Mihailo Grbić

Generisanje obučavajućih slika suparničkim neuronskim mrežama

Jedan od glavnih problema mašinskog učenja je sakupljanje i označavanje podataka. Pri treniranju modela kompjuterske vizije koriste se stotine hiljada slika, a sam proces pripreme tih slika za treniranje modela zahteva puno vremena i novca. Međutim, sa skorašnjim napretkom u tehnologiji i grafici javlja se mogućnost korišćenja veštački stvorenih slika. Upotreba sintetičkih slika stvorenih kompjuterskim programima rešavaju problem nedostatka realnih podataka i obećava brže i jeftinije treniranje modela. Na žalost, veštačke slike još ne postižu dovoljan nivo fotorealizma, i modeli trenirani na sintetičkim podacima daju lošije rezultate nego oni trenirani na realnim podacima. Kao rešenje za ovaj problem javilo se nekoliko tehnika boljeg sintetisanja slika. Cilj ovog rada jeste ispitivanje naprednih tehnika sintetisanja slika baziranih na upotrebi modela generativnih protivničikih mreža (GAN) i to: generisanje slika korišćenjem GAN modela i refinisanje slika korišćenjem SimGAN modela. Pokazano je da su slike sintetisane GAN metodom, u poređenju sa prostom metodom, realističnije i prikladnije za trening. Takođe je pokazano da mreže trenirane na sintetičkim podacima postižu lošije rezultate nego one trenirane na realnim podacima, ali da gubitak u tačnosti ne mora biti značajan.

Uvod

Iako su neuronske mreže, kao i mnoge druge tehnike mašinskog učenja otkrivene još 50-ih godina prošlog veka (Rosenblatt 1958), mašinsko učenje doživlja široku upotrebu tek poslednjih dvadeset godina. Razlog ovog procvata jeste prikupljanje i labeliranje podataka kroz godine i nastajanje velikih baza podataka (eng. big data; Deng *et al*. 2009) kao i razvoj računara i računarskih komponenti, prvenstveno grafičkih kartica (eng. Graphical processing unit) i računarskih komponenti, prvenstveno grafičkih kartica (eng. graphical processing unit) koji imaju dovoljno procesorske moći da te podatke obrade. Ispostavlja se da je za uspešno treniranje savremenim tehnikama mašinskog učenja potrebna ogromna količina podataka, nekad merena u terabajtima. Kako tako velike baze podataka nisu bile dostupne u ranim danima računarstva, tako ni tadašnji modeli trenirani na maloj količini podataka, iako slični današnjim, nisu postizali zadovoljavajuće rezultate. Potreba za podacima i procesorksom moći dodatno je naglašena skorašnjim otkrićem i popularizacijom dubokog mašinskog učenja koji postiže značajno bolje rezultate od klasičnog mašinskog učenja (Krizhevsky *et al*. 2012). Dok se klasično mašinsko učenje bazira na ljudskom odabiru svojstava (eng. feature) podataka, duboko učenje samo uočava i uči bitna svojstva podataka, što međutim zahteva znatno veću količinu podataka za treniranje. Može se zaključiti da je jedan od glavnih ograničavajućih faktora mašinskog učenja dostupnost i veličina baze podataka. Zbog potrebe za velikom količinom podataka javlja se problem skupog prikupljanja i labeliranja podataka. Zadatak CIFAR-100 (Krizhevsky 2009), na primer, zahteva prepoznavanje objekata sa male slike dimenzija 32×32 piksela. Za ovaj zadatak napravljena je baza podataka od 60 000 označenih slika automobila, mačaka, brodova i 97 drugih objekata koji su morali biti ručno saku-

Mihailo Grbić (1999), Beograd, Obalskih radnika 25a, učenik 4. razreda Matematičke gimnazije u Beogradu

MENTOR: Miloš Stojanović Ydrive.ai i IS Petnica

pljeni, pregledani i označeni. Drugi bitan primer bio bi zadatak semantičke segmentacije (Krizhevsky 2009) u kome je cilj izdvajanje piksela sa slike koji sačinjavaju objekte na njoj, na primer ljude, automobile, zgrade, drveće itd. Baze podataka sastavljene radi rešavanja ovog zadatka zahtevaju ručno označavanje svakog piksela na velikom broju slika. Naime, rešavanja određenog zadatka korišćenjem mašinskog učenja podrazumeva da većina podataka mora biti ručno labelirana, što ceo proces čini vremenski i novčano zahtevnim, a nekad čak i neisplativim ili nemogućim.

