Načrtovanje jezikovne komponente generativnega nasprotniškega nevronskega omrežja

Martin Pernuš, Simon Dobrišek

Fakulteta za elektrotehniko
Univerza v Ljubljani
Tržaška cesta 25, 1000 Ljubljana
{martin.pernus, simon.dobrisek}@fe.uni-lj.si

Povzetek

Področje ustvarjanja slik je kot pomembno področje umetne inteligence v zadnjih letih dobilo velik zagon. Vizualno prepričljivo ustvarjanje slik omogočajo generativni modeli, ki lahko sliko pogojijo tudi z jezikovnim opisom. Model, ki trenutno vrača najboljše rezultate na tem področju je pogojeno generativno nasprotniško omrežje. V tem članku opišemo uspešnejše postopke modeliranja tekstovnega opisa za generativna nasprotniška omrežja, način učenja takih modelov, podatkovne zbirke in metrike za vrednotenje.

Abstract

The field of image generation is an important area of artificial intelligence research that gained a lot of traction in the recent years. The most visually promising results are achieved by generative models that can be conditioned on language description. The predominant model in the generative modelling area is conditional generative adversarial network. In this paper we describe the most successful attempts on the combination of language modelling and generative adversarial networks, their training procedures, datasets and evaluation metrics.

1. Uvod

Področje računalniškega vida in umetne inteligence je v zadnjih letih doseglo velike uspehe na področju ustvarjanja slik. V ozadju teh rezultatov so generativni modeli globokih nevronskih omrežij (v nadaljevanju generativni modeli). Generativni modeli z modeliranjem ciljne verjetnostne porazdelitve podatkov omogočajo ustvarjanje novih vzorcev. V osnovi so generativni modeli nepogojeni, če pa izhod generativnega modela pogojimo z zunanjim signalom, pa govorimo o pogojenih generativnih modelih.

Danes najbolj uporabljeni generativni modeli so avtoregresivni modeli, pretočni modeli, variacijski avtokodirniki in generativni nasprotniški modeli (angl. Generative Adversarial Network, GAN). Slednji še posebej izstopajo, saj so zmožni ustvarjanja fotorealističnih in vizualno prepričljivih slik različnih objektov, kompleksnih scen in visokoresolucijskih slik, ki jih praktično ni več mogoče razlikovati od resničnih slik. Prav tako so se modeli GAN izkazali kot modeli, ki jim zlahka vnesemo zunanji signal kot pogoj za ustvarjanje ciljne slike z želeno vizualno semantično informacijo. Zunanji signali so lahko različne oblike, lahko je preprost, kot npr. razred izhodne slike, lahko pa tudi bolj kompleksen, kot npr. podroben opis želene izhodne slike. Slednji način obravnavamo v našem članku.

Glavni namen članka je pregled najnovejših metod tistih komponent generativnih modelov, ki skrbijo za opis informacije v jezikovni obliki. V članku se bomo osredotočili na tiste komponente, ki se uporabljajo v modelih GAN, saj ti prevladujejo na področju pogojenih generativnih modelov. Na koncu predlagamo še nekaj možnih rešitev za izboljšanje teh modelov.

2. Sorodna literatura

Generativni nasprotniški modeli so bili prvič opisani v članku (Goodfellow et al., 2014), kjer so se avtorji osredotočali na modeliranje obraznih slik brez zunanjih signalov. Nadgradnje prvotne zasnove modela so se v literaturi osredotočale predvsem na spremembo arhitektur modelov in spreminjanje prvotne kriterijske funkcije. V članku (Radford et al., 2015) so predlagali konvolucijsko nevronsko omrežje DCGAN, kar je omogočalo zmanjšanje števila parametrov v celotni arhitekturi ob hkratni izboljšani kvaliteti ustvarjenih slik. S postopnim načinom učenja modela ProGAN (Karras et al., 2018) so prvič ustvarili slike s kar milijon slikovnimi elementi. Model je bil nadgrajen v (Karras et al., 2019a) in (Karras et al., 2019b), kjer je model StyleGAN z vnosom stohastične raznolikosti in napredne arhitekture ustvaril slike obrazov, ki jih praktično ni več mogoče razlikovati od slik resničnih ljudi. Ti modeli so nepogojeni, saj gre za neposredno modeliranje ciljne verjetnostne porazdelitve brez kakršnekoli dodatne informacije in lastnostih na končni sliki. Pri spremembi kriterijske funkcije je bil napredek narejen na področju nenasičenih kriterijskih funkcij, kar v praksi omogoča stabilnejše učenje (Arjovsky et al., 2017), (Mao et al., 2017), (Nowozin et al., 2016), (Gulrajani et al., 2017).

Poleg nepogojenih GAN modelov so se razvili tudi pogojeni modeli GAN, ki izhodno sliko pogojujejo glede na vhodni signal. Model Pix2pix (Isola et al., 2017) je ciljni slog slike oblikoval glede na vhodno sliko. Model GauGAN (Park et al., 2019) je sliko pogojil z visokoresolucijsko sliko, ki lokacijsko opiše želeno semantično informacijo. BigGAN (Brock et al., 2019) je trenutno najboljši pogojeni generativni model, ki ciljno sliko ustvari le glede na informacijo o želenem razredu. Model GeNeVA-GAN (El-Nouby et al., 2019) je tekstovne opise uporabil za postopno dodajanje preprostih geometrijskih objektov na sliko, kot so kocka, krogle in piramide.

