



**fit@hcmus**

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG - HCM  
KHOA CÔNG NGHỆ THÔNG TIN

# XÂY DỰNG HỆ THỐNG GỢI Ý DỰA TRÊN MÔ HÌNH AUTOENCODER

Đào Đức Anh - Nguyễn Thành Nhân

GVHD: ThS. Trần Trung Kiên

# Nội dung

1. Tổng quan
2. Kiến thức nền tảng
3. Mô hình Variational Autoencoder cho Collaborative Filtering
4. Thực nghiệm
5. Kết luận

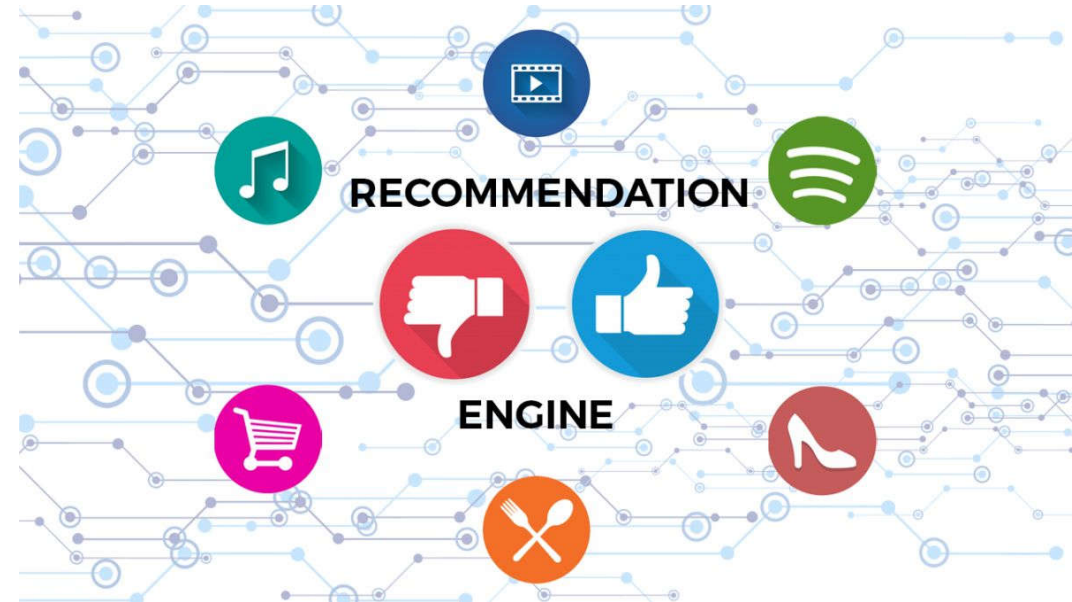
1.

# Tổng quan

**Giới thiệu về bài toán xây dựng hệ thống gợi ý sản phẩm**

# Giới thiệu

Hệ thống gợi ý được xây dựng để dự đoán những sản phẩm người dùng có thể thích, đặc biệt khi họ có nhiều lựa chọn.



# Giới thiệu



- Là một lĩnh vực trong khai thác dữ liệu và học máy.
- Là một phần quan trọng trong các doanh nghiệp

 **75%**

Các bộ phim được thuê

 **38%**

Số lượt click

 **35%**

Các sản phẩm được bán ra



# Phát biểu bài toán

- ❖ Cho input là dữ liệu về lịch sử tương tác của người dùng (user) với các sản phẩm (item) hoặc có thêm các mô tả của sản phẩm và thông tin về người dùng.
- ❖ Yêu cầu: đưa ra tập các item (không có trong lịch sử) được dự đoán là phù hợp với người dùng.

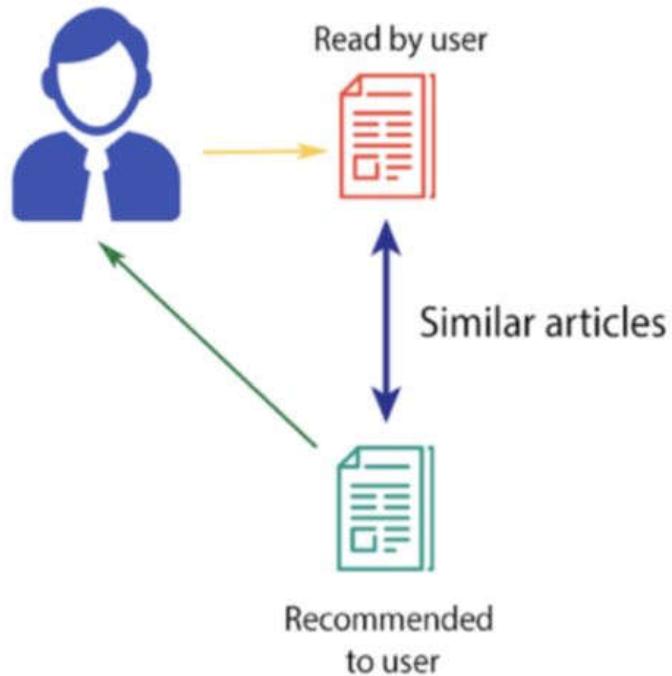
# Khó khăn gặp phải

- ❖ Khả năng tổng quát hóa còn hạn chế
- ❖ Sự thay đổi sở thích của người dùng
- ❖ Việc đánh giá hệ thống gợi ý còn khó khăn

