Picture-to-Amount (PITA): Predicting Relative Ingredient Amounts from Food Images

Jiatong Li\textsuperscript{1}, Fangda Han\textsuperscript{1}, Ricardo Guerrero\textsuperscript{2}, Vladimir Pavlovic\textsuperscript{1,2}

\textsuperscript{1} Rutgers University, NJ, USA, \textsuperscript{2} Samsung AI Center, Cambridge, UK

Problem Definition

- PITA deep learning framework to solve the problem and improve previous baselines.
- Ingredient substitution groups are constructed for the evaluation metric and loss function.
- Even in the presence of challenging test examples, the methods are still able to yield robust qualitative results.

Method

The three parts are trained sequentially.

- $L_{ret}$: triplet hard mining loss.
- $L_{ ingr}$: positive sample weighted binary cross entropy loss.
- $L_{am}$: Wasserstein distance using the ingredient distance matrix $M$.

Dataset

1362 ingredients canonically constructed from Recipe1M (8k recipes annotated with amounts and with ≥1 images)

- Frequency of ingredients (log scale)
- 6342 approved
- 3142 pairs

- 172 connected components (substitution groups)
- Ingredient Distance Matrix

High dimension and long-tail distribution call for the need to reduce the number of ingredients. Select ingredient pairs according to cosine similarity in Word2vec embedding space. Human annotators then approve or reject the selected pairs.

Contribution

- PITA deep learning framework to solve the problem and improve previous baselines.
- Ingredient substitution groups are constructed for the evaluation metric and loss function.
- Even in the presence of challenging test examples, the methods are still able to yield robust qualitative results.

Results

Demo at foodai.cs.rutgers.edu

Metrics:

Example:

- Predicted: pasta sauce (400 g), cheese (400 g), pasta (100 g), onion (100 g)
- GT: pasta sauce (440 g), colby cheese (400 g), penne (150 g), black pepper (10 g)

- #common ingredients=3.
- CVG: #common/#GT= 3/4=0.75.
- IOU: #common/(#GT+#pred-#common) = 3/(4+4-3)=0.6.
- EMD: Earth mover’s distance between ground truth and predicted amounts with the ingredient distance matrix. 40d(pasta sauce,onion)+50d(penne,onion)+10d(black pepper,onion) in this example.

<table>
<thead>
<tr>
<th>Method</th>
<th>CVG</th>
<th>IOU</th>
<th>EMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieval</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ATTEN</td>
<td>0.47</td>
<td>0.32</td>
<td>205.19</td>
</tr>
<tr>
<td>ACME</td>
<td>0.48</td>
<td>0.33</td>
<td>199.87</td>
</tr>
<tr>
<td>IE+AP</td>
<td>0.45</td>
<td>0.25</td>
<td>193.33</td>
</tr>
<tr>
<td>IE+RE+AP(Wass)</td>
<td>0.26</td>
<td>0.13</td>
<td>142.18</td>
</tr>
<tr>
<td>IE+RE+AP(CE)</td>
<td>0.50</td>
<td>0.28</td>
<td>145.30</td>
</tr>
<tr>
<td>IE+ID+AP</td>
<td>0.46</td>
<td>0.30</td>
<td>220.18</td>
</tr>
<tr>
<td>IE+RE+ID+AP(CE)</td>
<td>0.63</td>
<td>0.42</td>
<td>154.35</td>
</tr>
<tr>
<td>IE+RE+ID+AP(Wass)</td>
<td>0.63</td>
<td>0.42</td>
<td>147.29</td>
</tr>
</tbody>
</table>

Sample Results and Optimal Ingredient Flows:

- Nodes:
  - Ground truth: represented with different colors. Areas represent amounts.
  - Predicted ingredients: use sectional color representation.
  - The color sections depict predicted ingredients and their relative amounts that are transported to the ground truth ingredient when calculating the optimal earth mover’s distance.

- Edges:
  - Olive: exact ingredient match.
  - Red: ingredients from different groups. The darker the color, the larger the distance.

Cane Sauce
(For Dippin’ Chicken)