PIMMI

a command line interface to study image propagation

Nicolas Hervé, research department, Institut National de l’Audiovisuel, France

Béatrice Mazoyer, médialab, Sciences Po, France
How to study meme propagation?
What PIMMI does

- create clusters of identical images (total or partial copies) on millions of files
- deal with image transformations (crop, zoom)
- adapt to corpus characteristics (number and nature of images)
What PIMMI does not do not

- clustering of semantically similar images (e.g. separate cats from dogs)
- face recognition (e.g. find all images of Elizabeth II)
Use cases

Pimmi was designed to conduct studies on the use and re-use of images:

- Propagation of memes on social networks
- Usage of press agency photos in a press corpus
- Dissemination of fake news based on image montages
- Comparison of editorial choices between different media
- ...
What PIMMI does
How it works

Each image is represented as a set of keypoints using SIFT (Scale Invariant Feature Transform) algorithm. Local keypoints descriptors are vectors.
How it works

Local descriptors are indexed in a database (FAISS) optimized for similarity search. Different index structures are available, depending on corpus size.
How it works

PIMMI searches for matching local descriptors in the database.

For each pair of images who have matches, it filters only pairs compatible with some expected geometric transformations.
How it works

The clusters are detected on the graph of similarities between images.
How to use PIMMI

What you need:

- Python $\geq$ 3.7
- Images in .jpg or .png format saved on your computer
- A command-line shell
Demo

2023-02-03 19:08:02,095 : INFO : pimmi - index loaded ./index/demo.IVF1024,Flat.faiss
2023-02-03 19:08:02,096 : INFO : pimmi - meta loaded ./index/demo.IVF1024,Flat.meta
2023-02-03 19:08:02,096 : INFO : pimmi - index has 11/11 images with 1598 feature points
2023-02-03 19:08:02,096 : INFO : pimmi - type : IVF1024,Flat
2023-02-03 19:08:02,096 : INFO : pimmi - found 11 images in index
2023-02-03 19:08:02,096 : INFO : pimmi - total number of queries 11
2023-02-03 19:08:02,096 : INFO : pimmi - query 11 files from pack 0 -> ./index/demo.IVF1024,Flat.mining_000000.csv
2023-02-03 19:08:02,097 : INFO : pimmi - Using opencv : 4.7.0
2023-02-03 19:08:02,231 : INFO : pimmi - ~ [0 :: 0] (1000, 1, 0.1, 10, 1.6) extracting 375 sift points for (512 x 341, False) small_dataset/000001.jpg
(fosdem) bmazer@pt11206:~/Downloads/fosdem$
$ pimmi clusters demo > demo_clusters.csv
2023-02-03 19:08:23,522 : INFO : pimmi - Loading query results
2023-02-03 19:08:23,523 : INFO : pimmi - Number of vertices in the graph: 5
2023-02-03 19:08:23,523 : INFO : pimmi - Number of edges in the graph: 8
2023-02-03 19:08:23,523 : INFO : pimmi - Is the graph directed: 1
2023-02-03 19:08:23,523 : INFO : pimmi - Maximum degree in the graph: 4
2023-02-03 19:08:23,523 : INFO : pimmi - Connected components in the graph: 2
(path) $ xsv table demo_clusters.csv
<table>
<thead>
<tr>
<th>path</th>
<th>image_id</th>
<th>nb_points</th>
<th>degree</th>
<th>cluster_id</th>
<th>quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>000002.jpg</td>
<td>7</td>
<td>66</td>
<td>4</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>000005.jpg</td>
<td>1</td>
<td>66</td>
<td>4</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>000007.jpg</td>
<td>2</td>
<td>66</td>
<td>4</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>000008.jpg</td>
<td>10</td>
<td>133</td>
<td>2</td>
<td>1</td>
<td>0.16296296296296298</td>
</tr>
<tr>
<td>000009.jpg</td>
<td>5</td>
<td>139</td>
<td>2</td>
<td>1</td>
<td>0.16296296296296298</td>
</tr>
</tbody>
</table>

How long does it take?

It depends on:

- the number of indexed images
- the chosen index
- the number of cores

Query time for 3000 queries

24 cores, 256 Go RAM

Inverted file index | HNSW index

<table>
<thead>
<tr>
<th>Number of indexed images</th>
<th>Inverted file index</th>
<th>HNSW index</th>
</tr>
</thead>
<tbody>
<tr>
<td>3000</td>
<td>1m6s</td>
<td>10m30s</td>
</tr>
<tr>
<td>263,448</td>
<td>15m47s</td>
<td>1m48s</td>
</tr>
<tr>
<td>1,022,120</td>
<td>4m29s</td>
<td></td>
</tr>
<tr>
<td>4,076,954</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Future improvements

Detect parts of images - should solve text in images issues
Future improvements

Detect parts of images

Show images in their context (tweets, Instagram posts, etc.) or with additional metadata

Display the graph of image similarities
We need your use cases!

The development of PIMMI is still in progress

We need new use cases to improve the tool
Sources and references

Images

doge: https://knowyourmeme.com/memes/doge

small dataset: https://github.com/nrv/pimmi/tree/main/demo_dataset/small_dataset

copydays (INRIA):

https://lear.inrialpes.fr/~jegou/data.php#copydays

http://web.archive.org/web/20181015092553if_/http://pascal.inrialpes.fr/data/holidays/

faiss:

Sources and references

Papers
