An Image Dataset for Benchmarking Recommender Systems with Raw Pixels

Yu Cheng[1,2], Yunzhu Pan[2], Jiaqi Zhang[2], Yongxin Ni[2], Aixin Sun[3] , and Fajie Yuan[2]

[1] Zhejiang University; [2] Westlake University

[3] Nanyang Technological University

西湖大學
WESTLAKE UNIVERSITY

Overview

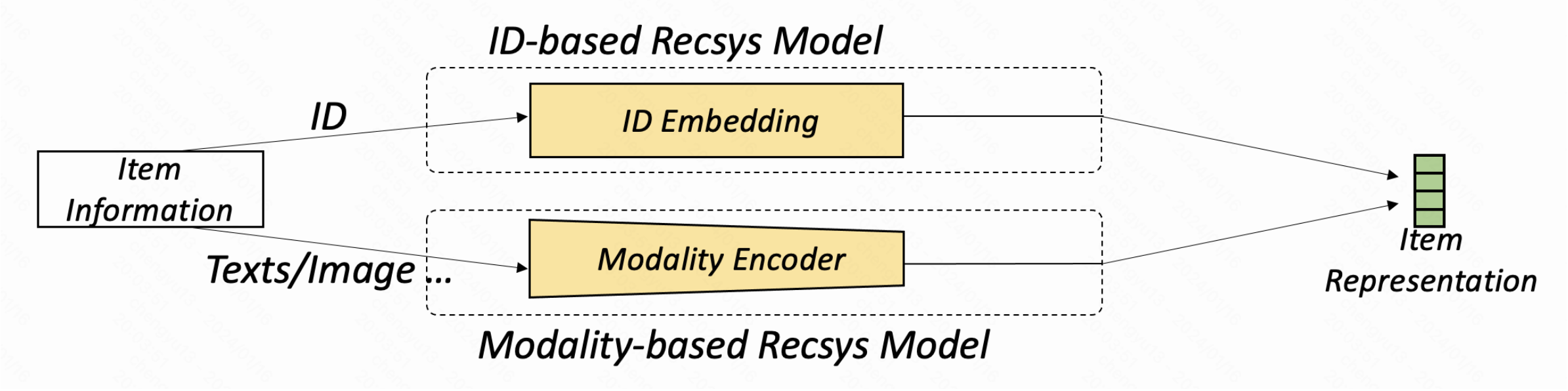
1. Background and Motivation

2. PixelRec Dataset

3. Contribution

4. Future Works

Background

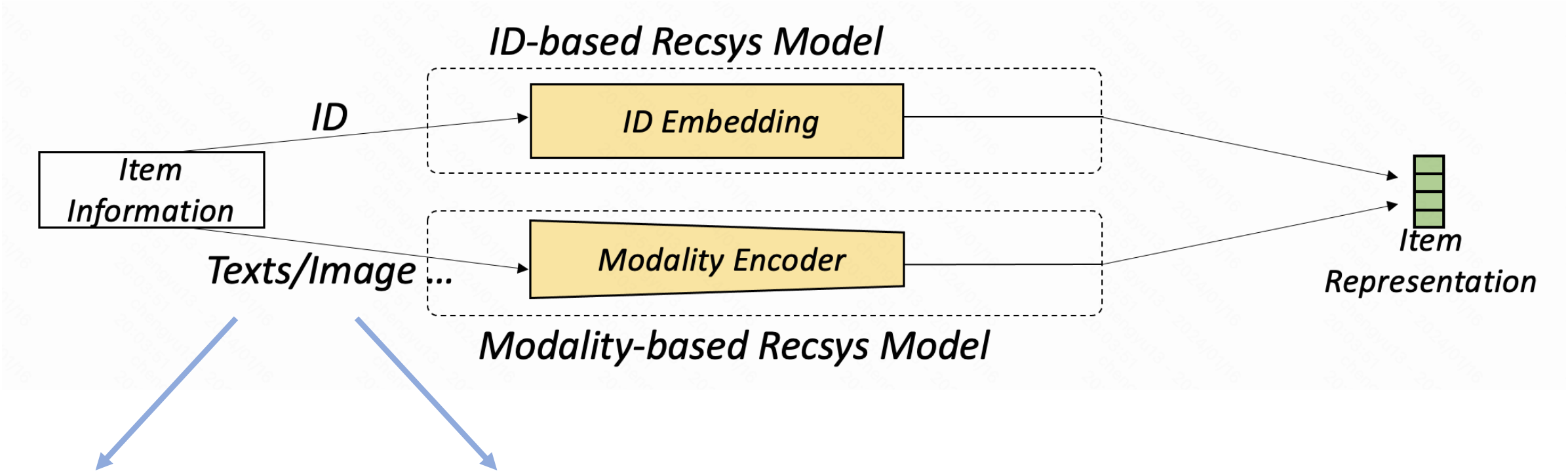


Encouraged by the success of LLMs

ID-based Recsys → Modality-based Recsys

Targeting at foundation models in Recsys

Background



Active trend

Limited research

BUT particular important in art, short videos ...

