Supplementary Material

Method

Since the total number of keywords/tags produced by the CNN is large (over 11,000), it is not computationally efficient to use all of them to construct a tag dictionary for a specific study due to the extremely low occurrence frequency of some tags. Therefore, we first built a tag dictionary using a large number of images acquired by eButton from free-living individuals. Then we computed a histogram of tags (Fig. A1) and removed those with low occurrence probabilities (e.g., p<0.2%) to decrease the computational load.

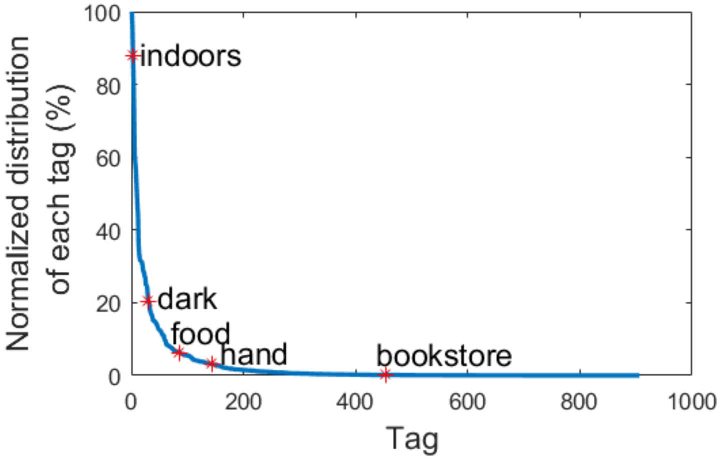

Fig.A1. An example of the histogram of the generated tags representing the occurrence frequency of each tag.

Next, we defined a tag dictionary ***D*** = {t_i, i=1,2,⋯,n}, where t_i represented the *ith* tag and n was the total number of tags in the dictionary. After ***D*** was defined, each image can be represented as a binary vector $\boldsymbol{x} = [x_1, \cdots, x_i, \cdots, x_n]$, as:

$$x_i = \begin{cases} 1, & \text{if } t_i \text{ is one of the image's tags} \\ 0, & \quad otherwise. \end{cases} \tag{A1}$$

Then we defined another binary vector $\boldsymbol{y} = [y_1, \ldots, y_i, \ldots, y_n]$ to represent the relatedness of each tag with food objects semantically:

$$y_i = \begin{cases} 1, & \text{if } t_i \text{ is related to food object(s)} \\ 0, & \quad otherwise, \end{cases} \tag{A2}$$

where the assignment as 1 or 0 was carried out by a two-step training process based on statistics from known training images.

In the first step, we calculated the probability of each tag t_i belonging to a food image by

$$p_i = \frac{N(t_i)_f}{N(t_i)_f + N(t_i)_n} \tag{A3}$$

where $N(t_i)_f$ and $N(t_i)_n$ denotes the number of occurrence of tag t_i in food and non-food images in the training set, respectively. In the extreme cases, $p_i = 1$ means that tag t_i only occurs in food images while $p_i = 0$ indicates that t_i only occurs in non-food images. Then, we obtained $\boldsymbol{y}$ by setting $y_i = 1$ if $p_i \geq$ P, where P is an empirical threshold, say, P=0.7. An example illustrating the calculation of p_i for a dictionary containing 20 tags is shown in Fig. A2. In this example, $\boldsymbol{y}$ =[0,0,0,0,0,0,0,1,1,0,0,1,0,1,1,1,1,0,0,0] when threshold is set to 0.7. It contains seven food-related tags which are "table", "food", "container", "coffee", "hand", "kitchenware", "cooking".