Pogojeni modeli, ki so v našem članku najbolj relevantni, pogojujejo zunanjo sliko glede na tekstovni opis. Prvi

večji preboj na tem področju so naredili v članku (Reed et al., 2016b), kjer so ustvarili slike velikosti 64×64. Nadaljnje delo se je osredotočalo na izboljšanje vizualne prepričljivosti slik in višanje resolucije, kot npr. StackGAN (Zhang et al., 2017), AttnGAN (Xu et al., 2018) in MirrorGAN (Qiao et al., 2019), ki bodo podrobneje opisani v 4. poglavju.

3. Ozadje

3.1. Generativna nasprotniška omrežja (GAN)

Omrežje GAN sestoji iz dveh modelov: generator G in diskriminator D. Celotno omrežje implicitno optimiziramo na tak način, da bo generator G modeliral ciljno verjetnostno porazdelitev (v našem primeru verjetnostno porazdelitev slik) čim boljše. Naloga diskriminatorja D je ločevanje med resničnimi slikami in umetnimi slikami, torej tistimi, ki jih je ustvaril generator. Naloga generatorja G je prelisičiti diskriminator tako, da bo ta domneval, da gre za resnični vzorec. Tako G kot D sta običajno visoko parametrizirani nelinearni parametrični funkciji, ki ju iterativno učimo s postopkom gradientnega spusta. V dani iteraciji bo funkcija G vrnila nov vzorec, funkcija D pa bo ovrednotila verjetnost, da je nek vzorec resničen. Z učenjem bo generator ustvarjal vse boljše slike, diskriminator pa vse bolje ločeval med resničnimi in umetnimi slikami. Po končanem učenju diskriminator zavržemo, generator pa uporabimo za ustvarjanje novih vzorcev.

Matematično lahko učni postopek zapišemo kot igro med dvema igralcema s kriterijsko funkcijo

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

kjer je p_{data} verjetnostna porazdelitev resničnih podatkov, z pa šumni vektor, vzorčen iz preproste verjetnostne porazdelitve p_z (običajno Gaussova porazdelitev, $z \sim \mathcal{N}(0, 1)$). V praksi se izkaže, da zaradi izginjajoče odvodne informacije pri gradientnem spustu ob predobro naučenem diskriminatorju generator boljše učimo z maksimizacijo $\log(D(G(z))$ namesto minimizacije $\log(1 - D(G(z))$.

3.2. Enodimenzionalna konvolucija

Konvolucija je pogosta operacija v globokih nevronskih omrežjih. Pretežno se uporablja na področju računalniškega vida, kjer uporabljamo dvodimenzionalno konvolucijo. Če za modeliranja teksta uporabljamo konvolucijo, se poslužimo enodimenzionalne različice. Definirana je kot

$$h(y) = \sum_{x=1}^{k} f(x)g(y \cdot d - x + c), \tag{2}$$

kjer sta f in g funkciji z diskretnih vhodom, k velikost jedra, d korak, $c = k - d + 1$ pa odmik. Kot je v nevronskih omrežjih običajno, ima celotno konvolucijsko omrežje množico funkcij $f_{ij}(x)$, imenovanih uteži, na množici vhodov $g_i(x)$ in izhodov $h_j(y)$.

Podobno kot pri nevronskih omrežjih v računalniškem vidu, tudi tu lahko definiramo združevanje aktivacij (angl. pooling), kot npr. maksimalno združevanje in povprečno združevanje, le da namesto na prostorski komponenti to naredimo na časovni komponenti.

3.3. Povratno nevronsko omrežje

Za modeliranje jezikovne informacije z nevronskimi omrežji se uporablja strukture, ki so primerne za diskreten tip podatkov. To so običajno povratna nevronska omrežja, ki se uporabljajo pri raznovrstnih nalogah obdelovanja naravnega jezika, npr. ustvarjanje teksta (Sutskever et al., 2011), prevajanje (Sutskever et al., 2014), analiza sentimenta (Tai et al., 2015), ... Povratna nevronska omrežja v primerjavi s klasičnimi nevronskimi omrežji niso omejena na podatke fiksne dolžine, temveč lahko delujejo z zaporedji poljubne dolžine.

Osnovna povratna nevronska omrežja imajo težave s pozabljanjem informacije in stabilnim učenjem, zato se večinoma uporabljata dva posebna modela povratnih nevronskih omrežij: Long Short-Term Memory (LSTM) (Hochreiter in Schmidhuber, 1997) in Gated Recurrent Unit (GRU) (Chung et al., 2014).

Model LSTM je matematično opisan z naslednjimi operacijami

$$\begin{aligned}i_t &= \sigma(W_{xi}^T x_t + W_{hi}^T h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}^T x_t + W_{hf}^T h_{t-1} + b_f) \\ o_t &= \sigma(W_{xo}^T x_t + W_{ho} h_{t-1} + b_o) \\ g_t &= \tanh(w_{xg}^T x_t + W_{hg}^T h_{t-1} + b_g) \\ c_t &= f_t \otimes c_{t-1} + i_t \otimes g_t \\ y_t &= h_t = o_t \otimes \tanh(c_t),\end{aligned} \tag{3}$$

model GRU pa z

$$\begin{aligned}z_t &= \sigma(W_{xz}^T x_t + W_{hz} h_{t-1} + b_z) \\ r_t &= \sigma(W_{xr}^T x_t + W_{hr} h_{t+1} + b_r) \\ g_t &= \tanh(W_{xg}^T x_t + W_{hg}^T (r_t \otimes h_{t-1}) + b_g) \\ h_t &= z_t \otimes h_{t-1} + (1 - z_t) \otimes g_t,\end{aligned} \tag{4}$$

kjer matrike W predstavljajo učene uteži modela, b učeni pristranski vektor modela, σ sigmoidno funkcijo, tanh hiperbolični tangens, $\otimes$ hadamardov produkt, x_t in y_t pa vhod in izhod modela ob času t. Vektorja c_t in h_t predstavljata notranje skrito stanje modela, ki omogoča prenos informacije čez daljše časovno obdobje.

