Implementacija softverskog rešenja za generisanje slika srpskih tradicionalnih jela analizom recepata i sastojaka

Bojan Mijanović, Vukašin Bogdanović, Jovan Jokić
Softversko inženjerstvo i informacione tehnologije
Univerzitet u Novom Sadu, Fakultet tehničkih nauka
Novi Sad, Srbija
mijanovic.r23.2024@uns.ac.rs bogdanovic.r212.2024@uns.ac.rs jokic.r219.2024@uns.ac.rs

I. Uvod

U eri digitalne transformacije, veštačka inteligencija (AI) i mašinsko učenje postaju ključni alati u različitim sferama ljudskog delovanja, uključujući i očuvanje kulturnog nasleđa. Gastronomija, kao fundamentalni deo kulturnog identiteta jednog naroda, doživljava novu renesansu kroz digitalizaciju. Generativni modeli, posebno u domenu sinteze slike iz teksta (eng. *text-to-image*), otvorili su revolucionarne mogućnosti za vizuelizaciju, edukaciju i promociju, omogućavajući kreiranje vizuelnog sadržaja iz prostog tekstualnog opisa.

Srpska tradicionalna kuhinja, sa svojom bogatom istorijom i raznovrsnošću, predstavlja značajan deo nacionalne baštine. Međutim, dok moderna kulinarska produkcija ima snažnu vizuelnu komponentu, ogroman deo tradicionalnog gastronomskog nasleđa, sačuvan u starim kuvarima, digitalizovanim arhivama i porodičnim beleškama, postoji isključivo u tekstualnoj formi. Ovi stariji recepti, prenošeni generacijama, često nemaju prateću sliku, što predstavlja ključnu prepreku za njihovu popularizaciju i očuvanje u digitalnom dobu. Bez vizuelne reprezentacije, korisnicima je teško da steknu pravu predstavu o krajnjem rezultatu jela, što smanjuje atraktivnost ovih recepata i otežava njihovo prenošenje novim generacijama koje su prevashodno vizuelno orijentisane.

Iako postoje moćni, opšte namenski generativni modeli trenirani na ogromnim skupovima podataka (npr. *Stable Diffusion*, *DALL-E*), oni često ne uspevaju da adekvatno generišu specifične, kulturološki nijansirane koncepte. Generisanje autentičnog prikaza lokalnih jela zahteva specifično poznavanje sastojaka i izgleda koji opšti modeli ne poseduju. Postoji jasna potreba za razvojem specijalizovanog rešenja koje može precizno interpretirati tekstualne opise (sastojke i korake pripreme) na srpskom jeziku i generisati fotorealistične prikaze koji odgovaraju tradicionalnoj srpskoj kuhinji.

Ovaj rad stoga predlaže razvoj i evaluaciju sistema koji transformiše tekstualni ulaz, specifično listu sastojaka i korake pripreme na srpskom jeziku u fotorealistični vizuelni izlaz koji predstavlja finalno jelo. Da bi se model obučio da razume semantičku vezu između teksta i slike u domenu srpske kuhinje, kreiraće se namenski skup podataka prikupljanjem recepata sa postojećih kulinarskih portala. Centralni deo istraživanja biće komparativna analiza različitih generativnih pristupa. Rad će uporediti performanse modela zasnovanih na arhitekturama kao što su kondicioni varijacioni autoenkoderi (cVAE) i kondicione generativne adversarijalne mreže (cGAN), treniranih na prikupljenom skupu podataka, sa rezultatima dobijenim finim podešavanjem (eng. *fine-tuning*) velikih, prethodno obučenih modela (*Stable Diffusion*) koristeći moderne i resursno efikasne tehnike poput LoRA (*Low-Rank Adaptation*).

II. Teorijske osnove

Ovo poglavlje pruža teorijski pregled generativnih modela i tehnika koje se koriste u ovom radu.

A. Varijacioni Autoenkoderi (VAE)

Varijacioni autoenkoderi (VAE), koje su uveli Kingma i Welling [1], su generativni modeli koji pripadaju klasi modela zasnovanih na verovatnoći. Osnovna ideja VAE je da nauči latentnu reprezentaciju (z) ulaznih podataka (x). Sastoje se od dve glavne komponente: enkodera $q(z|x)$ i dekodera $p(x|z)$.

VAE se trenira maksimizacijom donje granice verovatnoće (eng. *Evidence Lower Bound* - ELBO), što je ekvivalentno minimizaciji funkcije gubitka koja se sastoji od gubitka rekonstrukcije i Kullback-Leibler (KL) divergencije, koja deluje kao regularizator [1]. Funkcija gubitka (ELBO) za VAE je:

$$\mathcal{L}_{VAE} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x)||p(z)) \quad (1)$$

Gde prvi član predstavlja očekivani logaritam verovatnoće rekonstrukcije, a drugi član je KL divergencija.

