ContrastVAE: Contrastive Variational AutoEncoder for Sequential Recommendation. CIKM’22

Yu Wang, Hengrui Zhang, Zhiwei Liu, Liangwei Yang, Philip S. Yu

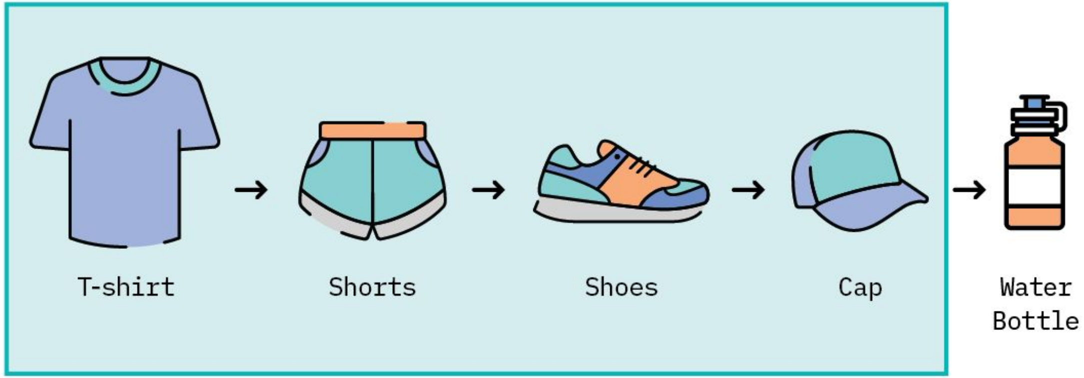
@2023/02 Chia-Jen, Yeh

Report Structure

1. *Cause & Effect*
2. *Proposed Method*
3. *Experiment Result*

Cause & Effect

Background



Sequential Recommendation (SR) has attracted increasing attention due to its ability to model the temporal dependencies in ***users' clicking histories***, which can help better understand user behaviors and intentions.

Background

Recent research justifies the promising ability of self-attention models in characterizing the temporal dependencies on real-world sequential recommendation tasks.

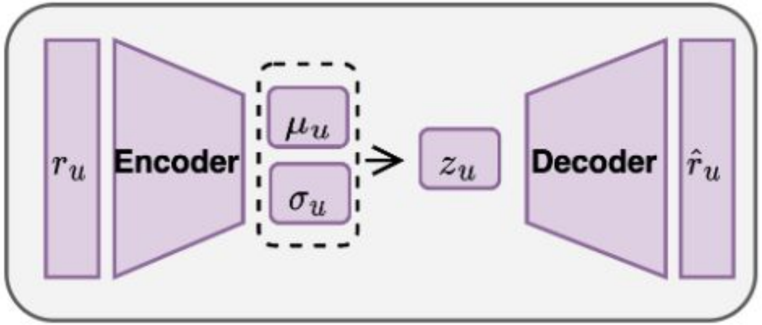
Research	Abstract
SASRec **ICDM'18 WC Kang**	SASRec is a pioneering work adopting the self-attention mechanism to learn transition patterns in item sequences
TiSASRec **WSDM'20 JC Li et al.**	TiSASRec is a time-interval aware version of SASRec
BERT4Rec **CIKM'19 Fei Sun et al.**	BERT4Rec extends it as a bi-directional encoder to predict the next item

Background

Despite their great representation power, both the **Uncertainty problem** and the **Sparsity Issue** impair their performance.

Problem	Definition	Example
Uncertainty Problem	Due to the rigorous assumption of sequential dependencies, which may be destroyed by unobserved factors in real-world scenarios.	For music recommendations, the genre of music that a user listens may vary according to different circumstances. Nevertheless, those factors are unknown and cannot be fully revealed in sequential patterns
Sparsity Issue	Sparsity Issue is a long-existing and not yet a well-solved problem in recommender systems	Supposing that a user only interacts with a few items, current methods are unable to learn high-quality representations of the sequences, thus failing to characterize sequential dependencies

VAE in Sequential Recommendation

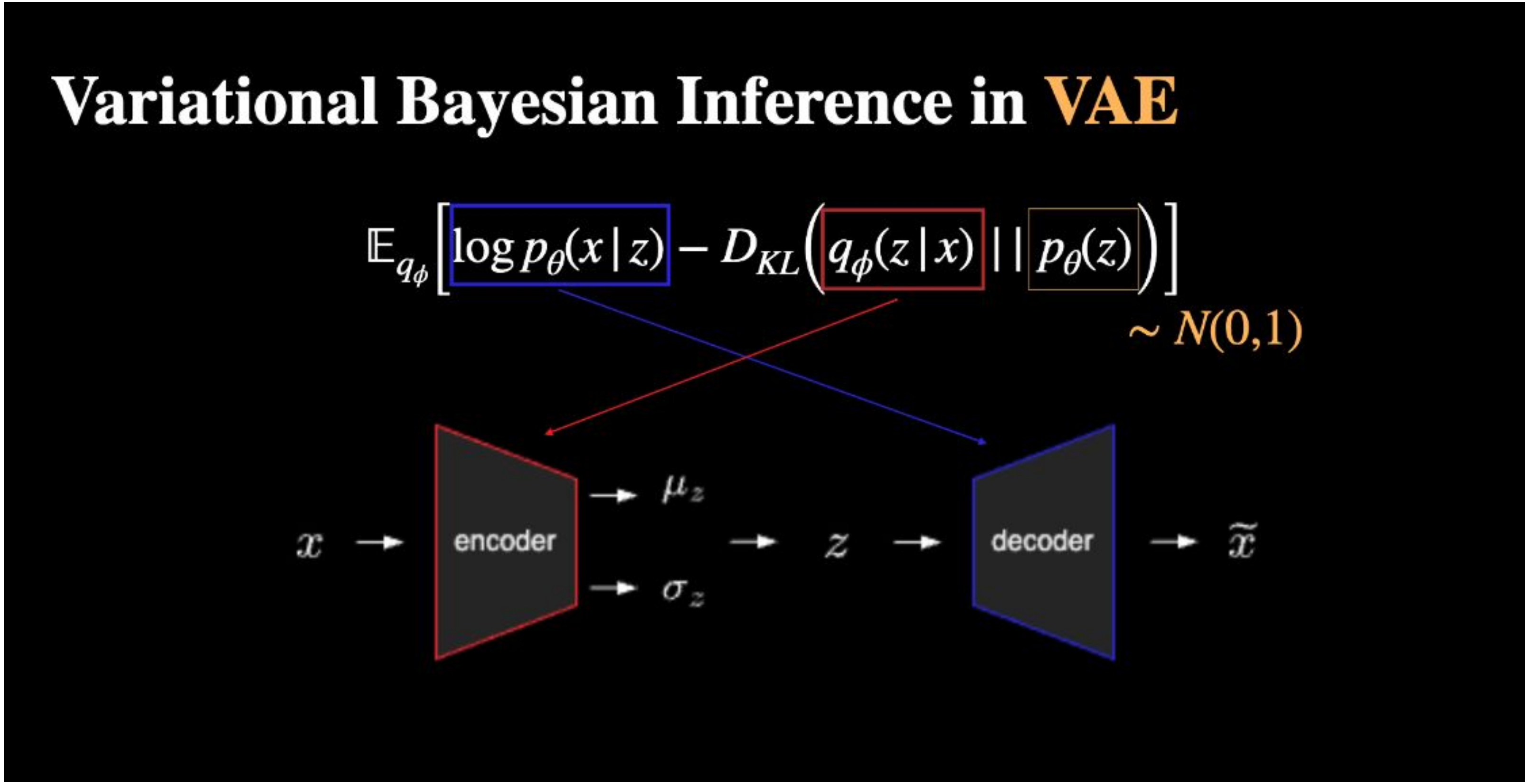


VAE can estimate the uncertainty of the input data.

More specifically, it characterizes the distributions of those **hidden representations** via an encoder-decoder learning paradigm, which assumes that those representations follow a **Gaussian distribution**.

