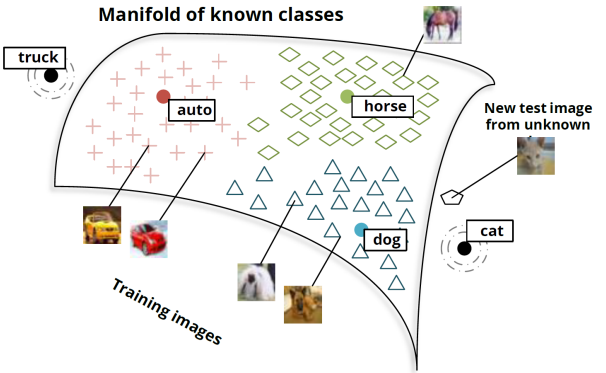
Modele generatywne 3: WAE

Jacek Tabor

13 października 2023

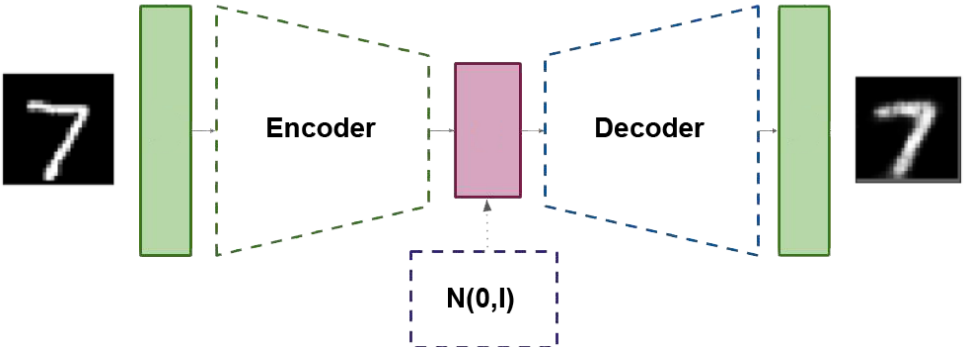


1 Rozkłady generatywne bazujące na AE

Twierdzenie 1. *Załóżmy, że mamy autoenkoder*

$$\mathbb{R}^D \xrightarrow{E} Z \xrightarrow{D} \mathbb{R}^D$$

który jest perfekcyjnie nauczony.

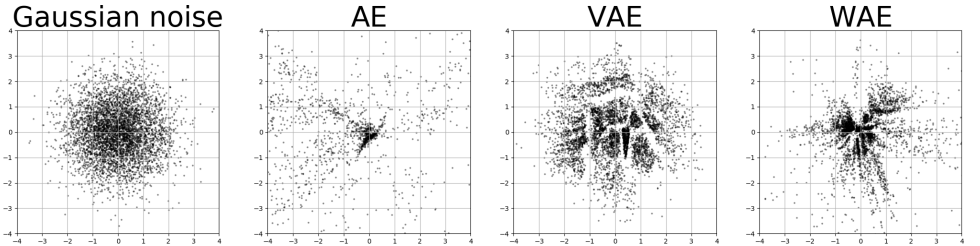


Rysunek 1: Typowy schemat działania modelu generatywnego na bazie AutoEnkodera. Dodatkowo żądamy, by zbiór danych po przejściu do przestrzeni latent miał rozkład normalny

Dodatkowo zakładamy, że zbiór danych X przerzucony przez enkoder E ma rozkład normalny $N(0, I)$

$$E(X) \sim N(0, I).$$

Samplowanie z rozkładu danych jest uzyskiwane przez samplowanie z $N(0, I)$ na przestrzeni Z.



Rysunek 2: Porównanie szumu gaussowskiego (pierwszy rysunek) z próbkami ze zbioru danych MNIST przerzuconych do latent dla modeli AE, VAE, WAE. Widzimy, że zarówno w VAE jak i WAE w latent dane mają rozkład zbliżony do rozkładu normalnego, podczas gdy AutoEnkoder nie wykazuje tego zachowania.

Najczęściej spotykane modele VAE, WAE.

Wizualizacja dla latenta dwuwymiarowego

2 Kernel density estimation (jądrowa/kernelowa estymacja gestości) jako metoda do porównywania próbek

Chcemy umieć policzyć odległość między rozkładami – próbkami. Esytmujemy gęstości na podstawie próbek.

próbka x_i

Mamy funkcję ϕ kernel (najczęściej gauss $N(0, h^2 I)$, h - to szerokość kernela) – domyślnie gęstość, dla której mamy wyliczoną wartość (funkcja kernelowa):

$$k(x, y) = \int \phi(s - x)\phi(s - y)ds$$

Często ϕ jest zadane rozkładem normalnym.