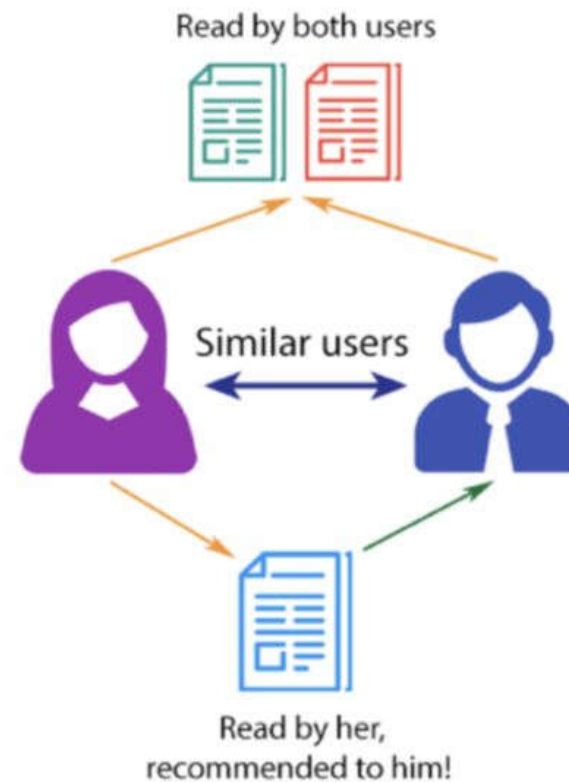


# Hướng tiếp cận

## CONTENT-BASED FILTERING



## COLLABORATIVE FILTERING



# Content-Based Filtering

## Ưu điểm

- ✓ Hiểu rõ về người dùng
- ✓ Điều chỉnh để phù hợp với sự thay đổi sở thích của người dùng

## Nhược điểm

- ✗ Hạn chế đối với sản phẩm ít thông tin
- ✗ Cần kiến thức nền để mô hình hóa dữ liệu

# Collaborative Filtering

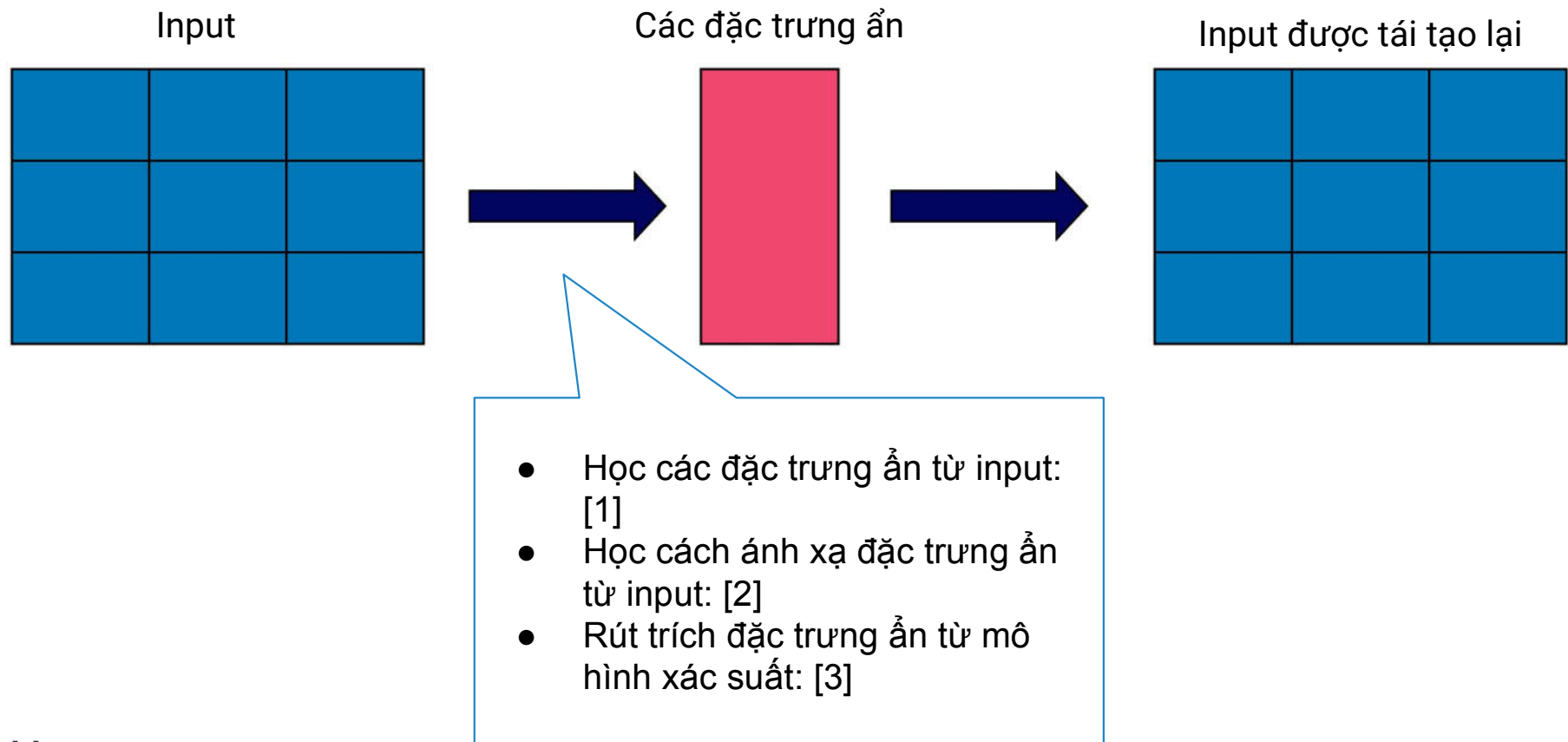
## Ưu điểm

- ✓ Tạo ra sự tình cờ
- ✓ Không cần các thuộc tính của sản phẩm
- ✓ Khả năng tổng quát hóa

