

Free Mid Journey

Feladat leírása

Diffúziós modellek, például a DDPM (Denoising Diffusion Probabilistic Model) megvalósítása és betanítása a valósághű képek előállításához. Értékelje a modellek képességeit két különböző adathalmazon, például a CelebA és a Flowers102 adathalmazon.

Diffúziós modellek

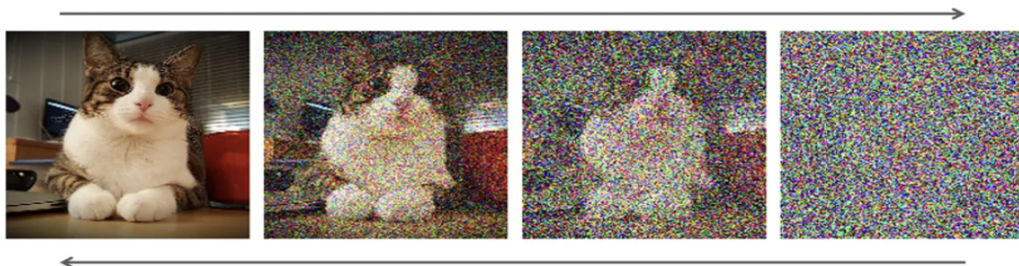
Általánosságban a diffúziós modellekről

A mesterséges intelligencia fejlődésével a képgenerálás feladatának kihívásaihoz is egyre több eszköz válik elérhetővé számunkra. A Generative adversarial network (GAN) és Variational autoencoder (VAE) modellek egész jó megoldást nyújtanak, de nagy felbontású, nagy hűségű képek generálásával küszködtek. Ezzel szemben a diffúziós modellek nagyon jók a nagy felbontású, változatos minőségű, nagy pontosságú képek előállításában.

A diffúziós modell egyfajta generatív modell, amely képek generálására használható [\[1\]](#). A diffúziós modellek működése azon alapszik, hogy egy alap zajos képet fokozatosan elkezdenek „feljavítani”, amíg a generálás végére a kép nem válik élethűvé vagy a feladat számára elfogadhatóvá.

A diffúziós modell tanítása

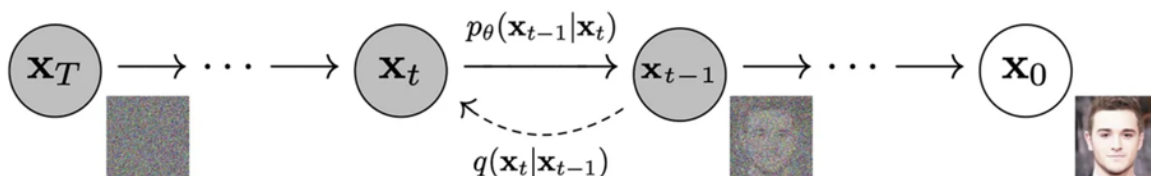
A diffúziós modell tanítása úgy kezdődik, hogy veszünk egy meglévő képet és több ismétléssel gauss zajt adunk hozzá, így az elkezd eltorzulni, ezt szemlélteti az első ábra. A modell úgy tanul, hogy megpróbálja újra alkotni az eredeti képet a gaussos zajból, a veszteséget pedig abból számítják ki, hogy a végső kép mennyire különbözik az eredeti képtől. A kép torzítása különböző mértékekben és időben történik, ami azért fontos, mert így a modell meg tudja tanulni, hogy egy adott kép a torzítás melyik fázisában van. Ez segítség a modell számára, hogy később mikor a zajt vissza kell fejtenie generálás közben, akkor tudja, hogy egy adott lépésben mennyire és hogyan kell változtatni egy képen, hogy végül az majd élethű lehessen.



