Exploring Food Embeddings using Semi-supervised Vocabulary-informed Learning

Xin Ming Fan, Bjorn Kwok, Jason Sinn, Zhenying Zhang
Department of Statistics and Actuarial Science, University of Waterloo

Abstract

Despite significant progress in image recognition and segmentation, the ability to associate a semantic representation with images of food is still extremely limited. Problems in this domain are difficult due to high intraclass variation and the desire to associate taste with visual appearance. This report discusses a visual-semantic embedding space where the relationships between images of food, recipes, and ingredients can be investigated. In order to accomplish this, we use semi-supervised vocabulary-informed learning, leveraging an existing ingredient embedding space to train a convolutional neural network that can properly associate an image with a flavour. We show that the resulting approach can be used for classification with accuracy comparable to well-established, open domain models. We also investigate the relationship between an image of a food and its associated ingredients, and discuss other potential uses, such as the inference of unobserved recipes based on the implicit prototypes created.

1 Introduction

In recent years, food photography has become the largest amateur photography genre on social media. As a result, photo sharing platforms face immense daily influxes of food photos which go largely unclassified when compared to other genres of photography. Additionally, with health and fitness technology on the rise, mobile food photography is quickly becoming a meal calorie gross estimation tool. The problem of food recognition and determining its ingredients has not been fully explored by the computer vision community. Addressing the problem can not only help the aforementioned groups, but also lead to discovering new models that are able to combat problems with high intraclass variation.

Due to the unusual spatial layout of a food dish, food classification problems have inherently high intraclass variation [7]. Although image classification has progressed immensely with the introduction of convolutional neural networks (CNNs), machine learning methods are still unable to understand the semantic meaning behind such images. In the context of food, one possible approach to tackling the intraclass variation problem is to use its flavour profile as a dimensionality reduction technique. If the proper flavour can be associated with the image, then the classifier may become invariant to some instances of intraclass variation in the context of food. This leads to a desire for an approach that can properly encode an image into a flavour embedding.

It must be noted that inferring a flavour from an image is not possible without contextual knowledge about the food that it is associated with. As a result, an existing knowledge base in the domain of food is required.

Furthermore, if intraclass variation is high, it would be desirable to be able to deduce a flavour profile of an image which was not explicitly trained. This falls under the semi-supervised learning paradigm - specifically, it is a transductive learning problem.

2 Background Information

Semi-supervised learning is a class of supervised learning where both labeled and unlabeled data is used. We introduce the following terminology to partition our dataset:

$$\begin{aligned} \mathcal{W}_s &= \text{source set, containing labeled data,} \\ \mathcal{W}_t &= \text{target set, for which no labels are available,} \\ \mathcal{W}_t \cap \mathcal{W}_s &= \emptyset \\ \mathcal{W}_s \cup \mathcal{W}_t &= \mathcal{W} = \text{complete dataset} \end{aligned}$$

Zero-shot learning trains with samples from $\mathcal{W}_s$, while aiming to identify unobserved classes in $\mathcal{W}_t$ by transferring knowledge through intermediate-level semantic representations [1].

Open-set recognition [1] decides if an object is from the source set or from the target set. Note that this task does not involve identifying the specific class, but only detecting which set the object originated from.

3 Approach

One naive approach to tackling the training of such an encoder would be to first train a simple image classifier that takes an input image and returns a label corresponding to the food. Afterwards, the label can be encoded using a word embedding. However, this approach is "backwards" in the sense that it defeats the purpose of computing the embedding in the first place. Furthermore, error in either the classification or the embedding is propagated rather than reduced. Instead, we choose to directly embed the image into a visual-semantic space using semi-supervised, vocabulary-informed learning (SS-Voc) (*Fu et al.*)[1].

SS-Voc is a learning method which allows for supervised learning, zero shot learning, and open-set recognition in a single unified framework. SS-Voc does this by leveraging the relationship between word embeddings to directly train an encoder that optimizes the globally warped maximal-margin between both supervised and unsupervised atoms.

SS-Voc builds on zero-shot learning in the following aspects: (1) zero-shot learning assumes that the test set is from the target set and they cannot be misclassified as categories in the source set while SS-Voc tests on objects of categories from both the source set and the target set; (2) SS-Voc learns from only a few number of training data while zero-shot learning requires a large training set for each source class; (3) SS-Voc takes into account both the source set and target set (semi-supervised) information in the training stage while zero-shot learning only considers information from the source set [1].

In the case of recognizing food, it is not sufficient for the classifier to be able to detect whether the food has been seen or not. What would be desirable is the ability to recognize images in both the source and target sets. As a result, SS-Voc is an appropriate technique for this situation.