Tworzymy kernelowe estymacje gęstości dla próbek:

$$F : s \to \frac{1}{n} \sum_i \phi(s - x_i),\ G : s \to \frac{1}{n} \sum_i \phi(s - x_i),.$$

Odległość między funkcjami f, g

$$\|f - g\|^2 = \int |f(s) - g(s)|^2 ds = \int f(s)f(s) - 2\int f(s)g(s) + \int g(s)g(s).$$

Podstawiając dostajemy

$$MMD_k((x_i),(y_j)) = \|F-G\|^2 = \int F(s)F(s) - 2\int F(s)G(s) + \int G(s)G(s)$$

Możemy przeliczyć każdy z 3 czynników:

$$\int F(s)F(s)ds = \int \frac{1}{n}\sum_i \phi(s-x_i)\frac{1}{n}\sum_{i'}\phi(s-x_{i'})ds = \frac{1}{n^2}\sum_{ii'}\int \phi(s-x_i)\phi(s-x_{i'}) = \frac{1}{n^2}\sum_{ii'}k(x_i,x_{i'}).$$

$$\int F(s)G(s)ds = \int \frac{1}{n}\sum_i \phi(s-x_i)\frac{1}{n}\sum_{j}\phi(s-y_j)ds = \frac{1}{n^2}\sum_{ij}\int \phi(s-x_i)\phi(s-y_j) = \frac{1}{n^2}\sum_{ij}k(x_i,y_j).$$

$$\int G(s)G(s)ds = \int \frac{1}{n}\sum_j \phi(s-y_j)\frac{1}{n}\sum_{j'}\phi(s-y_{j'})ds = \frac{1}{n^2}\sum_{jj'}\int \phi(s-y_j)\phi(s-y_{j'}) = \frac{1}{n^2}\sum_{jj'}k(y_j,y_{j'}).$$

Finalnie dla próbek (x_i), (y_j) dostajemy

$$MMD_k^2((x_i),(y_j)) = \frac{1}{n^2}\sum_{ii'}k(x_i,x_{i'}) - \frac{2}{n^2}\sum_{ij}k(x_i,y_j) + \frac{1}{n^2}\sum_{jj'}k(y_j,y_{j'})$$

My będziemy rozważać funkcję kernelową:

$$k(x,y) = \frac{1}{1+\|x-y\|^2}.$$

Python 1. *wylosuj dwie próbki z rozkładu normalnego $N(0,1)$, policz ich kernelową odległość, a następnie odległość $N(0,1), N(0,1/2), N(1,1)$.*

3 Model WAE-MMD

Rozpatrujemy model:

$$\mathbb{R}^D \xrightarrow{\mathcal{E}} Z \xrightarrow{\mathcal{D}} \mathbb{R}^D$$

Enkoder $\mathcal{E}$ i dekoder $\mathcal{D}$ jest zadany przez sieć.

Funkcja straty dla batcha x_i to

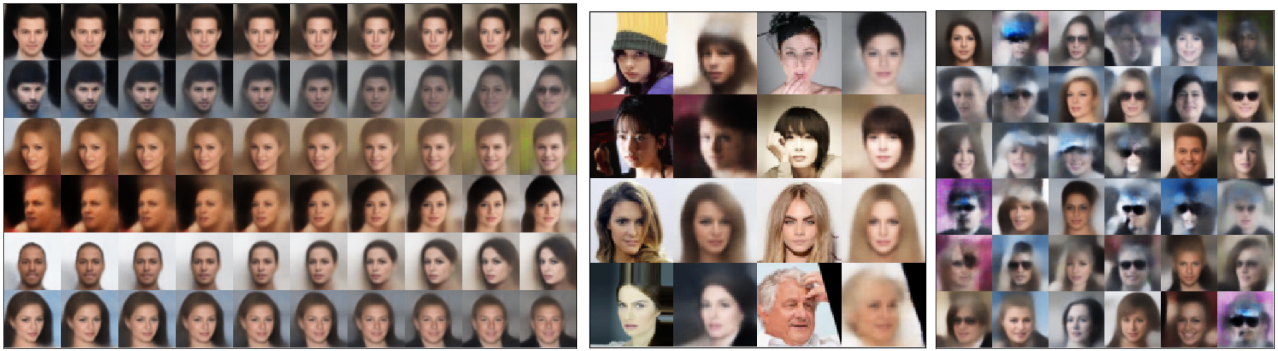
$$loss(x_i) = \frac{1}{n}\sum_i \|x_i - \mathcal{D}\mathcal{E}x_i\|^2 + \lambda MMD_k^2(x_i, y_i),$$

gdzie y_i losowo wybrana próbka z rozkładu normalnego (λ dobierana do zbioru danych).