B. Kondicioni Varijacioni Autoenkoderi (cVAE)

Kondicioni varijacioni autoenkoderi (cVAE) su proširenje VAE modela koje omogućava kontrolisano generisanje [2]. Ovo se postiže uslovljavanjem (eng. *conditioning*) kako enkodera tako i dekodera dodatnim informacijama c (npr. tekstualni embeding). Enkoder uči distribuciju $q(z|x, c)$, a dekoder $p(x|z, c)$. Funkcija gubitka se prilagođava da uključi ovaj uslov:

$$\mathcal{L}_{cVAE} = \mathbb{E}_{q(z|x,c)}[\log p(x|z,c)] - D_{KL}(q(z|x,c)||p(z|c)) \tag{2}$$

C. Generativne Adversarijalne Mreže (GAN)

Generativne adversarijalne mreže (GAN), koje je uveo Goodfellow et al. [3], predstavljaju pristup zasnovan na teoriji igara. GAN se sastoji od dve suprotstavljene neuronske mreže: Generatora (G) koji kreira podatke iz šuma z, i Diskriminatora (D) koji pokušava da razlikuje stvarne podatke (x) od lažnih ($G(z)$). Ove dve mreže igraju "minimax"igru definisanu ciljnom funkcijom:

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] \\ +\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]\end{aligned} \tag{3}$$

D. Kondicione Generativne Adversarijalne Mreže (cGAN)

Mirza i Osindero [4] su predložili kondicione generativne adversarijalne mreže (cGAN) kao proširenje koje omogućava kontrolu nad izlazom. I generatoru $G(z,c)$ i diskriminatoru $D(x,c)$ se prosleđuje dodatni uslov c. Ovo omogućava modelu da generiše slike koje odgovaraju specifičnom ulaznom tekstu, što je osnova za mnoge *text-to-image* modele pre pojave difuzije.

E. Difuzioni Modeli (Diffusion Models)

Difuzioni modeli [5], [6] su klasa generativnih modela koji postižu vrhunske rezultate (eng. *state-of-the-art*) u sintezi slika. Rade na principu dva procesa: (1) proces unapred, gde se Gausov šum postepeno dodaje slici x_0 kroz T koraka dok ne postane čisti šum x_T, i (2) proces unazad, gde se neuronska mreža (tipično U-Net) trenira da poništi ovaj proces, iterativno predviđajući i uklanjajući šum iz x_T da bi se rekonstruisala x_0. Generisanje je uslovljeno (npr. tekstom) ubacivanjem vektora uslova c u U-Net tokom procesa uklanjanja šuma.

F. Stable Diffusion

Standardni difuzioni modeli su računarski zahtevni jer rade u prostoru piksela. Rombach et al. [7] su predstavili *Latentne Difuzione Modele* (LDM), arhitekturu koja stoji iza *Stable Diffusion* modela. LDM primenjuje difuzioni proces ne u prostoru piksela, već u *latentnom prostoru* niže dimenzije. Slika se prvo kompresuje pomoću enkodera pre-treniranog VAE modela. Difuzija se vrši na ovoj latentnoj reprezentaciji, a zatim se rezultat dekodira nazad u prostor piksela pomoću VAE dekodera. Ovo drastično smanjuje računarsku složenost [7].

G. LoRA (Low-Rank Adaptation)

Fino podešavanje (*fine-tuning*) modela sa milijardama parametara, kao što je *Stable Diffusion*, je izuzetno skupo. LoRA (*Low-Rank Adaptation*) [8] je tehnika za efikasno fino podešavanje. LoRA zamrzava originalne težine modela (W_0) i ubacuje dve male, matrice za trening niskog ranga (A i B) u ključne slojeve (npr. *attention* slojeve). Tokom treninga, ažuriraju se samo parametri A i B, gde je $\Delta W = BA$. Pošto je $r \ll d$ (gde je r rang, a d dimenzija), broj parametara za treniranje je drastično smanjen [8].

H. FID (Fréchet Inception Distance)

Fréchet Inception Distance (FID) [9] je standardna metrika za evaluaciju kvaliteta i raznovrsnosti generisanih slika. FID meri sličnost između distribucije stvarnih slika (x) i distribucije generisanih slika (g). Obe grupe slika se propuštaju kroz pre-treniranu Inception-v3 mrežu da bi se dobili embedinzi (aktivacije). Zatim se te dve distribucije embedinga modeliraju kao multivarijantne Gausove distribucije, izračunavanjem njihovih srednjih vrednosti (μ) i kovarijansnih matrica (Σ). FID se izračunava kao Fréchet distanca između ove dve distribucije:

$$FID(x,g) = ||\mu_x - \mu_g||^2 + \text{Tr}(\Sigma_x + \Sigma_g - 2(\Sigma_x\Sigma_g)^{1/2}) \tag{4}$$

Niža FID vrednost ukazuje na veću sličnost između distribucija stvarnih i generisanih slika [9].

I. CLIPScore

CLIPScore je metrika za procenu semantičke usklađenosti između generisane slike i pripadajućeg tekstualnog opisa. Zasniva se na CLIP (*Contrastive Language–Image Pretraining*) modelu, koji zajednički uči reprezentacije teksta i slike mapirajući ih u isti latentni prostor [10].