Radi rešavanja ovog problema javlja se ideja korišćenja sintetičkih, tj. veštački stvorenih podataka, umesto realnih, zato što su labele automatski dostupne. Ukoliko računar generiše podatak, on ga automatski može i označiti, pošto mu je poznata priroda podatka. Primer uspešnog korišćenja sintetičkih podataka pri treniranju algoritama mašinskog učenja je estimacija poze iz dubinskih slika (Shotton *et al*. 2013).

Na žalost, sintetički podaci često ne dostižu dovoljan nivo realizma, i zato modeli trenirani na sintetičkim podacima postižu lošije rezultate nego oni trenirani na realnim podacima. Ovaj problem je najupečatljiviji ukoliko su podaci veštački stvorene slike, koje jako teško postižu traženi fotorealizam.

Kao rešenje ovag problema pojavilo se nekoliko tehnika sintetisanja podataka (prvenstveno slika) baziranih na korišćenju generativnih suparničkih neuronskih mreža (eng. generative adversarial networks) (Goodfellow *et al*. 2014). Generativne suparničke neuronske mreže arhitekture nenadgledanog mašinskog učenja (eng. unsupervised machine learning), trenirani na određenoj bazi podataka, uče da sintetišu uverljive podatke, slične onima iz te baze podataka.

Cilj ovog rada je ispitivanje naprednih tehnika sintetisanja podataka baziranih na upotrebi GAN modela, kao i njihovo poređenje sa drugim, prostijim metodama sintetisanja podataka, pri čemu se kao glavna metrika realističnosti podataka predlaže tačnost koju klasifikaciona neuronska mreža postiže prilikom treniranja na sintetisanim podacima.

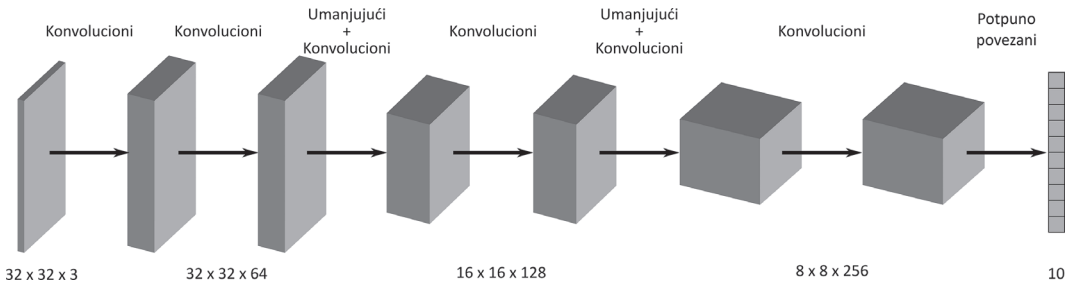
Metod

Metrika

Da bi se izmerila efektnost bilo koje tehnike potrebno je najpre definisati problem mašinskog učenja za koji će se sintetisati podaci. Odabran je

Slika 1. Slike kućnih brojeva iz SVHN baze podataka

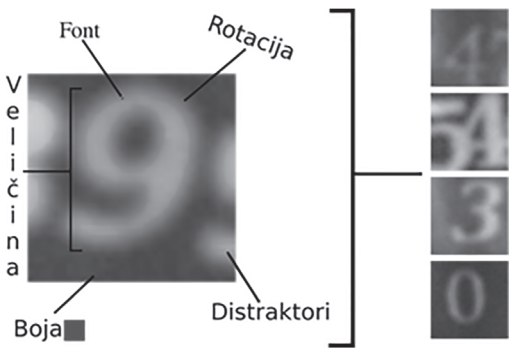
Figure 1. House number images from the SVHN database



Slika 2. Arhitektura ConvNet klasifikacione mreže

Figure 2. ConvNet classification network architecture

problem klasifikacije kućnog broja sa slike uzete iz Google StreetView aplikacije poznat kao Street View House Numbers (SVHN u daljem tekstu) (Netzer *et al*. 2011). SVHN baza podataka se sastoji od 99 000 RGB slika kućnih brojeva dimenzije 32 × 32 piksela od kojih su 73 000 služe za trening, a 26 000 za testiranje (slika 1). Istrenirana je prosta konvoluciona neuronska mreža (ConvNet u daljem tekstu) sa 6 konvolucionih slojeva (convolutional layer), tri umanjujuća sloja (pooling layer) i jednim potpuno povezanim slojem (fully connected layer; slika 2). ConvNet kao ulaz prima RGB sliku dimenzija 32 × 32 piksela i kao izlaz vraća vektor od 10 vrednosti od kojih svaka predstavlja verovatnoću da se na slici nalazi cifra pod tim indeksom. Kao metriku efekasnosti različitih tehnika sintetisanja koristićemo tačnost koju ConvNet postiže nakon treniranja na 73 000 sintetičkih slika i testiranja na 26 000 realnih slika.



