Autoencodere Variaționale (VAE)

Inteligență Artificială 3

Facultatea de Electronică, Telecomunicații și Tehnologia Informației

Studenți: Mălina-Cerasela Manolache și Radu-George Bolborici

Profesori: Șef Lucrări Dr. Ing. Ana-Antonia Neacșu și Drd. Ing. Vlad-Mihai Vasilescu

Data: decembrie 2024

Cuprins

- Scopul unui VAE
- Autoencodorul
  - Principii
  - Noițunea de spațiu latent
- Formulare matematică VAE
  - Bayes
  - Divergența Kullback-Leibler
  - Derivarea funcției de cost
- Antrenare și inferență
- Demo: model pentru generarea formelor geometrice
- Convergența VAE
- VAE pentru detecție de anomalii

Scopul unui autoencoder variațional (VAE)

- Model generativ.
- Introdus in 2013 de Diedrick P. Kigma si Max Welling in "Auto-Encoding Variational Bayes".
- **Generează date noi similare cu cele prezentate la antrenare.**

Imagini de antrenare

Imagini generate

fig. 1 VAE

Autoencoderele [1]. Noțiunea de spatiu latent

- Rețele neurale care învață o reprezentare compactă a datelor de intrare și apoi reconstruiesc datele pe baza acestei reprezentări.
- Conțin 2 componente: **encodor** si **decodor.**
- Aplicații: reducerea dimensionalității, eliminarea zgomotului din imagini, compresia datelor, extracția de trăsături.

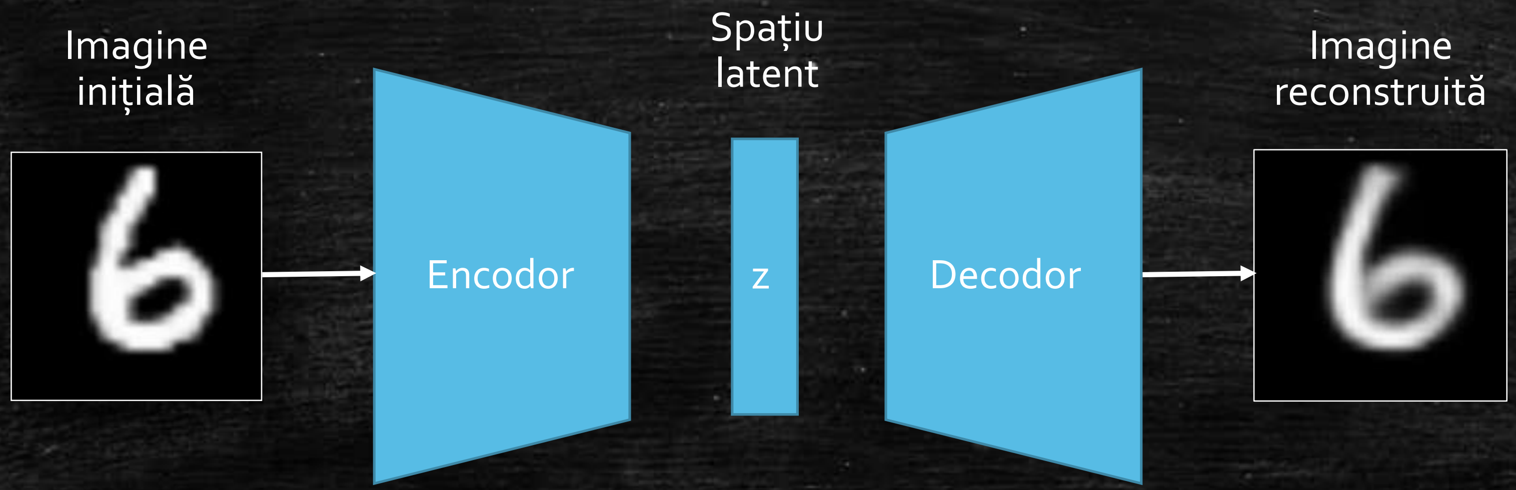

fig. 2 Schema generală a unui autoencoder

Autoencoderele. Formulare matematică

- Setul de date: $X = \{x_i | i = 1, \dots M\}, x_i \in \mathbb{R}^d$
- Spațiul latent: $Z \subseteq \mathbb{R}^k,\ k \ll d$
- Encodorul: $E_\phi : X \to Z \quad z = E_\phi(x)$
- Decodorul: $D_\theta : Z \to X \quad \hat{x} = D_\theta(z)$
- Funcția de cost: $L(x, \hat{x})$ (e.g. L2-norm)

MLP, CNN (ϕ, θ sunt ponderile rețelei)

$$min_{\phi,\theta} \frac{1}{M} \sum_{i=1}^{M} L(x, \hat{x}) = \mathbb{E}\left[L\left(x, D_\theta\left(E_\phi(x)\right)\right)\right] \quad (1)$$

Autoencoderele. Generarea datelor noi

- Cum se pot genera date noi cu un autoencodor?
  - Se generează spațiu latent aleator, apoi se decodează folosind decodorul antrenat
    - Din ce distribuție ar trebui extras vectorul latent?

