Mo' models, mo' problems

CLIP-like models and generalization in modern e-commerce platforms

Content Understanding and Generation, KDD 2022

Jacopo Tagliabue

Ciao!

ENGINEERING

- Founder of Tooso, acquired by TSX:CVO, Director of AI at Coveo for the past 3 years
- Passionate about MLOps, OS contributor

AI RESEARCH & EDUCATION

- 25+ papers in top NLP/ML venues (Best Paper NAACL21), co-organizer of SIGIR eCom
- Adj. Prof. of MLSys at NYU Tandon

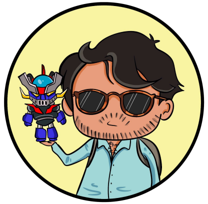
Today

- We will discuss ideas, models and projects originated within my team and collaborators at *CoveoLabs*.
- While I am the only speaker today, Patrick John, Federico and Ciro (and other people which unfortunately are without a chibi) reviewed these slides and share with me the credit for whatever value these ideas may have.
- **Obviously, all the remaining mistakes are theirs** 😁

Jacopo

Patrick John

Federico



Ciro

Today: 3 simple lessons

1. B2B eCommerce tech is hard - **we explain why is *so* hard**
2. The naive solution is not smart, nor efficient - **we detail what are the challenges involved**
3. More general models may help achieve sustainable unit economics, as well as unlock new use cases - **we show how CLIP-like models can be used and what we can learn from them**

The long tail wags the B2B dog*

* Credits to Arun at Eloquent AI!

B2C vs B2B in eCommerce tech

B2C Companies

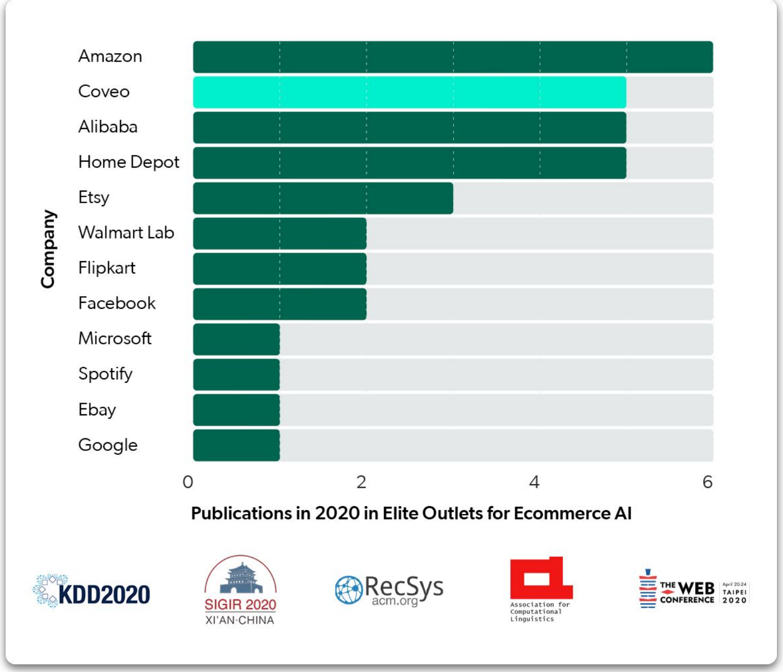
- **Business model**: they have a direct line to shoppers (shops / marketplaces).
- **Deployment target**: they deploy their technology on *their website* and control their data.
- **ROI on tech**: direct (precision comes first).
- **Examples**: Amazon, Alibaba, Etsy, etc.

B2B Companies

- **Business model**: they sell to shops, each with *their* own shoppers.
- **Deployment target**: they deploy technology on customer websites and have limited control on data they ingest.
- **ROI on tech**: indirect (robustness comes first).
- **Examples**: Coveo, Bloomreach, Algolia etc.

B2C vs B2B in eCommerce tech

- **Where does innovation come from?** The majority of innovation in e-commerce tech comes from very few, large, public B2C players (with one exception).



B2C vs B2B in eCommerce tech

- **Where does innovation come from?** The majority of innovation in e-commerce tech comes from very few, large, public B2C players (with one exception).
- **Why?**
    - Implementing ML is hard.
    - MLops is hard too.
    - Most literature is skewed towards the resources and problems of few companies.

ABSTRACT

You Do Not Need a Bigger Boat: Recommendations at Reasonable Scale in a (Mostly) Serverless and Open Stack

Author: Jacopo Tagliabue Authors Info & Claims

RecSys '21: Fifteenth ACM Conference on Recommender Systems • September 2021 • Pages 598–600 • https://doi.org/10.1145/3460231.3474604

Online: 13 September 2021 Publication History

0 547 Get Access

Even if we solve all of the above, there is still one problem that persists: chasing the tail.

Chasing the tail

“Think about search queries: few search queries account for the vast majority of traffic, and then thousands of queries are seen only once or twice.”

Taming the Tail: Adventures in Improving AI Economics

by Martin Casado and Matt Bornstein

AI, machine & deep learning • enterprise & SaaS • Company Building 101 • on the economics of AI/ML & data businesses

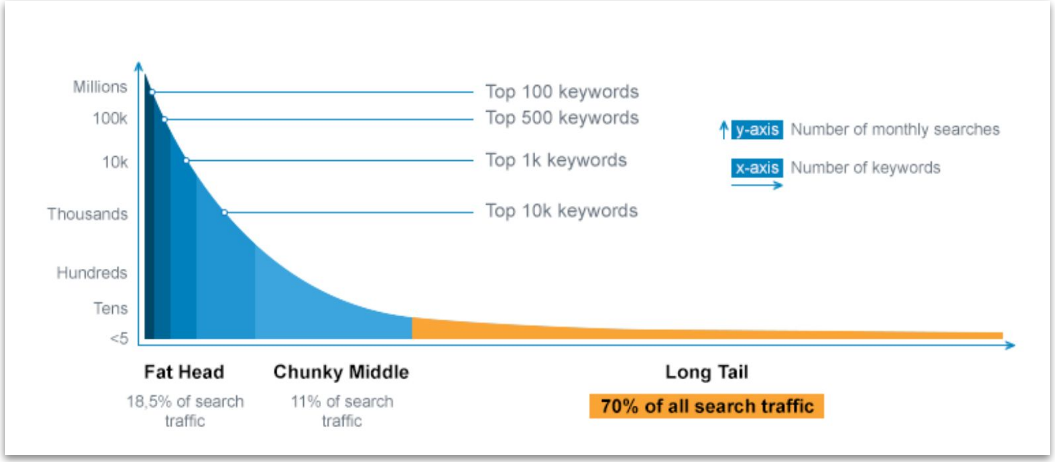
AI has enormous potential to disrupt markets that have traditionally been out of reach for software. These markets – which have relied on humans to navigate natural language, images, and physical space – represent a huge opportunity, potentially worth trillions of dollars globally.

However, as we discussed in our previous post The New Business of AI, building AI companies that have the same attractive economic properties as traditional software can be a challenge. AI companies often have lower gross margins, can be harder to scale, and don't always have strong defensive moats. From our experience, many of these challenges seem to be endemic to the problem space, and we've yet to uncover a simple playbook that guarantees traditional software economics in all cases.

That said, many experienced AI company builders have made tremendous progress in improving the financial profiles of their companies relative to a naive approach. They do this with a range of methods spanning data engineering, model development, cloud operations, organizational design, product management, and many other areas. The common thread

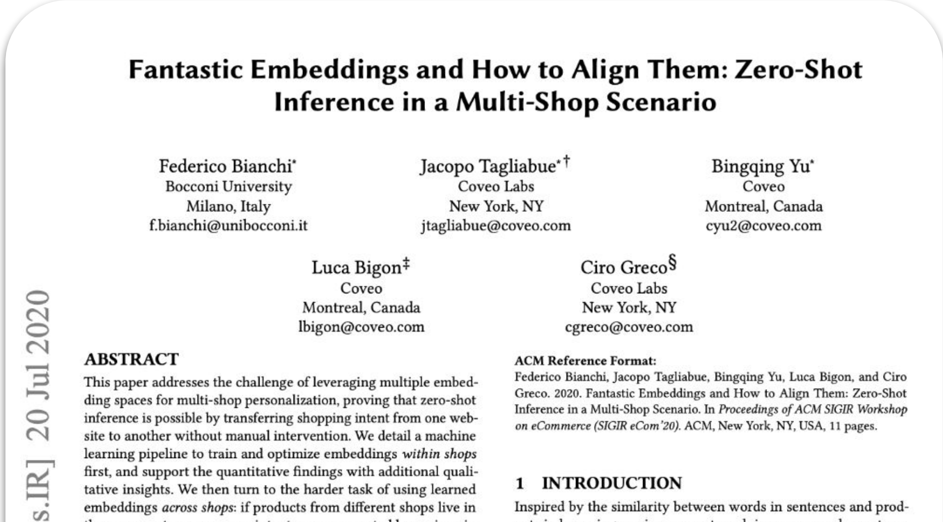
Chasing the tail

- In IR use cases, we have seen extreme values, e.g. top 2% queries account for ~50% query counts.
- This also applies to product in categories and behavioral shopping signals.