Hence, the variances in Gaussian Distribution can well **characterize the uncertainty of the input data**.

ELBO Review


Variational Bayesian Inference in VAE
$\mathbb{E}_{q_\phi}\left[\boxed{\log p_\theta(x|z)} - D_{KL}\left(\boxed{q_\phi(z|x)} \,||\, \boxed{p_\theta(z)}\right)\right]$
$\sim N(0,1)$
x
encoder
μ_z
σ_z
z
decoder
$\tilde{x}$

The pain point of VAE: Posterior Collapse

However, conventional VAE suffers from **posterior collapse** issues.

If the **decoder** is sufficiently expressive, **the estimated posterior distributions of latent factors tend to resemble the standard Gaussian distributions**

Specifically, the sequential input data consists of **long-tail items**, which refer to the infrequent items that rarely appear in the users' historical records.

These limitations prevent VAE from achieving satisfactory performance for SR tasks.

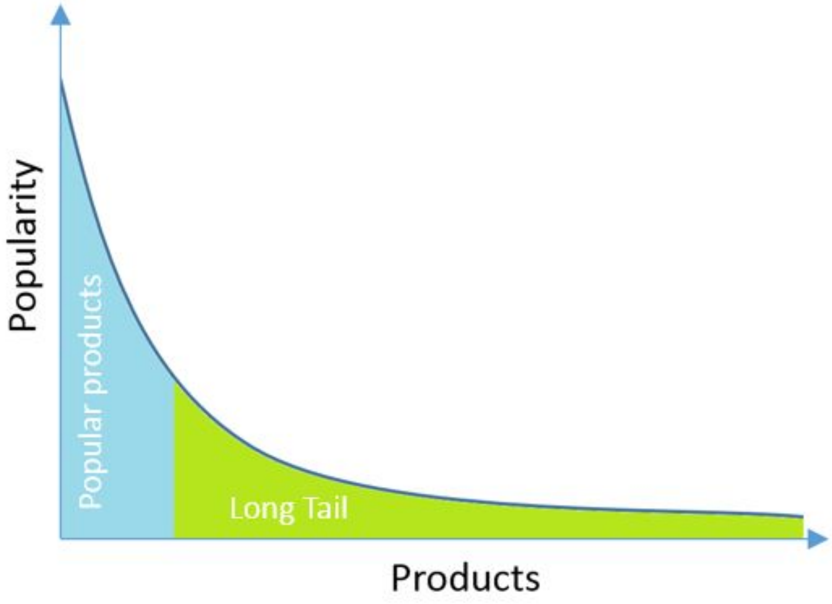


Illustration of long-tail items

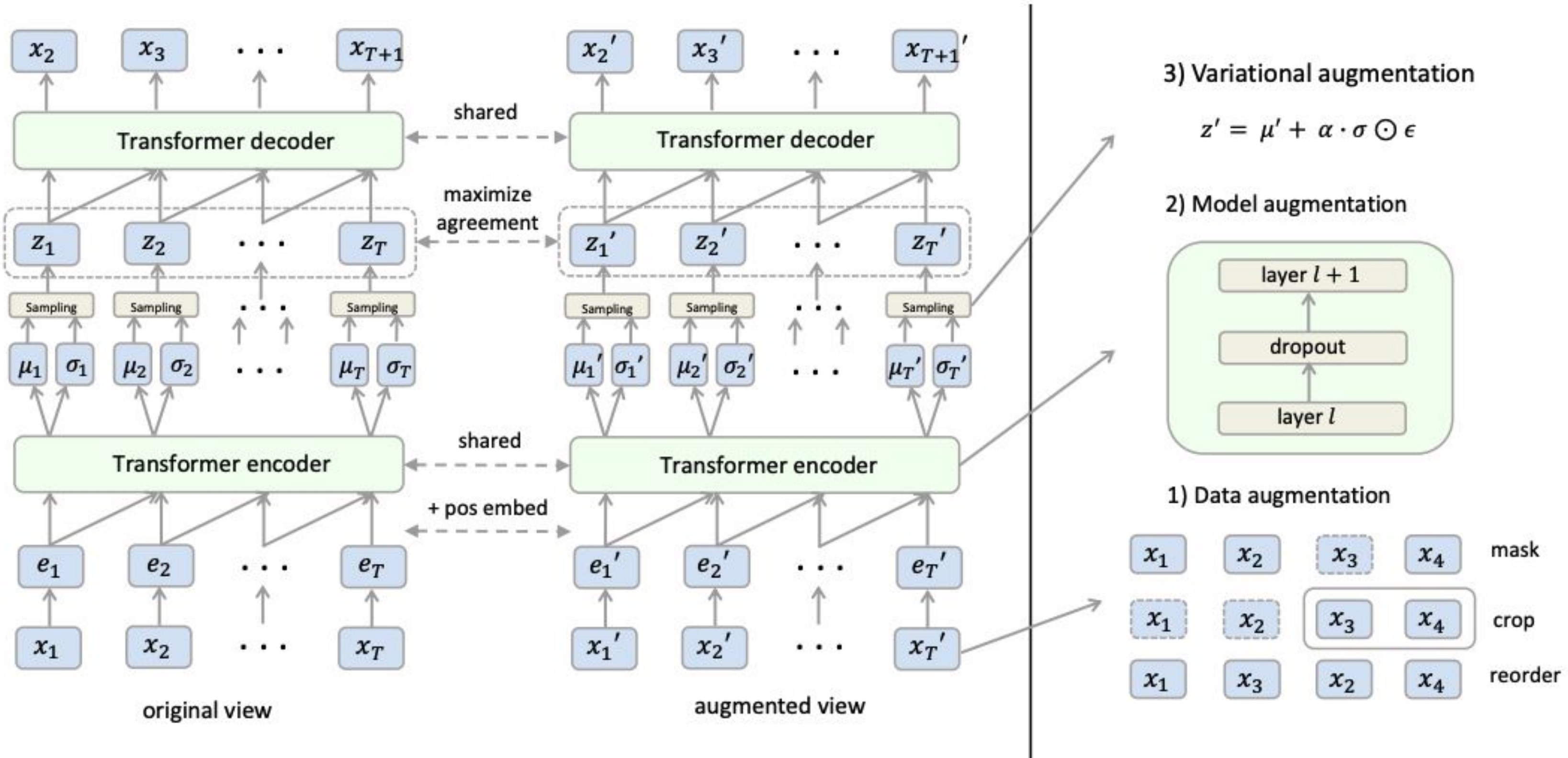
Previous studies to address Posterior Collapse

Method	Related Research
Reducing the impact of the KL-divergence term by Reducing its weight	● WWW'18. VAECF ● CIKM'21. ß-VAE ● ICDE'21. VSAN
Introducing an Additional Regularization Term that explicitly maximizes the mutual information between the input and latent representation	● ICML'20. A Simple Framework for Contrastive Learning of Visual Representations ● WSDM'22. DuoRec ● ICML'20. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere
Using Empirical Bayes that observed data to estimate the parameters of a prior distribution	● WSDM '21. BiVAE

But this paper find that these methods are **insufficient for better performance on the SR**

