SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Detekcija zatrovanih podataka korištenjem generativnih modela za slike

Josip Srzić
Voditelj: *Siniša Šegvić*

Zagreb, svibanj 2023.

SADRŽAJ

1. Uvod

S naglim razvojem računalnog vida i sve češćom primjenom sustava temeljenih na strojnom učenju i računalnom vidu, javlja se sve veća potreba osigurati sigurnost takvih sustava. To posebno vrijedi za kritične sustave poput autonomne vožnje, robotike i medicine. Kao korisnici želimo imati garanciju da će se sustav dobro ponašati u nepredviđenim i teškim situacijama, odnosno očekujemo da će biti stabilan i da neće naglo mijenjati odluku za male perturbacije ulaznog podatka. Veliku opasnost za duboke neuronske mreže predstavljaju napadi ubacivanjem stražnjih vrata (engl. *backdoor*) [7] u skup podataka za učenje. To je široka obitelj napada kod kojih maliciozni agent ubacuje vlastite podatke ili izmjenjuje postojeće u skupu za učenje. Posljedica je da naučeni model naizgled radi dobro, ali daje krive predikcije ako mu se na ulazu pojave određeni okidači (engl. *trigger*) postavljeni od strane napadača.

Detekciji zatrovanih podataka se može pristupiti kao problemu detekcije anomalija, odnosno pronalaženja stršećih vrijednosti u distribuciji ulaznih primjera. Također postoje i poveznice sa zadatkom klasifikacije u kojem modelu ostavljamo mogućnost da primjer ne klasificira u niti jednu od unaprijed određenih klasa (engl. *open-set recognition*). Klasični pristupi tom problemu se zasnivaju na nadziranom učenju i diskriminatvnim modelima. Postoje i hibridni pristupi [13] koji uz diskriminativne koriste i generativne modele kako bi ostvarili što kvalitetniju procjenu izglednosti potencijalno zatrovanih podataka. Nedavni razvoji difuzijskih modela i modela temeljenih na normalizirajućim tokovima daju zanimljiv pravac za istraživanje. Cilj ovog seminara je empirijski provjeriti u kojoj mjeri generativni modeli mogu otkriti zatrovane podatke. Ideja je, dakle, uzeti poznate skupove podataka, umjetno ih izmijeniti dodavanjem neprijateljskih primjera i vidjeti hoće li naučeni model pridavati manju izglednost neprijateljskim primjera u odnosu na ostale primjere iz skupa podataka.

2. Difuzijski modeli

Difuzijski modeli su klasa probabilističkih generativnih modela koja je stekla veliku popularnost u zadnjem vremenu. Pokazali su se kao vrlo uspješan pristup za generiranje slika. U mnogim zadacima postižu jednako dobre ili čak bolje rezultate [2] od nekih do sada popularnih pristupa, poput generativnih neprijateljskih mreža. Nedugo nakon prvih difuzijskih modela pojavile su se inačice koje generirane slike uvjetuju podacima različitih modalnosti, najčešće tekstom ili nekom drugom slikom. Na taj način se proces generiranja usmjerava prema željenoj slici. Neki od modela koji se bave tim zadatkom su *Stable Diffusion*, Googleov *Imagen* te *DALL-E* od OpenAI-a.

Difuzijski modeli su izvorno predstavljeni u [10]. Autori navode ideje iz neravnotežne termodinamike kao inspiraciju za dizajn modela. U [4] autori povezuju difuzijske modele s modelima zasnovanim na podudaranju mjere (engl. *score matching*). Također, prvi pokazuju da difuzijski modeli mogu generirati slike visoke kvalitete koje se mogu mjeriti s ostalim popularnim generativnim modelima. To potvrđuju postizanjem stanja tehnike za FID metriku na skupu CIFAR-10 [6]. Iako su pokazali da su difuzijski modeli usporedivim s ostalim generativnim modelima po pitanju kvalitete generiranih slika, još uvijek su zaostajali po pitanju ostalih metrika, poput log izglednosti. U [9] istraživači iz OpenAI-a uvode niz jednostavnih prijedloga i poboljšanja pomoću kojih znatno povećavaju log izglednost bez gubljenja na kvaliteti generiranih slika. Svoj rad nastavljaju u [2], u kojem pokazuju da difuzijski modeli mogu ostvariti bolju vjerodostojnost (engl. *fidelity*) kao i bolju raznolikost (engl. *diversity*) generiranih slika od generativnih neprijateljskih mreža.