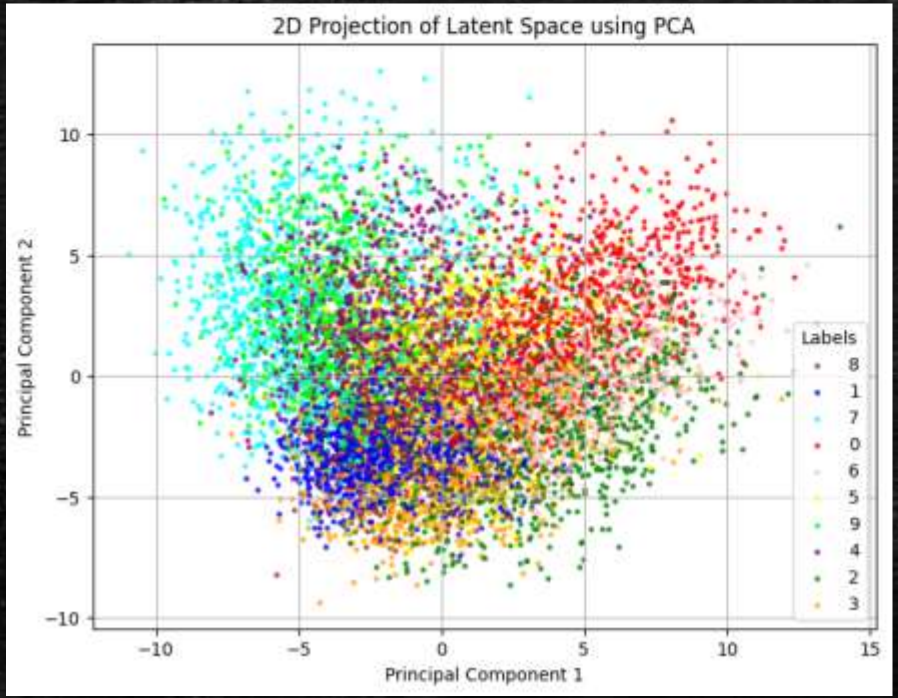

fig. 3 Spațiul latent al setului de test MNIST

- Spațiul latent este dezorganizat-> nu se pot genera date noi realiste.

- Autoencodor Variațional (VAE) -> spațiul latent este încurajat să tindă spre o anumită distribuție (a priori, uzual Gaussiană).

Teorema lui Bayes

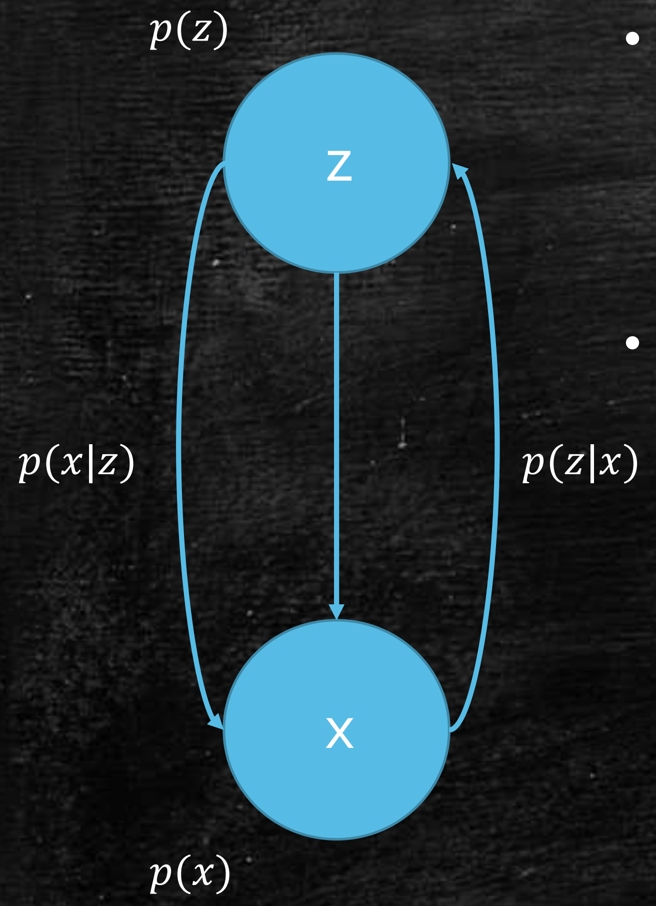

fig. 4

- Presupunem ca observația x este generată de un process aleatoriu care implică variabila aleatoare z (neobservată)
  1. z este extras din distribuția apriori $p(z)$
  2. x este extras din distribuția condiționată $p(x|z)$
- Ne interesează distribuția posterioară $p(z|x)$

