Uncertainty Modeling and Deep Learning Applied to Food Image Analysis

Eduardo Aguilar[1,2][a], Bhalaji Nagarajan[2][b], Rupali Khatun[2][c], Marc Bolaños[2][d] and Petia Radeva[2,3][e]

[1]*Departamento de Ingeniería de Sistemas y Computación, Universidad Católica del Norte, Avenida Angamos 0610, Antofagasta, Chile*

[2]*Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain*

[3]*Computer Vision Center, Cerdanyola, Barcelona, Spain*

eaguilar02@ucn.cl, {bhalaji.nagarajan, rupali.khatun, marc.bolanos, petia.ivanova}@ub.edu

Keywords: Uncertainty Modeling, Food Recognition, Deep Learning.

Abstract: Recognizing food images arises as a difficult image recognition task due to the high intra-class variance and low inter-class variance of food categories. Deep learning has been shown as a promising methodology to address such difficult problems as food image recognition that can be considered as a fine-grained object recognition problem. We argue that, in order to continue improving performance in this task, it is necessary to better understand what the model learns instead of considering it as a black box. In this paper, we show how uncertainty analysis can help us gain a better understanding of the model in the context of the food recognition. Furthermore, we take decisions to improve its performance based on this analysis and propose a new data augmentation approach considering sample-level uncertainty. The results of our method considering the evaluation on a public food dataset are very encouraging.

1 INTRODUCTION

In the present fast-paced world, unhealthy food habits are the basis of most chronic diseases (like obesity, diabetes, cardiovascular related diseases, thyroid, etc.). All over the world, problems regarding nutritional habits are related to the lack of knowledge about what people are eating on a daily basis. Unhealthy habits can more easily be prevented if they have the awareness about the nutritional value of the food they consume in their daily meals (Alliance, 2019). The problem is that more than 80% of people is not completely aware of how much they eat, what percentage of proteins, carbohydrates, salt, etc. are consumed in every plate. Moreover, it is quite difficult for people to calculate the nutritional aspects for every meal they consume (Sahoo et al., 2019). Manual calculation of this information is quite time-consuming and more often results in imprecise methods. This creates the need for automatic systems that would be able to log the food a person consumes everyday (Bruno and Silva Resende, 2017). This would enable both the patients and the health care professionals to better manage chronic conditions related to nutrition (El Khoury et al., 2019).

Figure 1: Example of high within-class variability belong to the ravioli food class.

Automatic food recognition is not only performed in the dietary management of patients, but has a wide variety of applications in the food and restaurant chains. Food detection in smart restaurants is becoming a practical application rather than a research problem (Aguilar et al., 2018). Automatic food recognition faces challenging computer vision and machine learning problems due to the nature of images that are used in this task (see Fig. 1).

Deep learning algorithms have become very pop-

[a] https://orcid.org/0000-0002-2463-0301
[b] https://orcid.org/0000-0003-2473-2057
[c] https://orcid.org/0000-0002-9682-5888
[d] https://orcid.org/0000-0001-9838-1435
[e] https://orcid.org/0000-0003-0047-5172

ular, and they own this popularity to their exceptional performance, enhanced processing abilities, large datasets, and outstanding classification abilities compared to the traditional machine learning methods (Subhi et al., 2019). However, despite the good performance shown, deep learning algorithms need huge amounts of data or they are prone to overfitting. To avoid it, one of the most difficult and general problems in this work is getting an adequate dataset, which not only means a large dataset, but also composed of very diverse and carefully curated samples.

Data augmentation is a popular strategy adopted to prevent deep learning methods from overfitting. It consists in applying transformations to the original data in order to increase the sample size and its variability. Examples of standard transformations in images are: random crops, image flips or reflections and color distortions. On the other hand, novel solutions have been provided by Generative Adversarial Network-based methods (GANs), which can generate synthetic, new and plausible images. However, the majority of data augmentation strategies have been applied indistinctly for all the images, without taking into account that in some cases, particular classes or images can be harder to classify and would require more particular data augmentation methods. On the other hand, uncertainty analysis can give us a good clue to understand what does the model learn and from this, we can expand the dataset to overcome the deficiencies we find. In this work, we propose to explore a combination of both fields: GANs and uncertainty modelling, with the aim of generating new data focusing on the samples that the model has not been able to learn well (with high uncertainty).