2.1. Teorijska podloga

Sama ideja difuzijskih modela je jednostavna: prilikom učenja uzorkujemo neku sliku x_0 iz skupa za učenje te joj kroz niz koraka T postepeno dodajemo šum modeliran Gaussovom distribucijom i na taj način uništavamo originalnu informaciju iz slike. Kad bi vrijedilo $T \to \infty$, odnosno kad bi imali beskonačno koraka, ono što bi se nalazilo

u x_T bila bi izotropna Gaussova distribucija, odnosno potpuni šum. U praksi će se, naravno, koristiti neki fiksan veliki broj koraka (npr. 4000). Sad želimo nekako obrnuti taj proces, odnosno krenuti od potpunog šuma i nizom koraka postepeno uklanjati šum sa slike. Nakon što se ukloni sav šum, ono što je preostalo je, nadamo se, neka potpuno nova slika koja izgleda kao da je mogla doći iz skupa za treniranje.

Malo formalnije, difuzija se sastoji od dva Markovljeva procesa: **unaprijednog** i **unatražnog**.

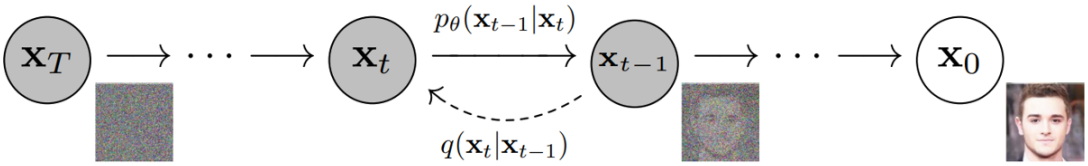
2.1.1. Unaprijedni proces

Unaprijedni proces naziva se još i difuzijski proces. On kreće od slike x_0 i dodavanjem šuma stvara niz latentnih varijabli od x_1 do x_T. Varijanca Gaussovog šuma prati fiksan raspored, definiran nizom hiperparametara $\beta_1, ..., \beta_T \in (0, 1)$. Matematički se to može zapisati ovako:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\mathbf{I}) \tag{2.1}$$

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \tag{2.2}$$

Iz izraza 2.1 je vidljivo da kako β_t raste od 0 prema 1, srednja vrijednost distribucije teži prema 0, a varijanca prema jediničnoj matrici.



Slika 2.1: Dva Markovljeva procesa difuzijskih modela. Isprekidanom strelicom prikazan je smjer unaprijednog procesa (oznaka q) koji kreće od čiste slike u x_0 i završava s potpunim šumom u x_T. Unatražni proces je prikazan punom strelicom (oznaka p) i kreće se s lijeva (potpuni šum) prema desno (slika s uklonjenim šumom). Slika je preuzeta iz [4].

Umnožak Gaussovih distribucija je i dalje Gaussova distribucija. Ta činjenica daje zgodno svojstvo unaprijednom procesu. Da bismo došli do zašumljene slike u koraku t, ne trebamo pratiti sve korake 0 do $t-1$ unaprijednog Markovljevog lanca. Umjesto

toga, možemo doći do proizvoljnog koraka difuzije t u jednom koraku, koristeći sljedeću parametrizaciju:

$$\alpha_t = 1 - \beta_t \tag{2.3}$$

$$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i \tag{2.4}$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}) \tag{2.5}$$