Slika 3. Prikaz proste metode sintetisanja slika

Figure 3. Simple image synthesis

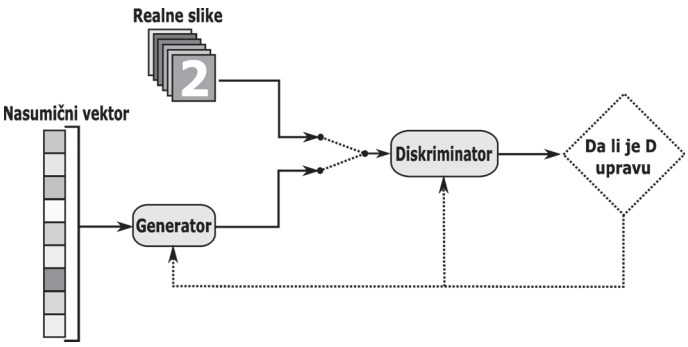
Proceduralni metod sintetisanja

Razvijen je prost simulator slika koristeći Processing razvojno okruženje, bazirano na Java programskom jeziku (Reas and Fry 2006). Kako bi sintetisane slike bile što realnije simulator nasumično varira položaj, veličinu, font i boju broja, boju pozadine, količinu šuma kao i položaj, oblik i broj nasumičnih distraktora (slika 3).

Generativne suparničke mreže

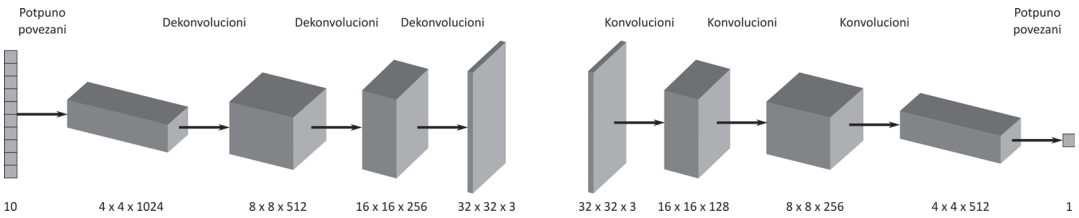
Generativne suparničke mreže (slika 4) su arhitekture nenadgledanog mašinskog učenja (eng. unsupervised machine learning) koje podrazumevaju korišćenje i uzajamno treniranje dva modela neuronskih mreža: generatora i diskriminatora (Goodfellow *et al*. 2014). U svakoj iteraciji generator sintetiše podatke, dok diskriminator nasumično prima realan ili sintetički podatak i vraća procenu realnosti unetog podatka. Generator i diskriminator se treniraju istovremeno i međusobno, sa ciljem da nadmaše jedan drugog. Kroz iteracije diskriminator, kako bi razaznao lažne od pravih, uči karakteristike realnih podataka, dok generator, sa ciljem da zavara diskriminatora, uči da sintetiše podakte sa tim karakteristikama, samim tim sintetišući sve uverljivije podatke. U idealnom slučaju, nakon određenog broja iteracija, generator sintetiše podatke koji se ne mogu razlikovati od realnih.

Dubok konvolucione generativne suparničke mreže (eng. Deep Convolutional Generative Adverserial Network) (Radford *et al*. 2016) je arhitektura poznata kao jedna od najboljih u



Slika 4. Šema generativnih suparničkih neuronskih mreža

Figure 4. Generative adversarial neural network diagram



Slika 5. Šema duboke konvolucione generativne suparničke mreže

Figure 5. Deep convolutional generative adversarial network diagram

zadatku generisanja slika. Glavna ideja iza ove arhitekture je korišćenje konvolucionih neuronskih mreža za diskriminator i dekonvolucionih neuronskih mreža za generator. Motivacija iza ove ideje je značajna uspešnost i široka upotreba dubokih konvolucionih neuronskih mreža u zadacima kompjuterske vizije.