## Nhược điểm

- ✗ Vấn đề cold-start

# Các nghiên cứu liên quan

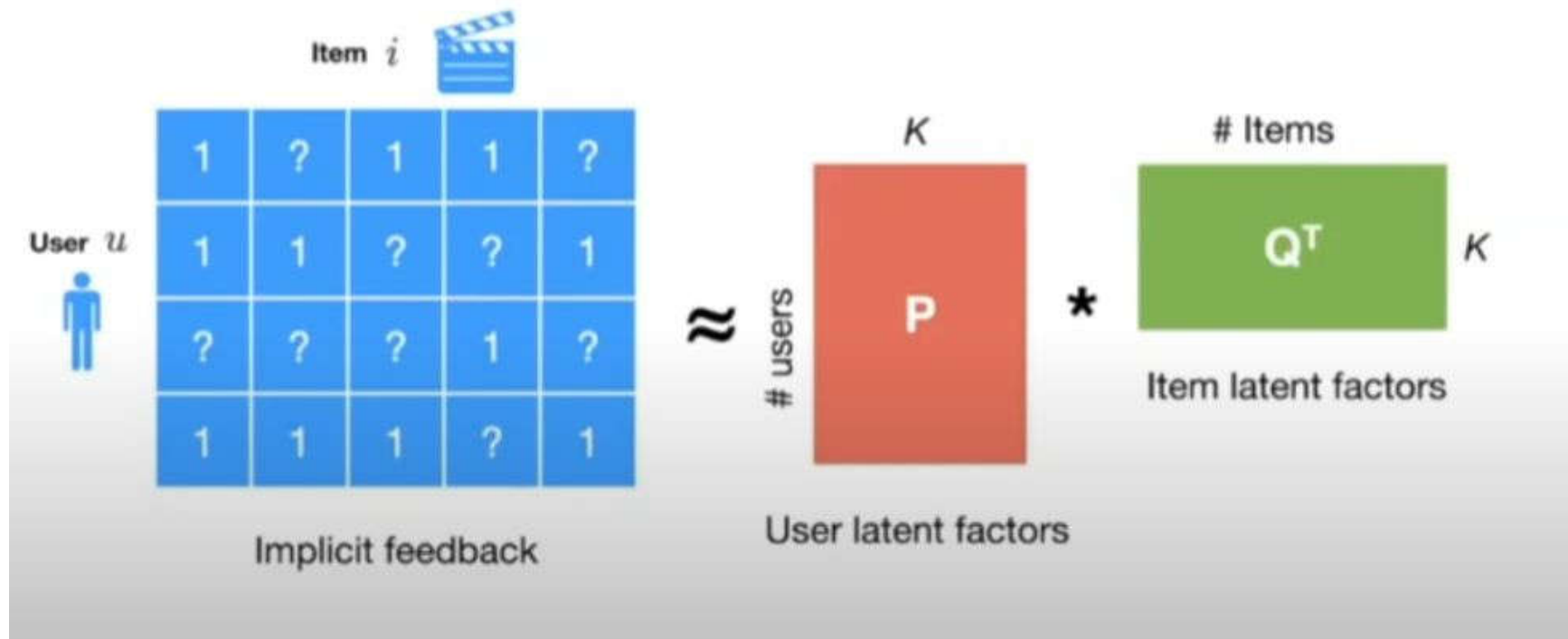


[1] Matrix factorization

[2] Yao Wu, Christopher DuBois, Alice X. Zheng, and Martin Ester. 2016. Collaborative denoising auto-encoders for top-n recommender systems. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. 153–162.

[3] “Variational Autoencoders for Collaborative Filtering” - Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, Tony Jebara

# Matrix factorization



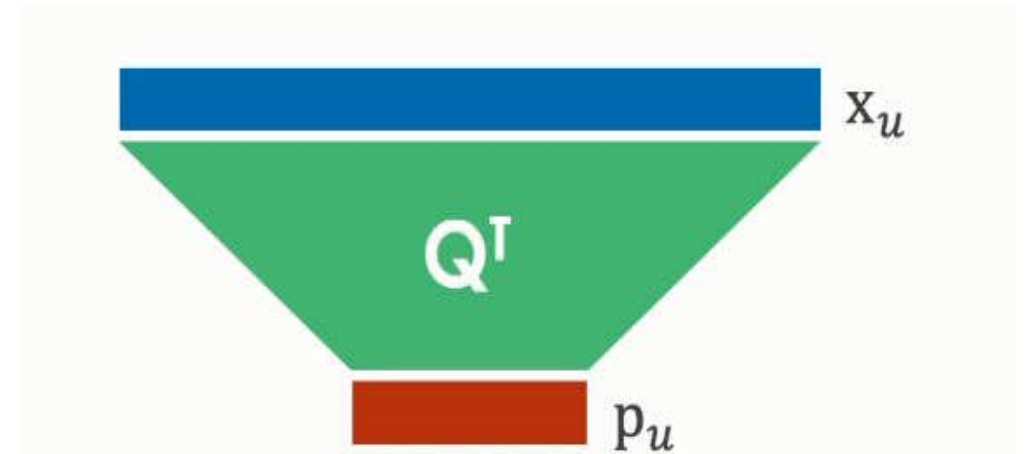
# Asymmetric Matrix Factorization

- Thêm một ma trận đại diện cho đặc trưng ẩn của item:  $V$
- Xây dựng đặc trưng ẩn của user dưới dạng **“trung bình” các item được user đó tương tác.**

