

Nội dung

1. Tổng quan
2. Xây dựng hệ thống gợi ý dựa trên mô hình Autoencoder
3. Xây dựng hệ thống gợi ý dựa trên mô hình Variational Autoencoder
4. Thí nghiệm
5. Kết luận

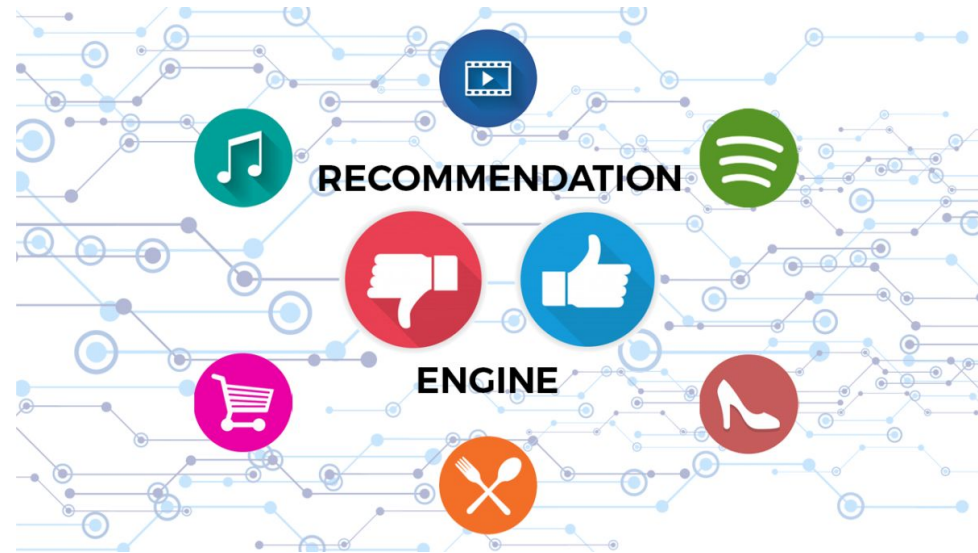
1.

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

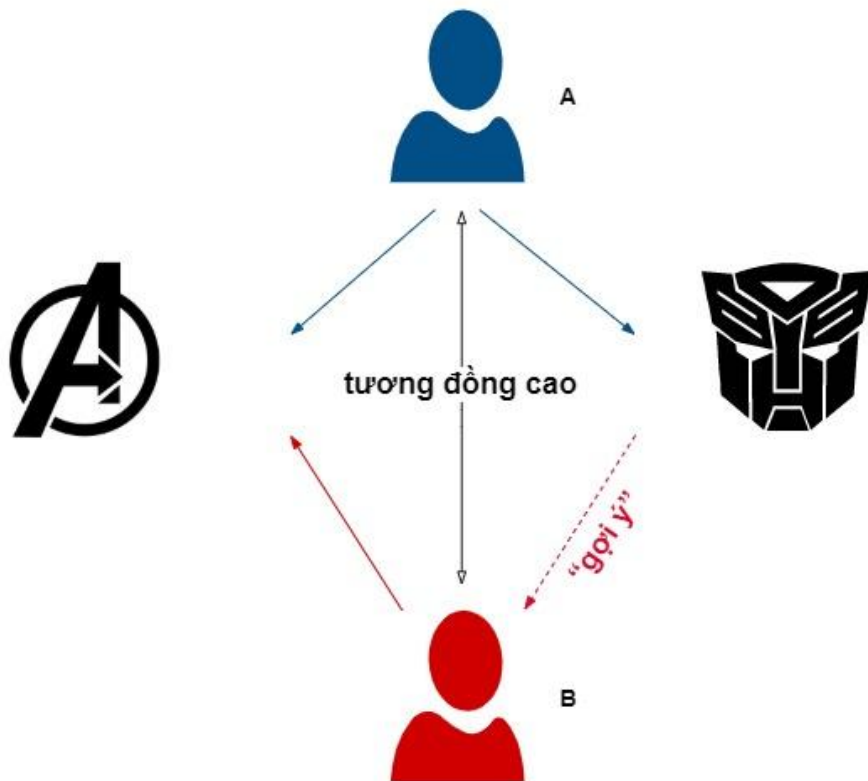


Phát biểu bài toán

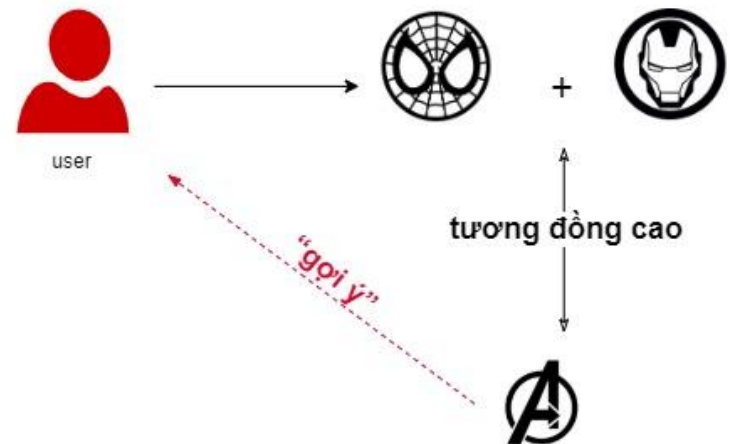
- ❖ Cho input là dữ liệu về lịch sử tương tác của người dùng (user) với các sản phẩm (item) hoặc có thêm các mô tả của sản phẩm
- ❖ Yêu cầu: đưa ra tập các item (không có trong lịch sử) được dự đoán là phù hợp với người dùng