Za datu sliku i tekstualni opis, CLIP enkoderi proizvode odgovarajuće vektorske reprezentacije koje se zatim upoređuju. CLIPScore se računa kao skalirana kosinusna sličnost između embeddinga slike i embeddinga teksta, pri čemu veće vrednosti ukazuju na jaču semantičku povezanost generisanog vizuelnog sadržaja i ulaznog opisa.

Za razliku od metrika koje mere samo vizuelni kvalitet ili statističku sličnost distribucija (npr. FID), CLIPScore direktno procenjuje koliko generisana slika odgovara značenju teksta, što ga čini posebno pogodnim za evaluaciju *text-to-image* modela.

J. CLIP Cosine Similarity

CLIP cosine similarity predstavlja osnovnu meru sličnosti između tekstualnog i slikovnog embeddinga dobijenih iz CLIP modela. Izračunava se kao kosinus ugla između dva vektora u zajedničkom latentnom prostoru:

$$\text{cos\_sim}(t,i) = \frac{t \cdot i}{\|t\|\|i\|} \tag{5}$$

gde su t i i embedding vektori teksta i slike, respektivno. Vrednosti kosinusne sličnosti se nalaze u intervalu $[-1,1]$, pri čemu veće vrednosti označavaju bolju semantičku usklađenost.

Ova metrika omogućava finiju analizu poravnanja teksta i slike i često se koristi zajedno sa CLIPScore-om kako bi se dobila stabilnija i interpretabilnija evaluacija generisanih rezultata.

III. Srodni radovi

Generisanje slika hrane na osnovu tekstualnog opisa predstavlja aktuelan izazov u domenu multimodalnog mašinskog učenja. Raniji pristupi, kao što je *ChefGAN* koji su predložili Pan et al. [11], uspešno su koristili kondicione generativne adversarijalne mreže (cGAN) za sintezu slika iz recepata, primenom kaskadnih modula za postizanje veće rezolucije i namenskih enkodera za tekst. Novija istraživanja se u velikoj meri oslanjaju na superiorne performanse difuzionih modela. Rad Ma et al. [12] predstavlja *MLA-Diff* model, koji koristi difuziju poboljšanu memorijskim modulima i unapređenim CLIP enkoderom za rešavanje *few-shot* problema generisanja na osnovu liste sastojaka. Pored samog generisanja, srodni radovi poput *Foodfusion* (Shi et al. [13]) bave se problemom kompozicije slika hrane, ali i, što je za ovaj rad posebno relevantno, rigoroznim metodologijama za kreiranje i preprocesiranje velikih skupova podataka, uključujući automatsku procenu i filtriranje slika lošeg kvaliteta. Ovaj rad se nadovezuje na ovu literaturu tako što direktno poredi klasične cGAN pristupe [11] sa modernim tehnikama finog podešavanja difuzionih modela (LoRA), koristeći CLIP kao centralni multimodalni enkoder po uzoru na [12], ali na novom, specifičnom skupu podataka tradicionalne srpske kuhinje.

IV. Skup podataka

Kvalitet i relevantnost skupa podataka su od presudnog značaja za uspešno treniranje generativnih modela, posebno u specifičnom domenu kao što je nacionalna kuhinja. S obzirom na to da ne postoji javno dostupan, anotiran skup podataka fokusiran na srpska tradicionalna jela pogodan za *text-to-image* zadatke, kreiran je sopstveni skup podataka.

A. Prikupljanje podataka

Podaci su prikupljeni sa popularnih domaćih kulinarskih portala *recepti.com* i *coolinarica.com*, koji poseduju bogatu bazu recepata koje su postavljali korisnici. Za potrebe automatskog preuzimanja podataka, razvijene su namenske skripte za struganje podataka (eng. *web scraping*) u programskom jeziku Python, koristeći biblioteke *BeautifulSoup* i *Selenium*.

Proces prikupljanja obuhvatio je preuzimanje slika jela, naziva recepata, liste sastojaka i tekstualnog opisa pripreme. Inicijalnim pokretanjem skripti prikupljen je sirovi skup podataka koji je brojao približno 30.000 unosa. Svi podaci su objedinjeni i strukturirani u jedinstveni JSON format, gde svaki objekat predstavlja instancu jela sa pratećim metapodacima.

B. Filtriranje i čišćenje podataka

Sirovi podaci prikupljeni sa interneta po prirodi sadrže veliku količinu šuma. Analizom inicijalnog skupa uočeni su sledeći problemi:

- Slike izuzetno niske rezolucije ili lošeg osvetljenja.
- Slike koje ne prikazuju hranu (npr. slike ambalaže, pribora ili lica).
- Generičke slike preuzete sa interneta (stock fotografije) koje ne odgovaraju receptu.
- Duplikati istih jela pod različitim nazivima.

Zbog toga je sproveden proces ručnog čišćenja i verifikacije. Cilj je bio zadržati samo one slike koje su vizuelno jasne, estetski prihvatljive i koje verodostojno predstavljaju koncept tradicionalne kuhinje. Nakon eliminacije neadekvatnih primera, veličina skupa podataka je redukovana sa 30.000 na finalnih 7.000 visokokvalitetnih parova slika i teksta. Iako je ovo značajno smanjenje kvantiteta, kvalitet podataka je drastično povećan, što je ključno za stabilnost treninga i vernost generisanih rezultata.

