RecSys 2019 Short Paper

Enhancing VAEs for Collaborative Filtering: Flexible Priors & Gating Mechanisms

Daeryong Kim & Bongwon Suh

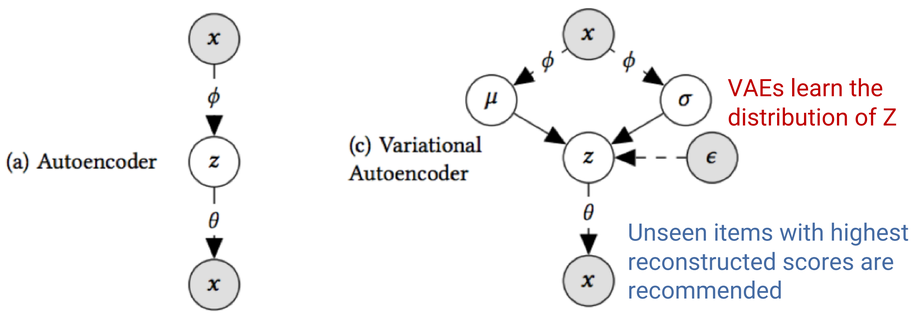
Human Centered Computing Lab,
Graduate School of Convergence Science and Technology,
Seoul National University

daeryong@snu.ac.kr, bongwon@snu.ac.kr

INTRODUCTION

- Variational Autoencoders for Collaborative Filtering (Liang et al. @WWW '18) showed State-Of-The-Art performance in the context of collaborative filtering.

VAEs for Collaborative Filtering (Liang et al. @WWW'18)



The baseline model of our research, our goal is to further improve model performance under this framework

- We aim to overcome some potentially problematic characteristics of VAEs in the task of collaborative filtering and appropriately tailor VAEs to further improve model performance and make high quality recommendations.

MOTIVATION

- The standard normal prior distribution used in VAEs may be too restrictive due to its simple unimodal nature, hindering the models from learning richer latent variables of user preference which is crucial to model performance.
- Preceding research using autoencoders for CF make use of relatively shallow networks. Learning from user-item interaction history has its own characteristics and may have more effective architectures to learn deeper latent representations.

Research Goal

Experiment the effect of **Flexible Priors**, **Hierarchical VAEs** & **Gating Mechanisms** in the context of collaborative filtering and further improve model performance under the VAE-CF framework.

ENHANCING VAES FOR CF

Flexible Priors

VampPrior. We experiment with a recently proposed flexible prior called the Variational Mixture of Posteriors prior (Tomzak et al.), originally proposed for image generation, to examine the effect of flexible priors in CF

- ELBO

$$\mathcal{L}(\phi,\theta,\lambda) = \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})]\right] + \quad (4)$$
$$+ \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\mathbb{H}[q_\phi(\mathbf{z}|\mathbf{x})]\right] + \quad (5)$$
$$- \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[-\ln p_\lambda(\mathbf{z})] \quad (6)$$

Prior Distribution chosen in advance

Cross Entropy $(q(z), p_\lambda(z))$

$q(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^{N} q_\phi(\mathbf{z}|\mathbf{x}_n)$ *Aggregated posterior*

- Optimal Prior

$$p^*_\lambda(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^{N} q_\phi(\mathbf{z}|\mathbf{x}_n).$$

- VampPrior

$$p_\lambda(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}|\mathbf{u}_k)$$

