Abstract

✓ Algorithm. Achieve food recognition by developing an Ingredient-Guided Cascaded Multi-Attention Network, which is capable of sequentially localizing multiple informative image regions with multi-scale from category-level to ingredient-level guidance in a coarse-to-fine manner.

✓ Dataset. Introduce a new dataset ISIA Food-200 with 200 food categories from the list in the Wikipedia, about 200,000 food images and 319 ingredients.

Motivation

✓ Image-level category labels only provide weak supervised information. CNNs trained with category labels can miss fine-grained food regions.

✓ Many types of food are non-rigid, and do not exhibit distinctive spatial configuration and fixed semantic patterns. It is hard to capture discriminative semantic information from food images.

✓ Ingredient attributes. Semantically meaningful ingredients, as basic units of food images, can offer one promising venue to empower a visual recognizer to arbitrary food images.

✓ Attentional regions. Diverse attentional regions over different image scales contain different level visual information.

Our Proposed Framework

Two Main Components:
- Category-supervised Attention Sub-network (CASN):
  Discover coarse-level attention regions with category-supervision
- Ingredient-supervised Attention Sub-network (IASN):
  Discover fine-grained attention regions with ingredient-supervision

CASN

- A category-supervised STN is utilized: one Spatial Transformer Layer is added into one CNN network.
- One LSTM is introduced to combine with the following LSTMs to construct stacked LSTMs for sequential dependency modeling of localized regions.

\[
\begin{align*}
    f_1 &= ST(f_1, M_0) \\
    x_1 &= \text{relu}(W_{f1}f_1 + b_x) \\
    h_1 &= \text{LSTM}(x_1) \\
    z_1 &= \text{relu}(W_{hz}h_1 + b_z) \\
    s_1 &= W_{zs}z_1 + b_s \\
    M_1 &= W_{zm}z_1 + b_m
\end{align*}
\]

IASN

- For each sub-network in IASN, it takes localized coarse region \( f_k \) as the reference and used updated parameters \( M_{k-1} \) to discover fine-grained attentional regions.

\[
\begin{align*}
    f_k &= ST(f_k, M_{k-1}) \\
    x_k &= \text{relu}(W_{f1}f_k + b_x) \\
    h_k &= \text{LSTM}(x_k) \\
    z_k &= \text{relu}(W_{hz}h_k + b_z) \\
    s_k &= W_{zs}z_k + b_s \\
    M_k &= W_{zm}z_k + b_m
\end{align*}
\]

Multi-scale Joint Representation

- Extract three types of features from the full image, coarse region and fine-grained regions and concatenate them as the final feature representation.

ISIA Food-200

<table>
<thead>
<tr>
<th>#Dataset</th>
<th>#Classes</th>
<th>#Images</th>
<th>#Ingredients</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETH Food-101</td>
<td>101</td>
<td>100,100</td>
<td>174</td>
</tr>
<tr>
<td>VireoFood-172</td>
<td>172</td>
<td>110,241</td>
<td>353</td>
</tr>
<tr>
<td>ISIA Food-200</td>
<td>200</td>
<td>197,323</td>
<td>319</td>
</tr>
</tbody>
</table>

Experiments

Comparison of our model and state-of-the-art methods on ETH Food-101, VireoFood-172, ISIA Food-200 (%).

<table>
<thead>
<tr>
<th>Method</th>
<th>ETH Food-101</th>
<th>ISIA Food-200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Top-1</td>
<td>Top-5</td>
</tr>
<tr>
<td>AlexNet</td>
<td>64.91</td>
<td>85.32</td>
</tr>
<tr>
<td>VGG-16</td>
<td>80.41</td>
<td>94.59</td>
</tr>
<tr>
<td>DenseNet-161</td>
<td>86.93</td>
<td>97.17</td>
</tr>
<tr>
<td>MultiTaskDCNN (VGG-16)</td>
<td>82.06</td>
<td>95.88</td>
</tr>
<tr>
<td>IG-CMAN (DenseNet-161)</td>
<td>90.63</td>
<td>98.4</td>
</tr>
</tbody>
</table>

Future Works

- We should build a large-scale ImageNet-level food dataset for providing critical training and benchmark data for food recognition algorithms.
- We should promote food computing in the multimedia community for its multifarious applications and services.