Background

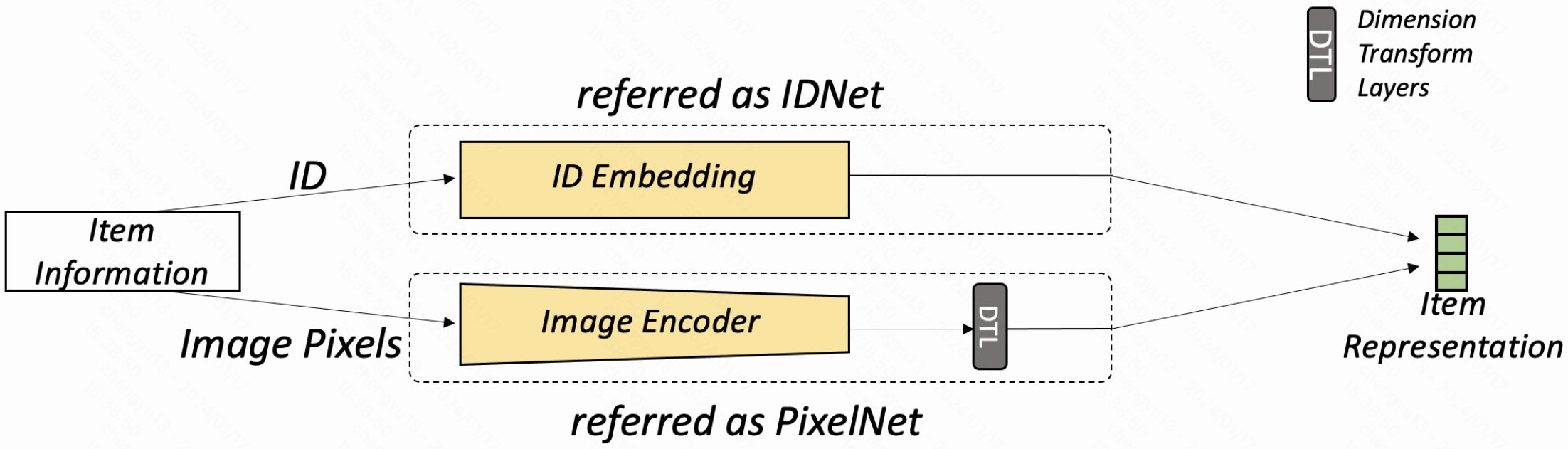


Fig. Item representation module in recommender model

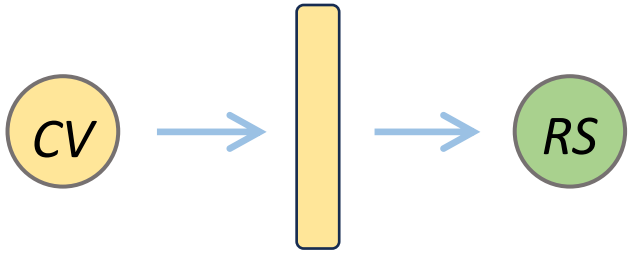
PixelNet

- Only take raw images as input of item (remove item ID in recommender model)
- Train recommender model and image encoder under end to end manner (guarantee high accuracy)

Motivation

Key weaknesses of existing visual recsys dataset

Pre-extracted feature vectors

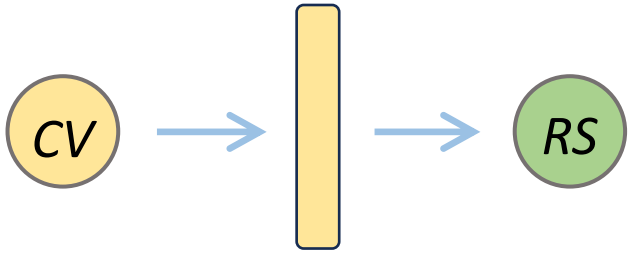


(1) Mismatch in tasks and vocabs

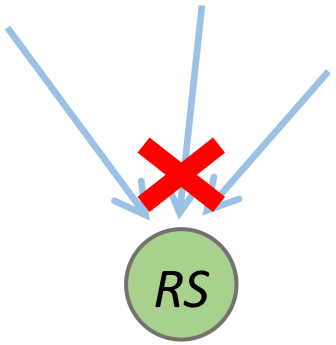
Motivation

Key weaknesses of existing visual recsys dataset

Pre-extracted feature vectors



Updated vision models



(1) Mismatch in tasks and vocabs

(2) Hindering technological advancement

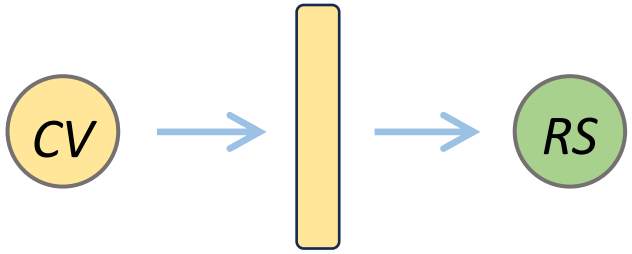
Motivation

Key weaknesses of existing visual recsys dataset

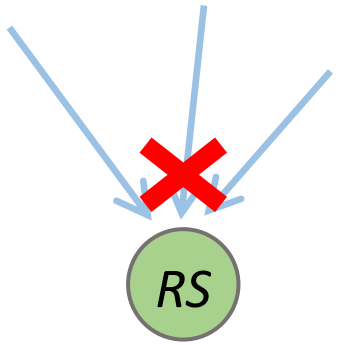
Behance[1], Flickr[2], KuaiRec[3]