Hướng tiếp cận

Collaborative Filtering

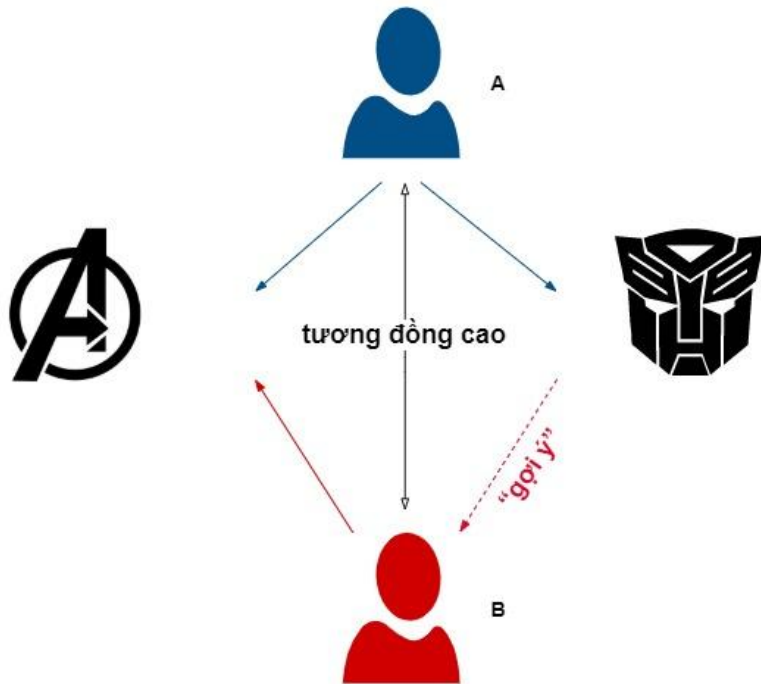


Content-based Filtering



Hướng tiếp cận tìm hiểu

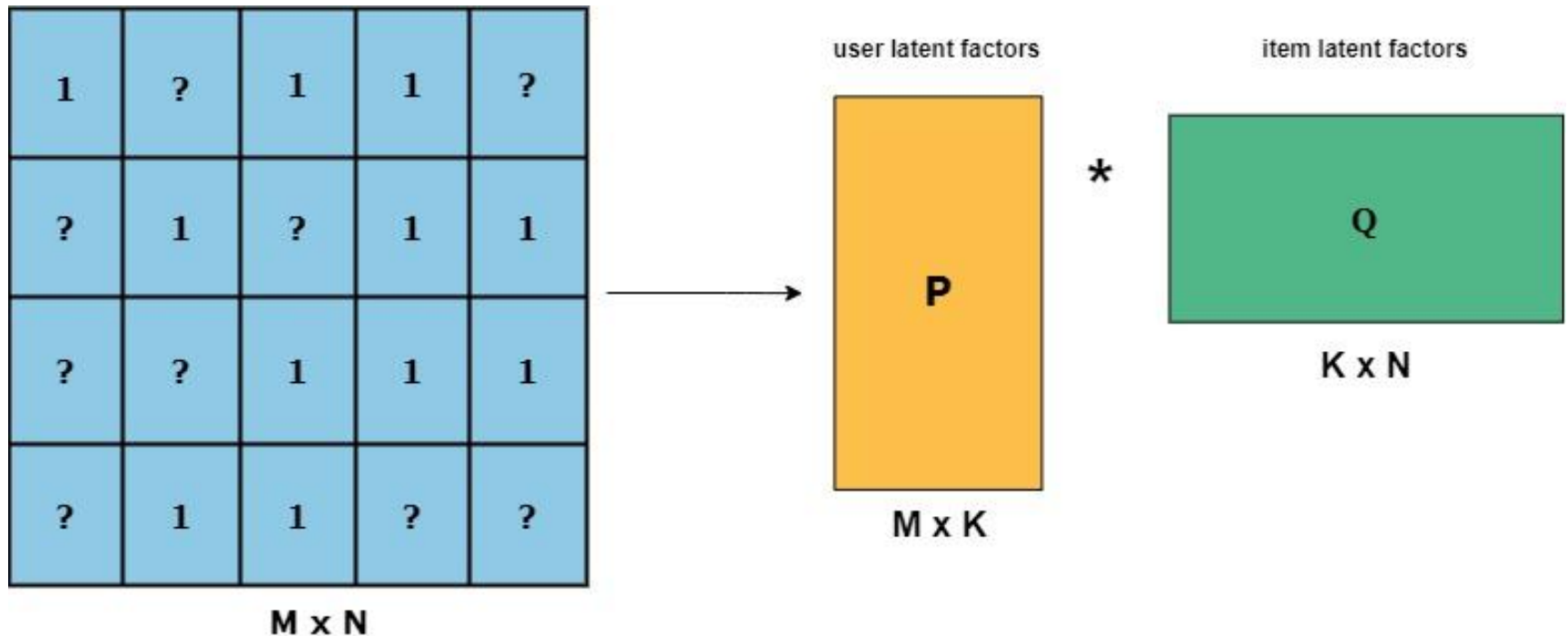
Collaborative Filtering



- ❖ Tổng quát hóa tốt hơn
- ❖ Đưa ra gợi ý đa dạng hơn, tạo ra sự tình cờ cho người dùng
- ❖ Là hướng nghiên cứu đang được quan tâm

Các nghiên cứu liên quan

Matrix Factorization



➡ Đây là một phương pháp khá tốn kém chi phí

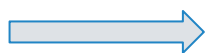
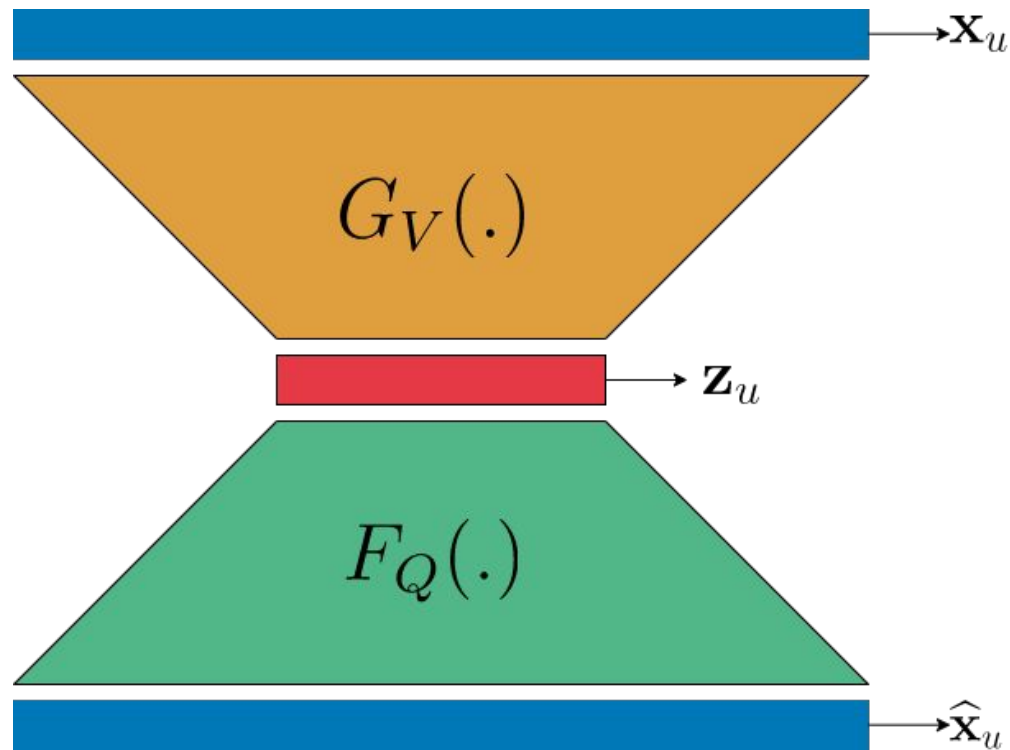
Các nghiên cứu liên quan

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 đặc trưng ẩn của item được tương tác**
- Phát sinh các gợi ý tương tự *Matrix Factorization*
- Số lượng tham số của mô hình tỉ lệ với số lượng sản phẩm

Các nghiên cứu liên quan

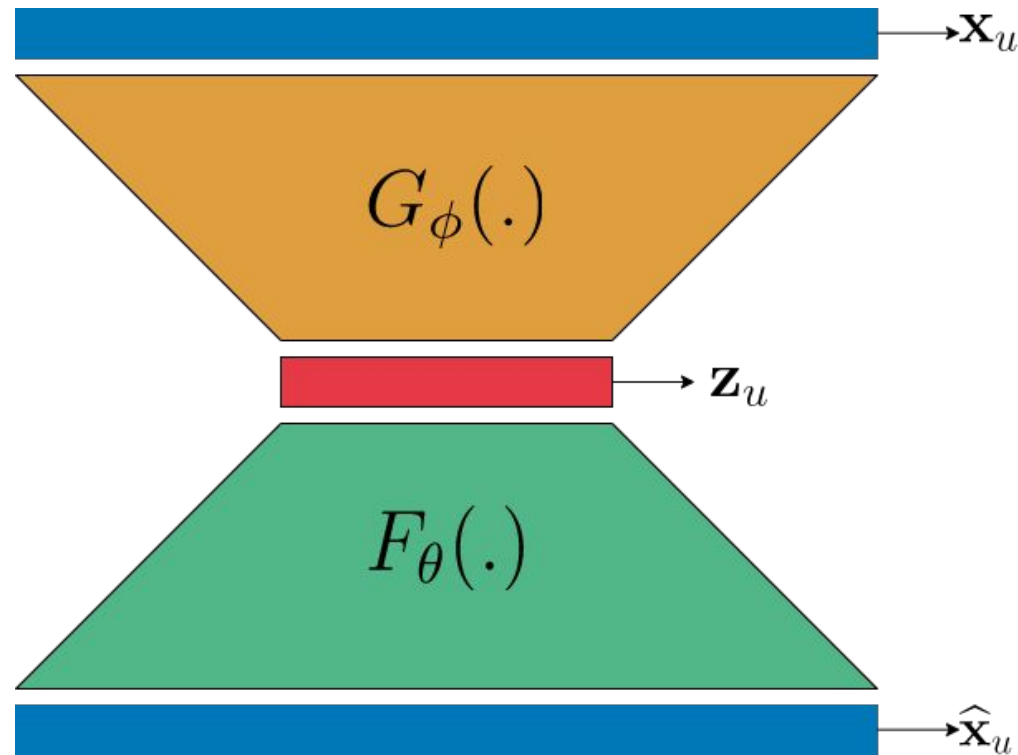
Asymmetric Matrix Factorization



Đây là kiến trúc của một Linear Autoencoder

Các nghiên cứu liên quan

Autoencoder



➡ Tận dụng sức mạnh của hàm phi tuyến (mạng nơ-ron)