$$p_u = \frac{1}{\sqrt{|I_u|}} \sum_{i \in I_u} v_i$$
$$= \tilde{x} \times V$$

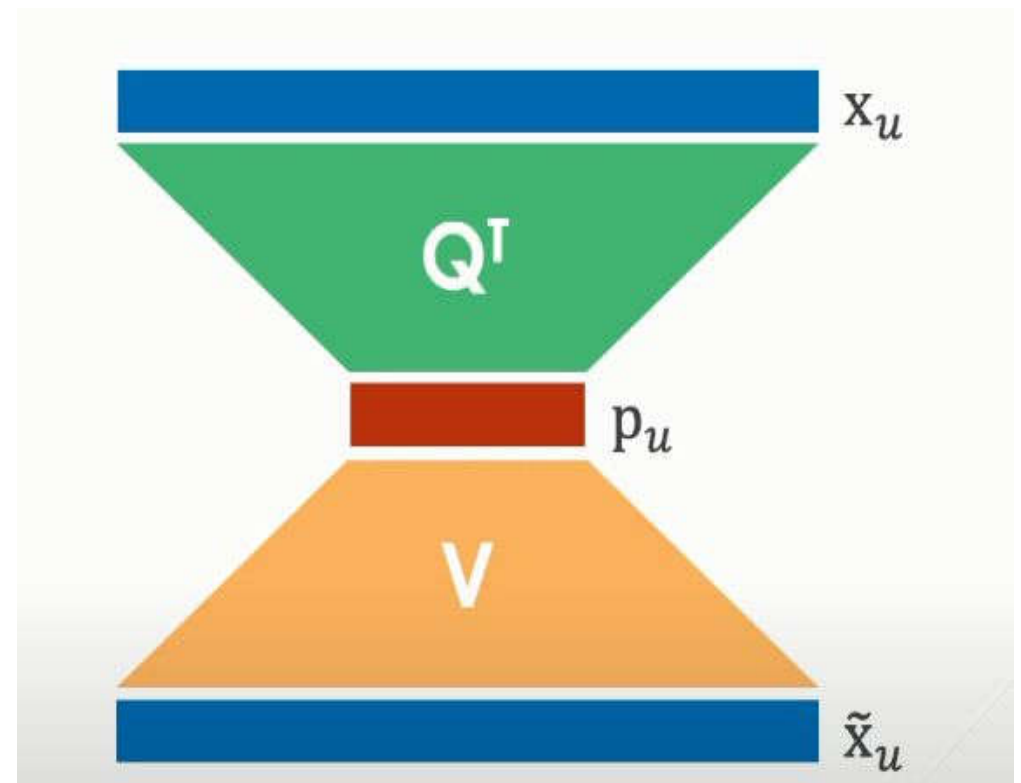
# Asymmetric Matrix Factorization

$$x_u \approx p_u \times Q^T$$



# Asymmetric Matrix Factorization

$$\begin{aligned}x_u &\approx p_u \times Q^T \\ &= \tilde{x} \times V \times Q^T\end{aligned}$$



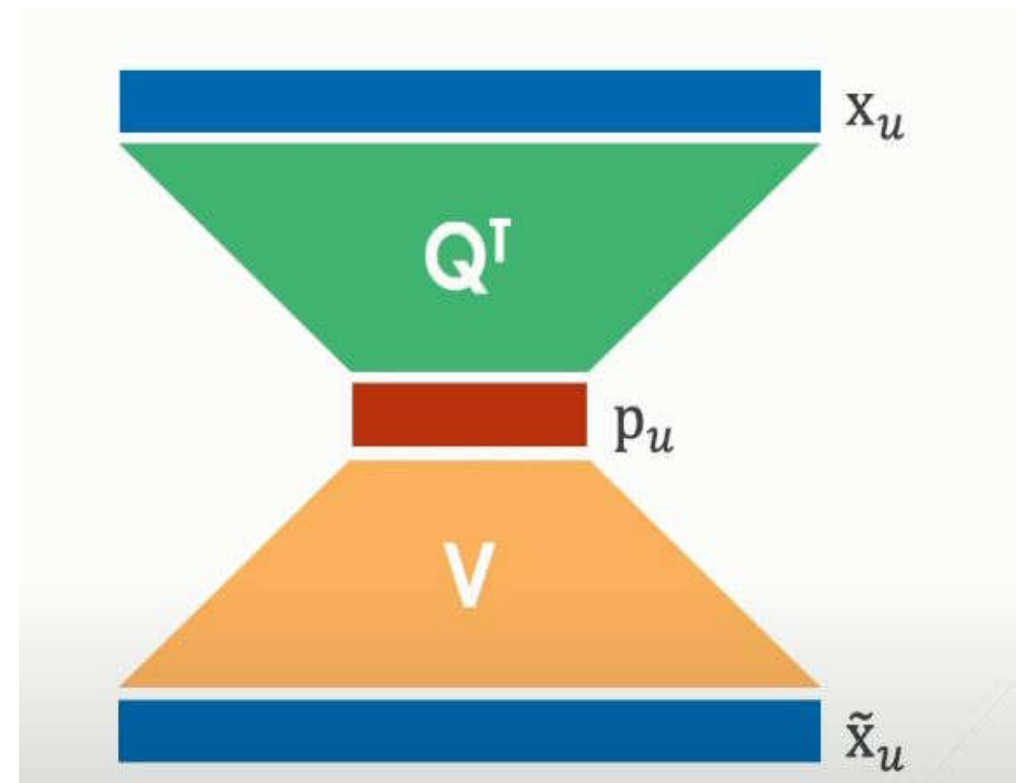


# Asymmetric Matrix Factorization

$$\begin{aligned}x_u &\approx p_u \times Q^T \\ &= \tilde{x} \times V \times Q^T\end{aligned}$$