Chasing the tail

- In IR use cases, we have seen extreme values, e.g. top 2% queries account for ~50% query counts.
- This also applies to product in categories and behavioral shopping signals.



Fantastic Embeddings and How to Align Them: Zero-Shot Inference in a Multi-Shop Scenario

Federico Bianchi*
Bocconi University
Milano, Italy
f.bianchi@unibocconi.it

Jacopo Tagliabue*†
Coveo Labs
New York, NY
jtagliabue@coveo.com

Bingqing Yu*
Coveo
Montreal, Canada
cyu2@coveo.com

Luca Bigon‡
Coveo
Montreal, Canada
lbigon@coveo.com

Ciro Greco§
Coveo Labs
New York, NY
cgreco@coveo.com

s.IR] 20 Jul 2020

ABSTRACT

This paper addresses the challenge of leveraging multiple embedding spaces for multi-shop personalization, proving that zero-shot inference is possible by transferring shopping intent from one website to another without manual intervention. We detail a machine learning pipeline to train and optimize embeddings *within shops* first, and support the quantitative findings with additional qualitative insights. We then turn to the harder task of using learned embeddings *across shops*: if products from different shops live in

ACM Reference Format:
Federico Bianchi, Jacopo Tagliabue, Bingqing Yu, Luca Bigon, and Ciro Greco. 2020. Fantastic Embeddings and How to Align Them: Zero-Shot Inference in a Multi-Shop Scenario. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR eCom'20)*. ACM, New York, NY, USA, 11 pages.

1 INTRODUCTION

Inspired by the similarity between words in sentences and prod-

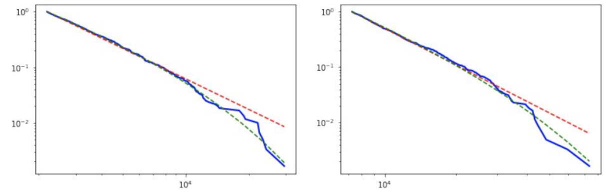


Figure 3: Shop A (left) and Shop B (right) log plots for product views: *empirical* distribution is in blue, *power-law* in red and *truncated power-law* in green. Truncated power-law is a better fit than standard power-law for both shops ($p < .05$), with $\alpha = 2.32$ for A and $\alpha = 2.72$ for B. Power-law analysis and plots are made with [1].

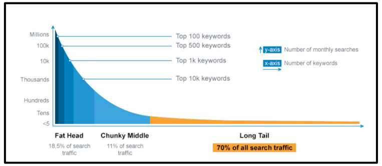
The B2C tail

- In B2C, your model needs to account for *one* distribution, over which you can often exercise *some* level of control:
  - Data tracking
  - Change in UI/UX
  - Data quality
- Improvements will have a cumulative effect and (hopefully) help with the generalization in the long-tail:
  - Improvement in the modelling code
  - Improvement in the underlying data (e.g. catalog quality)
  - Improvement in quantity / quality of data collection

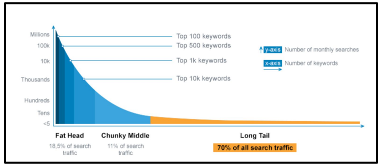
The B2B tail(s)

- In **B2B**, each customer / shop will have its own distribution - some sell shoes, some sell electronics:
  - Queries are different
  - Target items / catalogs are different (also in meta-data and quality!)
  - All behavioral data is also site-specific

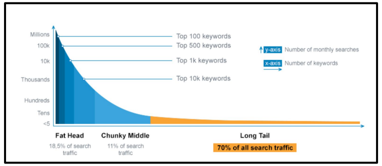
Shop 1



Shop 2



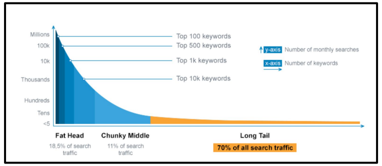
Shop 3



...

...

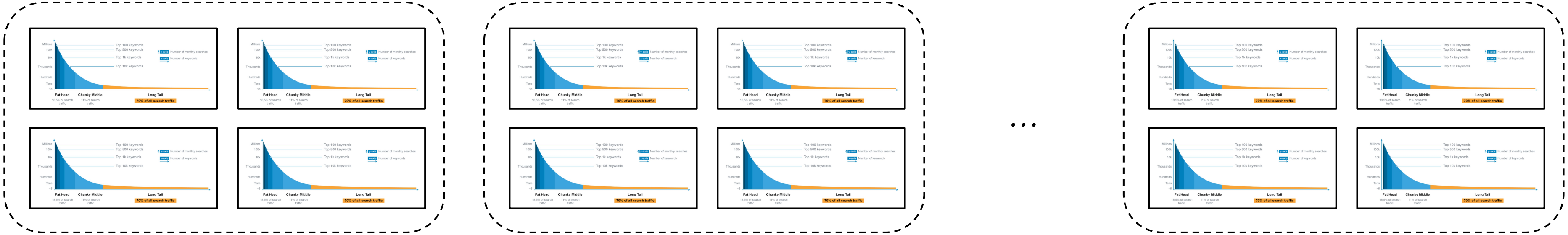
Shop n



The B2B tail(s)

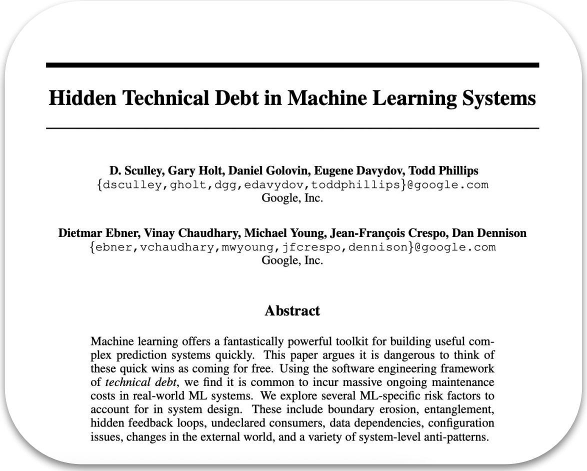
- In **B2B**, provider typically has different APIs to fulfil different needs - in a “naive” one-use-case-one-model scenario you may have:
  - A recommender system
  - A query ranking model
  - A type-ahead model
  - A catalog classification model

Shop 1 *Shop 2* … *Shop n*



The B2B tail(s)

- Even if learning was “solved”, operating thousands of per-use-case and per-client models come with its own costs.
    - [MLOps is still more art than science]



Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

Abstract

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of *technical debt*, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration issues, changes in the external world, and a variety of system-level anti-patterns.

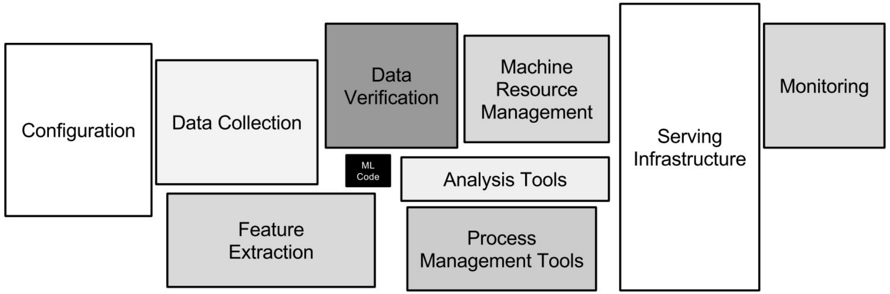


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

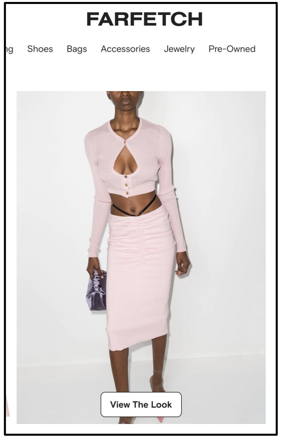
Mo' models, mo' problems

Isn't there a better way?

