CLIK | Contrastive Learning of text and Image for ranKing

CLOVA MD, CLOVA ML X

Jee Hyung Ko / 고 지 형

Problem Definition - AS-IS

Men's Clothing

Winter Wide Pants for Men

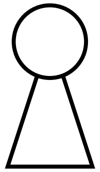
2022. 01. 16 ~ 2022. 01. 22

TBD

#MensWidePants #FleeceLinedPants

< An Example of Online Special Exhibition Service, NAVER Shopping >

Generated Special Exhibition & Included items



Human Service Operator

Select representative image

Drop unrelated images

Topic - 'Winter Wide Pants for Men'

&

Choose the image that looks the most appealing

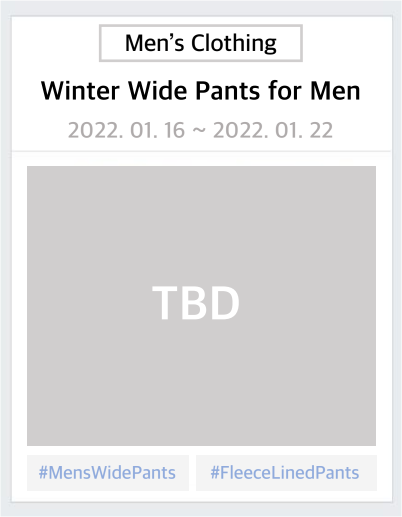
Men's Clothing

Winter Wide Pants for Men

2022. 01. 16 ~ 2022. 01. 22

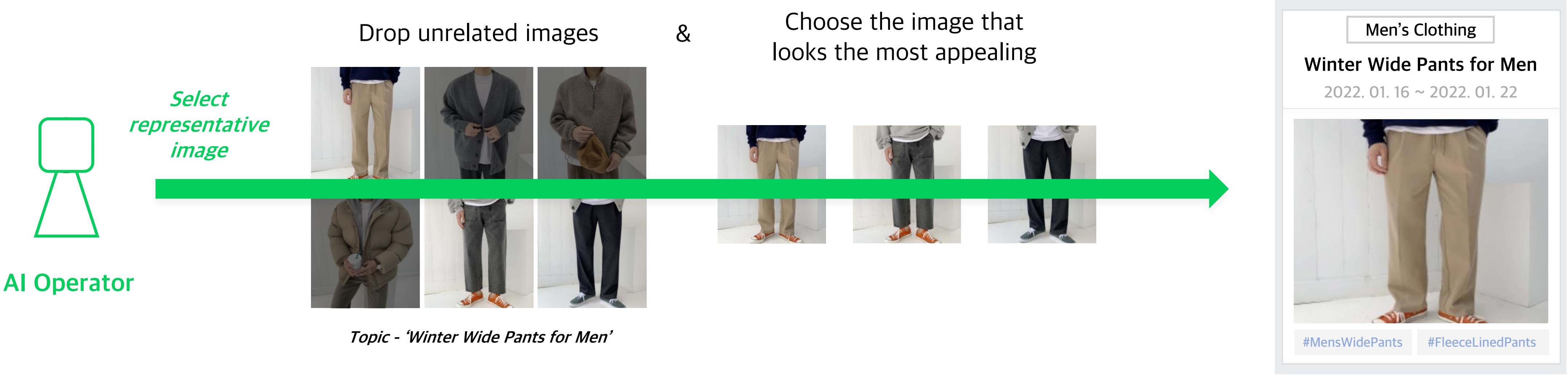
#MensWidePants #FleeceLinedPants

Problem Definition - TO-BE



< An Example of Online Special Exhibition Service, NAVER Shopping >

Generated Special Exhibition & Included items



Problem Definition

What is ‘Representative Image’?

- Contextual: suitable for given context
- Attractive: draw the attention of users

Problem Definition

What is 'Representative Image'?

- Contextual: suitable for given context
- Attractive: draw the attention of users

What is 'Representative Image Selection'?

- To find the most compatible image as a representative among item images included to a special exhibition
- Compatibility - contextual & attractive

Problem Definition

What is 'Representative Image'?

- Contextual: suitable for given context
- Attractive: draw the attention of users

What is 'Representative Image Selection'?

- To find the most compatible image as a representative among item images included to a special exhibition
- Compatibility - contextual & attractive

Define it more generally.

- Our goal is to find a function f as follows:

$$
\begin{aligned}
(x_{i*}, c_{i*}) &= \max_{c}\big(f(s_i, X_i)\big) \\
&= \max_{c}(\{(x_{i1}, c_{i1}), \ldots, (x_{i|X_i|}, c_{i|X_i|})\})
\end{aligned}
$$

$(s_i, X_i) \coloneqq g_i$: ith instance group
(=given special exhibition)

s_i : shared context of g_i
(=theme of given special exhibition)

$X_i = \{x_{ij}\}_{j=1}^{|X_i|}$: set of instances of g_i
(=item images included to given special exhibition)

c_{ij} : (predicted) compatibility of instance x_{ij}

Problem Definition

Representation Learning for ‘Contextual’

- Contrastive approach-based self-supervised learning
- Good to learn ‘contextual’ aspect,
  but hard to consider how attractive each instance is

CTR Prediction/Creative Ranking for ‘Attractive’

- Predict CTR for an item or compare ranking scores among items
- Good to learn ‘attractive’ aspect,
  but hard to consider shared context of instance group directly