$$p_u = g(x_u)$$

$$x_u = f(p_u)$$

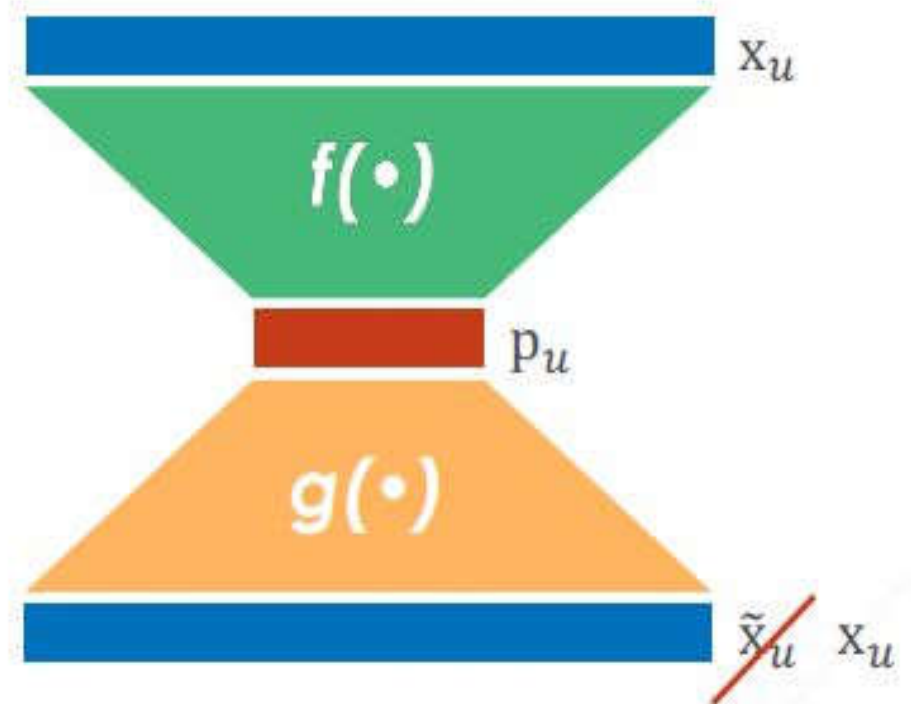


# Non-linear Autoencoder

$$\begin{aligned}x_u &\approx p_u \times Q^T \\ &= \tilde{x} \times V \times Q^T\end{aligned}$$