Mô hình khóa luận tìm hiểu

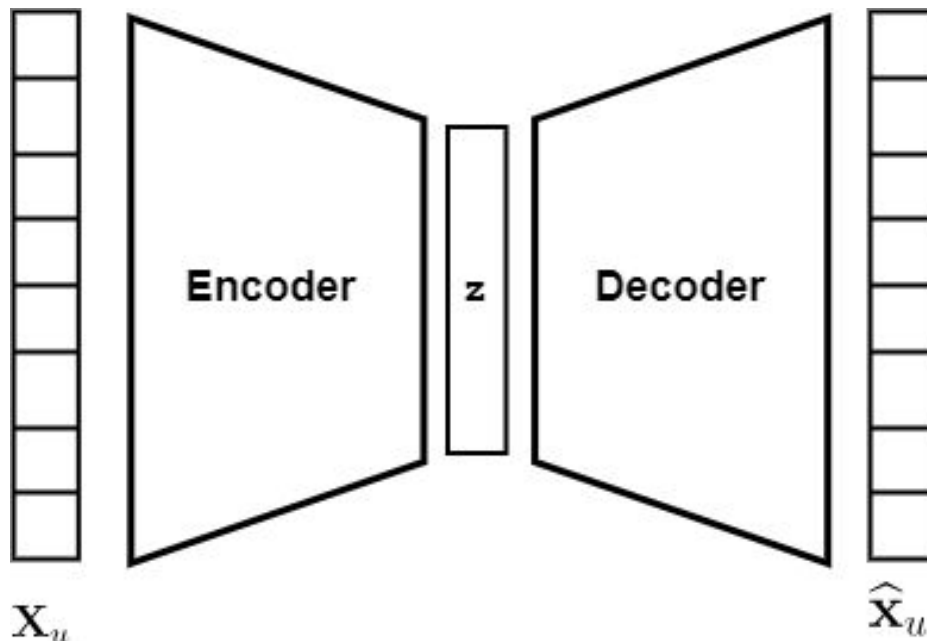
- ❖ Bài báo “Variational Autoencoders for Collaborative Filtering” được công bố tại hội nghị “International World Wide Web Conference Committee 2018”
- ❖ Mô hình nổi bật nhất trong nhóm các mô hình Autoencoder
- ❖ Đặc trưng ẩn của người dùng sẽ được phát sinh từ một phân phối xác suất

2.

Xây dựng hệ thống gợi ý dựa trên mô hình Autoencoder

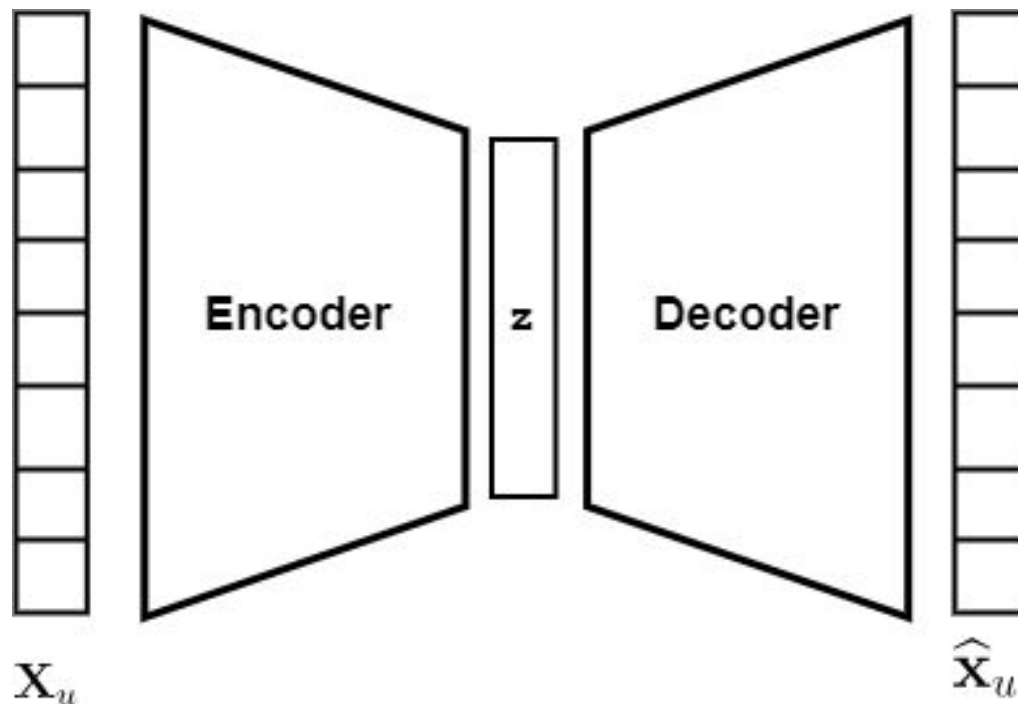
Mô hình Autoencoder

- Mô hình Autoencoder là một mô hình có khả năng học cách biểu diễn đặc trưng ẩn của dữ liệu
 - Giảm chiều dữ liệu
 - Trích xuất các đặc trưng ẩn



- Encoder: ánh xạ vector input \mathbf{x} sang vector biểu diễn ẩn \mathbf{z}
- Decoder: tái tạo lại input \mathbf{x} từ vector biểu diễn ẩn \mathbf{z}
- Thông thường, \mathbf{z} sẽ có kích thước nhỏ hơn so với \mathbf{x}

Mô hình Autoencoder cho bài toán xây dựng hệ thống gợi ý sản phẩm



- Đầu tiên, encoder sẽ rút trích các đặc trưng ẩn của người dùng từ vector tương tác của họ
- Decoder sẽ tái tạo lại tương tác của người dùng từ đặc trưng ẩn z

Hàm mục tiêu của Autoencoder

$$\mathcal{L}(\phi, \theta) = \frac{1}{U} \sum_{x_u \in X} (\|x_u - h_{\phi, \theta}(x_u)\|_2^2) + \frac{\lambda}{2} \times (\|\phi\|_2^2 + \|\theta\|_2^2)$$

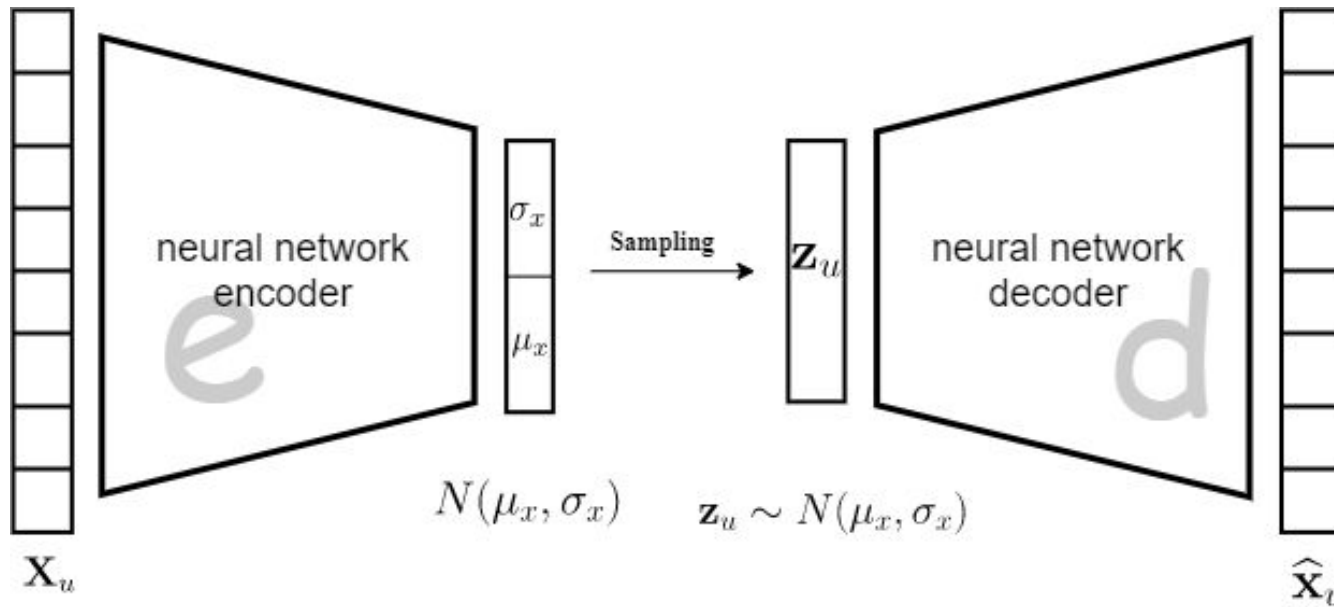
Trong đó:

- U là số lượng người dùng, x_u là tương tác của người dùng
- λ là hệ số regularization

3.