Problem Definition

Representation Learning for 'Contextual'

- Contrastive approach-based self-supervised learning
- Good to learn 'contextual' aspect,
  but hard to consider how attractive each instance is

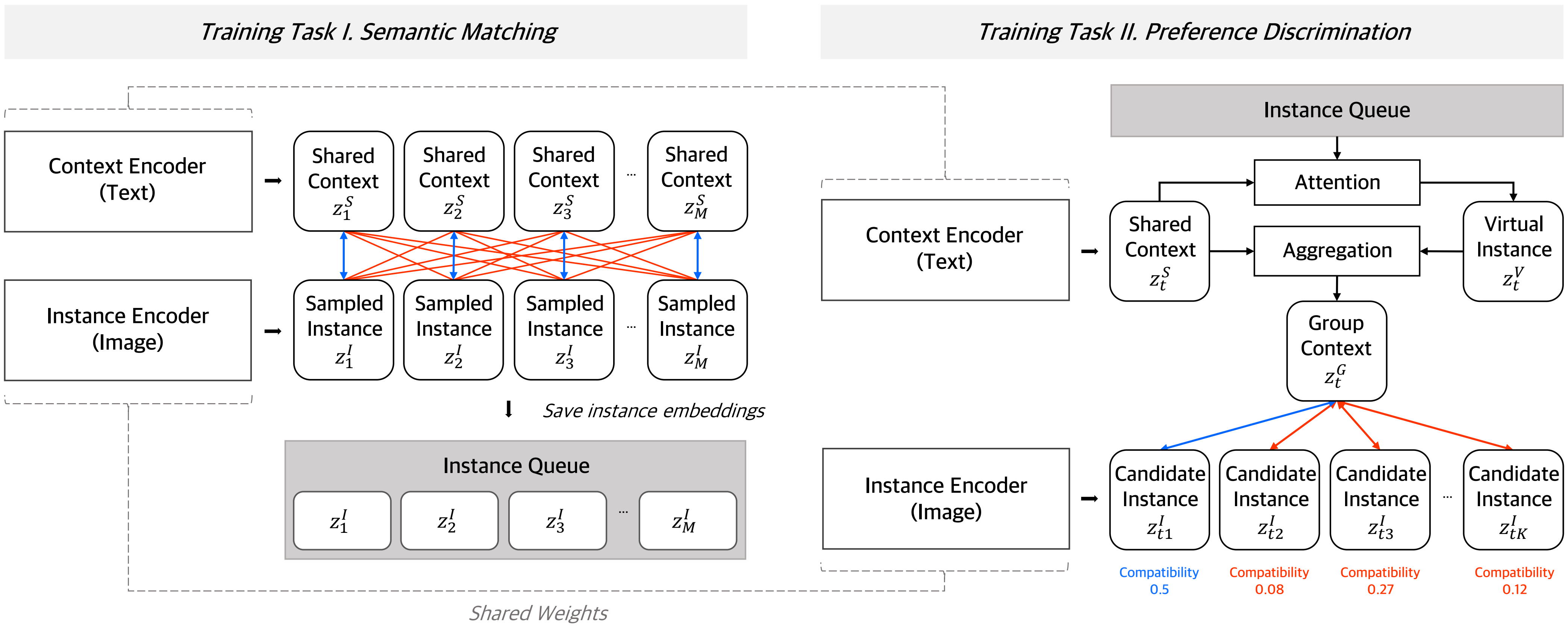
➡ Goal: Representation Learning + Ranking

CTR Prediction/Creative Ranking for 'Attractive'

- Predict CTR for an item or compare ranking scores among items
- Good to learn 'attractive' aspect,
  but hard to consider shared context of instance group directly