Proposed Method

Proposed Framework



(a) Framework of ContrastVAE

(b) Strategies of augmentations

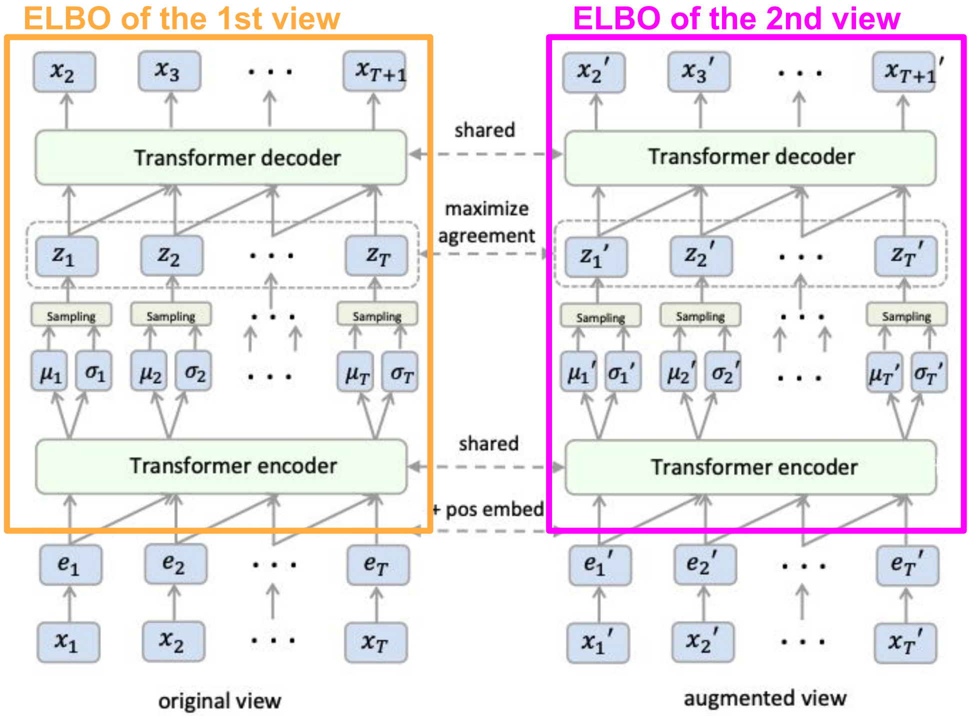
ConstrastELBO

$$
\begin{aligned}
\log p(x, x') \geq\ & \mathbb{E}_{q(z|x)} \log p(x|z) - D_{KL}[q(z|x)||p(z)] \\
&+ \mathbb{E}_{q(z'|x')} \log p(x'|z') - D_{KL}[q(z'|x')||p(z')] \\
&+ \mathbb{E}_{q(z,z'|x,x')} \log \left[\frac{p(z,z')}{p(z)p(z')} \right]
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{L} =\ & \mathcal{L}_{CE} - \mathcal{L}_{KL} \quad \text{(ELBO of the 1st view)} \\
&+ \mathcal{L}'_{CE} - \mathcal{L}'_{KL} \quad \text{(ELBO of the 2nd view)} \\
&+ \lambda \cdot \mathcal{L}_{InfoNCE} \quad \text{(InfoNCE)}
\end{aligned}
$$

The proof of this equation can be found in **Appendix 1**

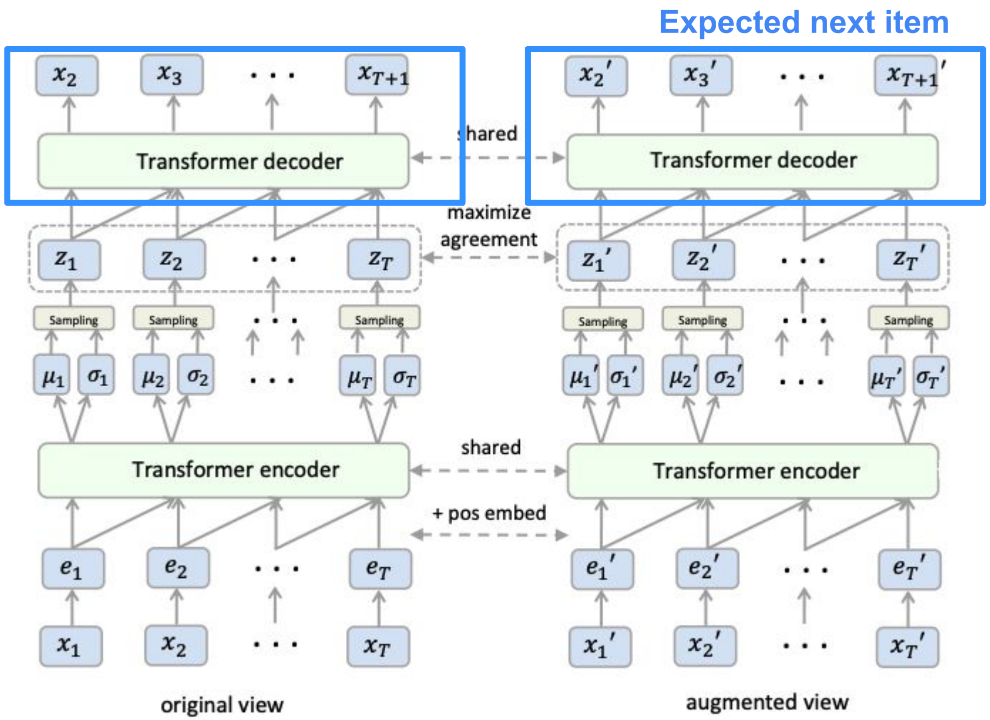
Proposed Framework



(a) Framework of ContrastVAE

$$\begin{aligned}\mathcal{L} = &\ \boxed{\mathcal{L}_{CE} - \mathcal{L}_{KL} \quad \text{(ELBO of the 1st view)}} \\ &+ \boxed{\mathcal{L}'_{CE} - \mathcal{L}'_{KL} \quad \text{(ELBO of the 2nd view)}} \\ &+ \lambda \cdot \mathcal{L}_{InfoNCE} \quad \text{(InfoNCE)}\end{aligned}$$

Proposed Framework



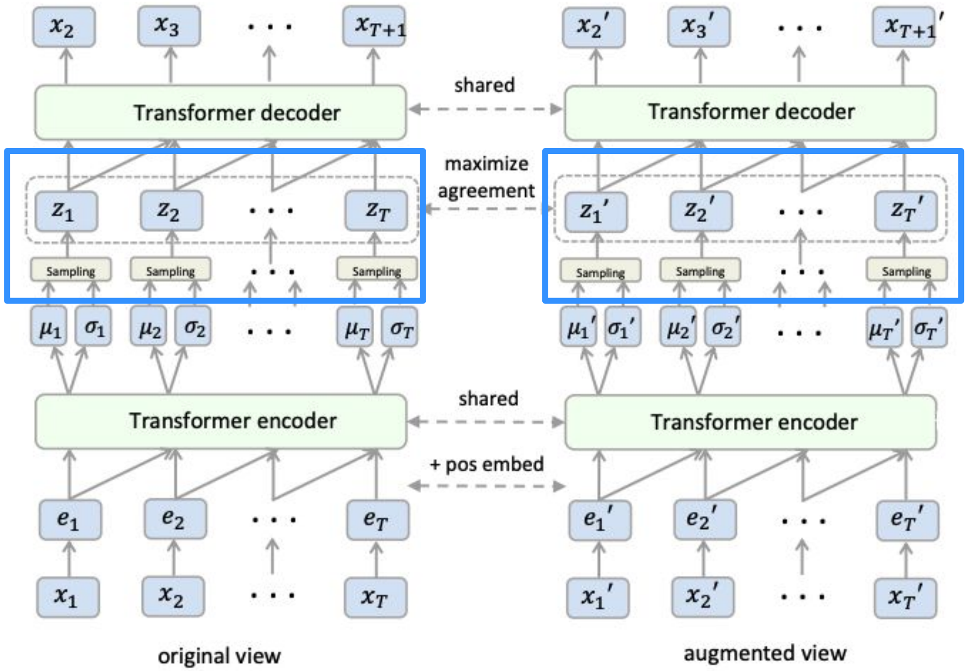
(a) Framework of ContrastVAE

$$\mathcal{L}_{CE} - \mathcal{L}_{KL}$$

$$\mathcal{L}_{CE} = \sum_{t=1}^{T} \left[\log(\sigma(\mathbf{D}_t^T \mathbf{M}_t)) + \log(\sigma(1 - \mathbf{D}_t^\top \mathbf{M}_n))\right]$$

Calculate the **decoder output's** cross entropy between the **expected next item** and a randomly sample negative item.