C. Preprocesiranje teksta i augmentacija

Sirovi tekstualni podaci (sastojci i priprema) zahtevali su značajnu obradu pre upotrebe u modelima. Prvi korak podrazumevao je čišćenje teksta od HTML tagova, suvišnih razmaka i nestandardnih karaktera.

S obzirom na to da su osnovni modeli korišćeni u ovom radu (kao što je *Stable Diffusion*) i njihovi tekstualni enkoderi (CLIP) primarno trenirani na engleskom jeziku, direktna upotreba srpskog jezika dovela bi do suboptimalnih rezultata. Takođe, originalni recepti su često predugački i sadrže narativne elemente nepotrebne za vizuelno generisanje.

Za rešavanje ovog problema korišćen je veliki jezički model GPT-4o (OpenAI). Implementiran je automatizovani *pipeline* u kojem se modelu prosleđuje originalni naziv, sastojci i opis na srpskom jeziku, uz sistemsku instrukciju (eng. *system prompt*) da izvrši dva zadatka:

1) Prevod ključnih vizuelnih elemenata na engleski jezik.
2) Sažimanje (eng. *summarization*) teksta u koncizan opis (eng. *caption*) koji je optimizovan za *text-to-image* modele (fokus na boje, teksture, oblike i serviranje, a ne na proces kuvanja).

D. Generisanje embedinga

Kao finalni korak pripreme podataka, tekstualni opisi dobijeni od GPT-4o modela konvertovani su u vektorski prostor (embedinge). Za ovu svrhu korišćen je CLIP (*Contrastive Language-Image Pre-Training*) model [14]. CLIP projektuje tekst i sliku u zajednički latentni prostor, omogućavajući modelu da razume semantičku povezanost između vizuelnih i tekstualnih podataka. Ovi pre-komputirani embedinzi su korišćeni kao uslovni ulaz (kondicioniranje) tokom treniranja generativnih modela opisanih u sekciji Metodologija.

V. Metodologija

Ovo poglavlje opisuje tehničku implementaciju predloženog sistema. Proces se sastoji od tri ključne faze: (1) kodiranje tekstualnih opisa u vektorski prostor, (2) treniranje osnovnih generativnih modela (cVAE i cGAN) od nule na prikupljenom skupu podataka, i (3) fino podešavanje (eng. *fine-tuning*) savremenog difuzionog modela korišćenjem tehnike LoRA.

A. Pregled arhitekture sistema

Ulaz u sistem predstavljaju preprocesirani tekstualni opisi jela na engleskom jeziku (dobijeni metodom opisanom u poglavlju 4). Ovi opisi se prvo propuštaju kroz pre-trenirani

enkoder teksta kako bi se dobila bogata semantička reprezentacija (embeding). Dobijeni vektori služe kao uslov (kondicioniranje) za generativne modele koji na izlazu produku finalnu sliku jela.

B. Kodiranje teksta i kondicioniranje

Za razumevanje semantike teksta korišćen je *CLIP (Contrastive Language-Image Pre-Training)* model [14]. Konkretno, korišćena je varijanta ViT-L/14 koja tekstualni opis preslikava u vektor fiksne dužine od 768 dimenzija. Ovaj vektor se koristi kao uslovni ulaz c u svim eksperimentima. S obzirom na to da je CLIP treniran na stotinama miliona parova slika-tekst, on omogućava modelima da razumeju koncepte (npr. črvena boja", "tanjir", "čorba") mnogo bolje nego što bi to bilo moguće učenjem samo na našem ograničenom skupu podataka.

C. Implementacija osnovnih modela

U prvoj fazi eksperimenta, implementirani su i trenirani cVAE i cGAN modeli "od nule"(eng. *from scratch*) na prikupljenom skupu podataka. Cilj je bio uspostaviti osnovnu liniju performansi (eng. *baseline*). Slike su za potrebe ovih modela skalirane na rezoluciju od 128x128 piksela.

1) Arhitektura cVAE: Za implementaciju kondicionog varijacionog autoenkodera je korišćena konvuoluciona arhitektura zasnovana na CLIP embedinzima kao uslovnom signalu. Model se sastoji iz enkodera i dekodera, i adaptirani su za generisanje slika dimenzija 64x64. U daljem tekstu ćemo se osvrnuti na komponente *cVAE* modela:

a) Enkoder: Arhitektura enkodera ima za cilj da transformiše ulazne slike u latentni prostor zadate dimenzionalnosti.*CLIP* embedinzi se prostorno repliciraju i konkateniraju sa ulaznom slikom duž kanalne dimenzije. Enkoder se sastoji od četiri konvoluciona bloka. Nakon svakog konvolucionog bloka izuzev prvog, prisutna je *batch* normalizacija i kao i deaktivacija određenog procenta neurona. Korišćena je ReLU aktivacija. Dva linearna sloja generišu parametre latentne distribucije (i log ²). Uzorkovanje iz latentnog prostora vrši se pomoću reparametrizacionog trika: z = + · .