Model Description - CLIK: Contrastive Learning of text and Image for ranKing

CLIK: Contrastive Learning of text and Image for ranKing



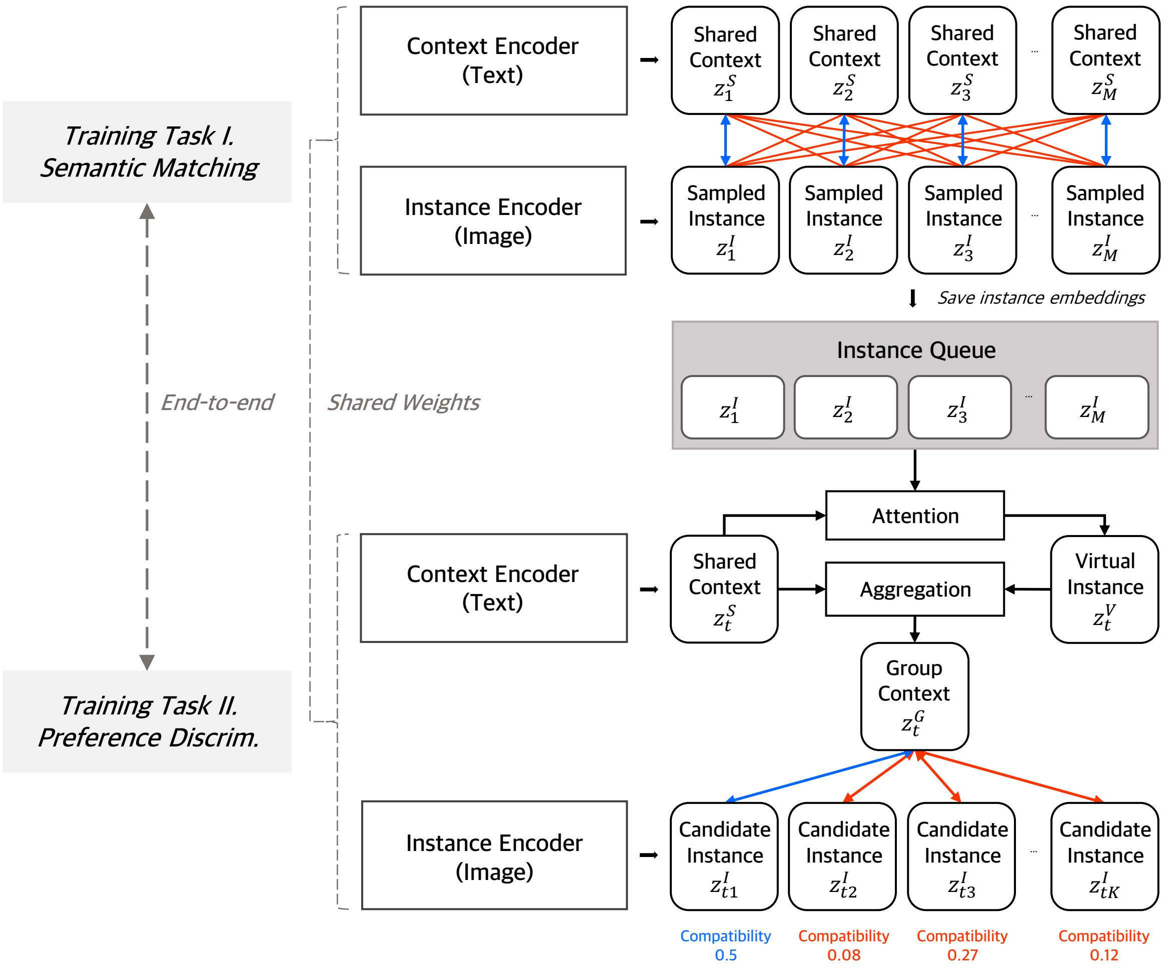
Model Description - CLIK: Contrastive Learning of text and Image for ranKing

Overview

- Perform two training tasks

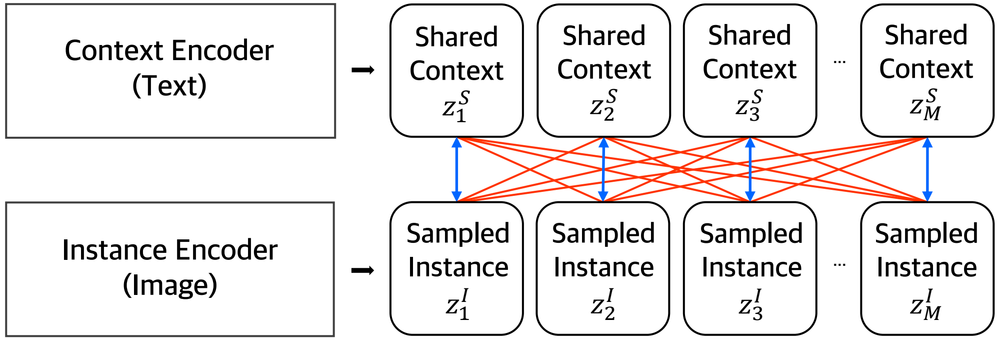
- *Task I. Semantic Matching* (Representation Learning)
  understand 'shared context ↔ instance' relation
- *Task II. Preference Discrimination* (Ranking)
  select the best of instances in given instance group

- Both tasks are performed
  mainly by dual encoder with shared weights,
  in end-to-end manner



Model Description - CLIK: Contrastive Learning of text and Image for ranKing

Task I. Semantic Matching

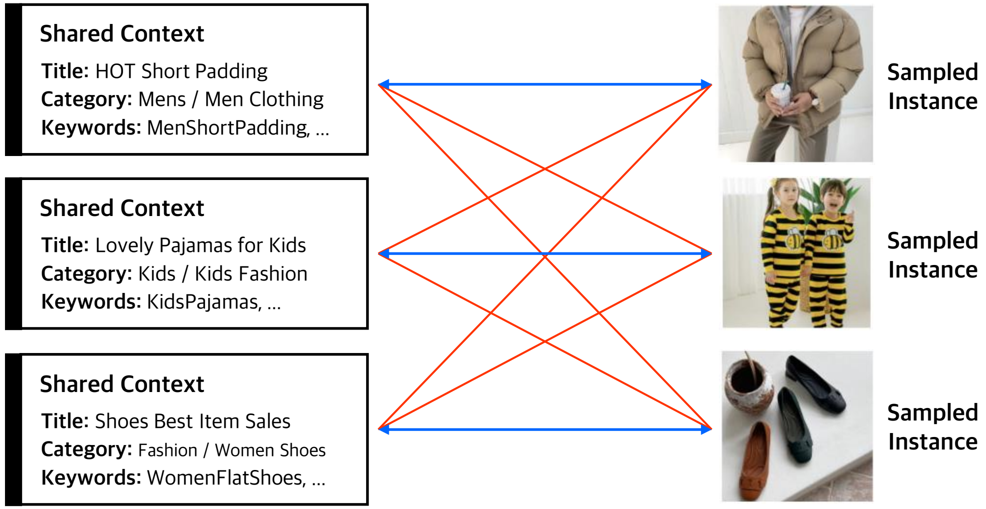


- Understand 'shared context ↔ instance' relation
- Same as the training task of CLIP, SSL with contrastive approach so, it need NO human annotations

