Regional Attention Based Deep Feature for Image Retrieval

Jaeyoon Kim (김 재 윤)

Advisor: Prof. Sung-Eui Yoon
Background – Image Retrieval (Object Retrieval)

• Given a query image, try to find visually similar images from an image database.
Background – General Pipeline of Image Retrieval

Database

Query

Ranked image list

Top-1

Top-2

Top-3

…

…

Image Encoding

Similarity Search

e.g., L1, L2 and cosine distance

…

…
Background – Image Encoding (Image Embedding)

• Need to describe an image $I$ into a **single feature vector** $f_I$.

$$f_I = \begin{bmatrix} f_{I,1} \\ f_{I,2} \\ \vdots \\ f_{I,k} \end{bmatrix}$$
Problem – Common Challenges in IR

**Background:** Acted as a *distractor* while doing the aggregation

**Object clutter:** Where do we focus on?

Many buildings in this image
Problem – R-MAC, Tolias et al., 2015

• A commonly used technique for image retrieval.
• **Region-based aggregation** method with deep local features.

Convolutional feature map

Uniformly sampling in a rigid-grid manner

• **Vulnerable to backgrounds and object clutters** due to the uniform region sampling.

Tolias et al., "Particular object retrieval with integral max-pooling of CNN activations", *CoRR, abs/1511.05879*, 2015.
Problem – R-MAC, Tolias et al., 2015

• A Conflict between below claim 1 and 2 in terms of region size

1. **Vulnerability** becomes worse as small regions are sampled.

2. **Important to consider the small regions** in order to get a good result, especially in small object retrieval.

![Diagram](image)
Problem – R-MAC, Tolias et al., 2015

• Accuracy variation with different scales.
  • Verify the conflict is valid or not.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale (S)</th>
<th>Oxford 5k(mAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S=3</td>
<td>69.9</td>
</tr>
<tr>
<td></td>
<td>S=4</td>
<td><strong>70.7</strong></td>
</tr>
<tr>
<td></td>
<td>S=5</td>
<td>70.1</td>
</tr>
<tr>
<td></td>
<td>S=6</td>
<td>69.0</td>
</tr>
</tbody>
</table>

No longer increase from the scale of 4 or more due to the background noise.
Objectives

• Limitations of R-MAC are:
  – Not to use a finer scale parameter due to the background noise.
  – Not to consider varying importance among objects in object-cluttered images

• Goals are to:
  – Use our regional attention network for filtering the background noise.
  – Use our context-awareness strategy for considering varying importance.
Our Approach – Regional Attention

• Propose regional attention network to filter the background noise.

• An existing pixel-based attention

  Feature map → Attention network → Weighted feature map

• Our region-based attention

  Regional feature map
  \[ s = 1 \]
  \[ s = 2 \]
  \[ s = 3 \]

  Regional attention weights
  \[ s = 1 \]
  \[ s = 2 \]
  \[ s = 3 \]
Our Approach – Regional Attention with R-MAC

• How our regional attention collaborates with R-MAC.
Our Approach – Context-Awareness

• Typically, people see an overall context of an image and then determine whether an object is salient or not [1].

Which of objects is more important?
w/o context: ambiguity
w/ context: red one

• Consider a global context to get high-quality attention weights of local regions.

Our Approach – Regional Attention Architecture

- Use **regional** and **global** features to reflect the context-awareness
- Two linear layers and two non-linear layers

\[
\Phi(k) = \text{softplus}(W_c \pi(k) + b_c),
\pi(k) = \tanh(W_r k + b_r).
\]

\(k\): Regional feature vector, \(\Phi(k)\): Attention weight of \(k\)
Our Approach – Ablation Study of Regional Attention

• Performance variation when our methods are added.

<table>
<thead>
<tr>
<th>Method</th>
<th>Oxford5k</th>
<th>Paris6k</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline + PCA Landmark</td>
<td>70.1</td>
<td>85.4</td>
<td>0.095</td>
</tr>
<tr>
<td>+ Regional attention</td>
<td>74.9</td>
<td>86.0</td>
<td>0.115</td>
</tr>
<tr>
<td>+ Context awareness</td>
<td>76.8</td>
<td>87.5</td>
<td>0.123</td>
</tr>
</tbody>
</table>

Our approach **consistently improved** when each module was added.
Our Approach – Image Encoding Pipeline

• Goal: Extract a global feature vector $\hat{f}_I$ from an input image $I$
Our Approach – Image Encoding Pipeline

1. Extract a CNN feature map and sample regional feature maps in an R-MAC manner.
Our Approach – Image Encoding Pipeline

2.1 Produce R-MAC feature vectors with the regional feature maps

2.2 Calculate regional attention weights
Our Approach – Image Encoding Pipeline

3. Obtain a global feature vector, $\hat{f}_I$, through combining R-MAC features with regional attention weights.
Our Approach – Summary of Main Contributions

• Propose context-awareness as well as regional attention network.
• Improve R-MAC with a large gap of accuracy.
• Achieve a new state-of-the-art performance in IR.
Result & Analysis

• Experiment
  • Training dataset: ImageNet 1M – 1000 classes
  • Benchmark datasets
    • Oxford 5k: landmark (building) images in Oxford
    • Oxford 105k: Oxford 5k + 100k distractor images
    • Paris 6k: landmark (building) images in Paris
    • Paris 106k: Paris 6k + 100k distractor images
Result & Analysis

• Comparison with the state-of-the-arts
Our approach consistently set a state-of-the-art accuracy for each 4 dataset with large gaps!