3.4. Podatkovne zbirke

Modeli, ki pogojujejo izhodno sliko z jezikovno informacijo, so vizualno prepričljive rezultate dosegli na podatkovnih bazah z omejeno raznolikostjo. Dve taki podatkovni bazi, ki sta obširno uporabljeni, sta Caltech-UCSD Birds (Welinder et al., 2010), ki vsebuje 11.788 slik 200 različnih vrst ptic s desetimi opisi na sliko in Oxford-102 Flowers (Nilsback in Zisserman, 2008), ki vsebuje 8.189 slik 102 različnih vrst rož s petimi opisi na sliko. Podatkovna zbirka, ki vsebuje vsakdanje prizore na slikah, je Microsoft COCO (Lin et al., 2014) s 82.783 slikami, ki vsebuje pet tekstovnih opisov na sliko. Primeri slik s pripadajočimi opisi vsake podatkovne zbirke so prikazani na sliki 1.

Slika 1: Primer para tekstovnega opisa in pripadajoče slike za zbirko Oxford-102 Flowers (na vrhu), Caltech-UCSD Birds (na sredini) in Microsoft COCO (spodaj).

3.5. Vrednotenje

Za vrednotenje generativnih modelov se običajno uporabljata metriki Inception Score (Salimans et al., 2016) in Fréchet Inception Distance (Heusel et al., 2017), ki sta namenjeni ocenjevanju vizualne kakovosti in raznovrstnosti nastalih slik. Za ocenjevanje ustreznosti semantične informacije na izhodni sliki glede na tekstovni opis pa se v člankih najpogosteje zatekajo k človeškemu vrednotenju preko raznih spletnih platform. Pomanjkanje avtomatskih metod za vrednotenje semantične skladnosti v tekstovnem opisu in na ustvarjeni sliki je velik problem, saj onemogoča ovrednotenje prispevkov posameznih komponent modela h končnem rezultatu.

Izmed dveh omenjenih metrik se v tekstovno pogojenem ustvarjanju slik večinoma uporablja metrika Inception Score. Definirana je kot

$$\mathrm{IS} = \exp(\mathbb{E}_{\boldsymbol{x}} D_{KL}(p(y|x), p(y))), \tag{5}$$

kjer $x = G(z)$ predstavlja umetno ustvarjeno sliko iz naključnega vektorja $z \sim p_z$, $p(y|x)$ verjetnostna porazdelitev razvrščevalnika Inception (Szegedy et al., 2016) na posamezne razrede, $p(y) = \int p(y|x = G(z))dz$ pa mejna porazdelitev razredov. Običajno se validacijska množica podatkov razdeli na deset podmnožic, IS pa oceni na vsaki podmnožici. Rezultat nato podamo v obliki povprečja in standardne deviacije.

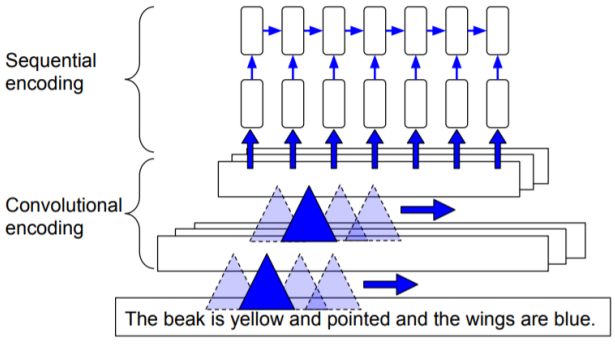
4. Jezikovne komponente v generativnih modelih

V nadaljevanju bosta opisani dve najpomembnejši metodi, ki sta uporabljeni kot jezikovna komponenta generativnega nasprotniškega modela. Globoke simetrične strukturne skupne vloženke je metoda, ki za osnovno enoto teksta vzame posamezno črko, globok pozornostni multimodalno podobnostni model pa za osnovno enoto teksta vzame besedo. Po opisu teh metod na kratko opišemo še doprinose najnovejših člankov s tega področja, možnosti uporabe slovenščine, rezultate trenutno najboljših modelov in možne izboljšave na tem področju.

4.1. Globoke simetrične strukturne skupne vloženke

Prvi članek, ki je prikazal prepričljivo ustvarjanje slik, pogojenih na njihovih tekstovnih opisi, je članek (Reed et al., 2016b). Za modeliranje tekstovne komponente so predlagali postopek, ki je združeval konvolucijska in povratna nevronska omrežja, imenovan globoke simetrične strukturne skupne vloženke (angl. Deep symmetric structured joint embedding). Metoda učenja tekstovnega modela je osnovana na članku (Reed et al., 2016a).