**Xây dựng hệ thống gợi ý
dựa trên mô hình
Variational Autoencoder**

Mô hình Variational Autoencoder



- Variational Autoencoder là một Autoencoder mà quá trình huấn luyện được chuẩn hoá để tránh overfitting và đảm bảo được đặc trưng ẩn có thể được dùng để phát sinh dữ liệu mới
- Đặc trưng ẩn của mô hình là một phân phối xác suất thay vì là một điểm dữ liệu

Hàm mục tiêu của VAE

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

Độ lỗi tái tạo lại dữ liệu ban đầu

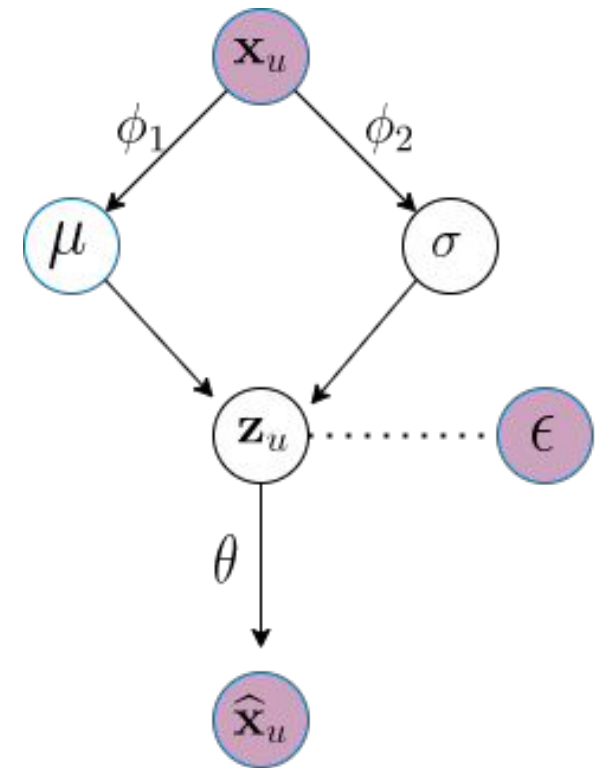
Chuẩn hóa dữ liệu, đảm bảo tính chất của phân phối xác suất

Trong đó:

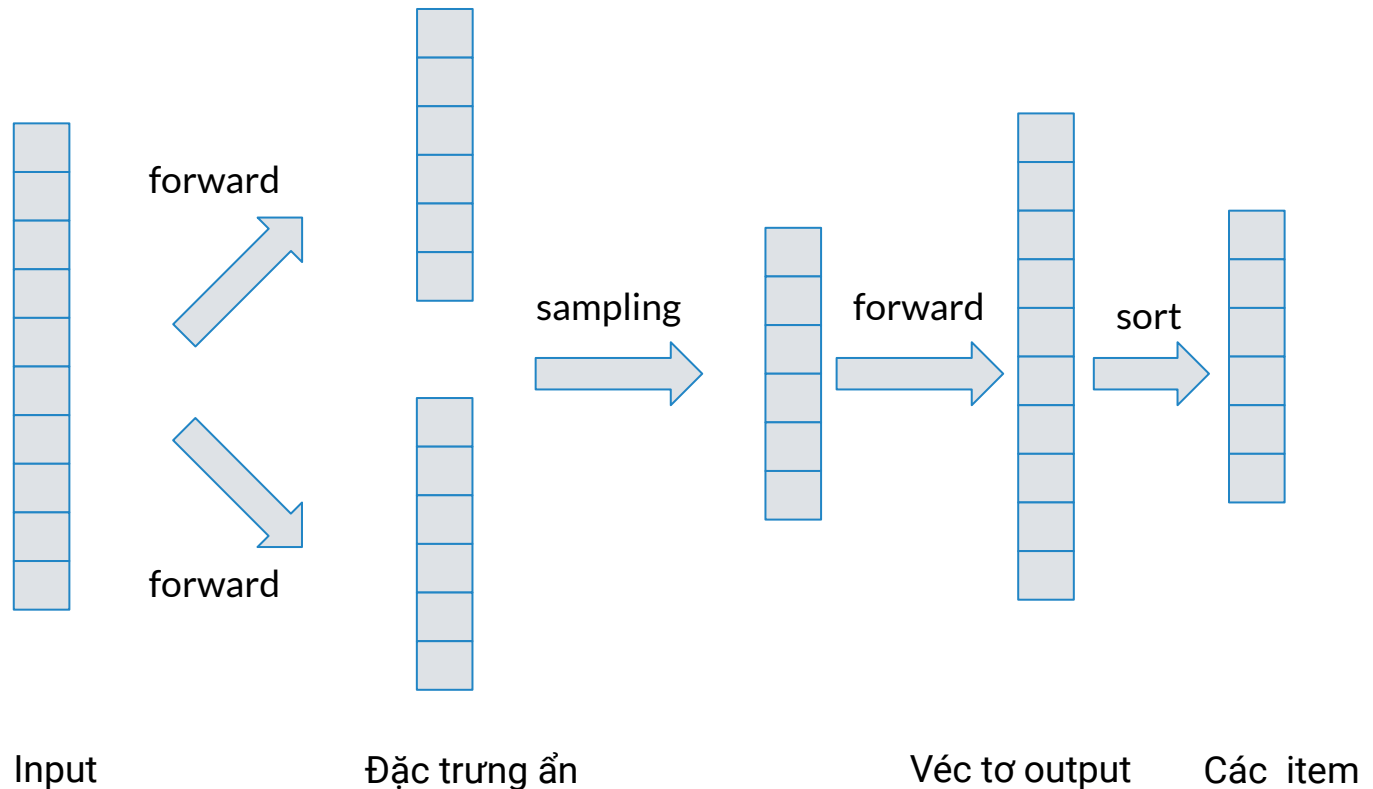
- p là phân phối giả định của chúng ta về đặc trưng ẩn
- q là phân phối trả về từ mô hình

Mô hình Variational Autoencoder cho bài toán xây dựng hệ thống gợi ý sản phẩm

- Input của mô hình là dữ liệu tương tác của người trong quá khứ x_u
- Encoder trả về phân phối xác suất được định nghĩa bởi $\mu_{\phi_1}(x_u)$, $\sigma_{\phi_2}(x_u)$
- 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 tương tác được tái tạo lại của người dùng
- Dựa trên tương tác được tái tạo lại, mô hình sẽ đưa ra gợi ý cho người dùng



