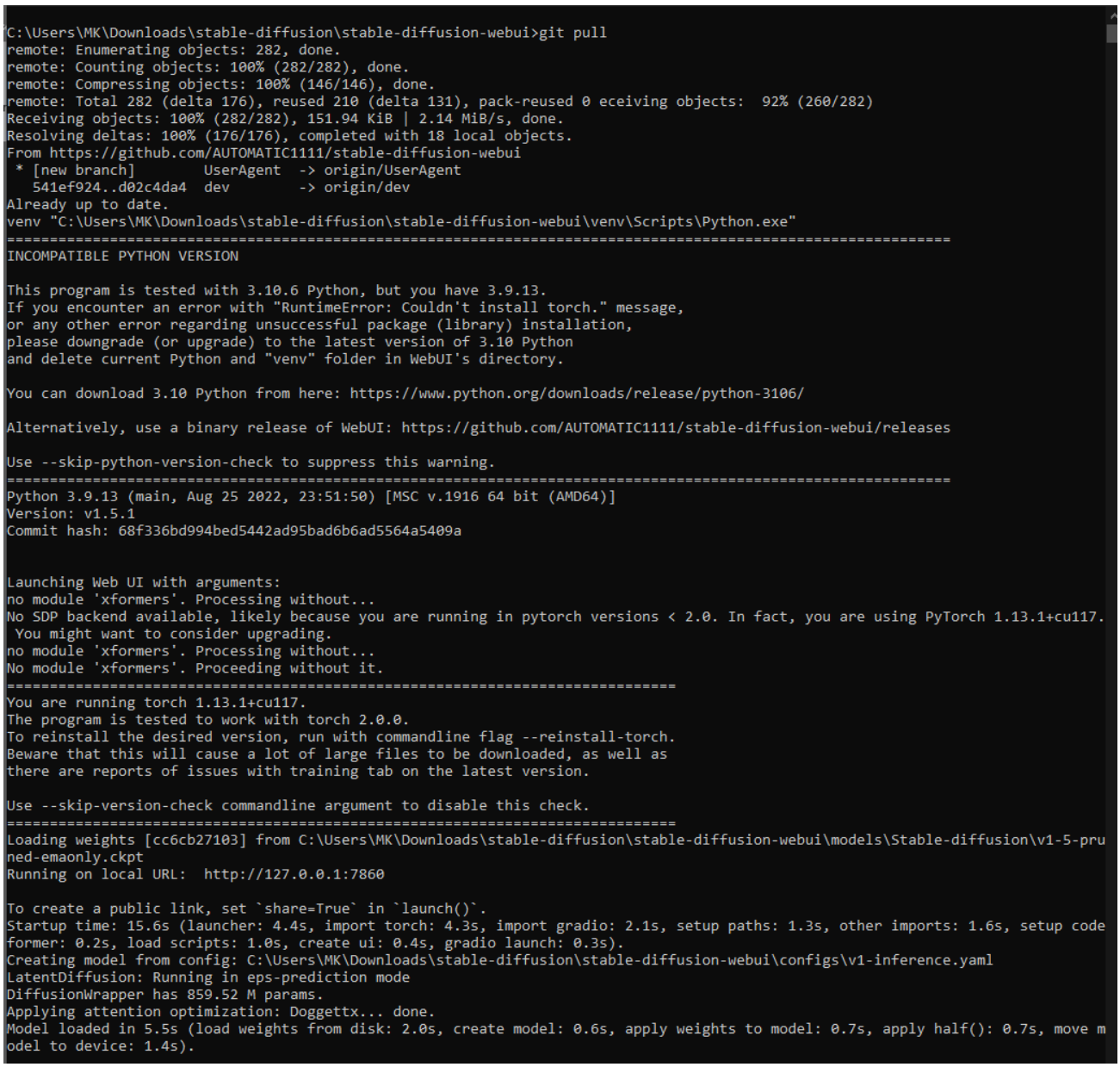
6.2.1. Pokretanje Stable Diffusion

Za pokretanje Stable Diffusion 1.5 modela koristeći AUTOMATIC1111 sučelje, unutar mape „stable-diffusion-webui“ u mapi u kojoj su instalirani model i sučelje, potrebno je pokrenuti datoteku naziva „webui-user.bat“. Slika prikazuje primjer putanje te sadržaj spomenute mape. U ovom slučaju, „stable-diffusion“ naziv je mape unutar koje su instalirani model te sučelje. Datoteka koju je potrebno pokrenuti na slici 6.5. je selektirana.



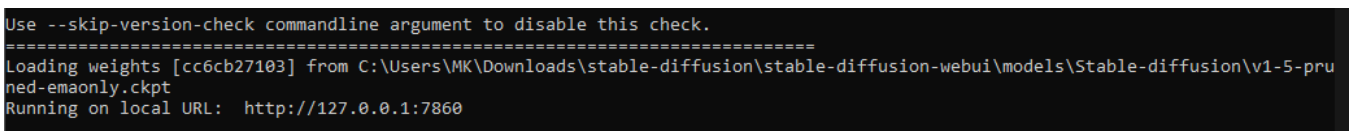
Slika 6.5. Prikaz mape u kojoj je instaliran Stable Diffusion preko sučelja AUTOMATIC1111
Izvor: Autor

Pri prvom pokretanju spomenute datoteke, instalirat će se potrebne ovisnosti za pokretanje modela. Za kasnija pokretanja trebat će manje vremena. Na slici 6.6 je prikazan izlaz u komandnoj liniji nakon što je pokrenuta datoteka „webui-user.bat“. Izlaz sa slike 6.6. primjer je pokretanja nakon što je procedura inicijalnog pokretanja već izvršena.



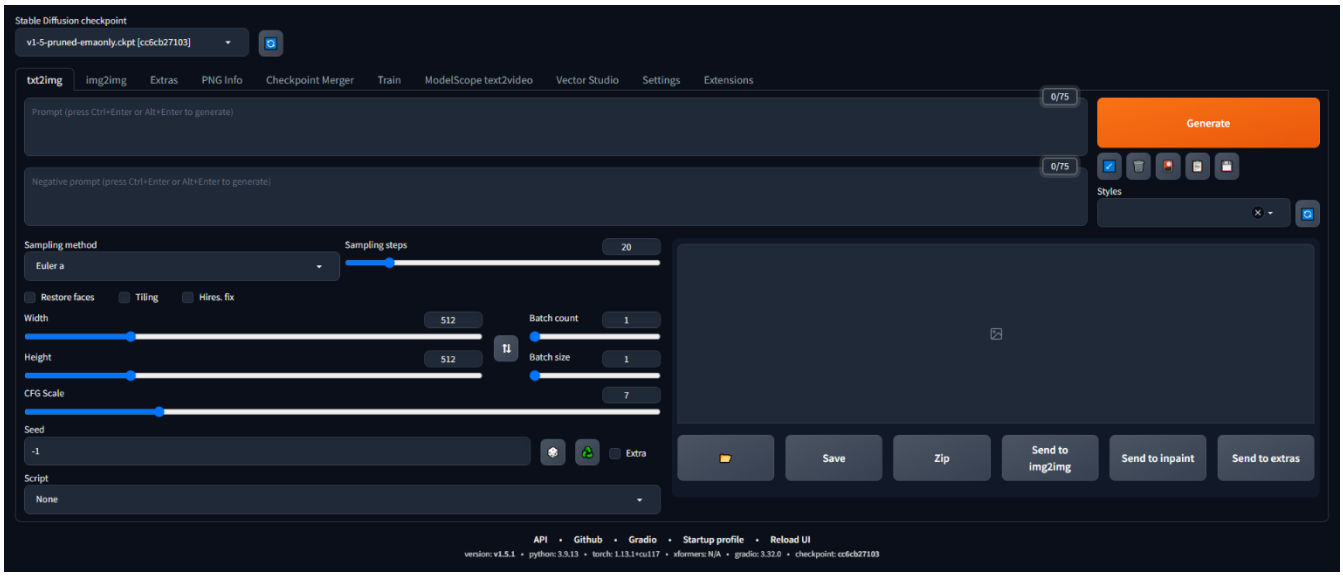
Slika 6.6. Prikaz izlaza komandne linije nakon pokretanja datoteke webui-user.bat
Izvor: Autor

Izlaz u komandnoj liniji sadrži lokalnu URL adresu na kojoj je pokrenuto AUTOMATIC1111 Stable Diffusion sučelje. Toj adresi je potrebno pristupiti u web pregledniku za otvaranje sučelja. Slika 6.7. prikazuje dio izlaza komandne linije sa lokalnom URL adresom na kojoj je pokrenuto sučelje.



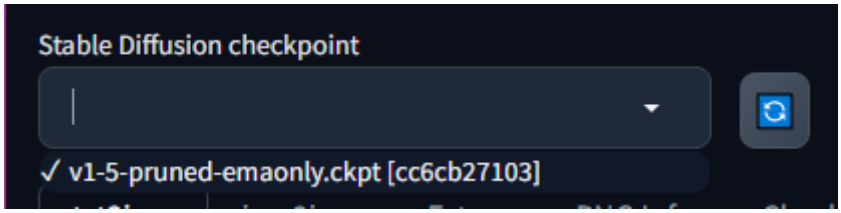
Slika 6.7. Prikaz dijela izlaza komandne linije sa URL adresom na kojoj je pokrenuto sučelje
Izvor: Autor

Sučelje se otvara pristupanjem lokalnoj URL adresi. Na slici 6.8. je prikazan izgled sučelja.



Slika 6.8. AUTOMATIC1111 sučelje
Izvor: Autor

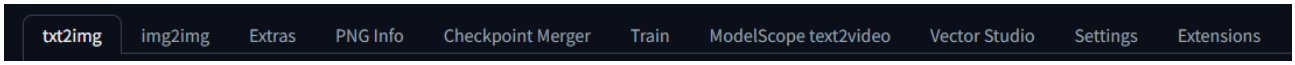
Na vrhu sučelja nalazi se meni za odabir predtreniranog modela koji će biti učitan. Klikom na tipku pored naziva modela osvježava se lista mogućih modela za odabir. Slika 6.9. prikazuje izbornik za odabir modela. U ovom slučaju moguće je odabrati samo jedan od modela sa liste, a to je Stable Diffusion 1.5.



Slika 6.9. Prikaz menija za odabir predtreniranog modela
Izvor: Autor

Ispod izbornika za odabir predtreniranog modela nalazi se traka za odabir načina rada. Pri otvaranju sučelja automatski se odabire način rada za generiranje slike iz teksta („txt2img“). Od ostalih načina rada moguće je odabrati „img2img“, za način rada iz slike u sliku ili nadopunjavanja nedostajećih dijelova slike te „Extras“ preko kojeg je moguće pristupiti načinu rada za super-rezoluciju. „PNG info“ iz generirane slike učitava informacije o parametrima koji su korišteni za generiranje slike [76]. „Checkpoint merger“ služi za spajanje do tri različitih predtreniranih modela u jedan. „Train“ način rada je za treniranje vlastitih modela. „Settings“ su postavke sučelja, dok je „Extensions“ za upravljanje dodacima za sučelje i Stable Diffusion. Na slici 6.10. je prikazana traka za odabir načina rada. Na prikazanoj traci se nalaze i opcije „ModelScope

text2video“ i „Vector Studio“. To su naknadno instalirani dodaci koji se ne pojavljuju na svježe instaliranim verzijama sučelja.



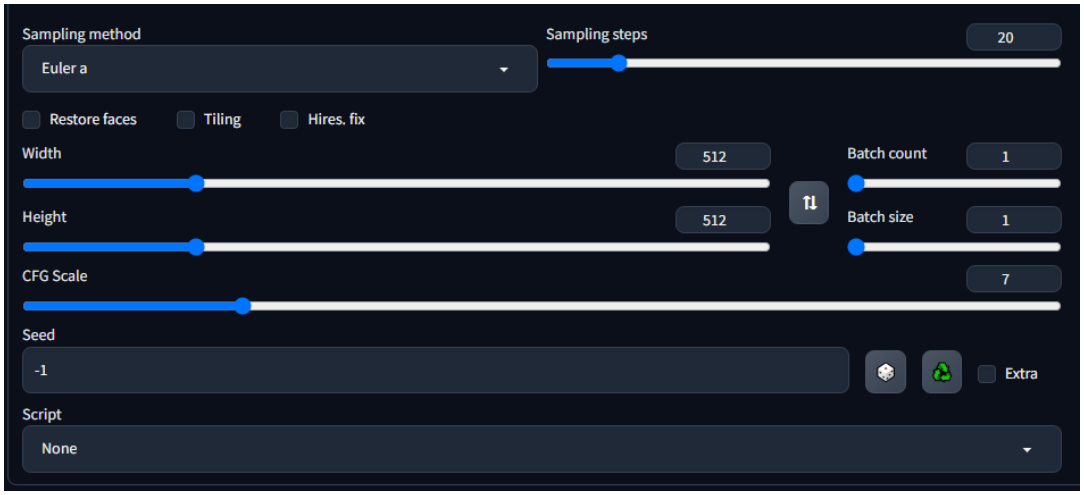
Slika 6.10. Prikaz trake za odabir načina rada
Izvor: Autor

Daljnji dijelovi sučelja ovise o odabranom načinu rada. U načinu rada za generiranje slike iz teksta dva su tekstualna polja. Prvo tekstualno polje služi za upisivanje opisa iz kojeg se želi generirati slika. Drugo tekstualno polje služi za negativne opise. Negativnim opisima specificiraju se stvari koje model treba izbjegavati pri generiranju slike [76]. Ispod tipke „Generate“ kojom se započinje proces generiranja slike, nalaze se dodatne opcije, te izbornik za stilove. Stilovi označavaju spremljene opise. Klikom na tipku sa sličicom diskete, koji se nalazi sa desne strane ispod gumba „Generate“, spremaju se trenutačno upisani opisi kao stil pod željenim nazivom. Spremljeni stil kasnije se može učitati klikom na izbornik sa lijeve strane od gumba sa sličicom diskete. Ispod spomenutih tipki nalazi se izbornik kojim se odabire stil koji se želi koristiti. Na slici 6.11. se nalazi opisani dio sučelja za opisivanje opisa za generiranje slike iz teksta.