Python 2. *Nauczyć WAE dla MNISTa.*

4 Wizualna ocena modelu: próbkowanie, interpolacja

Reprezentacja zbudowana przez modele generatywne AutoEnkoderowe mają dwie dodatkowe cechy w stosunku do samych AutoEnkoderów (rysunek 3):



Rysunek 3: Na poniższym rysunku pokazujemy typowy sposób wizualnej oceny, czy dany model WAE-MMD jest poprawnie nauczony. Pierwszy rysunek pokazuje liniowe interpolacje w przestrzeni latent przerzucone do przestrzeni danych. Drugi rysunek pokazuje rekonstrukcje przykładowych punktów. Ostatni zaś losowe próbki z rozkładu normalnego w przestrzeni latent przerzucone do przestrzeni danych. W tym przypadku widzimy, że tak nauczony model osiągnał prawidłowe wyniki.

- pozwalają na próbkowanie z rozkładu danych,
- pozwalają na dokonywanie interpolacji.

Przez próbkowanie z rozkładu danych rozumiemy wykonanie próbki z rozkładu normalnego w przestrzeni latent i następnie przerzucenie do przestrzeni danych. Interpolacji (liniowej) dokonujemy także w przestrzeni latent, następnie przerzucamy do przestrzeni danych. Niestety przykłady generowane przez WAE-MMD często odbiegają jakością od prawdziwych danych, szczególnie dla dużych obrazów, rysunek 3. W kolejnym rozdziale przedstawimy koncepcję GANów, które są rodzajem modelu generatywnego pozwalającym na tworzenie lepszej jakości zdjęć.

Manipulacje na zdjęciach Okazuje się, że w modelach typu VAE możemy wykonywać podobnego manipulacje. Załóżmy bowiem, że mamy zdjęcie osoby, która jest smutna. Domyślnie pracujemy tu na zbiorze danych CelebA, w którym mamy poetykietowane twarze. Możemy wtedy do niej dodać uśmiech, by się uśmiechnęła. Oczywiście pytanie jak uzyskać wektor uśmiechu – po prostu bierze się różnicę między uśrednionym człowiekiem usmiechniętym a smutnym (oczywiście wszystkie operacje wykonujemy w przestrzeni latent):

$$SmutnaTwarz + [\mathbf{mean}(wesola) - \mathbf{mean}(smutna)] = WesolaTwarz.$$

5 FID score (Frechet Inception Distance)

Mamy zbiór danych X i próbkę y_i. Chcemy określić jak blisko jest próbka do rozkładu danych X.

Przechodzimy do przestrzeni reprezentacji, przedostatnia warstwa w sieci:

- wyłapuje istotne cechy obrazów

- „znośna” wymiarowość

Oznaczam $f(X)$, $f(y_i)$.

FID score Dla dwóch wielowymiarowych rozkładów Gaussa

$$d_F(N(\ mu, \Sigma), N(\mu', \Sigma'))^2 = \|\mu - \mu'\|_2^2 + \operatorname{tr}\left(\Sigma + \Sigma' - 2\left(\Sigma^{\frac{1}{2}} \cdot \Sigma' \cdot \Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)$$

To pozwala nam zdefiniować FID w formie pseudokodu:

- WPROWADŹ funkcję [reprezentacja]
$$f : \Omega_X \to \mathbb{R}^n.$$
- WPROWADŹ dwa zestawy danych
$$S, S' \subset \Omega_X.$$
- Oblicz
$$f(S), f(S') \subset \mathbb{R}^n.$$
- Dopasuj dwa rozkłady Gaussa, odpowiednio – $\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma')$

- Zwróć

$$d_F(N(\mu, \Sigma), N(\mu', \Sigma')^2.$$

W większości praktycznych zastosowań FID, Ω_X to przestrzeń obrazów, a f to model Inception v3 wyszkolony w ImageNet, ale bez końcowej warstwy klasyfikacyjnej. Technicznie rzecz biorąc, jest to 2048-wymiarowy wektor aktywacji ostatniej warstwy.