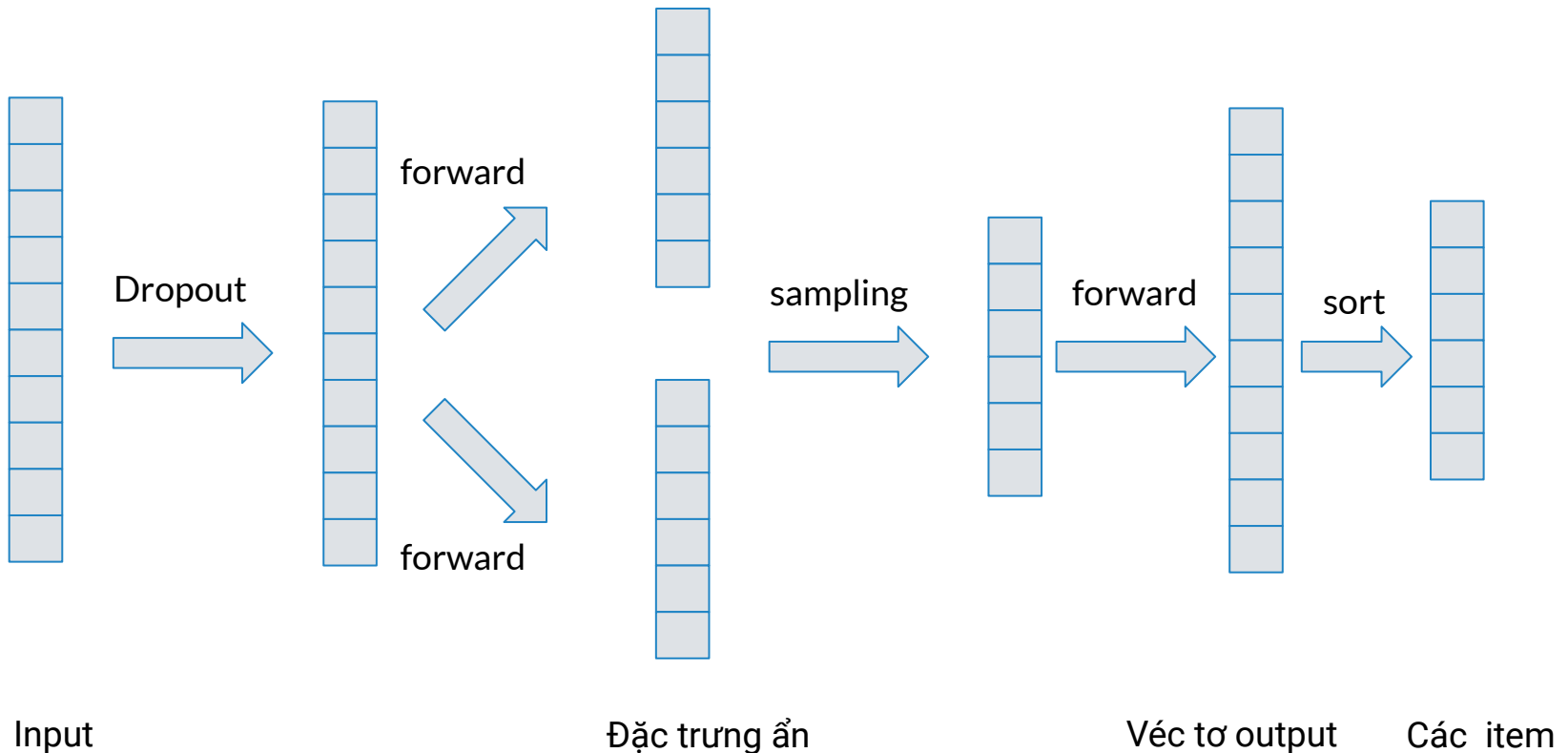
Mô hình Variational Autoencoder cho bài toán xây dựng hệ thống gợi ý sản phẩm



Thay đổi input trong quá trình huấn luyện bằng kỹ thuật dropout

- Trong quá trình huấn luyện, ta thêm nhiễu vào dữ liệu bằng cách ngẫu nhiên ẩn một vài tương tác hay còn được gọi là dropout
 - Sử dụng dropout như một kỹ thuật tránh tình trạng overfitting
 - Mô hình sẽ phải học cách dự đoán các sản phẩm cho người dùng


Thay đổi input trong quá trình huấn luyện bằng kỹ thuật dropout



Hàm mục tiêu khi huấn luyện VAE cho bài toán gợi ý sản phẩm

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

Siêu tham số

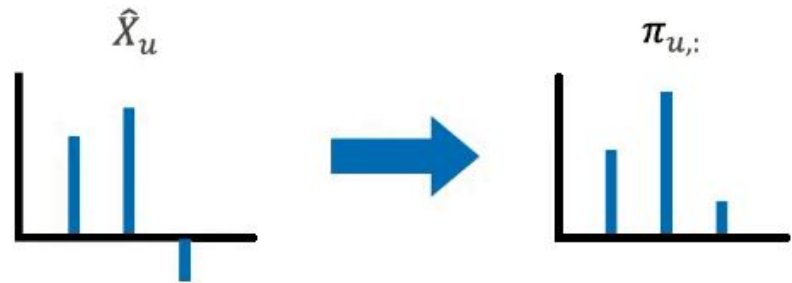


- Với hệ thống gợi ý sản phẩm, siêu tham số β được dùng để kiểm soát việc mô hình dữ liệu và việc xấp xỉ phân phối xác suất theo giả định của mô hình
 - Đối với hệ thống đưa ra sản phẩm, ta quan tâm tới việc đưa ra gợi ý hơn là việc đảm bảo dữ liệu tuân theo các tính chất của phân phối xác suất

Hàm lỗi Multinomial likelihood

$$\pi_{ui} = \frac{\exp(\hat{X}_{ui})}{\sum_j \exp(\hat{X}_{uj})} \quad \square$$

$$\mathcal{L} = - \sum_{u,i} X_{ui} \log \pi_{ui}$$



- Với Multinomial likelihood, ta giả định rằng:

$$\mathbf{x}_u \sim \text{Mult}(N_u, \pi(\mathbf{z}_u))$$

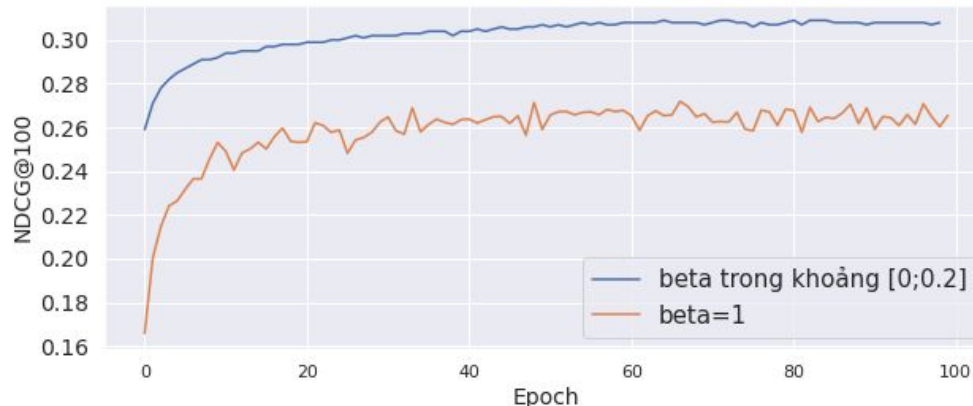
- Có nghĩa là các item được tương tác sẽ tuân theo xác suất tương ứng trong vector $\pi(\mathbf{z}_u)$

Tại sao chọn Multinomial log-likelihood

- Với bài toán gợi ý sản phẩm ta quan tâm đến việc tìm xếp hạng giữa các item hơn
- Với Multinomial log-likelihood, các tương tác của người dùng, hay các giá trị trả về sẽ có tính chất xác suất trên toàn bộ các item
 - Tổng xác suất xuất của các item sẽ bằng 1
 - Phù hợp hơn so với các hàm likelihood khác trong việc xếp hạng các item
 - Gaussian log-likelihood [1]
 - Logistic log-likelihood [2]

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

Kỹ thuật KL - Annealing



- Là phương pháp “heuristic” để lựa chọn siêu tham số β thay vì thực hiện huấn luyện trên nhiều mô hình với các giá trị β khác nhau
- Ta tăng dần giá trị β qua mỗi lần cập nhật trọng số, cho đến một giá trị giới hạn của β
 - Ban đầu β khởi tạo bằng 0
 - Tăng dần β sau mỗi lần cập nhật trọng số

4.

