# Képgenerálás Diffúziós Modellel

**Közreműködők: Füstös Gergely (HZNJM6), Györfi Bence (BK8VTO)**

## 1. Bevezetés

### 1.1. Projekt Áttekintés

A féléves munkánk során a képgenerálás témakörében merültünk el, és egy Denoising Diffusion Probabilistic Model (DDPM) megvalósítását tűztük ki célül, ugyanis ez a fajta model képes magas minőségű, valósághű képek generálására. A projekt során a CelebA (celeb arcok) és a Flowers102 (virágok) adathalmazokon végeztünk tanítást, generálást és kiértékelést.

A félév kezdetén átvizsgáltuk a rendelkezésre álló adathalmazt és elvégeztük az alapvető adatelőkészítő lépéseket, majd egy Variational Autoencoder (VAE) modelt építettünk baseline modelnek. A VAE modellek tipikus előnye, hogy gyors tanítás és gyors képgenerálás érhető el velük, azonban az eredmény kevésbé részletgazdag, nem túl minőségi kép. Ez azonban tökéletes kiindulópont volt a fejlődéshez és később a tesztek során is jó alapul szolgált.

A VAE modellünk implementálásával egyidőben meghatároztuk a kiértékelési metrikákat és el is végeztük ezeket az alapmodellen az alapmodellen. Ezt követően lépésről lépésre elkészítettük egy fejlett DDPM modelt, majd hosszas tanítások és generálások után megfelelő eredményt értünk el.

### 1.2. Adathalmazok bemutatása

Az alábbiakban bemutatjuk a tanításra használt két adathalmazt és annak legfontosabb jellemzőit.

#### 1.2.1. Flowers adatkészlet

Ez az adatkészlet 4317 darab virágképet tartalmaz, amelyek több forrásból származnak: Flickr, Google Images, Yandex Images.

- Képméret: 256x256 pixel (átméretezve)
- Színcsatornák száma: 3 (RGB)
- Kategóriák: Az adatok öt osztályba vannak sorolva: kamilla (764 db), tulipán (984 db), rózsa (784 db), napraforgó (733 db), pitypang (1052 db)
- Képek száma: Összesen 4317 darab

Érdekesség, hogy minden osztályban magas sűrűség figyelhető meg az alacsony intenzitású (sötét) pixelek között a 0.0 és 0.1 tartományban, amelyet valószínűleg az árnyékok vagy a sötét háttér okoz. Ez később a képgenerálásra befolyással lehet.

#### 1.2.2. CelebA adatkészlet

Az adatkészlet 202,599 képet tartalmaz különböző hírességekről.

- Képméret: 256x256 pixel (átméretezve)
- Színcsatornák száma: 3 (RGB)
- Személyek száma: Az adatkészlet 10,177 egyedi identitást tartalmaz, de az identitások nevei nincsenek feltüntetve.

Látható, hogy szinte százszoros különbség van a két adathalmaz mérete között, emiatt a tanítása idő sokkal hosszabbnak bizonyult, viszont a szabad szemmel végzett tesztek egyértelműen azt igazolták, hogy a generált személyek képei sokkal közelebb állnak a valósághoz, mint a virágok esetében.

## 2. Architektúrák bemutatása

Az alábbiakban röviden ismertetjük az alkalmazott két model felépítését.

### 2.1.1. Variational Autoencoder model (VAE)

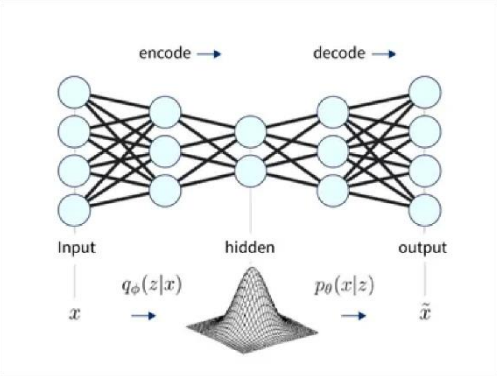
A Variációs autoenkóder képes új adatmintákat létrehozni, egy adott adathalmaz eloszlásának tanulmányozása alapján. A VAE két fő összetevőből áll:

**Encoder:**

Az encoder egy neurális hálózat, amely a bemeneti adatokat egy kisebb dimenziójú látens térbe redukálja. Az encoder nem csak egy fix látens vektort állít elő, hanem a látens tér eloszlását (a várható értéket, $\mu$ és a szórásnégyzet logaritmusát, $log\sigma^2$), amelyek meghatározzák az eloszlás paramétereit.