Measurement unit: mAP (mean Average Precision)

<table>
<thead>
<tr>
<th>Method</th>
<th>Dim.</th>
<th>Oxford5k</th>
<th>Paris6k</th>
<th>Oxford105k</th>
<th>Paris106k</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDCF [8]</td>
<td>2048</td>
<td>69.1</td>
<td>81.7</td>
<td>65.4</td>
<td>74.3</td>
</tr>
<tr>
<td>CroW [14]</td>
<td>2048</td>
<td>68.7</td>
<td>82.8</td>
<td>62.7</td>
<td>75.1</td>
</tr>
<tr>
<td>R-MAC [27]</td>
<td>2048</td>
<td>70.1</td>
<td>85.4</td>
<td>66.9</td>
<td>80.8</td>
</tr>
<tr>
<td>CAM [13]</td>
<td>2048</td>
<td>69.9</td>
<td>84.3</td>
<td>64.3</td>
<td>77.1</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>2048</td>
<td>76.8</td>
<td>87.5</td>
<td>73.6</td>
<td>82.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query expansion (QE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDCF+QE [8]</td>
</tr>
<tr>
<td>CroW+QE [14]</td>
</tr>
<tr>
<td>R-MAC+QE [27]</td>
</tr>
<tr>
<td>CAM+QE [13]</td>
</tr>
<tr>
<td><strong>Ours+QE</strong></td>
</tr>
</tbody>
</table>

Query expansion: commonly used in IR as an existing material

Average improvement gap: 4.1 mAP
Other methods: 2~3 mAP

The larger gap with query expansion!
Result & Analysis - Qualitative Comparison

• Show one example (query) where ours outperforms R-MAC
• Note that ours surpasses R-MAC in 54 queries out of 55 queries.
Result & Analysis - Qualitative Comparison

• Qualitative results – comparison with R-MAC
  • Only single case where R-MAC outperforms ours

<table>
<thead>
<tr>
<th>More similar</th>
<th>Less similar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>R-MAC</td>
</tr>
<tr>
<td>4-8</td>
<td>5-6</td>
</tr>
<tr>
<td>6-9</td>
<td>8-5</td>
</tr>
<tr>
<td>7-9</td>
<td>11-15</td>
</tr>
<tr>
<td>12-10</td>
<td></td>
</tr>
</tbody>
</table>

Small gap!

Ranking number
Ours -> R-MAC

R-MAC -> Ours

Diagram showing the comparison between Ours and R-MAC.
Result & Analysis

• Ablation Study – Region Proposal Network (RPN)
  • RPN has been employed in various computer vision tasks.

Our approach can cooperate with another region-sampling method (RPN) as well as R-MAC.

<table>
<thead>
<tr>
<th>Method</th>
<th>Oxford5k</th>
<th>Paris6k</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPN + PCA Landmark</td>
<td>64.7</td>
<td>75.5</td>
</tr>
<tr>
<td>+ Regional attention</td>
<td>66.6</td>
<td>75.8</td>
</tr>
<tr>
<td>+ Context awareness</td>
<td>67.9</td>
<td>76.4</td>
</tr>
</tbody>
</table>

Measurement unit: mAP (mean Average Precision)
Conclusion

• Introduced regional attention network to handle the background noise and object clutter.

• Set the state-of-the-art search accuracy, especially in query expansion by:
  • Combining R-MAC with our regional attention network.
  • Utilizing a context-awareness strategy for regional attention network.

• Showed a generality of our regional attention network by cooperating with region proposal network (RPN).
Future Work

• Apply our-method-equipped RPN into other tasks of computer vision
• Consider the fine-tuning using metric learning
Publication

• J. Kim, S. Yoon, “Regional Attention Based Deep Feature for Image Retrieval”
  • Published in BMVC, 2018

• J. Kim, S. Um, D. Min, “Fast 2D Complex Gabor Filter with Kernel Decomposition”
  • Published in IEEE TIP, April, 2018
Thank you for listening!

Q&A

Acknowledgements
Advisor Sung-Eui Yoon & SGVR members
Problem – R-MAC, Tolias et al., 2015

• The **bigger the scale we choose, the smaller objects** can be detected.

\[ S = L \text{ means } s = 1, 2, ..., L \]

\[ s=1 \quad s=2 \]

\[ S = 2 \]

\[ s=1 \quad s=2 \quad s=3 \]

\[ S = 3 \]

\[ s=1 \quad s=2 \quad s=3 \quad s=4 \]

\[ S = 4 \]

We can detect this sized object.
Result & Analysis

• Ablation Study – R-MAC vs ours in terms of scale
  • Our approach works robustly on **finer scale**.

Our approach **can see more details** of an image thanks to finer scale

\[ S = 4 \]

Measurement unit: mAP(mean Average Precision)

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale (S)</th>
<th>( S=3 )</th>
<th>( S=4 )</th>
<th>( S=5 )</th>
<th>( S=6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-MAC</td>
<td>69.9</td>
<td>70.7</td>
<td>70.1</td>
<td>69.0</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>75.1</td>
<td>76.7</td>
<td><strong>76.8</strong></td>
<td>76.4</td>
<td></td>
</tr>
</tbody>
</table>