Thí Nghiệm

Các kết quả thí nghiệm của mô hình

Độ đo để đánh giá

- Recall@k
 - Đánh giá tỉ lệ item đúng trong top-k item được mô hình dự đoán
 - Miền giá trị $[0,1]$, giá trị càng cao thì càng tốt
- NDCG@k
 - Đánh giá chất lượng xếp hạng của mô hình
 - Miền giá trị $[0,1]$, giá trị càng cao thì càng tốt

Tập dữ liệu sử dụng

	MovieLens-20M (ML-20M)	MSD
Số lượng user	136,667	571,355
Số lượng item	20,108	44,140
Số lượng tương tác	10M	33.6M
% tương tác	0.36%	0.14%
Số lượng user trong tập kiểm thử và kiểm tra (held-out)	10,000	50,000

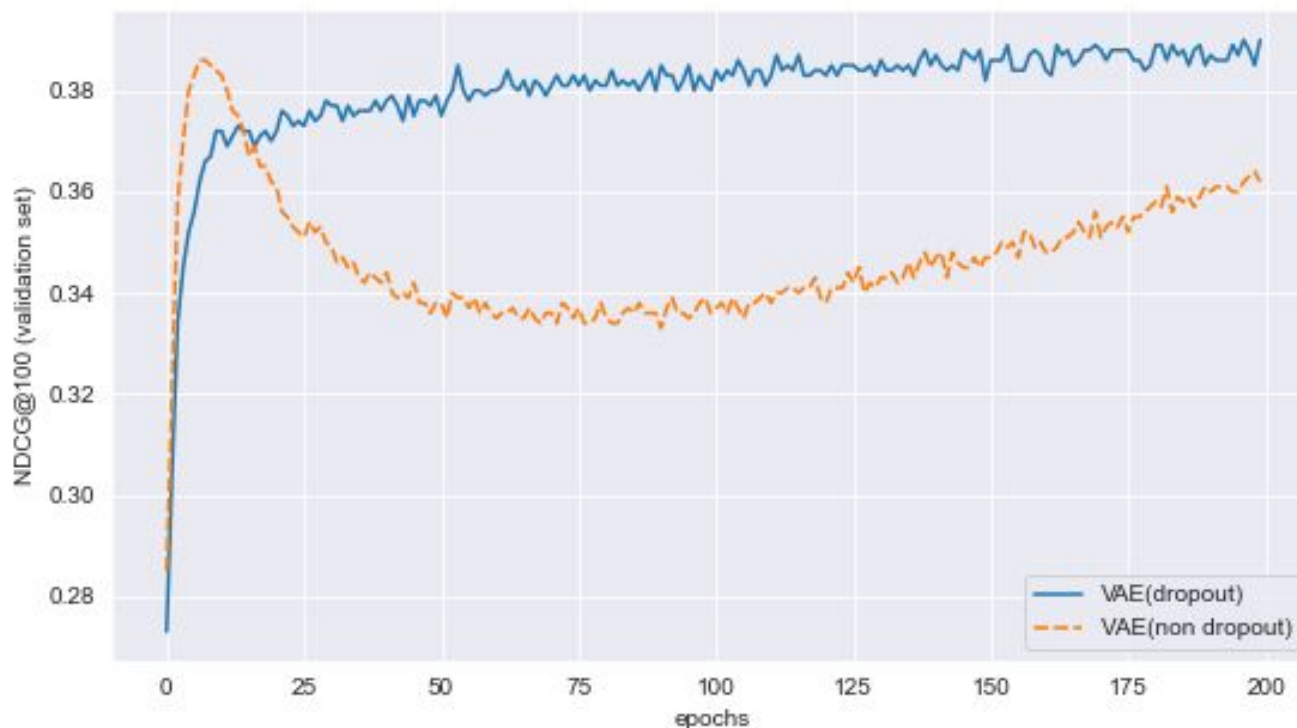
Kết quả thí nghiệm

	NDCG@100	Recall@50	Recall@20	Dữ liệu
Cài đặt của tác giả [*] 10,000 held-out user	0.426	0.537	0.395	ML-20M
Cài đặt của chúng tôi (10,000 held-out user)	0.429	0.537	0.395	
Cài đặt của tác giả [*] 50,000 held-out user	0.266	0.364	0.316	MSD
Cài đặt của chúng tôi (50,000 held-out user)	0.257	0.353	0.308	

Kết quả trên mô hình VAE có áp dụng dropout và thay đổi hàm mục tiêu khi huấn luyện

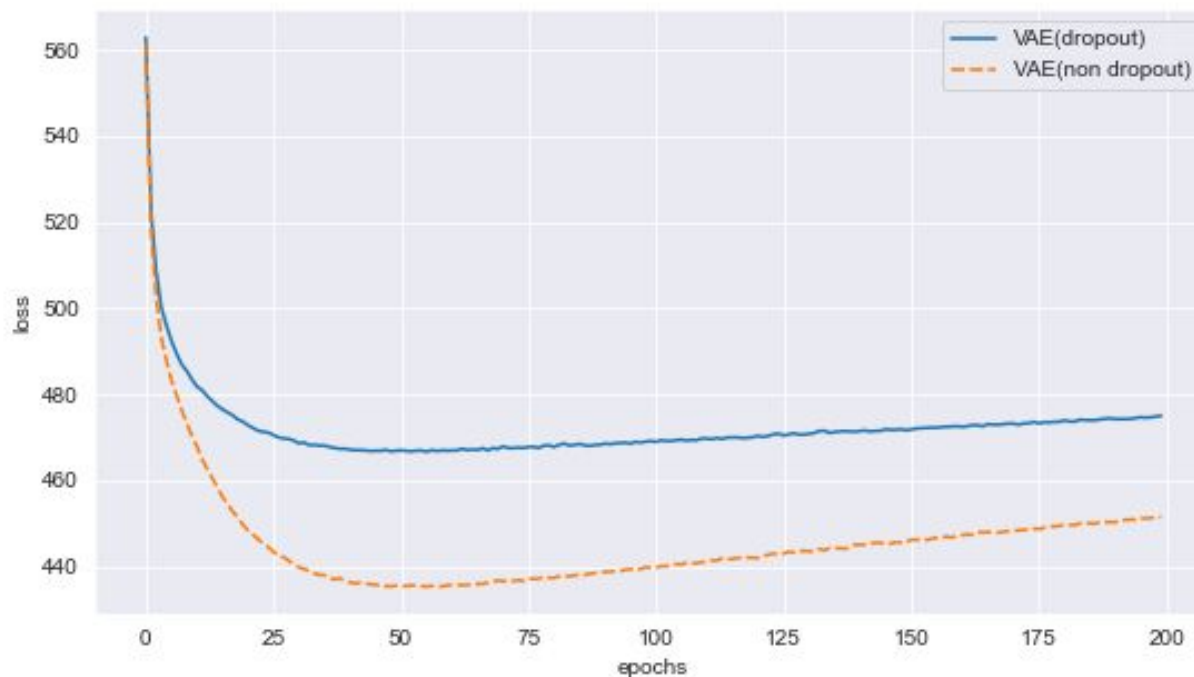
[*] "Variational Autoencoders for Collaborative Filtering" - Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, Tony Jebara

Thí nghiệm 1: hiệu quả của kỹ thuật dropout trong quá trình huấn luyện



Kết quả của mô hình VAE khi áp dụng dropout và không áp dụng dropout trên tập dữ liệu ML-20M với độ đo NDCG@100