2.1.2. Unatražni proces

Unatražni proces također modeliramo Gaussovom distribucijom. To je opravdano jer u svakom koraku unaprijednog prolaza dodajemo samo jako malu količinu šuma. Štoviše, može se pokazati da u limesu kada β_t infinitezimalno teži prema 0, unatražni proces ima identičan funkcijski oblik kao i unaprijedni proces [3]. Idealni unatražni proces je opisan s $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. Međutim, on ovisi o cijeloj distribuciji podataka, stoga ga aproksimiramo:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) := \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)) \tag{2.6}$$

$$p_\theta(\boldsymbol{x}_{0:T}) := p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \tag{2.7}$$

pri čemu je $p(\boldsymbol{x}_T) := \mathcal{N}(\boldsymbol{x}_t; \mathbf{0}|\mathbf{I})$, odnosno jedinična Gaussova distribucija. U [4] varijancu $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ postavljaju na fiksne vrijednosti koje ovise samo o vremenskom koraku t, a za predviđanje srednje vrijednosti koriste neuronsku mrežu. Za model odabiru modificiranu U-Net arhitekturu. Specifičnost U-Neta je da su izlazi iz modela jednakih prostornih dimenzija kao i ulazi. To je zgodno jer su sve latentne reprezentacije difuzijskog modela također jednakih prostornih dimenzija. U rezidualne blokove modela se dodatno ubacuje informacija o vremenskom koraku t, budući da se u različitim vremenskim koracima dodaje različita količina šuma, ovisno o odabranom rasporedu. Vremenski korak t se pritom pretvara u vektor koristeći pozicijska ugrađivanja (engl. *positional embeddings*) na istovjetan način kao u originalnom radu o transformerskim modelima [11].

2.1.3. Gubitak

Difuzijski modeli, poput varijacijskih autoenkodera [5] i ostalih probabilističkih generativnih modela s latentnim varijablama, ne optimiraju direktno negativnu log-izglednost podataka. Razlog tomu je što bi za to bilo potrebno uzeti u obzir sve moguće vrijednosti latentnih varijabli, a to nije traktabilno. Stoga se umjesto log-izglednosti optimira donja ograda na log-izglednost, još poznata kao i varijacijska donja ograda (engl. *variational/evidence lower bound*):

$$\mathbb{E}\left[-\log p_\theta(\boldsymbol{x}_0)\right] \leq \mathbb{E}_q\left[-\log \frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] \tag{2.8}$$

Ako posteriornu distribuciju dodatno uvjetujemo početnim podatkom x_0, može se pokazati da takva distribucija $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ postaje traktabilna. Ta pretpostavka vrijedi za proces učenja modela, jer nam je uvijek poznat ulazni podatak. Uz tu pretpostavku, varijacijska donja ograda (gubitak) se dalje može raspisati ovako:

$$\mathcal{L}_{vlb} := \mathcal{L}_0 + \mathcal{L}_1 + \ldots + \mathcal{L}_{T-1} + \mathcal{L}_T \tag{2.9}$$

$$\mathcal{L}_0 := -\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) \tag{2.10}$$

$$\mathcal{L}_{t-1} := D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \tag{2.11}$$

$$\mathcal{L}_T := D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T)) \tag{2.12}$$

D_{KL} označava mjeru različitosti dviju distribucija koja se naziva Kullback-Leiblerova (KL) divergencija. Svaki član osim $\mathcal{L}_0$ je KL divergencija dviju Gaussovih distribucija, što znači da se može izračunati u zatvorenoj formi. L_0 se evaluira zasebnim dekoderom, a $\mathcal{L}_T$ je konstanta, stoga ne igra ulogu u procesu učenja i može se zanemariti. U [4] odlučuju prilikom učenja uniformno uzorkovati vremenske korake t i na taj način učiti svaki član gubitka zasebno. Pokazuju da se u tom slučaju gubitak može izraziti kao skalirana i kvadrirana L2 norma razlike između stvarnog šuma koji je dodan slici i šuma kojeg predviđa model:

$$\mathcal{L}_t = \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\epsilon},t}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)||^2\right] \tag{2.13}$$

Empirijski su pokazali da dobivaju još bolje rezultate ako zanemare skalarni član i tako dolaze dolaze do konačne formulacije jednostavnog gubitka:

$$\mathcal{L}_{simple} = \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\epsilon},t}\left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)||^2\right] \tag{2.14}$$