Postopek uči tekstovni model φ, slikovni model θ pa je prednaučen na nalogi razvrščanja podatkovne zbirke ImageNet (Deng et al., 2009) in je v postopku učenja zamrznjen. Tekstovni model φ je sestavljen iz začetnih konvolucijskih slojev, ki jim sledi višjenivojski LSTM sloj. Arhitektura modela je prikazana na sliki 2. Motivacija za tako arhitekturo izhaja iz dejstva, da je konvolucijska operacija hitra z nizko računsko kompleksnost, vendar pa se ne zaveda celotnega zaporedja tekstovnega opisa. S kombinacijo obeh arhitektur dosežemo hitro procesiranje zaradi začetnih konvolucijskih operacij, za časovno komponento pa skrbi LSTM, ki obdeluje višjenivojske značilke. Tekstovni model φ za osnovno enoto v tekstu vzame posamezno črko, ki jo kodira kot višjedimenzionalni vektor.



Slika 2: Arhitektura konvolucijskega in sekvenčnega kodiranja teksta. Za osnovno enoto vzamemo posamezne črke, ki jih kodiramo s konvolucijsko operacijo, višje-nivojske predstavitve pa s povratnim nevronskim omrežjem. Slika je povzeta po (Reed et al., 2016a).

Končno predstavitev teksta so modelirali kot

$$\varphi(t) = \frac{1}{L}\sum_{i=1}^{L} h_i, \tag{6}$$

kjer je h_i skriti vektor končnega sloja povratnega nevronskega omrežja, L pa dolžina zaporedja.

Ob danih podatkih $\mathcal{S} = \{(v_n, t_n, y_n), n = 1, ..., N\}$, kjer $v \in \mathcal{V}$ predstavlja slikovno informacijo, $t \in \mathcal{T}$ tekstovno informacijo, $y \in \mathcal{Y}$ pa razred, želimo ob učenju funkcij $f_v : \mathcal{V} \to \mathcal{Y}$ in $f_t : \mathcal{T} \to \mathcal{Y}$ minimizirati empirično tveganje, ki ga zapišemo kot

$$\frac{1}{N}\sum_{n=1}^{N} \Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n)), \tag{7}$$

kjer je Δ 0-1 kriterijska funkcija, z N pa smo označili število vseh parov v podatkovni zbirki.

Funkcijo kompatibilnosti slikovne in tekstovne informacije so definirali kot

$$F(v, t) = \theta(v)^T \varphi(t), \tag{8}$$

razvrščevalnika f_v in f_t pa sta definirana kot

$$\begin{aligned} f_v(v) &= \arg\max_{y \in \mathcal{Y}} \mathbb{E}_{t \sim \mathbb{T}(y)} F(v, t) \\ f_t(t) &= \arg\max_{yin\mathcal{Y}} \mathbb{E}_{v \sim \mathbb{V}(y)} F(v, t), \end{aligned} \tag{9}$$

kjer je $\mathcal{T}(y)$ podmnožica tekstovne informacije $\mathcal{T}$ razreda y, $\mathcal{V}(y)$ pa podmnožica slikovne informacije $\mathcal{V}$ razreda y.

Ker je 0-1 kriterijska funkcija v enačbi 7 nezvezna in nima definiranega odvoda, namesto te optimiziramo pomožno funkcijo, ki je zvezna:

$$\begin{aligned} &\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} l_v(v_n, t_n, y_n) + l_t(v_n, t_n, y_n) \\ &l_v(v_n, t_n, y_n) = \\ &\sum_{y \in \mathcal{Y}} \max(0, \Delta(y_n, y) + \mathbb{E}_{t \sim \mathcal{T}(y)}[F(v_n, t) - F(v_n, t_n)] \\ &l_t(v_n, t_n, y_n) = \\ &\sum_{y \in \mathcal{Y}} \max(0, \Delta(y_n, y) + \mathbb{E}_{t \sim \mathcal{T}(y)}[F(v, t_n) - F(v_n, t_n)] \end{aligned} \tag{10}$$

Z učenjem tako dobimo naučeno nevronsko omrežje φ, ki je ustrezno za kodiranje jezikovne informacije.

V članku (Reed et al., 2016b) so kodirali tekstovni opis s kodirnikom φ, nato pa so zakodiran opis s polnopovezano plastjo zmanjšali na 128-dimenzionalni vektor in dodali nelinearno aktivacijo. Ta vektor so zaporedno dodali naključnemu vektorju z, vse skupaj pa nato tvori začetni signal za generator G.

Diskriminator prav tako potrebuje informacijo o tekstovnemu opisu, saj želimo z diskriminatorjem oceniti, ali se dana slika in njen opis ujemata ali ne. Četudi je generator ustvaril prepričljivo sliko, moramo preveriti, ali ta slika ustreza našemu tekstovnemu opisu. Zato so v tem članku prej omenjeni 128-dimenzionalni vektor prostorsko ponovili in dodali v konvolucijski sloj prostorske dimenzije 4×4. Celotna arhitektura je prikazana na sliki 3.