Slika 6.11. Prikaz dijela sučelja za upisivanje opisa
Izvor: Autor

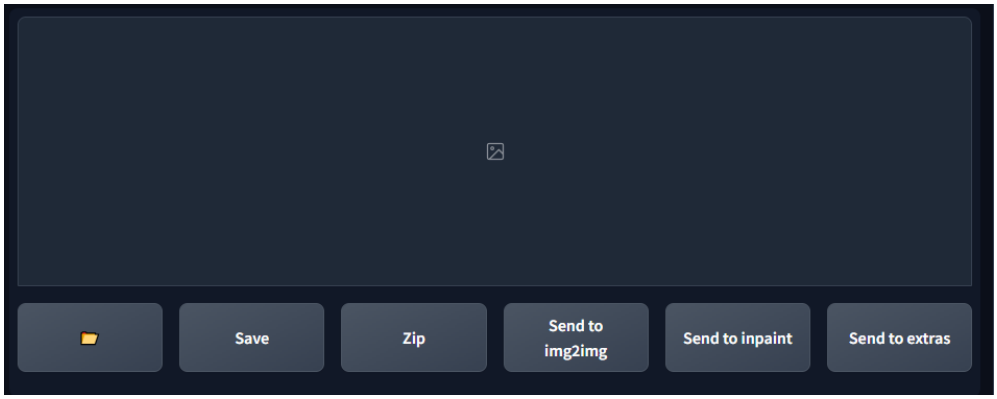
Ispod dijela sučelja za upisivanje opisa nalazi se dio sučelja za specificiranje parametara generirane slike. „Sampling method“ izbornikom bira se algoritam za stvaranje slike. „Sampling steps“ polje je za specificiranje broja koraka kojim će se generirana slika poboljšavati. Uz više vrijednosti će generiranje slike trajati duže, dok će niže vrijednosti stvarati rezultate lošije kvalitete. Polja „Width“ i „Height“ služe za postavljanje širine i visine slike. „Batch count“ vrijednost označava koliko serija slika će se generirati, dok je „Batch size“ broj slika u svakoj seriji. Slike unutar serije generiraju se pararelno. „CFG Scale“ specificira Classifier Free Guidance Scale vrijednost. Ta vrijednost kontrolira u kolikoj mjeri proces generiranja slike treba pratiti tekstualni opis. Uz veću vrijednost slika snažnije prati opis [77], ali niže vrijednosti stvaraju kreativnije rezultate. „Seed“ je broj iz kojeg Stable Diffusion model generira šum. Generirana slika može se reproducirati kroz nekoliko sesija ako se specificira isti „seed“ broj uz isti opis i parametre korištene za prvo stvaranje slike [78]. Ukoliko je „seed“ postavljen na -1, tada će se za svako

generiranje slike koristiti nasumičan „seed“ broj. U „Script“ izborniku moguće je odabrati dodatnu skriptu koja utječe na generiranje slike. Primjeri skripta su matrica opisa kojom se mogu generirati slike iz različitih kombinacija opisa, te X/Y/Z plot kojim je moguće generiranje slika iz različitih kombinacija parametara [76]. Slika 6.12. prikazuje dio sučelja za specificiranje parametara generirane slike.



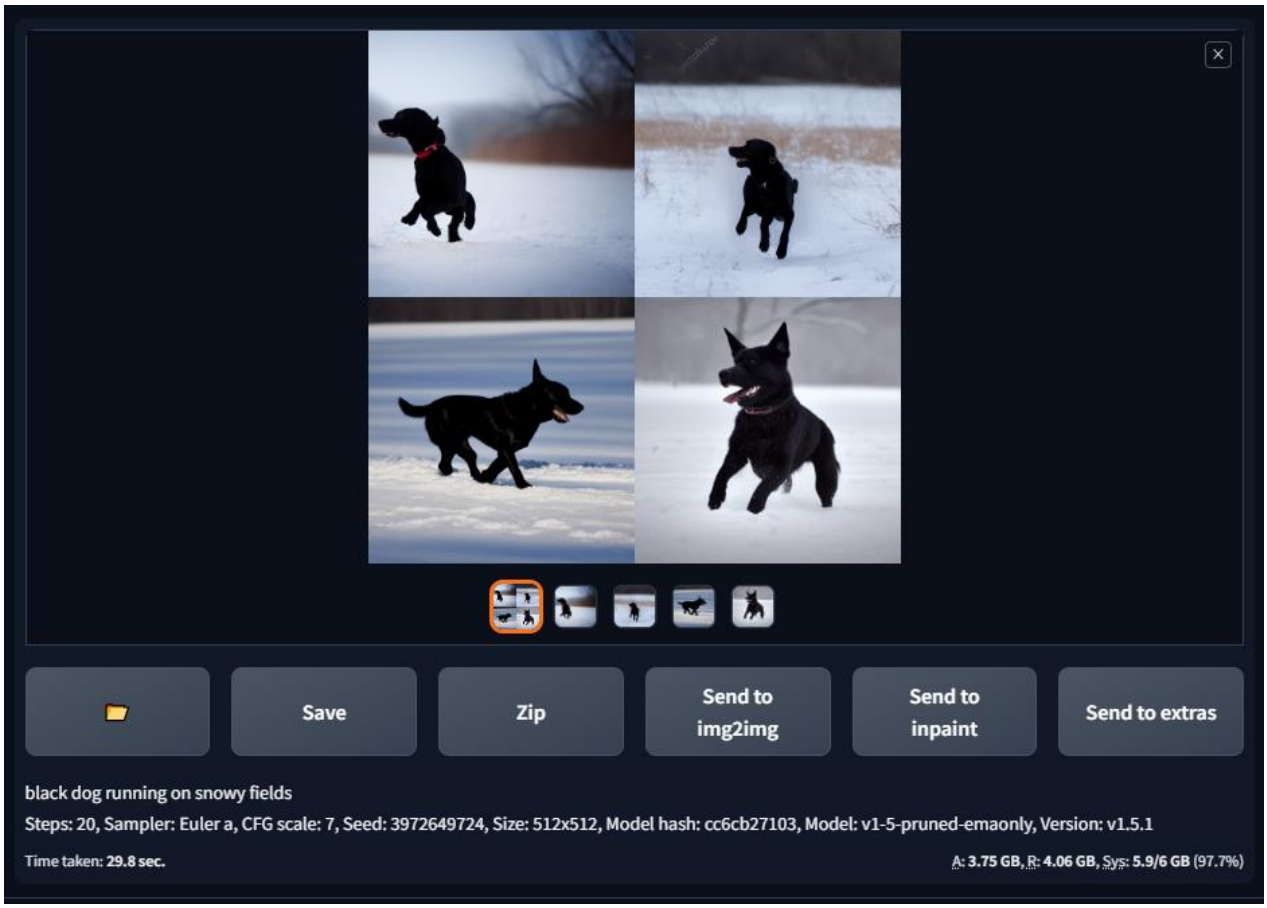
Slika 6.12. Prikaz dijela sučelja za specificiranje parametara generirane slike
Izvor: Autor

Pored dijela sučelja za specificiranje parametara generirane slike nalazi se dio sučelja za prikaz generirane slike. Tipke ispod prikaza slike nude mogućnosti za otvaranje mape u kojoj su spremljene generirane slike, mogućnost spremanja generirane slike, mogućnost izrade ZIP arhive, te mogućnosti slanja slike na ostale načine rada (način rada iz slike u sliku „img2img“, način rada nadopunjavanja nedostajećih dijelova slike „inpaint“, te dodatni način rada „extras“ koji uključuje način rada za super-rezoluciju. Na slici 6.13. je prikaz dijela sučelja za prikaz generirane slike.



Slika 6.13. Prikaz dijela sučelja za prikaz generirane slike
Izvor: Autor

Za generiranje četiri slike sa opisom „black dog running on snowy fields“ veličina 512x512, potrebno je upisati spomenuti opis u prvo tekstualno polje na dijelu sučelja za upisivanje opisa. Na dijelu sučelja za specificiranje parametara, polja „Width“ i „Height“ potrebno je postaviti na 512, te polje „Batch count“ na 4. Za dobivanje četiri slike moguće je i postavljanje vrijednosti „Batch count“ na 2, te „Batch size“ na 2. Klikom na tipku „Generate“ započinje proces generiranja slika. Na slici 6.14. je prikazan izgled dijela sučelja za prikaz generirane slike nakon uspješnog generiranja slika sa spomenutim parametrima. Automatski je kreiran kolaž sa svim generiranim slikama, te je generirane slike moguće pregledavati pojedinačno.

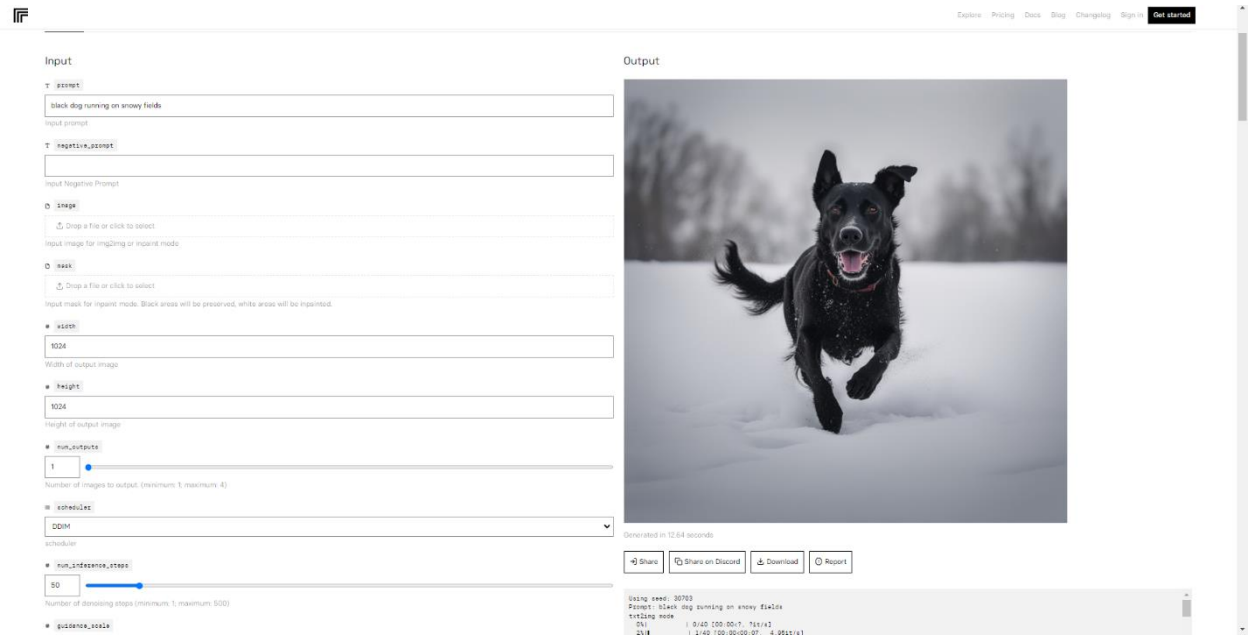


Slika 6.14. Prikaz dijela sučelja za prikaz generirane slike nakon uspješnog generiranja slika
Izvor: Autor

6.2.2. Pokretanje SDXL

Za pokretanje Stable Diffusion SDXL modela koristit će se web servis Replicate. Stvaranje četiri slike sa opisom istim kao i u prijašnjem primjeru moguće je upisivanjem opisa u tekstualno polje „prompt“, pod sekcijom „Input“. Ispod polja za upis parametara nalazi se tipka „Submit“, koja kada se klikne započinje proces stvaranja slika. Generirane slike nalaze se pored forme za

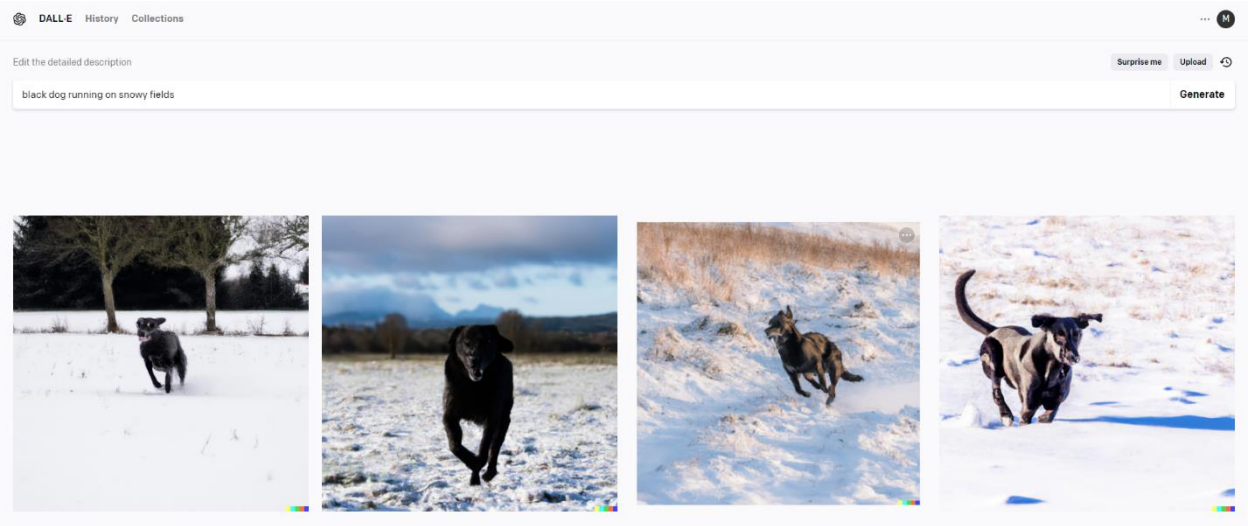
upis parametara. Moguće je pregledati izlaz s povratnim informacijama tijekom pokretanja modela. Izlaz uključuje „seed“ vrijednost slike. Slike je moguće podijeliti ili spremiti na računalo. Na slici 6.15. je prikaz ekrana sustava Replicate nakon što su generirane slike.