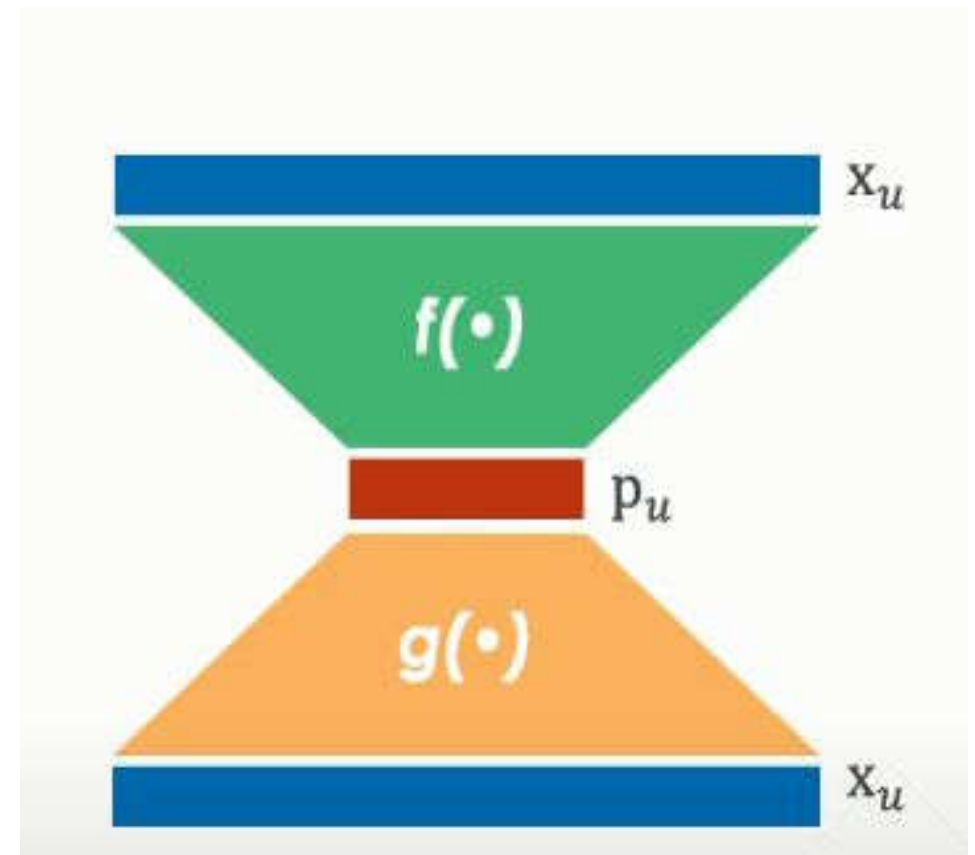
$$p_u = g(x_u)$$

$$x_u = f(p_u)$$



# Denosing autoencoder

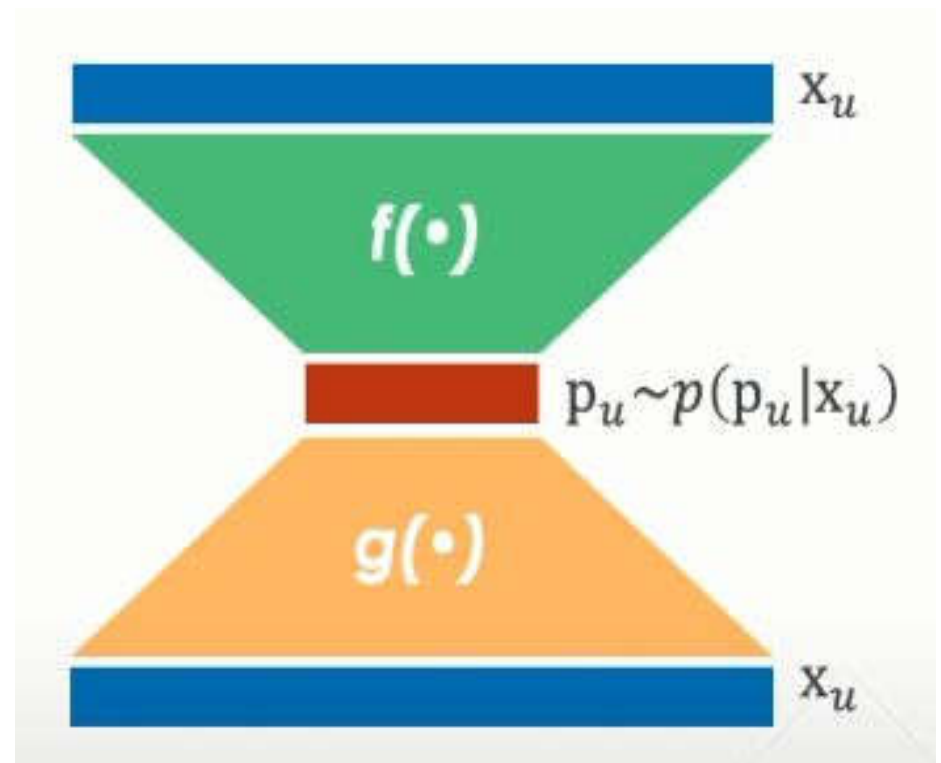
- Thay vì input là vec-tơ tương tác của người dùng, ta thêm nhiễu bằng cách ngẫu nhiên ẩn một vài tương tác hay còn được gọi là drop out
- Drop out khiến cho mô hình phù hợp hơn với bài toán xây dựng gợi ý sản phẩm
  - Tránh tình trạng overfitting
  - Trong quá trình huấn luyện mô hình sẽ cần dự đoán các item mới từ các tương tác trong quá khứ



Yao Wu, Christopher DuBois, Alice X. Zheng, and Martin Ester. 2016. Collaborative denoising auto-encoders for top-n recommender systems. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. 153–162.

# VAE

- Thay vì  $p_u$  là một điểm dữ liệu thông qua hàm phi tuyến  $f(\cdot)$  thì  $p_u$  là một điểm được phát sinh từ phân phối xác suất có điều kiện  $p_u \sim p(z_u | x_u)$
- Dữ liệu đầu vào thường thưa
  - Người dùng thường tương tác với một lượng nhỏ các item



# Mô hình tìm hiểu

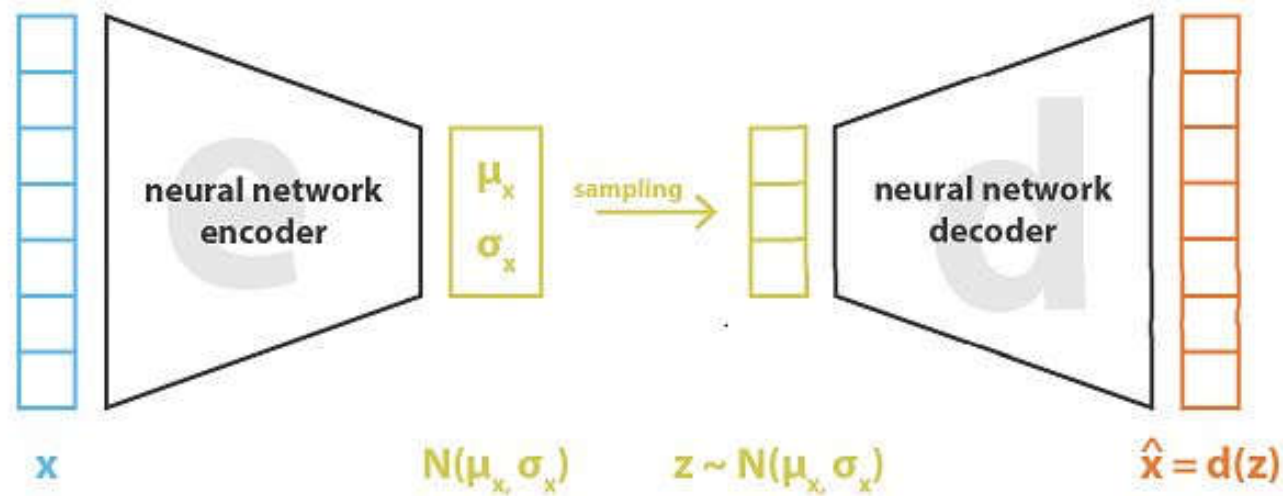
- “Variational Autoencoders for Collaborative Filtering” - Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, Tony Jebara
- Đây là một phương pháp tổng quan hơn so với các phương pháp đã trình bày trước đó
- Mô hình thể hiện ý tưởng chính của Collaborative Filtering
- Phù hợp với dữ liệu thưa

2.