E pluribus unum

Content understanding in the age of ML

- Catalogs have two types of meta-data: **images** and **text**.
- Text usage is widespread:
  - search engine;
  - content-based recSys;
  - item classification.
- Image usage is way less popular:
  - visual search;
  - item classification.



New Season

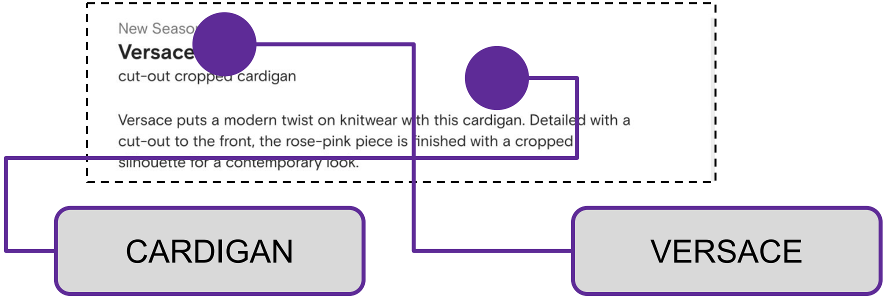
Versace

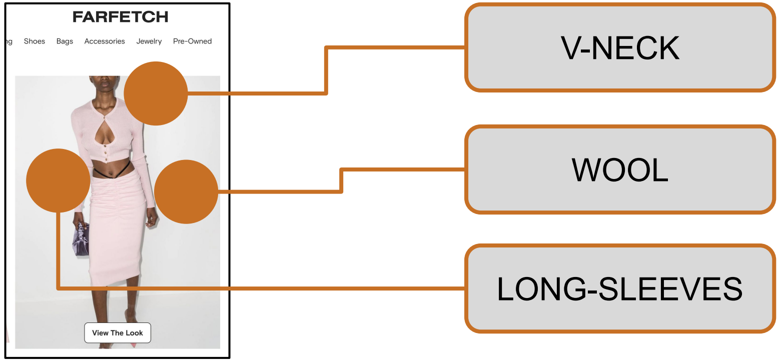
cut-out cropped cardigan

Versace puts a modern twist on knitwear with this cardigan. Detailed with a cut-out to the front, the rose-pink piece is finished with a cropped silhouette for a contemporary look.

Content understanding in the age of ML

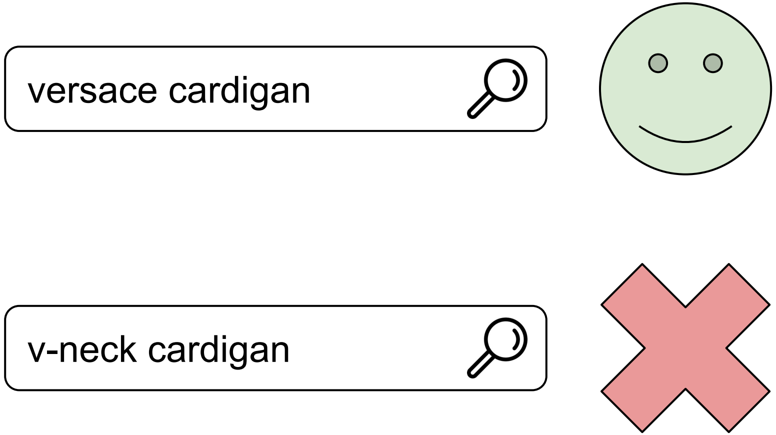
- Even when both are used, the typical scenario is *one-use-case-one-model*:
    - Text representations will result in a vector-space for textually-sourced concepts.
    - Image representations will result in a vector-space for visually-sourced concepts.





Content understanding in the age of ML

- Even when both are used, the typical scenario is *one-use-case-one-model*:
  - Text representations will result in a vector-space for textually-sourced concepts.
  - Image representations will result in a vector-space for visually-sourced concepts.

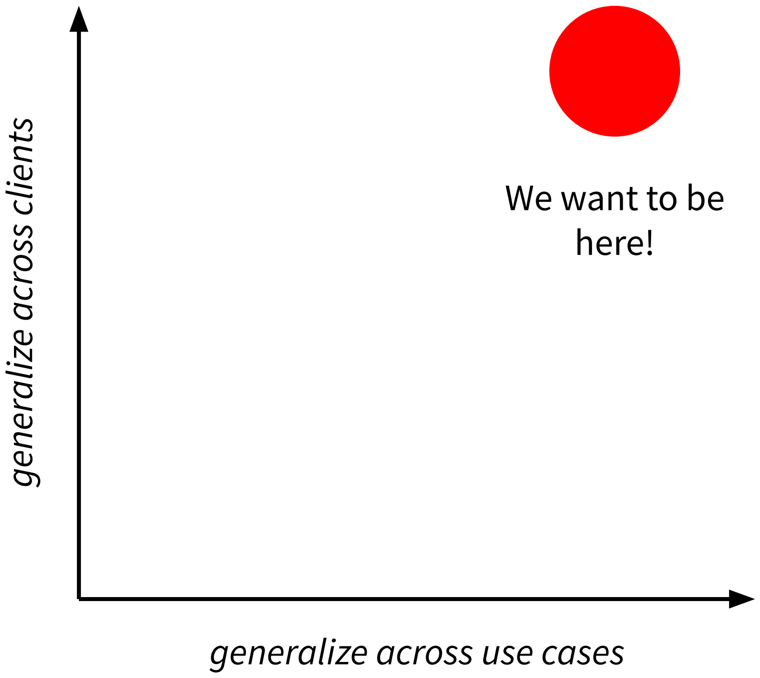


Those two spaces are *not* the same: in fact, you cannot search for visual concepts.

Content understanding is siloed

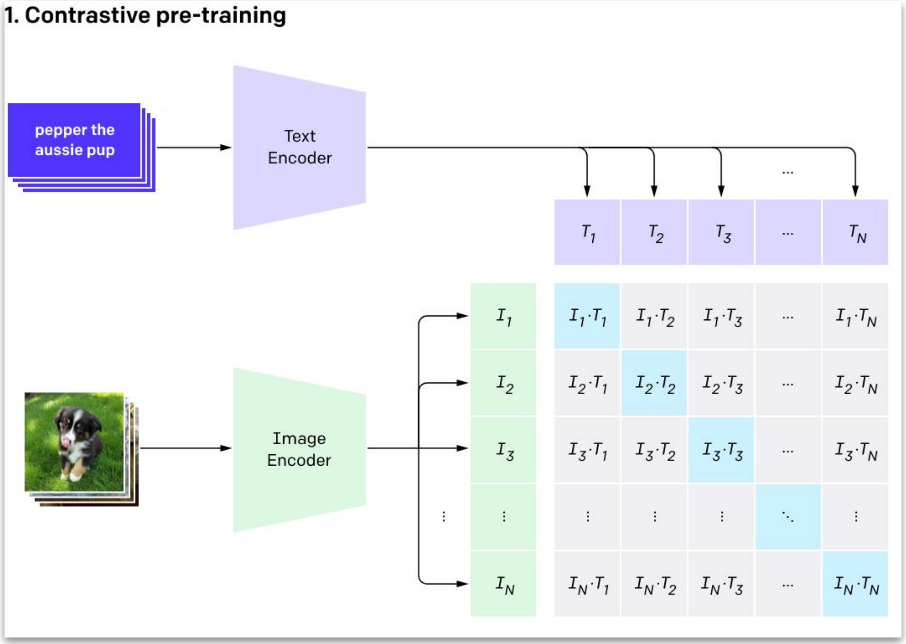
Content understanding in the age of ~~ML~~ CLIP

- The naive strategy has two “for loops”:
  - For each shop in shops
    - For each use case in shop
- Ideally, we wish to remove both
  - Can we re-use the same model across shops?
  - Can we re-use the same concepts across use cases?