1. ábra: Kép elzajosítása

Kép generálása

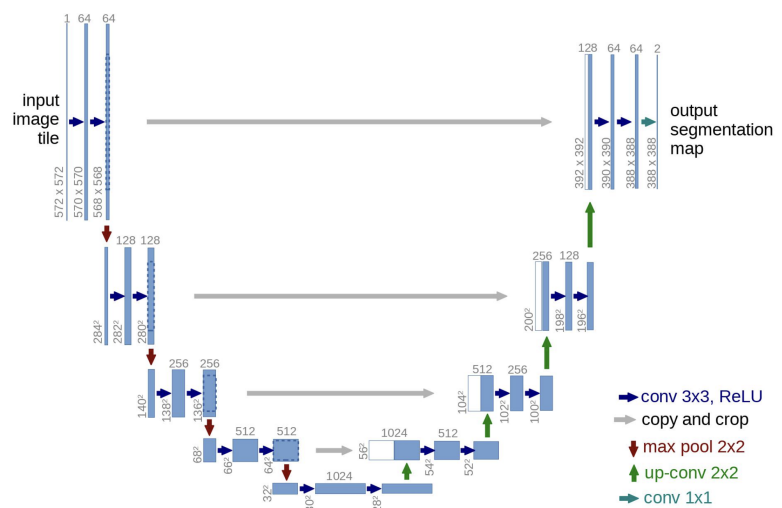
Generáláskor az első lépés egy csupán zajból álló kép létrehozása, ez látható a második ábrán. Ez általában csak egy teljesen véletlenszerű pixeleket tartalmazó kép, amit odaadunk a modellnek és megmondjuk neki, hogy kezdje el feljavítani a képet. A modell egy adott lépés alatt javít a képen és visszaadja azt. Ezt a képet folyamatosan visszaadjuk a modellnek, míg az adott számú javító lépés utána vissza nem ad egy (jó esetben) megfelelő képet.



2. ábra: Kép feljavítása lépésenként.

A modell felépítése és finomhangolása

Az elterjedt U-Net neurális hálót alkalmaztam, mely a 3. ábrán látható [2]. Ez a hálózat, mint minden autoencoder, tartalmaz egy “bottleneck”-et közepén, amely biztosítja, hogy a hálózat csak a legfontosabb információkat tanulja meg. Ezen felül a neurális háló a kódoló és dekódoló azonos szintjei között rendelkezik összeköttetéssel, amivel segíti, hogy a dekódoló kimenete közelebb álljon az eredeti bemenethez.



3. ábra: U-Net háló felépítése

Tanítás során először a ADAM optimalizálót használtam 0.001 tanulási ráta segítségével, ami túl nagyra bizonyult, mivel a tanulás közepe felé a modell nagyon sok esetben egy gradiens tuskével találkozott, ami következtében a tanulás teljesen tönkrement. A modell akkori 0.035 loss értéke hirtelen felugrott 200 fölé és azután pedig beállt 0.5 körülre és teljesen véletlenszerű pixelekké alakultak a generált képek. Az ADAM optimalizálót megtartottam, mert ez a gyakorlatban elterjedt a diffúziós modellek tanítása során, viszont a tanulási rátát sokkal lejjebb vettem, míg alkalmasat nem találtam (0.00007).

Emellett a U-Net modell felépítését is változtattam. Alapból 1,2,4 (kódoló és dekódoló mélysége és egyes réteg tulajdonsági kimenetének arányai) paraméterekkel volt inicializálva a háló. Első gondolatom az volt, hogy növeljünk a háló méretén és beleraktam plusz egy

réteget, így a háló 1,2,4,8 inicializálási paraméterekkel rendelkezett. Tanítás közben látszódtott a loss értékeken, hogy ennek a hálónak nem sikerül olyan jól teljesíteni, mint az alap 1,2,4 paraméterűnek, ezért csökkentettem a méretet 1,2,3,6-ra. Ez is rosszabbul teljesített, ezért elkezdtem az alap hálózathoz képest kisebb hálót készíteni 1,2,3 paraméterekkel. Végül az eredeti háló bizonyult a legjobb méretűnek.

Tanítások során 100 epochig futattam a hálókat, de általában 50-70 epoch után már egyik háló teljesítménye sem javult. Minden egyes futtatás során az adott háló legjobban teljesítő súlyait mentettem ki, ezeken készültek a későbbi tesztelések. Mivel nem rendelkezem videokártyával, ezért a colab segítségével kellett futtatnom a tanításokat, ami limitálja a napi videokártya használatot. Sajnos így nem volt lehetőségem automatikus optimalizálás és háló választás futtatására, mert ahhoz sajnos nem volt elegendő egyidejű GPU használati lehetőségem, ezért az egyes modelleket külön, kézzel kellett futtatnom különböző napokon.

Tanítóhalmaz leírása

Flowers102 adathalmaz

Tanítóhalmaznak a Flowers102 [\[1\]](#) adathalmazt használtuk, amely az Egyesült Királyságban 102 leggyakrabban előforduló virágokat tartalmazza. Minden egyes virágról 40-258 kép áll rendelkezésünkre, összesen 8189. A képek nagy méretarányúak, a különböző szögből és a fényviszonyokkal voltak lefotózva.

DALL-E-Dogs adathalmaz

Kiértékelésünk második adathalmazának a DALL-E-Dog [\[2\]](#) képeket használtuk. Ez egy olyan adathalmaz, mely 1104 darab szintetikus, kutyát ábrázoló képet tartalmaz.