Pinterest[4], WikiMedia[5]

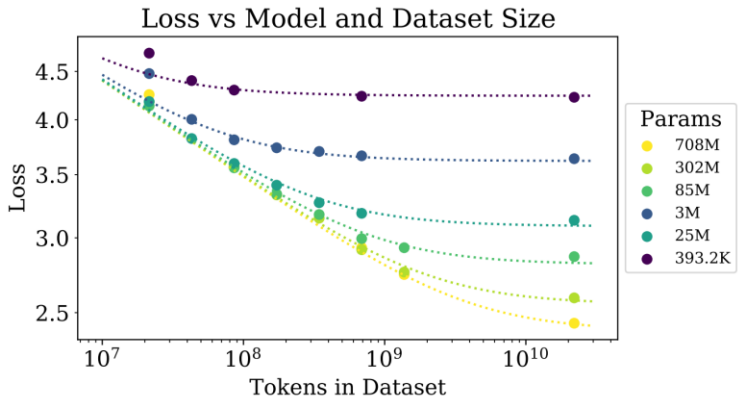
Pre-extracted feature vectors



Updated vision models



Scaling law in LLMs



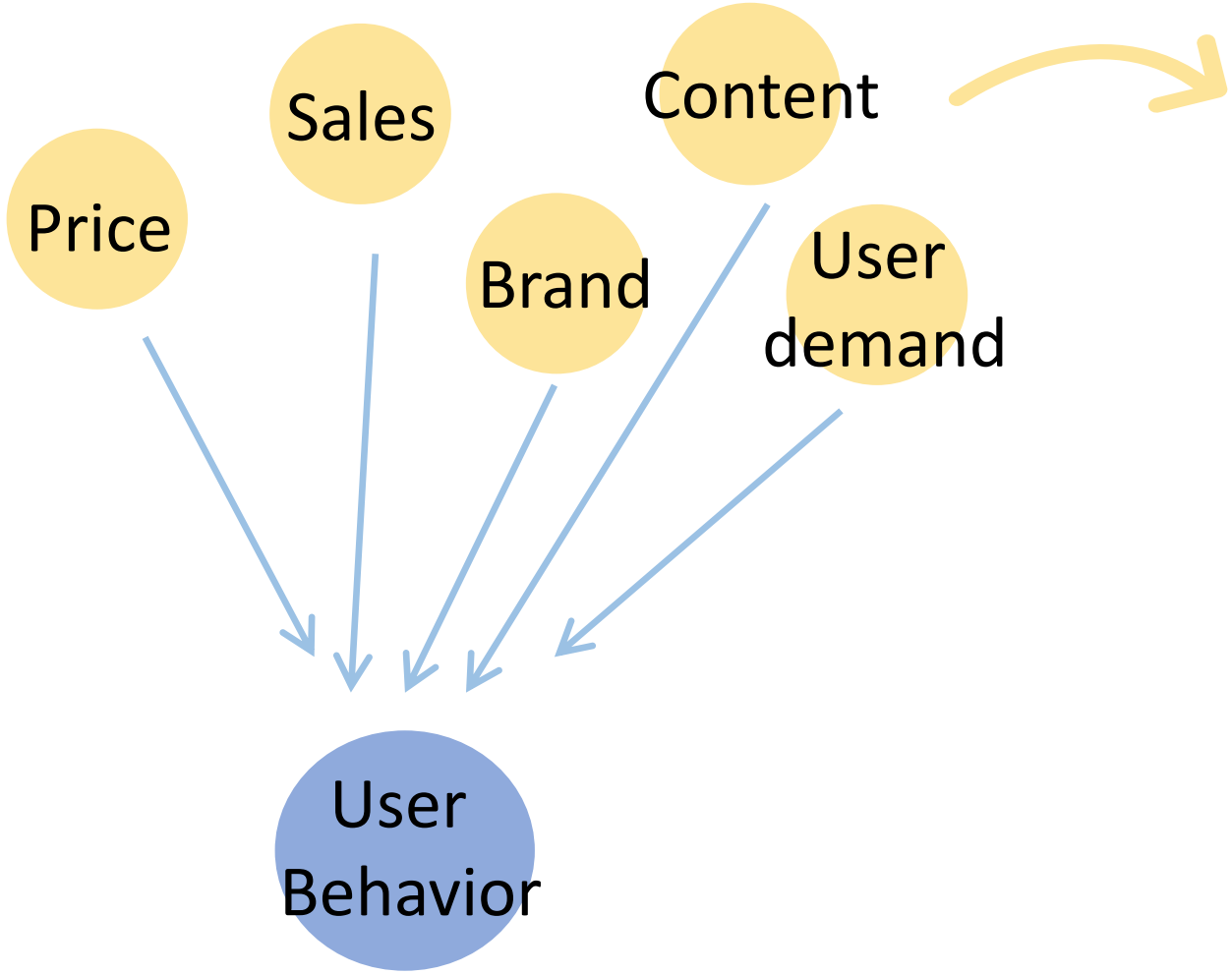
(1) Mismatch in tasks and vocabs

(2) Hindering technological advancement

(3) Limited scale

E-commence dataset ⇨ Not a typical visual recommendation scenario

Amazon[6], HM[7], GEST[8]



