ACM MM 27th
Nice, France
21 - 25 October 2019

Ingredient-Guided Cascaded Multi-Attention Network for Food Recognition

Weiqing Min[1,2], Linhu Liu[1,2], Zhengdong Luo[1,2], Shuqiang Jiang[1,2]
[1]Intelligent Information Processing, Institute of Computing Technology, CAS, China
[2]University of Chinese Academy of Sciences, China
minweiqing@ict.ac.cn

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.

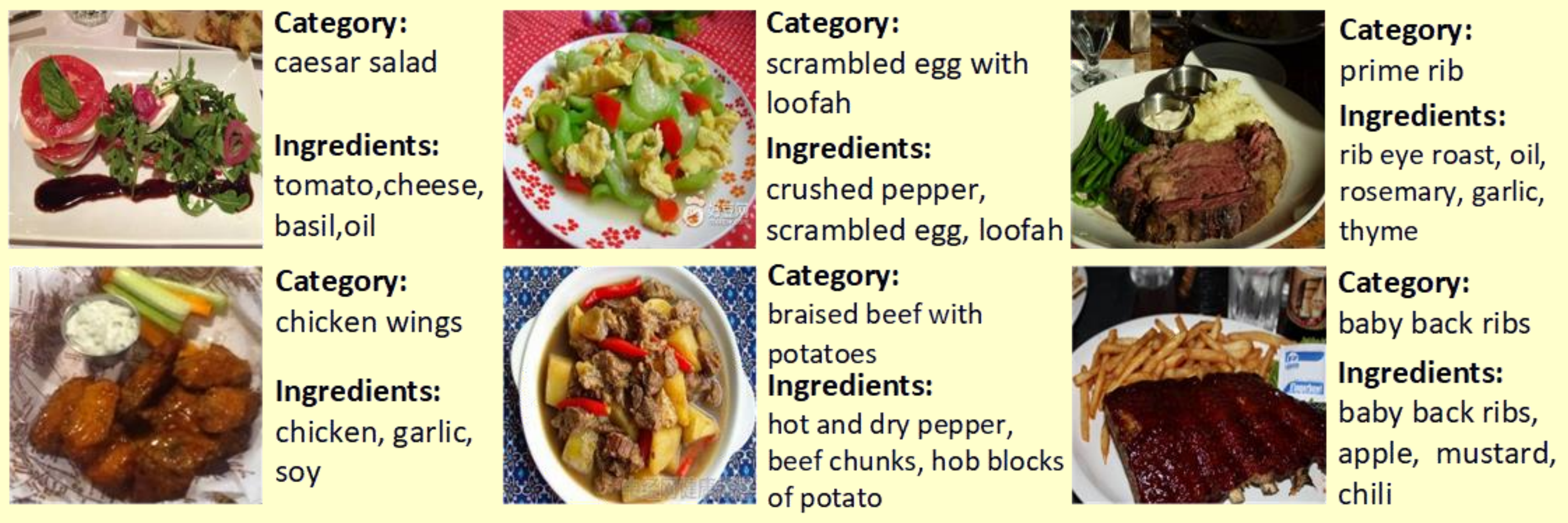

Figure.1 Some food samples with rich ingredients

- ✓ **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

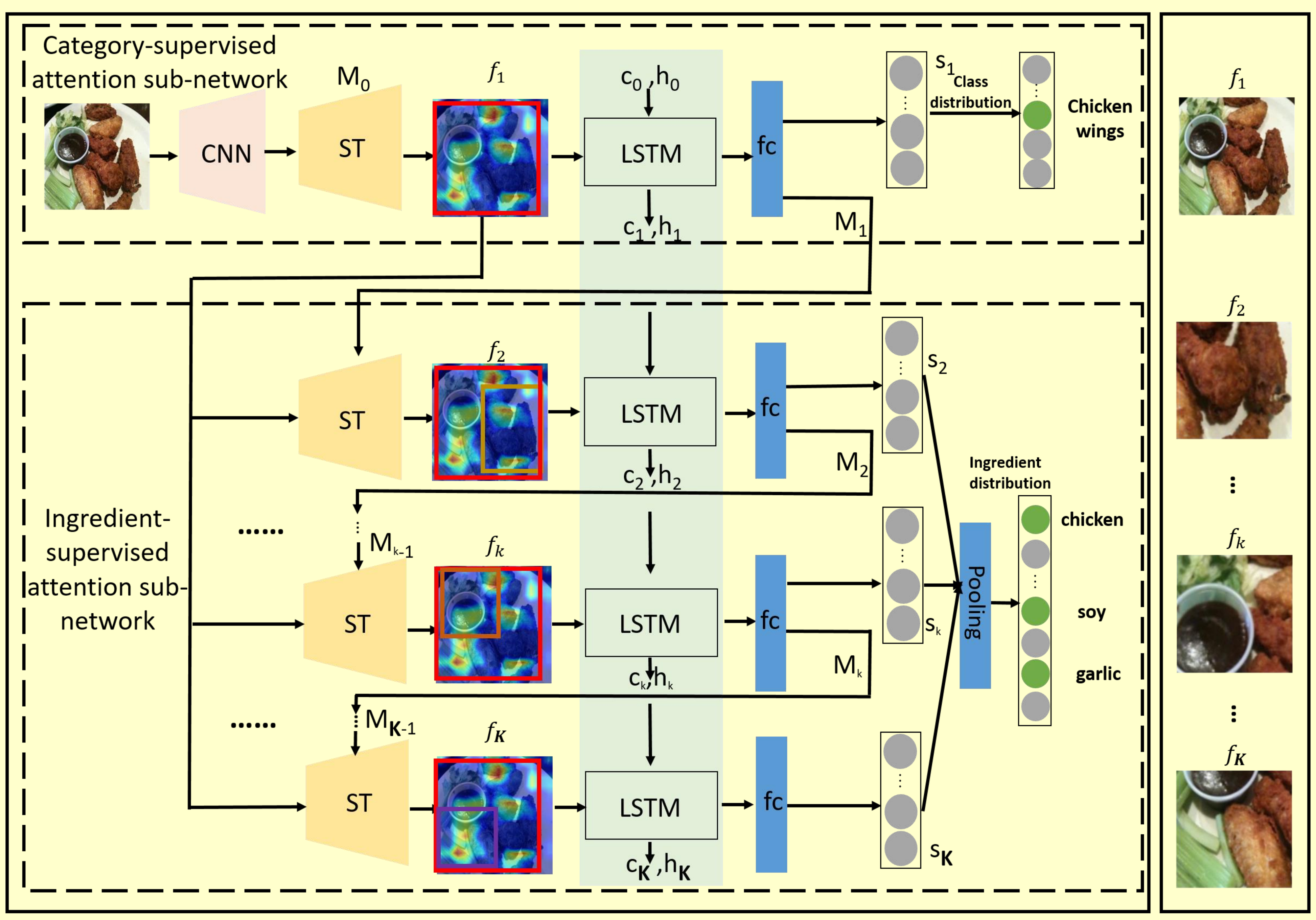

Figure 2: Overview of proposed framework for food recognition

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.

$$f_1 = \mathrm{ST}(f_1, M_0) \quad x_1 = \mathrm{relu}(W_{fx} f_1 + b_x) \quad h_1 = LSTM(x_1)$$
$$z_1 = \mathrm{relu}(W_{hz} h_1 + b_z) \quad s_1 = W_{zs} z_1 + b_s \quad M_1 = W_{zm} z_1 + b_m$$

IASN

- ➢ For each sub-network in IASN, it takes localized coarse region f_1 as the reference and used updated parameters M_{k-1} to discover fine-grained attentional regions.

$$f_k = \mathrm{ST}(f_1, M_{k-1}) \quad x_k = \mathrm{relu}(W_{fx} f_k + b_x) \quad h_k = LSTM(x_k)$$