Proposed Framework



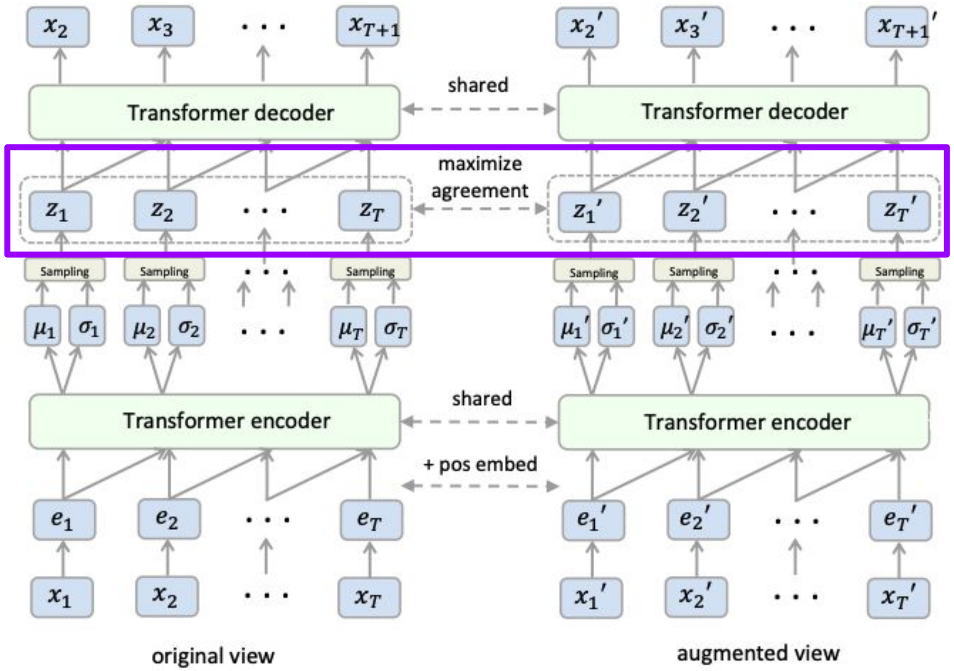
(a) Framework of ContrastVAE

$$\mathcal{L}_{CE} - \mathcal{L}_{KL}$$

$$\mathcal{L}_{KL} = D_{KL}[q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})]$$
$$= \sum_{t=1}^{T} D_{KL}[q(\boldsymbol{z}_t|\boldsymbol{x}_t)||p(\boldsymbol{z}_t)]$$

Calculate the **KLD** between encoded distribution **and** prior distribution

Proposed Framework



(a) Framework of ContrastVAE

Contrastive Learning

$$
\begin{aligned}
\mathcal{L} = \; & \mathcal{L}_{CE} - \mathcal{L}_{KL} \quad \text{(ELBO of the 1st view)} \\
+ \; & \mathcal{L}'_{CE} - \mathcal{L}'_{KL} \quad \text{(ELBO of the 2nd view)} \\
+ \; & \lambda \cdot \boxed{\mathcal{L}_{InfoNCE} \quad \text{(InfoNCE)}}
\end{aligned}
$$

Maximize the mutual information

$$I(z, z')$$

The proof of the relationship between **infoNCE** and **mutual information** can be found in **Appendix 2**

InfoNCE

The InfoNCE loss is proposed from Contrastive Predictive Coding (CPC), which uses `categorical cross-entropy loss` to identify the positive sample amongst a set of `unrelated noise samples`.
The formula mentioned in Moco; He, Kaiming as below:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(q \cdot k_+/\mathcal{T})}{\sum_{i=0}^{K} \exp(q \cdot k_i/\mathcal{T})}$$

Property	Description
$\mathcal{T}$	A temperature hyper-parameter
q	An encoded query
$k_i = \{k_0, k_1, k_2, ...\}$	A set of encoded sample
k_+	Positive sample
$\sum_{i=0}^{K} \exp(q \cdot k_i/\mathcal{T})$	The sum is over 1 postive and K negative samples

$$I(k_+, q)$$

InfoNCE is used to **maximize** the mutual information of $I(k_+, q)$

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748 (2018).

InfoNCE

The InfoNCE loss is proposed from Contrastive Predictive Coding (CPC), which uses `categorical cross-entropy loss` to identify the positive sample amongst a set of `unrelated noise samples`.
The formula mentioned in Moco; He, Kaiming as below:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(q \cdot k_+/\mathcal{T})}{\sum_{i=0}^{K} \exp(q \cdot k_i/\mathcal{T})}$$

Property	Description
$\mathcal{T}$	A temperature hyper-parameter
q	An encoded query
$k_i = \{k_0, k_1, k_2, ...\}$	A set of encoded sample
k_+	Positive sample
$\sum_{i=0}^{K} \exp(q \cdot k_i/\mathcal{T})$	The sum is over 1 postive and K negative samples

Input

Target

$$I(\boxed{k_+}, \boxed{q})$$

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748 (2018).

InfoNCE

The InfoNCE loss is proposed from Contrastive Predictive Coding (CPC), which uses `categorical cross-entropy loss` to identify the positive sample amongst a set of `unrelated noise samples`.
The formula mentioned in Moco; He, Kaiming as below:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(q \cdot k_+ / \mathcal{T})}{\sum_{i=0}^{K} \exp(q \cdot k_i / \mathcal{T})}$$

Property	Description
$\mathcal{T}$	A temperature hyper-parameter
q	An encoded query
$k_i = \{k_0, k_1, k_2, ...\}$	A set of encoded sample
k_+	Positive sample
$\sum_{i=0}^{K} \exp(q \cdot k_i / \mathcal{T})$	The sum is over 1 postive and K negative samples

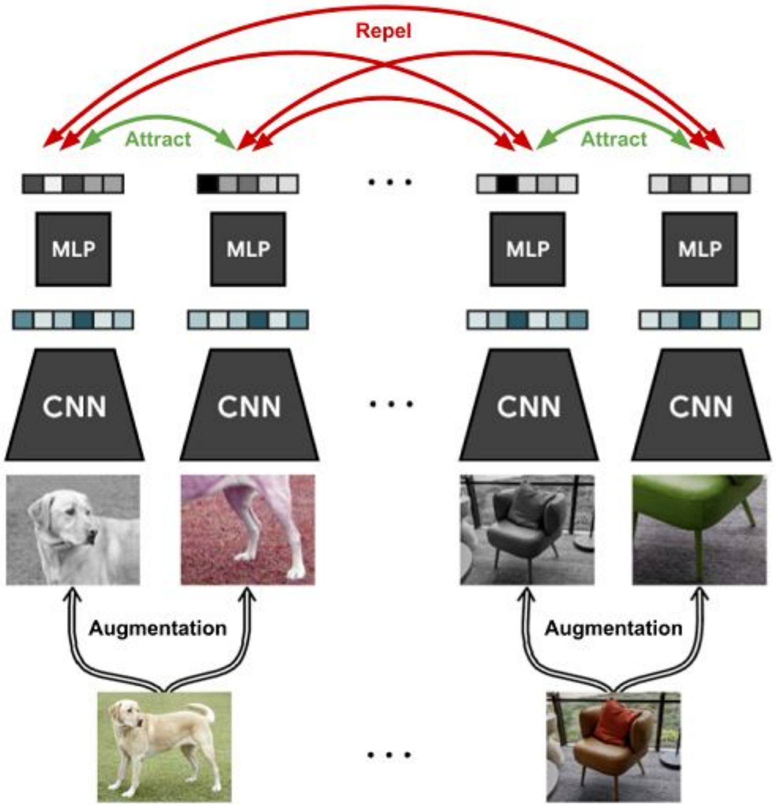
Input

Target

$$I(\boxed{z}, \boxed{z'})$$

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748 (2018).

InfoNCE



InfoNCE is the contrastive loss used in **SimCLR**

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.

Question 1 By Ying-Jia

Why can contrastive learning be used to alleviate posterior collaspe?