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \quad (2)$$

- $p(x) - distributia\ datelor$
- $p(y) - distributia\ a\ priori$
- $p(x|z) - distributia\ conditionata$
- $p(z|x) - distributia\ posterioara$

Teorema lui Bayes

- Corespondența cu autoencodorul
- Dacă știm $p(x)$ putem genera date noi
- Obiectiv:

$$max_\theta\ p_\theta(x) = \int p_\theta(x|z)p(z)dz \qquad (3)$$

Complex de calculat

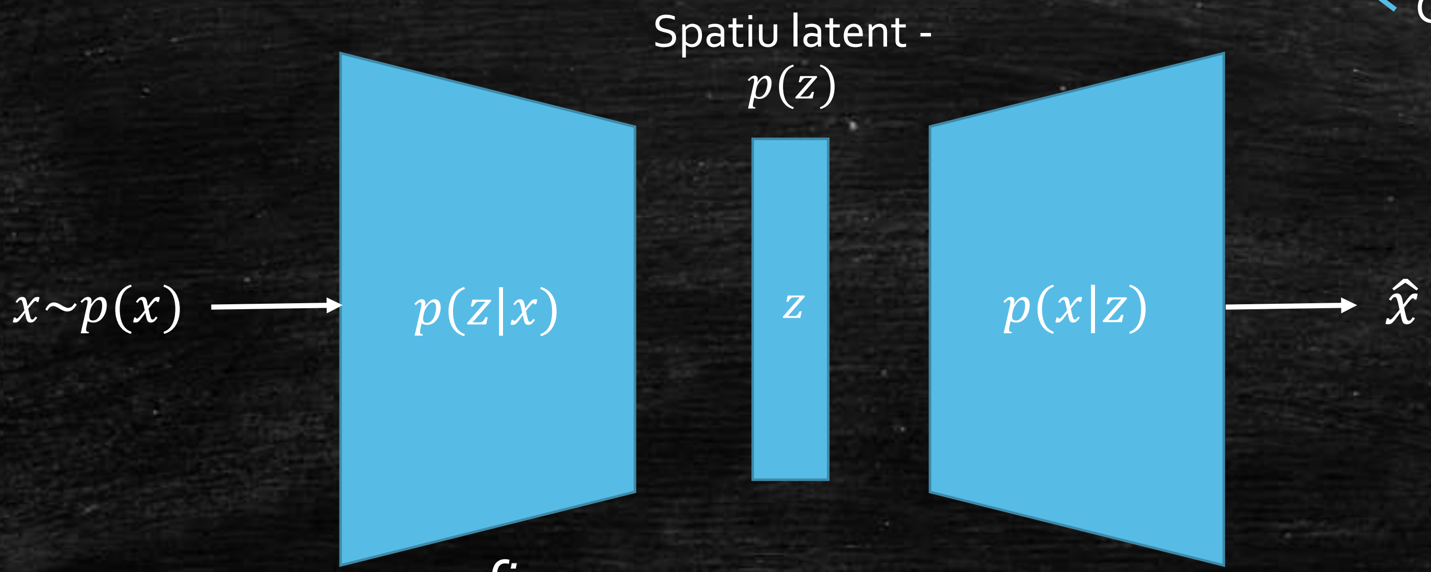

fig. 5 Procesul generativ

Bayes Variațional [2]

- Distribuția a posteriori $p(z|x)$- encodorul

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\int p(z)p(x|z)dz} \quad (4)$$

Complex de calcul
(intractable integral)

- Bayes Variațional = aproximăm $p(z|x)$ cu o distribuție cat mai apropiată $q_{\phi}(z|x)$ (e.g. Gaussiană)
- Problema de optimizare:

$$min_{\phi}\, D_{KL}(q_{\phi}(z|x)||p(z|x)) \quad (5)$$

Ce este distana Kullback-Leibler (KL)

- Divergența KL indică **distanța dintre două distribuții:** **q**(dist. post.) **p**(dist. a priori)

$$D_{KL}(q(z|x)||p(z)) = \int q(z|x) log \frac{q(z|x)}{p(z)} dz \quad (6)$$

- Dacă ambele distribuții sunt Gaussiene:

$$D_{KL}(q(z)||p(z)) = \frac{1}{2}\left[\frac{\sigma_q^2}{\sigma_p^2} + \frac{(\mu_p - \mu_q)^2}{\sigma_p^2} - 1 + \log\frac{\sigma_p^2}{\sigma_q^2}\right] \quad (7)$$