" A black and white football "

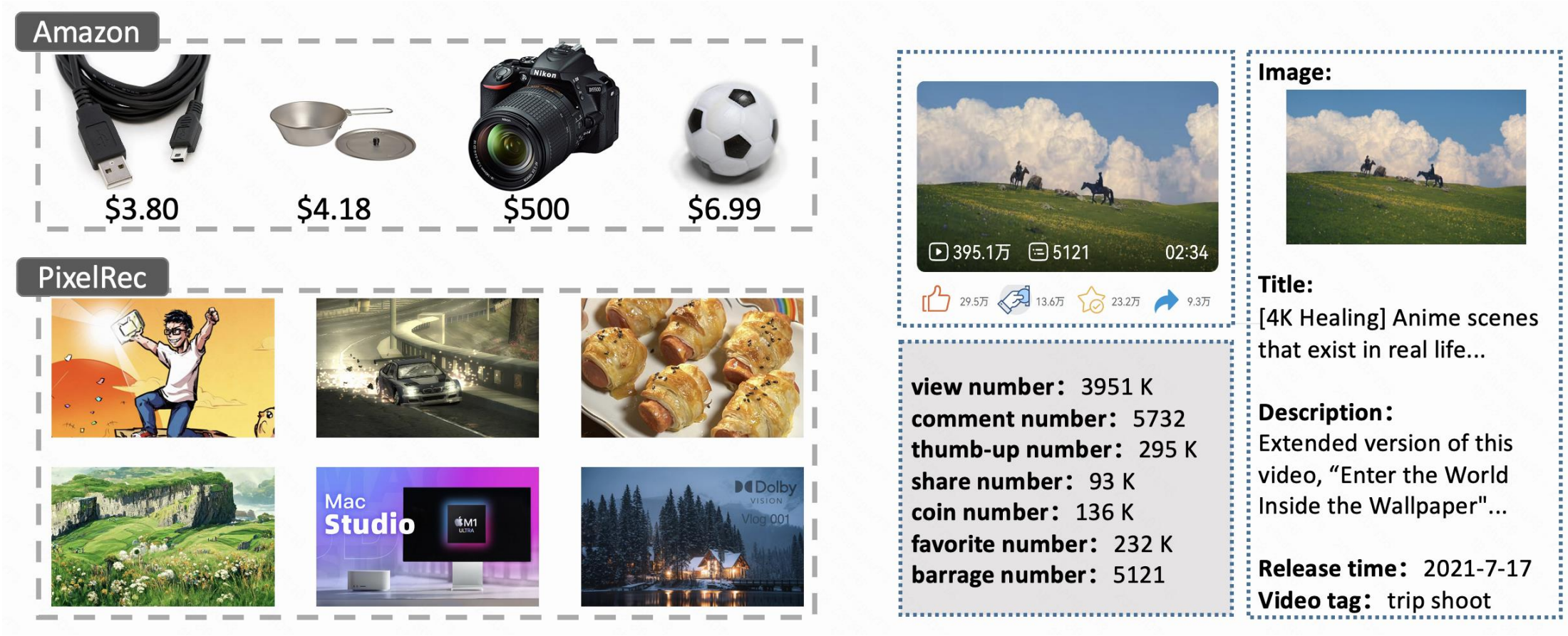
- E-commence is not a content-driven scenario

