Recognition of Edible Vegetables and Fruits for Smart Home Appliances

Marco Buzzelli*, Federico Belotti, Raimondo Schettini
Dipartimento di Informatica, Sistemistica e Comunicazione
Università degli Studi di Milano-Bicocca
Viale Sarca 336, Milan 20126, Italy
marco.buzzelli@disco.unimib.it, f.belotti8@campus.unimib.it, schettini@disco.unimib.it

***Abstract*—We present a state of the art method for vegetable and fruit recognition based on convolutional neural networks. We developed our solution around the concept of a smart kitchen/refrigerator equipped with an on-board camera. With this objective in mind, we adopted a dataset that was specifically collected and annotated according to the eating characteristics of the portrayed items. We performed two types of experiment: we first trained and evaluated different state-of-the-art neural architectures on the task of vegetable and fruit recognition. Secondly, we designed and tested a solution that exploits the hierarchical nature of such classes to further improve the final performance of our system. Experimental results demonstrate the quantitative superiority of the proposed solution compared to existing approaches.**

***Index Terms*—vegetable, fruit, smart fridge, internet of things, deep learning**

I. Introduction

Object recognition in real-life digital images is a challenging task. Several difficulties arise from non-controlled acquisition environments, such as different illumination conditions, different poses of the subject, and cluttered scenes [1]. State of the art solutions typically address these problems through deep learning techniques, which take advantage from large quantities of training data to build robust and well-generalizing classifiers [2].

Despite recent advances in the field, the problem of fine-grained image classification is still difficult in some application domain. In fact, fine-grained classification requires the identification of classes that exhibit high inter-class similarity (e.g. distinguishing different types of pear: Crown pear, Bergamot pear, Sand pear), and as such it depends on the ability to spot very subtle differences that might help in class discrimination. This family of problems can be encountered in our everyday-life, therefore having great potential application in the Internet of Things, and enabling common-life objects to acquire greater value through the introduction of smart capabilities [3].

In this work we focus on automated fine-grained classification of edible plants, fruits, and mushrooms. We envision its application to smart kitchen appliances such as the refrigerator depicted in Fig. 1. Based on the recognition of available products, this device would be able to perform tasks that include:

* Corresponding author

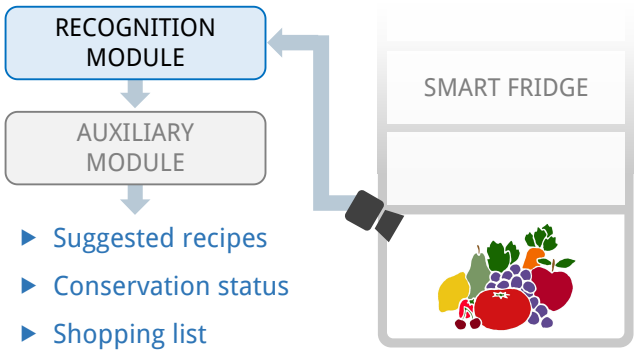

Fig. 1. Envisioned architecture for the application of vegetable and fruit recognition. The smart fridge will exploit an embedded camera, a recognition module, and an auxiliary module that processes the classification results to provide recommendations and alerts. The recognition module, which is the subject of this publication, can either be implemented on-board with GPU support, or through a client-server architecture.

- Proposing recipes. Through application of global and user-customizable ingredient networks [4], it is possible to suggest recipes that comply with daily diet monitoring schedules [5].
- Analyzing conservation status. Texture-based processing [6] and analysis of features related to cooking states [7] can help tracking the conservation status of stored goods.
- Suggesting shopping items. By combining recommender systems with methods for visual object counting [8], it is possible to provide the user with meaningful hints for items to buy.

All these features can be addressed with, or inspired by, existing literature work. However, we argue that the first step in this direction is a proper recognition of the raw food itself. To this extent, we exploit a very recent dataset containing vegetables, fruits, and other types of food in their uncooked status. Classes in this dataset are organized in a hierarchical structure, following standard agricultural taxonomies. We address the classification task with different neural architectures, and propose a solution that outperforms the state of the art. Finally, we investigate and propose different types of hierarchical training, in order to exploit the taxonomy provided with the dataset and further improve the classification performance.