Slika 6.15. Replicate web sustav nakon generiranja slike
Izvor: Autor

6.2.3. Pokretanje DALL-E 2

DALL-E 2 model moguće je pokrenuti preko web sustava tvrtke OpenAI. Za generiranje četiri slika s istim opisom iz prijašnjeg primjera, potrebno je opis upisati u tekstualno polje web sustava. Klikom na tipku „Generate“ započinje proces generiranja četiri slike dimenzija 1024x1024. Nakon uspješnog generiranja slika, slike se prikazuju te ih je moguće preuzimati, uređivati u web uređivaću te stvarati varijacije iz odabrane slike. Stvaranje varijacija generira četiri nove slike bazirane na odabranoj slici. Na slici 6.16. je prikaz ekrana DALL-E 2 web sustava nakon uspješnog generiranja slika.



Slika 6.16. Prikaz DALL-E 2 web sustava nakon uspješnog generiranja slika
Izvor: Autor

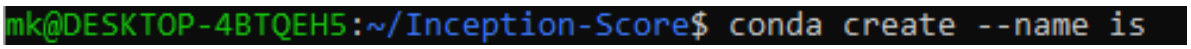
6.3. Priprema modela za objektivnu evaluaciju

Implementacije modela za objektivnu evaluaciju zahtijevaju instalaciju specifičnih biblioteka za specifične verzije Python programskog jezika, ovisno o vremenu kada je stvorena implementacija. U ovom radu korištena je Windows Subsystem for Linux (WSL) značajka Windows operativnog sustava za pripremu i pokretanje modela za objektivnu evaluaciju. Korištena su Python virtualna okruženja kreirana upraviteljem Python paketa Conda.

6.3.1. Priprema za ICS mjeru

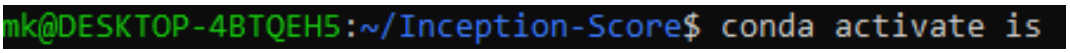
Implementacija ICS mjere u vrijeme izrade rada posljednji puta je ažurirana 2017. godine [79]. Potrebne biblioteke za pokretanje implementacije starije su u odnosu na biblioteke potrebne za ostale modele. Zbog toga je potrebno kreirati Python virtualno okruženje za ICS mjeru. Virtualno okruženje je izolirana kopija Pythona sa posebnim nazivom, koja sadrži svoje datoteke, mape i putanje koje rade sa specifičnim verzijama biblioteka ili Pythona bez utjecanja na ostale instalirane Python projekte [80]. Za izradu virtualnog okruženja u ovom radu koristi se upravitelj Python paketa miniconda.

Preduvjeti za pokretanje Python implementacije za ICS mjeru su instalirana biblioteka tensorflow 1.3.0, programski jezik Python 2.7.12, biblioteka numpy 1.13.1, te biblioteka scipy 0.17.0 [79]. Pomoću upravitelja paketa miniconda, izrađuje se virtualno okruženje. Naredba za izradu virtualnog okruženja, gdje je naziv okruženja „is“, je na slici 6.17.



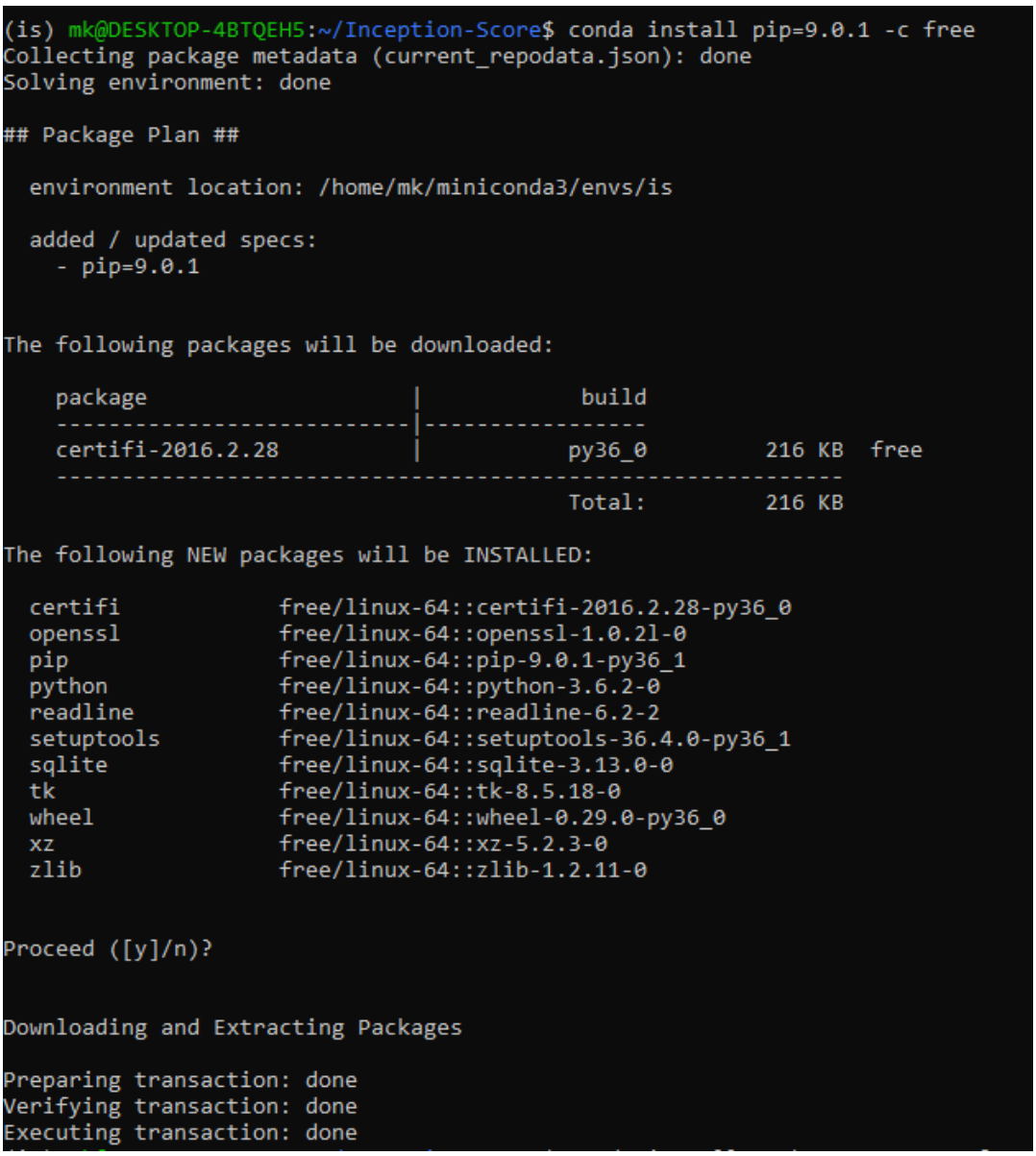
Slika 6.17. Naredba za izradu virtualnog okruženja
Izvor: Autor

Virtualno okruženje potrebno je aktivirati kako bi se instalacije paketa izvršile na željenom okruženju. Aktivacija je moguća naredbom sa slike 6.18.



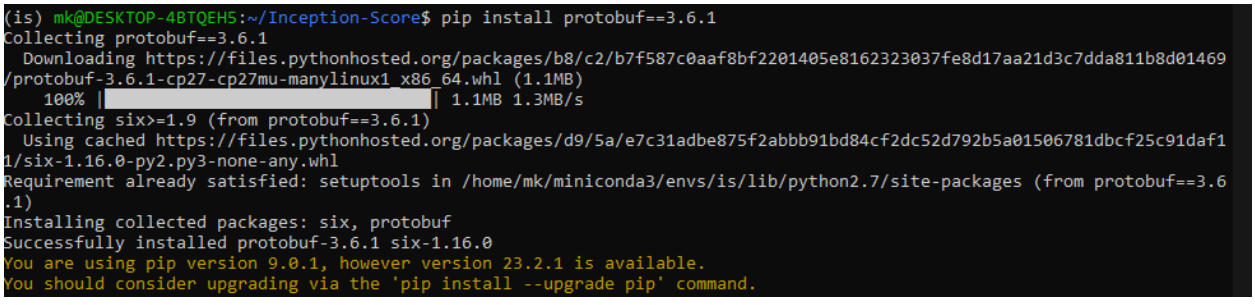
Slika 6.18. Naredba za aktivaciju virtualnog okruženja
Izvor: Autor

Biblioteke koje je potrebno instalirati ne nalaze se unutar kolekcije standardnih Anaconda biblioteka. Neke od njih potrebno je instalirati pomoću upravitelja paketa pip, a neke preko alternativnog Anaconda kanala za pakete. Potrebna verzija pip upravitelja paketa je 9.0.1, koju je moguće instalirati preko Anaconda kanala za pakete pod nazivom „free“. Na slici 6.19. je naredba za instalaciju pip upravitelja paketa preko kanala za pakete „free“, gdje je nakon -c specificiran naziv kanala. Prikazan je i izlaz nakon uspješne instalacije paketa. Znakovi „is“ unutar zagrada označavaju da je aktivirano virtualno okruženje naziva „is“.



Slika 6.19. Instalacija pip upravitelja paketa
Izvor: Autor

Pomoću upravitelja biblioteka pip potrebno je instalirati određene pakete. Instalacija jednog od potrebnih paketa prikazana je na slici 6.20., kao i prikaz nakon uspješne instalacije.



Slika 6.20. Instalacija paketa pomoću pip upravitelja paketa

Izvor: Autor

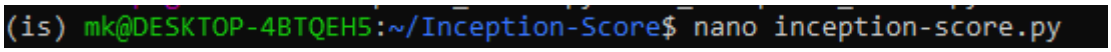
Na idućoj slici (6.21.) prikazana je lista paketa potrebnih za pokretanje modela za izračunavanje ICS mjere. Paketi za koje je naveden „Channel“ naziva „pypi“ instalirani su preko pip upravitelja paketa, što znači da ih je potrebno instalirati na način prikazan na slici 6.20. Paketi za koje je naveden „Channel“ naziva „free“ instalirani su preko Anaconda upravitelja za pakete, što znači da ih je potrebno instalirati na način prikazan na slici 6.19.

```
(inception_score) ┌──(mk㉿ DESKTOP-4BTQEH5)-[~/Inception-Score]
└─$ conda list
# packages in environment at /home/mk/miniconda3/envs/inception_score:
#
# Name                    Version                   Build  Channel
backports-functools-lru-cache 1.6.4                    pypi_0    pypi
backports-weakref         1.0.post1                 pypi_0    pypi
bleach                    1.5.0                     pypi_0    pypi
certifi                   2016.2.28                 py27_0    free
cycler                    0.10.0                    pypi_0    pypi
funcsigs                  1.0.2                     pypi_0    pypi
html5lib                  0.9999999                 pypi_0    pypi
imageio                   2.1.2                     pypi_0    pypi
kiwisolver                1.1.0                     pypi_0    pypi
markdown                  3.1.1                     pypi_0    pypi
matplotlib                2.2.5                     pypi_0    pypi
mock                      3.0.5                     pypi_0    pypi
numpy                     1.13.1                    pypi_0    pypi
openssl                   1.0.2l                         0    free
pillow                    6.2.2                     pypi_0    pypi
pip                       9.0.1                     py27_1    free
protobuf                  3.6.1                     pypi_0    pypi
pyparsing                 2.4.7                     pypi_0    pypi
python                    2.7.12                         1    free
python-dateutil           2.8.2                     pypi_0    pypi
pytz                      2023.3                    pypi_0    pypi
readline                  6.2                            2    free
scipy                     0.17.0                    pypi_0    pypi
setuptools                36.4.0                    py27_1    free
six                       1.16.0                    pypi_0    pypi
sqlite                    3.13.0                         0    free
subprocess32              3.5.4                     pypi_0    pypi
tensorflow                1.3.0                     pypi_0    pypi
tensorflow-tensorboard    0.1.8                     pypi_0    pypi
tk                        8.5.18                         0    free
werkzeug                  1.0.1                     pypi_0    pypi
wheel                     0.29.0                    py27_0    free
xz                        5.2.3                          0    free
zlib                      1.2.11                         0    free
```

Slika 6.21. Prikaz liste paketa virtualnog okruženja za ICS mjeru
Izvor: Autor

Nakon što su instalirani potrebni paketi, potrebno je kreirati mapu u kojoj će se nalaziti datoteke potrebne za računanje ICS mjere. Unutar kreirane mape potrebno je kreirati podmapu naziva

„data“, u koju će se postaviti željene slike za evaluaciju. Slike moraju biti PNG formata. Izvan podmape „data“, potrebno je kreirati novu datoteku formata „.py“, koja će sadržavati kod potreban za računanje ICS mjere. Na slici 6.22. je prikaz naredbe za kreiranje nove Python datoteke, gdje je „inception-score.py“ naziv datoteke. Izvršenjem naredbe na slici otvara se uređivač teksta.