b) Dekoder: Ulaz u dekoder predstavlja latentni vektor koji se konkatenira sa *CLIP* embedingom. Dekoder se sastoji iz četiri konvoluciona transponovana bloka koji uvećavaju prostornu dimenzionalnost. Ovde se, kao i u enkoderu, koriste ReLU aktivacija i *batch* normalizacija. Izuzetak predstavlja finalni sloj koji koristi Tanh aktivaciju. Trening je izvršen kroz 35 epoha, dok funkcija gubitka kombinuje *MSE* (*eng. Mean Squared Error*) i Kullback-Leibler divergenciju

2) Arhitektura cGAN: Za implementaciju kondicionog GAN-a korišćena je arhitektura zasnovana na rezidualnim blokovima sa CLIP embedinzima kao uslovom, uz primenu naprednih tehnika stabilizacije.

a) Generator: Sastoji se od niza ResNet blokova sa uzorkovanjem naviše (4×4 $\rightarrow$ 128×128). Ključna komponenta je Conditional Batch Normalization (CondBN) gde se parametri γ i β generišu direktno iz 768-dimenzionalnih tekstualnih embedinga. Koristi se ReLU aktivacija u skrivenim slojevima i Tanh na izlazu.

b) Diskriminator: Implementiran je kao *Projection Discriminator* sa ResNet blokovima. Za obezbeđivanje Lipschitz-ovog ograničenja, ključnog za stabilnost GAN treninga, primenjena je spektralna normalizacija (eng. *Spectral Normalization*) na svim slojevima.

c) Trening: Koristi Hinge Loss funkciju i Adam optimizator ($\beta_1 = 0.0$, $\beta_2 = 0.9$). Radi sprečavanja preprilagođavanja diskriminatora na malom skupu podataka, primenjena je diferencijabilna augmentacija (DiffAug) koja uključuje nasumične promene boja i geometrijske transformacije tokom samog treninga. Dodatno, koristi se R1 *gradient penalty* i eksponencijalni pokretni prosek (EMA) težina generatora za stabilniju inferenciju.

D. Fino podešavanje latentnog difuzionog modela

Zbog visoke računske zahtevnosti potpunog treniranja difuzionih modela, u ovom radu je primenjen pristup efikasnog finog podešavanja (eng. *Parameter-Efficient Fine-Tuning* - PEFT) korišćenjem tehnike LoRA na *Stable Diffusion v1.5* modelu.

Umesto ažuriranja svih težina, LoRA ubacuje trenabilne matrice niskog ranga u slojeve unakrsne pažnje (eng. *cross-attention*). Definisana su dva eksperimentalna okvira kako bi se ispitala uloga semantičkog razumevanja:

- UNet-only: Fino podešavanje isključivo vizuelnog dela modela (U-Net), dok tekstualni enkoder ostaje zamrznut.
- UNet + Text Encoder:e Istovremeno ažuriranje vizuelnog modela i CLIP tekstualnog enkodera, kako bi se model bolje prilagodio specifičnoj terminologiji srpske kuhinje.

Ovaj pristup omogućava komparativnu analizu između stabilnosti vizuelnog generisanja i fleksibilnosti semantičkog razumevanja.

E. Eksperimentalno okruženje

Treniranje je sprovedeno u hibridnom hardverskom okruženju. Modeli trenirani od nule (cVAE, cGAN) razvijani su na lokalnoj radnoj stanici sa NVIDIA GeForce RTX 5060 Ti grafičkom karticom. Za fino podešavanje *Stable Diffusion* modela korišćeno je *cloud* okruženje (Google Colab) sa NVIDIA T4 GPU-om, radi optimizacije korišćenjem *xFormers* biblioteke i efikasnije manipulacije memorijom. Implementacija se oslanja na *PyTorch* i *Diffusers* biblioteke.

Hiperparametri treninga za cVAE model su bili sledeći:

- Trajanje treninga: 35 epoha sa priodom čuvanja kontrolnih tačaka na svakih 5 epoha.
- Veličina paketa (eng. *batch size*): 64.
- Stopa učenja (eng. *learning rate*): 2×10^{-4}, sa AdamW optimizatorom i weight decay faktorom od 1×10^{-5} za regularizaciju .
- Optimizator: Korišćen je AdamW optimizator sa parametrima $\beta_1 = 0.9$ i $\beta_2 = 0.999$.
- Regularizacija: Korišćeno je sečenje gradijenta (eng. *gradient clipping*) sa normom 1.0 kao i deaktivacija neurona sa stopom 1.0 u enkoderu.
- Beta parametar: $\beta = 0.5$ za ponderisanje *KL* divergencije u funkciji gubitka.

- *KL* zagrevanje: Primenjeno je linearno zagrevanje u periodu od 10 epoha za postepenu aktivaciju *KL* divergencije u cilju sprečavanja kolapsa latentnog prostora u ranim fazama treninga.
- CLIP kondicioniranje: Korišćeni su 512-dimenzianalnih iz ViT-B/32 .