- Dacă în plus distribuția teoretică (a priori) p este Standard (m=0; μ=1):

$$D_{KL}(q(z)||p(z)) = \frac{1}{2}(\sigma_q^2 + \mu_q^2 - 1 - \log(\sigma_q^2)) \quad (8)$$

- Formula ce cuantifică diferențele dintre distribuții poate fi utilizată ca funcție de cost pentru a forța distribuțiille învățate să tindă spre gaussiene standard

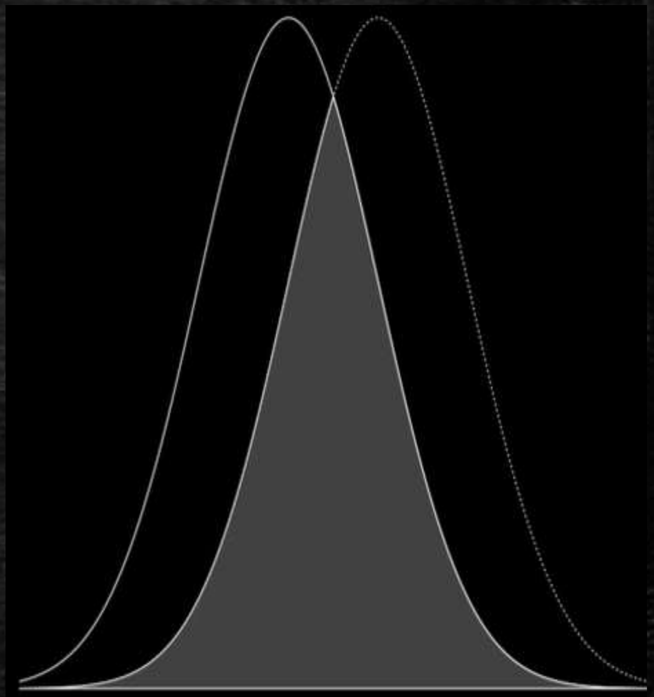

fig. 6 distribuții normale

$Obiectiv: max_\theta\, p_\theta(x)$

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

Derivarea functiei de cost

- $D_{KL}(q_\phi(z|x)||p(z|x)) = \int q_\phi(z|x) log \frac{q_\phi(z|x)}{p(z|x)} dz \quad (9)$ — Se scrie Bayes pentru p(z|x)
- $D_{KL}(q_\phi(z|x)||p(z|x)) = \int q_\phi(z|x) log \frac{q_\phi(Z|X)p(x)}{p(X|Z)p(z)} dz \quad (10)$ — Se separă logaritmul
- $D_{KL}(q_\phi(z|x)||p(z|x)) = \int q_\phi(z|x) \left(log \frac{q_\phi(Z|X)}{p(X|Z)p(z)} + \log p(x) \right) dz \quad (11)$ — Se separă integrala
- $D_{KL}(q_\phi(z|x)||p(z|x)) = \int q_\phi(z|x) log \frac{q_\phi(Z|X)}{p(X|Z)p(z)} dz + \int q_\phi(z|x) \log p(x)\, dz \quad (12)$ — $p(x)$ nu depinde de z
- $D_{KL}(q_\phi(z|x)||p(z|x)) = \int q_\phi(z|x) log \frac{q_\phi(Z|X)}{p(X|Z)p(z)} dz + \log p(x) \int q_\phi(z|x) dz \quad (13)$ — Integrala unei distribuții e 1
- $D_{KL}(q_\phi(z|x)||p(z|x)) = \int q_\phi(z|x) log \frac{q_\phi(Z|X)}{p(X|Z)p(z)} dz + \log p(x) \quad (14)$

Derivarea funcției de cost

- $D_{KL}(q_\phi(z|x)||p(z|x)) = \int q_\phi(z|x) log \frac{q_\phi(Z|X)}{p(X|Z)p(z)} dz + \log p(x)$ (15) se inversează fracția din logaritm
- $D_{KL}(q_\phi(z|x)||p(z|x)) = -\int q_\phi(z|x) log \frac{p(X|Z)p(z)}{q_\phi(Z|X)} dz + \log p(x)$ (16)
- $D_{KL}(q_\phi(z|x)||p(z|x)) = -L(\phi) + \log p(x)$ (17)
- $min_\phi\, D_{KL} \rightarrow max_\phi L(\phi)$ (pentru encodor)