Slika 6.22. Kreiranje Python datoteke
Izvor: Autor

Potreban kod koji kreirana datoteka treba sadržavati nalazi se na slici 6.23.

```
from math import floor
from numpy import ones
from numpy import expand_dims
from numpy import log
from numpy import mean
from numpy import std
from numpy import exp
from numpy.random import shuffle
from keras.applications.inception_v3 import InceptionV3
from keras.applications.inception_v3 import preprocess_input
from keras.datasets import cifar10
from skimage.transform import resize
from numpy import asarray

from PIL import Image
import os.path
import numpy as np
import glob

def scale_images(images, new_shape):
    images_list = list()
    for image in images:
        new_image = resize(image, new_shape, 0)
        images_list.append(new_image)
    return asarray(images_list)

def calculate_inception_score(images, n_split=10, eps=1E-16):
    model = InceptionV3()
    scores = list()
    n_part = floor(images.shape[0] / n_split)
    for i in range(n_split):
        ix_start, ix_end = i * n_part, (i+1) * n_part
        subset = images[ix_start:ix_end]
        subset = subset.astype('float32')
        subset = scale_images(subset, (299,299,3))
        subset = preprocess_input(subset)
```

```
        p_yx = model.predict(subset)
        p_y = expand_dims(p_yx.mean(axis=1), 0)
        kl_d = p_yx * (log(p_yx + eps) - log(p_y + eps))
        sum_kl_d = kl_d.sum(axis=1)
        avg_kl_d = mean(sum_kl_d)
        is_score = exp(avg_kl_d)
        scores.append(is_score)
    is_avg, is_std = mean(scores), std(scores)
    return is_avg, is_std

if __name__=='__main__':
    (images_all, _), (_, _) = cifar10.load_data()
    shuffle(images_all)
    images=images_all[0:100]
    print('loaded', images.shape)
    is_avg, is_std = calculate_inception_score(images)
    print('score', is_avg, is_std)

    def get_images(filename):
        return np.asarray(Image.open(filename))

    filenames = glob.glob(os.path.join('./data', '*.*'))
    images2 = [get_images(filename) for filename in filenames]
    images3=np.asarray(images2)
    print('loaded', images3.shape)
    is_avg3, is_std3 = calculate_inception_score(images3)
    print('score', is_avg3, is_std3)
```

Slika 6.23. Kod za računanje ICS mjere
Izvor: Autor

Prikazan kod potrebno je spremiti u kreiranu datoteku.

6.3.2. Priprema za FID mjeru

Implementaciju modela za FID mjeru moguće je instalirati preko pip upravitelja paketa. Paketi koji su preduvjeti za instalaciju implementacije su python3, pytorch, torchivison, pillow, numpy i scipy [81]. Specifične verzije paketa preduvjeta uz koje je implementacija u vrijeme izrade rada bila funkcionalna su torch 2.0.0 (noviji naziv za pytorch), torchvision 0.15.1, numpy 1.23.5, scipy 1.10.1, te Pillow 9.5.0. Nakon kreiranja i aktivacije virtualnog okruženja na način prikazan na slikama 6.17. i 6.18., spomenute pakete potrebno je instalirati pip upraviteljem paketa na način prikazan na slici 6.20. Naredba za instalaciju algoritma FID mjere nalazi se na idućoj slici (6.24.).

```
(fid) mk@DESKTOP-4BTQEH5:~/fid$ pip install pytorch-fid
```

Slika 6.24. Naredba za instalaciju implementacije za FID mjeru
Izvor: Autor

Slike za evaluaciju trebaju se nalaziti u dvije različite mape. Slike moraju biti jednakih veličina. Kako bi se veličina slika promijenila programski, potrebno je napraviti modifikaciju u kodu implementacije FID mjere. Datoteka koju je potrebno modificirati nalazi se u mapi na putanji sa slike 6.25., gdje je „fid“ naziv virtualnog okruženja za FID score.

```
miniconda3 › envs › fid › lib › python3.11 › site-packages › pytorch_fid
```

Slika 6.25. Putanja mape za pytorch_fid implementaciju
Izvor: Autor

Naziv datoteke koju je potrebno modificirati je „fid_score.py“. Slijedeću liniju koda prikazanu na slici 6.26. potrebno je izmijeniti.

```
dataset = ImagePathDataset(files, transforms=TF.ToTensor())
```

Slika 6.26. Linija koda koju je potrebno izmijeniti
Izvor: Autor

Spomenuti kod treba izmijeniti u kod na idućoj slici (6.27.). U ovom primjeru dimenzije u koje će se promijeniti slike je 512x512, brojeve je potrebno izmijeniti ovisno o željenim dimenzijama.

```
    #transforming image size
        dataset = ImagePathDataset(files,
        transforms=TF.Compose([
            TF.Resize((512,512)),
            TF.ToTensor(),
    ]))
```

Slika 6.27. Izmjena koda za programsku izmjenu veličina slika
Izvor: Autor

6.3.3. Priprema za CLIP mjeru

Za instalaciju implementacije za CLIP mjeru, potrebno je preuzeti kod za implementaciju, na način kao što je prikazano na slici 6.1., uz izmjenu poveznice nakon naredbi „git clone“ u iduću

poveznicu: https://github.com/jmhessel/clipscore.git. Nakon toga potrebno je preuzimanje potrebnih paketa. U repozitoriju za CLIP mjeru nalazi se datoteka naziva „requirements.txt“ koja sadrži poveznice na potrebne pakete. Sve potrebne pakete spomenute u toj datoteci moguće je instalirati preko naredbe na slici 6.28.

```
(clipscore) mk@DESKTOP-4BTQEH5:~/clipscore/clipscore$ pip install -r requirements.txt
```

Slika 6.28. Naredba za instalaciju paketa iz liste unutar datoteke
Izvor: Autor

Nakon instalacije paketa iz „requirements.txt“ potrebno je još instalirati paket „scikit-learn“ na način prikazan na slici 6.29. Unutar mape sa kodom za CLIP mjeru nalazi se podmapa naziva „example“. U spomenutoj mapi u podmapu „images“ potrebno je staviti željene slike za evaluaciju. Unutar mape „example“ nalazi se datoteka „good_captions.json“, unutar koje je potrebno upisati opise koji se žele evaluirati. Opisi se upisuju u formatu prikazanom na slici, gdje su „image1“, „image2“ i „image3“ nazivi datoteka slika parirani sa opisima koji se odnose na sliku.

```
{
  "image1": "dog walking on the street",
  "image2": "cat walking on grass",
  "image3": "person sitting in a restaurant"
}
```

Slika 6.29. Opisi slika za CLIP mjeru
Izvor: Autor

6.3.4. MSCOCO

MSCOCO je skup slika kompleksnih scena koje sadrže česte objekte u njihovim prirodnim kontekstima. Skup slika sadrži 91 vrstu objekata kroz 328 tisuća slika [82], od kojih više od 200 tisuća sadrži oznake. Označene slike imaju 5 opisa po slici [83].

Računanje FID vrijednosti zahtijeva dva seta slika, od kojih jedan set predstavljaju slike iz stvarnosti, dok su u drugom setu slika generirane slike. U ovom radu, slike iz stvarnosti preuzet će se iz MSCOCO seta, specifično seta validacijskih slika 2017. Iz opisa preuzetih slika generirat će se nove slike koje će predstavljati set generiranih slika. Isti opisi koristit će se pri računanju CLIPScorea. Na slici 6.30. je prikaz opisa slika u MSCOCO setu, gdje „imgId“ predstavlja broj opisane slike, a „Caption“ stupci predstavljaju opise slike.

	A	B	C
1	imgId	Caption1	Caption2
2	397133	A man is in a kitchen making pizzas.	Man in apron standing on front of oven with pans and bakeware
3	37777	The dining table near the kitchen has a bowl of fruit on it.	A small kitchen has various appliances and a table.
4	252219	a person with a shopping cart on a city street	City dwellers walk by as a homeless man begs for cash.
5	87038	A person on a skateboard and bike at a skate park.	A man on a skateboard performs a trick at the skate park
6	174482	a blue bike parked on a side walk	A bicycle is chained to a fixture on a city street
7	403385	A bathroom that has a broken wall in the shower.	A bathroom looks clean but is missing tile at the shower stall.
8	6818	a couple of buckets in a white room	A bathroom with no toilets and a red and green bucket.
9	480985	The shiny motorcycle has been put on display.	The new motorcycle on display is very shiny.
10	458054	A row of white toilets sitting on top of a dirt ground.	A bunch of dirty looking white toilets in a row outside.
11	331352	A small closed toilet in a cramped space.	A tan toilet and sink combination in a small room.
12	296649	People are walking and riding motorcycles on the street	A group of motorists pass very large buildings in asia.
13	386912	a person sitting at a desk with a keyboard and monitor	a woman at her desk sits intently and happy
14	502136	A building wall and pair of doors that are open, along with vases of flowers on the outside of the building.	a building with dirty walls and dirty doors
15	491497	A fluffy white chair that faces away fro a television.	A pillow covered reading chair in the corner of the living room
16	184791	A painting of a table with fruit on top of it.	Painting of oranges, a bowl, candle, and a pitcher
17	348881	A large jetliner sitting on top of an airport runway.	Airline employees by an aircraft parked at the gate
18	289393	Set of toy animals sitting in front of a red wooden wagon.	Several toy animals - a bull, giraffe, deer and parakeet.
19	522713	a bench sitting in the grass facing the water and boats	A wooden bench overlooking a harbor beneath a cloudy sky.
20	181666	a flock of goats and some men watching them	a man and a large herd of goats in a desert setting.
21	17627	Several cars are parked in front of a building.	A green car has parked on the curb in a parking lot
22	143931	A political candidate advertisement on the side of a coach bus.	Here is Massachusettes candidate Scott Brown's campaign trailer.
23	303818	A bus traveling down a curvy road behind a black car.	A group of people cross the curved street.
24	463730	Two buses parked in a parking lot next to cars.	People crossing a busy city street full of vehicles.
25	460347	A bus traveling on a freeway next to other traffic.	A bus and other cars driving down a multi-laned street .
26	322864	A yellow taxi cab sitting below tall buildings.	A picture of an animal is on a pole and next to it is a yellow taxi.
27	226111	A No bicycles, skates or skateboards sign on a pole.	a close up of a vandalized street sign
28	153299	two giraffes are standing together outside a barn	A mother giraffe is standing with a baby giraffe.
29	308394	An older woman riding a train while sitting under it's window.	A woman with an umbrella on a commuter train takes a snooze
30	456496	A woman sitting in front of the Eiffel tower near pigeons.	A woman sitting on ledge with three pigeons, with gate railing, trees, and base of the Eiffel Tower behind.
31	58636	A large wooden pole with a green street sign hanging from it.	street signs on the corner of Gladys and Detroit
32	41888	Three birds walking around a dry grass field.	There are 3 female peacocks together walking around.
33	184321	A train traveling down tracks next to lights.	A blue and silver train next to train station and trees.
34	565778	A train going back to its coarse filled with people.	A blue commuter train traveling towards a train tunnel.
35	297343	A stop sign has been placed upside-down in the grass beside a building.	A stop sighn is propped up against the side of a building.
36	336587	a stop sign sittin on a pole that is somewhat broken	A stop sign on a broken post across the street from houses.
37	122745	A stop sign is lit up in the dark of night.	A red stop sign sitting on the side of a dark road.
38	219578	A dog and cat lying together on an orange couch.	A dog and a cat curled up together on a couch.
39	555705	Two cats sitting on top of a pair of shoes.	Two cats are outside and perched on someone's sneakers.
40	443303	A cat laying on clothes that are in a suitcase.	A cat that is laying in a piece of luggage.