- Images are relatively simple and easy to describe

Overview

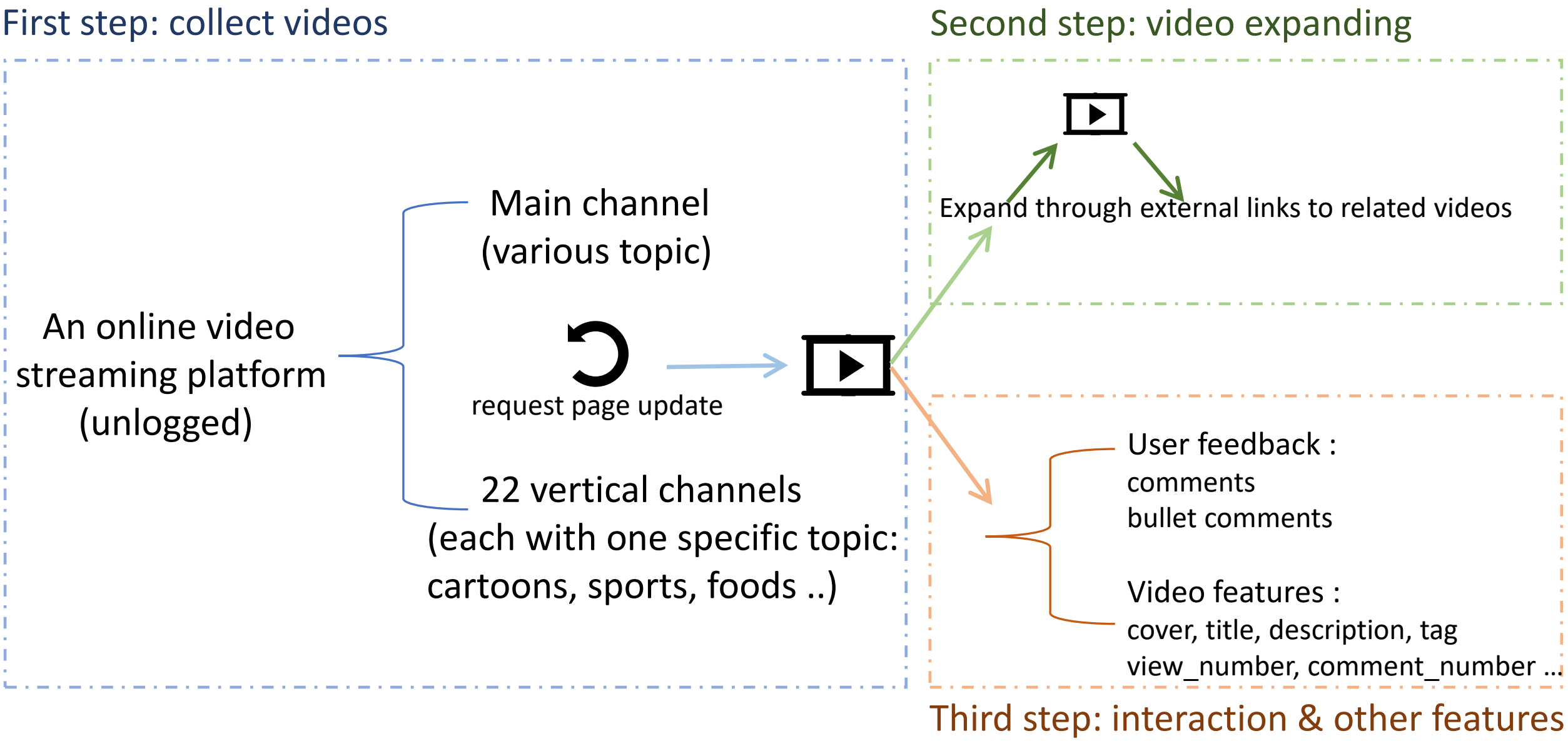
1. Background and Motivation
2. PixelRec Dataset
3. Contribution
4. Future Works

PixelRec (Overview)



- Raw images
- Diversity of visual elements
- Rich features
- Content-driven scenario
- Large scale
- Pivot role of image

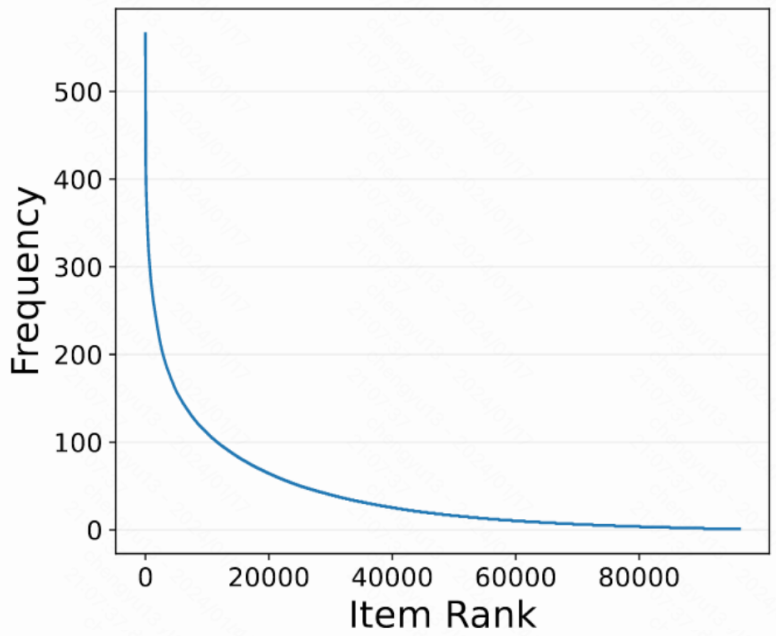
PixelRec (Details)