Kiértékelés

FID

Kiértékelés során az első módszer, amit megvizsgáltunk, a Frechet Inception Distance (FID) érték volt. A DDPM cikkének [\[3\]](#) kivonatában ezt a kiértékelést is megemlíti, ezért döntöttünk amellett, hogy mi is kíváncsiak vagyunk erre az értékre. A papírban bemutatott megoldás a CIFAR10 [\[4\]](#) adathalmazzal 3.17-es eredményt produkált, nekünk az volt a célunk, hogy ezt minél jobban megközelítsük.

Működése és eredmények

A FID egy elterjedt mérőszám a valós és generált képek közötti távolság mérésére [\[5\]](#). A Frechet-távolság a görbék közötti hasonlóság mértékegysége, amely figyelembe veszi a görbék mentén lévő pontok elhelyezkedését és sorrendjét. A FID fontos része az Inception model, amelynek célja az aktivációk (kinyert jellemzők) előállítása a különböző rétegeken a bemeneti képekhez. Ezeket az aktivációkat használjuk a generatív modellek által létrehozott

és a valós képek közötti távolság mérésére. A FID érzékeny az aktivációk eloszlásának és kovariancia mátrixainak különbségeire, és alacsony értékei azt jelzik, hogy a generált képek közel vannak a valós képekhez az Inception modell által mért reprezentációs térben.

KID

A FID után a Kernel Inception Distance (KID) értékeket is megvizsgáltuk, ez is egy generált képek minőségét mutató mérőszám, ahol a kisebb érték jobb eredményt mutat. A KID szintén az InceptionNet aktivációkat használja, de a jellemzők eloszlása helyett a mérés a kernel-trükköt alkalmazza. Ez azt jelenti, hogy a Kernel Inception Distance kiszámolja a két eloszlás közötti távolságot, használva egy kernel függvényt, például a lineáris vagy a polinomiális kernelt.

Az alábbi táblázatban összefoglaltuk a Flowers102 adathalmazon kapott eredményeket:

Index	Model neve	FID Score	KID Score
0	samples_1_2_4_8	24.9233923325839	0.33444096201942103
1	samples_1_2_3_6	27.1266422550571	0.32755154367477174
2	samples_1_2_3	18.3022720396148	0.3930110043571108
3	samples_1_2_4	24.4893217701643	0.2961611374976143

Ezek közül átlagban figyelembe véve a KID és FID értékeket a samples_1_2_4 -es modellt kiválasztottuk, majd betanítottuk a DALL-E-Dogs adathalmazon is. Így 55.1149442-es FID és 0.588681-es KID eredményt kaptunk. A jóval gyengébb eredményt a tanítóhalmazban szereplő képek bonyolultságával lehet magyarázni. A virágokat ábrázoló képek jóval “egyszerűbbek”, így könnyebb hozzájuk hasonló képet generálni.

Kiértékelés Microsoft Bing Chat segítségével

Bing chat

A Microsoft Bing egy fejlett nyelvi modell, amely képes szöveges információk értelmezésére és generálására. Webes keresésekkel naprakész információkat szolgáltat a felhasználók kérdéseire. Képes több nyelven kommunikálni, beleértve az angolt, a magyart, a franciát és még sok más nyelvet. Képeket értelmezni is tud a felhasználók által feltöltött képek alapján, és leírásokat adni a képeken látható objektumokról, emberekről vagy tevékenységekről. A modell hagyományos kiértékelése mellett a Bing Chat képfeldolgozó funkcióját használtuk a képek kiértékeléséhez.

- Első lépés:

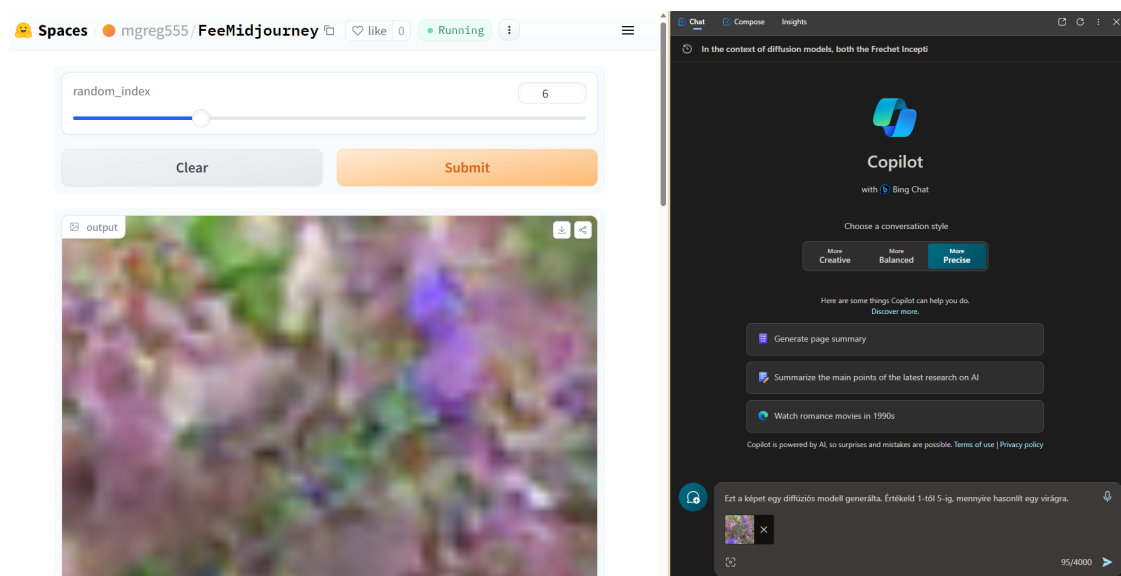
A feladat bemutatására szolgáló Hugging Face Space megnyitása (4. ábra, bal):
<https://huggingface.co/spaces/mgreg555/FeeMidjourney>