Za učenje, ki bo upoštevalo semantično skladnost opisa in slike, moramo spremeniti kriterijsko funkcijo. Namesto enačbe 1 se diskriminator uči na parih podatkov. Označimo z x resnično sliko, s h kodiran tekst, ki semantično pripada trenutni sliki, z $\hat{x}$ sliko, ki jo je ustvaril generator in s $\hat{h}$ kodiran tekst, ki semantično ne ustreza trenutni sliki. Kriterijska funkcija za diskriminator, ki jo ta skuša maksimizirati, se tako glasi

$$\mathcal{L}_D = \log(D(x, h)) + \frac{\log(1 - D(x, \hat{h}) + \log(1 - D(\hat{x}, h))}{2}, \tag{11}$$

kriterijska funkcija za generator, ki jo ta skuša maksimizirati, pa se sedaj glasi

$$\mathcal{L}_G = \log(D(\hat{x}, h). \tag{12}$$

V članku (Zhang et al., 2017), ki je prikazal vizualno prepričljivejše in višje-resolucijske slike, kot (Reed et al., 2016b), so uporabljali isti način kodiranja φ, vendar skušajo doseči večjo razpršenost tekstovnih kodiranj. Ob upoštevanju, da je količina tekstovnih opisov omejena in da so dimenzije kodiranega tekstovnega opisa običajno visokodimenzionalne (običajno > 100 dimenzij), se v prostoru mnogoterosti kodiranega teksta pojavlja nezveznost. Zaradi tega v članku predlagajo tehniko pogojene avgmentacije (angl. Conditioning Augmentation), ki je zelo podobna tehniki kodiranja pri variacijskem avtokodirniku (Kingma in Welling, 2013).

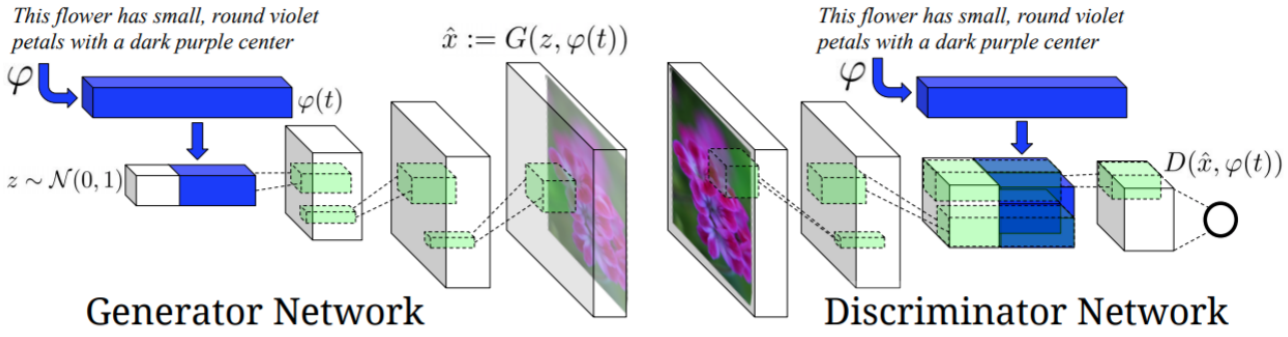
Metoda tekstovno kodiranje c vzorči iz Gaussove porazdelitve $\mathcal{N}(\mu(\varphi(t)), \Sigma(\varphi(t)))$, kjer je kovariančna matrika Σ diagonalna. Tako vektor povprečij μ kot diagonalna kovariančna matrika Σ sta funkciji tekstovnega kodiranja $\varphi(t)$. Za preprečevanje prenaučenosti modela h kriterijski funkciji dodajo še izraz za regularizacijo te verjetnostne porazdelitve s Kullback-Leibler divergenco $D_{KL}(\mathcal{N}(\mu(\varphi(t)), \Sigma(\varphi(t))), \mathcal{N}(0, I))$. S to tehniko zagotovimo, da posamezni kodirani tekst zavzame večji prostor na mnogoterosti kodiranih tekstov, obenem pa generativnimi model spodbuja k robustnosti na manjše perturbacije kodiranega teksta. Tehnika pogojne avgmentacije je postala popularna metoda za vse nadaljnje generativne modele, ki so zunanji signal pogojili s tekstovnim opisom.

Poleg tega so v (Zhang et al., 2017) namesto arhitekture enega generativnega modela uporabili dva GAN modela, zložena en na drugega. Prvi je ustvaril nižjeresolucijsko 64×64 sliko, drugi pa je glede na to sliko ustvaril višjeresolucijsko 256×256 sliko.

4.2. Globok pozornostni multimodalni podobnostni model

Motivacija za modeliranje po črkah namesto modeliranja po besedah je predvsem omogočanje večje robustnosti na tipkarske napake v opisih. Ob zadostni količini lepo označenih podatkov pa imajo prednost modeli, ki kot samostojno enoto modelirajo posamezno besedo. Besede so modelirane kot vektorske zloženke, ki so pogosto prednaučene na nenadzorovan način, kot npr. GloVe (Pennington et al., 2014). Celotni tekstovni opis nato običajno modeliramo s povratno nevronsko mrežo.