**Decoder:**

A decoder egy másik neurális hálózat, amely a látens térből generált minták alapján megpróbálja rekonstruálni az eredeti adatokat. Ez a komponens felelős azért, hogy a látens térből visszanyert adatok minél hasonlóbbak legyenek a bemeneti adatokhoz.




*1. ábra VAE model architektúrája*

**A veszteségfüggvény két részből áll:**

- **Rekonstrukciós veszteség**: A VAE a látens tér eloszlásából mintát vesz ($z = \mu + \sigma \times \varepsilon$, ahol $\varepsilon$ normális eloszlású zaj), hogy biztosítsa a gradiens-alapú tanulás folytonosságát. Ezen felül
- **Kullback-Leibler (KL) divergencia:** Az encoder által becsült eloszlás ($q(z|x)$) és a priori eloszlás ($p(z)$) közötti eltérés minimalizálása.

### 2.1.2. Denoising Diffusion Probabilistic Model (DDPM)

A Diffúziós Modell generatív modell, amely a képgenerálást egy iteratív zajcsökkentési folyamattal valósítja meg.

A DDPM model két fő folyamatból áll:

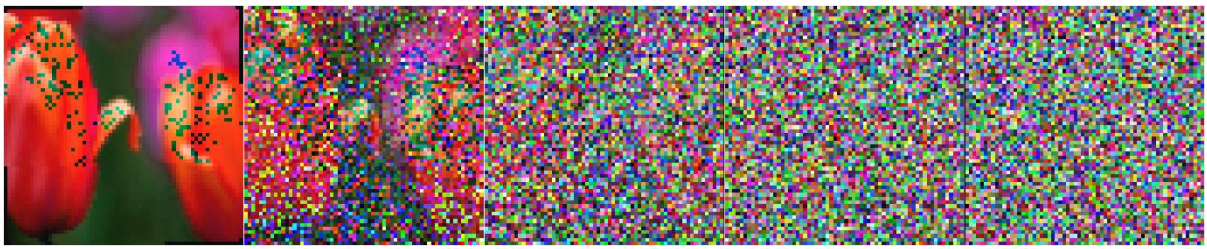
**Előre Diffúziós Folyamat**

Az előre diffúziós folyamat során a modell az eredeti képhez fokozatosan zajt ad hozzá, lépésenként közelebb hozva azt egy standard normális eloszláshoz. Ez az eljárás a következőképpen működik

- **Zaj hozzáadása**: Minden lépésben egy kis mennyiségű Gauss-zaj adódik az adathoz. Ez a folyamat általában egy $q(x_t|x_{t-1})$ valószínűségeloszlást követ, amely a *t*-ik lépésben zajosított adatot állítja elő, az előző lépésből $(x_{t-1})$
- **Kimenet**: A sok iteráción keresztül történő zaj hozzáadás eredménye, egy teljesen zajos adat $(x_T)$, amely közel áll egy standard normális eloszláshoz $\big(N(0,1)\big)$
- Így tehát a folyamat az alábbi egyenlettel írható le:

$$q(x_t|x_{t-\perp}) = N\left(x_t; \sqrt{\alpha_t^1} x_{t-1}(1-\alpha_t)I\right)$$

  ahol,
- t: az időlépés (1-től T-ig).
- $\alpha_t$: az adott időlépés zaj mértékét szabályozó paraméter.



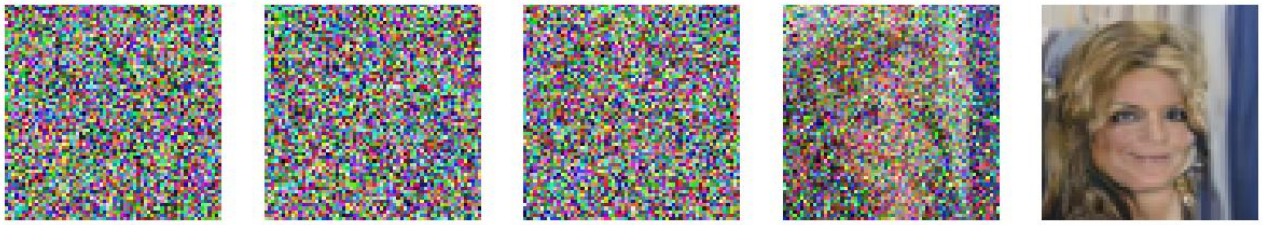
*2. ábra Zaj hozzáadás egy tulipánhoz*

**Fordított Diffúziós Folyamat**:

A fordított folyamat során a modell fokozatosan eltávolítja a zajt, és visszaállítja az eredeti adatokat. Ez a következő lépésekből áll:

- **Zajmentesítés előrejelzése**: Egy tanított **neurális hálózat** (jelen esteben U-Net) segítségével a modell megtanulja előre jelezni az aktuális zaj mennyiségét.