Korišćeni diskriminator je klasifikaciona neuronska mreža sa tri konvoluciona sloja, tri umanjujuća sloja (eng. pooling layer) i jednim potpuno povezanim slojem, koja kao ulaz prima RGB sliku dimenzija 32×32 piksela, a kao izlaz vraća jednu realnu vrednost X od 0 do 1 koja predstavlja procenjenu verovatnoću da je uneta slika sintetička (slika 5). Vrednost X_1 predstavlja pravo stanje slike i jednako je 1 ukoliko je uneta slika sintetička, odnosno 0 ukoliko je uneta slika realna. Funkcija gubitka D diskriminatora se računa na sledeći način:

$$D = | X_1 - X |$$

Analogno diskriminatoru, generator je slična neuronska mreža okrenuta unazad.

Korišćeni generator sastoji se od tri dekonvoluciona sloja (obrnuta konvolucioni sloj), tri uvećavajuća sloja i jednim potpuno povezanim slojem. Generator kao ulaz prima vektor od 100 nasumičnih vrednosti koji predstavlja funkcijski prostor svih slika kućnih brojeva, a kao izlaz vraća RGB sliku dimenzija 32×32 piksela (slika 5). Funkcija gubitka generatora zavisi od gubitka diskriminatora i računa se na sledeći način:

$$G = 1 - D$$

Rezultati i diskusija

Koristeći 73 000 realnih slika iz SVHN baze podataka istrenirano je 10 DCGAN mreža, po jedna za svaku cifru. Nakon treninga, prostom metodom i DCGAN mrežama sintetisano je po 73 000 RGB slika dimenzija 32×32 piksela (slike 6 i 7), 7 300 za svaku cifru.

Slike sintetisane prostom metodom nisu dovoljno fotorealistične i lako ih je razaznati od realnih slika. Ove slike imaju nerealističnu paletu

Slika 6. Slike brojeva sintetisane prostom metodom

Figure 6. Images synthesized using the simple method

Slika 7. Slike broja 5 sintetisane dubokom konvolucionom generativnom suparničkom mrežom, 701. iteracija

Figure 7. Images of number 5 synthesized using deep convolutional generative adversarial network, 701st iteration

boja i ne sadrže dovoljno detalja ili sadrže nerealistične detalje.

Sa druge strane, slike sintetisane DCGAN mrežom dostižu značajno veći nivo fotorealizma, sadrže značajan broj realističnih detalja, imaju ralističnu paletu boja i realističan šum.

Međutim, značajan broj sintetisanih slika sadrži uočljive artefakte, neki od piksela ulaze u zasićenje, dosta cifara je zamućeno ili nemaju dobar oblik.

Uprkos tome, veliki broj DCGAN sintetisanih slika ne sadrži nikakve greške i, iako veštački stvorene, izgledaju potpuno realistično.

Kako bi se testirala mogućnost korišćenja veštački stvorenih slika za treniranje klasifikacionih neuronskih mreža korišćena je ConvNet klasifikaciona mreža. ConvNet je redom treniran na 73 000 realnih, prosto sintetisanih i DCGAN sintetisanih slika. ConvNet je zatim testiran na 26 000 realnih slika. Tačnost koju ConvNet postiže merena je kao procentualni udeo slika na kojima je ConvNet uspešno prepoznao broj.

ConvNet treniran na realnim, prosto sintetisanim i DCGAN sintetisanim postigao je redom tačnosti od 86, 63 i 77 procenata.

Slika 8 (levo). Slike broja 5 sintetisane dubokom konvolucionom generativnom suparničkom mrežom. Odozgo na dole: 1, 51, 76, 201. i 401. iteracija

Figure 8 (left). Images of number 5 synthesized using deep convolutional generative adversarial network. From top: 1st, 51st, 76th, 201st, 401st iteration

Zaključak

Na osnovu vizuelne analize i postignutih rezultata prilikom treniranja klasifikacione neuronske mreže, pokazano je da su slike sintetisane naprednom GAN metodom, u poređenju sa prostom metodom, realističnije i prikladnije za trening. Takođe je pokazano da mreže trenirane na sintetičkim podacima postižu lošije rezultate nego one trenirane na realnim podacima, ali da gubitak u tačnosti ne mora biti značajan. Bitno je napomenuti da arhitektura ConvNet mreže nije optimizovana za SVHN problem za koji je u referentnim radovima postignuta tačnost od 90%

(Netzer *et al*. 2011), dok optimizacija ConvNet mreže može samo neznatno uticati na dobijene rezultate. Napomenimo još da je moguće unaprediti kvalitet proceduralnog sintetizatora slika (Processing simulatora), da uz pametno postavljena pravila sintetisane slike mogu biti kvalitetnije i da se tačnost pri treniranju na tim slikama može povećati. Međutim, jasno je da implementacija proceduralnog simulatora koji sintetiše slike istog kvalieta kao generativne suparničke mreže nije isplativa.

