MultiGap: Multi-Pooled Inception Network with Text Augmentation for Aesthetic Prediction of Photographs

Yong-Lian Hii, John See, Magzhan Kairanbay, Lai-Kuan Wong
Main contribution

1. Deep neural network architecture called MultiGAP that exploits features from multiple inception modules pooled by global average pooling (GAP), evaluated by prediction of (10-class) categorical distribution, before performing score binarization

2. The incorporation of textual features trained simultaneously with MultiGAP using a recurrent neural network (RNN) with gated recurrent unit (GRU) layers

3. The ability to leverage the GAP block for visualization of activation maps based on the aesthetic category and rating
Related work

Handcrafted low-level features
- color, hue, saturation, light exposure, and also other heuristics driven by rule of thumbs used by professional photographers

Generic features
- SIFT, feature encoding method such as a Fisher Vector

Deep learning models
- CNN, Double Column CNN, Multi-modal CNN
Binary classification

High vs Low
Regression (Score)

6.7  5.4  7.2
Rating distribution

1: 0.2
2: 0.04
3: 0.04
4: 0.02
5: 0.25
6: 0.24
7: 0.01
8: 0.12
9: 0.07
10: 0.01
AVA
250K images

230K training

20K testing

2012, N. Murray
AVA
250K images

230K training

20K testing

\( \delta \) used to filter out from noisy images

\( \delta = \{2, 1.5, 1, 0.5, 0\} \)

threshold = 5

\( \delta = \{0\} \)
AVA
250K images

DPChallenge

AVA Comments
1.5 million comments

Zhou et al.
Proposed method

Visual Information

Textual Information

Aesthetic Prediction

- This image is amazing!
- Lacks creativity
- A new favourite
- I feel as if a deeper DOF would have worked better
- Beautiful capture!
Visual features

GoogleNet

Convolution
Pooling
Softmax
Concat/Normalize
Visual features

Inception module
GAP (Global Average Pooling) Layer
Textual Features

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Glove embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 2</td>
<td>Glove embedding</td>
</tr>
<tr>
<td>Word 3</td>
<td>Glove embedding</td>
</tr>
<tr>
<td>Word 4</td>
<td>Glove embedding</td>
</tr>
<tr>
<td>...</td>
<td>Glove embedding</td>
</tr>
<tr>
<td>Word 100</td>
<td>Glove embedding</td>
</tr>
</tbody>
</table>

300

Recurrent Neural Network
Proposed method
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCNN [2]</td>
<td>73.25</td>
</tr>
<tr>
<td>RDCNN [2]</td>
<td>74.46</td>
</tr>
<tr>
<td>Kao et al. [12]</td>
<td>74.51</td>
</tr>
<tr>
<td>AlexNet [10] – finetuned</td>
<td>75.11</td>
</tr>
<tr>
<td>DMA [5]</td>
<td>75.41</td>
</tr>
<tr>
<td>GoogLeNet [16] – finetuned</td>
<td>75.60</td>
</tr>
<tr>
<td>MultiGAP</td>
<td>75.76</td>
</tr>
<tr>
<td>SingleGAP</td>
<td>76.31</td>
</tr>
<tr>
<td>BDN [13]</td>
<td>76.80</td>
</tr>
<tr>
<td>word2vec [19]</td>
<td>78.40</td>
</tr>
<tr>
<td>1D-CNN [20]</td>
<td>79.48</td>
</tr>
<tr>
<td>Naive Bayes SVM [14]</td>
<td>80.90</td>
</tr>
<tr>
<td>RNN (1-layer GRU)</td>
<td>81.09</td>
</tr>
<tr>
<td>RNN (2-layer GRU)</td>
<td>81.79</td>
</tr>
<tr>
<td>Multimodal DBM [14]</td>
<td>78.88</td>
</tr>
<tr>
<td>SingleGAP + RNN (2-layer GRU)</td>
<td>80.54</td>
</tr>
<tr>
<td>MultiGAP + RNN (2-layer GRU)</td>
<td>82.27</td>
</tr>
</tbody>
</table>
Results
Confusion matrix
Class Activation Map (CAM)
Class Activation Maps (CAM)
Thank you for your attention

Q&A