Executive Summary

- ***Visually Similar Items:***
  Visual Recommendation Feature at Mercari

- Visual Recommendation (baseline approach):
  Embeddings from product images via MobileNetV2 [1] for ANN
  → **Limitations:**
  - **Image feature representation**
  - **Recommendation performance**

- Visual recommendation system using **SigLIP [2] image encoder**
  - Offline Evaluation: **nDCG@5 +9.1%**
  - Online Evaluation: **CTR +50%, CVR +14%**

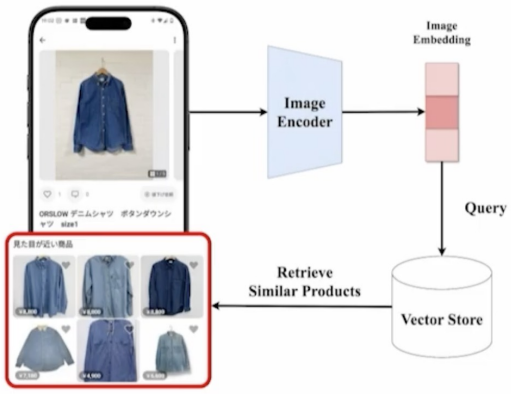

Fig.1. Visual Recommendation System

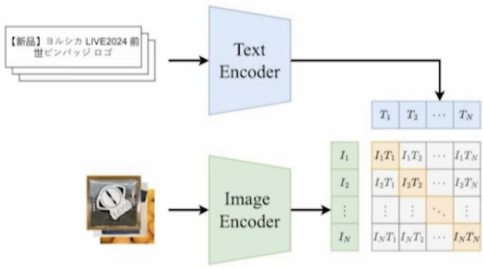

Fig. 2. Model Training

Visual Recommendation at Mercari

- ***Mercari***
  A major **C2C marketplace with 20+ million MAU**, the largest in Japan

- Visual Recommendation
  - Recommendations based on visual features such as color, shape, and patterns
  - Implemented across numerous e-commerce platforms [3-5]

- **"Visually Similar Items" on item detail page:**
  Visual Recommendation at Mercari

Figure 3. Visually Similar Item at Mercari

“Visually Similar Items”

Visual Recommendation System (Fig.1.):

1) Transform product image into a vector via image encoder
2) ANN search is performed on the Vector Store to obtain visually similar products

- Existing Visual Recommendation: CNN-based MobileNetV2 [1] as the Image Encoder
  → **Computationally efficient, but limited in feature expressiveness**

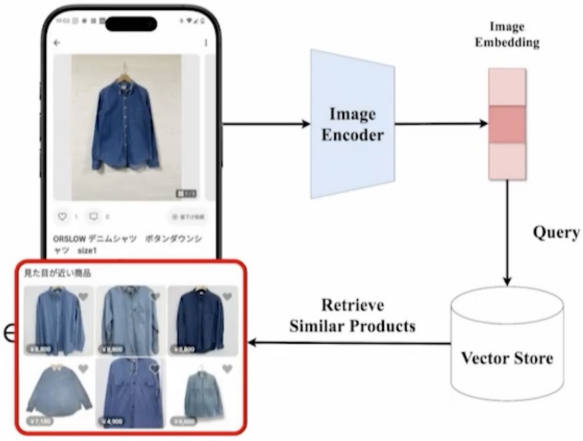

Figure 1. Visual Recommendation System (reproduced)

Contribution

- Applied **Vision Language Model (VLM) Image Encoder** for visual recommendations on e-commerce platforms
- Visual recommendations using VLM demonstrated **superior performance in offline evaluation with key retrieval metrics** compared to conventional methods
- Deployed the VLM-based visual recommendation to an e-commerce platform serving over 20+ million MAU
  → Online evaluation via A/B testing showed **significant business performance (CTR/CVR) improvements**

SigLIP: Sigmoid Loss for Language Image Pre-Training [2]

- Models pre-trained on large-scale image-text pair datasets (e.g., CLIP [6])
  → High performance on zero-shot classification and retrieval tasks

- **SigLIP[2]**
  - Employs Sigmoid Loss function (CLIP uses Softmax Loss)
  - **SigLIP demonstrated substantial performance improvements** over existing methods across multiple benchmarks.