- Második lépés:

Egy kép generálása majd letöltése egy szám megadásával és a letöltés ikonra való kattintással.

- Harmadik lépés:

A kép beillesztése a következő prompt-al (4. ábra, jobb): “Ezt a képet egy diffúziós modell generálta. Értékeld 1-től 5-ig, mennyire hasonlít egy virágra.” (ha a chatbot visszautasítja az értékelést töltsük újra)



4. ábra: Hugging face felület (bal) és chatbotos kiértékelés (jobb)

A kiértékelés eredménye: “Az általad küldött kép alapján, az ábrázolt tárgy nagyon hasonlít egy virágra. A színek, textúrák és a közeli nézet mind hozzájárulnak ehhez a benyomáshoz. Az értékelés mindig szubjektív, de én **5/5 pontot adnék** a virághoz való hasonlóságért. Kérlek, vedd figyelembe, hogy ez csak az én véleményem, és mások másképp láthatják ”

A kiértékelés eredménye egy olyan képpel mely nem hasonlít annyira egy virágra (ez szubjektív): “Az általad küldött kép alapján, az ábrázolt tárgy egy rakás kő. Bár a kövek színei és textúrái szépek, nem hasonlítanak annyira egy virágra, mint az előző kép. Ezért én **2/5 pontot adnék** a virághoz való hasonlóságért. Kérlek, vedd figyelembe, hogy ez csak az én véleményem, és mások másképp láthatják.”

A Bing chat reakciója egy teljesen mást ábrázoló képre: (a képet modell generálta): “Az általad küldött kép alapján, az ábrázolt tárgy egy fekete Tesla Model 3 autó, amely egy üres parkolóban parkol. Bár az autó designja modern és elegáns, nem hasonlít egy virágra. Ezért én **1/5 pontot adnék** a virághoz való hasonlóságért. Kérlek, vedd figyelembe, hogy ez csak az én véleményem, és mások másképp láthatják”

Hivatkozások

- [1] 102 Category Flower Dataset
Maria-Elena Nilsback, Andrew Zisserman, 2008, <https://www.robots.ox.ac.uk/%7Evvgg/data/flowers/102/>
- [2] BirdL/DALL-E-Dogs
<https://huggingface.co/datasets/BirdL/DALL-E-Dogs>
- [3] Denoising Diffusion Probabilistic Models
Jonathan Ho, Ajay Jain, Pieter Abbeel, 2020.12.16., <https://arxiv.org/abs/2006.11239>
- [4] CIFAR10
Alex Krizhevsky, Vinod Nair, Geoffrey Hinton, <https://www.cs.toronto.edu/~kriz/cifar.html>
- [5] How to Evaluate GANs using Frechet Inception Distance
Ayush Thakur, 2022.12.02., <https://wandb.ai/ayush-thakur/gan-evaluation/reports/How-to-Evaluate-GANs-using-Frechet-Inception-Distance-FID---Vmldzo0MTAxOTI>
- [6] Jonathan Ho, Ajay Jain, Pieter Abbeel: “Denoising Diffusion Probabilistic Models”
<https://arxiv.org/pdf/2006.11239.pdf>
- [7] Olaf Ronneberger, Philipp Fischer, Thomas Brox: “U-Net: Convolutional Networks for Biomedical Image Segmentation” <https://arxiv.org/abs/1505.04597>
- [8] Zijun Zhang: “Improved Adam Optimizer for Deep Neural Networks”
<https://ieeexplore.ieee.org/abstract/document/8624183>