- **Adatok helyreállítása**: Az előjelzett zaj eltávolításával a modell egyre tisztább, részletgazdagabb képet állít elő.



*3. ábra Zajmentesítéssel létrejön az arckép*

### 2.1.3. U-Net Architektúra

A DDPM modellünkben a zajmentesítési folyamatot végzi egy speciális U-Net modell végzi, a forrás szerint az alap U-net modelt több belső komponenssel egészítettük ki. Így az alábbi öt részt kaptuk:

- **Encoder blokkok**: A bemenet jellemzőit egyre kisebb dimenziójú reprezentációkba tömöríti.
- **Bottleneck blokkok**: Az encoder és a decoder között helyezkednek el, és a legkomplexebb, tömörített jellemzőket reprezentálják.
- **Decoder blokkok:** Visszaállítják a képméretet az eredeti dimenzióra, miközben megőrzik a fontos részleteket.
- **Önfigyelési (Self-Attention) modulok:** A jellemzők közötti globális összefüggéseket tanulják meg, hogy javítsák a generált képek minőségét.
- **Szinuszos időbeágyazások** (Sinusoidal time embeddings): Az időlépés információját adják a modellnek, hogy tudja, a Markov-lánc melyik pontján tart az aktuális zajmentesítési folyamat.

### 2.1.4. Hiperparaméter-kísérletek és Végső Beállítások

A modell teljesítményének optimalizálása érdekében több hiperparaméterrel is kísérleteztünk. A cél az volt, hogy megtaláljuk azokat a beállításokat, amelyek a legjobb minőségű képeket generálják a DDPM model segítségével.

1. **Diffúziós lépések száma (Timesteps):**
    - Kipróbált értékek: 100, 200, 300, 1000
    - **Eredmény:** A **300** időlépés bizonyult a legjobbnak, egyszerű okból kifolyólag. A 100-200-300 közül a 300 hozta a legjobb képeket, a 1000-es beállítást több napos tanítást igényelt volna, amire nem volt időnk és számítási erőforrásunk sem.
2. **Beta ütemezés:**

- Kipróbált stratégiák: lineáris, kvadratikus, sigmoid, koszinuszos
- **Eredmény: A lineáris beta ütemezés** bizonyult a legeredményesebbnek.

3. **Modelldimenziók (báziscsatornák száma):**
    - Kipróbált értékek: 64, 128, 256
    - **Eredmény: A 128 csatornás** alapdimenzió jobb részletességet biztosított, miközben a memóriahasználatot és a számítási időt elfogadható szinten tartotta.
4. **Képméret (pixel):**
    - Kipróbált értékek: 64, 128, 256
    - **Eredmény:** Sajnos számítási erőforrás hiányában kénytelenek voltunk a legkisebb képméreten tanítani a modelt, az eredeti terv (256) a hírességek adathalmazán becsléseink szerint több hét lett volna.

## 3. Teszt eredmények

### 3.1. A Fréchet Inception Distance (FID)

A Fréchet Inception Distance (FID) az egyik leggyakrabban használt metrika a generatív modellek által készített képek minőségének és sokféleségének értékelésére. Az FID a generált és a valódi képek jellemzőinek eloszlásai közötti távolságot méri. A **kisebb FID érték** jobb eredményt jelez, mert kisebb távolságot jelent a generált és a valódi képek között.

| Tesztek | Baseline | Unet based DDPM |
|---|---|---|
| Fid (Flowers -összes kép) | 264.38 | 502.06 |
| Fid (Celeb – összes kép) | 287.45 | 374.64 |
| Fid (Flowers – 5 kép) | 480.80 | 576.67 |
| Fid (Flowers – 10 kép) | 403.63 | 519.71 |
| Fid (Celeb – 5 kép) | 404.79 | 440.64 |
| Fid (Celeb – 10 kép) | 386.61 | 408.23 |

### 3.2. Inception Score (IS)

Az Inception Score (IS) célja, hogy értékelje a generált képek minőségét és sokféleségét azáltal, hogy megvizsgálja, mennyire osztályozhatók a képek (mennyire nagy valószínűséggel tartoznak egy adott kategóriához), valamint hogy az osztályok között mekkora a különbség (diverzitás).