Model AttnGAN (Xu et al., 2018) je besede modeliral z učenimi vektorskimi zloženkami. Niz besed je bil opisan z besednimi značilkami $e \in \mathbb{R}^{D \times T}$, kjer je T dolžina niza, D pa dimenzionalnost značilk. Iz niza besed modeliramo



Slika 3: Predlagana arhitektura modela iz članka (Reed et al., 2016b). Kodiran tekst $\varphi(t)$ se zaporedno doda naključnemo vektorju z, skupaj pa tvorita začetni signal za naslednje sloje generatorskega omrežja. Kodiran tekst $\varphi(t)$ se prav tako doda diskriminatornem omžreju, s čimer lahko prisilimo generatorsko omrežje k semantični povezanosti tekstovnega opisa in nastale slike. Slika je povzeta po (Reed et al., 2016a).

globalno tekstovno predstavitev $\bar{e}$ z uporabo dvosmernega LSTM modela. Dvosmerni LSTM model je sestavljen iz dveh LSTM komponent, kjer vsaka komponenta modelira niz besed v svoji smeri, prva od začetka povedi proti koncu, druga pa od konca povedi proti začetku. Globalni vektor $\bar{e}$ dobimo z združenjem končnega skritega vektorja h obeh komponent.

Podobno kot pri modelu StackGAN, tudi ta model uporabi več diskriminatorjev, kjer je vsak diskriminator zadolžen za svojo resolucijo. Bistveno se razlikuje v tem, da AttnGAN tekst ne modelira le z globalno tekstovno predstavitev $\hat{e}$, temveč uporabi tudi posamezne besede e za implementacijo mehanizma pozornosti v modelu. Pozornost predstavlja mehanizem, prvič predstavljen v članku (Bahdanau et al., 2014), kjer se model sam skuša naučiti, koliko posamezna vhodna komponenta prispeva h končnemu izhodu.

Tekstovni kodirnik φ v nasprotju s prejšnimi modeli učijo po postopku, imenovanem globok pozornostni multimodalni podobnostni model (angl. deep attentional multimodal similarity model). Model maksimizira podobnost slike in posameznih besed v pripadajočih tekstovnih opisih. Označimo z $v \in \mathbb{R}^{D \times M}$ vizualne značilke, pridobljene iz prednaučene konvolucijske mreže, kjer je D njihova dimenzionalnost, M pa število prostornih elementov globokih značilk. Posamezno kodirano besedo označimo z e_i. Izračunamo podobnostno matriko

$$s = e^T v, \tag{13}$$

kjer velja $s \in \mathbb{R}^{T \times M}$, element matrike $s_{i,j}$ pa ponazarja podobnost med i-to besedo in j-to regijo slike. To matriko normaliziramo

$$\bar{s}_{i,j} = \frac{\exp(s_{i,j})}{\sum_{k=0}^{T-1} \exp(s_{k,j})}. \tag{14}$$

Vektorsko predstavitev slike glede na i-to besedo dobimo z

$$c_i = \sum_{j=0}^{M-1} \alpha_j v_j, \tag{15}$$

kjer je

$$\alpha_j = \frac{\exp(\gamma_1 \bar{s}_{i,j})}{\sum_{k=0}^{M} \exp(\gamma_1 \bar{s}_{i,k})}. \tag{16}$$

S faktorjem γ_1 uravnavamo prispevke drugih regij na vektorsko predstavitev posamezne regije.

Ustreznost med i-to besedo e_i in predstavitvijo celotne slike c_i definiramo z mero kosinusne podobnosti. Funkcijo ustreznosti definiramo kot

$$U(c_i, e_i) = \frac{c_i^T e_i}{||c_i|| ||e_i||}, \tag{17}$$

ustreznost med celotno sliko Q in celotnim tekstovnim opisom pa kot

$$U(Q, D) = \log(\sum_{i=0}^{T-1} \exp(\gamma_2 U(c_i, e_i)))^{\frac{1}{\gamma_2}}, \tag{18}$$

kjer je γ_2 faktor, s katerim uravnavamo prispevke vektorskih predstavitev drugih besed.

Funkcija $U(Q, D)$ nam torej vrne podobnost med tekstovnim opisom in besedo. Edino, kar nam še preostane, je definiranje kriterijske funkcije za učenje. Učenje definiramo tako, da povečujemo logaritmsko verjetnost ustreznih parov opisov in slik. Konkretno lahko za sveženj parov $\{(Q_i, D_i)\}_{i=1}^{M}$ zapišemo, da je verjetnost ustreznosti opisa D_i in slike Q_i

$$P(D_i|Q_i) = \frac{\exp(\gamma_3 U(Q_i, D_i))}{\sum_{j=1}^{M} \exp(\gamma_3 U(Q_i, D_i))}. \tag{19}$$

V danem svežnju označimo, da si Q_i in D_i ustrezata, vsi ostali opisi pa ne ustrezajo. Kriterijska funkcija, ki jo minimiziramo, je definirana kot

$$L_1 = -\sum_{i=1}^{M} \log P(D_i|Q_i), \tag{20}$$

simetrično pa minimiziramo tudi

$$L_2 = -\sum_{i=1}^{M} \log P(Q_i|D_i). \tag{21}$$

S to optimizacijo si zagotovimo primerne vloženke e_i, ki se uporabljajo v sklopu mehanizma pozornosti v konvolucijskih slojih modela GAN. Prav tako si želimo še kvalitetne globalne predstavitve celotnega opisa, za kar poskrbi model LSTM. Globalna predstavitev tekstovnega opisa $\bar{e}$ in globalna predstavitev celotne slike kot povprečenje aktivacij konvolucijskih filtrov na zadnjem sloju $\bar{v}$ se uporabita kot funkciji enačbe 17, nato pa postopek sledi vsem nadaljnjim enačbam. S tem dobimo še globalni doprinos h kriterijski funkciji.