2.1.4. Učenje i generiranje slika

Učenje se može opisati ovim jednostavnim algoritmom:

Algoritam 1 Učenje modela

1: **Ponavljaj**
2: $\quad \boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$
3: $\quad t \sim Uniformno(\{1, ..., T\})$
4: $\quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: $\quad$ Odradi korak gradijentnog spusta na temelju
$$\nabla_\theta ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||^2$$
6: **Do** konvergencije

Generiranje/uzorkovanje slike kreće od zadnjeg koraka i prolazi kroz sve korake difuzije. U posljednjem unatražnom koraku $t = 1$ ne dodajemo varijancu (šum) jer bi to samo moglo pokvariti finalnu generiranu sliku $\boldsymbol{x}_0$:

Algoritam 2 Uzorkovanje iz naučene distribucije

1: $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **Za** $t = T, \ldots, 1$ **radi**
3: $\quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ako $t > 1$, inače $\boldsymbol{z} = \mathbf{0}$
4: $\quad \boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)) + \sigma_t \boldsymbol{z}$
5: **kraj Za**
6: **Vrati** $\boldsymbol{x}_0$

2.1.5. Poboljšanja

Već smo spomenuli kako je [9] uveo niz poboljšanja za difuzijske modele. Ovdje ćemo ih samo navesti:

- ne koriste fiksnu varijancu u unatražnom procesu, već ju također uče
- uvode hibridni gubitak, kao kombinaciju jednostavnog gubitka $\mathcal{L}_{simple}$ i gubitka varijacijske donje ograde $\mathcal{L}_{vlb}$
- ne koriste linearni nego kosinusni raspoređivač šuma
- koriste veći broj koraka (4000 umjesto 1000)
- pokazuju da model generira dobre slike i sa smanjenim brojem koraka prilikom uzorkovanja

3. Eksperimenti

Sav kod vezan uz difuziju preuzet je i po potrebi izmijenjen iz službene implementacije za [9]. Sav kod vezan uz trovanje podataka i napade dolazi iz [8].

3.1. Generiranje slika

Cilj prvog eksperimenta bio je praktično primijeniti difuzijske modele za generiranje slika iz različitih skupova podataka i pritom vizualizirati unatražni proces, odnosno postepeno uklanjanje šuma iz slike. Težine modela koji su korišteni za generiranje također dolaze iz službenog repozitorija za [9].

Slika 3.1: Generiranje slika iz modela naučenog na skupu CIFAR-10 [6]. Model koristi hibridni gubitak i naučene varijance. Za regularizaciju se koristi izostavljanje neurona (engl. *dropout*) s parametrom 0.3. Difuzija se provodi kroz 4000 koraka uz kosinusni raspoređivač šuma.

Slika 3.2: Generiranje slika iz modela naučenog na skupu ImageNet64 [1]. Model koristi hibridni gubitak i naučene varijance. Difuzija se provodi kroz 4000 koraka uz kosinusni raspoređivač šuma.

Slika 3.3: Vizualizacija unatražnog procesa za generiranje slike iz skupa LSUN Bedroom 256x256 [12]. Model koristi hibridni gubitak i naučene varijance. Difuzija se provodi kroz 1000 koraka uz linearni raspoređivač šuma.

3.2. Eksperimenti nad zatrovanim podacima

3.2.1. Trovanje skupa CIFAR-10 s jednim ciljnim razredom

Trovanje podataka provodi se jednostavnom metodom *BadNets* koja se može opisati na sljedeći način:

Algoritam 3 Trovanje skupa CIFAR-10 metodom *BadNets* s jednim ciljnim razredom

1: učitaj cijeli skup podataka X, udio trovanja p i ciljni razred napada y_{target}
2: uniformno uzorkuj pX slika iz skupa X i dodaj ih u skup $X' \subset X$
3: **Za** svaku sliku i pripadnu oznaku $(\boldsymbol{x}, y)$ iz X' **radi**
4: $\quad \boldsymbol{x}_{[:,-3:,-3:]} = 255$ ▷ postavi piksele u donjem desnom kvadratu slike dimenzija 3x3 na maksimalnu vrijednost
5: $\quad y = y_{target}$ ▷ postavi oznaku razreda na y_{target}
6: **kraj Za**
7: **Vrati** $X[y = y_{target}] \cup X'$ ▷ vrati sve slike koje pripadaju ciljnom razredu

Korak 4 zapravo opisuje dodavanje jednostavnog okidača (engl. *trigger*) u sliku. U početnom eksperimentu ovaj algoritam provodimo sa sljedećim postavkama:

- X je cijeli skup za treniranje iz CIFAR-10 (50000 slika jednoliko raspoređenih u 10 razreda)
- $p = 0.05$
- $y_{target} = 1$ (razred "auto")

Novonastali skup $X[y = y_{target}] \cup X'$ sastoji se od 7270 slika s oznakom "auto". Kvantitativno, u njemu postoje 3 tipa slika kao što se vidi u prikazu 3.4.

(a) Čista (nezatrovana) slika auta

(b) Zatrovana slika auta

(c) Zatrovana slika druge klase

Slika 3.4: Tri tipa slika nastalih trovanjem skupa CIFAR-10 s jednim ciljnim razredom.

3.2.2. Prvi (grubi) eksperiment

Hipoteza ovog seminara je: difuzijski modeli će pridavati veću izglednost čistim (nezatrovanim) podacima nego zatrovanim. Da bismo to provjerili, nakon modificiranja skupa za treniranje na njemu je istreniran difuzijski model na 800k iteracija. Jedna iteracija označava jedan korak gradijentnog spusta za N podataka, gdje je N veličina mini-grupe. Hiperparametri modela: veličina mini-grupe 64, stopa učenja $1e-4$, 4000 koraka difuzije, linearni raspoređivač šuma, jednostavni gubitak L_{simple}.

Zatim se provodi evaluacija izglednosti slika mjerenjem bitova po dimenziji (engl. *bits per dimension*). Ta metrika govori koliko bitova je potrebno modelu za kodiranje svake dimenzije podataka. Drugim riječima, to je mjera koja pokazuje koliko dobro model može sažeti informacije iz podataka. Manja vrijednost znači da modelu ne treba puno bitova da predstavi neki podatak, što znači da mu pridaje veću izglednost. Kad se radi o slikama, svaka dimenzija podataka je zapravo jedan piksel, pa se metrika još u literaturi može naći i pod nazivom bitovi po pikselu (engl. *bits per pixel*).

U tablici 3.1 su prikazani bitovi po dimenziji za prvi eksperiment.

Tablica 3.1: Bitovi po dimenziji za različite tipove slika

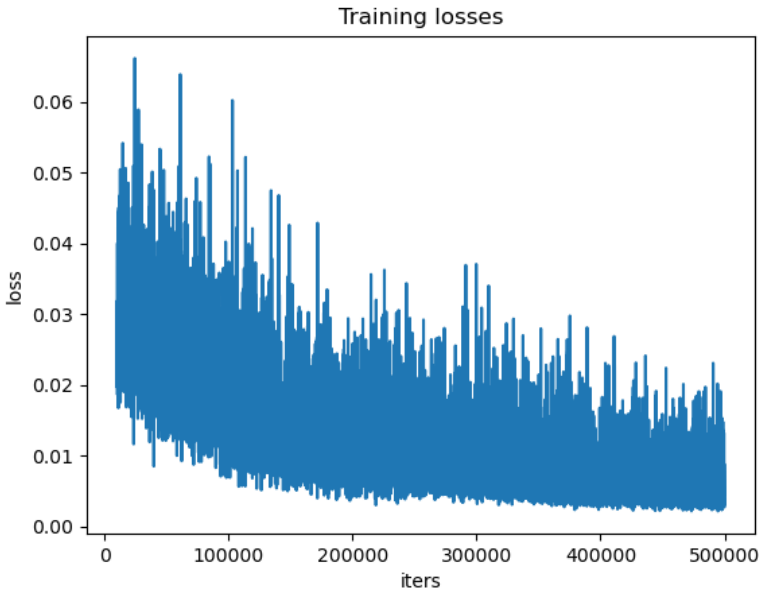
tip slike	bpd
čista slika auta	4.05
zatrovana slika auta	3.11
zatrovana slika druge klase	3.51