Derivarea funcției de cost

- $L(\phi) = \int q_\phi(z|x) log \frac{p(X|Z)p(z)}{q_\phi(Z|X)} dz,\qquad D_{KL} = -L(\phi) + \log p(x)$
- Pentru decodor fixăm q, parametrizăm $p(x|z)$ cu θ
- $L(\theta) = \int q(z|x) log \frac{p_\theta(X|Z)p(z)}{q(Z|X)} dz\,,\ D_{KL} = -L(\theta) + \log p(x) \geq 0 \quad (18)$
- $\log p(x) \geq L(\theta)$ ("Evidence Lower Bound"-ELBO)
- Am pornit de la faptul că nu putem calcula $p(x)$ dar daca maximizăm ELBO -> maximizăm $p(x)$
- Obiectivul pentru decodor: $max_\theta L(\theta)$

Derivarea funcției de cost

- Obiectivul VAE: $max_{\phi,\theta} L(\phi,\theta)$

- $L(\phi,\theta) = -\int q_{\phi}(z|x) \log \frac{q_{\phi}(z|x)}{p_{\theta}(x|Z)p(z)} dz \quad (19)$

- $L(\phi,\theta) = -\int q_{\phi}(z|x) \log \frac{q_{\phi}(Z|x)}{p(z)} dz + \int q_{\phi}(z|x) \log p(x|z) dz \quad (20)$

- $L(\phi,\theta) = -D_{KL} + \mathbb{E}_{\sim q_{\phi}(Z|x)}[\log p(x|z) dz] \quad (21)$

Costul pentru regularizarea spațiului latent

Costul pentru reconstrucție

Distribuțiile VAE

- Posteriorul aproximat $q_\phi(z|x)$ și procesul generativ $p_\theta(x|z)$ sunt rețele neurale

- Posteriorul este o distribuție Gaussiană multidimensională
- $q_\phi(z|x) = \mathcal{N}(z; \mu_\phi, \sigma_\phi \boldsymbol{I}) \quad (22)$

- A priori este Gaussiană standard
- $p(z) = \mathcal{N}(0, \boldsymbol{I}) \quad (23)$

Schema generală VAE: antrenare

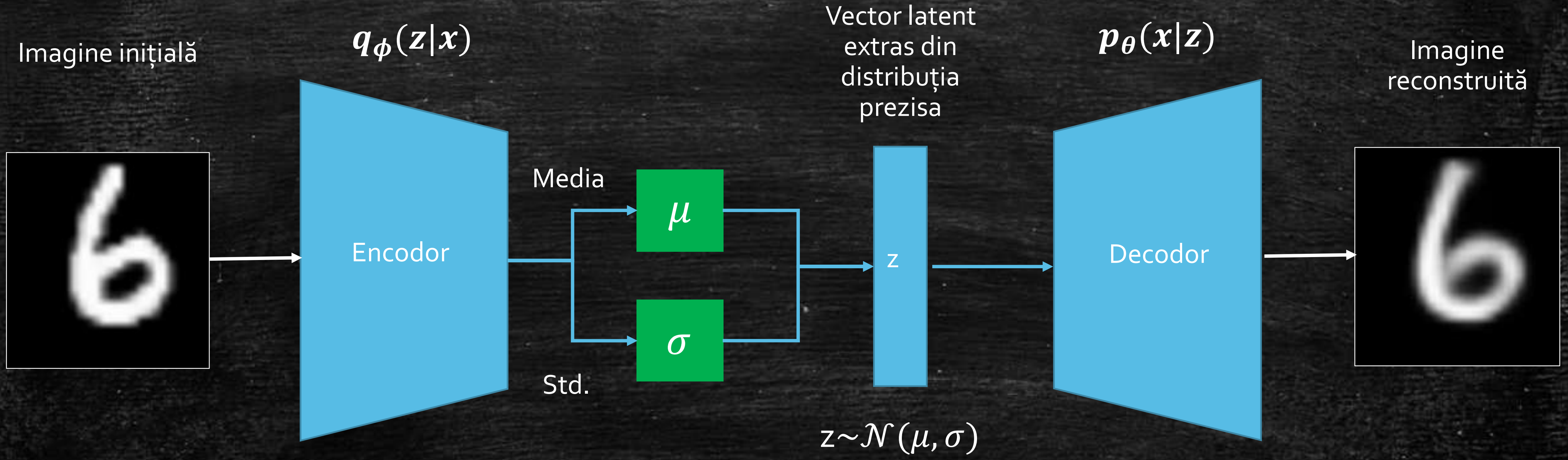

fig. 7 antrenarea VAE

$$L = L_{reconstructie} + L_{KL} \quad (6)$$

$L_{reconstructie}$ *poate fi* $L2 - norm$, *BCELoss*

Eșantionarea spațiului latent (prin *Reparametrization trick*)

- *Reparametrization trick este* folosit pentru a transforma un vector extras dintr-o distribuție sursă (standard de medie o si dispersie 1), într-un vector dintr-o distribuție destinație (cea de la ieșirea encoder-ului):

$$z = \mu + \sigma\varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, 1)$$