From September 2021 to October 2022, 13 months in total

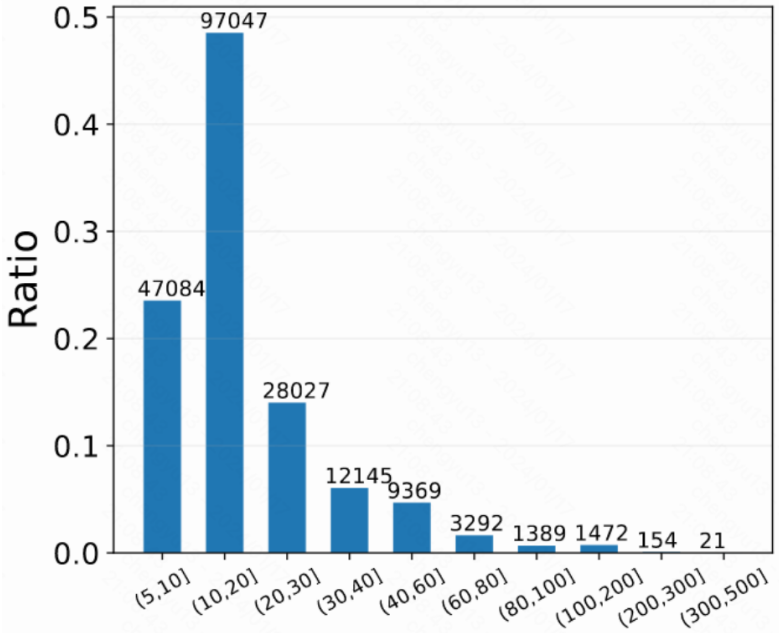
PixelRec (Statistics)

	Pixel1M	Pixel8M	PixelRec
#User	1,001,822	8,886,078	29,845,039
#Item	100,541	407,082	408,374
#Interaction	19,886,579	158,488,652	195,755,320



Statistics of Pixel200K

#User	200,000	#Item	96,282	#Inter.	3,965,656
#User.avg	19.83	#Item.avg	41.19	Sparsity	99.97%



(b) Session Length

Overview

1. Background and Motivation

2. PixelRec Dataset

3. Contribution

4. Future Works

Contribution

- PixelRec dataset
- PixelRec benchmark
- Exploratory results
- Baseline algorithms & Operating pipeline

Contribution (PixelRec dataset)

- PixelRec dataset
  - High-resolution raw images
  - Effective and precise image-based recommendation

Contribution (PixelRec dataset)

- PixelRec dataset
  - High-resolution raw images
  - Rich features
  - Short video domain
  - Short-video/ multimodal recommendation

Contribution (PixelRec dataset)

- PixelRec dataset
  - High-resolution raw images
  - Diversity of visual elements
  - Pivot role of image in user decision making
  - Challenging and high-quality benchmark for image recommendation tasks
  - Facilitate Bridging of RS and CV domains

Contribution (PixelRec dataset)

- PixelRec dataset
  - High-resolution raw images
  - Large scale
  - Diversity of visual elements
  - Pivot role of image in user decision making

- Pre-training resource for foundation vision recommendation models

Contribution (PixelRec dataset)

- PixelRec dataset

  - Content-driven scenario
  - Developing recommender models that prioritize item contents.

Contribution (PixelRec dataset)

- PixelRec dataset

- Pivot role of image in user decision making
- Studying preference models founded solely on images

Contribution (PixelRec benchmark)

- 9 recommender models
  - non-sequential : MF, DSSM, FM
  - sequential : GRU4Rec, NextItNet, SR-GNN, SASRec, BERT4Rec, LightSANs

- 9 image encoders
  - Transformer backbone: CLIP-ViT, Swin Transformer tiny, Swin Transformer base, BEiT
  - CNN backbone: ResNet50, CLIP-RN50, CLIP-RN50x4, CLIP-RN50x16, CLIP-RN50x64

- Exhaustive search on hyper-parameters of IDNet baselines
  - embedding size [128, 512, 1024, 2048, 4096, 8192]
  - batch size [64, 128, 512, 1024]
  - …

Contribution (Exploratory results)

ItemEnc	Metrics	Non-Sequential Recommender			Sequential Recommender					
		MF	FM	DSSM	SRGNN	GRU4Rec	BERT4Rec	NextItNet	SASRec	LightSANs
ID	Recall@10	1.013	1.357	1.401	1.597	1.833	1.972	2.187	2.500	2.578
	NDCG@10	0.490	0.679	0.701	0.808	0.937	0.994	1.153	1.350	1.384
RN50	Recall@10	0.357	1.024	0.960	2.224	2.294	2.391	2.140	2.633	2.417
	NDCG@10	0.169	0.501	0.475	1.132	1.138	1.199	1.073	1.321	1.226
ViT	Recall@10	0.472	1.124	1.242	2.152	2.102	2.450	2.215	2.583	2.461
	NDCG@10	0.229	0.543	0.617	1.065	1.031	1.230	1.106	1.292	1.224

Observation:

For Non-Sequential Recommender:
PixelNet << corresponding IDNet counterparts

For Sequential Recommender:
PixelNet ≈ corresponding IDNet counterparts

Contribution (Exploratory results)

ItemEnc	Metrics	Non-Sequential Recommender			Sequential Recommender					
		MF	FM	DSSM	SRGNN	GRU4Rec	BERT4Rec	NextItNet	SASRec	LightSANs
ID	Recall@10	1.013	1.357	1.401	1.597	1.833	1.972	2.187	2.500	2.578
	NDCG@10	0.490	0.679	0.701	0.808	0.937	0.994	1.153	1.350	1.384
RN50	Recall@10	0.357	1.024	0.960	2.224	2.294	2.391	2.140	2.633	2.417
	NDCG@10	0.169	0.501	0.475	1.132	1.138	1.199	1.073	1.321	1.226
ViT	Recall@10	0.472	1.124	1.242	2.152	2.102	2.450	2.215	2.583	2.461
	NDCG@10	0.229	0.543	0.617	1.065	1.031	1.230	1.106	1.292	1.224