4.3. Ostali modeli

Trenutno najboljše rezultate daje model MirrorGAN (Qiao et al., 2019), kjer skušajo modelu zagotoviti konsistentnost originalnih tekstovnih opisov in tekstovnih opisov ustvarjenih slik. Postopek se zgleduje po konceptu ciklične konsistence modela CycleGAN (Zhu et al., 2017). Konsistentnost tekstovnih opisov zagotovijo z predučenjem modela, ki glede na sliko ustvari ustrezen tekstovni opis. Celotni sistem učenja generativnega modela nato temelji na tem, da mora slika generatorja, pogojena na določenem tekstu, dobiti isti tekst kot rezultat novo predstavljenega modela ustvarjanja tekstovnega opisa. Ostali deli arhitekture se močno zgledujejo po modelu AttnGAN.

Opazimo, da vsi novejši modeli sliko gradijo postopno, torej z uporabo več generatorjev in diskriminatorjev za posamezne resolucije. V nasprotju z njimi pa v članku (Souza et al., 2020) model učijo direktno na najvišji resoluciji 256×256 brez uporabe vmesnih generatorjev in diskriminatorjev. To storijo z modifikacijo modela BigGAN (Brock et al., 2019), ki ga posodobijo na tak način, da namesto informacije o razredu izhodnega modela dobi kodiran tekstovni opis $\varphi(t)$.

4.4. Uporaba slovenščine

Domnevamo, da je zaradi odsotnosti podatkovnih zbirk in večje narave pregibnosti besed slovenščina problematičen jezik za učinkovito tekstovno pogojeno ustvarjanje slik. Mogoča rešitev se skriva v skrbnem predprocesiranjem podatkov ali uporabi predhodno naučenih besednih zloženk s postopki, kot je npr. GloVe (Pennington et al., 2014) in FastText (Bojanowski et al., 2017). Slednji tudi ponuja prednaučen model za slovenščino (Grave et al., 2018). Ti postopki temeljijo na nenadzorovanem učenju na velikanskih bazah podatkov (npr. Wikipedija nekega jezika), s čimer pridobijo besedne vloženke, ki besede s podobnim pomenom preslika bližje v višjedimenzionalnem prostoru. To lahko s pridom izkoristijo različni modeli za kodiranje tekstovnega opisa (Collobert et al., 2011).

4.5. Rezultati

V tabeli 1 so prikazani rezultati v obliki metrike Inception Score za nekaj najboljših modelov na področju tekstovno pogojenega ustvarjanja slik. Rezultati so ovrednoteni na validacijski množici podatkovnih zbirk, kjer ustvarimo 30.000 slik glede na tekstovni opis. Glede na ustvarjene slike izračunamo metriko Inception Score. Rezultati so prikazani za podatkovni zbirki Caltech-UCSD Birds (CUB) (Welinder et al., 2010) in Microsoft COCO (Lin et al., 2014). Ovrednoteni modeli so GAN-INT-CLS (Reed et al., 2016b), StackGAN (Zhang et al., 2017), AttnGAN (Xu et al., 2018) in MirrorGAN (Qiao et al., 2019).

Tabela 1: Inception Score posameznih modelov. Višji rezultat pomeni kvalitetnejši model.

Model	CUB	COCO
GAN-INT-CLS	2.88	7.88
StackGAN	3.70	8.45
AttnGAN	4.36	25.89
MirrorGAN	4.56	26.47

Na sliki 4 pa so prikazani primeri rezultatov posameznih modelov, ko izhod modela pogojimo z nekim tekstovnim opisom.

Slika 4: Rezultati posameznih modelov GAN na podatkovni zbirki Caltech-UCSD Birds. Od zgoraj navzdol: GAN-INT-CLS (Reed et al., 2016b), StackGAN (Zhang et al., 2017), AttnGAN (Xu et al., 2018) in MirrorGAN (Qiao et al., 2019).

4.6. Možnosti nadgradnje

Potencialne možnosti nadgradnje vidimo predvsem v naslednjih izboljšavah:

- **Tekstovni model**
  Trenutni tekstovni modeli delujejo na principu nenadzorovanega učenja za učenje značilk posameznih be-

sed v kombinaciji s povratno nevronsko mrežo za modeliranje celotnega stavka. Novejši tekstovni modeli (Devlin et al., 2018) delujejo brez povratnega mehanizma, temevač na podlagi mehanizma pozornosti z nenadzorovanem učenjem na velikanskih podatkovnih bazah pridobijo značilke, ki ustrezajo celotni povedi. Z doučevanjem na posameznih podatkovnih zbirkah so ti modeli na klasičnih podatkovnih zbirkah za vrednotenje razumevanje teksta presegli rezultate, ki so jih dosegale povratne nevronske mreže. Z ustreznim načinom doučevanja domnevamo, da lahko izkoristimo predznanje, zajeto v prednaučenem tekstovnem modelu, s tem pa pridobimo boljšo pogojenost glede na tekstovni opis.

- **Bogatenje podatkov**
  Podatkovne zbirke, opisane v članku, so zaradi zamudnega zbiranja podatkov manj obsežne kot podatkovne zbirke, namenjene drugim izključno eni modaliteti. Za generativna nasprotniška omrežja je znano, da boljše delujejo večanju količine podatkov. Nedavni članek (Zhao et al., 2020) pa je pokazal, da je mogoče kvalitetne generativne modele naučiti z veliko manj podatki ob uporabi odvodljivih transformacij za bogatenje podatkov.