A common term used in the paper by *Fu et al.* [1] is the concept of a prototype. The prototype of a word is its "correct" location in the embedding space. In the original paper, all prototypes are known, or explicit, as the existing embedding space contains the corresponding label. Meanwhile, this report refers to both implicit and explicit prototypes. An implicit prototype in this report refers to the fact that the word does not already exist in the vocabulary of the embedding space, but instead must be computed in another fashion. The implicit prototype of a food will be denoted as $g(z_i)$, while the explicit prototype is denoted as u_{z_i}.

One possible formulation of the SS-Voc objective in the context of food is as follows:

$$\min_{W,V} \alpha\mathcal{L}(x, g(y_i), V; W) + (1-\alpha)(\mathcal{M}_t(x, g(z_i), V; W) + \mathcal{M}_s(x, u_{z_i}, V; W))$$

where $\mathcal{L}$ is the $\epsilon - SSVR$ distance function, $\mathcal{M}_t$ is the pairwise maximal margin term to each food, and $\mathcal{M}_s$ is the pairwise maximal margin term to each ingredient.

4 Implementation

4.1 Data Collection

In order to train a visual-semantic embedding for food images, recipes and ingredients, we first explored existing datasets such as ImageNet, Food-101 [6] and FooDD [8]. However, these datasets were too noisy due to either non-standardized photography methodology or the lack of labels. Instead, we decided to use a proof-of-concept dataset consisting of 5 training classes and 5 testing classes, making a total of 10 food classes and their corresponding ingredients for our open set. We trained on images of baked potatoes, cheesecakes, hamburgers, lemongrass chicken, and shrimp curry. Meanwhile, our test set included images of egg rolls, spaghetti, chicken on rice, simmered pork, and hot dogs. We gathered 100 images of each training class by scraping Google, and referred to the UEC FOOD 100 dataset for our test set. All images were resized to 48 by 48 pixels. Our training pipeline then split images from each class into sets of 50, 25, 25 for training, testing, and validation respectively.

4.2 Semantic Vocabulary Space

The paper by *Fu et al.* [1] used a Google word2vec model trained on a large text corpus of around 7 billion words to initialize the semantic space and generate the vocabulary. The source and target classes were mapped onto the semantic space under the word2vec assumption that words appearing in similar contexts have similar meanings and are, therefore, represented by similar vectors.

For the approach discussed in this report, food2vec [9] - an analogous, pre-trained word2vec model built on a collection of recipes scraped from the Internet - was used to represent a semantic ingredient space. Following the word2vec assumption that words appearing in similar contexts have similar meanings, food2vec assumes that ingredients appearing in similar recipes have similar flavour profiles. As a result, the dimensions of the food2vec vectors are a latent representation of an ingredient's flavour profile.

Since this report's approach is focused on a recipe-level representation of food, the assumption is made that a recipe is represented by the centroid of the ingredient vectors it is composed of. Note that this centroid assumption does not take into account the different weighting in amount or volume of each of the ingredients in a recipe.

4.3 Neural Network Architecture

In order to map directly from the image space to the semantic ingredient space, a convolutional neural network was trained. This network consisted of two 64 channel convolutional layers, followed by two feedforward layers. Both convolutional layers used a feature map of size 5x5 with a stride length of 1. After each convolution, a bias was added and the result was fed through a rectified linear activation function. Finally, max pooling with window size 3x3 and stride length 2 was used. The result was flattened and fed into a two layer feedforward network. No softmax function was used in the architecture as the resultant is not a probability. Instead, normalization of the 100 dimensional vector was computed manually using the L2 norm.

Implementation of the cost function was done manually using tensorflow matrix operations. At its core, the cost function can be summarized to be the globally warped maximal margin from the computed food embedding to its implicit prototype. This distance is augmented by a pairwise term to other foods as well as to other ingredients. One major difference between the implementation discussed in *Fu et al.* [1] is a simplification of the pairwise term to include all classes rather than the k-nearest neighbours. This was justified by the fact that the domain of the project was small and a full computation of the distances was feasible.

