Extending CLIP for Category-to-image Retrieval in E-commerce

Mariya Hendriksen m.hendriksen@uva.nl

University of Amsterdam

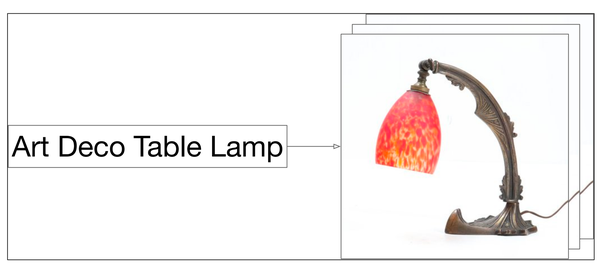
Motivation

Product category tree:
• Assist customers when navigating product catalogue.
• Ability to retrieve an image for a given category is a challenge due to:
  • noisy category and product data
  • size and dynamic character of product catalogues

Multi-modal data in e-commerce:
• Current e-commerce search focuses on textual and behavioural signals.
• Multimodal product data is barely used.
• Prior work mainly on Fashion retrieval.
• Knowledge gap: multimodal retrieval in general e-commerce domain.

Task

Category-to-image retrieval task
Given a category and a collection of products, retrieve a list of images of products that belong to a given category.



Task characteristics:
• Categories in category tree vary in granularity
• The category tree is not fixed, hence, we aim to generalise towards unseen categories.
• Modalities: text, image, attribute information, category tree

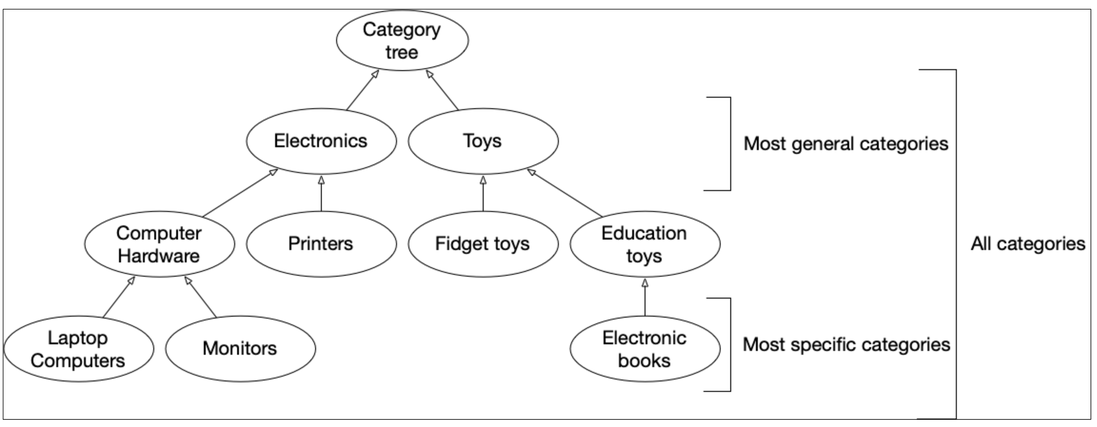
Research Questions

RQ1 How do baseline models perform on the category-to-image retrieval task?
  - unimodal vs. bi-modal models performance
  - performance w.r.t. category granularity
RQ2 How does combining information from multiple modalities impact the performance on the task?
RQ3 How can we improve performance on the task by leveraging product attribute and category tree information?

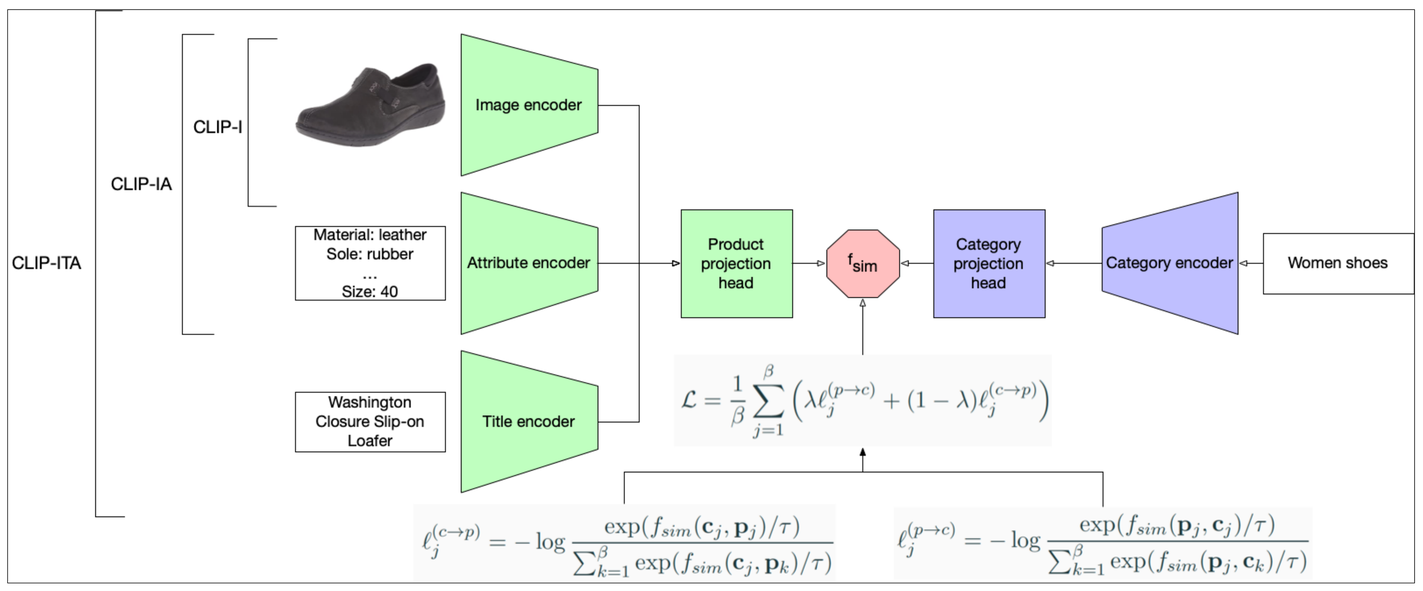
Approach

Metrics
Precision@K where K = {1; 5; 10}, mAP@K where K = {5; 10}, and R-precision.
Dataset
• Amazon XMarket dataset [1]
• Textual, visual, attribute information, category tree
• Modalities: text, image
Baselines
• Text-only: BM25, and MPNet
• Multimodal: CLIP

Evaluation w.r.t. category type



Model



Experiments

Evaluation:
1. Baselines, BM25, CLIP, MPNet.
2. Image-based product representations, CLIP-I.
3. Image and attribute-based product representations, CLIP-IA.
4. Image, attribute, and title-based product representations, CLIP-ITA.

Conclusion

• Introduced category-to-image retrieval task and the model for the task.
• Evaluated the model in three settings: all categories, most general categories, most specific categories.
• Multimodal models tend to outperform unimodal models.
• Combining textual, visual, and attribute information when building product represetations produces best results on the task.

The research is supported by Ahold Delhaize