978-1-5386-6095-9/18/$31.00 ©2018 IEEE

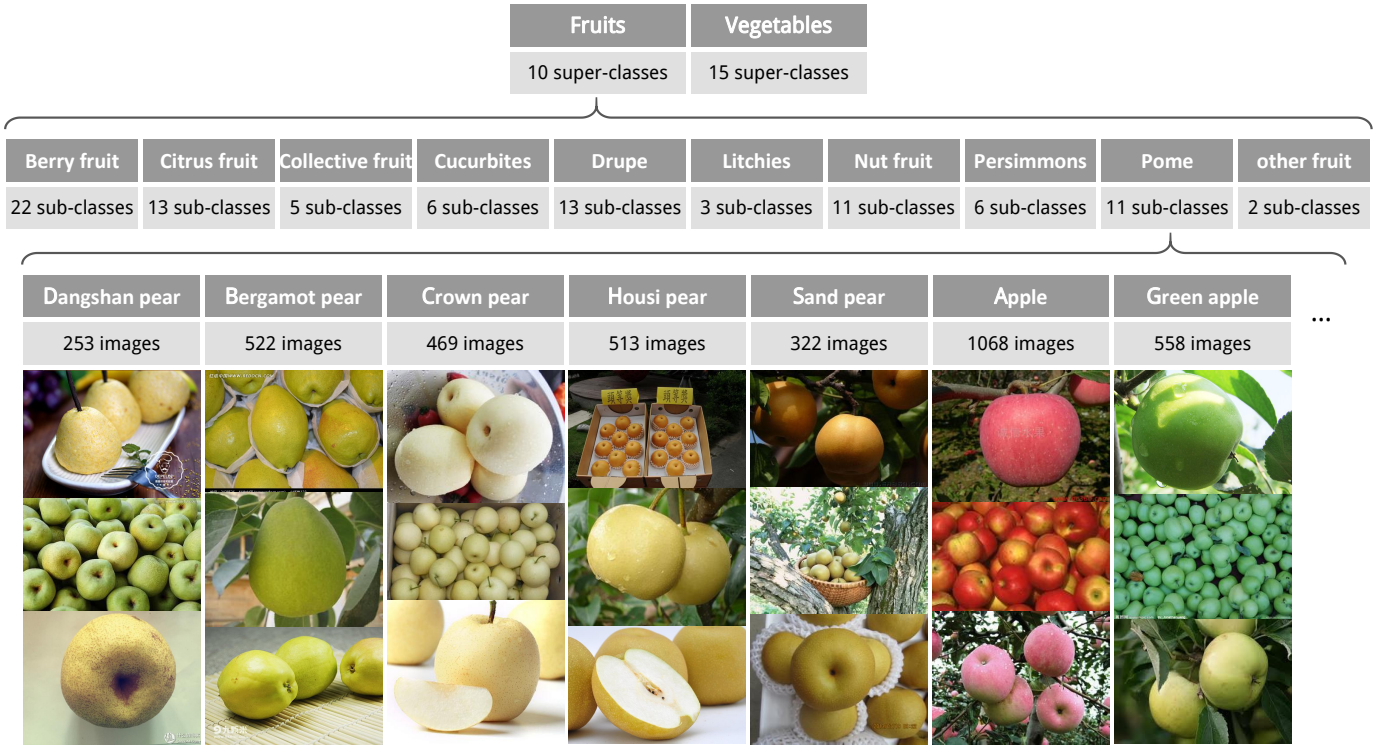

Fig. 2. Detail of the hierarchical annotation provided with the VegFru dataset. This figure shows a subset of the Fruit/Pome sub-class, highlighting the subtle interclass differences.

II. Datasets and related works

The literature offers several food-related datasets, such as UNIMIB2016 [9], Food-101 [10], UNICT889 [11], but very few of them focus on fruits and vegetables, specifically Fruits 360 [12] and VegFru [13]. VegFru is a dataset for fine-grained visual categorization specific to fruits (92 classes) and vegetables (200 classes, including several non-vegetable "mushroom" classes). Images belonging to each class are also labeled with an additional super-class following official agricultural and horticultural taxonomies, as depicted in Fig. 2, for a total of respectively 10 and 15 super-classes. This hierarchical partitioning is also used to separate different edible parts of the same item, which are potentially cooked in disparate ways, e.g. "soybean", "soybean sprout", and "soybean seed". We chose this dataset for its high cardinality (more than 160000 images), and variety of acquisition conditions, which suggest a successful translation to our envisioned domain involving a smart refrigerator.