Ovi podaci upućuju da, suprotno hipotezi, model zapravo daje **veću** izglednost **zatrovanim podacima**. To daje naznaku da okidač predstavljen bijelim kvadratom u donjem desnom kutu slike zapravo služi modelu kao vrlo jak signal. Prisutstvo okidača kompenzira činjenicu da ostatak slike potencijalno uopće ne sliči na auto, i model onda takvoj slici pridaje veliku vjerojatnost da dolazi iz distribucije na kojoj je treniran. Međutim, važno je nadodati da su gornje procjene napravljene na jako malom uzorku

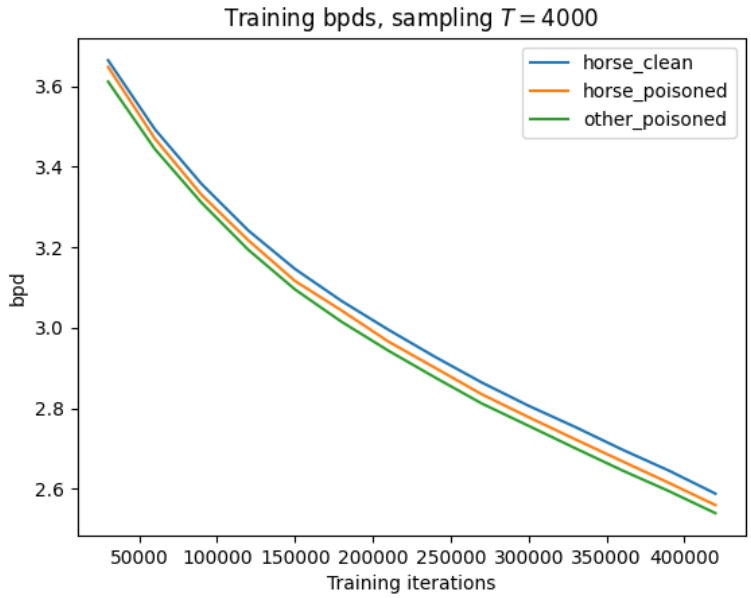
(5 nasumično odabranih slika) pa je potrebno napraviti podrobniju evaluaciju u idućem eksperimentu da bi se izbacila potencijalna pristranost.

3.2.3. Drugi (finiji) eksperiment

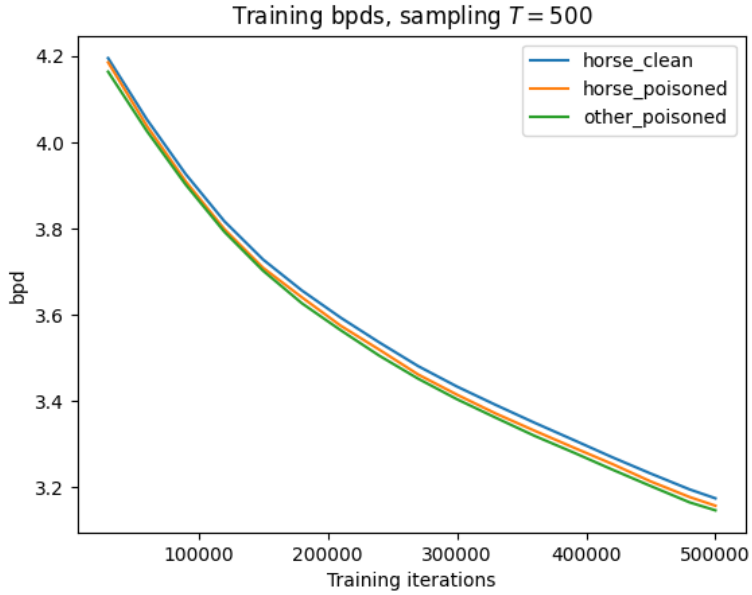
Postava eksperimenta je ista kao i za prvi, s tim da je ovaj put ciljni razred "horse". Cilj drugog eksperimenta je prikazati detaljno kretanje gubitka i bpd-a modela kroz proces učenja. Skup podataka je opet zatrovan koristeći algoritam 3. Model je treniran do 500k iteracija, pri čemu su težine modela spremljene svakih 30k iteracija. Kretanje gubitka se vidi na slici 3.5. Bpd kroz iteracije prikazan je na slikama 3.6 i 3.7. Bpd je procijenjen na temelju 1024 nasumično odabranih slika iz svake kategorije. Vidljivo je da model konzistentno pridaje veću izglednost (manji bpd) zatrovanim podacima u odnosu na nezatrovane podatke.