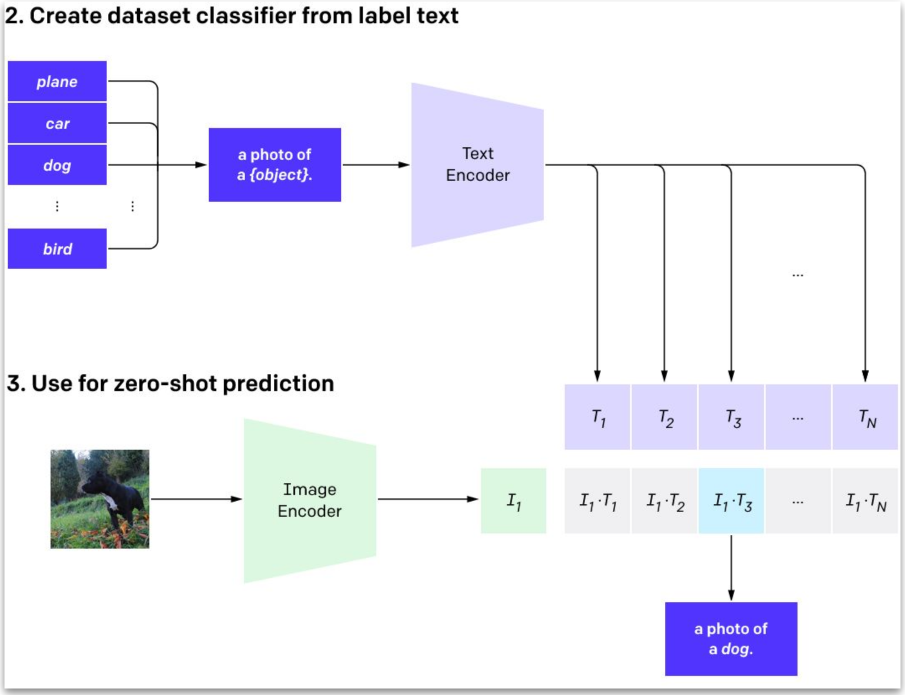
Content understanding in the age of ~~ML~~ CLIP

- Enter OpenAI CLIP:
  - "CLIP pre-trains an image encoder and a text encoder to predict which images were paired with which texts in our dataset"

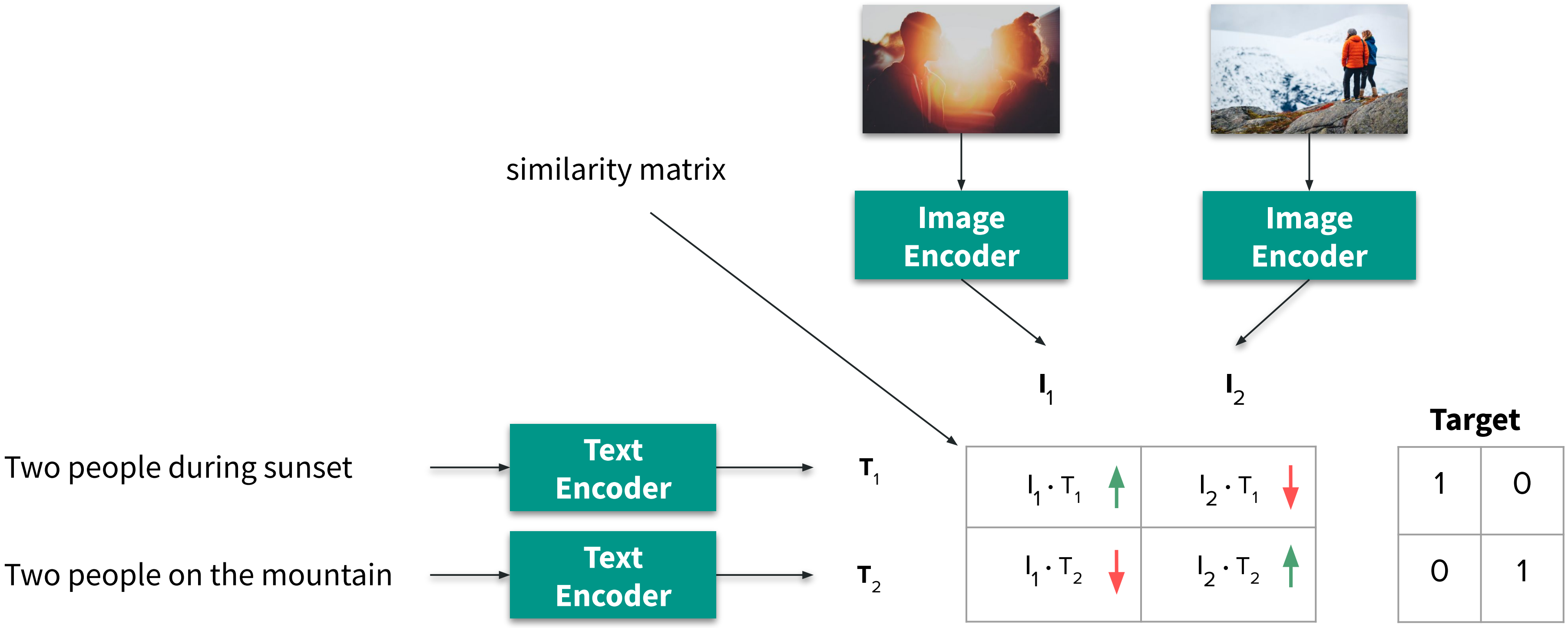


Content understanding in the age of ~~ML~~ CLIP

- Enter OpenAI CLIP:
  - “CLIP pre-trains an image encoder and a text encoder to predict which images were paired with which texts in our dataset”
- CLIP learns to place together in **one** space images and strings that are related, and far apart those that are not.



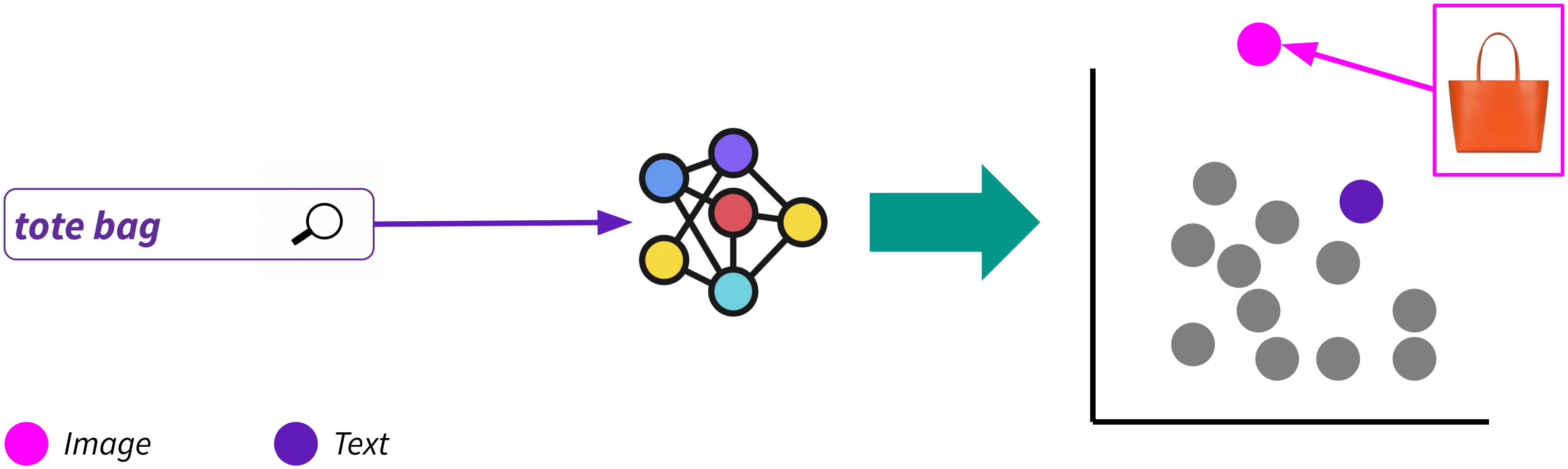
Content understanding in the age of ~~ML~~ CLIP*



* Credits to Fede at Stanford!