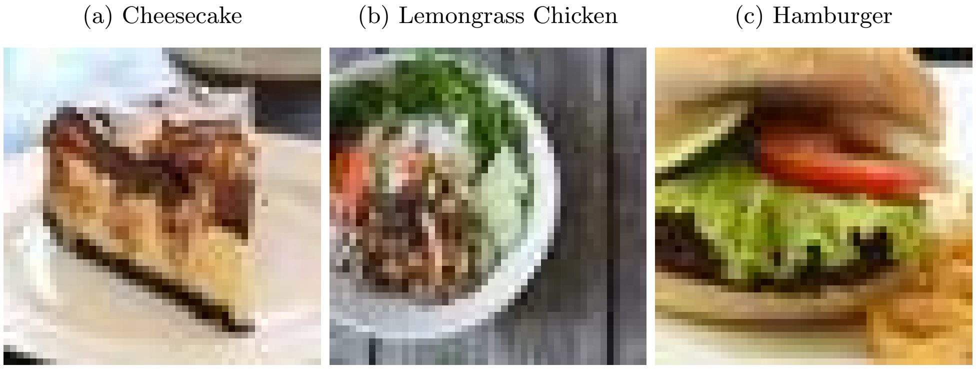
Another difference between the original implementation and the one discussed in this report is in the computation of the global warping matrix, V. In the original report, the global warping V was initialized as a separate weight matrix and was trained using a method similar to the expectation maximization algorithm. Instead, we make an assumption that a food can be computed as an average of its ingredients. In an open domain, recent NLP developments such as the seq2seq model [3] have shown that this assumption is quite naive. However, in the context of food, the flavour profile of a food is often the average of its ingredients. As a result, we argue that the semantic embedding space is already well-formed, and as a result, the global warping term can simply be omitted. We also argue that this allows us to create an implicit prototype in the ingredient space that can be done in a fully unsupervised fashion rather than the semi-supervised training of the prototype that is described in the original paper.

The paper by *Fu et al.* [1] uses the L-BFGS optimization method. This method is not implemented natively in tensorflow and the Adam algorithm [2] is used instead with step size $\eta = 10^{-4}$. Furthermore, instead of using the ϵ-SSVR distance function, the least squares distance is used. With a batch size of 25, the network was trained for 50 epochs.

5 Results and Discussion

After training an embedding space containing semantic definitions of food and ingredients as well as images of the aforementioned, inference can be conducted in a number of ways. It should be noted that due to the lack of existing work in the domain, quantitative evaluation of the results was somewhat difficult to attain.

A first naive method of evaluation that was conducted was simple classification. A pre-trained version of VGG-19 for image recognition was first used to find a baseline accuracy



(a) Cheesecake (b) Lemongrass Chicken (c) Hamburger

Samples of observed images. The nearest ingredients to each image are - Figure 1a: peppermint oil, butter, pastry flour, chocolate chips, cake mix. Figure 1b: teriyaki, prawns, rice, nuoc nam, sirloin steak. Figure 1c: bread, pepper, beef, salad dressing, salt.

for the context of food. The results of the network are shown in the confusion matrix found in Table 1. To compare, we used the embedding space created to compute a k-nearest neighbours graph for each resulting image embedding in order to find the top-1 and top-2 foods that might correspond to the given image. Meanwhile, the results of the top-1 and top-2 accuracy for our embedding space are found in Table 2.

Table 1: Baseline Results, VGG-19

1-NN: 67.2% 2-NN: 85.2%	Predicted				
Actual	14	3	2	5	1
	2	18	3	2	0
	2	5	16	1	1
	0	1	1	19	6
	1	1	3	3	17

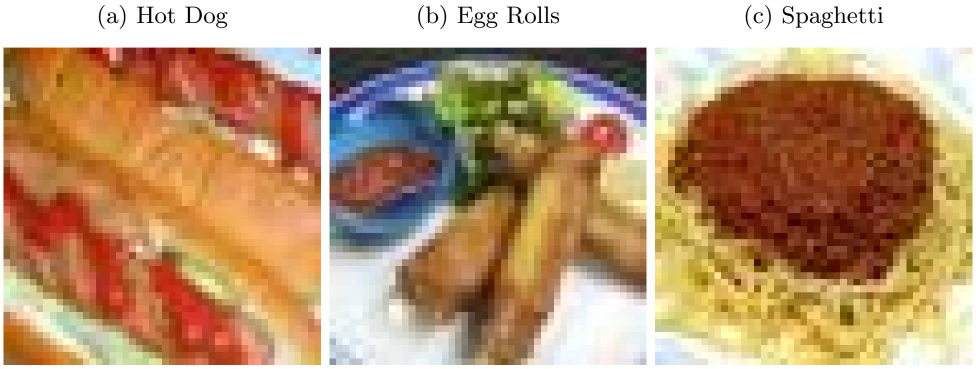
Table 2: Results from Embedding Space

1-NN: 75.2% 2-NN: 89.6%	Predicted				
Actual	17	1	2	3	2
	1	18	3	3	0
	1	1	19	4	1
	1	1	1	21	1
	1	0	1	4	19

Ordered Labels from left to right, top to bottom: Baked potato, Cheesecake, Hamburger, Lemongrass Chicken, Shrimp Curry

Comparing the two tables, we can see that the domain-specific embedding space that was created generates similar results to a well-established, open domain classifier [4], and performs better than other primitive methods for food recognition [5,6,7]. We suspect that the domain knowledge inferred by food2vec as well as through training of domain-specific examples allowed us to achieve this level of performance on foods with a relatively simple network. With more work on hyperparameter and network architecture optimization, as well as more computation power, we suspect that the results achieved could easily be improved.