$$z_k = \mathrm{relu}(W_{hz} h_k + b_z) \quad s_k = W_{zs} z_k + b_s \quad M_k = W_{zm} z_k + b_m$$

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

#Dataset	#Classes	#Images	#Ingredients
ETH Food-101	101	101,000	174
VireoFood-172	172	110,241	353
ISIA Food-200	**200**	**197,323**	**319**

Figure 3: Some food samples from this dataset.
The dataset is available via Github

Experiments

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

ETH Food-101

Method	Top-1	Top-5
AlexNet-CNN	56.4	-
DCNN-FOOD	70.41	-
DeepFood	77.4	93.7
FCAN	86.5	-
CurriculumNet	87.3	-
Inception V3	88.28	96.88
ResNet-200	88.38	97.85
DenseNet-161	86.94	97.03
WRN	88.72	97.92
WISeR	90.27	98.71
IG-CMAN(DenseNet-161)	90.37	98.42

VireoFood-172

Method	Top-1	Top-5
AlexNet	64.91	85.32
VGG-16	80.41	94.59
DenseNet-161	86.93	97.17
MultiTaskDCNN (VGG-16)	82.06	95.88
MultiTaskDCNN (DenseNet-161)	87.21	97.29
IG-CMAN(DenseNet-161)	90.63	98.4

ISIA Food-200

Method	Top-1	Top-5
AlexNet	49.34	79.3
VGG-16	59.05	86.53
ResNet-152	61.07	87.87
DenseNet-161	62.62	88.28
IG-CMAN(DenseNet-161)	67.47	91.75

Top-1 Accuracy : State-of-the-art-performance in three datasets

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.