Slika 6.30. Opisi slika unutar MSCOCO seta
Izvor: Autor

6.4. Pokretanje modela za objektivnu evaluaciju

Ukoliko su kreirana posebna virtualna okruženja za svaki model objektivne evaluacije, za uspješno pokretanje modela potrebno je aktivirati virtualno okruženje za odabrani model, kao što je prikazano na slici 19. Daljnji koraci ovise o samom modelu koji se koristi.

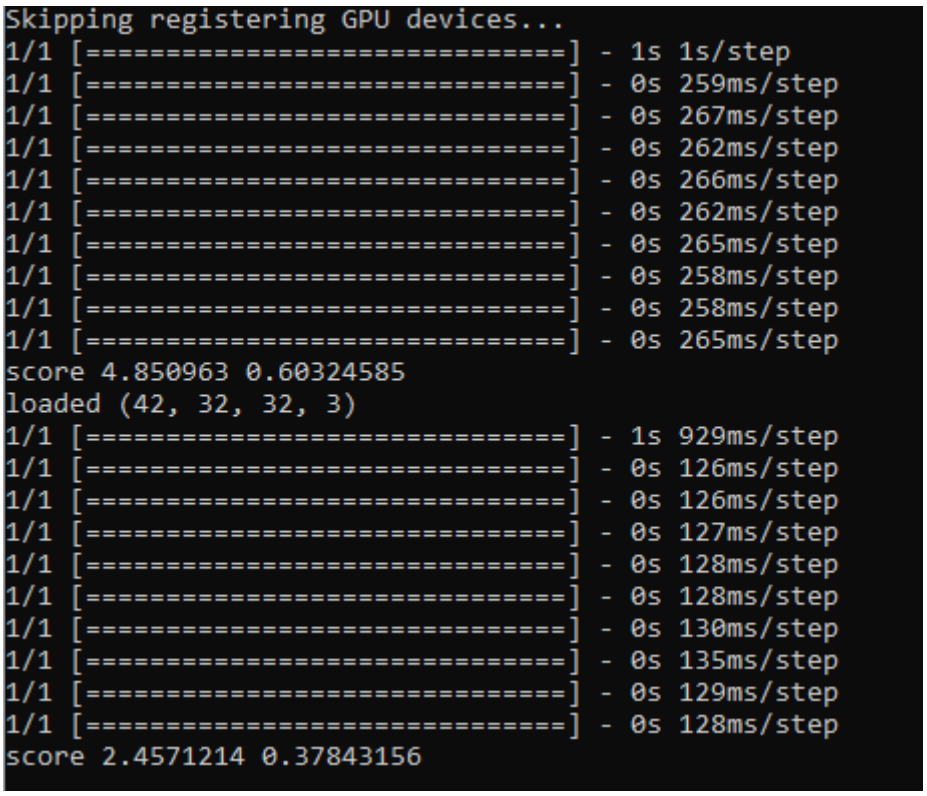
6.4.1. Izračunavanje ICS mjere

Za pokretanje modela za ICS mjeru, potrebno je preseliti se u mapu u kojoj se nalazi kod za implementaciju ICS mjere. Zatim je potrebno pokrenuti kreiranu skriptu. Na slici 6.31. je primjer naredbe za pokretanje Python skripte, gdje je „inception-score.py“ naziv datoteke.

```
(is) mk@DESKTOP-4BTQEH5:~/Inception-Score$ python3 inception-score.py
```

Slika 6.31. Naredba za pokretanje Python datoteke
Izvor: Autor

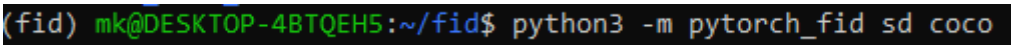
Slika 6.32. prikazuje primjer izlaza nakon uspješnog računanja ICS mjere. Prvi ispisani rezultat (u ovom slučaju 4.85 i 0.60) vrijednosti su izračunate ICS mjere na nasumično učitanim slikama iz CIFAR-10 podatkovnog seta, dok drugi ispisani rezultat (2.45 i 0.37) predstavlja vrijednosti izračunate na slikama iz „data“ direktorija. Prva vrijednost rezultata predstavlja srednji rezultat, tj. samu ICS mjeru za evaluirane slike, dok je druga vrijednost standardna devijacija ICS mjere [59].



Slika 6.32. Izlaz nakon pokretanja ICS mjere
Izvor: Autor

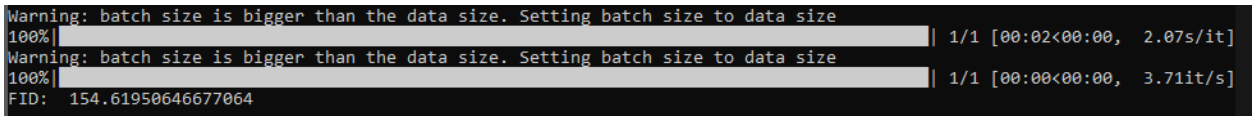
6.4.2. Izračunavanje FID mjere

Model za FID score pokreće se naredbom sa slike 6.33., gdje „sd“ i „coco“ predstavljaju nazive direktorija sa slikama koje se evaluiraju.



Slika 6.33. Naredba za pokretanje modela za FID mjeru
Izvor: Autor

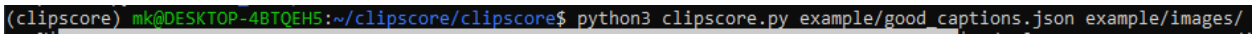
Izlaz nakon uspješnog računanja FID mjere prikazan je na slici 6.34. U ovom primjeru, rezultat je 154.619.



Slika 6.34. Izlaz nakon pokretanja modela za FID mjeru
Izvor: Autor

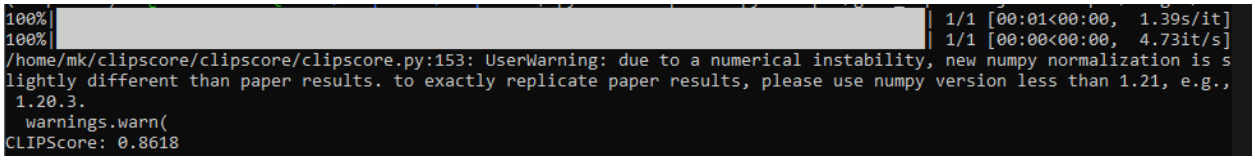
6.4.3. Izračunavanje CLIP mjere

Model za CLIP mjeru pokreće se pomoću naredbe na slici 6.35., gdje je „example/good_captions.json“ putanja do datoteke sa opisima za evaluaciju, dok je „example/images/“ putanja direktorija sa slikama koje će se evaluirati.



Slika 6.35. Naredba za pokretanje modela za CLIP mjeru
Izvor: Autor

Slika 6.36. prikazuje izlaz nakon uspješnog računanja CLIP mjere, koja u ovom primjeru iznosi 0.861.



Slika 6.36. Izlaz nakon pokretanja modela za CLIP mjeru
Izvor: Autor

6.5. Objektivna evaluacija

U nastavku će se opisati proces objektivne evaluacije odrađene za potrebe rada. Modelima za stvaranje slika generirati će se slike, koristeći opise iz MSCOCO skupa slika. Generirane slike će se evaluirati modelima za evaluaciju, dok će pripadajuće MSCOCO slike predstavljati skup slika iz stvarnog svijeta.

6.5.1. Stvaranje slika

Iz MSCOCO skupa slika nasumično se odabire 10 slika. Spremaju se opisi odabranih slika. Odabrane slike iz MSCOCO skupa slika spremaju se u zaseban direktorij. Iz opisa odabranih slika Stable Diffusion 1.5 modelom kreiraju se četiri slike dimenzija 1024x1024. Iz istih opisa stvaraju se slike i DALL-E 2 modelom. DALL-E 2 model stvara slike isključivo dimenzija 1024x1024. Stable Diffusion model treniran je na slikama dimenzija 512x512 [84]. Zbog toga će se radi usporedbe Stable Diffusion 1.5 modelom iz istih opisa stvoriti još četiri slike dimenzija 512x512. Stable Diffusion SDXL modelom stvoriti će se slike dimenzija 1024x1024.

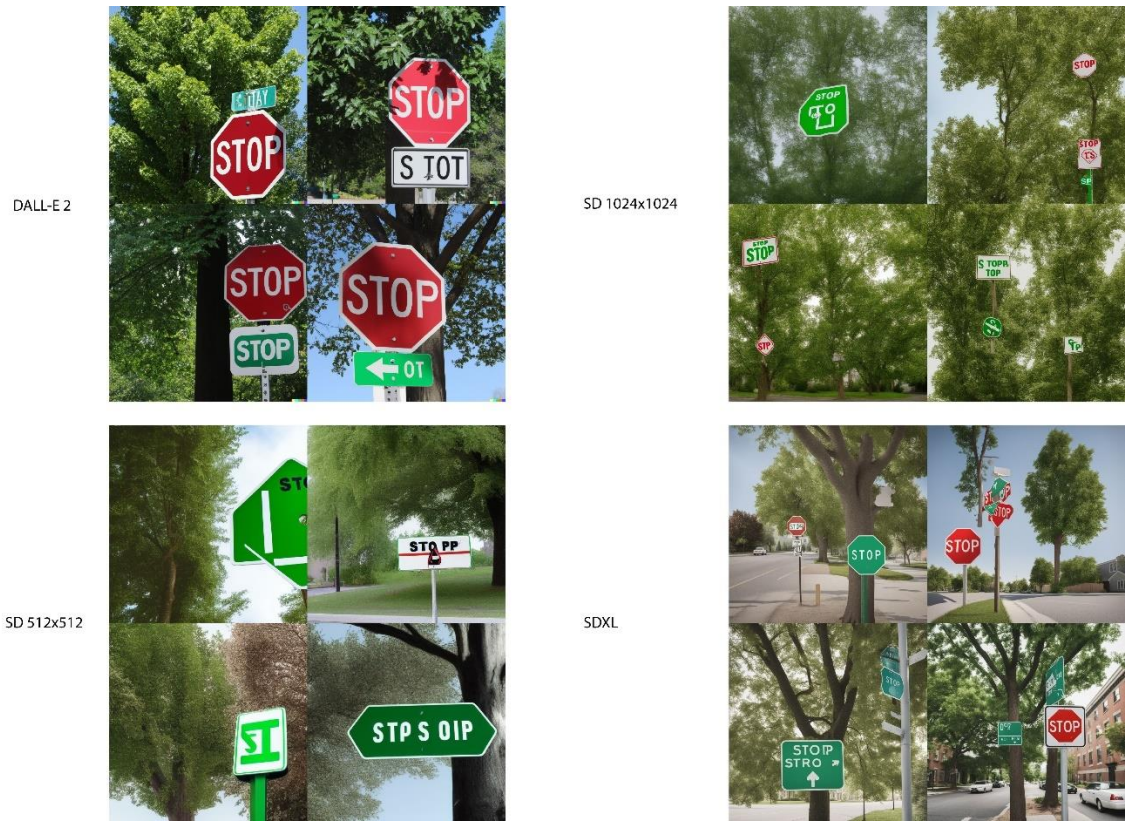
Na slici 6.37. je referentna slika iz MSCOCO skupa za opis „Living room with white chairs and couches, fireplace and books on bookshelves.“ što prevedeno znači „dnevna soba sa bijelim stolicama i kaučevima, kaminom i knjigama na policama za knjige“.

Slika 6.37. Primjer slike iz MSCOCO skupa
Izvor: Autor

Iduća slika (6.38.) prikazuje rezultate generiranja slika za spomenuti opis. U gornjem lijevom kutu nalaze se 4 slike generirane DALL-E 2 modelom. Ostale mreže slika sadrže slike generirane Stable Diffusion modelom. Prva mreža slika u donjem lijevom kutu sadrži slike dimenzija 512x512 generirane Stable Diffusion 1.5 modelom, dok druga mreža slika u gornjem desnom kutu sadrži slike dimenzija 1024x1024 generirane istom verzijom modela. U donjem desnom kutu nalazi se mreža slika dimenzija 1024x1024 generiranih SDXL modelom.

Slika 6.38. Primjer generiranih slika
Izvor: Autor

Slika broj 6.39. prikazuje rezultate generiranja slika za opis „a stop sign and a white and green street sign and a tree“, što u prijevodu znači „znak stop, bijeli i zeleni ulični znak te drvo“.



Slika 6.39. Primjer generiranih slika
Izvor: Autor

Iduća slika (6.40.) prikazuje rezultate generiranja slika za opis „Two people stand near a bike wearing helmets.“, što prevedeno znači „dvoje ljudi stoji blizu bicikla noseći kacige“.