The major contributions of this work are as follows: a) to use Epistemic Uncertainty to find the samples that are the hardest for the model to learn; and b) to use Generative Adversarial Networks to perform data augmentation to create visually similar images to the hard samples in the dataset. The rest of the work is organized as follows. Next section details the recent relevant literature. Section 3 explains the proposed methodology. Experimental details are provided in Section 4, followed by conclusions in the last Section.

2 RELATED WORK

Food image analysis is an active area of research, which analyses food data from various sources and applies it to solve different food-related tasks. Here, the most relevant recent literature is discussed.

2.1 Food Recognition

Food recognition is a challenging computer vision task, due to the complex nature of food images. The images could contain dishes that are mixed or could contain many food items (Wang et al., 2019). The task is of a fine-grained nature, where the classes have high intra-class variability and high inter-class similarity.

The initial works related to the recognition task used different hand-crafted features such as color, texture and shape (Matsuda et al., 2012; Chen et al., 2009; Joutou and Yanai, 2009; Bosch et al., 2011). These works were primarily concerned with tackling the problem in a constrained environment. The datasets during these studies had less number of images or classes and are restrictive in the conditions in which the images were taken (Ciocca et al., 2017a; Matsuda et al., 2012; Jing-jing Chen, 2016).

With the advent of Convolutional Neural Networks (CNN), food recognition tasks of complex nature were also tackled. CNNs were able to massively outperform by far the traditional food recognition algorithms. The datasets started to have large numbers of images and a large number of dishes were being recognized (Bossard et al., 2014; Ciocca et al., 2017b; Donadello and Dragoni, 2019; Kaur et al., 2019). Different CNNs have been successfully applied to food recognition task, as AlexNet (Yanai and Kawano, 2015), GoogLeNet (Wu et al., 2016; Meyers et al., 2015; Liu et al., 2016a), Network-In-Networks (Tanno et al., 2016), Inception V3 (Hassannejad et al., 2016), ResNet-50 (Ming et al., 2018), Ensemble NN (Nag et al., 2017), Wide Residual Networks (Martinel et al., 2018), and CleanNet (Lee et al., 2018).

Food images in the wild often contain more than one food class. Therefore multi-label food recognition and detection have an increased complexity. Also, different food can be located very close to each other or even mixed. In this case, food recognition is usually preceded by food detection (Anzawa et al., 2019). Earlier works involved using colour and texture-based food segmentation (Anthimopoulos et al., 2014). (Aguilar et al., 2019) proposed a semantic food framework, covering food segmentation, detection and recognition. (Chen et al., 2017) focused on multi-label ingredient recognition, while a multi-task learning has been proposed in the works of (Aguilar et al., 2019; Zhou et al., 2016).

The high inter-class similarity of the food images makes it difficult to train models that could be used to recognize dishes in the wild. Although large datasets are created with more classes, the images do not represent the complex nature of the food. Therefore, generative models could be used to create new synthetic

data that are similar to the real world data.

2.2 Generative Adversarial Network

A Generative Adversarial Network (Goodfellow et al., 2014) is a deep learning method that generates very realistic synthetic images in the domain of interest. In the GAN framework, two different networks compete with each other. And these two networks work as thief (generator) and police (discriminator). The Generator, as its name states, generates fake samples from random noise and tries to fool the Discriminator. On the other hand, the Discriminator has the role of distinguishing between real and fake samples.

They both compete with each other in the training phase. The steps are repeated several times in order for the Generator and Discriminator to get better in their respective jobs after each iteration.

One of the most popular extensions of Generative Adversarial Nets is the conditional model. Note that in an unconditioned generative model, there is no control over modes of the data being generated. However, in the Conditional GAN (CGAN) (Mirza and Osindero, 2014), the generator learns to create new samples with a specific condition or set of characteristics. Such conditioning could be based on class labels, on some part of data for inpainting like (Goodfellow et al., 2013), or even on data from different modalities. Thus, in CGAN both the generator and the discriminator are conditioned on some extra information y, where y could be any kind of auxiliary information such as a label associated to an image or more detailed tag, rather than a generic sample from an unknown noisy distribution.