Thí nghiệm 1: hiệu quả của kỹ thuật dropout trong quá trình huấn luyện



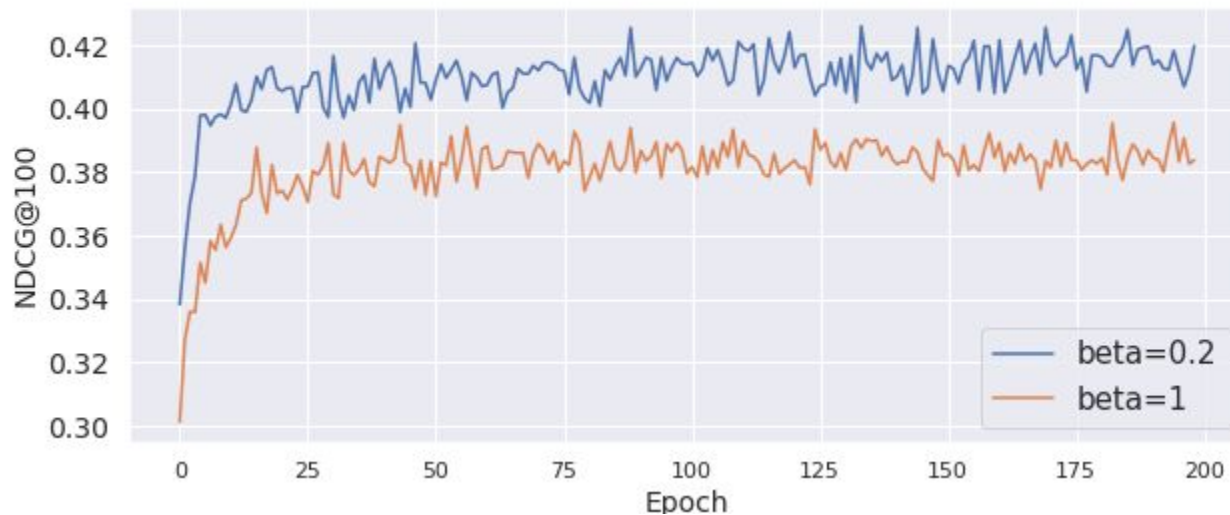
Độ lỗi của mô hình VAE trong quá trình huấn luyện khi áp dụng dropout và không áp dụng dropout trên tập dữ liệu ML-20M

Thí nghiệm 2: hiệu quả của hàm loss Multinomial likelihood

	Multinomial	Gaussian	Logistic	Dữ liệu
VAE + dropout	0.429	0.422	0.419	ML-20M
AE + dropout	0.423	0.409	0.412	

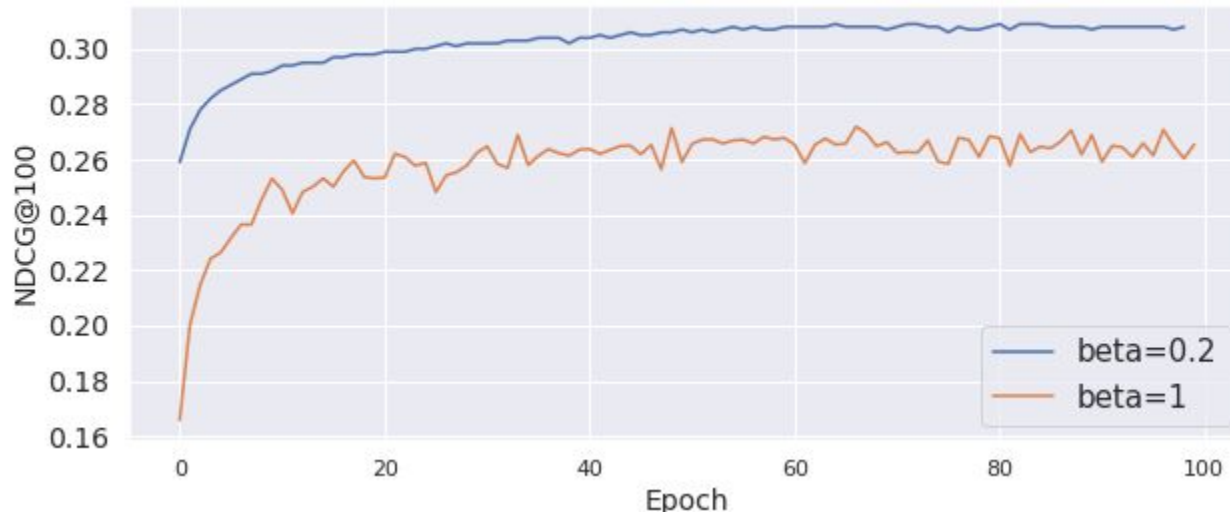
Kết quả trên mô hình VAE có áp dụng dropout với các hàm loss khi huấn luyện trên tập dữ liệu MovieLens với độ đo NDCG@100

Thí nghiệm 3: hiệu quả của việc có thêm siêu tham số β



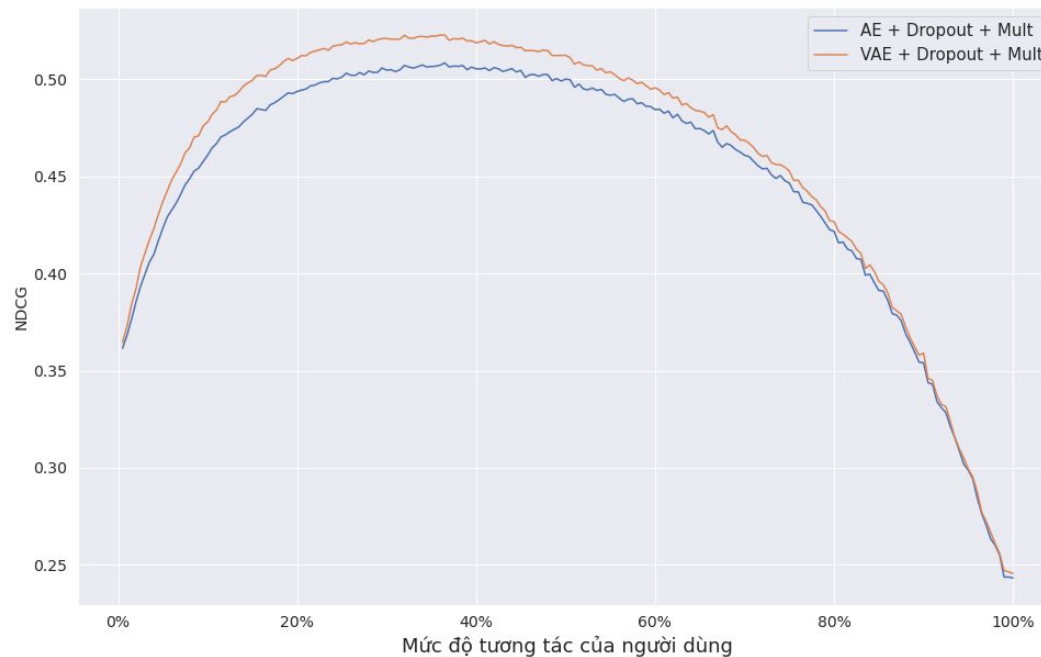
So sánh hiệu quả của siêu tham số β trên tập dữ liệu ML-20M

Thí nghiệm 3: hiệu quả của việc có thêm siêu tham số β



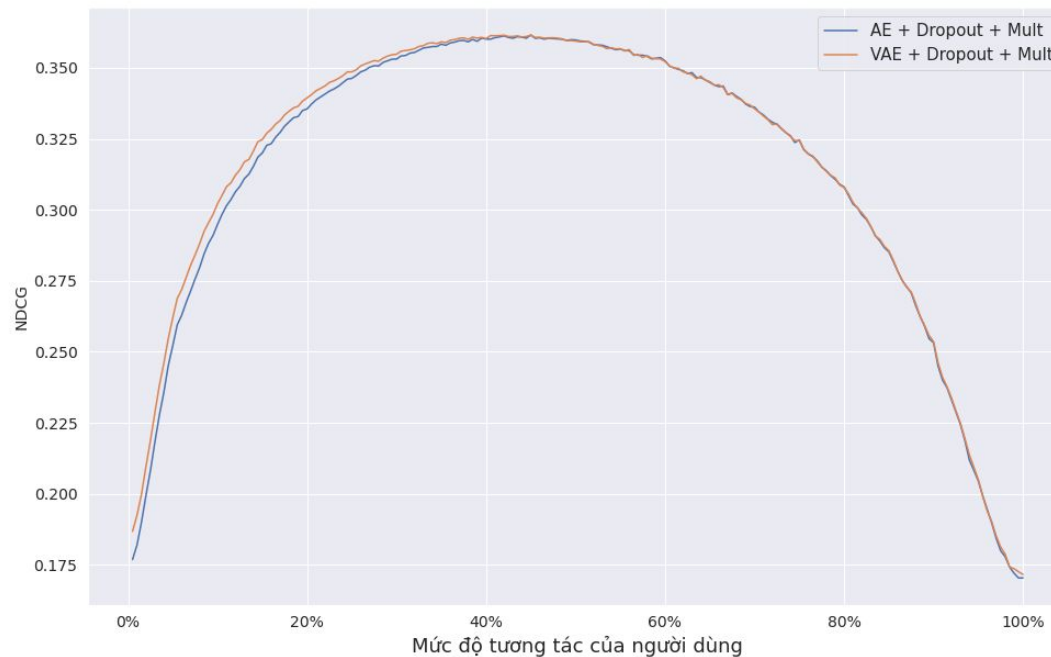
So sánh hiệu quả của siêu tham số β trên tập dữ liệu MSD