- Maximize similarities of 'shared context ↔ instance' pairs from the same instance group,
- Minimize similarities for the other pairs

- Contrast pairs are composed as follows

$$a\ pair\ of\ g_i = (s_i, x_{ij}), \qquad x_{ij} \in X_i\ (randomly\ sampled)$$



Model Description - CLIK: Contrastive Learning of text and Image for ranKing

Task I. Semantic Matching

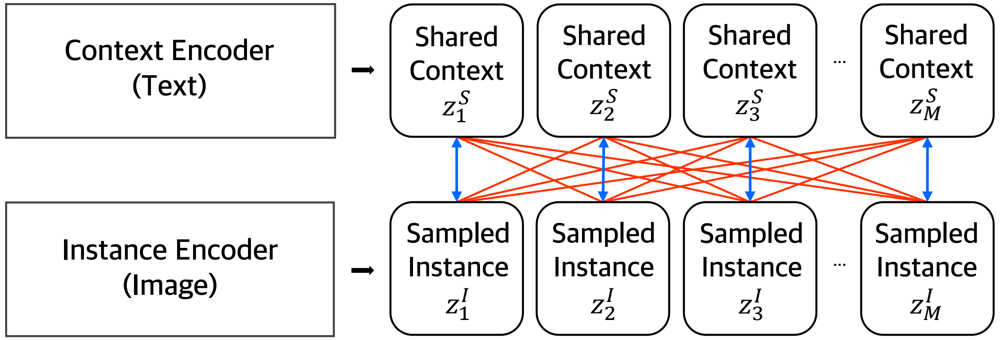
- Loss function - the same as that of CLIP, NT-Xent Loss

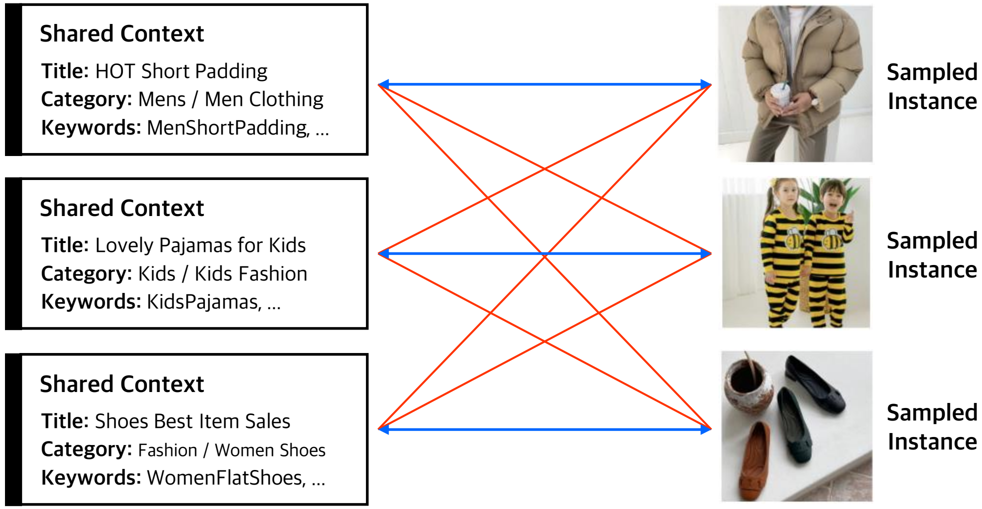
$$L_{matching} = (L_{S2I} + L_{I2S}) \;/\; 2$$

$$L_{S2I} = -\frac{1}{M}\sum_{m=1}^{M} \log \frac{\exp(sim(z_m^S, z_m^I) \;/\; \tau)}{\sum_{i=1}^{M} \exp(sim(z_m^S, z_i^I) \;/\tau)}$$

$$L_{I2C} = -\frac{1}{M}\sum_{m=1}^{M} \log \frac{\exp(sim(z_m^I, z_m^S) \;/\; \tau)}{\sum_{i=1}^{M} \exp(sim(z_m^I, z_i^S) \;/\tau)}$$

M	: batch size for *Semantic Matching*
z_i^I	: instance embedding sampled from g_i
z_i^S	: shared context embedding of g_i
$sim(\cdot)$	: cosine similarity
τ	: temperature parameter