A further extension of the GAN architecture, which is built upon the CGAN extension, is the Auxiliary Classifier GANs (ACGAN) (Odena et al., 2017). In ACGAN, the input is the latent space along with a class label. Furthermore, every generated sample has a corresponding class label. The Generator model receives as input, a random point from the latent space and a class label, and gives as output the generated image. The Discriminator model receives as input an image and returns as output the probability that the provided image is real, or the probability of the image belonging to each known class. As well known, unbalanced data is a big problem for object recognition, where models tend to classify much better in the dominant classes. In the case of the GANs, this problem is also present producing low quality synthetic images in classes with few samples. Some proposals have addressed this problem (Mariani et al., 2018; Ali-Gombe and Elyan, 2019), which are discussed in the following paragraphs.

In BAGAN (Mariani et al., 2018), an augmentation framework is proposed to restore balance in unbalanced datasets by creating new synthetic images for minority classes. The proposed approach requires two training steps: the first corresponds to initializing the GAN with the features learned by means of an auto-encoder, and then the entire model is retrained. Another approach to restore balance is MFC-GAN (Ali-Gombe and Elyan, 2019). Oppositely to BAGAN, MFC-GAN is simpler to train and just needs one training step. This model uses multiple fake classes to ensure a fine-grained generation and classification of the minority class instances.

A novel method titled SINGAN (Shaham et al., 2019) has been recently published. Differently to previous GANs, this model is an unconditional generative model that can be learned from a single natural image. SINGAN model is trained to capture the internal distribution of the image patches, then it generates high quality, diverse samples that contain the same visual content as the image. The pyramid structure of fully convolutional layers of SINGAN learns the patch distribution of the image at a different scale in each layer. This results in generating new samples of arbitrary size and aspect ratios that have significant variability, yet maintain both the global structure and the fine textures of the training image.

SINGAN requires to train a separate model for each new sample that one desires to generate from a particular image, thus, becoming a very expensive technique. However, this can be very useful if we only need to increase a small subset of the images. Uncertainty modeling can help us decide which subset is best suited to improve model performance.

2.3 Uncertainty Modeling

Uncertainty can be explained simply as a state of doubt about what the model has or has not learned from the input data. In Bayesian modeling, uncertainty mainly can be presented in two different ways (Kendall and Gal, 2017): aleatory uncertainty, which captures noise inherent in the observations; and epistemic uncertainty, which can be explained away given enough data. The uncertainty can be captured from a Bayesian Neural Network (BNN). However, in a Deep Learning scheme it becomes intractable in this way (Blundell et al., 2015; Gal and Ghahramani, 2016; Sensoy et al., 2018). Instead, variational Bayesian methods have been adopted in the literature (Blundell et al., 2015; Gal and Ghahramani, 2016; Molchanov et al., 2017; Louizos and Welling, 2017), where MC-dropout (Gal and Ghahramani, 2016) is the most popular technique to estimate the uncertainty