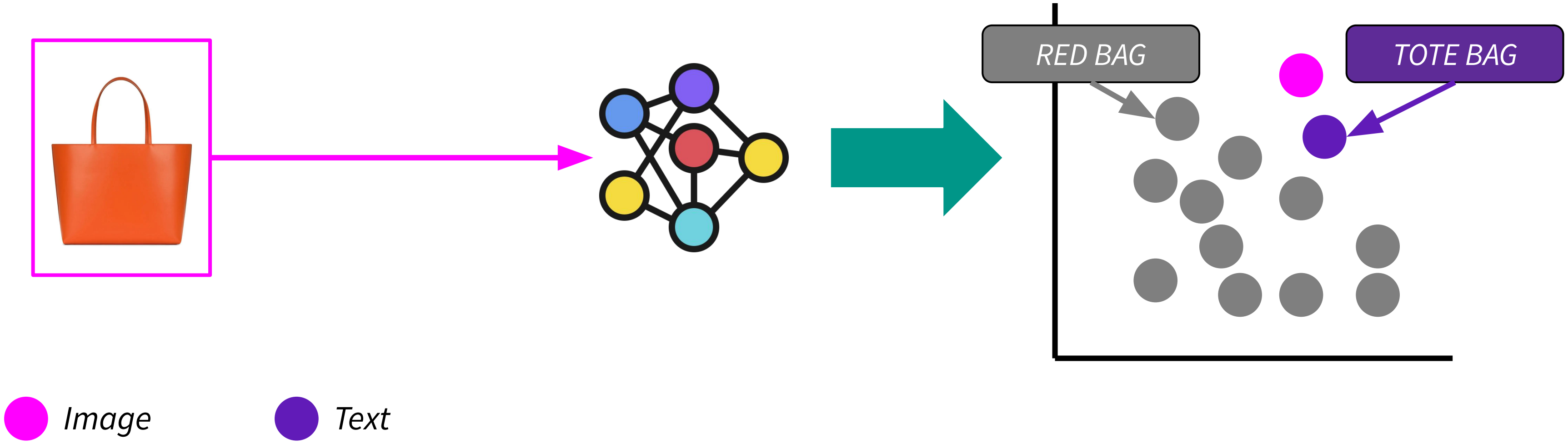
Content understanding in the age of ~~ML~~ CLIP

- After training, you can use **text** to find **images**: you just have to look into the space and find images close to your query!



Content understanding in the age of ~~ML~~ CLIP

- After training, you can use **images** to rank **captions**: you just have to look into the space and find captions close to your image!



FashionCLIP

Teaching Fashion to CLIP

- In an international collaboration across 4 time-zones, 5 countries and 5 institutions (Bocconi University, Stanford University, Coveo, Farfetch, Telepathy Labs), we trained FashionCLIP, a “fashion-aware” model obtained by **fine-tuning CLIP** on 800k pairs of fashion products (images + text).

FashionCLIP: Connecting Language and Images for Product Representations

Patrick John Chia* Coveo Montreal, Canada pchia@coveo.com

Giuseppe Attanasio Bocconi University Milan, Italy

Federico Bianchi Bocconi University Milan, Italy

Silvia Terragni Telepathy Labs, Zurich

Ana Rita Magalhães Farfetch Porto, Portugal

Diogo Goncalves Farfetch Porto, Portugal

Ciro Greco Coveo Labs New York, United States

Jacopo Tagliabue Coveo Labs New York, United States

Abstract

The steady rise of online shopping goes hand in hand with the development of increasingly complex ML and NLP models. While most use cases are cast as specialized supervised learning problems, we argue that practitioners would greatly benefit from more transferable representations of products. In *this* work, we build on recent developments in contrastive

classification (Chen et al., 2021) and many other use cases (Tsagkias et al., 2020).

As a standard practice before the rise of capable zero-shot alternatives, e-commerce models are typically trained over task-specific datasets, directly optimizing for individual metrics: for example, a product classification model might be trained on $< product\ description, category >$ pairs derived from catalog data (Gupta et al., 2016). In-

Teaching Fashion to CLIP

- **Fact 1**: FashionCLIP beats CLIP in several fashion-related benchmarks (held-out sets and *other* out-of-distribution fashion datasets).
- **Fact 2**: training FashionCLIP is relatively cheap.

Model	Dataset	HITS@5
F-CLIP	TEST	**0.61**
CLIP		0.22
F-CLIP	HOUT-C	**0.57**
CLIP		0.28
F-CLIP	HOUT-B	**0.55**
CLIP		0.27

Table 2: Comparing FashionCLIP (F-CLIP) vs CLIP on the multi-modal retrieval task.

LR	Loss	Time(m)	USD	$kgCO_2eq$
1e-4	16.0	618	31$	0.77
1e-5	1.73	617	31$	0.77
1e-6	**2.83**	**621**	**31$**	**0.78**

Table 1: Comparing training time, performance, costs and carbon emission on variants of the FashionCLIP architecture on the *Farfetch* catalog. Cost is calculated with the AWS pricing for a *p3.2xlarge*; estimations were conducted using the Machine Learning Impact calculator from Lacoste et al. (2019). Model used for testing in **bold**.

Model	Dataset	F1
F-CLIP	TEST	**0.39**
CLIP		0.31
F-CLIP	KAGL	**0.67**
CLIP		0.63
F-CLIP	F-MNIST	**0.71**
CLIP		0.66
F-CLIP	DEEP	**0.47**
CLIP		0.45

Table 3: Comparing the performance of FashionCLIP (F-CLIP) on product classification task over several datasets (**F1** is *weighted macro F1*).

FashionCLIP for product search

FashionCLIP

Text-only

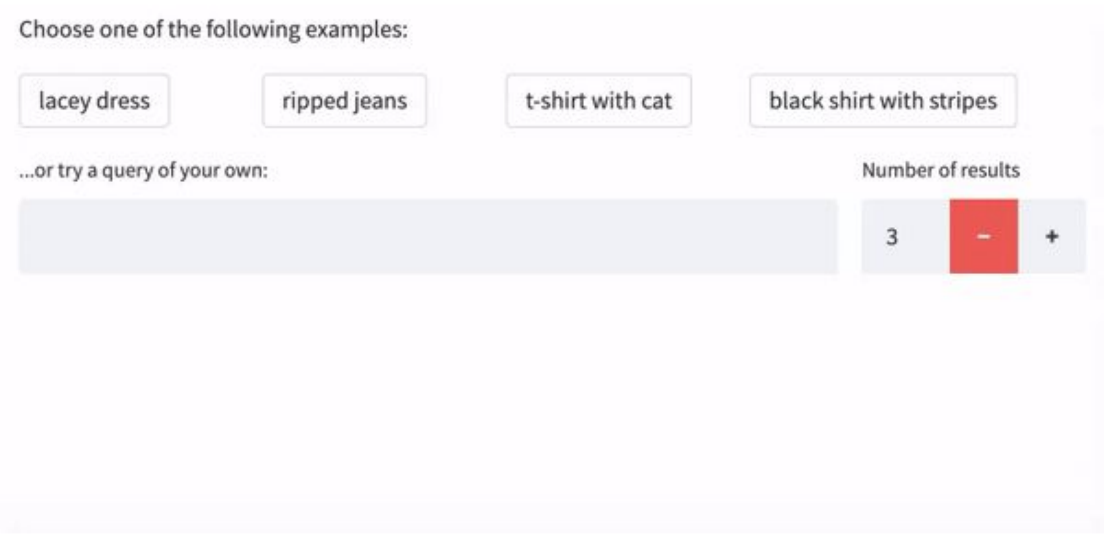
t-shirt with cat

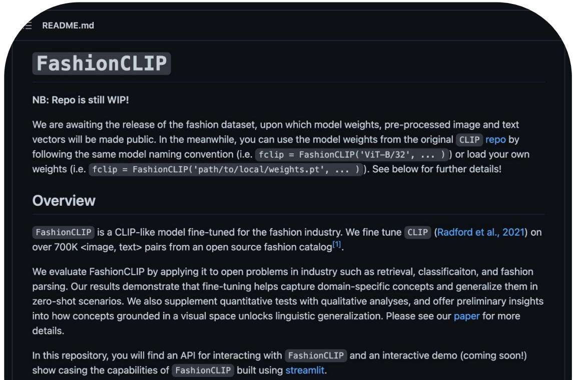
light red dress

dark red dress

FashionCLIP for product search

- FashionCLIP is available as an open source project, with a built-in app for visualizing results.




README.md

FashionCLIP

NB: Repo is still WIP!

We are awaiting the release of the fashion dataset, upon which model weights, pre-processed image and text vectors will be made public. In the meanwhile, you can use the model weights from the original `CLIP` repo by following the same model naming convention (i.e. `fclip = FashionCLIP('ViT-B/32', ... )`) or load your own weights (i.e. `fclip = FashionCLIP('path/to/local/weights.pt', ... )`). See below for further details!

Overview

`FashionCLIP` is a CLIP-like model fine-tuned for the fashion industry. We fine tune `CLIP` (Radford et al., 2021) on over 700K <image, text> pairs from an open source fashion catalog[1].