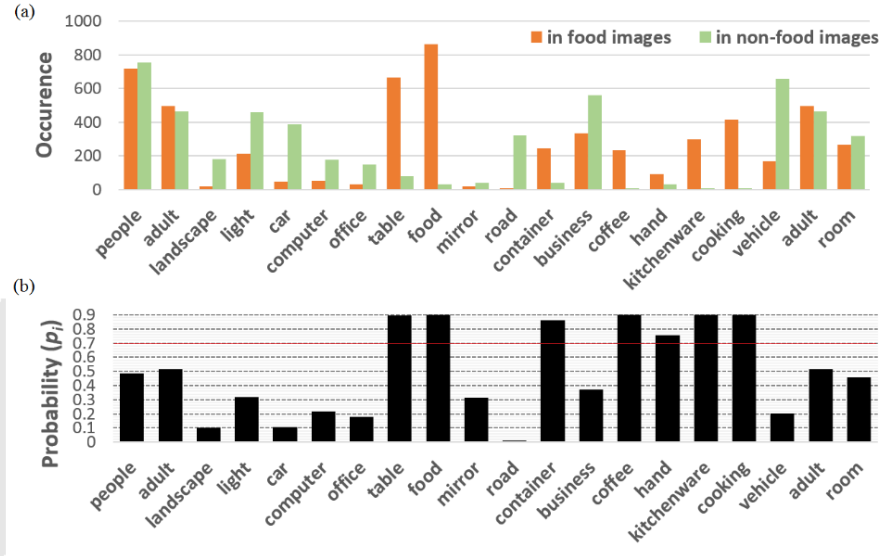

Fig. A2. An example showing the occurrence frequency of 20 tags in food images and non-food images(a), and the calculation of their probabilities in food images(b).

In the second step, we measured the similarity of each food-related tag in $\boldsymbol{y}$ with tag "food" in semantic domain. There have been a number of semantic similarity measures in the field of computational linguistics(1-4). Here we used the Jaccard and Dice measure to define the similarity between tag t_i and tag t_j as follows:

$$\text{Jaccard} = \frac{N(t_i \cap t_j)}{N(t_i\ ) + N(t_j) - N(t_i \cap t_j)}$$

$$\text{Dice} = \frac{2 * N(t_i \cap t_j)}{N(t_i) + N(t_j)} \quad , \tag{A4}$$

where $N(t_i)$ and $N(t_j)$ denotes the numbers of occurrences of tag t_i and tag t_j, respectively; and $N(t_i \cap t_j)$ denotes the co-occurrences of these two tags. Here t_j is the "food" tag because we need to calculate the relatedness of "food" with all other tags. An example of the semantic measures between several tags and "food" is shown in Fig. A3. It can be seen that, for several tags (e.g., round, sound, candle), its Jaccard or Dice measure is very small, meaning that these tags are not related with food semantically. Thus the elements in $\boldsymbol{y}$ corresponding to those tags are set to 0 to further remove the un-related tags. A threshold ε=0.05 determined experimentally is used in this study.

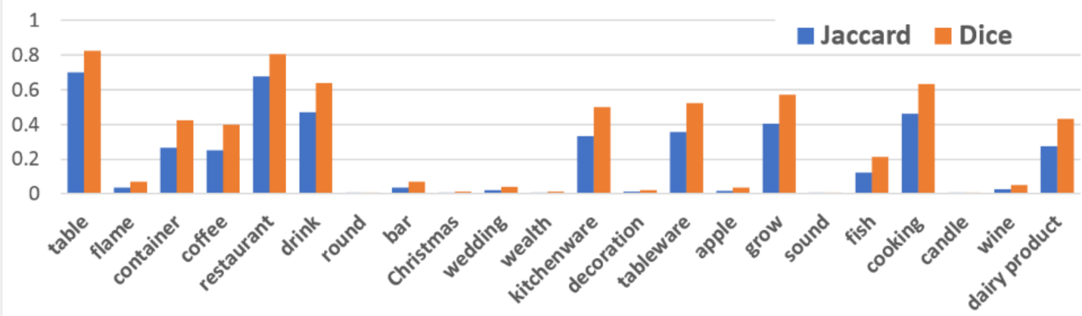

Fig. A3. Example of the semantic measures between several tags and the "food" tag.

Because y_i =1 means that tag t_i in the tag dictionary is related with food, we can calculate the total number of food-related tags in each image by computing the inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$, defined as an evidence index:

$$e(\boldsymbol{x}) = \boldsymbol{x} \cdot \boldsymbol{y} = \sum_{i=1}^{n}(x_i \cdot y_i). \tag{A5}$$

For any image, if $e(\boldsymbol{x})$ is higher than a threshold k, it is classified as a food image. Otherwise, it is a non-food image. If the number of tags generated for each image is not the same, a normalization factor has to be added as:

$$e(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{n} x_i \cdot y_i, \tag{A6}$$

where N is the number of non-zero elements in $\boldsymbol{x}$.

References:

1. Xu Z, Luo XF, Liu YH *et al.* (2014) Measuring semantic relatedness between flickr images: from a social tag based view. *Sci World J*, 758089.
2. Levandowsky M, Winter D (1971) Distance between sets. *Nature* **234**, 34-35.

3. Bhattacharjee S, Ghosh SK (2016) Measurement of semantic similarity: a concept hierarchy based approach. *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, 407-416.
4. Bollegala D, Matsuo Y, Ishizuka M (2011) A web search engine-based approach to measure semantic similarity between words. *IEEE Transactions on Knowledge and Data Engineering* **23**, 977-990.