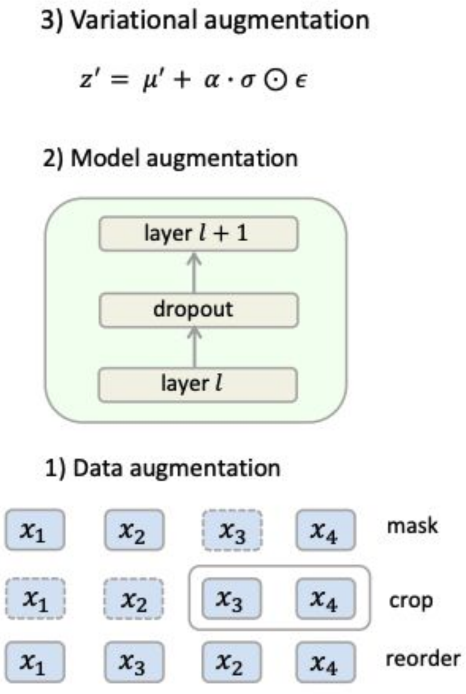
Terms of ELBO

$$
\begin{aligned}
\mathcal{L} = \; & \boxed{\begin{array}{ll} \mathcal{L}_{CE} - \mathcal{L}_{KL} & \text{(ELBO of the 1st view)} \\ + \; \mathcal{L}'_{CE} - \mathcal{L}'_{KL} & \text{(ELBO of the 2nd view)} \end{array}} \\
& + \; \lambda \cdot \mathcal{L}_{InfoNCE} \quad \text{(InfoNCE)}
\end{aligned}
$$

$$I(z, z')$$

Method
Reducing the impact of the KL-divergence term by Reducing its weight
Introducing an Additional Regularization Term that explicitly maximizes the mutual information between the input and latent
Using Empirical Bayes that observed data to estimate the parameters of a prior distribution

Strategies of Augmentations

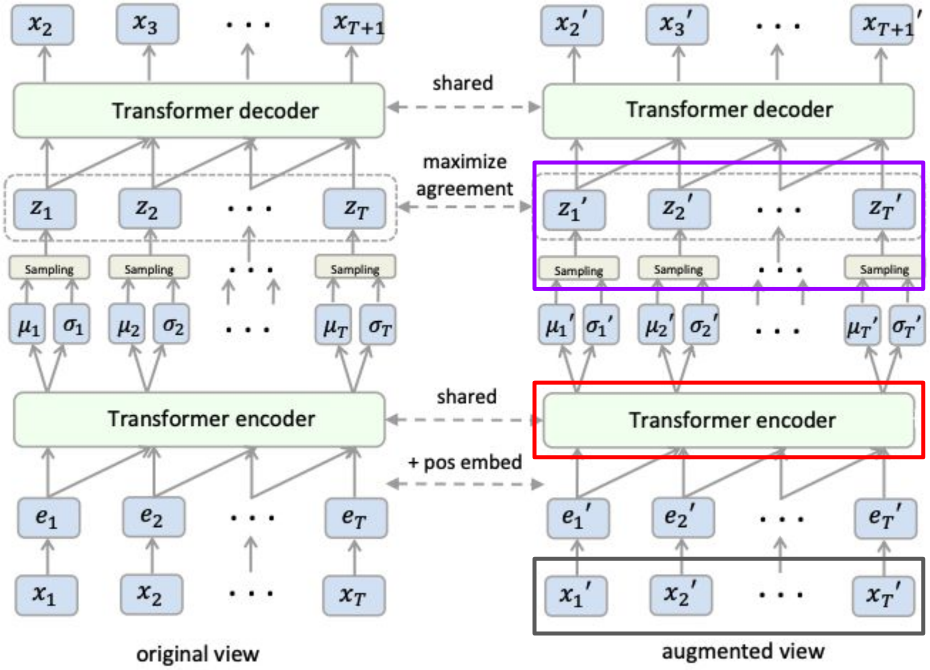


(b) Strategies of augmentations

Optional Strategy

Strategy	Description
`DA`, Data Augmentation in input space	Such as random masking, cropping, or reordering of a sequence. [Qiu et al.] argues that this may lead to inconsistency problems between 2 augmented views, especially when the sequences are very short
`MA`, Model Augmentation	Adding a `random dropout` in each intermediate layer of the Transformer encoder. Recent studies show that the simple dropout operation is powerful enough to generate informative views for CL.
`VA`, Variational Augmentation	Adding a `learnable Gaussian dropout` rate at sampling step

Strategies of Augmentations



(a) Framework of ContrastVAE

Variational Argumentation: Dropout in reparameterization

Model Argumentation: Dropout in Transformer

Data Argumentation

Experiment & Result

Research Questions

RQ1. How does **ContrastVAE** perform compared with **state-of-the-art** SR models?

RQ2. Are the key components in **ContrastVAE**, such as augmentations and contrastive learning, necessary and beneficial for satisfactory improvement?

RQ3. How is the performance of **ContrastVAE** on items with different frequencies and sequences with different lengths? Does **ContrastVAE** improve the performance on long-tail items, and what are the reasons?

RQ4. How is the robustness of **ContrastVAE** w.r.t. noisy input sequences, and is **ContrastVAE** sensitive to some key model hyperparameters?

Datasets

Table 1: Statistics of datasets, we report the number of users, number of items, number of interactions, number of interactions per item, and the averaged sequence length.

Dataset	#Users	#Items	#Interactions	#Ints / item	Avg. seq. len.
Beauty	22,363	12,101	198,502	16.40	8.3
Toys	19,412	11,924	167,597	14.06	8.6
Tools	16,638	10,217	134,476	13.16	8.1
Office	4,905	2,420	53,258	22.00	10.8

Table 2: Number of sequences end at items of different frequency groups.

Dataset	[≤10]	[10, 20]	[20, 30]	[30, 40]	[≥40]
Beauty	17,353	3,152	1,065	367	426
Toys	16,345	2,320	476	130	141
Tools	13,929	1,769	400	230	310
Office	3,150	1,028	547	97	83