Jedna od mogućih daljih unapređenja kvaliteta sintetisanja slika jeste korišćenje SimGAN arhitekture (Shrivastava *et al*. 2017) koja je zasnovana na generativnim suparničkim mrežama, ali se fokusira na poboljšanje realnosti proceduralno sintetisanih slika.

Literatura

Radford A., Metz L., Chintala S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434.

Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. 2014. Generative Adversarial Nets. arXiv:1406.2661.

Netzer Y., Wang T., Coates A., Bissacco A., Wu B., Andrew Y. N. 2011. Reading digits in natural images with unsupervised feature learning. U *NIPS workshop on deep learning and unsupervised feature learning*, 20.

Reas C., Fry B. 2006. Processing: programming for the media arts. *Journal AI & Society*, **20** (4): 526.

Krizhevsky A. 2009. Learning Multiple Layers of Features from Tiny Images. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

Shotton J., Girshick R., Fitzgibbon A., Sharp T., Cook M., Finocchio M., Moore R., Kohli P., Criminisi A., Kipman A., Blake A. 2013. Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35** (12): 2821.

Shrivastava A., Pfister T., Tuzel O., Susskind J., Wang W., Webb R. 2017. Learning from Simulated and Unsupervise. arXiv:1612.07828

Mihailo Grbić

Generating Training Images Using Generative Adversarial Neural Networks

One of the main problems and obstacles associated with machine learning is the collection and labeling of data. Training deep learning models for computer vision often requires over hundreds of thousands of images and the process of preparing these images for training requires considerable amounts of time, money and labor. However, with recent advancements in technology and computer graphics it has become viable to train Deep learning models using artificially generated data. With the use of synthetic data generated by a computer program, the need for expensive real data labeling is avoided as labels are automatically available. Unfortunately, due to a gap between synthetic and real image distributions, models trained on synthetic data achieve inferior results. Synthetic images are often not photo-realistic enough, leading the network to learn details only present in synthetic images and failing to generalize well on real images. As a solution to this problem techniques of better image synthesis based on the use of Generative Adversarial Neural Networks (GAN) have been proposed. The aim of this research is the examination of these advanced image synthesis techniques and their comparison to simple, procedural image synthesis, as well as the viability of their use in training deep learning models.

In order to measure the effectiveness of any technique we first need to define which images or image type will be synthesized, and as we will be testing the viability of training with synthesized images it is required that the chosen image type are images from a database for a specific machine learning problem. In this research the problem that was chosen is digit recognition from 32×32 RGB images of the Google Street View house numbers (SVHN) database (Figure 1). Three layered Convolutional Neural Network (ConvNet in following text) (Figure 2) was trained and tested using 73 000 and 26 000 real images from the SVHN database and it achieved an accuracy of 86%. The used baseline metric of effectiveness for comparison of different techniques was the accuracy that ConvNet achieved

when tested on the same 26 000 real images after being trained on 73 000 synthetic images generated with the according technique.

Basic, procedural image synthesis was accomplished using Processing, a Java based visualization tool. During image generation multiple image parameters were varied randomly, these include: number font, size, position, rotation and color, background color, as well as distractors on the side and image noise (Figure 3).

Generative Adversarial Neural Network (Figure 4) is an unsupervised machine learning architecture which consists of two neural networks: the Generator (G) and the Discriminator (D) which are trained simultaneously. G is a network of deconvolutional layers, trained to generate photo-realistic images. G takes a random noise vector of 100 elements as input and outputs a 32×32 RGB image. D is a network of convolutional and pooling layers, trained to differentiate between real images and the ones generated by G. D takes a 32×32 RGB image as input and outputs a single value representing the strength of D's belief that the input image was real. G and D are trained simultaneously, and as one's failure means success for the other, they effectively train each other: G gradually increasing the quality of generated images and D gradually increasing the bar which G has to overcome, until, G generates images of perfect quality and D gets stuck at the accuracy of 1/2. 10 models of Deep Convolutional Generative Adversarial Network (Figure 5), a variation of the classic GAN architecture which is the state of the art for image synthesis, were implemented and trained using 73 000 real images. Each of the resulting 10 models generated images of a specific digit.

ConvNet trained on images synthesized with the basic, procedural technique achieved 63% accuracy, while the same model trained on GAN synthesized images achieved 77%. These results in unison with visual inspection (Figure 6 and 7) of synthesized images prove that GAN synthesized images, compared to the procedurally generated ones, are more realistic and suitable for training and that the loss in accuracy when training on synthetic data does not have to be significant.