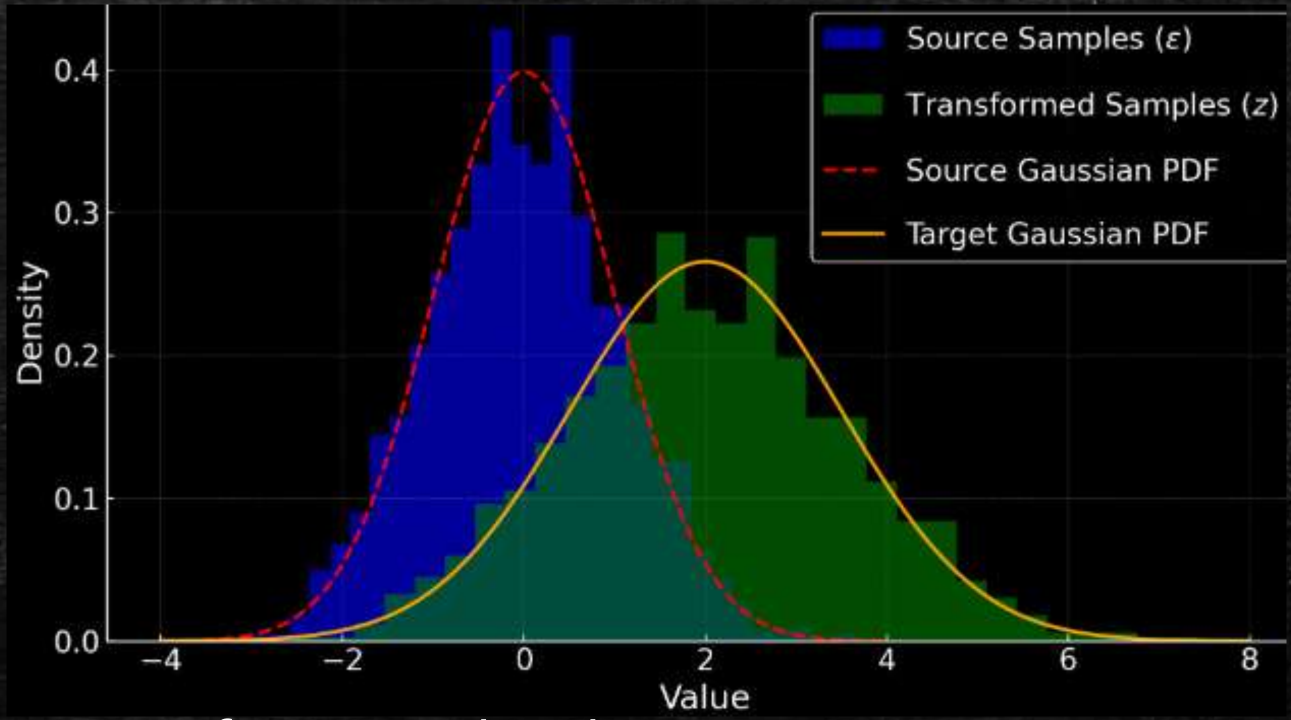

fig. 8 transformarea distribuției prin reparametrization trick

Schema generală VAE: inferență

- La inferență encodorul nu mai este folosit: se extrag eșantioane din distribuția **a PRIORI**
- Datorită regularizării spațiului latent, distribuția învățată (a posteriori) este similară celei a priori