Hiperparametri treninga za cGAN model su bili sledeći:

- Trajanje treninga: 40,000 iteracija.
- Veličina paketa (eng. *batch size*): 4. Ova vrednost je odabrana zbog optimizacije za memorijska ograničenja RTX 5060 Ti GPU-a.
- Stopa učenja (eng. *learning rate*): Za generator je korišćena konzervativna stopa od 5×10^{-5}, dok je za diskriminator korišćena nešto veća stopa od 1.5×10^{-4} radi održavanja stabilnosti treninga.
- Optimizator: Korišćen je Adam optimizator sa parametrima $\beta_1 = 0.0$ i $\beta_2 = 0.9$.
- Regularizacija: Primenjen je R1 penal nad gradijentima ($\lambda = 2.0$) radi stabilizacije diskriminatora, kao i sečenje gradijenta (eng. *gradient clipping*) sa normom 1.0.
- EMA (Exponential Moving Average): Korišćen je faktor opadanja od 0.999 za težine generatora tokom inferencije, što je omogućilo stabilniju generaciju.
- Kapacitet modela: Osnovni broj kanala (eng. *base channels*) je postavljen na 32 (umanjeno sa standardnih 64) kako bi se zadovoljila memorijska ograničenja hardvera.
- CLIP kondicioniranje: Korišćeni su 768-dimenzionalni embedinzi iz ViT-L/14 modela umesto standardnih 512-dimenzianalnih iz ViT-B/32 za bogatiju semantičku reprezentaciju.

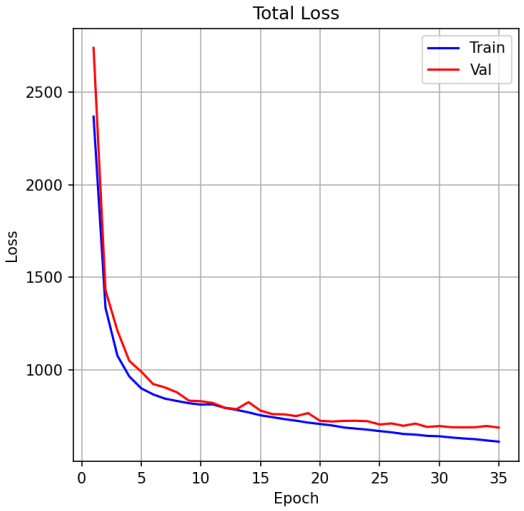
Za fino podešavanje difuzionog modela korišćeni su sledeći parametri:

- Osnovni model: Stable Diffusion v1.5.
- Trajanje treninga: 200-1000 iteracija.
- Stopa učenja: Konstantna stopa od 1×10^{-4}.
- Rezolucija: Slike su redimezionirane na 512×512 piksela.
- Optimizacija memorije: Korišćena je *mixed-precision* (fp16) tehnika kako bi se smanjilo zauzeće memorije tokom treninga.

VI. Rezultati i diskusija

A. cVAE rezultati

Proces treniranja cVAE modela kroz 35 epoha pokazuje uspešnu konvergenciju i stabilnost varijacionog učenja. Ukupan gubitak (eng. *total loss*) na početku treninga iznosi približno 2700 i rapidno opada tokom prvih 10 epoha, dostigavši vrednost oko 800. Nakon toga, model nastavlja da se postepeno poboljšava, sa finalnim vrednostima od približno 600 za trening skup i 680 za validacioni skup. Evaluacija modela na test skupu od 500 slika pokazuje sledeće performanse: FID (eng. *Fréchet Inception Distance*) skor od 204.09 ukazuje na značajnu razliku u distribuciji između generisanih i realnih slika. CLIP skor od 0.1858 ± 0.0243 meri semantičku usklađenost generisanih slika sa tekstualnim opisima, dok CLIP kosinusna sličnost od 0.1695 ± 0.0352 pokazuje umerenu korelaciju između ulaznih tekstualnih embedinga i generisanih slika. Ove metrike ukazuju da model uspešno uči uslovljenu generaciju, ali sa prostorom za dalja poboljšanja



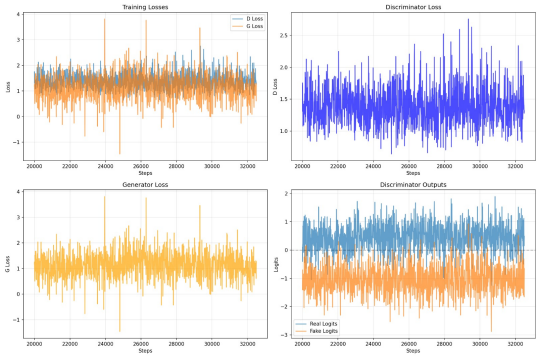
Slika 1: Funkcija gubitka tokom treniranja cVAE kroz 35 epoha

Slika 2: Primeri slika srpskih jela nastalih korišćenjem cVAE modela.

B. cGAN rezultati

Proces treniranja cGAN modela kroz 40.000 iteracija pokazuje uspešnu adversarijalnu stabilizaciju. Generator loss se kreće u opsegu 0.5-2.0, dok diskriminator loss osciluje oko 1.0-1.5, što ukazuje na zdravu konkurenciju između komponenti. Ključni faktor za uspeh bila je upotreba bogatijih, 768-dimenzionalnih CLIP embeddinga (ViT-L/14) umesto standardnih 512-dimenzionalnih, što je omogućilo preciznije semantičko uslovljavanje.