In [13] authors trained a neural network architecture based on VGG-16 VGG-16 [14] (from Visual Geometry Group) to address both the sub-class and super-class problem of VegFru. They showed how it is possible to improve performance on the sub-classes, which is inherently more challenging, by fusing features extracted through the processing of both problems with a Compact Bilinear Pooling layer (CBP) [15]. In this work we propose and evaluate other fusion strategies that exploit the available hierarchical labeling, and compare them to the results obtained with CBP. The VegFru dataset is very recent, for this reason no other methods have yet been published that tackle its specific classification problem. A subset of this dataset was used in [16] as an auxiliary task to latently train a color constancy neural network without the need for explicit illuminant annotations, which was made possible thanks to the importance of color in vegetable discrimination.

III. Proposed method for vegetable and fruit recognition

In the following we present our approach to automated recognition of edible fruits and vegetables. The same techniques can also be easily extended to a broader set of food types thanks to our focus on the hierarchical nature of such classes. Specifically, we first introduce the different neural architectures that will be fine-tuned and evaluated, i.e. ResNet-34, ResNet-50 [17], and NASNet [18], and then we propose different solutions to exploit the hierarchical labeling through a specialization of the final fully-connected layer of the trained neural networks.

A. Neural baseline architectures

In [13] authors presented a baseline evaluation of several off-the-shelf neural networks on the VegFru dataset, including AlexNet [1], GoogLeNet [19], and VGG-16 [14]. They report best results on the use of GoogLeNet, followed closely by VGG-16. The analysis offered in [20] highlights how GoogLeNet and VGG-16 produce comparable results on

the ImageNet challenge [1], with VGG-16 requiring a much larger set of parameters ($\sim$125millions vs. $\sim$10 millions). A promising alternative suggested by the same study appears to be the ResNet architecture [17]: in particular, ResNet-34 and ResNet-50 produced a 5% performance leap compared to GoogLeNet, at the expense of only double the number of parameters ($\sim$20 millions). For this reason, we will evaluate the accuracy of both ResNet-34 and ResNet-50 on the task of vegetable and fruit recognition by fine-tuning and testing such architectures.

Authors of [18] developed a training framework that learns the appropriate neural architecture directly from observing the training data. They were able to automatically design (and, subsequently, train) state of the art architectures on a multiplicity of tasks, as well as prove the generalization capability of the learned architectures on tasks that were not seen during the search for an optimal neural structure. Here we extend this study by exploiting a network that was automatically designed for object recognition on CIFAR-10 [21], which we will refer to as NASNet (from Network Architecture Search).

B. Hierarchical labeling

The VegFru dataset is supplied with a two-level-hierarchy labeling, as described in Section II. Authors of [13] demonstrated the advantage of exploiting top-level annotations (so-called "super-classes") to improve the bottom-level predictions (or "sub-classes"), which was shown to be a more challenging task. In the following, we describe two techniques to exploit hierarchical labeling through the specialization of fully-connected layers.

The last block of a Convolutional Neural Network (CNN) for classification is typically characterized by a fully-connected layer, i.e. a linear combination mapping a N-dimensional feature vector to the representation of the final problem size. This is also preceded by other fully-connected blocks, building an internal hierarchy of interpretation used by the neural network. Our idea is to infuse this implicit representation with information coming from the explicit multilevel annotation, as depicted in Fig. 3:

1) We train, or fine-tune, the selected neural architecture on the super classes problem (lower cardinality).
2) We further fine-tune the whole model on the sub-classes problem, using one of the following approaches:
   - We replace the final fully-connected layer.
   - We append to the existing final fully-connected layer a new one.

In both cases, the final neural network will take as input the image, and produce a prediction vector with the same dimensions as the number of sub-classes.

IV. Experiments

In this section we systematically evaluate our proposed solutions under the standard evaluation criteria posed by authors of the VegFru dataset [13].

