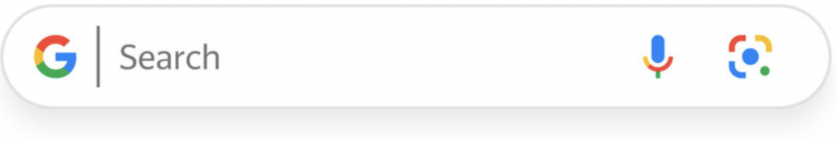
COL764 - Project

Final Project Proposal

Atharv Dabli & Gurarmaan S. Panjeta
2020CS10328 & 2020CS50426



Feat. Multi-modal Querying, Visual Feature Extraction and Word Embeddings

October 22, 2023

October 22, 2023

Objective

To build an efficient, fast and resilient query system for textual and image data.

This Domain Agnostic Search fits applications like web-shopping, searching for diagnostics evaluation, and simply aiding a directed visual search by incorporating textual evidence. We seek to experiment and tune the best architecture, and the relative weights to be given to image features, text inferred from the image, and text passed in by the user.

Key steps

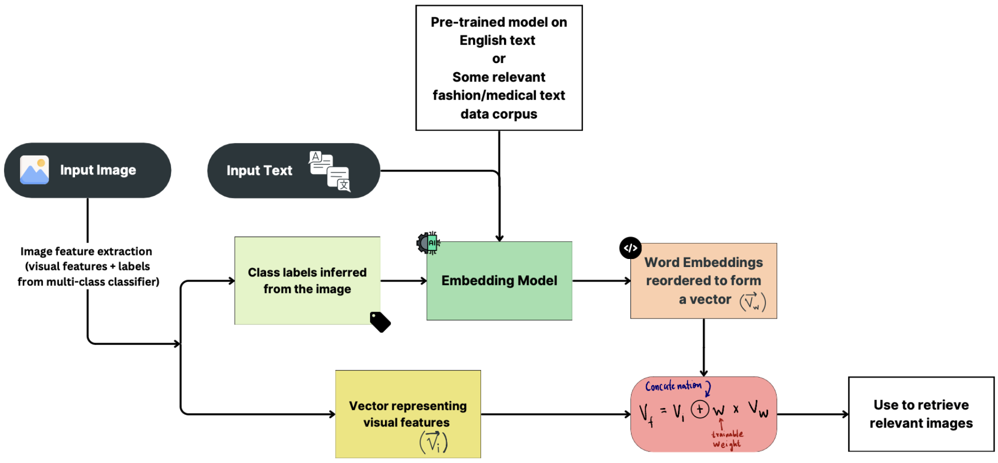
a) User feeds in Image and/or Text.

b) Vectorization of Image into feature vector.

c) Obtaining class labels from the image , combining with user's text and obtaining a suitable word embedding.

d) Augment the image fetures and text vector. Search through the data by a cosine distance metric or some cluster based optimisation.

Proposed Architecture


Pre-trained model on English text or Some relevant fashion/medical text data corpus
Input Image
Input Text
Image feature extraction (visual features + labels from multi-class classifier)
Class labels inferred from the image
Embedding Model
Word Embeddings reordered to form a vector $(\vec{V_w})$
Vector representing visual features $(\vec{V_i})$
Concatenation
$V_f = V_i \oplus w \times V_w$
trainable weight
Use to retrieve relevant images

Novelty and Experimentation

- Architectures for Image Feature Extractions - CNNs/MLPs/Inception, analysis of size of features required (256 vs 512 vs 1024 etc.).
- Using linguistic semantics incorporated as word embeddings to help with search. Figuring out a uniform-sized representation for the non-uniform length input that may be passed in.

- An experiment to perform some way of Clustering on data to reduce the number of pairwise distance measurements.
- Deciding on the weightage to be given to between the text evidence and image features - some user may not be sure about the text they're providing - or may recognise the poor relevance of the image with their information need. Present a default value that fits the data well, if not provided by user.

Insights

- In case the user provides only one of the modes of data, the system, via it's dot-product approach will still be able to perform a search by filling in 0 vectors for absent modes, which makes dot product distance ignore these modes
- Using word embeddings of the passed text and inferred class labels from the image - adds a dimension of linguistic semantics into the computed similarity scores.
- The visual features can be extracted from the penultimate layers of a well-trained network classifier. The choice of architecture is an important investigation point.
- Until now, conventional models have typically utilized distinct search processes for images and text, necessitating subsequent fusion of their separate results. This fusion and reordering step is complex and often falls short in effectively conveying relevance across both modalities. In contrast, our approach seamlessly performs a concurrent search on both modalities, treating them as a unified query, and as a result, produces seamlessly integrated outcomes.

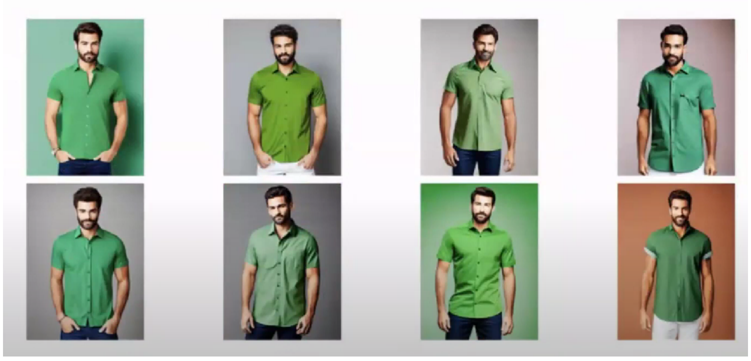
Expected Results

Our query would have two modes, image and text.

Text : Short sleeves shirt

Image :

Query Result :



The retrieved images we expect would be short sleeve shirts with a similarity of colour as given in the image. We would try to come up to such similar results.

Evation Metric

- Assessing the outcomes of similar item searches within a dataset of comparable items poses a significant challenge in terms of objective scoring.
- Furthermore, the sheer scale of the database renders manual identification of the most similar items impractical.
- To address this, we propose a manual evaluation process where we review the top 5 or top 10 images retrieved for numerous queries and manually label their relevance.
- Additionally, we may explore the development and application of similarity metrics between query images and retrieved images as part of our evaluation methodology.

Data Sets

- DeepFashion Dataset, includes 44,096 high-resolution human images, manually annotated for cloths, textures and accessories.
- Fashion200 Dataset - Includes 300k images, with annotations for all wearables.
- Stanford Online Products - Non-annotated set of images. To Test and tune the class label identifier network.