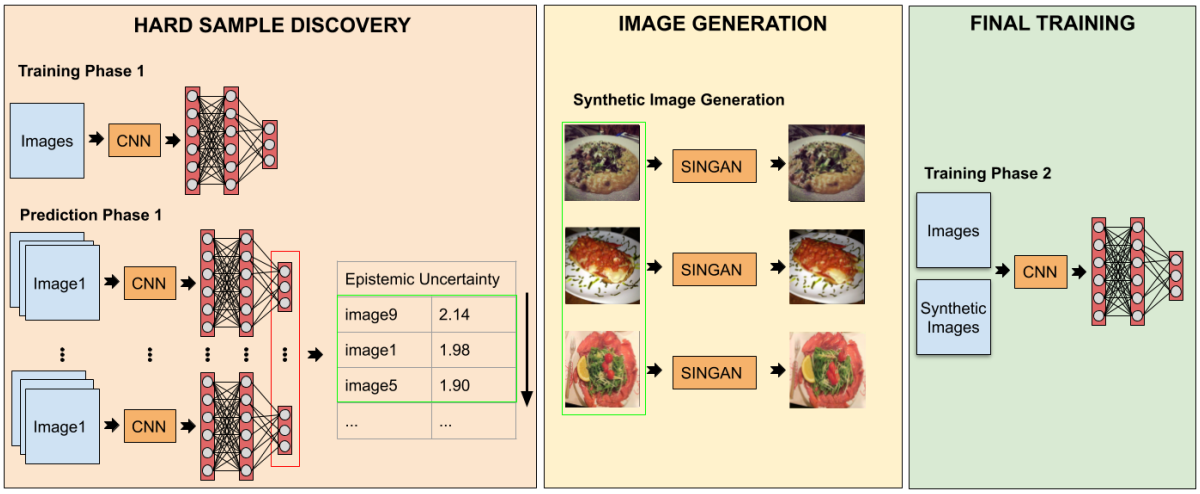

Figure 2: Main scheme of the our UAGAN food recognition method.

due to its simplicity regarding to the implementation.

Recent methods of image classification have adopted this technique to estimate the uncertainty in their scheme (Aguilar et al., 2019; Khan et al., 2019; Aguilar and Radeva, 2019b; Aguilar and Radeva, 2019a; Nielsen and Okoniewski, 2019). In the case of (Aguilar et al., 2019), the aleatory uncertainty is used in order to weigh dynamically different kinds of losses for multi-label and single-label food-related tasks. On the other hand, in (Khan et al., 2019), the authors deal with the imbalanced object classification problem. They redefined the large-margin softmax loss (Liu et al., 2016b), incorporating uncertainty at the class-level and sample-level based on the Bayesian uncertainty measure to address the rarity of the classes and the difficulty level of the individual samples. Regarding (Aguilar and Radeva, 2019b; Aguilar and Radeva, 2019a), the analysis of the epistemic uncertainty has been applied for different purposes: to identify the best data augmentation that will be applied in a particular class (Aguilar and Radeva, 2019a) and to judge when a flat or hierarchical classifier is used (Aguilar and Radeva, 2019b). A work closer to our proposal, but not in the food recognition field, is that published by (Nielsen and Okoniewski, 2019), that proposes an active learning scheme based on acquisition function sampling. This mechanism considers the prediction uncertainty of the classifier to determine the GAN samples to incorporate in the training set, which are labeled by an external oracle.

The main differences between our proposal and (Nielsen and Okoniewski, 2019) are the following: a) our aim is completely different, we apply the uncertainty analysis to discover complex samples to perform data augmentation, and not to apply it after the data augmentation to select the sample that will be used during the training, b) our training scheme is done in two phases, and not several phases, which do not require an external oracle to do it, because the labels are automatically assigned, and c) we adopt a GAN that generates a new sample keeping a high quality content of the input image, instead of generating a sample by merging different input images that in some cases can be very noisy and insert a bias towards the most frequent content.

3 UNCERTAINTY-AWARE GAN-AUGMENTED FOOD RECOGNITION

In this section, we describe all phases involved in the Uncertainty-Aware GAN-Augmented (UAGAN) method to perform food recognition using uncertainty modeling and GANs. As you can see in Fig. 2, the method contemplates 3 main phases with the following purposes: a) hard samples discovery, b) synthetic image generation and c) final training.

3.1 Hard Sample Discovery

The first step of our proposed approach involves the analysis of the food images of the training set, with the aims of identifying those that are difficult to classify. To do this, our criterion is based on the analysis of Epistemic Uncertainty (EU) through the calculation of the entropy. The samples with high uncertainty are those in which the model has not been able to learn well their discriminant features and, therefore, are considered hard samples. On the other hand, we adopt the method called MC-dropout (Gal and

Ghahramani, 2016) for EU estimation, mainly due to its simple implementation. Basically, we need to add a dropout layer before each fully connected layer, and after the training, we perform K predictions with the dropout turned on. The K probabilities (softmax outputs) are averaged and then the entropy is calculated to reflect the EU. Finally, the images are ordered with respect to their EU, and we select the top n images with higher EU to perform the next step.