Slika 6.40. Primjer generiranih slika
Izvor: Autor

6.5.2. Evaluacija slika

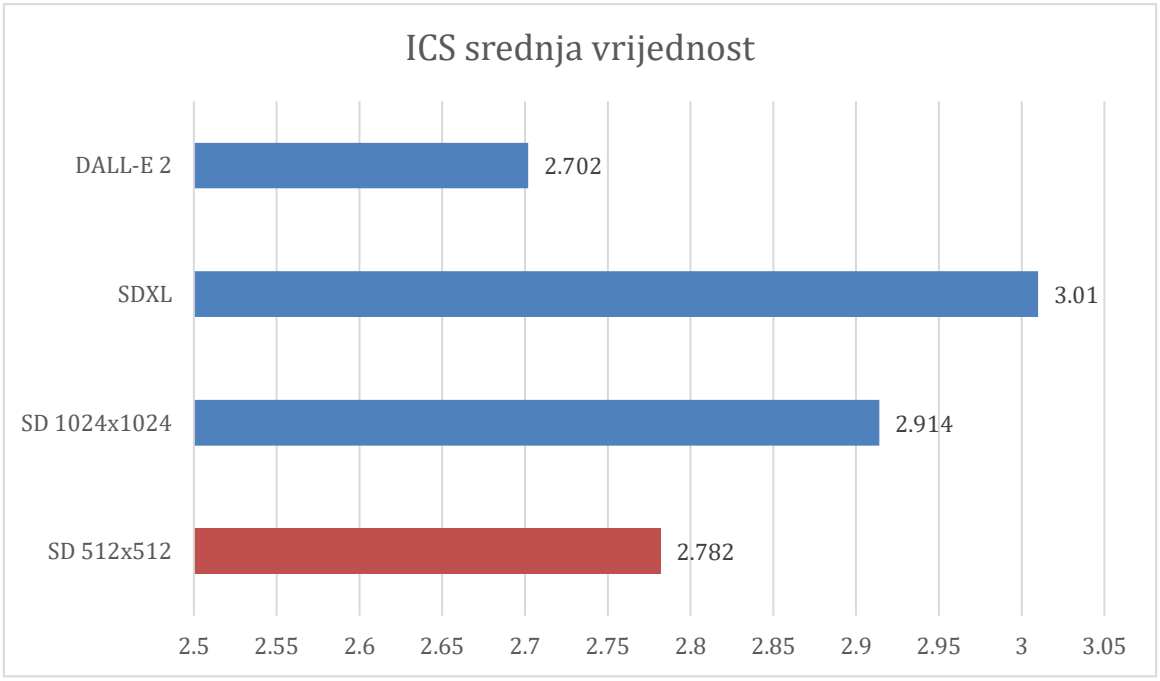
Generirani skupovi slika (DALL-E, Stable Diffusion 1.5 512x512, Stable Diffusion 1.5 1024x1024 te SDXL) bit će evaluirani modelima za objektivnu evaluaciju. Slike iz MSCOCO skupa predstavljat će skup slika iz stvarnog svijeta pri računanju FID mjere. FID mjera će se računati više puta, jednom za slučaj u kojem se slike iz MSCOCO skupa smanjuju pomoću programa za uređivanje slika, te drugi puta za slučaj u kojem se slike smanjuju programski na dimenzije 512x512. Potrebno je promijeniti dimenzije slika zbog zahtijeva modela za računanje FID mjere da sve slike unutar skupa budu istih dimenzija. Slike iz MSCOCO skupa dolaze u različitim dimenzijama, dok su slike unutar specifičnih skupova generiranih slika jednakih dimenzija. Za računanje CLIP mjere, za svaki opis nasumično će se odabrati jedna od generirane slike iz skupova. U drugom računanju CLIP mjere, koristiti će se sve generirane slike za svaki od opisa.

6.6. Analiza rezultata

Izračunati su rezultati evaluacije za skupove generiranih slika, u obliku mjera za objektivnu evaluaciju. Svaka od mjera sadrži više vrsta rezultata, ovisno o korištenim podacima, načinu pripreme podataka ili samoj vrsti objektivne mjere. Dimenzije skupova slika SDXL i DALL-E 2 modela su 1024x1024. Jedan od skupova slika modela Stable Diffusion 1.5 je dimenzija 512x512, dok je drugi skup dimenzija 1024x1024. Potrebno je napomenuti da zbog dimenzija slika rezultat skupa Stable Diffusion 1.5 dimenzija 512x512 nije direktno usporediv sa ostalim skupovima slika koji su dimenzija 1024x1024.

6.6.1. Analiza ICS mjere

Viša ICS mjera pokazuje da je model sposoban stvaranja mnogo raznolikih slika. Računanjem ICS mjere kao rezultat dobivaju se dvije različite vrijednosti, od kojih prva predstavlja „mean Inception score“, tj. srednju vrijednost ICS mjere izračunatu za skup slika. Slika 6.41. prikazuje grafički prikaz prve vrijednosti rezultata.

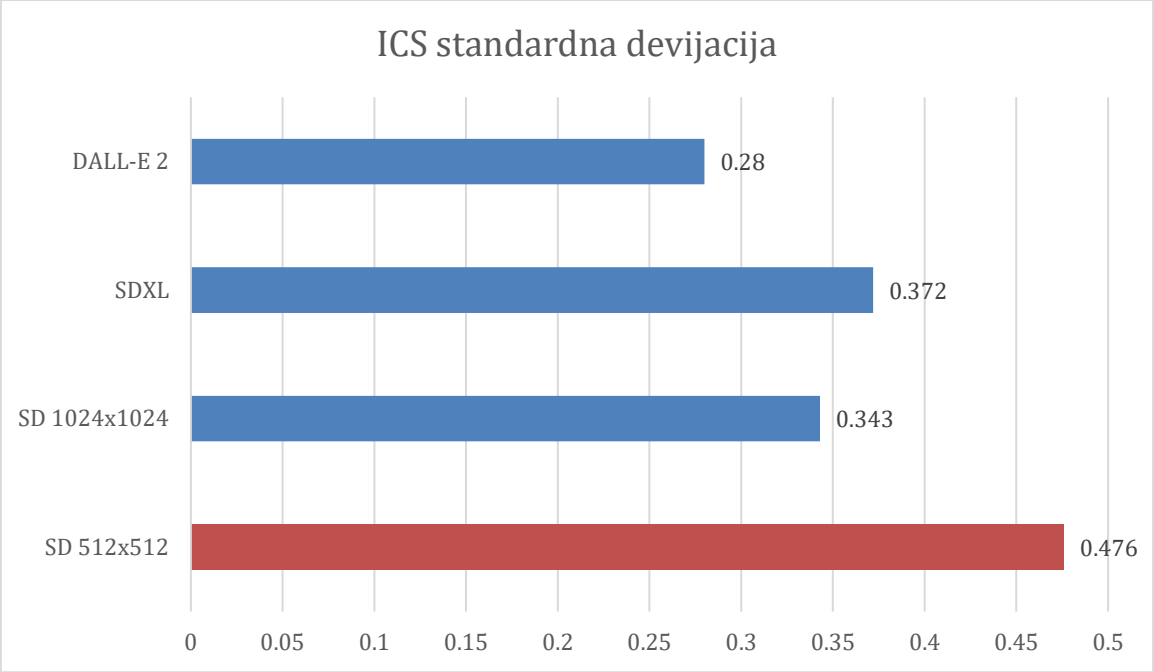


Slika 6.41. ICS srednja vrijednost
Izvor: Autor

Stable Diffusion 1.5 model treniran je na slikama dimenzija 512x512. Skup slika dimenzija 512x512 generiranih Stable Diffusion 1.5 modelom ostvaruje niži rezultat ICS mjere (2.782) u odnosu na slike dimenzija 1024x1024 generiranih istim modelom (2.914). Najniži rezultat

ostvaruje skup slika generiran DALL-E 2 modelom (2.702). Najnoviji od evaluiranih modela, Stable Diffusion model SDXL ostvaruje najveći rezultat (3.01).

Druga vrijednost rezultata ICS mjere predstavlja njezinu standardnu devijaciju. Na slici 6.42. je grafički prikaz vrijednosti standardne devijacije.

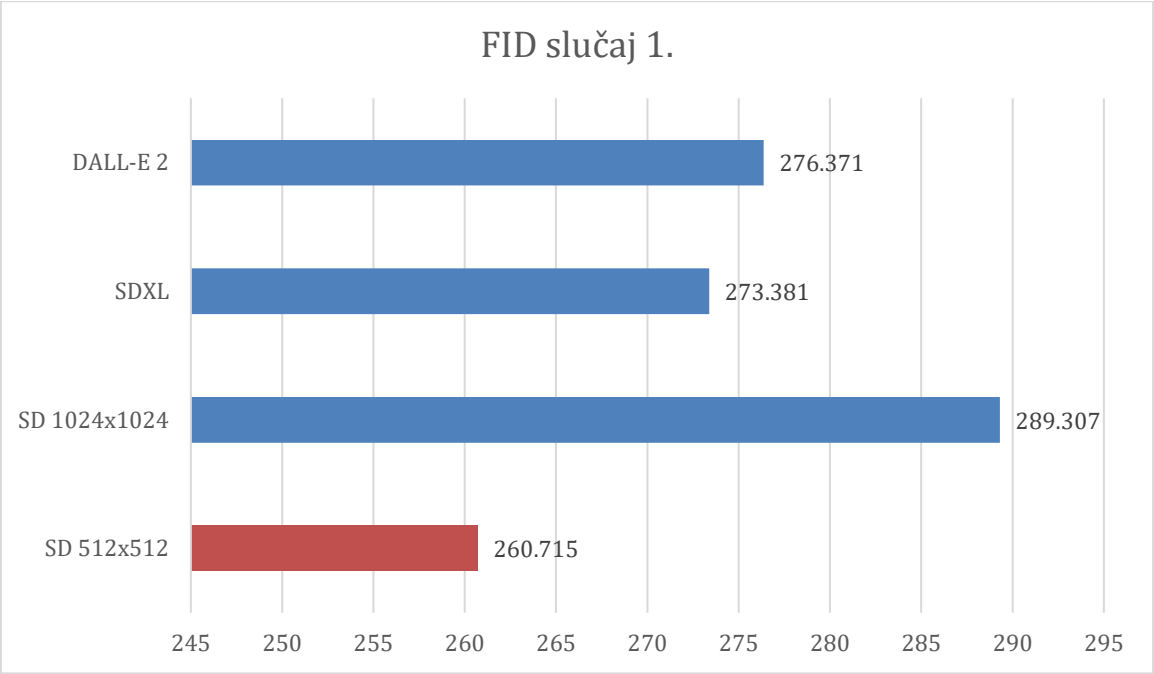


Slika 6.42. ICS standardna devijacija
Izvor: Autor

Rezultati ukazuju da skup slika dimenzija 512x512 generiranih Stable Diffusion 1.5 modelom ima najveću raznolikost u vrijednostima ICS mjere (0.476). Slike dimenzija 1024x1024 generirane istim modelom imaju manju standardnu devijaciju (0.343). Skup slika SDXL modela sa 0.372 ima nešto veću standardnu devijaciju u odnosu na skup slika istih dimenzija Stable Diffusion 1.5 modela. Najmanju raznolikost u vrijednostima imaju slike generirane DALL-E 2 modelom (0.28). To ukazuje da slike iz DALL-E 2 skupa imaju najkonzistentniju raznolikost.

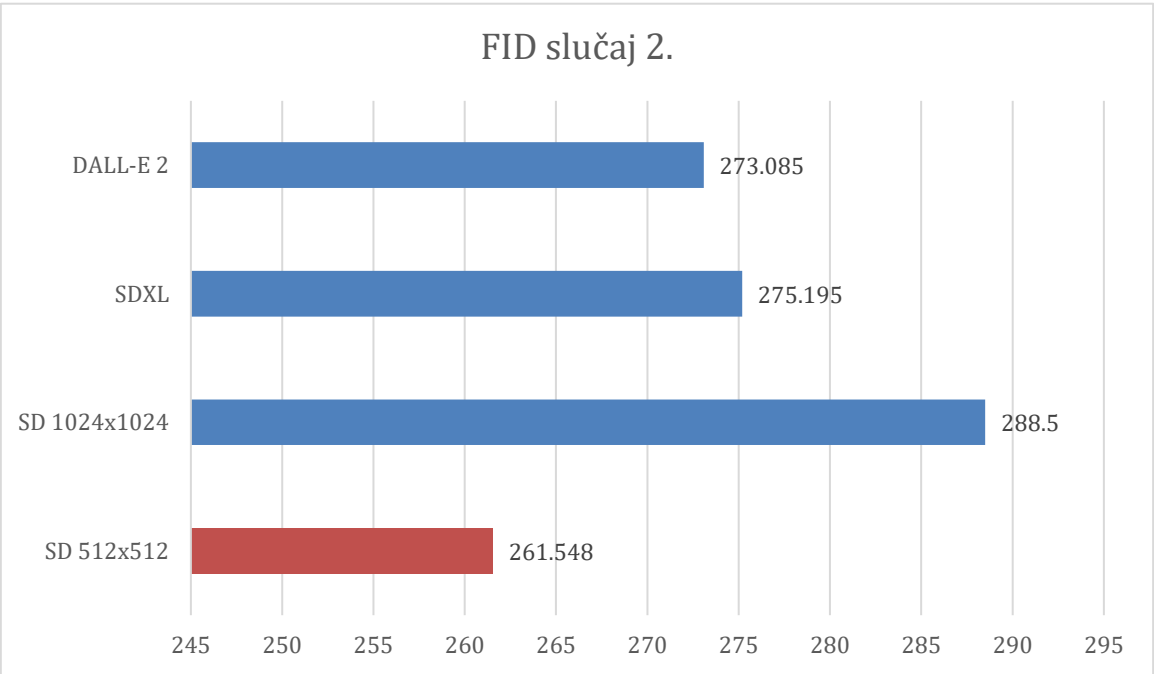
6.6.2. Analiza FID mjere

FID rezultat prikazuje koliko su dvije grupe slika statistički slične. Niži rezultati ukazuju na veću sličnost između grupa slika, dok savršen 0.0 rezultat ukazuje da su dvije grupe slika jednake. Niži rezultati pokazali su korelaciju sa slikama više kvalitete. Izračunata su dva skupa rezultata FID mjere. Prvi skup rezultata izračunat je za slučaj u kojem je veličina slika izmijenjena pomoću programa za uređivanje slika, dok drugi skup rezultata predstavlja slučaj u kojem je veličina slika izmijenjena programski. Na slici 6.43. je grafički prikaz prvog skupa rezultata.