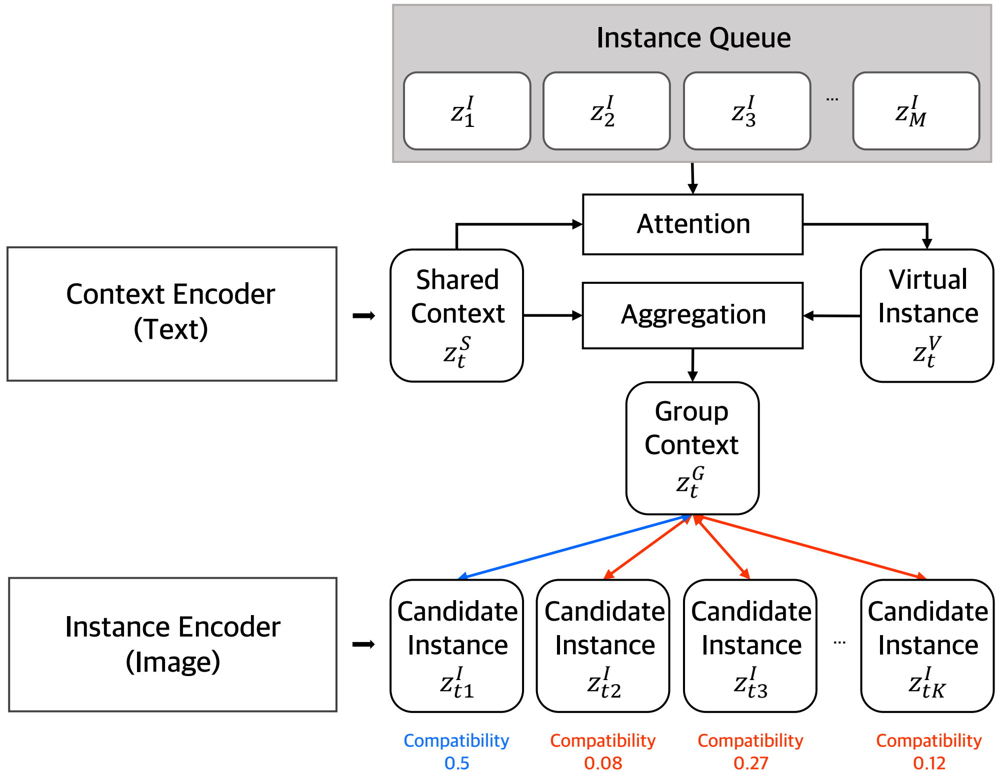




Model Description - CLIK: Contrastive Learning of text and Image for ranKing

Task II. Preference Discrimination

- Select the best of instances in given instance group
- Anchor: a special embedding called 'Group Context'
- Positive/Negative: instances of given group



Model Description - CLIK: Contrastive Learning of text and Image for ranKing

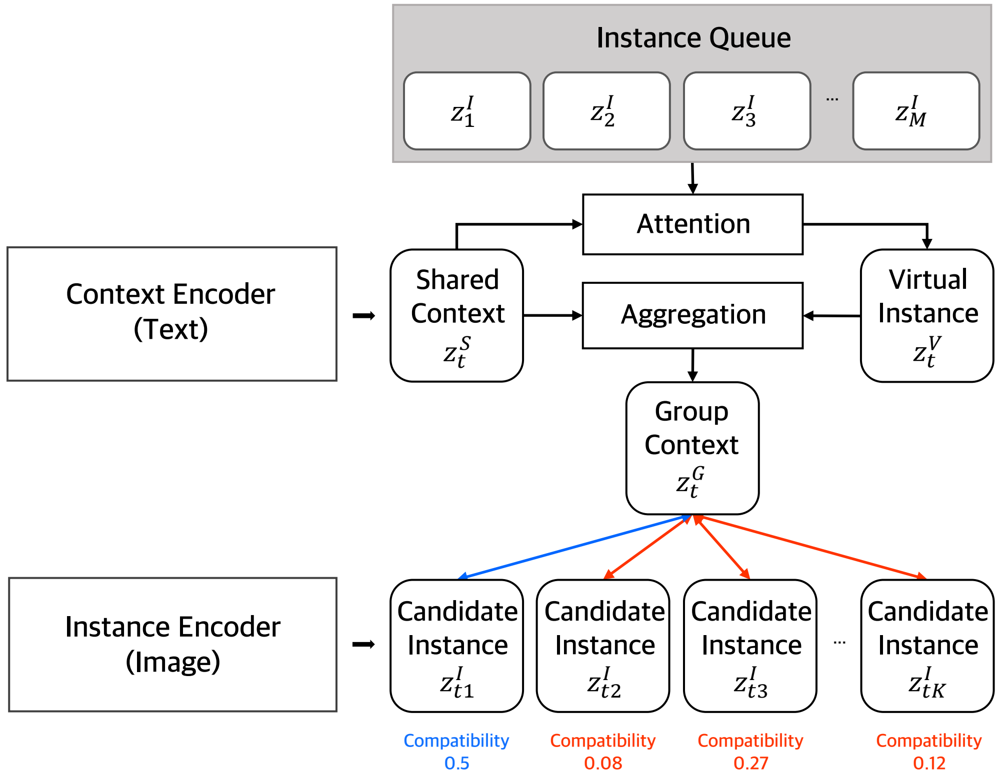
Task II. Preference Discrimination

- Loss function

$$L_{discrim} = -\frac{1}{D}\sum_{i=1}^{D}\log\frac{\exp(sim(z_i^G, z_{i*}^I)\ /\ \tau)}{\sum_{k=1}^{K}\exp(sim(z_i^G, z_{ik}^I)\ /\ \tau)}$$

D	: # discrimination iterations
K	: sampling size of instances from group
z_i^G	: group context embedding of g_i
z_{ik}^I / z_{i*}^I	: kth sampled instance/best instance of g_i
τ	: temperature parameter
$sim(\cdot)$	: cosine similarity

- Maximize similarity of 'group context ↔ instance' pair when compatibility of the instance is greater than the others
- Minimize similarities for the other pairs