3.2 Image Generation

Once the images have been chosen, the next step corresponds to increasing the data with nearby images in terms of visual appearance. We believe that one of the determining factors that does not allow the model to learn the features of hard images corresponds to the fact that they differ from most images that represent a particular class. This hard images may be present in the training set due to the complexity of the acquisition and also after dividing the data for the training. The latter is due to the fact that during the generation of subsets only the sample size is considered and not the variability of the sample. Therefore, we propose to make new images by applying small changes to the original ones. The best method for this purpose is the recent GAN-based method called SINGAN (Shaham et al., 2019), which can learn from a single image and generate different samples carrying the same visual content of the input image. In this step, we adopt the SINGAN to generate one synthetic image for each chosen image according to the uncertainty criterion.

3.3 Final Training

Finally, in the last step, the whole CNN model is trained with both types of images: the synthetic images obtained with SINGAN and the original images.

4 VALIDATION

In this section, we first describe the dataset used to evaluate the proposed approach, which is composed of public images of food belonging to Italian cuisine.

Next, we describe the evaluation metric and experimental setup. Finally, we present the results obtained with the baseline methods and our proposal.

4.1 Dataset

From the dataset MAFood-121 (Aguilar et al., 2019), we use all the images of the dishes that belong to the Italian cuisine. In total, 11 dishes were chosen, which are composed of 2468 images with a maximum, minimum and average of 250, 104 and 224 images, respectively. The data is distributed as 72% of the images for training, 11% for validation and 17% for test.

4.2 Metric

In order to evaluate our proposal, we use the standard metric used for object recognition named overall Accuracy (Acc). We evaluate our experiment 5 times and show the result in terms of average accuracy and the respective standard deviation.

4.3 Experimental Setup

For classification purposes, ResNet-50 (He et al., 2016) was adopted as the base CNN architecture. We adapted this model to be able to apply MC-dropout by removing the output layer, and instead, we added one fully connected layer of 2048 neurons, followed by a dropout layer with a probability of 0.5, and ended up with an output layer of 11 neurons with softmax activation. For simplicity, we call this architecture the same as the original (ResNet50). As for training, we use the categorical cross-entropy loss and the Adam optimizer to train all models during 40 epochs with a batch-size of 32, initial learning rate of 0.0002, decay of 0.2 every 8 epochs and patience of 10 epochs.

Three different training strategies of the same model are used for a benchmark purpose:

- ResNet50, baseline model training with the original images without data augmentation.
- ResNet50+SDA, baseline model with standard data augmentation applying during the training, like random crops and horizontal flips.
- UAGAN , ResNet50+SDA using the real and synthetic images.

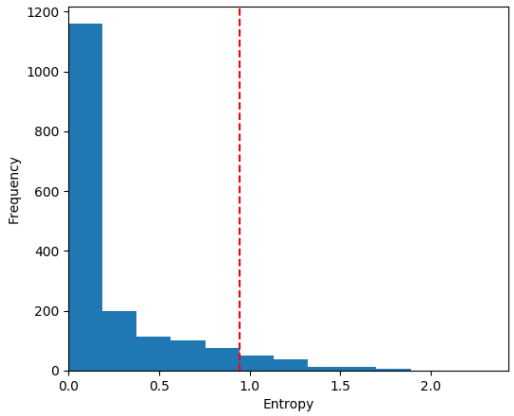

Figure 3: Histogram for the entropy of the predicted images.

With respect to the image generation, we use the default parameters proposed by the authors of SINGAN.

4.4 Results

In this section, we present the results obtained by the proposed method. The first step of our method corresponds to selecting those images difficult to classify (with high uncertainty). After training the model with the original images, we determine the EU and build a histogram for the training images (see Fig. 3). The right side of the histogram corresponds to all the images considered to generate the new ones. The criterion applied corresponds to selecting all images with EU equal to or greater than the uncertainty calculated by the average between the maximum and minimum uncertainty predicted for all images. A total of 120 images was selected.