Overall Comparison

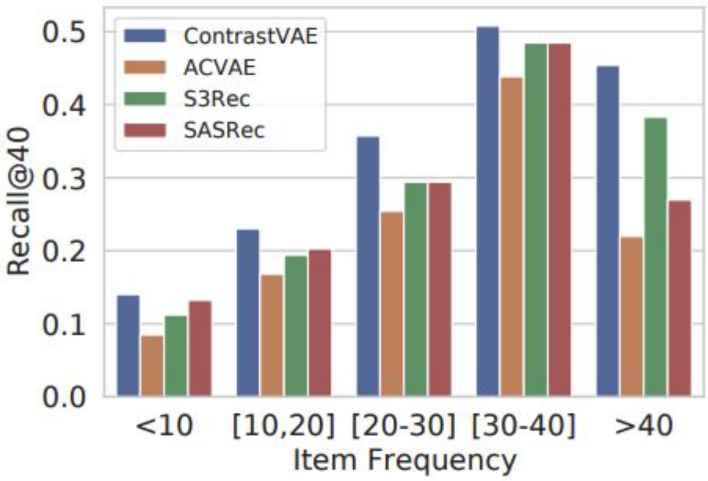
Dataset	Metric	SVAE	ACVAE	S3Rec	CL4Rec	LightGCN	BPRMF	Bert4Rec	SASRec	STOSA	DT4SR	ContrastVAE	Improv.
Beauty	R@20	0.0268	0.0951	0.0946	0.0398	0.0759	0.0739	0.0890	0.0952	0.0975	0.0982	**0.1095**	11.51%
	R@40	0.0417	0.1294	0.1348	0.0554	0.1112	0.1089	0.1285	0.1389	0.1337	0.1404	**0.1541**	9.76%
	N@20	0.0102	0.0467	0.0424	0.0168	0.0306	0.0311	0.0395	0.0420	0.0469	0.0446	**0.0496**	5.76%
	N@40	0.0132	0.0537	0.0505	0.0200	0.0378	0.0383	0.0476	0.0509	0.0542	0.0533	**0.0587**	8.30%
Office	R@20	0.0988	0.1327	0.1335	0.0646	0.0532	0.0483	0.1350	0.1478	0.1578	0.1429	**0.1708**	8.24%
	R@40	0.1647	0.2075	0.2112	0.1025	0.0797	0.0718	0.2230	0.2251	0.2391	0.2186	**0.2617**	9.45%
	N@20	0.0389	0.0560	0.0571	0.0291	0.0243	0.0218	0.0551	0.0657	0.0694	0.0643	**0.0741**	6.77%
	N@40	0.0523	0.0713	0.0729	0.0368	0.0297	0.0266	0.0729	0.0815	0.0859	0.0797	**0.0925**	7.68%
Toy	R@20	0.0178	0.0722	0.0973	0.0392	0.0671	0.0692	0.0699	0.1112	0.1008	0.1130	**0.1164**	3.01%
	R@40	0.0260	0.1030	0.1307	0.0596	0.0977	0.1007	0.0982	0.1479	0.1357	0.1478	**0.1610**	8.86%
	N@20	0.0069	0.0359	0.0467	0.0182	0.0287	0.0304	0.0318	0.0539	0.0496	0.0515	**0.0547**	1.48%
	N@40	0.0086	0.0421	0.0536	0.0224	0.0349	0.0369	0.0376	0.0614	0.0567	0.0560	**0.0638**	4.42%
Tool	R@20	0.0340	0.0537	0.0632	0.0443	0.0537	0.0505	0.0508	0.0640	0.0615	0.0601	**0.0731**	14.21%
	R@40	0.0521	0.0759	0.0849	0.0634	0.0751	0.0715	0.0777	0.0879	0.0867	0.0861	**0.1049**	19.34%
	N@20	0.0149	0.0249	0.0286	0.0194	0.0238	0.0219	0.0213	0.0294	0.0295	0.0289	**0.0326**	10.51%
	N@40	0.0186	0.0294	0.0330	0.0233	0.0282	0.0262	0.0268	0.0345	0.0346	0.0342	**0.0381**	10.12%

Comparison of the performance of different augmentation strategies

Dataset	Metric	AVAE	DA	MA	VA
Beauty	R@20	0.0448	0.1059	**0.1095**	0.1066
	R@40	0.0709	0.1561	0.1541	**0.1578**
	N@20	0.0180	0.0459	**0.0496**	0.0464
	N@40	0.0233	0.0562	**0.0587**	0.0568
Office	R@20	0.1093	**0.1745**	0.1708	0.1672
	R@40	0.1918	**0.2658**	0.2617	0.2599
	N@20	0.0419	0.0739	**0.0741**	0.0722
	N@40	0.0586	0.0924	**0.0925**	0.0911
Toy	R@20	0.0423	0.1112	0.1130	**0.1164**
	R@40	0.0700	0.1554	0.1548	**0.1610**
	N@20	0.0171	0.0503	0.0566	**0.0547**
	N@40	0.0227	0.0593	**0.0652**	0.0638
Tool	R@20	0.0380	0.0671	0.0691	**0.0731**
	R@40	0.0603	0.1004	0.0986	**0.1049**
	N@20	0.0164	0.0295	0.0310	**0.0326**
	N@40	0.0209	0.0364	0.0370	**0.0381**

Strategy	WinRate	Description
`DA`	12.5% (2/16)	DA performs best on the Office dataset. This might be because the Office dataset has the largest `average sequence length` (see Table 1) and thus is less sensitive to the perturbation of data augmentations.
`MA`	37.5% (6/16)	`MA` method performs competitively compared with DA but is much simpler.
`VA`	50.0% (8/16)	`VA` achieves comparable or even better results, especially on Toy and Tool datasets with smaller average sequence length and testing item frequency. This shows that the proposed VA can effectively benefit the prediction of short sequences and long-tail items.

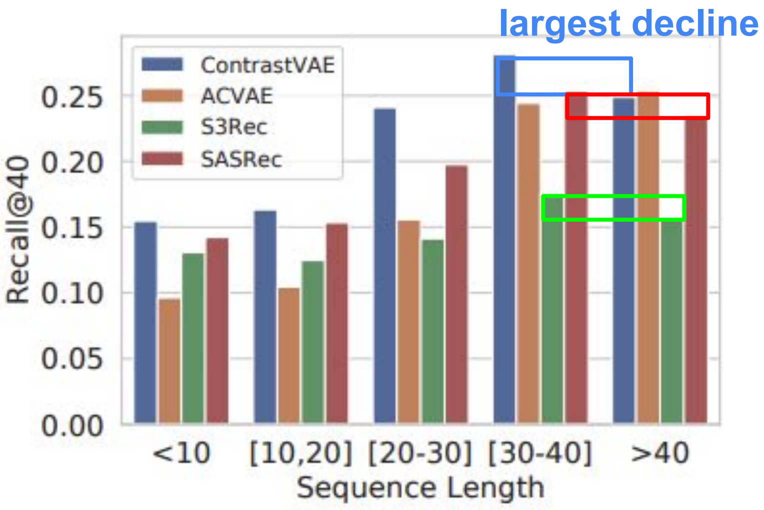
Comparison of Item frequencies



(a) Item frequencies

- Categorizing the user sequences **into 5 groups** according to **the frequencies of their last clicked items**
- Reporting **Recall@40** of ContrastVAE and representative baseline methods on the Toy dataset
- **ContrastVAE** achieves the highest Recall@40 sores on all groups of sequences.
  - Specifically, on long-tail items (i.e., [≤ 10] and [10, 20]), **ContrastVAE** outperforms other baseline models by a large margin.

Comparison of Sequence lengths



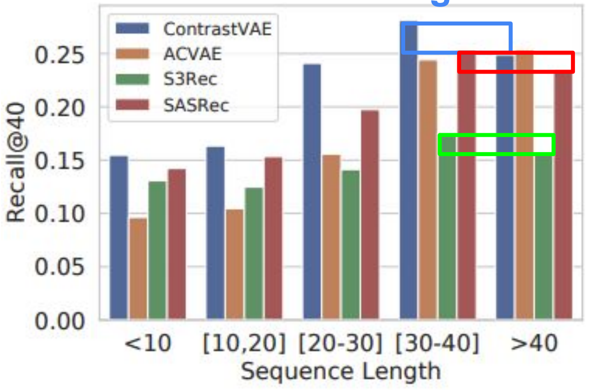
(b) Sequence lengths

- Studying how the **sequence length** affects the model's performance
- Reporting Recall@40 of **ContrastVAE** and representative baseline methods on the Toy dataset
- **ContrastVAE** consistently exhibits good performance on sequences with various lengths
- **ContrastVAE** greatly improves the performance of short sequences (i.e., less than 20 interactions).
- On **longer sequences** (e.g., [≥ 40]), **ContrastVAE** doesn't show superior performance compared with other models, this may because **for long sequences, the users' preferences tend to become certain and easy to predict, in which case the uncertainty and randomness** introduced by **ContrastVAE** would not help the prediction results.

Question 2 By Yi-Ting

(1) If the long sequences is certain and easily to predict, why the performance of other attention-based model decline together?

largest decline



(b) Sequence lengths

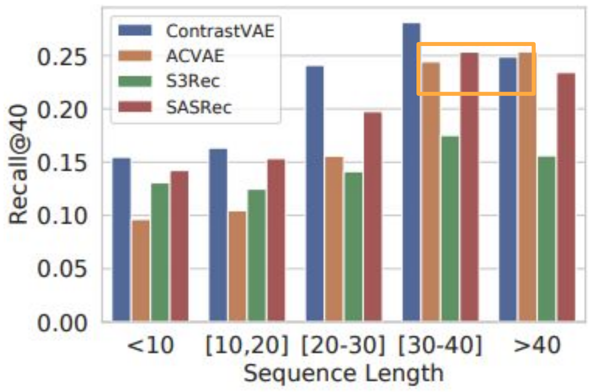
Table 2: Number of sequences end at items of different frequency groups.