On the other hand, we were also interested in the "semi-supervised"-ness of the embedding space. It should be noted that for this specific problem (where none of the class prototypes exist in the vocabulary space), quantitative evaluation is extremely difficult. One method of inference that might be interesting would be to find a k-nearest neighbours graph for the inferred image embedding on the nearest ingredients. Figure 1 depicts the $k = 5$


(a) Hot Dog

(b) Egg Rolls

(c) Spaghetti

Figure 2: Samples of unobserved images. The corresponding nearest ingredients are - Figure 2a: bread, salt, sausage, ground beef, italian sauce. Figure 2b: rice, chile, prawns, teriyaki, soy. Figure 2c: pepper, tandoori paste, cumin, cilantro, curry paste.

nearest ingredients to a given image of foods that were trained. These figures already show that the existing vocabulary information provides a basis for semi-supervised inference as training did not contain information about all of the ingredients that were listed at inference time.

Although not all of the ingredients were correct, it must be noted that the ingredients produced share a common flavour profile, and would often be found together in the same culture of food. This is likely due to the existing knowledge and training methods of the food2vec embedding space. We suspect that as food2vec continues to grow as a knowledge base for recipes and ingredients, the inferences will also become closer to the actual recipe.

Another experiment that was conducted was in the semi-supervised nature of the implicit prototypes. Since the original embedding space does not contain any information about the foods, we wanted to see whether or not we could infer new implicit prototypes based on the ones that we had trained. In order to do this, we initially trained on the 5 foods listed previously, and then created a test set consisting of 10 foods (the 5 foods listed, and then an additional 5). Afterwards, the inferred test embeddings were compared to implicit prototypes of the 10 foods, and results were recorded.

It should be noted that for this experiment, we could not find a suitable quantitative measure of performance. Overall, it was found that typically, the embeddings were quite inaccurate. In some cases, such as the example shown in Figure 2a, extremely accurate results were found. However, inferred ingredients were generally more similar to the example in Figure 2b for the most part.

The authors suspect that the small number of prototypes trained was an issue. Semi-supervised learning requires a large knowledge base from which it can use to infer gaps in its existing knowledge. In the current context, the amount of knowledge available to the embedding space about the implicit prototypes of food is extremely small. As a result, the decision boundaries between each of the foods is also not well-defined. If a new food is presented that is not ”covered” in the semantic space, then it will not be recognizable. This would be one possible explanation of why the results were poor in this situation. Furthermore, when the new food presented is strongly covered by the implicit prototypes, good results are created.

In order to increase the level of performance on this task, the authors suspect that more data should be collected. By having more implicit prototypes and more information, the decision boundaries between classes will become tighter, and as a result, inference on the areas in gaps of knowledge will be more precise.

6 Conclusions

This report applies the problem of semi-supervised vocabulary-informed learning to the domain-specific application of food. The approach outlined in the paper by *Fu et al.* [1] was used to train a better image classifier for both observed and unobserved classes by leveraging the semantic space represented by the pre-trained food2vec model [9]. The approach also allowed for semi-supervised inference of ingredients from images of food. For supervised image classification, the experimental results outline that a simple network leveraging food2vec performs comparably to well-established, open domain models. On the other hand, for inference of ingredients from an image of an observed food, the experimental results demonstrate that this approach does not effectively infer the exact ingredients that the recipe is composed of, rather ingredients with similar flavour profiles as the recipe are shown. Furthermore, the implementation was not effective in identifying unobserved classes due to the small number of prototypes trained. In the future, experimental results can be improved by gathering data for more implicit prototypes. Additionally, other ingredient-to-recipe assumptions can be explored.

References

[1] Y. Fu, L. Sigal. Semi-supervised Vocabulary-informed Learning. In CVPR, 2016.

[2] D. Kingma, J. Lei Ba. Adam: A method for stochastic Optimization. In ICLR, 2015.

[3] K. Cho et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In EMNLP, 2014.

[4] K. Simonyan, A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In ICLR, 2015.

[5] H. Kagaya et al. Food Detection and Recognition Using Convolutional Neural Network. In IEEE, 2015.

[6] L. Bossard et al. Food-101–mining discriminative components with random forests. In ECCV, 2014.

[7] M. Chen et al. Pfid: Pittsburgh fast-food image dataset. In IEEE, 2009.

[8] P. Pouladzadeh, A. Yassine, S. Shirmohammadi. FooDD: Food Detection Dataset for Calorie Measurement Using Food Images. In ICIAP, 2015.

[9] J. Altosaar. food2vec - Augmented cooking with machine intelligence. https://jaan.io/food2vec-augmented-cooking-machine-intelligence, 2017.