Conclusion:

1. Adopting sequential recommender backbone and end2end training strategy, PixelNet perform satisfactorily in regular recommendation setting

2. The performance of PixelNet may be significantly influenced by the specific recommendation backbone network and training approach used

Contribution (Exploratory results)

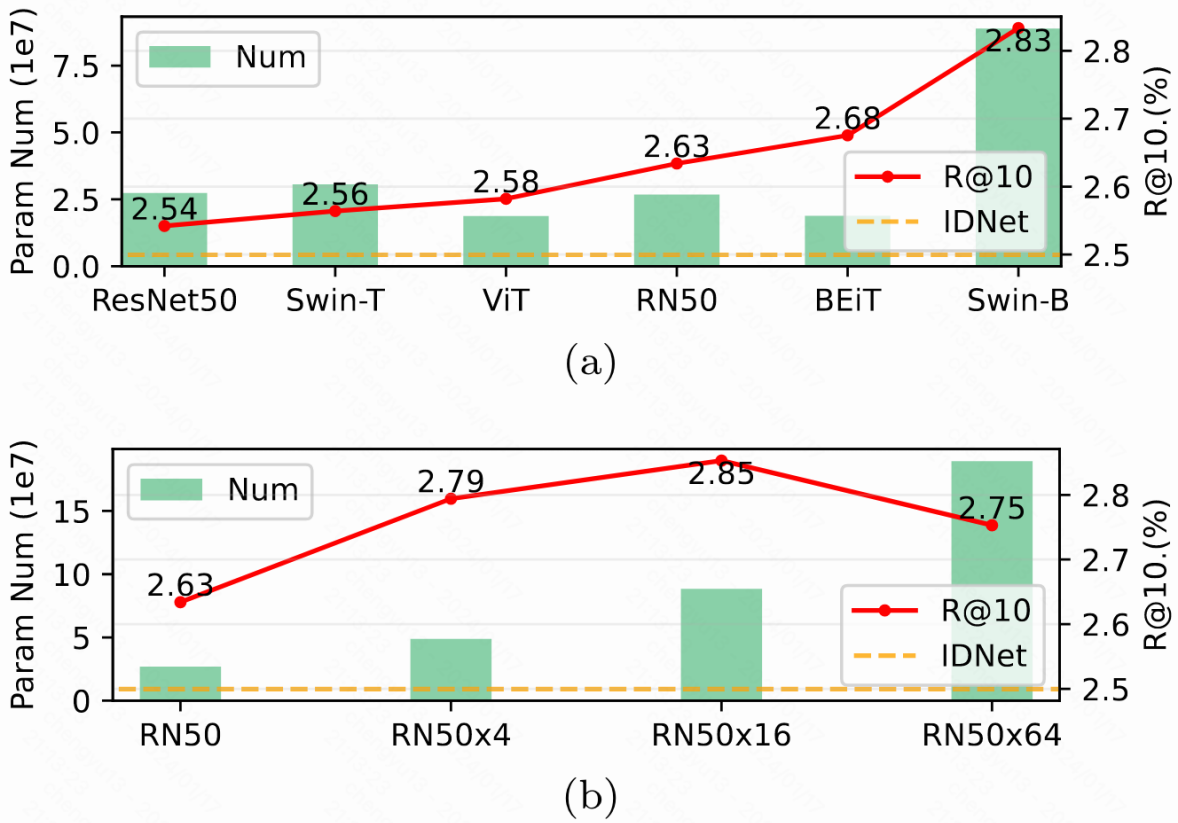


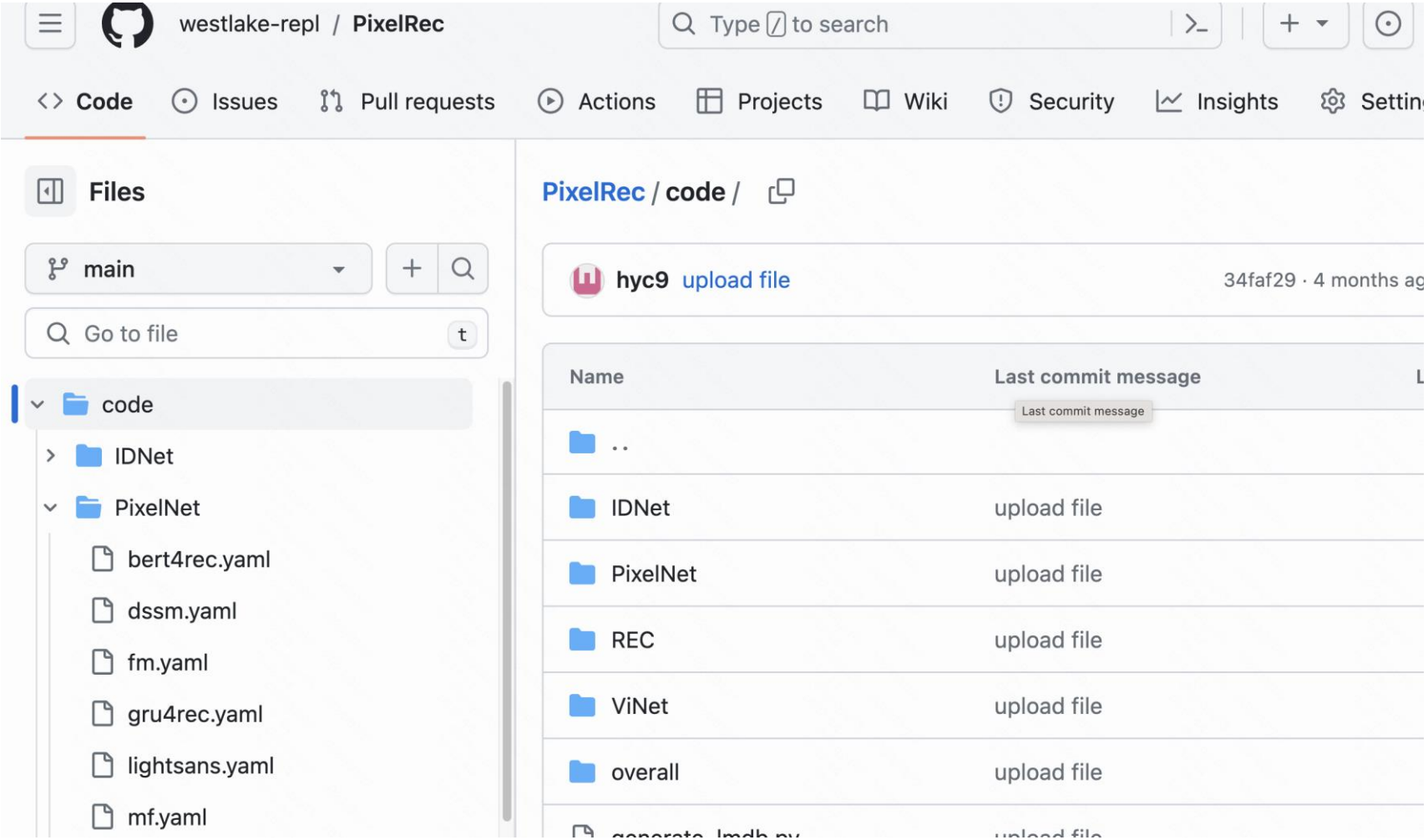
Figure 4: Benchmark image encoders on Pixel200K. The dashed yellow line is the accuracy of IDNet. The green bar chart is the number of trainable parameters. The red line chart is the recall@10.

Conclusion:

1. Larger image encoders do lead to improved performance, but only up to a certain point

2. Both recommendation architectures and image encoders play important roles in the effectiveness of PixeNet

Contribution (Baseline algorithms & Operating pipeline)



Abundant Baseline algorithms

Traditional ID-based recommender
Traditional visual recommender
Pixel-based recommender

Complete operating pipeline

Data processing
Model loading
Model training
Model inference
Hyper-parameter record
....

Access link : https://github.com/westlake-repl/PixelRec

Overview

Background and Motivation

PixelNet Dataset

Contribution

Future Works

Future works

- Reducing computation consumption of end2end training
- Effective hyper-parameter tuning of PixelNet
- Building foundation vision recommender models

An Image Dataset for Benchmarking Recommender Systems with Raw Pixels