Model Description - CLIK: Contrastive Learning of text and Image for ranKing

Task II. Preference Discrimination

- Group Context z_t^G (core component of CLIK)
  anchor embedding representing given g_t

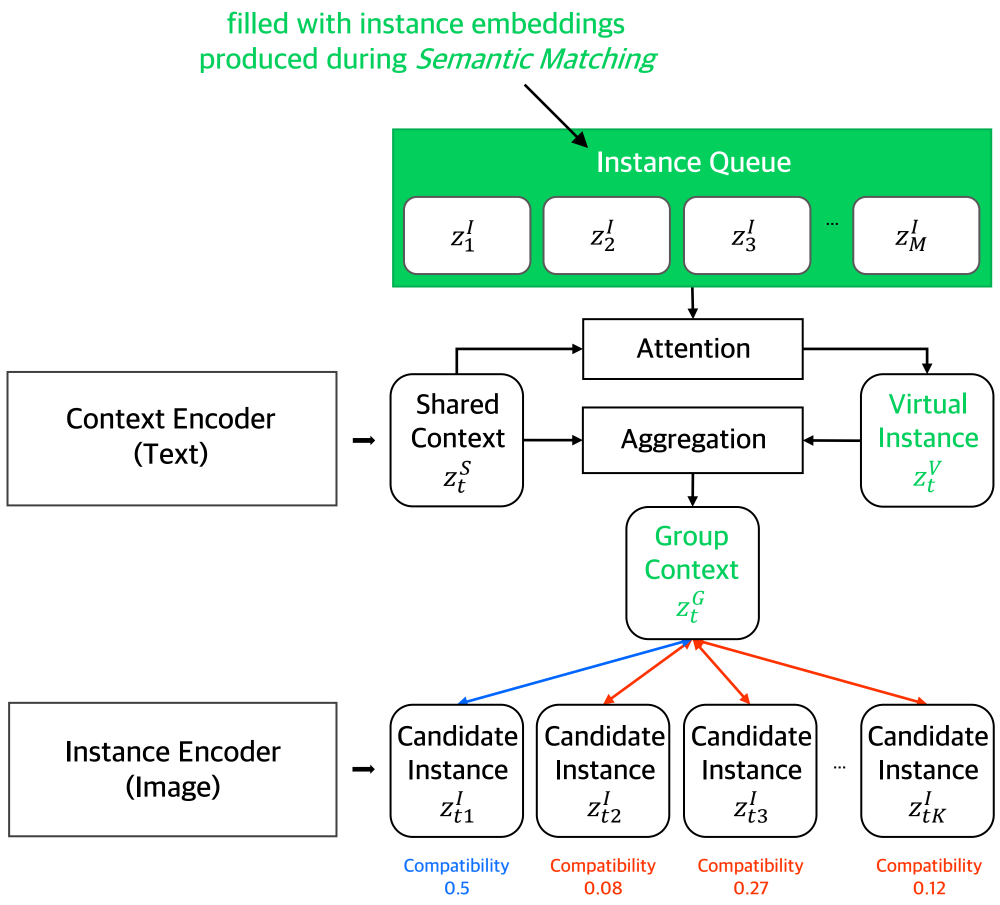
$$z_t^G = Agg(z_t^S, z_t^V)$$

- Virtual Instance z_t^V
  virtual embedding likely to fit with z_t^S semantically
  generated by attention between z_t^S and external instances

$$z_t^V = Attention\left(z_t^S, \left\{ z_{\sim t}^I \mid x_{\sim t} \in \bigcup_{i \neq t} X_i \right\}\right)$$

- External Instances (Instance Queue)
  instance sources used to generate z_t^V
  right after *task I*, store their instance embeddings to the queue