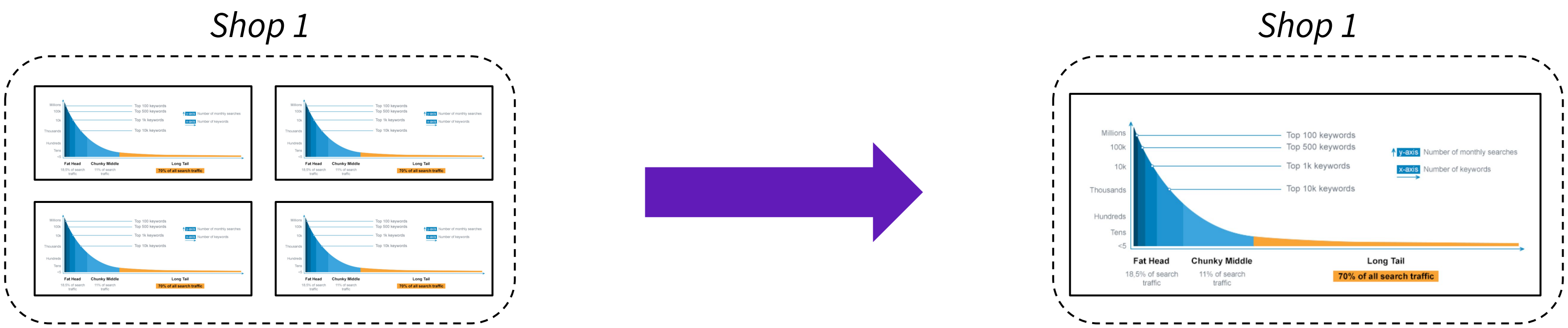
We evaluate FashionCLIP by applying it to open problems in industry such as retrieval, classificaiton, and fashion parsing. Our results demonstrate that fine-tuning helps capture domain-specific concepts and generalize them in zero-shot scenarios. We also supplement quantitative tests with qualitative analyses, and offer preliminary insights into how concepts grounded in a visual space unlocks linguistic generalization. Please see our paper for more details.

In this repository, you will find an API for interacting with `FashionCLIP` and an interactive demo (coming soon!) show casing the capabilities of `FashionCLIP` built using streamlit.

One **model** to rule them all

FashionCLIP for B2B players

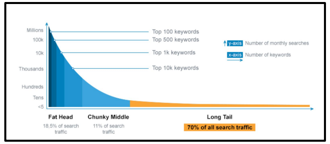
- **Inner For Loop**: since FashionCLIP is not trained on retrieval or classification, it can be used across use-cases without extra-work: for example, we show how to classify a product in **styles** (streetwear, elegant, etc.) to scale-up manual merchandisers' work.



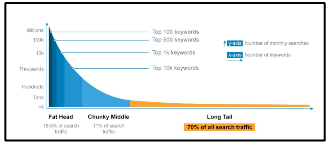
FashionCLIP for B2B players

- **Outer For Loop**: since FashionCLIP is not trained on Gucci (or Armani etc.), any customer that needs “fashion understanding” can re-use it to provide **out-of-the-box content understanding**.

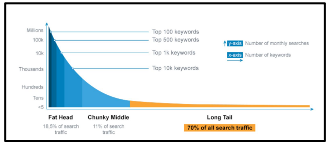
Shop 1



Shop 3

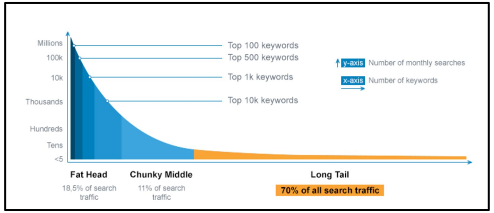


Shop n



…

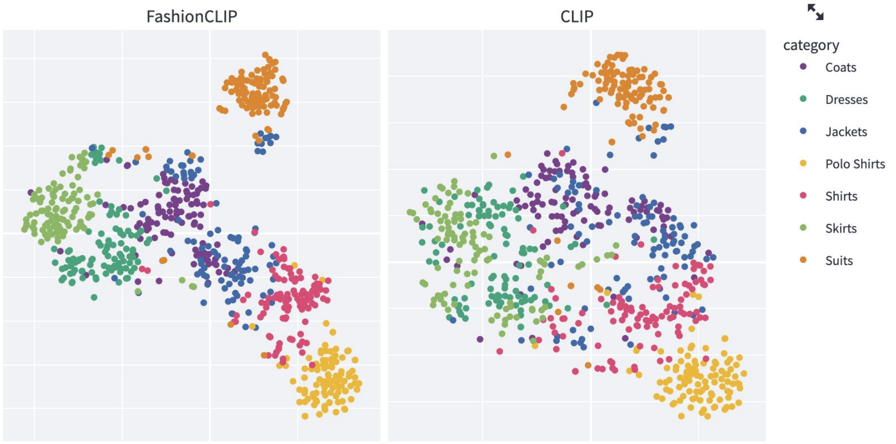
“Fashion”



Slip inside the eye of CLIP mind

Generalization matters

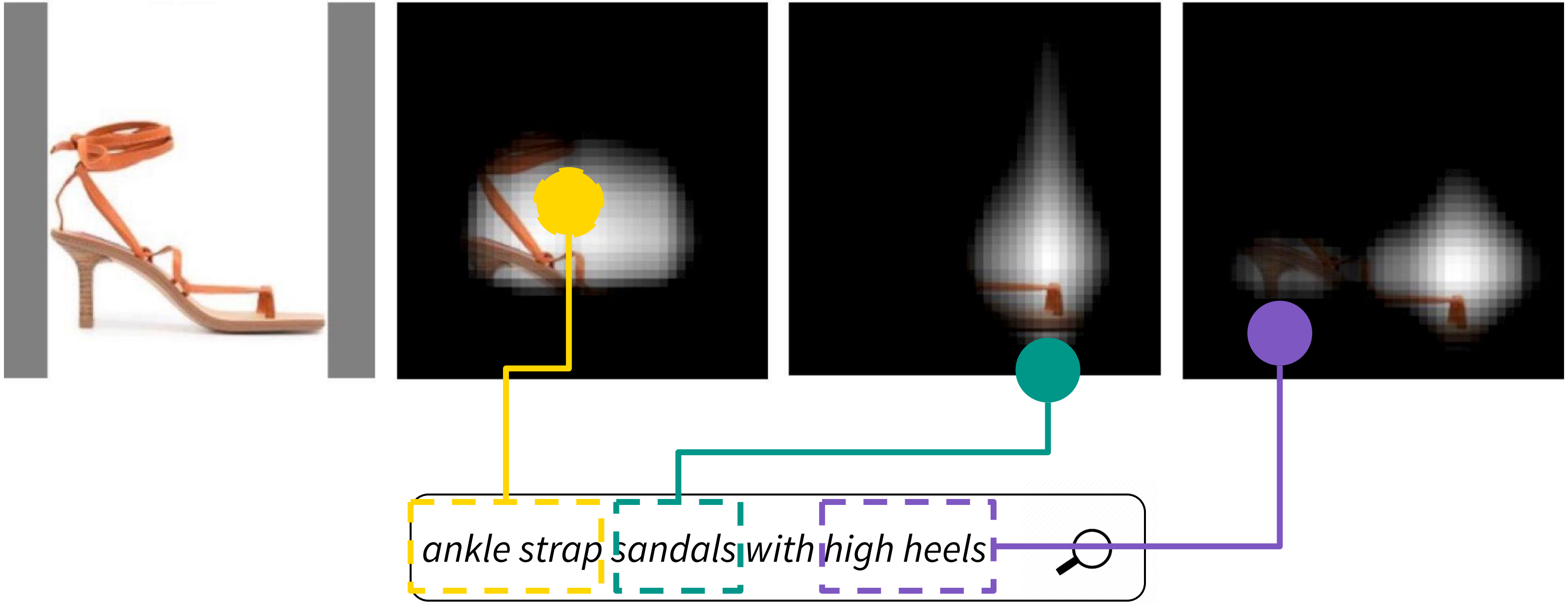
- Our entire argument rests on one **major premise,** that is, FashionCLIP does not understand *this or that dataset*, but **product content across two modalities**.
- Opening the “black-box” is therefore necessary to make sure the model behaves as it should.



t-sne plot for clothing concepts, FashionCLIP vs CLIP

Grounding content

- Occlusion maps may help revealing how FashionCLIP is composing complex concepts out of simpler ones.