- Adopted pre-trained model: **Multilingual mSigLIP pre-trained on WebLI [8]**
  - **Image encoder: ViT-B/16**

Fine-tuning SigLIP with Product Data

Dataset

- Trained on a dataset constructed from product image-title pairs from Mercari

- 1,000,000 items from products listed from April 29 to July 29, 2024

Training Method

- Contrastive Learning on product image-title pairs

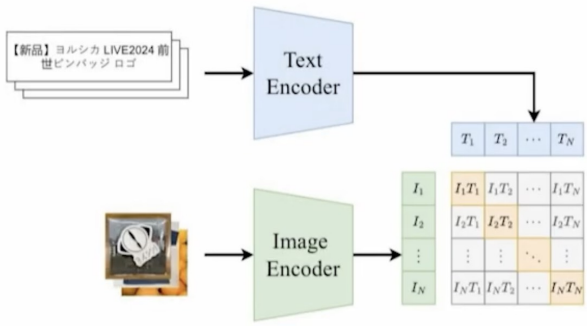

Figure 2. Fine-Tuning SigLIP (reproduced)

Performance Evaluation

Evaluation Method

- Evaluated embedding model performance through both offline analysis and online A/B testing:
  - MobileNetV2 (Baseline)
  - SigLIP / SigLIP + PCA (768 dim to 128 dim)

Offline Evaluation

- Evaluate retrieval metrics using user tap events from the existing system: **nDCG@K (K=5), Precision@K (K=1,3)**

Online Evaluation

- A/B test for 2 weeks by releasing to the "Visually Similar Items" feature on Mercari product detail pages.

Offline Evaluation

Table.1. Performance comparison of image encoders

	nDCG@5	Precision@1	Precision@3
MobileNetV2	0.607	0.356	0.601
SigLIP	0.662	0.412	0.660
SigLIP + PCA	0.647	0.406	0.658

Offline Evaluation: Dimensionality Reduction by PCA (768 to 128)

Table.1. Performance comparison of image encoders

	nDCG@5	Precision@1	Precision@3
MobileNetV2	0.607	0.356	0.601
SigLIP	0.662	0.412	0.660
SigLIP + PCA	**0.647**	**0.406**	**0.658**

- Dimensionality compression **from 768 dim to 128 dim** using PCA
  - **83% storage cost savings** for the vector database
  - **Minimal performance degradation of 2.3%** in nDCG@5 relative to original SigLIP (0.662 → 0.647)

mercari

Online Evaluation via A/B test

- 2 weeks A/B test on "Visually Similar Items" on Mercari product detail pages
  - Control (Baseline): MobileNetV2
  - Treatment: SigLIP + PCA
- A/B test results:
  - Click-through rate (CTR): **+50%**
  - Buy count from detail pages (CVR): **+14%**

Figure 3. Visually Similar Item at Mercari (reproduced)

mercari

Conclusion

- Applied **SigLIP VLM's image encoder for visual recommendation on Mercari**, the marketplace platform serving 20+ million MAU.

- Evaluated model performance for visual recommendations through offline analysis of tap logs and online A/B test evaluation:
  - MobileNetV2 (Baseline)
  - SigLIP / SigLIP + PCA (768 dim to 128 dim)

- Offline evaluation on retrieval metrics (nDCG@K, Precision@K):
  SigLIP-based visual recommendations outperformed the baseline method.

- 2-week A/B test on "Visually Similar Items" on Mercari's product detail pages:
  → CTR +50%, Buy count from detail page (CVR) +14%