Thí nghiệm 4: hiệu quả của mô hình VAE đối với dữ liệu thưa



Kết quả so sánh VAE và DAE trên dữ liệu thưa ở tập MovieLens với độ đo NDCG@100

Thí nghiệm 4: hiệu quả của mô hình VAE đối với dữ liệu thưa



Kết quả so sánh VAE và DAE trên dữ liệu thưa ở tập Million Song Datasets với độ đo NDCG@100

5. Tổng kết

Tổng kết và hướng phát triển

Tổng kết

❖ Tổng kết

- Tìm hiểu và cài đặt mô hình có kết quả xấp xỉ công bố trong bài báo “Variational Autoencoders for Collaborative Filtering”
- Thực hiện các thí nghiệm để phân tích khả năng gợi ý sản phẩm mô hình

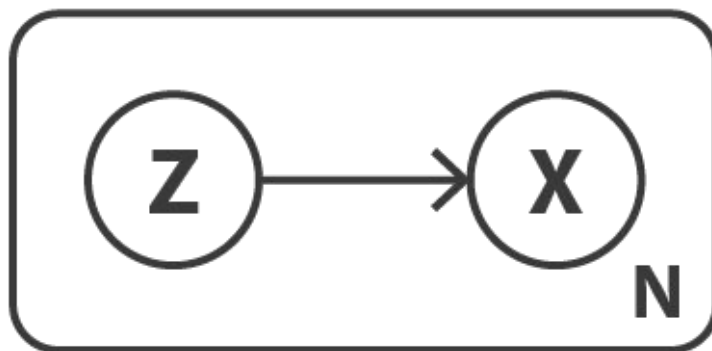
❖ Hướng phát triển

- Tìm hiểu cách kết hợp thêm thông tin mô tả của sản phẩm
- Cải thiện khả năng đưa ra gợi ý của mô hình



Cảm ơn quý thầy cô đã lắng nghe!

Phụ lục: Mô hình xác suất



- Xét mô hình xác suất dùng để phát sinh dữ liệu như sau:
 - Đầu tiên, ta phát sinh đặc trưng ẩn z từ một phân phối $p(z)$
 - Tiếp theo, x được phát sinh từ phân phối xác suất có điều kiện $p(x|z)$
- x là dữ liệu ta có được, z ta không quan sát được
 - Ta cần suy diễn z dựa vào dữ liệu x có được hay tìm phân phối xác suất có điều kiện $p(z|x)$

Phụ lục: Phương pháp Variational Inference

Posterior Likelihood Prior

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

Evidence Chi phí tính toán cao

➡ Thay vì tính trực tiếp, ta sẽ xấp xỉ phân phối $p(z|x)$

Phụ lục: Phương pháp Variational Inference

Posterior Likelihood Prior

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

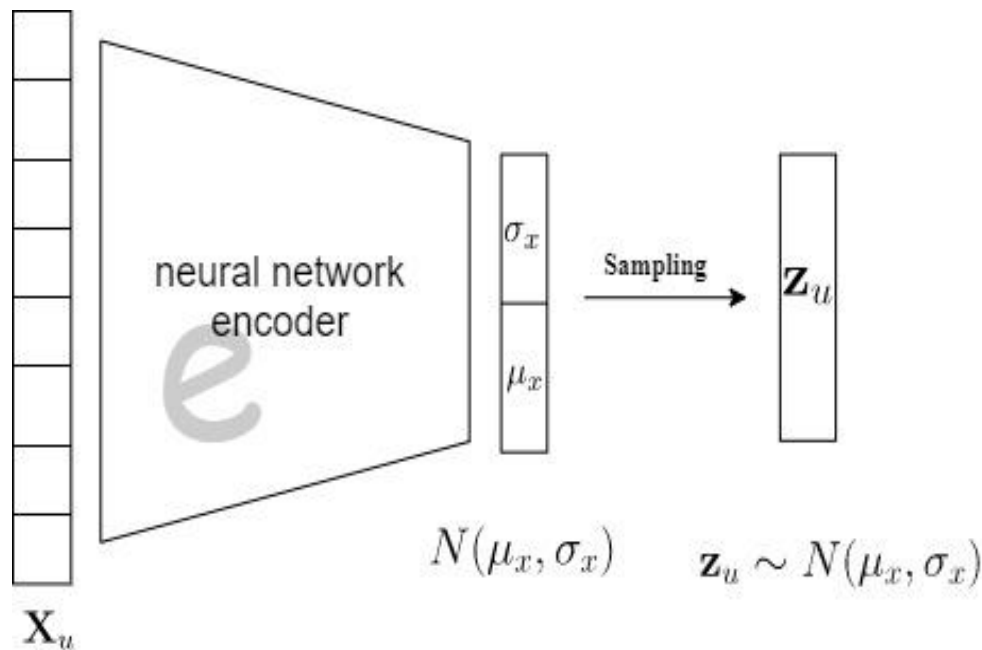
Evidence Chi phí tính toán cao

➡ Thay vì tính trực tiếp, ta sẽ xấp xỉ phân phối $p(z|x)$

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

- **Ý tưởng:** ta học bộ tham số ϕ theo dữ liệu x có được sao cho phân phối $q_\phi(z|x)$ “gần” với $p(z|x)$

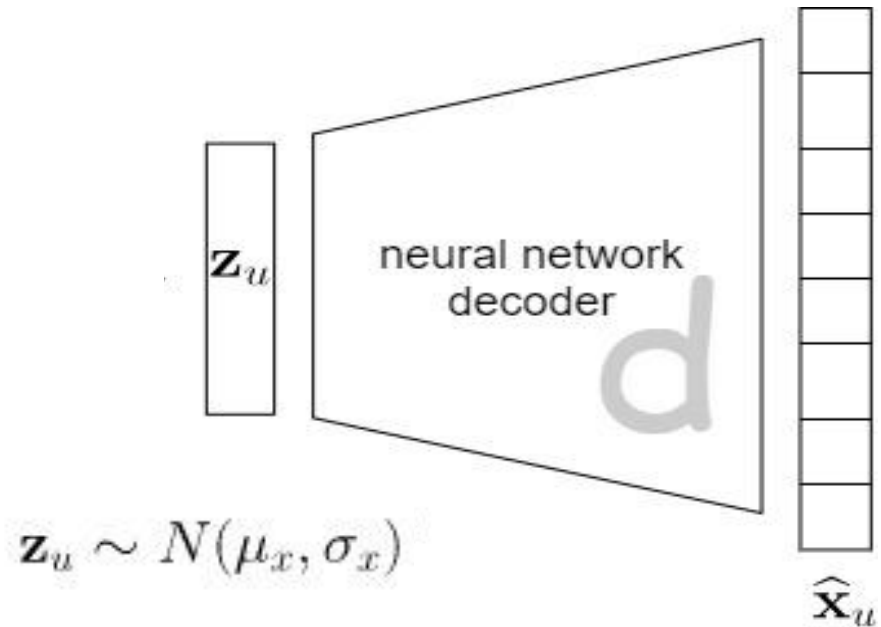
Phụ lục: Tiến trình suy diễn của mô hình VAE



- Encoder (Inference model) trả về phân phối xác suất cho đặc trưng ẩn dựa trên dữ liệu quan sát được

Phụ lục: Tiến trình phát sinh của mô hình VAE

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



Phụ lục: Độ sai biệt Kullback-leibler

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

- 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
 - $p(z)$ Sẽ được giả định là một phân phối cơ bản để có thể dễ dàng tính được