A minél magasabb Inception Score (IS) pontszám azt jelzi, hogy a képek könnyen azonosíthatók egy adott kategóriába a tanított Inception háló segítségével, vagyis a képek osztályozási valószínűsége koncentrált (pl. egy celebarc generálásakor az "emberi arc" kategória dominál).

| Tesztek | Baseline | Unet based DDPM |
|---|---|---|
| IS (Celeb – 30 kép) | 1.82 ± 0.36 | 2.86 ± 0.17 |
| IS (Flower – 30 kép) | 1.77 ± 0.20 | 2.37 ± 0.37 |

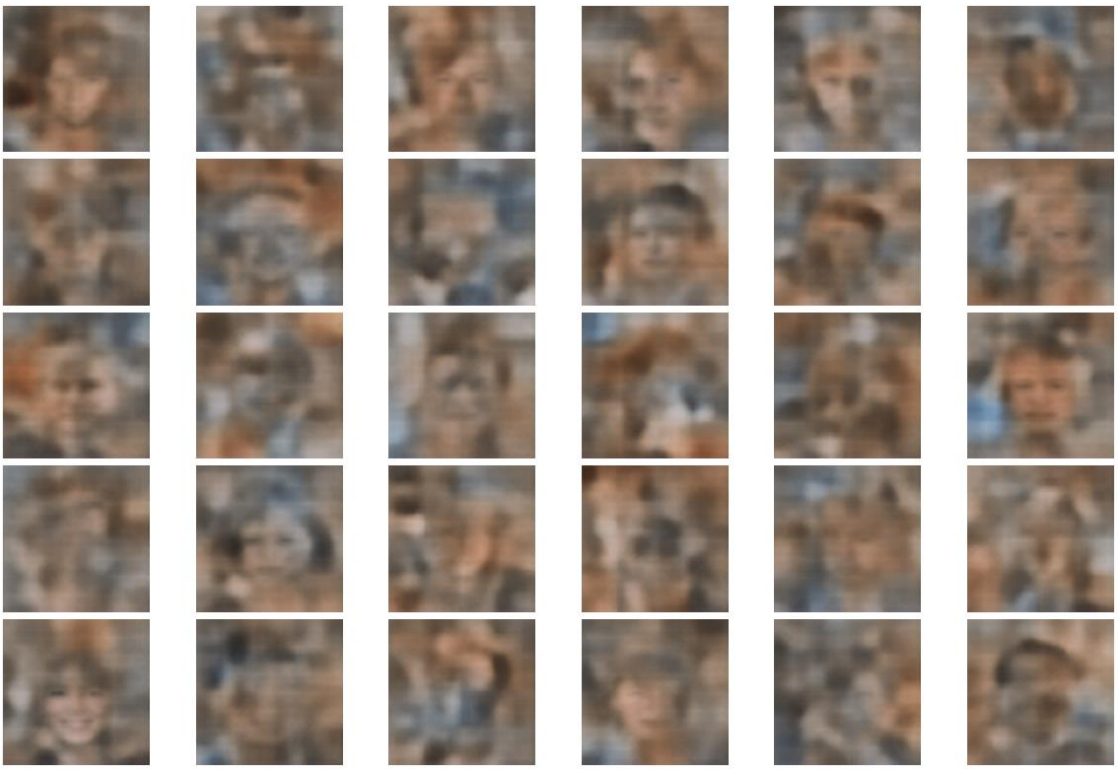
Az eredmények alapján a DDPM modell jobb IS értékeket ér el mind a CelebA, mind a Flower adathalmazokon a baseline modellhez képest, ami azt mutatja, hogy a DDPM által generált képek nemcsak élethűbbek, hanem változatosabbak is.

### 3.3. Vizuális Turing-teszt

A vizuális Turing-tesztet arra használtuk, hogy megvizsgáljuk, mennyire képesek a megkérdezettek felismerni, hogy a generált képek valódiak-e vagy mesterségesek. A teszt során két különböző modell által generált képeket mutattunk: egy VAE (Variational Autoencoder) és egy DDPM (Denoising Diffusion Probabilistic Model) által készített mintákat.

**VAE által generált képek**:

- Az értékelések alapján a VAE által készített képeket sokan „homályosnak” és „kevésbé részletgazdagnak” találták.
- Több megjegyzés utalt arra, hogy az emberek a képekben torzításokat vagy természetellenes textúrákat véltek felfedezni



*4. ábra A VAE modellel generált arcképek*

**DDPM által generált képek**:

- A megkérdezettek nagy része úgy vélte, hogy ezek a képek közel állnak a valódihoz, de szintén egy kis homályosságot vagy elmosódottságot említettek.
- A DDPM által készített képek esetében (főleg a celeba model által generált képeknél) az emberek a részletgazdagságot és a természetes textúrákat emelték ki.