Dataset	[≤10]	[10, 20]	[20, 30]	[30, 40]	[≥40]
Beauty	17,353	3,152	1,065	367	426
Toys	16,345	2,320	476	130	141
Tools	13,929	1,769	400	230	310
Office	3,150	1,028	547	97	83

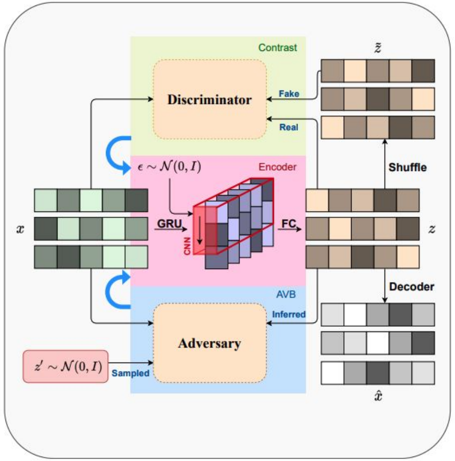
The performance may be a trade-off between the benefit from the long sequences and the harm from the lack of data.

Question 2 By Yi-Ting

(2) Why has ACVAE's performance risen instead of fallen?



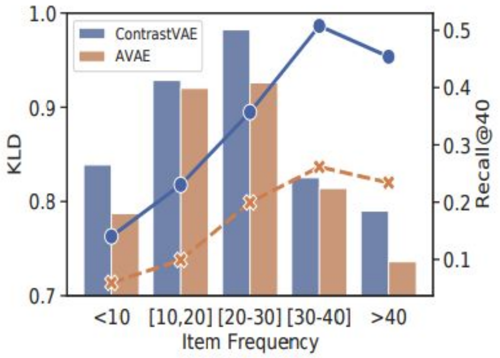
(b) Sequence lengths



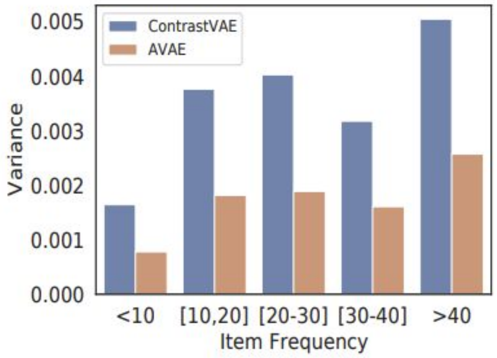
ACVAE use GRU as basic model, which is typically less sensitive to the size of the dataset than attention models, as they do not rely on explicit attention mechanisms to weight the input features.

ContrastVAE alleviates posterior collapse and point-estimation in latent space

As we know, the **posterior collapse** is caused by the **estimated posterior distribution** becoming too similar to the prior **Standard Normal Distribution**, which limited the decoder's capacity to generate diverse outputs



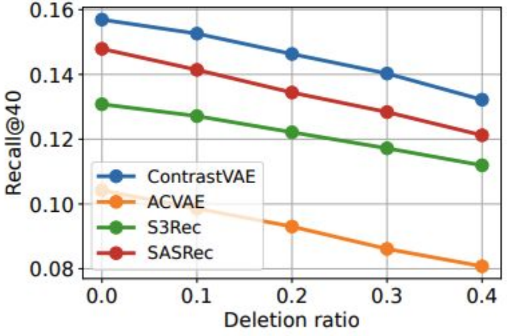
(a) Recall@40 (line graph) and KL-divergence (bar graph)



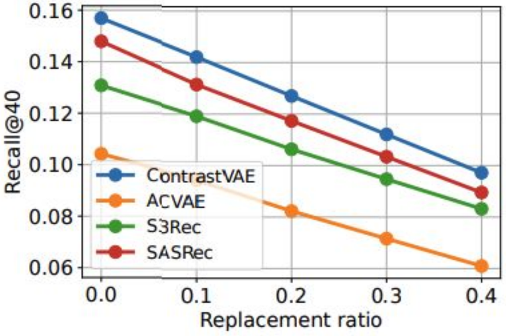
(b) Variance of latent variable estimation

- **Average KLD** between the posterior distribution p(z|x) of sequences and the standard Gaussian distribution N(0,I) , which reflects the extent of posterior collapse problem (posterior collapse induces small KLD)
- **Average variance** of latent variables, which reflect the extent of variance vanishing.

Robustness analysis



(a) Random deletion



(b) Random replacement

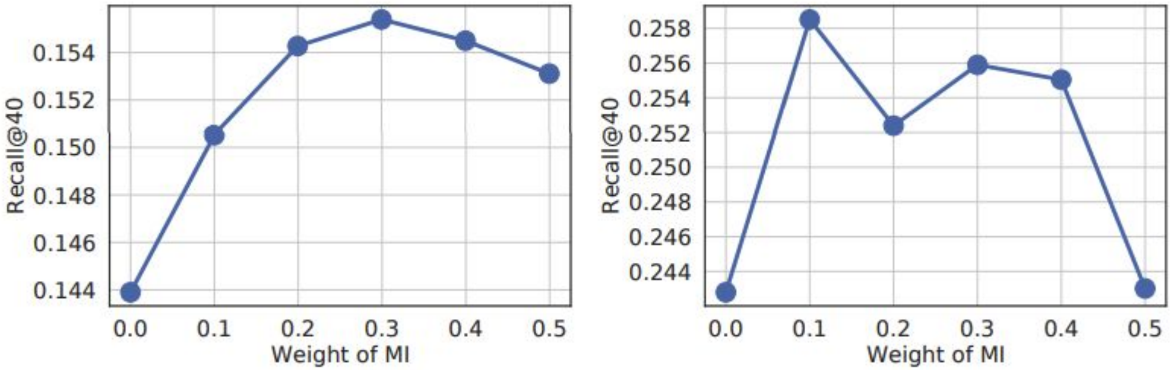
Measure robustness by two corrupting strategies:

- **(a) randomly deleting a proportion of items** in each sequence (random deletion)
- **(b) randomly replacing proportion items** with other items in each sequence (random replacement).

The performance of all models exhibits a **drop** as we increase the corruption ratio.

However, **ContrastVAE** always outperforms other baseline models by a large margin whatever the corruption method and the corruption ratio, which indicates that ContrastVAE **can still exhibit good performance for noisy input data.**

Hyper-parameter sensitivity analysis



(a) Beauty **(b) Office**

$$
\begin{aligned}
\mathcal{L} = \quad & \mathcal{L}_{CE} - \mathcal{L}_{KL} && \text{(ELBO of the 1st view)} \\
+ \quad & \mathcal{L}'_{CE} - \mathcal{L}'_{KL} && \text{(ELBO of the 2nd view)} \\
+ \quad & \boxed{\lambda} \cdot \mathcal{L}_{InfoNCE} && \text{(InfoNCE)}
\end{aligned}
$$

Conclusion

The contributions of this paper are summarized as follows:

- Deriving **ContrastELBO**, which is an extension of conventional single-view ELBO to two-view case and naturally incorporates contrastive learning into the framework of VAE.
- Proposing **ContrastVAE**, a two-branched VAE framework guided by **ContrastELBO** for sequential recommendation.
- Introducing **model augmentation** and **variational augmentation** to avoid **the semantic inconsistency problem** led by conventional data augmentation.
- Conducting comprehensive experiments to evaluate our method.
- Extensive ablation studies and empirical analysis verify the effectiveness of the proposed components.

SWOT

Strengths

- The idea that use contrastive learning on VAE can be used to solve the posterior collapse problem

Weaknesses

- Doesn’t perform well with longer sequence lengths, which means it may not perform well on dataset with high density, i.e. Movielen.

Opportunities

- The experiment metrics can well define the posterior collapse problem, which may help us to complete our study
- Adding contrastive learning to our methods

Threats

- Currently this paper is the one of the SOTA of sequential recommender system which means it is the potential opponent of our research.