We fine-tuned the ResNet and NASNet architectures on the VegFru training subset, composed of 29200 examples, by applying data augmentation techniques that include random crop, random rotation, and random horizontal flip. These models, previously trained on the ImageNet task [1], were here fine-tuned via Stochastic Gradient Descent (SGD) with momentum, decaying learning rate of an order of magnitude every 30 epochs starting from 0.01, setting batch size at 4, for a total of 100 epochs. The epoch that produced the best accuracy on the validation set (14600 images) was then selected for evaluation, in order to avoid overfitting both on the training and final test data. The models were evaluated on the VegFru test set (116931 images) using mean top-1 accuracy, i.e. accuracy is computed for each class separately, and subsequently averaged. Results are presented in Table I.

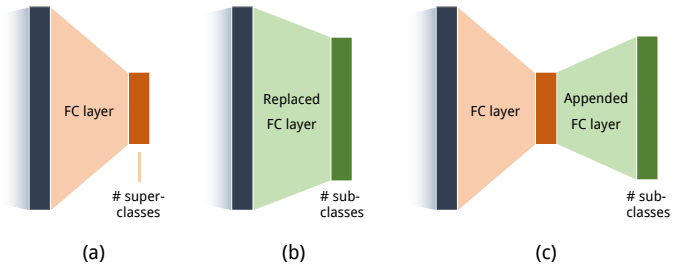

Fig. 3. Our fully-connected (FC) specialization exploits hierarchical labeling to improve performance on the sub-classes, with respect to directly training on the sub-classes themselves. (a) The initial FC always maps to the number of super-classes. (b) The "replace" specialization substitutes the original layer with a new one. (c) The "append" specialization concatenates a new FC layer.

TABLE I
RESULTS FOR BASELINE ARCHITECTURES ON BOTH THE SUPER- AND SUB-CLASS PROBLEM. THE SOLUTIONS WE PROPOSE, BASED ON RESNET AND NASNET, OUTPERFORM THE RESULTS PRESENTED IN [13].

Neural architecture		Super-classes (25 classes)	Sub-classes (292 classes)
[13]	AlexNet	72.87%	66.40%
	GoogLeNet	82.52%	79.22%
	VGG-16	82.45%	77.12%
	VGG-16 (CBP)	-	83.51%
This paper	ResNet-34	82.06%	81.14%
	ResNet-50	85.33%	79.78%
	NASNet	**87.41%**	**84.50%**

Coherently with the analysis reported in Section III-A, on the sub-class problem ResNet-34 performed better than all previously investigated baselines [13] that don't use hierarchical labels (AlexNet, GoogLeNet, and VGG-16), while ResNet-50 improved performance on the super-class problem. For both tasks, however, the best results were obtained through the fine-tuning and application of NASNet, respectively reaching 87.41% accuracy on the super-classes problem, and 84.50% on the sub-classes. This solution, based on a single neural model, outperformed also the 83.51% accuracy obtained by authors of [13] (VGG-16 CBP), who used Compact Bilinear Pooling [15] to merge the features of two neural networks independently trained on the super- and sub-class task. Example mistakes of the NASNet architecture that we fine-tuned to the task of vegetable and fruit recognition are shown in Fig. 4.

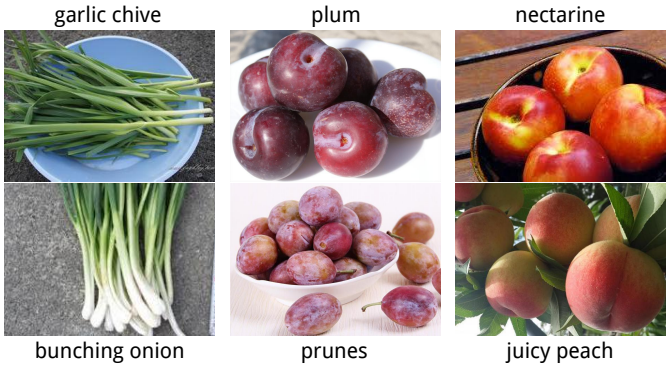

Fig. 4. Often-confused class pairs for our best performing model, based on NASNet. The mistakes are mainly due to extreme similarities between involved classes.

TABLE II
RESULTS FOR HIERARCHICAL LABELING EXPLOITATION. OUR PROPOSED SOLUTIONS BASED ON FULLY-CONNECTED SPECIALIZATION PROVIDE A FURTHER IMPROVEMENT ON THE SELECTED NEURAL BASELINES.