5. ábra A DDPM modellel generált arcképek

## 4. Összefoglalás

A projekt során elkészültek alapján kijelenthetjük, hogy DDPM valóban sokkal élethűbb és meggyőzőbb képeket képes generálni, mint a VAE. Azonban fontos megjegyezni, hogy egyszerű kis számítási kapacitással rendelkező laptopokkal lehetetlen sokkal kifinomultabb modelleket tanítani. Érdekes lenne a későbbiekben megvizsgálni, hogy egy sok RAM-os a100-as GPU segítségével képesek lehetünk-e még jobb modelt tanítani. Ez azonban költséges manapság.

## 5. LLM model használata

A munka során mind kódgeneráláshoz (rövid szkriptek), mind a dokumentáció készítéséhez használtunk LLM-et (ChatGPT). Tapasztalataink alapján ezzel kb. 30-40%-kal sikerült meggyorsítanunk a munkafolyamatot.

## 6. Telepítési útmutató

Nyiss egy terminált a projekt gyökérkönyvtárában. Az alkalmazás indításához kövesd az alábbi lépéseket:

- Mielőtt futtatnád az alkalmazást, először meg kell építeni a Docker Image-et:

```
docker-compose build
```

- Ha a **run** és **stop** scripteket használod, biztosítsd, hogy azok futtathatóak legyenek az alábbi parancsokkal:

```
chmod +x run.sh
chmod +x stop.sh
```

- **Gradio felület indítása a localhost:7860 címen:**

```
./run.sh
```

- **Modellek betanítása:**

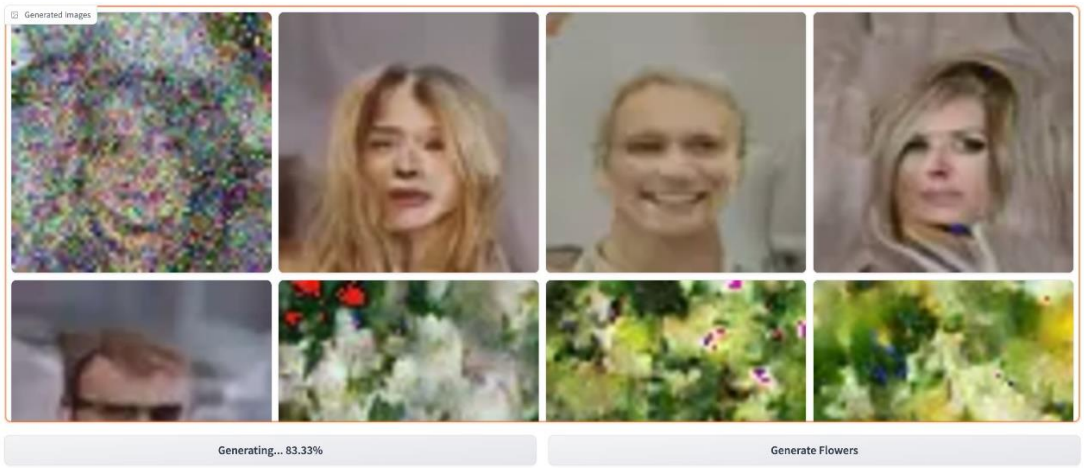
```
./run.sh --train-flowers                # Virág modellek betanítása
./run.sh --train-celebs                 # Celeb modellek betanítása
```

- **Virágképek generálása:**

```
./run.sh --generate-flowers             # Legjobb modell használata
./run.sh --generate-flowers --latest    # Legutóbbi modell használata a legjobb helyett
```

- **Celeb arckép generálása:**

```
./run.sh --generate-celebs              # Legjobb modell használata
./run.sh --generate-celebs --latest     # Legutóbbi modell használata a legjobb helyett
```




*6. ábra A Gradio felülete*

## 7. Felhasznált Irodalom


- Hugging Face. Diffusers Library. Elérhető: https://github.com/huggingface/diffusers
- Mallick, S. Guide to Training DDPMs from Scratch: Generating Flowers Using DDPMs. LearnOpenCV. Elérhető: https://github.com/spmallick/learnopencv/blob/master/Guide-to-training-DDPMs-from-Scratch/Generating_flowers_using_DDPMs.ipynb
- LearnOpenCV. Denoising Diffusion Probabilistic Models. Elérhető: https://learnopencv.com/denoising-diffusion-probabilistic-models/#What-Are-Diffusion-Probabilistic-Models
- Keras Documentation. Denoising Diffusion Implicit Models (DDIM). Elérhető: https://keras.io/examples/generative/ddim/