# Kiến thức nền tảng

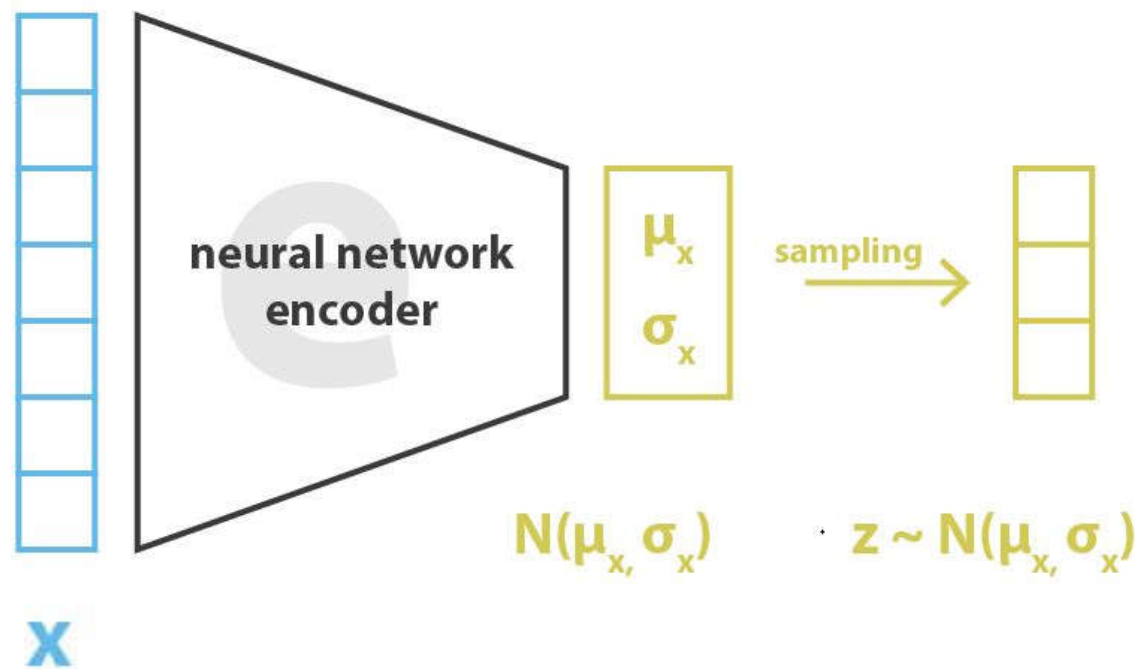
Variational inference và mô hình VAE

# Mô hình Variational Autoencoder



- Mô hình Variational Autoencoder là một mô hình phát sinh dữ liệu từ đặc trưng ẩn.
- Mô hình sẽ gồm 2 tiến trình:
  - Tiến trình suy diễn đặc trưng ẩn từ dữ liệu ta quan sát được (dữ liệu đầu vào)
  - Tiến trình phát sinh dữ liệu từ đặc trưng ẩn

# Tiến trình suy diễn



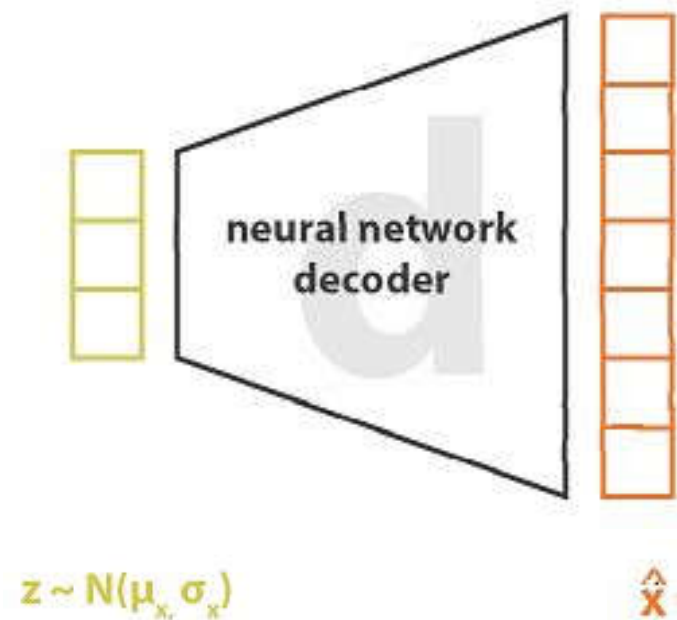
- Encoder (Inference model) trả về phân phối xác suất cho đặc trưng ẩn dựa trên dữ liệu quan sát được
- Giả định rằng  $\mathbf{z}_u \sim \mathcal{N}(0, \mathbf{I}_K)$

$$p(\mathbf{z}|\mathbf{x}) \approx q_{\phi}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_{\phi}(\mathbf{x}), \sigma_{\phi}^2(\mathbf{x}))$$



# Tiến trình phát sinh

- Decoder (Generative model) phát sinh dữ liệu từ đặc trưng ẩn có được



# Hàm mục tiêu của VAEs

$$\mathcal{L}_u(\theta, \phi) = \underbrace{\mathbb{E}_{q_\phi(z_u|x_u)} [\log p_\theta(x_u|z_u)]}_{\text{Độ lỗi tái tạo lại dữ liệu ban đầu}} - \underbrace{\mathcal{D}_{KL}(q_\phi(z_u|x_u) || p(z_u))}_{\text{Chuẩn hóa dữ liệu, đảm bảo tính chất của phân phối xác suất}}$$

- Trade-off: Đánh đổi giữa khả năng đặc trưng mô tả cho dữ liệu và đặc trưng ẩn tuân theo phân phối xác suất.
  - Đây là trade off thường gặp của VAEs khi dùng để phát sinh dữ liệu

# Kullback-leibler Divergence

$$\mathcal{D}_{KL}(q_\phi(z_u|x_u)||p(z_u)) = \mathbb{E}_{z_u \sim q_\phi}(\log q_\phi(z_u|x_u)) - \mathbb{E}_{z_u \sim q_\phi}(\log p(z_u))$$

- **Kullback-Leibler divergence** là (còn được gọi là Entropy tương đối) dùng để đo mức độ lệch của một phân bố đối với một phân bố được chỉ định.
  - $p(z_u)$  Sẽ được giả định là một phân phối cơ bản để có thể dễ dàng tính được

3.