Neural architecture	Specialization technique	Before specialization	After specialization
ResNet-34	append	81.14%	74.93%
ResNet-34	replace	81.14%	82.06%
ResNet-50	replace	79.78%	84.10%
NASNet	replace	84.50%	84.08%

The second set of experiments aims at assessing how hierarchical label exploitation impacts the recognition performance on the sub-classes problem. Following Section III-B, we first fine-tuned ResNet-34 on the super-class problem, and subsequently applied fully-connected specialization on the sub-classes problem, by either replacing the last layer or appending a new one to it. Results reported in Table II show that the "append" technique highly degraded the final performance, while "replace" produced an appreciable improvement. This could be explained by the fact that, by constraining the network activations to a lower dimensionality compared to the final problem size, we are in fact compressing the data with a possible loss of information. The promising "replace" technique was then applied to the ResNet-50 and NASNet models as well. While the ResNet-50 architecture gained the most from this specialization, NASNet showed essentially no improvement at all. This behaviour possibly indicates that such a powerful neural model is already able to infer most of the hierarchy-related correlations directly from data annotated with only the sub-classes labels.

These experiments demonstrate the superiority of our approach for vegetable and fruit classification compared to existing solutions [13], based on the training of more powerful neural models, further enhanced by techniques for exploitation of hierarchical labeling.

V. CONCLUSIONS

Automated recognition of vegetables, fruits, and other types of food, can be the first step towards creating *actually smart* home appliances, and for the development of the Internet of Things in general. In this work we addressed such a challenging task through the training and application of state-of-the-art neural networks. By exploiting the hierarchical nature of the involved classes, we were able to outperform the previously existing solutions on a standard and publicly-available dataset. As a direction for future work, we will consider further exploration of hierarchical labeling, as well as the introduction of saliency-based segmentation techniques [22], in order to handle multiple items present in the same picture.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[2] S. Bianco, M. Buzzelli, D. Mazzini, and R. Schettini, "Deep learning for logo recognition," *Neurocomputing*, vol. 245, pp. 23–30, 2017.
[3] S. Poslad, *Ubiquitous computing: smart devices, environments and interactions*. John Wiley & Sons, 2011.
[4] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic, "Recipe recommendation using ingredient networks," in *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 2012, pp. 298–307.
[5] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition and leftover estimation for daily diet monitoring," in *International Conference on Image Analysis and Processing*. Springer, 2015, pp. 334–341.
[6] C. Cusano, P. Napoletano, and R. Schettini, "Evaluating color texture descriptors under large variations of controlled lighting conditions," *JOSA A*, vol. 33, no. 1, pp. 17–30, 2016.
[7] A. B. Jelodar, M. S. Salekin, and Y. Sun, "Identifying object states in cooking-related images," *arXiv preprint arXiv:1805.06956*, 2018.
[8] V. Lempitsky et al., "Learning to count objects in images," in *Advances in neural information processing systems*, 2010, pp. 1324–1332.
[9] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition: a new dataset, experiments, and results," *IEEE journal of biomedical and health informatics*, vol. 21, no. 3, pp. 588–598, 2017.
[10] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *European Conference on Computer Vision*. Springer, 2014, pp. 446–461.
[11] G. M. Farinella, M. Moltisanti, and S. Battiato, "Classifying food images represented as bag of textons," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5212–5216.
[12] H. Mureşan and M. Oltean, "Fruit recognition from images using deep learning," *arXiv preprint arXiv:1712.00580*, 2017.
[13] S. Hou, Y. Feng, and Z. Wang, "Vegfru: A domain-specific dataset for fine-grained visual categorization," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 541–549.
[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[15] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 317–326.
[16] M. Buzzelli, J. van de Weijer, and R. Schettini, "Learning illuminant estimation from object recognition," in *Image Processing (ICIP), 2018 IEEE International Conference on*. IEEE, 2018.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[18] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," *arXiv preprint arXiv:1707.07012*, 2017.
[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich *et al.*, "Going deeper with convolutions." Cvpr, 2015.
[20] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," *arXiv preprint arXiv:1605.07678*, 2016.
[21] A. Krizhevsky and G. Hinton, "Convolutional deep belief networks on cifar-10," *Unpublished manuscript*, vol. 40, p. 7, 2010.
[22] S. Bianco, M. Buzzelli, and R. Schettini, "Multiscale fully convolutional network for image saliency," *Journal of Electronic Imaging*, vol. 27, no. 5, p. 051221, 2018.