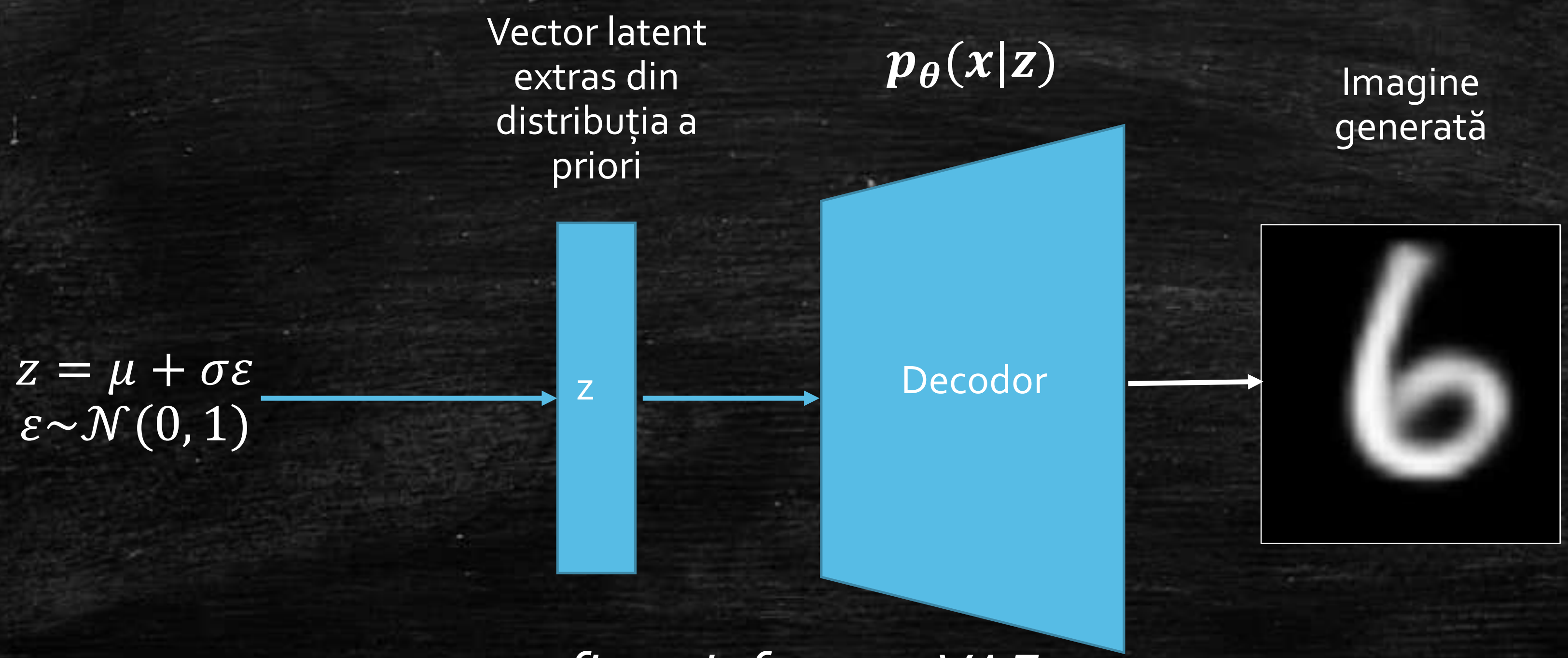

fig. 9 inferența VAE

Condiționarea VAE

- Se pune **problema condiționării** conținutului generat (Ex: controlarea claselor de obiecte prezente într-o imagine generată)
- Condiția se introduce la antrenare, în funcție de clasele/caracteristicile prezente în datele de antrenare, apoi, la inferență, va încuraja prezența acelorași trăsături în eșantionul de date generat
- Ce anume poate fi o condiție? Ex: un *token*, un *embedding*
- Unde și cum se introduce condiția?
  - în *encodor* (la intrarea datelor *x* sau concatenată în spațiul trăsăturilor extrase)
  - în *decodor* (de obicei concatenată cu reprezentarea latentă *z*)
- Formulare matematică: $\boldsymbol{q_\phi(z|x) \rightarrow q_\phi(z|x,c)}$

  $\boldsymbol{p_\theta(x|z) \rightarrow p_\theta(x|z,c)}$

Probleme asociate antrenării (C)VAE
Ponderarea loss-ului KL și dimensiunea spațiului latent

- Problema coeficientului de regularizare

$$L = L_{reconstructie} + \beta L_{KL} \quad (24)$$

- β prea mic: regularizare slabă: la generare, când se eșantionează din distribuția a priori, există șanse mici să se extragă un vector reprezentativ dpdv. semantic
- β prea mare: se poate ajunge la soluția banală ca distribuția a posteriori învățată să coincidă cu cea a priori, iar decodorul poate ignora acest spațiul latent (poate cauza **colaps posterior**)
- Eventuală soluție: creșterea treptată a coeficientului β
- Alegerea dimensiunii spațiului latent:
  - Dimensiunea spațiului latent prea mică: resurse insuficiente pentru modelarea distribuției datelor
  - Dimsniunea spațiului latent prea mare: supraînvățare / ignorarea condiției

Probleme asociate antrenării (C)VAE Colapsul Modelului

- Colapsul posterior (*posterior collapse*) [5]
  - Reconstrucția (decodorului) are loc preponderent pe baza distribuției a priori și nu mai ține cont suficient de mult de spațiul latent.
  - Cauză: un **coeficient prea mare** pentru termenul de cost Kullback-Leiber -> se poate forța ca distribuțiile sa fie prea apropiate de standard (soluție banală).

- Colapsul modelului (*model collapse/ posterior overfitting*)
  - Spațiul latent nu reflectă diversitatea distribuției datelor (caz mai general).
  - Datele generate devin repetitive

Decuplarea spațiului latent (*disentenglement*)

- Decuplare (*disentenglement*) = variabilele latente să fie independente și să fie asociate unor trăsături distincte pentru a putea modela distribuții complexe

- SOLUȚIE: încurajarea decorelării între componentele spațiului latent prin minimizarea corelației totale (*TC*)

  - $TC(z) = D_{KL}(q(z)||\prod_i q(z_i)), unde\ z_i\ sunt\ componentele\ z$ (25)
  - $L = L_{reconstructie} + \gamma TC(z)$ (26)

- Interpretare: când variabilele spațiului latent sunt independente, distribuția comună și produsul distribuțiilor marginale coincid, deci distanța dintre ele trebuie minimizate ca parte a funcției de cost.