# Mô hình Mul-VAE

Mở rộng Variational Autoencoder cho  
Collaborative Filtering

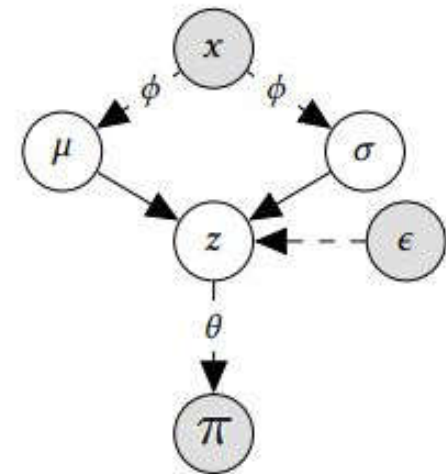
# Variational Autoencoder cho Collaborative Filtering

- $X \in \mathbb{N}^{U \times I}$  Là ma trận tương tác của người dùng
- $x_u = [x_{u1}, \dots, x_{uI}] \in \mathbb{N}^I$  Là vec-tơ tương tác của user  $u$  trên tập  $I$  item

		user			
		1	2	3	4
item	1	✓	?	?	✓
	2	✓	?	✓	✓
	3	?	✓	✓	?
	4	?	✓	?	✓

# Variational Autoencoder cho Collaborative Filtering

- Input của mô hình là dữ liệu tương tác của user trong quá khứ  $x_u$
- Encoder trả về phân phối xác suất được định nghĩa bởi  $\mu_\phi(x_u), \sigma_\phi(x_u)$
- Đặc trưng ẩn  $z_u$  được phát sinh từ phân phối xác suất trên
- Decoder nhận input là đặc trưng ẩn và trả về vector xác suất các item phù hợp với user  $\pi(z_u) \in \mathbb{R}^I$



# Variational Autoencoder cho collaborative Filtering

Siêu tham số

$$\mathcal{L}_u(\theta, \phi) = \mathbb{E}_{q_\phi(z_u|x_u)} [\log p_\theta(x_u|z_u)] - \beta \times \mathcal{D}_{KL}(q_\phi(z_u|x_u) || p(z_u))$$

- Với bài toán recommendation, siêu tham số  $\beta$  được dùng để kiểm soát việc mô hình dữ liệu và việc xấp xỉ phân phối xác suất theo giả định của mô hình.
  - Đối với hệ thống đưa ra sản phẩm, ta quan tâm tới việc đưa ra gợi ý hơn là việc đảm bảo dữ liệu tuân theo các tính chất của phân phối xác suất

# Multinomial likelihood

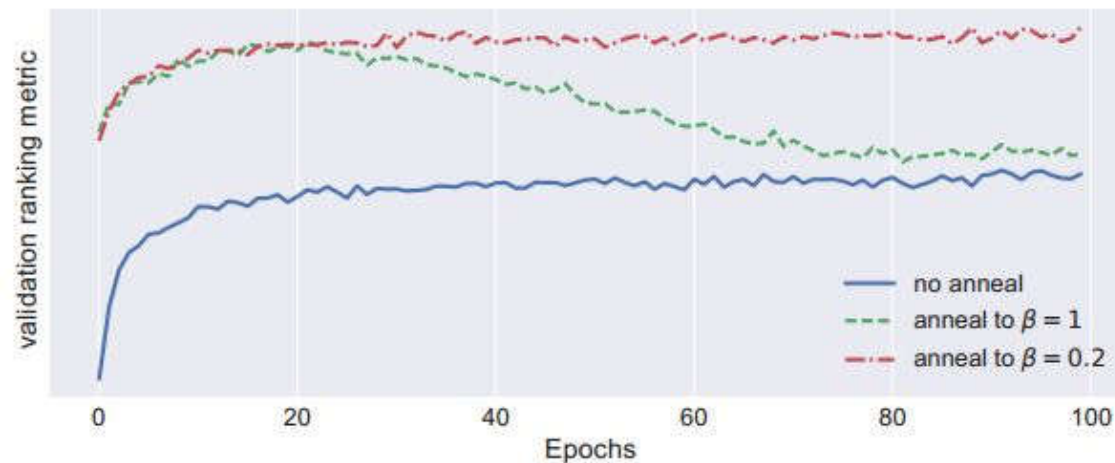
$$\log p_{\theta}(x_u | z_u) = \sum x_{ui} \log \pi_i(z_u)$$

- Là độ lỗi trong việc tái tạo lại dữ liệu tương tác của người dùng
- Với Multinomial likelihood , các tương tác của người dùng, hay các giá trị sẽ có tính chất xác suất trên toàn bộ các item
  - Tổng xác suất xuất của các item sẽ bằng 1
  - Phù hợp với bài toán gợi ý sản phẩm hơn các hàm likelihood khác
    - Gaussian log-likelihood <sup>1</sup>
    - Logistic log-likelihood <sup>2</sup>

- 1 Còn được biết đến là Least Square Error.  
- 2 Logistic log-likelihood chính là Binary Cross-entropy



# KL - Annealing



- Ta tăng dần giá trị  $\beta$  qua mỗi lần cập nhật trọng số, thay vì huấn luyện mô hình với các giá trị siêu tham số khác nhau.
  - Ban đầu  $\beta$  khởi tạo bằng 0
  - Tăng dần  $\beta$  sau mỗi lần cập nhật trọng số

4.

# Thí Nghiệm

Các kết quả thí nghiệm của mô hình

# Thanks!

## Any questions?

You can find me at:

@username

user@mail.me