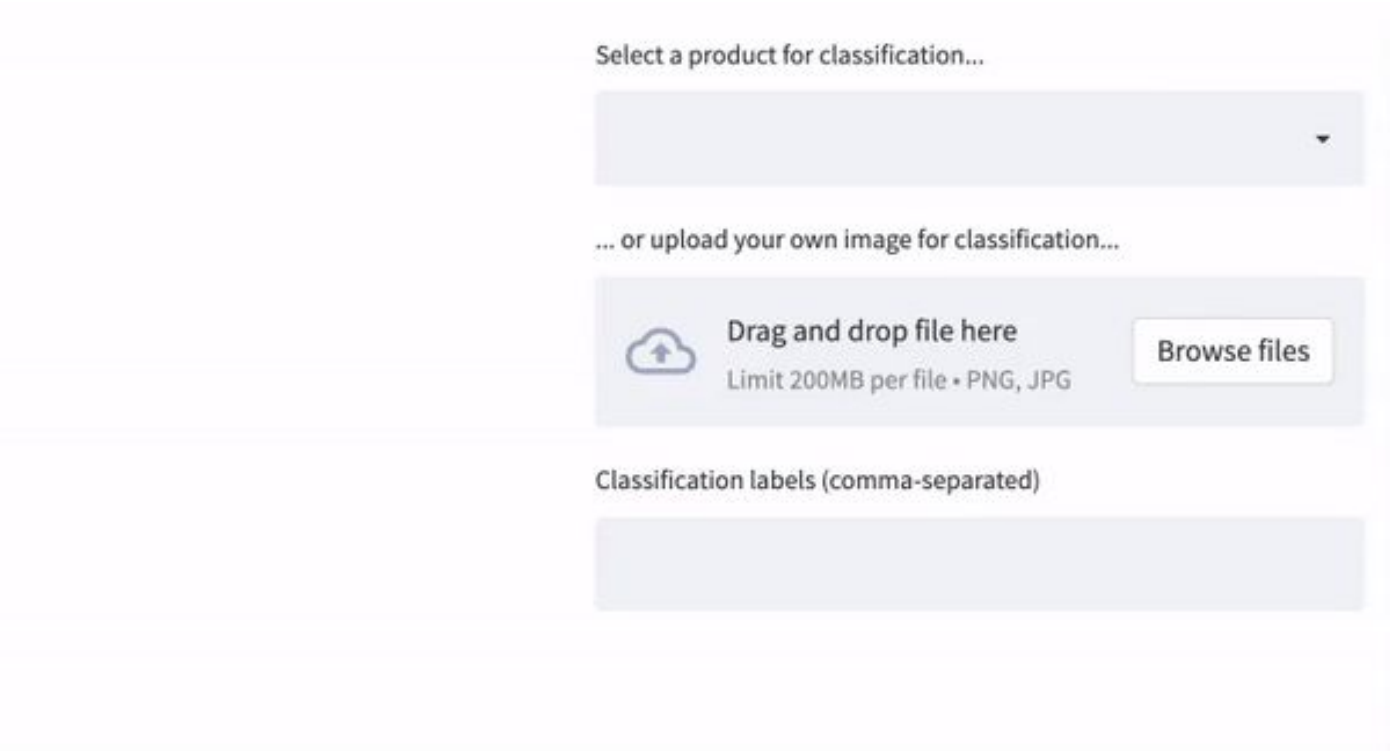
Improbable products

- Testing FashionCLIP on different datasets is a first step in generalization, but can we go **further**?
  - After all, even small kids are pretty good at coming up "on-the-fly" with new concepts!

Improbable products

- Testing FashionCLIP on different datasets is a first step in generalization, but can we go **further**?
  - After all, even small kids are pretty good at coming up “on-the-fly” with new concepts!
- We can test FashionCLIP on “improbable products”, i.e. products that are by definition not to be found in the training distribution as *they don’t exist in the real world.*

Improbable products



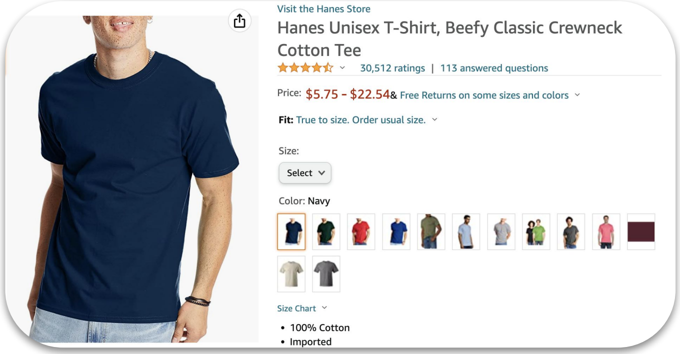
Does it come in black?

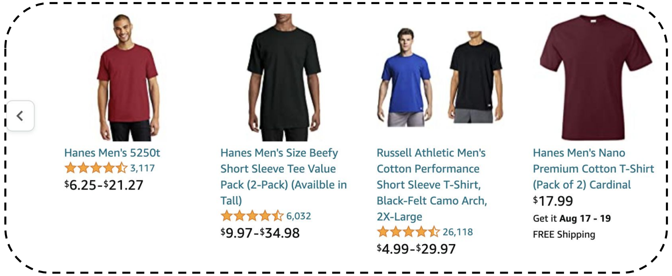
- How many times do we like a *t-shirt*, but we wish it was darker? Or a skirt, but wish it was *longer*?

A request from the World's Greatest Detective

Does it come in black?

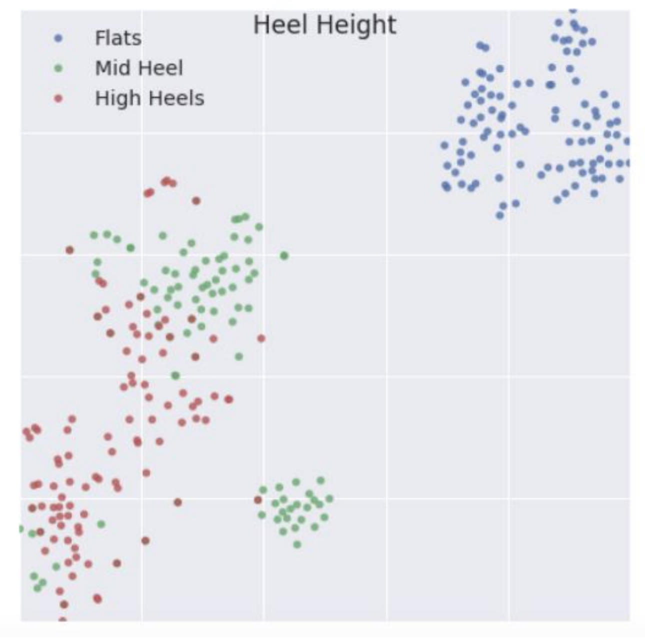
- How many times do we like a *t-shirt*, but we wish it was darker? Or a skirt, but wish it was *longer*?
- RecSys today gives you “Items like X”, but don’t allow you to move in the space along *one relevant attribute*.





Does it come in black?

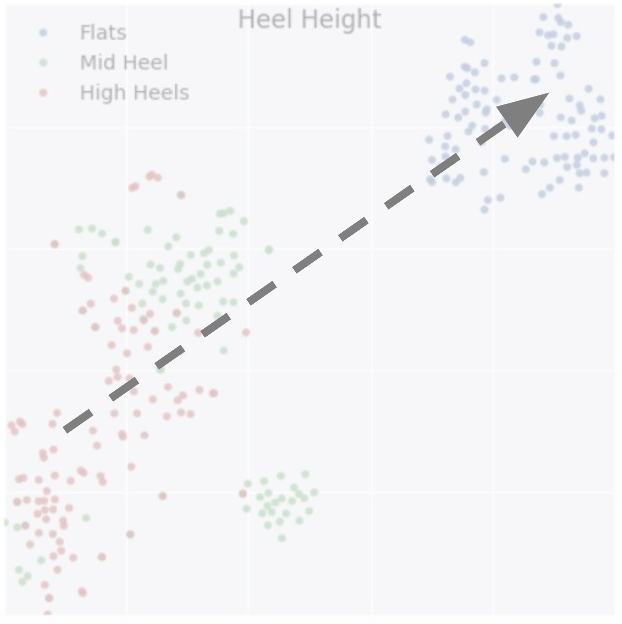
- How many times do we like a *t-shirt*, but we wish it was darker? Or a skirt, but wish it was *longer*?
- RecSys today gives you “Items like X”, but don’t allow you to move in the space along *one relevant attribute*.
- **Can we use CLIP latent multi-modal space to move in the catalog with English (“darker”, “longer”)?**



FashionCLIP latent space encodes geometric gradients

Does it come in black?