Model dostiže stabilnu konvergenciju oko 32.000 koraka bez znakova kolapsa moda (eng. *mode collapse*), zahvaljujući primeni spektralne normalizacije i konzervativnih stopa učenja.



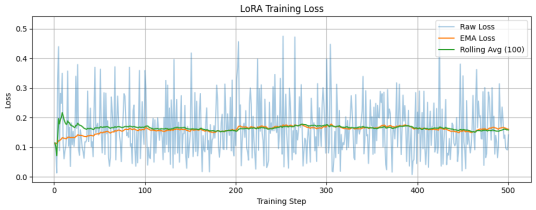
Slika 3: Dinamika treninga cGAN modela kroz finalne korake (32k iteracija). Prikazani su generator loss, discriminator loss i discriminator outputs (real/fake logits).

Slika 4: Primeri generisanih slika srpskih jela i slika je korišćenjem cGAN modela. Grid prikazuje raznovrsnost generisanih jela uključujući sarmu, ćevape, gulaš, tradicionalne čorbe i pečena mesa.

C. LoRA rezultati

LoRA fine-tuning Stable Diffusion v1.5 modela evaluiran je kroz različite korake treninga, poredeći dve strategije: treniranje samo UNet-a i zajedničko treniranje UNet-a i tekstualnog enkodera. Svi osnovni eksperimenti sprovedeni su na skupu od približno 1.000 slika, koji se pokazao kao optimalan balans između učenja domenski specifičnih vizuelnih karakteristika i očuvanja sposobnosti generalizacije.

Dinamika treninga prikazana na slici 5 pokazuje stabilnu i brzu konvergenciju, pri čemu se većina korisnih reprezentacija uči u prvih 200–300 koraka. Odsustvo oscilacija u EMA i rolling loss krivama ukazuje na stabilan proces optimizacije i adekvatno izabrane hiperparametre.



Slika 5: Dinamika treninga LoRA modela. Prikazani su raw loss, EMA loss i rolling average kroz 500+ koraka treninga, demonstrirajući stabilnu konvergenciju.

Rezultati prikazani u Tabeli I ukazuju da zajedničko treniranje UNet-a i tekstualnog enkodera ostvaruje najbolje kvantitativne performanse u ranoj fazi treninga (FID = 155.17 na 200 koraka), što se može pripisati boljem semantičkom poravnanju tekstualnih opisa tradicionalnih jela. Međutim, produženo treniranje ove konfiguracije dovodi do pogoršanja metrika, što ukazuje na pojavu preprilagođavanja (*overfitting*) usled ograničene veličine skupa podataka.

Nasuprot tome, treniranje samo UNet LoRA adaptera pokazuje veću stabilnost kroz veći broj koraka, uz manje varijacije FID i CLIP metrika. Ovi rezultati sugerišu da je UNet-only pristup robusniji izbor za manje skupove podataka, dok treniranje tekstualnog enkodera zahteva rano zaustavljanje.

Tabela I: Rezultati LoRA finog podešavanja

Konfiguracija	Korak	FID ↓	CLIPScore ↑	CLIP cos sim ↑
UNet only	200	165.43	64.52	0.2904
UNet only	400	158.30	64.45	0.2890
UNet + Text Encoder	200	**155.17**	**64.68**	**0.2936**
UNet + Text Encoder	400	164.53	64.42	0.2884

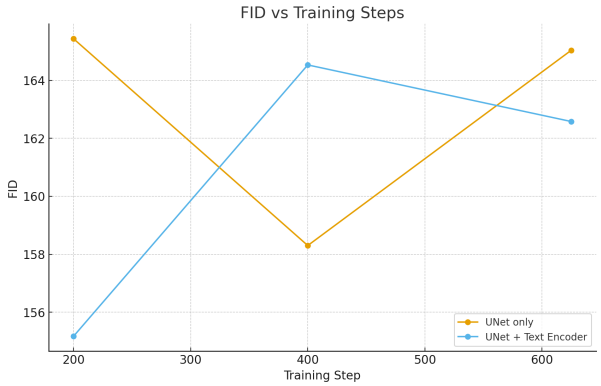
U dodatnom eksperimentu sa proširenim skupom od približno 7.000 slika i 1.000 koraka treninga postignut je znatno niži FID (113.24). Iako numerički superioran, ovaj rezultat je praćen izraženim *overfitting*-om: generisane slike pokazuju visoku sličnost sa trening primerima i smanjenu varijaciju. Ovaj nalaz potvrđuje da se FID metrika mora interpretirati u kombinaciji sa kvalitativnom analizom.

Evolucija FID metrike na slici 6 potvrđuje da konfiguracija UNet + Text Encoder postiže optimalne performanse u ranoj fazi treninga, dok dalje treniranje ne donosi dodatna poboljšanja. Vizuelno poređenje prikazano na slici 7 dodatno ilustruje prednost LoRA fino podešenog modela u odnosu na bazni Stable Diffusion model u kontekstu razumevanja tradicionalnih srpskih jela.