Slika 6.43. FID slučaj 1.
Izvor: Autor

Rezultati ukazuju da u slučaju izmjene veličine slika pomoću programa za uređivanje slika, slike dimenzija 512x512 generirane Stable Diffusion 1.5 modelom ostvaruju najniži rezultat FID mjere (260.715), što ukazuje da su slike iz tog skupa najviše kvalitete. Od skupova slika dimenzija 1024x1024, SDXL skup ostvaruje najniži rezultat FID mjere (273.381). Slike generirane DALL-E 2 modelom ostvaruju još veći (276.371), dok je najveći rezultat ostvaren za skup Stable Diffusion 1.5 slika dimenzija 1024x1024 (289.307). Grafički prikaz drugog skupa rezultata je na slici 6.44.

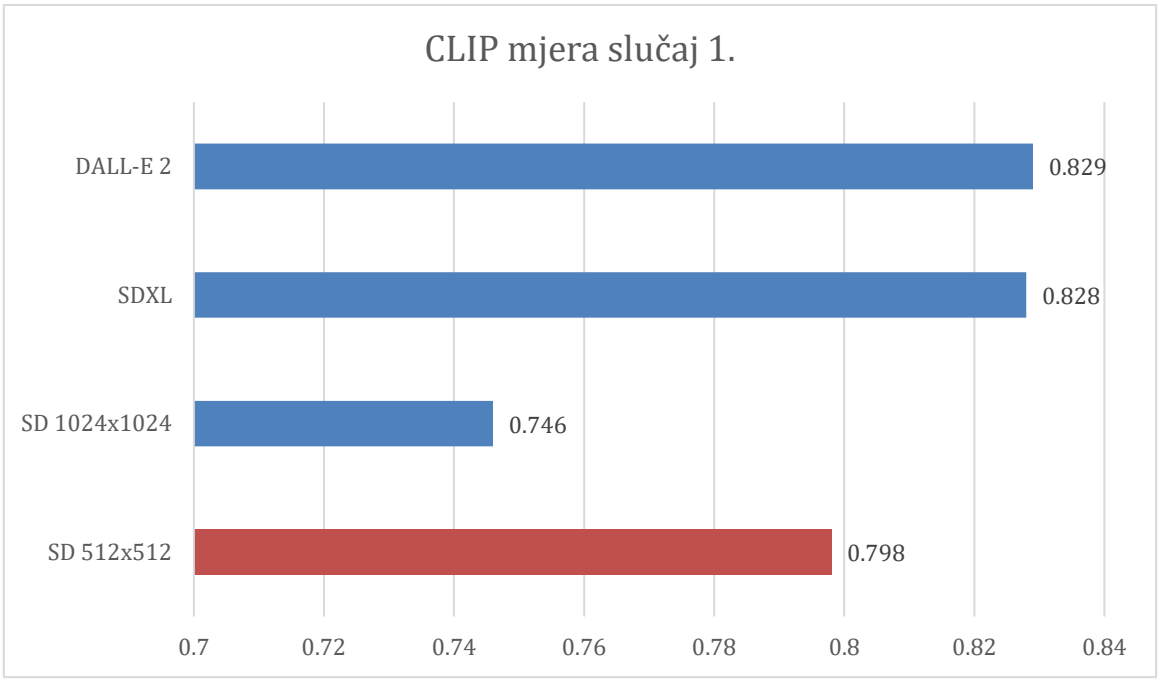


Slika 6.44. FID slučaj 2.
Izvor: Autor

Za slučaj izmjene veličine slika programski, dolazi do izmjene poredka između DALL-E 2 i SDXL skupa slika. U ovom slučaju, niži rezultat ostvaruje DALL-E 2 skup slika u odnosu na SDXL. Skup Stable Diffusion 1.5 512x512 ostvaruje najniži rezultat FID mjere, koji je u odnosu na rezultat sa slučaja 1. malo viši (261.548). Najniži rezultat FID mjere od slika dimenzija 1024x1024 ostvaruje skup slika DALL-E 2 modela, i to nešto manji u odnosu na FID slučaj 1. (273.085). Slike iz SDXL skupa ostvaruju veći FID score u odnosu na prvi slučaj (275.195). Malo niži rezultat u odnosu na slučaj 1. ostvaruje skup slika Stable Diffusion 1.5 1024x1024 (288.5).

6.6.3. Analiza CLIP mjere

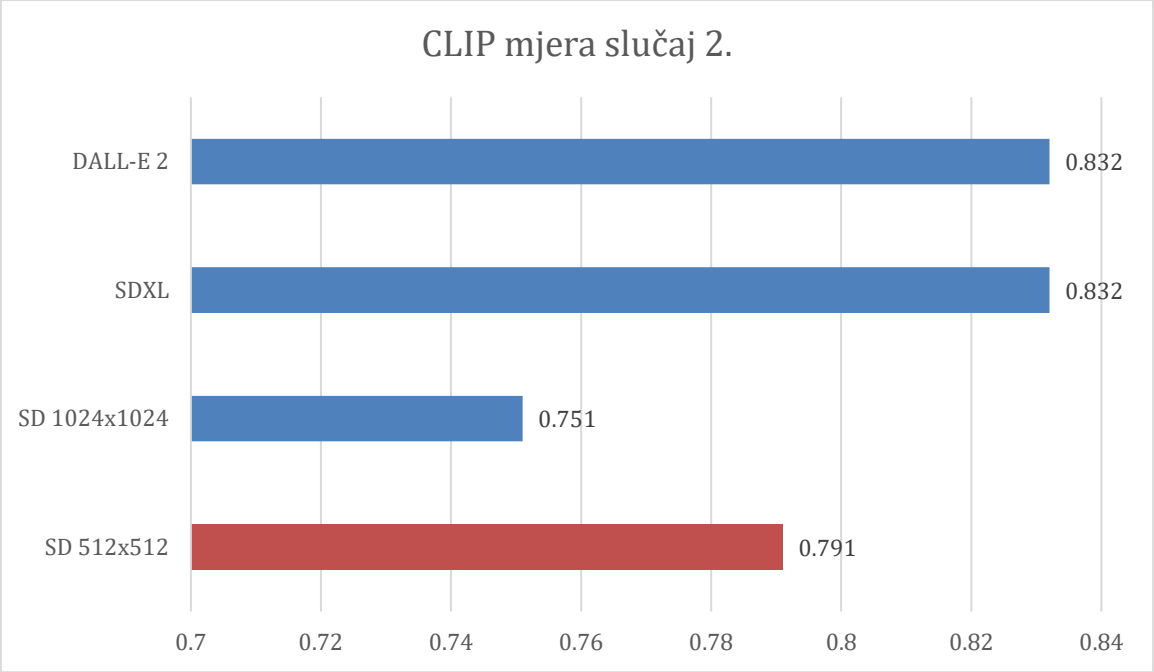
Viši rezultat CLIP mjere ukazuje da su slika i tekst više semantički povezani, dok niži rezultat ukazuje da nisu. Izračunata su dva skupa rezultata CLIP mjere. Prvi skup rezultata izračunat je za slučaj u kojem su evaluirane sve slike unutar skupova slika, dok je drugi skup rezultata za slučaj u kojem je evaluirana po jedna nasumično odabrana slika za svaki od opisa. Grafički prikaz prvog skupa rezultata je na slici 6.45.



Slika 6.45. CLIP mjera slučaj 1.
Izvor: Autor

Rezultati ukazuju da su u semantičkoj povezanosti sa opisima skup slika koje je generirao DALL-E 2 model i skup slika koje je generirao Stable Diffusion SDXL model gotovo jednaki.

Razlika je za 0.001 u korist DALL-E 2 (0.829 i 0.828). Skup slika Stable Diffusion 1.5 dimenzija 512x512 ima nešto niži rezultat (0.798), dok znatno manji CLIP rezultat ostvaruje skup slika Stable Diffusion 1.5 dimenzija 1024x1024. Na idućoj slici (6.46.) nalazi se grafički prikaz drugog skupa rezultata, za slučaj u kojem su nasumično odabrane generirane slike za svaki od opisa.



Slika 6.46. CLIP mjera slučaj 2.
Izvor: Autor

DALL-E 2 model i Stable Diffusion SDXL i u drugom slučaju ostvaruju gotovo jednake rezultate. Razlika u rezultatima iznosi 0.0004, i to u korist SDXL modela (0.8328 i 0.8324 za DALL-E 2). CLIP mjera za skupove slika spomenutih modela neznatno je veća u odnosu na slučaj 1. (razlika oko 0.004). Male su razlike u rezultatima i za Stable Diffusion 1.5 skupove slika. Skup slika 512x512 ostvaruje malo manji rezultat (0.791) u odnosu na slučaj 1., dok skup slika 1024x1024 ostvaruje neznatno veći rezultat (0.751) u odnosu na slučaj 1.

7. Zaključak

Modeli za stvaranje slika iz teksta snažan su alat koji pospješuje kreativnost te širem broju ljudi daje mogućnost izražavanja na načine koji im dosada nisu bili dostupni. Područje stvaranja slika iz prirodnog jezika koristeći modele dubokog učenja značajno napreduje, te su viđeni veliki pomaci u proteklim godinama. Napredovanjem modela razlikovanje stvarnog sadržaja pronađenog na internetu od onog generiranog modelima dubokog učenja postaje sve teže. Sve bitnijim se pokazuje sadržaj koji se dosada dijelio na internetu, jer je svaki objavljen sadržaj potencijalno iskorišten za treniranje neke vrste modela dubokog učenja, što stvara zabrinutosti vezane za privatnost, pa i autorska prava (u slučaju umjetnika čiji su radovi korišteni za treniranje modela). Modeli za generiranje slika imaju mnogo potencijalnih pozitivnih učinaka, ali je bitno da se društvo pravilno i pravovremeno prilagodi novim izazovima koje ova snažna tehnologija sa sobom donosi.

Opisani su neki od najznačajnijih modela, kao i njihovi zadaci vezani uz generiranje slika iz teksta. Spomenuti su i modeli za generiranje videozapisa iz teksta. Kroz tehnike kao što su generativne suparničke mreže, varijacijski autokoderi, transformeri i difuzijski modeli, istraživači stvaraju slike visoke kvalitete iz tekstualnih opisa uz sve veću preciznost i točnost. Uz razvoj modela za stvaranje videozapisa iz teksta, raste potencijal za kreiranje sadržaja pomoću dubokog učenja. Napredak u ovim područjima naglašava potencijal dubokog učenja da transformira način na koji stvaramo i koristimo vizualne medije.

Tri modela za stvaranje slika iz teksta, DALL-E 2, Stable Diffusion 1.5 i Stable Diffusion SDXL evaluirana su pomoću mjera za objektivnu evaluaciju ICS, FID i CLIP mjera. Za evaluaciju su Stable Diffusion 1.5 modelom generirane slike dimenzija 512x512, poput podatkovnog seta na kojem je model treniran, te 1024x1024. Modelom Stable Diffusion SDXL generirane su slike dimenzija 1024x1024. Set slika Stable Diffusion 1.5 dimenzija 1024x1024 ostvaruje veći ICS rezultat u odnosu na set slika dimenzija na kojima je model treniran (512x512), pa čak i u odnosu na set slika generiranih DALL-E 2 modelom, koji u spomenutoj mjeri ostvaruje najmanji rezultat. Prema subjektivnoj procjeni, SD set slika 1024x1024 pun je slika sa pogreškama, deformacijama i scenama koje ne izgledaju realistično. Unatoč tome, spomenuti set slika ostvaruje drugi najveći ICS rezultat, slabiji samo od rezultata najnovijeg Stable Diffusion SDXL modela. Relativno visok ICS rezultat SD 1024x1024 seta može se objasniti činjenicom da ICS mjeri raznolikost generiranih slika, pa set slika unatoč pogreškama uz dovoljnu raznolikost može ostvariti visok rezultat. Ostale objektivne mjere, FID i CLIP bolje reflektiraju subjektivnu procjenu kvaliteta slika. SD 1.5 set slika dimenzija 1024x1024 u spomenutim mjerama ostvaruje najlošiji rezultat. U CLIP mjeri SDXL se pokazuje podjednakim sa DALL-E 2 modelom, dok su Stable Diffusion 1.5 rezultati

zamjetno slabiji, posebice za slike dimenzija 1024x1024. Varijacije u pristupu računanja FID i CLIP vrijednosti ne uzrokuju bitne promjene u vrijednostima, ali uzrokuju promjene u redoslijedima. Rezultat FID mjere za prvi slučaj (izmjena veličine slika programom za uređivanje slika) manji je kod SDXL modela u odnosu na DALL-E 2 model, dok je za drugi slučaj (izmjena veličine slika programski) obrnuto. Isti modeli podjednaki su u oba slučaja CLIP mjere, no u prvom slučaju (evaluacija svih generiranih slika) DALL-E 2 ostvaruje neznatno veći rezultat, dok je u drugom slučaju (evaluacija nasumično odabranih slika) neznatno veći rezultat za SDXL model.