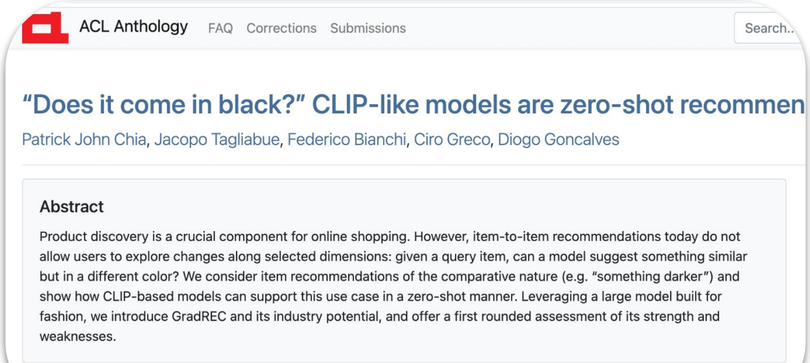
- How many times do we like a *t-shirt*, but we wish it was darker? Or a skirt, but wish it was *longer*?
- RecSys today gives you “Items like X”, but don’t allow you to move in the space along *one relevant attribute*.
- **Can we use CLIP latent multi-modal space to move in the catalog with English (“darker”, “longer”)?**



FashionCLIP latent space encodes geometric gradients

GradRecs

- To everybody's surprise (ours included!), our work on ["zero-shot" recommendations]() proved that we can disentangle FashionCLIP latent space (to some extent).

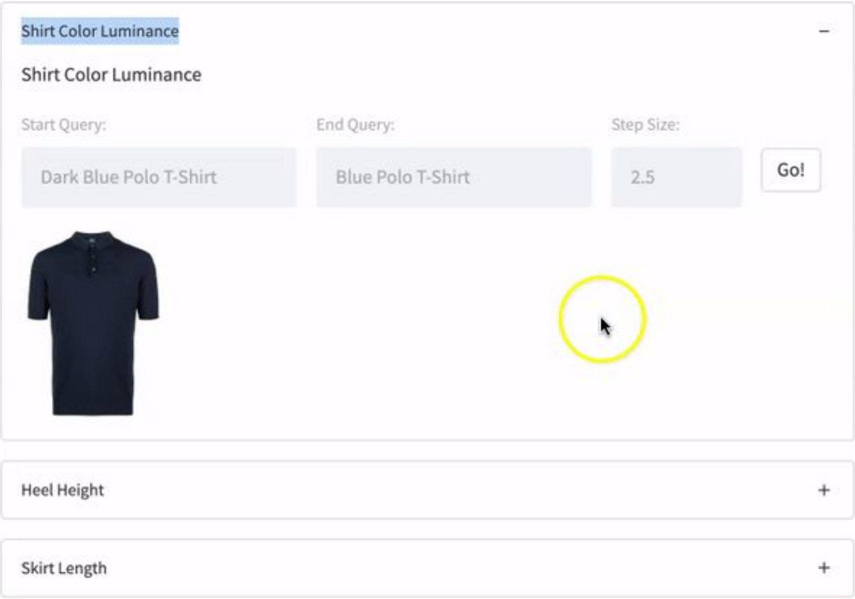

ACL Anthology FAQ Corrections Submissions Search...

"Does it come in black?" CLIP-like models are zero-shot recommen

Patrick John Chia, Jacopo Tagliabue, Federico Bianchi, Ciro Greco, Diogo Goncalves

Abstract

Product discovery is a crucial component for online shopping. However, item-to-item recommendations today do not allow users to explore changes along selected dimensions: given a query item, can a model suggest something similar but in a different color? We consider item recommendations of the comparative nature (e.g. "something darker") and show how CLIP-based models can support this use case in a zero-shot manner. Leveraging a large model built for fashion, we introduce GradREC and its industry potential, and offer a first rounded assessment of its strength and weaknesses.



What’s next?

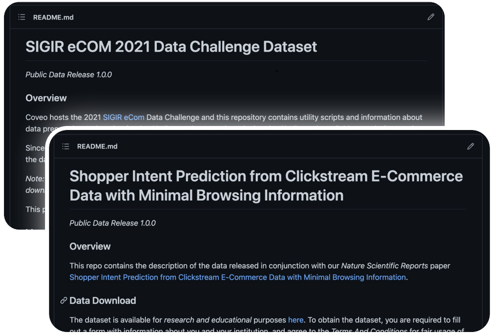
From CLIP to DALL-E

- The advent of accurate text-to-image generative models such as DALLE-2 opens up interesting possibilities:
  - data augmentation
  - synthetic data
  - testing.

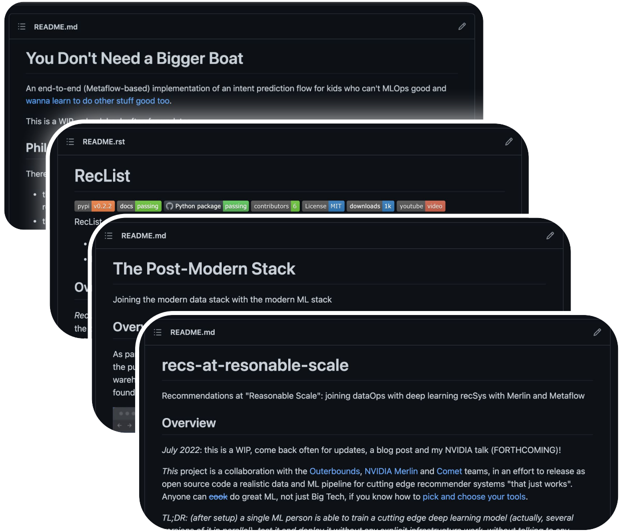
Caveat: DALLE-2 struggles with compositionality so some work needs to be done there!

It takes an open village

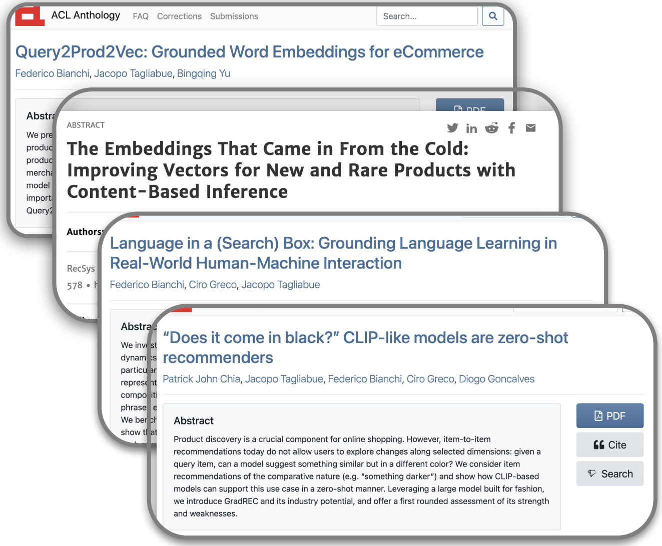
Open Data



Open Source



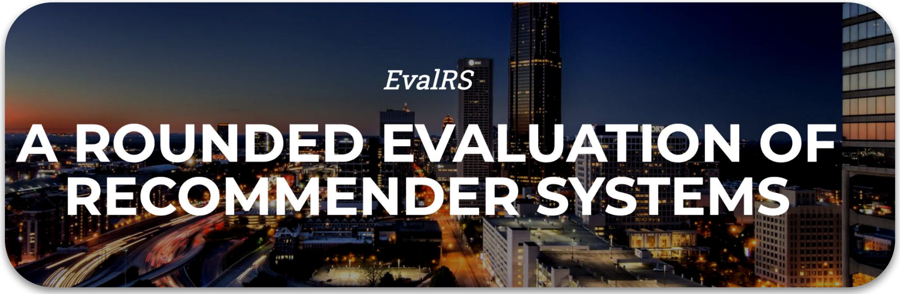
Open Science



It takes an open village

JOIN US NOW AT THE CIKM DATA CHALLENGE!

https://reclist.io/cikm2022-cup/



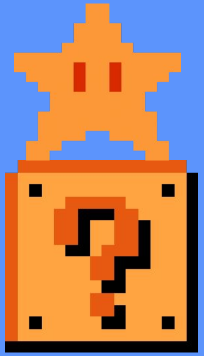
RecList

RecList is an open source library providing behavioral, "black-box" testing for recommender systems. Inspired by the pioneering work of Ribeiro et al. 2020 in NLP, we introduce a general plug-and-play procedure to scale up behavioral testing, with an easy-to-extend interface for custom use cases.

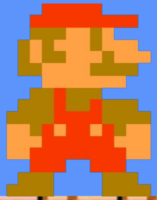
To streamline comparisons among existing models, RecList ships with popular datasets and ready-made behavioral tests: read the our TDS blog post as a gentle introduction to the main use cases, and try out our colab to get started with the code.

We are actively working towards our beta, with new

Check out / share / add a star to our open source **projects**!



Wanna work with us? Get in touch!



“We can only see a short distance ahead, but we can see **plenty** there that needs to be done.”

See you, (vector) space cowboys