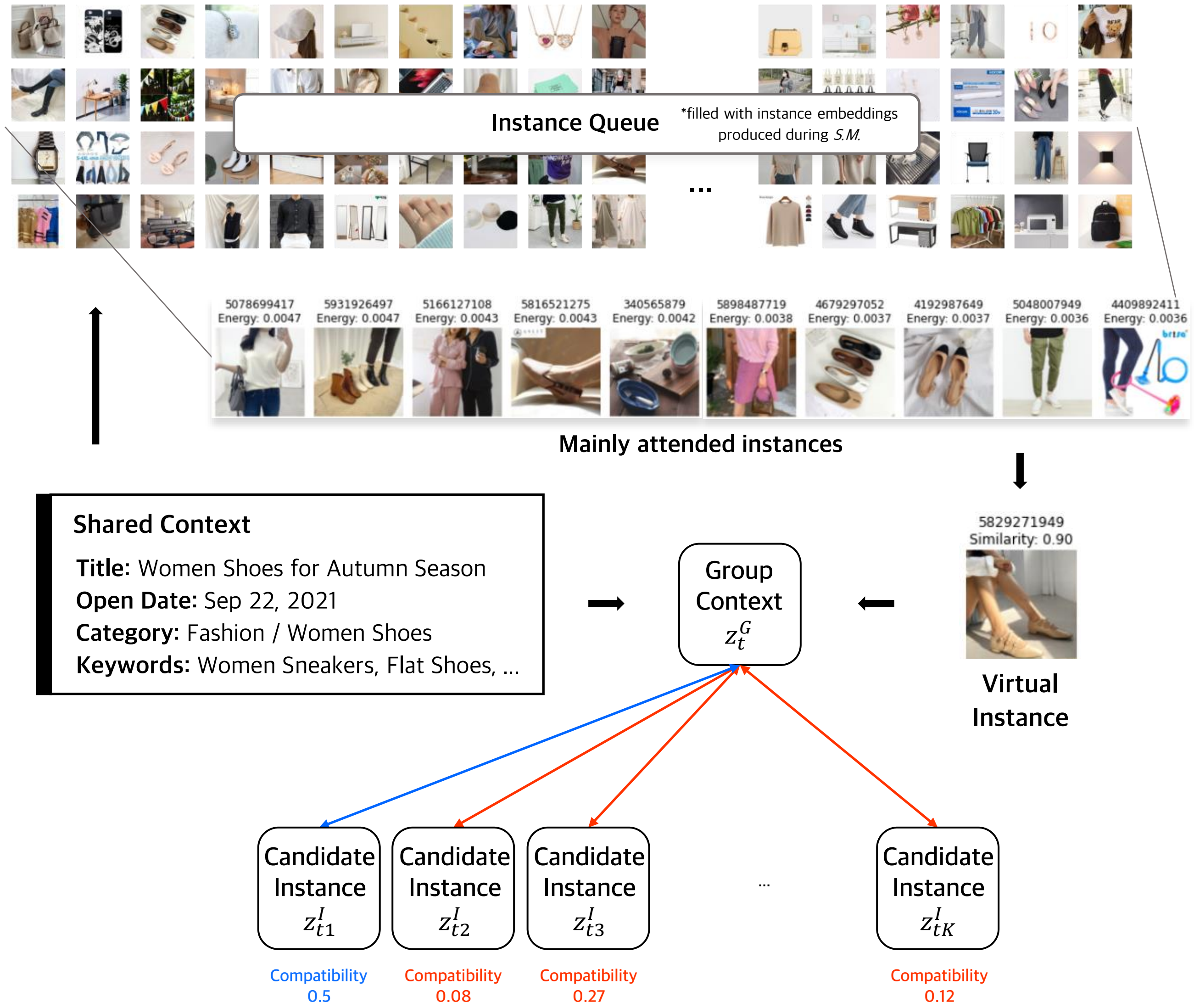
$$Instance\ Queue = \{z_{ij}^I \mid x_{ij} \in X_{i \neq t}\}$$



Model Description - CLIK: Contrastive Learning of text and Image for ranKing

Task II. Preference Discrimination

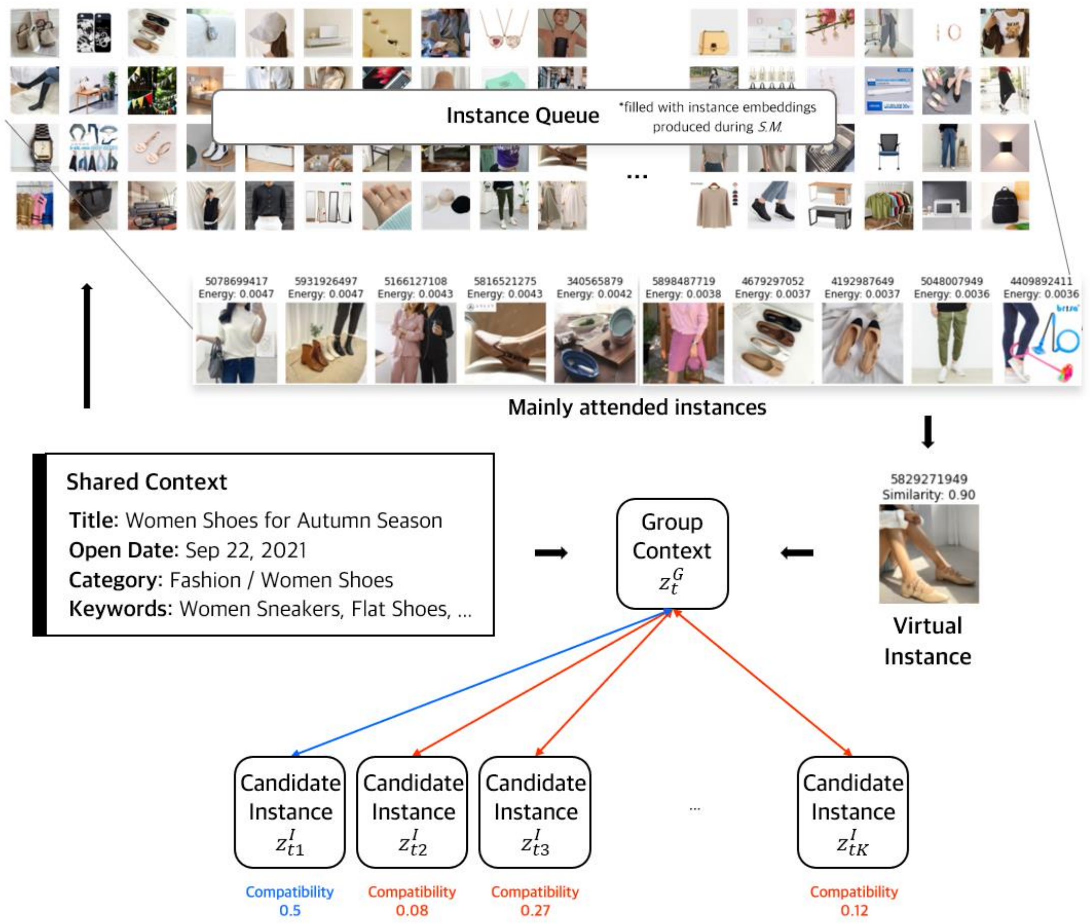
- Intuitive Example



Model Description - CLIK: Contrastive Learning of text and Image for ranKing

Task II. Preference Discrimination

- Intuitive Example



- Role of Group Context z_t^G
  - **With additional information, make anchor representation rich**
    helpful for contents-only-based recommendation
  - **Prevent confusion between two training tasks**
    with new modality of z_t^G, ironic positive/negative aligning can be avoided (without z_t^G, train/generalization performance both become poor)