Thank you for listening

Appendix

Appendix 1 Proof of ContrastELBO

- First, we use a variational distribution $q(z, z'|x, x')$ to approximate the posterior distribution $p(x, x'|z, z')$, which could be factorized as:

$$q(z, z'|x, x') = q(z|x)q(z'|x')$$

- Then:

$$\begin{aligned}
\log p(x, x') &= \log p(x)p(x') \\
&= \log \int p(x, x', z, z') \cdot dz \cdot dz' \\
&= \log \mathbb{E}_{q(z,z'|x,x')} \left[\frac{p(x, x', z, z')}{q(z, z'|x, x')} \right] \\
&\geq \mathbb{E}_{q(z,z'|x,x')} \log \left[\frac{p(x, x', z, z')}{q(z, z'|x, x')} \right]
\end{aligned}$$

Appendix 1 Proof of ContrastELBO

- And the probability distribution of x and x' is only related to z and z', x and x' are independent of each other, so we can derive that:

$$
\begin{aligned}
&\geq \mathbb{E}_{q(z,z'|x,x')} \log \left[\frac{p(x,x',z,z')}{q(z,z'|x,x')} \right] \\
&= \mathbb{E}_{q(z,z'|x,x')} \log \left[\frac{p(x|z)p(x'|z')p(z,z')}{q(z|x)q(z'|x')} \right] \\
&= \mathbb{E}_{q(z|x)} \log[p(x|z)] + \mathbb{E}_{q(z'|x')} \log[p(x'|z')] + \mathbb{E}_{q(z,z'|x,x')} \log \left[\frac{p(z,z')}{q(z|x)q(z'|x')} \right]
\end{aligned}
$$

- The red term can be expand as:

$$
\begin{aligned}
&\mathbb{E}_{q(z,z'|x,x')} \log \left[\frac{p(z,z')}{q(z|x)q(z'|x')} \right] \\
&= \mathbb{E}_{q(z,z'|x,x')} \log \left[\frac{p(z,z')p(z)p(z')}{p(z)p(z')q(z|x)q(z'|x')} \right] \\
&= \mathbb{E}_{q(z,z'|x,x')} \log \left[\frac{p(z,z')}{p(z)p(z')} \right] + \mathbb{E}_{q(z,z'|x,x')} \log \left[\frac{p(z)p(z')}{q(z|x)q(z'|x')} \right] \\
&= \mathbb{E}_{q(z,z'|x,x')} \log \left[\frac{p(z,z')}{p(z)p(z')} \right] - \left(D_{KL}[q(z|x)||p(z)] + D_{KL}[q(z'|x')||p(z'))] \right)
\end{aligned}
$$

Appendix 1 Proof of ContrastELBO

- The green term can be derived by the definition of KL divergence
- Finally, expand the red term, we get:

$$
\begin{aligned}
&= \mathbb{E}_{q(z|x)} \log[p(x|z)] + \mathbb{E}_{q(z'|x')} \log[p(x'|z')] + \mathbb{E}_{q(z,z'|x,x')} \log\left[\frac{p(z,z')}{q(z|x)q(z'|x')}\right] \\
&= \mathbb{E}_{q(z|x)} \log[p(x|z)] + \mathbb{E}_{q(z'|x')} \log[p(x'|z')] + \mathbb{E}_{q(z,z'|x,x')} \log\left[\frac{p(z,z')}{p(z)p(z')}\right] - \Big(D_{KL}[q(z|x)||p(z)] + D_{KL}[q(z'|x')||p(z'))]\Big) \\
&= \mathbb{E}_{q(z|x)} \log[p(x|z)] - D_{KL}[q(z|x)||p(z)] \\
&+ \mathbb{E}_{q(z'|x')} \log[p(x'|z')] - D_{KL}[q(z'|x')||p(z'))] \\
&+ \mathbb{E}_{q(z,z'|x,x')} \log\left[\frac{p(z,z')}{p(z)p(z')}\right]
\end{aligned}
$$

Appendix 2 The relationship between InfoNCE and Mutual information

First we need to derive the probability of positive sample when given a context vector c:

$$p(x_+|X,c) = \frac{p(x_+)\prod_{i=1,\dots,N;i\neq +} p(x_i)}{\sum_{j=1}^{N}\left[p(x_j|c)\prod_{i=1,\dots,N;i\neq j} p(x_j)\right]} = \frac{\frac{p(x_+|c)}{p(x_+)}}{\sum_{j=1}^{N}\frac{p(x_j|c)}{p(x_j)}} = \frac{f(x_+,c)}{\sum_{j=1}^{N} f(x_j,c)}$$

- where the scoring function $f(x,c) \propto \frac{p(x|c)}{p(x)}$

$$\boxed{\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(q\cdot k_+/\mathcal{T})}{\sum_{i=0}^{K}\exp(q\cdot k_i/\mathcal{T})}}$$

For brevity, let us write the loss of InfoNCE as:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{f(x,c)}{\sum_{x'\in X} f(x',c)}\right]$$

And then we dervie the mutual information with In terms of PMFs for discrete distributions:

$$I(x;c) = \sum_{x,c} p(x,c)\log\frac{p(x,c)}{p(x)p(c)} = \sum_{x,c} p(x,c)\log\frac{p(x|c)}{p(x)}$$

Where the term in blue is estimated by f, which means when we maximize the f between input x_+ and context vector c, we maximize the `mutual information` between input x_+ and context vector c

Appendix 3 The relationship between Cross entropy and InfoNCE

The cross entropy formular is:

$$\mathcal{L}_{CE}(p(y_i)) = -\sum_{i\in K} y_i \log(p(y_i))$$

The softmax formula is:

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=0}^{K}\exp(x_j)}$$

To calculate the probability of y_i, we have to use softmax on the model ouput logits x_i, so:

$$\boxed{\begin{aligned} p(y_+) &= \text{softmax}(x_+) = \frac{\exp(x_+)}{\sum_{j=0}^{K}\exp(x_j)} \\ \mathcal{L}_{CE}(p(y_+)) &= -\sum_{i\in K} y_i \log(p(y_+)) \\ &= -\sum_{i\in K} y_i \log(\frac{\exp(x_+)}{\sum_{j=0}^{K}\exp(x_j)}) \\ &= -\log(\frac{\exp(x_+)}{\sum_{j=0}^{K}\exp(x_j)}) \end{aligned}}$$

In above formula, K is the total classes in the dataset.

- Let's take the `ImageNet dataset` in the CV field as an example. There are **1.28 million* images in the dataset. We use data augmentation techniques (such as `random cropping`, `random color distortion`, and `random Gaussian blur`) to generate `positive sample` pairs for contrastive learning. Each image is a separate category, so K is **1.28 million categories**. The more images there are, the more categories there are.
  But calculating with softmax on such a large number of categories is very time-consuming, especially with `exponential operations`. When the dimension of the vector is several million, the computation complexity is quite high. So the $\mathcal{L}_{CE}$ is not suitable for use on the Contrastive learning.

Then we look back to $\mathcal{L}_{\text{InfoNCE}}$:

$$\mathcal{L}_{\text{InfoNCE}} = -\log\frac{\exp(q\cdot k_+/\mathcal{T})}{\sum_{i=0}^{K}\exp(q\cdot k_i/\mathcal{T})}$$

If we ignore the temperature hyper-parameter $\mathcal{T}$, the loss function became:

$$\mathcal{L}_{\text{InfoNCE}} = -\log\frac{\exp(q\cdot k_+)}{\sum_{i=0}^{K}\exp(q\cdot k_i)}$$

As we can see, The `InfoNCE` loss is actually a `cross entropy loss`, and it performs a classification task with $k+1$ classes.

Appendix 4 My Experiments

Testing on **Toy Datasets**

Metric	**Score**
Recall@1	0.00283
Recall@20	0.04224
Recall@40	0.06959
NDCG@20	0.01600
NDCG@40	0.02158

It seems not to match the performance in the paper with default parameter.