DEMO

Generarea de forme colorate

Set de date generat automat: fiecare imagine conține o figură geometrică: pătrat ◻(0) , cerc ◯(1) , triunghi △(2) sau hexagon ⬡(3)

Fiecare figură geoetrică are o culoare:
roșu (0), albastru(1), verde (2), galben (3)

Atribute codiționate: ***formă*** și ***culoare***
Atribute necondiționate: ***poziționare*** și ***dimensiune***

Link proiect (branch: one_single_shape_improve_4):
https://github.com/RaduBolbo/Conditional-VAE-for-generating-geometric-shapes

Discuție: garanții teoretice de convergență

- La ce se referă convergența?
- Maximizarea ELBO (Evidence Lower BOund) -> adică **distribuția învățată trebuie să se asemene distribuției datelor**

$$log\ p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[logp_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)) \quad (27)$$

- **Ponderile converg** la valori care **maximizează ELBO**
- **Gradienții converg** spre **zero**.
- Articolul "Theoretical Convergence Guarantees for Variational Autoencoders" [4] arată că diverse variante de VAE converg cu o rată de **O(log(n)/sqrt(n))** în cazuri non-asimptotice (date limitate, iterații finite)

Discuție: garanții teoretice de convergență

- Tot articolul [4] arată că:
  - β mic -> crește viteza convergenței (dar reduce regularizarea spațiului latent)
  - K mare -> crește viteza de convergență pentru modelele IWAE (*Importance Weighted Autoencodes*: *discuție*!), K este numărul de eșantioane extrase pentru un anumit exemplu, la antrenare
  - funcția de activare "*generalised soft-clipping*" (28) ->introdusă pentru a îmbunătăți stabilitatea convergenței la antrenare

$$f(x) = \frac{1}{s}\log\left(\frac{1+e^{s(x-s_1)}}{1+e^{s(x-s_2)}}\right) + s_1 \quad (28)$$

Funcția (28) este:

1) continuă Lipschitz (nu are variații foarte bruște) $|f(x) - f(y)| \leq L|x - y| \quad (29)$

unde L e constanta Lipschitz, iar x, y două puncte

2) are derivata continuă

Alte exemple de aplicații ale (C)VAE

- Generare

  - generare de imagini. Ex: forme geometrice (link)

  - image morphing. Ex: fețe umane (link);

  - sinteza vorbirii (TTS). Ex: VITS (link)

- Detecție de anomalii: Ex: (link)

VAE în detecție de anomalii

- **Detecția de anomalii** = identificare de *pattern-uri* rare, care nu se regăsesc în mod normal în distribuția de date sau se gasesc cu o frecvență mai redusă (*outliers*)
- Aplicații: detecție de fraudă, de intruziuni, alerte în supravegherea video, detecția anomaliilor medicale
- Cum ar putea fi folosit VAE pentru detecția de anomalii?

Ex: detecție de anomalii [6]

- Intuiție: datele din afara distribuției (*outliers*) probabil că nu vor putea fi recosntruite la fel de bine la trecerea prin VAE
- Procedeu[6]: datele se trec prin encodor care produce o medie și o dispersie utilizate pentru a extrage mai mulți vectori ce pot fi decodați, astfel estimând intrarea originală.
- O eroare de reconstrucție mare poate indica o anomalie

References

- [1] D. Bank, N. Koenigstein, and R. Giryes, *Autoencoders*. 2021. [Online]. Available: https://arxiv.org/abs/2003.05991
- [2] D. P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*. 2022. [Online]. Available: https://arxiv.org/abs/1312.6114
- [3] D. P. Kingma and M. Welling, “An Introduction to Variational Autoencoders,” *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019, doi: 10.1561/2200000056.
- [4] Surendran, Sobihan, Antoine Godichon-Baggioni, and Sylvain Le Corff. "Theoretical Convergence Guarantees for Variational Autoencoders." *arXiv preprint arXiv:2410.16750* (2024).
- [5] Dai, Bin, Ziyu Wang, and David Wipf. "The usual suspects? Reassessing blame for VAE posterior collapse." *International conference on machine learning*. PMLR, 2020.
- [5] An, Jinwon, and Sungzoon Cho. "Variational autoencoder based anomaly detection using reconstruction probability." *Special lecture on IE* 2.1 (2015): 1-18.