D. Komparativna analiza i diskusija

Poređenje cVAE, cGAN i LoRA pristupa otkriva jasne kompromise između resursa, kontrole i kvaliteta:

a) Kvalitet slike i realizam: LoRA je ubedljivo superiorna, generišući slike visoke rezolucije (512x512) sa realističnim teksturama, zahvaljujući transferu znanja sa LAION-5B skupa. cGAN i cVAE, ograničeni treniranjem od nule na malom skupu, bore se sa generisanjem finih detalja i često proizvode geometrijske deformacije.



Slika 6: Evolucija FID metrike kroz korake treninga za različite LoRA konfiguracije. UNet + Text Encoder konfiguracija postiže najbolje performanse na 200 koraka.

(a) Bazni Stable Diffusion (b) LoRA fine-tuned

Slika 7: Poređenje generisanja sarme: bazni model vs LoRA fine-tuned model. LoRA model pokazuje bolje razumevanje srpskog jela i tradicinalnog načina serviranja.

b) Stabilnost treninga: cVAE se pokazao kao najstabilniji, ali po cenu oštrine slike. cGAN zahteva pažljivo podešavanje hiperparametara (R1 penalty, learning rate) kako bi se izbegao kolaps moda, ali nudi bolju oštrinu od cVAE. LoRA je stabilna i brzo konvergira, ali nosi rizik od katastrofalnog zaboravljanja ili overfittinga ako se predugo trenira na malom skupu.

c) Efikasnost resursa: Iako deluje kontraintuitivno, LoRA je najefikasnija za postizanje visokog kvaliteta. Da bi cGAN dostigao sličan nivo realizma, zahtevao bi eksponencijalno više podataka i vremena za trening. LoRA omogućava demokratizaciju generativnih modela, postižući vrhunske rezultate na potrošačkom hardveru.

Zaključno, za specifične domene sa ograničenim podacima poput nacionalne kuhinje, adaptacija velikih modela (LoRA) predstavlja optimalnu strategiju, dok su arhitekture trenirane od nule (cGAN, cVAE) primerenije za scenarije gde je dostupna ogromna količina specifičnih podataka ili gde su hardverski resursi za inferenciju ekstremno ograničeni.

VII. Zaključak

Ovaj rad je predstavio sveobuhvatnu studiju o primeni generativnih modela veštačke inteligencije za digitalno očuvanje srpskog kulinarskog nasleđa. Temeljni doprinos istraživanja predstavlja razvoj prvog kuriranog multimodalnog skupa podataka od 7.000 parova slika i teksta. Analiza je pokazala da je semantički kvalitet opisa, postignut integracijom GPT i CLIP modela, bio presudan za uspešno uslovljavanje generativnih procesa.

Rezultati nedvosmisleno potvrđuju primat *transfer learning* pristupa kroz LoRA fino podešavanje. Međutim, studija je otkrila važne nijanse u procesu treniranja: dok je uključivanje tekstualnog enkodera donelo inicijalna poboljšanja, dugotrajno treniranje na ograničenom skupu dovelo je do gubitka generalizacije. Posebno je značajan nalaz da puko povećanje skupa podataka i produženo treniranje mogu dovesti do "memorisanja"uzoraka, gde numerički bolji FID skor prikriva kvalitativni pad u raznovrsnosti generisanih jela. Ovo naglašava potrebu za balansom između adaptacije modela i očuvanja njegovog prethodnog znanja.

S druge strane, arhitekture trenirane od nule (cGAN i cVAE) su, uprkos nižoj vizuelnoj vernosti, demonstrirale važnost arhitektonskih izbora. Implementacija 768-dimenzionalnih CLIP embeddinga u cGAN modelu pokazala se kao ključna za stabilizaciju adversarijalnog treninga i precizno semantičko mapiranje, nudeći putokaz za buduće lake modele specifične namene.

Zaključno, dok LoRA nudi trenutno najviši kvalitet za vizuelizaciju kulturne baštine, dugoročno rešenje leži u hibridnom pristupu: korišćenju moćnih pre-treniranih osnova uz pažljivo kurirane, manje skupove podataka i rigoroznu kvalitativnu, a ne samo kvantitativnu evaluaciju.

LITERATURA

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014. [Online]. Available: https://arxiv.org/abs/1312.6114

[2] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015)*, 2015.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS 2014)*, 2014.

[4] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014. [Online]. Available: https://arxiv.org/abs/1411.1784

[5] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, 2015.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS 2020)*, 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://arxiv.org/abs/2106.09685

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[11] S. Pan, L. Dai, X. Hou, H. Li, and B. Sheng, "Chefgan: Food image generation from recipes," in *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. Association for Computing Machinery (ACM), 2020, p. 4244–4252.

[12] J. Ma, Y. Wan, and Z. Ma, "Memory-based learning and fusion attention for few-shot food image generation method," *Applied Sciences*, vol. 14, no. 18, 2024.

[13] C. Shi, X. Wang, S. Shi, X. Wang, M. Zhu, N. Wang, and X. Gao, "Foodfusion: A novel approach for food image composition via diffusion models," 2024.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.