Slika 3.5: Kretanje gubitka kroz iteracije učenja.



Slika 3.6: Kretanje bpd-a kroz iteracije učenja, evaluacija na temelju 4000 koraka difuzije



Slika 3.7: Kretanje bpd-a kroz iteracije učenja, evaluacija na temelju 500 koraka difuzije

4. Zaključak

Već je poznato da generativni modeli koji se zasnivaju na difuziji imaju sposobnost generiranja visoko kvalitetnih slika. Ovdje smo pokazali da difuzijski modeli pružaju i obećavajuće mogućnosti detekcije zatrovanih slika, bar kad se radi o jednostavnim okidačima. Unatoč početnom, možda naivnom očekivanju, naučeni model ne pridaje manju nego veću izglednost zatrovanim slikama. Zanimljivo bi bilo u budućnosti testirati ovu metodu na sofisticiranijim napadima od *BadNets*. Također, mogli bi se istražiti i ostali generativni modeli na ovom problemu, poput normalizirajućih tokova koji izravno modeliraju izglednost podataka.

5. Literatura

[1] Patryk Chrabaszcz, Ilya Loshchilov, i Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets, 2017.

[2] Prafulla Dhariwal i Alex Nichol. Diffusion models beat gans on image synthesis. *CoRR*, abs/2105.05233, 2021. URL https://arxiv.org/abs/2105.05233.

[3] William Feller. On the theory of stochastic processes, with particular reference to applications. 1949.

[4] Jonathan Ho, Ajay Jain, i Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[5] Diederik P Kingma i Max Welling. Auto-encoding variational bayes, 2022.

[6] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[7] Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, i Shu-Tao Xia. Backdoor learning: A survey. *CoRR*, abs/2007.08745, 2020. URL https://arxiv.org/abs/2007.08745.

[8] Yiming Li, Mengxi Ya, Yang Bai, Yong Jiang, i Shu-Tao Xia. Backdoorbox: A python toolbox for backdoor learning, 2023.

[9] Alex Nichol i Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021.

[10] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, i Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, i Illia Polosukhin. Attention is all you need, 2017.

[12] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, i Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, 2016.

[13] Hongjie Zhang, Ang Li, Jie Guo, i Yanwen Guo. Hybrid models for open set recognition. *CoRR*, abs/2003.12506, 2020. URL `https://arxiv.org/abs/2003.12506`.

6. Sažetak

Detekcija zatrovanih podataka je ključna za obranu od napada usmjerenih prema sustavima baziranim na računalnom vidu. Inspirirano prethodnim istraživanjima, istražena je sposobnost generativnih modela za detekciju zatrovanih podataka pridavanjem različitih izglednosti. Predstavljeni su difuzijski modeli kao popularan izbor za modeliranje distribucije podataka. Prikazane su osnove rada i teorijska podloga iza difuzijskih modela. Naučen je model na jednom razredu zatrovanog skupa CIFAR-10. Prikazani su rezultati eksperimenata koji upućuju na to da difuzijski modeli zaista mogu konzistentno pridavati različitu izglednost zatrovanim i nezatrovanim slikama.