- **Način učenja**
  V članku (Brock et al., 2019) so pokazali, da generativni modeli dosežejo boljše rezultati ob večanju velikosti svežnja (angl. batch size). Slabost te tehnike pa je velika računska potratnost.

5. Zaključek

Tematika tekstovno pogojenega ustvarjanja slik je aktivna in privlačna veja raziskovanja, ki v zadnjih letih dobiva velik zagon. V članku smo opisali metode, podatkovne zbirke in trenutne rezultate na področju tekstovno pogojenega generiranja slik z generativnimi nasprotniškimi omrežji. Posvetili smo se arhitekturi modelov za kodiranje tekstovnega opisa, načinom kodiranja tekstovnega opisa in ustreznim kriterijskim funkcijam. Na koncu smo predstavili še možnosti uporabe slovenščine in nekaj obetajočih novejših metod.

6. Zahvala

Predstavljeno delo je delno financirano s strani ARRS v okviru raziskovalnega programa P2-0250 Metrologija in biometrični sistemi ter aplikativnega raziskovalnega projekta L7-9406 OptiLEX.

7. Literatura

Martin Arjovsky, Soumith Chintala in Léon Bottou. 2017. Wasserstein generative adversarial networks. V: *International Conference on Machine Learning*, str. 214–223.

Dzmitry Bahdanau, Kyunghyun Cho in Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Piotr Bojanowski, Edouard Grave, Armand Joulin in Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

A. Brock, J. Donahue in K. Simonyan. 2019. Large scale GAN training for high fidelity natural image synthesis. V: *ICLR*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho in Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu in Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li in L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. V: *CVPR09*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee in Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio in Graham W Taylor. 2019. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. V: *Proceedings of the IEEE International Conference on Computer Vision*, str. 10304–10312.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville in Yoshua Bengio. 2014. Generative adversarial nets. V: *Advances in neural information processing systems*, str. 2672–2680.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin in Tomas Mikolov. 2018. Learning word vectors for 157 languages. V: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin in Aaron C Courville. 2017. Improved training of wasserstein gans. V: *Advances in neural information processing systems*, str. 5767–5777.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler in Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. V: *Advances in neural information processing systems*, str. 6626–6637.

Sepp Hochreiter in Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou in Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. V: *Proceedings of the IEEE conference on computer vision and pattern recognition*, str. 1125–1134.

T. Karras, T. Aila, S. Laine in J. Lehtinen. 2018. Progressive growing of GANs for improved quality, stability, and variation. V: *ICLR*.

Tero Karras, Samuli Laine in Timo Aila. 2019a. A style-based generator architecture for generative adversarial networks. V: *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition, str. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen in Timo Aila. 2019b. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*.

Diederik P Kingma in Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár in C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. V: *European conference on computer vision*, str. 740–755. Springer.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang in Stephen Paul Smolley. 2017. Least squares generative adversarial networks. V: *Proceedings of the IEEE International Conference on Computer Vision*, str. 2794–2802.

Maria-Elena Nilsback in Andrew Zisserman. 2008. Automated flower classification over a large number of classes. V: *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec.

Sebastian Nowozin, Botond Cseke in Ryota Tomioka. 2016. f-gan: Training generative neural samplers using variational divergence minimization. V: *Advances in neural information processing systems*, str. 271–279.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang in Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. V: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, str. 2337–2346.

Jeffrey Pennington, Richard Socher in Christopher D Manning. 2014. Glove: Global vectors for word representation. V: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, str. 1532–1543.

Tingting Qiao, Jing Zhang, Duanqing Xu in Dacheng Tao. 2019. Mirrorgan: Learning text-to-image generation by redescription. V: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, str. 1505–1514.

A. Radford, L. Metz in S. Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Scott Reed, Zeynep Akata, Honglak Lee in Bernt Schiele. 2016a. Learning deep representations of fine-grained visual descriptions. V: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, str. 49–58.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele in Honglak Lee. 2016b. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford in Xi Chen. 2016. Improved techniques for training gans. V: *Advances in neural information processing systems*, str. 2234–2242.

Douglas M Souza, Jônatas Wehrmann in Duncan D Ruiz. 2020. Efficient neural architecture for text-to-image synthesis. *arXiv preprint arXiv:2004.11437*.

Ilya Sutskever, James Martens in Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. V: *Proceedings of the 28th international conference on machine learning (ICML-11)*, str. 1017–1024.

Ilya Sutskever, Oriol Vinyals in Quoc V Le. 2014. Sequence to sequence learning with neural networks. V: *Advances in neural information processing systems*, str. 3104–3112.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens in Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. V: *Proceedings of the IEEE conference on computer vision and pattern recognition*, str. 2818–2826.

Kai Sheng Tai, Richard Socher in Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie in P. Perona. 2010. Caltech-UCSD Birds 200. Tehnično poročilo CNS-TR-2010-001, California Institute of Technology.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang in Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. V: *Proceedings of the IEEE conference on computer vision and pattern recognition*, str. 1316–1324.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang in Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. V: *Proceedings of the IEEE international conference on computer vision*, str. 5907–5915.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu in Song Han. 2020. Differentiable augmentation for data-efficient gan training. *arXiv preprint arXiv:2006.10738*.

Jun-Yan Zhu, Taesung Park, Phillip Isola in Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. V: *Proceedings of the IEEE international conference on computer vision*, str. 2223–2232.