8. Literatura

[1] Zhang, Chenshuang, et al. Text-to-image diffusion model in generative ai: A survey. arXiv preprint arXiv:2303.07909, 2023.

[2] Mansimov, Elman, et al. Generating images from captions with attention. arXiv preprint arXiv:1511.02793, 2015.

[3] Ramesh, Aditya, et al. Zero-shot text-to-image generation. In: International Conference on Machine Learning. PMLR, str. 8821-8831, 2021.

[4] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X, Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (str. 1316-1324), 2018.

[5] Zhang, G., Ji, J., Zhang, Y., Yu, M., Jaakkola, T. S., & Chang, S. Towards Coherent Image Inpainting Using Denoising Diffusion Implicit Models., https://openreview.net/pdf?id=17YbAlc1tW, 2023.

[6] Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. Text-Guided Neural Image Inpainting. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, str. 1302–1310, 2020.

[7] Shi, K., Alrabeiah, M., & Chen, J. Progressive With Purpose: Guiding Progressive Inpainting DNNs Through Context and Structure, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10005125, 2023.

[8] Hanyu Xiang, Qin Zou, Muhammad Ali Nawaz, Xianfeng Huang, Fan Zhang, Hongkai Yu, Deep learning for image inpainting: A survey, https://www.sciencedirect.com/science/article/abs/pii/S003132032200526X, 2023

[9] Dogra, A., Goyal, B., Sharma, A., Kukreja, V., & Vig, R. Exploring image inpainting for seamless restitution, https://www.laserfocusworld.com/detectors-imaging/article/14295241/exploring-image-inpainting-for-seamless-restitution, 2023.

[10] Zheng, C., Cham, T. J., & Cai, J. Pluralistic image completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (str. 1438-1447), 2023.

[11] Zhang, Y., Zhang, K., Chen, Z., Li, Y., Timofte, R., Zhang, J., ... & Busch, C. NTIRE 2023 challenge on image super-resolution (x4): Methods and results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (str. 1864-1883), 2023.

[12] Saharia, Chitwan, et al. Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45.4: 4713-4726, 2022.

[13] Zamfir, E., Conde, M. V., & Timofte, R. Towards real-time 4k image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (str. 1522-1532), 2023.

[14] Dong, C., Loy, C.C., He, K., Tang, X. Learning a Deep Convolutional Network for Image Super-Resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8692. Springer, Cham, 2014.

[15] Kim, J., Lee, J. K., & Lee, K. M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (str. 1646-1654), 2016.
[16] Yu, L., Li, X., Li, Y., Jiang, T., Wu, Q., Fan, H., & Liu, S. DIPNet: Efficiency Distillation and Iterative Pruning for Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (str. 1692-1701), 2023.
[17] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. International Conference on Machine Learning, 2021.
[18] https://huggingface.co/tasks/unconditional-image-generation, dostupno 18.2.2023.
[19] Zhang, T., Fu, H., Zhao, Y., Cheng, J., Guo, M., Gu, Z., ... & Liu, J. SkrGAN: Sketching-rendering unconditional generative adversarial networks for medical image synthesis. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22 (str. 777-785). Springer International Publishing, 2019.
[20] Engel, J., Hoffman, M., & Roberts, A. Latent constraints: Learning to generate conditionally from unconditional generative models. arXiv preprint arXiv:1711.05772, 2017.
[21] Kang, Minguk; Park, Jaesik. Contragan: Contrastive learning for conditional image generation. Advances in Neural Information Processing Systems, 33: 21357-21369, 2020.
[22] Odena, Augustus; Olah, Christopher; Shlens, Jonathon. Conditional image synthesis with auxiliary classifier gans. In: International conference on machine learning. PMLR. str. 2642-2651, 2017.
[23] Mirza, M., & Osindero, S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.
[24] Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." Advances in Neural Information Processing Systems 34: 8780-8794, 2021.
[25] Ho, J., & Salimans, T. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022.
[26] https://sander.ai/2022/05/26/guidance.html#fn:cf, dostupno 31.8.2023.
[27] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., & Norouzi, M. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. ArXiv, abs/2205.11487, 2022.
[28] https://learnopencv.com/image-generation-using-diffusion-models/, dostupno 1.8.2023.
[29] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning (str. 2256-2265). PMLR, 2015.
[30] https://towardsdatascience.com/understanding-u-net-61276b10f360, dostupno 3.8.2023.
[31] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia, "Attention Is All You Need", Advances in Neural Information Processing Systems 30 (NIPS), 2017.
[32] https://medium.com/@geetkal67/attention-networks-a-simple-way-to-understand-self-attention-f5fb363c736d, dostupno 4.8.2023.

[33] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[34] Petrović, Darko. Transformer neuronske mreže za obradu prirodnog jezika. PhD Thesis. University North. University centre Varaždin. Department of Multimedia, Design and Application, 2022.
[35] https://github.com/CompVis/stable-diffusion, dostupno 16.2.2023.
[36] https://stats.stackexchange.com/questions/442352/what-is-a-latent-space, dostupno 5.8.2023.
[37] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (str. 10684-10695), 2022.
[38] https://vaclavkosar.com/ml/cross-attention-in-transformer-architecture, dostupno 5.8.2023.
[39] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., ... & Rombach, R. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv preprint arXiv:2307.01952, 2023.
[40] https://replicate.com/stability-ai/sdxl, dostupno 31.8.2023.
[41] https://sde-image-editing.github.io/, dostupno 31.8.2023.
[42] https://github.com/simonsanvil/DALL-E-Explained/blob/main/README.md, dostupno 20.2.2023.
[43] https://www.kdnuggets.com/2021/03/beginners-guide-clip-model.html, dostupno 10.8.2023.
[44] Liu, Anting; Zhang, Shichen. Taking Text to Image for Spin via DALL-E., https://www.researchgate.net/profile/Anting-Liu-3/publication/366485840_Taking_Text_to_Image_for_Spin_via_DALL-E/links/63a3756e5ed88950503f5341/Taking-Text-to-Image-for-Spin-via-DALL-E.pdf, 2023.
[45] https://openai.com/dall-e-2/, dostupno 16.02.2023.
[46] Ramesh, Aditya, et al. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1.2: 3, 2022.
[47] https://medium.com/augmented-startups/how-does-dall-e-2-work-e6d492a2667f, dostupno 10.8.2023.
[48] Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, David J. Fleet, "Video Diffusion Models",NeurIPS, https://arxiv.org/pdf/2204.03458v2.pdf, 2022
[49] Skorokhodov, Ivan et al. “StyleGAN-V: A Continuous Video Generator with the Price, Image Quality and Perks of StyleGAN2.” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 3616-3626, 2021.
[50] Singer, Uriel, et al. Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792, 2022.
[51] https://www.theverge.com/2022/9/29/23378210/meta-text-to-video-ai-generation-make-a-video-model-dall-e, dostupno: 19.2.2023.
[52] https://venturebeat.com/ai/google-announces-ai-advances-in-text-to-video-language-translation-more/, dostupno 19.2.2023.
[53] https://video-diffusion.github.io/, dostupno 10.8.2023.

[54] https://www.louisbouchard.ai/make-a-video/, dostupno 12.8.2023.
[55] Cho, J., Zala, A., & Bansal, M. DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Models. arXiv preprint arXiv:2202.04053, 2022.
[56] Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., ... & Lin, W. AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. arXiv preprint arXiv:2306.04717, 2023.
[57] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017.
[58] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. Improved techniques for training gans. Advances in neural information processing systems, 29, 2016.
[59] Barratt, S., & Sharma, R. A note on the inception score. arXiv preprint arXiv:1801.01973, 2018.
[60] Yu, Yu; Zhang, Weibin; Deng, Yun. Frechet inception distance (fid) for evaluating gans. China University of Mining Technology Beijing Graduate School: Beijing, China, 2021.
[61] https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score-372dff6a8c7a, dostupno 16.8.2023.
[62] https://www.techtarget.com/searchenterpriseai/definition/inception-score-IS, dostupno 17.8.2023.
[63] Dimitrakopoulos, Panagiotis; Sfikas, Giorgos; Nikou, Christophoros. Wind: Wasserstein inception distance for evaluating generative adversarial network performance. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. str. 3182-3186, 2020.
[64] Parmar, G., Zhang, R., & Zhu, J. Y. On aliased resizing and surprising subtleties in gan evaluation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (str. 11410-11420), 2022.
[65] https://unimatrixz.com/blog/latent-space-clip-score/#how-to-compute-clip-score, dostupno 17.8.2023.
[66] Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., & Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718, 2021.
[67] Ahmadi, S., & Agrawal, A. An Examination of the Robustness of Reference-Free Image Captioning Evaluation Metrics. arXiv preprint arXiv:2305.14998, 2023.
[68] https://stable-diffusion-ai.art/stable-diffusion-system-requirements/, dostupno 21.8.2023.
[69] https://github.com/AUTOMATIC1111/stable-diffusion-webui, dostupno 18.8.2023.
[70] https://github.com/invoke-ai/InvokeAI, dostupno 18.8.2023.
[71] https://github.com/abhishekkrthakur/diffuzers, dostupno 18.8.2023.
[72] https://huggingface.co/runwayml/stable-diffusion-v1-5, dostupno 18.8.2023.
[73] https://replicate.com/docs/how-does-replicate-work, dostupno 31.8.2023.
[74] https://help.openai.com/en/articles/6399305-how-dall-e-credits-work, dostupno 21.8.2023.
[75] https://www.datacamp.com/tutorial/how-to-run-stable-diffusion, dostupno 21.8.2023.
[76] https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Features, dostupno 21.8.2023.
[77] https://getimg.ai/guides/interactive-guide-to-stable-diffusion-guidance-scale-parameter, dostupno 21.8.2023.

[78] https://onceuponanalgorithm.org/guide-what-is-a-stable-diffusion-seed-and-how-to-use-it/, dostupno 21.8.2023.
[79] https://github.com/nnUyi/Inception-Score, dostupno 22.8.2023.
[80] https://uoa-eresearch.github.io/eresearch-cookbook/recipe/2014/11/20/conda/, dostupno 22.8.2023.
[81] https://github.com/mseitzer/pytorch-fid, dostupno 23.8.2023.
[82] Lin, Tsung-Yi, et al. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014. p. 740-755, 2014.
[83] https://cocodataset.org/#home, dostupno 25.8.2023.
[84] https://huggingface.co/blog/stable_diffusion, dostupno 28.8.2023.

Popis slika

Sveučilište
Sjever

SVEUČILIŠTE SJEVER

IZJAVA O AUTORSTVU

Završni/diplomski rad isključivo je autorsko djelo studenta koji je isti izradio te student odgovara za istinitost, izvornost i ispravnost teksta rada. U radu se ne smiju koristiti dijelovi tuđih radova (knjiga, članaka, doktorskih disertacija, magistarskih radova, izvora s interneta, i drugih izvora) bez navođenja izvora i autora navedenih radova. Svi dijelovi tuđih radova moraju biti pravilno navedeni i citirani. Dijelovi tuđih radova koji nisu pravilno citirani, smatraju se plagijatom, odnosno nezakonitim prisvajanjem tuđeg znanstvenog ili stručnoga rada. Sukladno navedenom studenti su dužni potpisati izjavu o autorstvu rada.

Ja, MATIJA KRAJAČIĆ *(ime i prezime)* pod punom moralnom, materijalnom i kaznenom odgovornošću, izjavljujem da sam isključivi autor/ica diplomskog *(obrisati nepotrebno)* rada pod naslovom STVARANJE SLIKA IZ PRIRODNOG JEZIKA KORIŠTEĆI MODELE DUBOKOG UČENJA *(upisati naslov)* te da u navedenom radu nisu na nedozvoljeni način (bez pravilnog citiranja) korišteni dijelovi tuđih radova.

Student/ica:
(upisati ime i prezime)
Matija Krajačić M-
(vlastoručni potpis)

Sukladno čl. 83. Zakonu o znanstvenoj djelatnost i visokom obrazovanju završne/diplomske radove sveučilišta su dužna trajno objaviti na javnoj internetskoj bazi sveučilišne knjižnice u sastavu sveučilišta te kopirati u javnu internetsku bazu završnih/diplomskih radova Nacionalne i sveučilišne knjižnice. Završni radovi istovrsnih umjetničkih studija koji se realiziraju kroz umjetnička ostvarenja objavljuju se na odgovarajući način.

Sukladno čl. 111. Zakona o autorskom pravu i srodnim pravima student se ne može protiviti da se njegov završni rad stvoren na bilo kojem studiju na visokom učilištu učini dostupnim javnosti na odgovarajućoj javnoj mrežnoj bazi sveučilišne knjižnice, knjižnice sastavnice sveučilišta, knjižnice veleučilišta ili visoke škole i/ili na javnoj mrežnoj bazi završnih radova Nacionalne i sveučilišne knjižnice, sukladno zakonu kojim se uređuje znanstvena i umjetnička djelatnost i visoko obrazovanje.