Yu Cheng[1,2], Yunzhu Pan[2], Jiaqi Zhang[2], Yongxin Ni[2], Aixin Sun[3] , and Fajie Yuan[2]

Q&A

西湖大學
WESTLAKE UNIVERSITY

Reference

[1] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. 2016. Vista: A visually, socially, and temporally-aware model for artistic recommendation. In Proceedings of the 10th ACM conference on recommender systems. 309–316.

[2] Le Wu, Lei Chen, Richang Hong, Yanjie Fu, Xing Xie, and Meng Wang. 2019. A hierarchical attention model for social contextual image recommendation. IEEE Transactions on Knowledge and Data Engineering (2019).

[3] C. Gao, S. Li, W. Lei, J. Chen, B. Li, P. Jiang, X. He, J. Mao, and T.-S. Chua, Kuairec: A fully-observed dataset and insights for evaluating recommender systems, in Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 540–550.

[4] Xue Geng, Hanwang Zhang, Jingwen Bian, and Tat-Seng Chua. 2015. Learning image and user features for recommendation in social networks. In ICCV.

Reference

[5] Denis Parra, Antonio Ossa-Guerra, Manuel Cartagena, Patricio Cerda-Mardini, and Felipe del Rio. 2021. VisRec: A Hands-on Tutorial on Deep Learning for Visual Recommender Systems. In 26th International Conference on Intelligent User Interfaces-Companion. 5–6.

[6] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In SIGIR.

[7] https://www.kaggle.com/competitions/h-and-m-personalized-fashionrecommendations

[8] An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, and Julian McAuley. 2022. Personalized Showcases: Generating Multi-Modal Explanations for Recommendations.