Trainable Pseudo Inputs $(K \ll N)$

VampPrior is an approximation of the optimal prior maximizing the ELBO

RESULTS 01) *Model Performance*

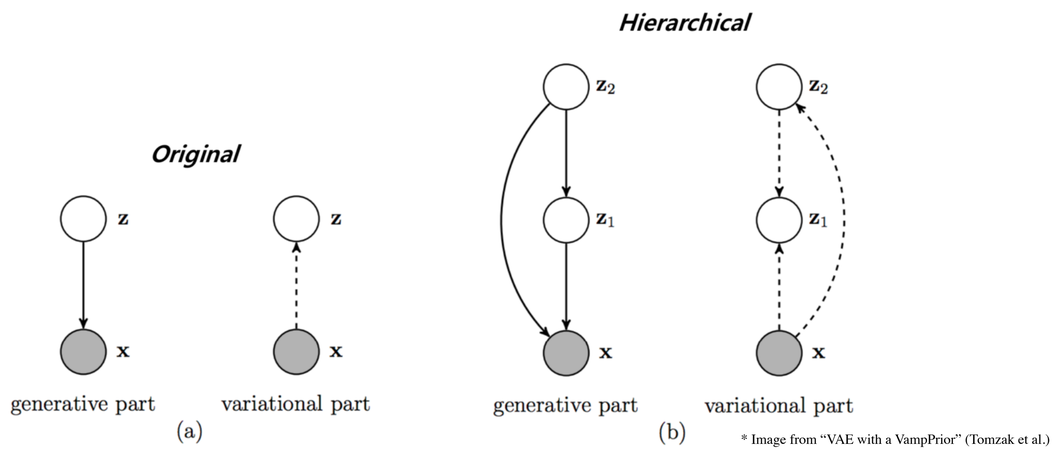
- Adopting the VampPrior, HVAE and Gating Mechanism **sequentially improves model performance**
- Our final model *H+Vamp (Gated)* significantly outperforms the strongest baseline *Mult-VAE* producing new **state-of-the-art results**

	MovieLens 20M			Netflix		
Models	NDCG@100	Recall@50	Recall@20	NDCG@100	Recall@50	Recall@20
WMF †	0.386	0.498	0.360	0.351	0.404	0.316
SLIM †	0.401	0.495	0.370	0.379	0.428	0.347
CDAE †	0.418	0.523	0.391	0.376	0.428	0.343
Mult-VAE	0.42700	0.53524	0.39569	0.38711	0.44427	0.35255
Vamp	0.43433	0.53933	0.40310	0.39589	0.44907	0.36327
H+Vamp	0.43684	0.53974	0.40524	0.40242	0.45605	0.37090
Mult-VAE (Gated)	0.43515	0.54498	0.40558	0.39241	0.44958	0.35953
H+Vamp (Gated)	**0.44522**	**0.55109**	**0.41308**	**0.40861**	**0.46252**	**0.37678**

Table 1: Results for MovieLens 20M and Netflix dataset. Standard errors are around 0.002 for ML-20M and 0.001 for Netflix.
†Results are taken from (Liang et al.), note that our datasets, metrics and experimental settings are consistent with (Liang et al.).

Hierarchical VAE

We also adopt **Hierarchical Stochastic Units** to learn even richer latent representations. The original stochastic latent variable $\boldsymbol{z}$ is replaced by a stacked hierarchical structure of $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$



* Image from "VAE with a VampPrior" (Tomzak et al.)

Gating Mechanism

Gated Linear Units. We experiment with a non-recurrent gating mechanism proposed in Gated CNNs (Dauphin et al.) which was suggested to help information propagation in deeper networks :

Element-wise Product

$$h_l(\mathbf{X}) = (\mathbf{X} * \mathbf{W} + \mathbf{b}) \otimes \sigma(\mathbf{X} * \mathbf{V} + \mathbf{c})$$

Linear Transformation *Gates*

EXPERIMENTAL SETUP

Experiments were conducted on the MovieLens-20M and Netflix prize dataset. Both datasets were binarized by keeping only ratings of 4 or higher. All models are fully tuned with grid search on possible hyperparameter values.

Models

Mult-VAE	Regular VAE for CF proposed by Liang et al. (Baseline Model)
Vamp	VAE + VampPrior
H + Vamp	VAE + VampPrior + HVAE
H + Vamp (Gated)	VAE + VampPrior + HVAE + Gated Linear Units (Final Model)

RESULTS 02) *Effect of Gating*

- For models with no gates, increasing the depth does not bring performance gain while for gated models it does
- Adding gates without additional layers also boost performance due to the higher-level interactions the self-attentive gates allow

Netflix (NDCG@100)	No-Gate	Gated
Mult-VAE (1 Layer)	**0.38711**	0.39229
Mult-VAE (2 Layer)	0.38359	**0.39241**
Vamp (1 Layer)	**0.39589**	0.40169
Vamp (2 Layer)	0.39346	**0.40277**
H + Vamp (1 Layer)	**0.40242**	0.40728
H + Vamp (2 Layer)	0.37970	**0.40861**

Table 2: Comparison of performance between Gated and Un-Gated for models of different depth. The model with better performance (1 Layer vs 2 Layers) is marked in bold.