Summary

- Loss function

$$L = L_{matching} + \lambda \cdot L_{discrim} \quad (default\ \lambda = 0.05)$$

Experiments

Dataset

- Collected from Online Special Exhibition, NAVER Shopping
- De-duplicate & removed data leakage

Type	# Train Exhibitions	# Eval Exhibitions	# Train Items	# Eval Items	Target
1	3.7K	290	2.6M	14K	Review Counts
2	1.2K	343	83K	17K	CTR

Metric

- Evaluation for Ranking
  - $MRR(X_n) = \frac{1}{Predicted\ Rank\ of\ x_{n*}}$
  - $TopK - Top1\ Accuracy$
    $$= \frac{1}{N}\sum_{n=1}^{N}\begin{cases}1 & if\ (true\ rank\ of\ the\ instance\ predicted\ as\ 1^{st}) \leq K \\ 0 & others\end{cases}$$
- Evaluation for context ↔ instance understanding
  - Linear Evaluation Protocol with frozen dual encoders
    Target: Exhibitions & products category
    ex) Label of laptop at digital exhibition: 'laptop-digital'
    Epochs / LR / Batch Size: 10 / 1e-4 / 32

Experiments

Implementation/Train Detail

context enc. (text)	instance enc. (image)	emb. dim	queue size	agg. method	attn. method	# params
BERT 70M	ViT S/16	128	512	concat	dot-product	92.9M

Epochs	LR (text)	LR (image)	optim	scheduler	batch size(M) (task I)	batch size (task II)
10	5e-5	1e-4	AdamW (decay 0.01)	cos. annealing 1% warm-up	512	20(K), 12(D)

Baselines

- Triplet Loss (# Params: 92.9M)
- Pairwise Loss (# Params: 93.0M)
- Pointwise Loss (# Params: 92.9M)

*Backbones of dual encoders are the same as CLIK

Epochs	LR	optim	scheduler	batch size
10	1e-4	AdamW (decay 0.01)	cos. annealing 1% warm-up	32

Experiments

Main Results

- Result for Data Type 1 (3.7K Exhibitions, Review Counts)

Model	Target	Linear Eval.	MRR	Top1-Top1	Top3-Top1	Top5-Top1
CLIK	Review Count	**0.8347**	<u>0.1561</u>	**0.0621**	**0.1345**	**0.1897**
Triplet		0.2541	**0.1645**	0.0448	<u>0.1</u>	<u>0.1414</u>
Pairwise		<u>0.2784</u>	0.1380	0.0448	0.0828	0.131
Pointwise		0.2509	0.1078	0.0207	0.0517	0.0724
Random		0.0103	0.0899	0.02	0.06	0.1

- Result for Data Type 2 (1.2K Exhibitions, CTR)

Model	Target	Linear Eval.	MRR	Top1-Top1	Top3-Top1	Top5-Top1
CLIK	CTR	**0.7528**	**0.1226**	**0.0496**	0.0729	<u>0.1283</u>
Triplet		<u>0.2682</u>	0.102	0.0233	<u>0.0758</u>	0.1254
Pairwise		0.2192	0.1063	0.0379	0.0641	0.1195
Pointwise		0.239	<u>0.121</u>	0.0379	**0.0947**	**0.1457**
Random		0.0119	0.0899	0.02	0.06	0.1

Additional Experiments

- *Usage of Group Context z^G*

Model	Target	Linear Eval.	MRR	Top1-Top1	Top3-Top1	Top5-Top1
CLIK	Review Count	**0.8347**	**0.1561**	**0.0621**	**0.1345**	**0.1897**
CLIK w/o z^G		0.8280	0.0909	0.0207	0.0621	0.1
Random		0.0103	0.0899	0.02	0.06	0.1

- *Make contrastive representation learning more powerful*

  Limitation of exhibition dataset: too few texts compared to images

  With more texts, performance can be improved with NO annotations

Model	Target	Linear Eval.	MRR	Top1-Top1	Top3-Top1	Top5-Top1
CLIK	Review Count	0.8347	0.1561	0.0621	0.1345	0.1897
CLIK w txt aug.		**0.8752**	**0.1596**	**0.069**	**0.1345**	**0.2138**
Random		0.0103	0.0899	0.02	0.06	0.1

Appendix

Inference Comparison

Shared Context

Title: Trousers, Bending, Spandex Pants for Men
Open Date: Aug 30, 2021
Category: Fashion / Men Clothing
Keywords: MensBendingPants, MensTrousers, …

Inference Result of CLIK

Most of tops ranked at bottom

Pairwise Loss

Triplet Loss

Pointwise Loss

Thank you