In Fig. 4, we can see the distribution of the training images along each dish, the average entropy in all the images, the proportion of the selected images and the average entropy for the selected images. Unlike the evidence shown in (Khan et al., 2019) for CIFAR-10, for this type of data, the frequency of the images is not a factor that determines a high or low uncertainty for a specific class. In our case, we believe that uncertainty occurs due to the great variability of visual appearance that may be present in the images belong to the same class of dish, where the factor to consider is the diversity of the collected sample and not only the size of the sample. To fill the gap of poorly represented sample for a class, we duplicate the presence of images with high uncertainty through a generation of synthetic images with the SINGAN method. In Fig. 5, some examples of the generated images are shown.

With a total of 1889 training images, 1679 originals and 120 synthetic ones, we train the final model. The results obtained by three different training strategies of the same model are shown in the Table 1. All models were fine-tuned from ImageNet (Krizhevsky et al., 2012) and retrained the whole network using the target training set. For each strategy, 5 models were trained with random initialization values and random order of images. Then, we calculated the average accuracy and the standard deviation achieved for the best model obtained on each iteration according to the performance on the validation set. For the results achieved, we can see that UAGAN improved the performance in terms of accuracy with respect to the rest of strategies. Specifically, the improvement is 3,17% on ResNet50 and 1,32% on ResNet50+SDA.

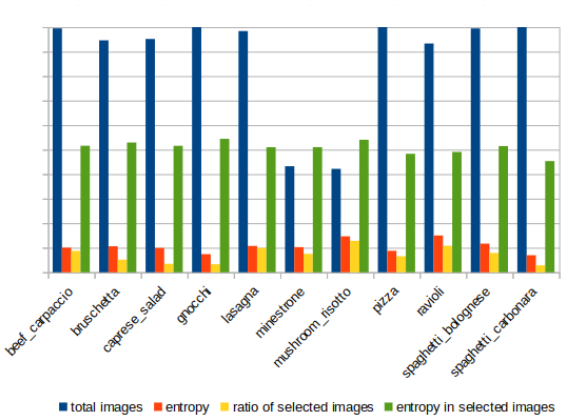

Figure 4: Training images vs epistemic uncertainty.

Table 1: Results obtained on the test set in term of accuracy with the standard deviation.

Method	Acc	Std
ResNet50	79.15%	0,60%
ResNet50 + SDA	81.00%	0,78%
UAGAN (our proposal)	82.32%	0,96%

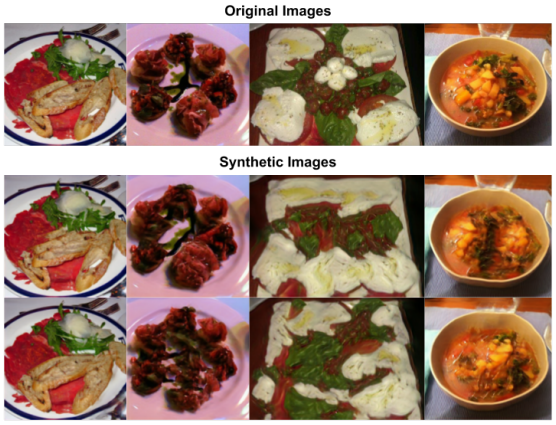

Figure 5: Synthetic image generated on the selected images from the training set.

5 CONCLUSIONS

In this paper, we presented a novel method for sample-level uncertainty-aware data augmentation composed of three phases: 1) identification of hard samples, by means of analysis of the epistemic uncertainty; 2) generating new data from identified samples; and 3) performing the final training with the original and synthetic images. We demonstrated the effectiveness of the approach proposed on the Italian dishes from MAFood121 public dataset. The result obtained shows that our proposal outperforms the

classification by incorporating only 120 synthetic images based on the uncertainty analysis (5% of the total). As future work, we will explore both sample-level and class-level uncertainty to increase deep learning datasets in an active learning framework.

ACKNOWLEDGEMENTS

This work was partially funded by TIN2018-095232-B-C21, SGR-2017 1742, Nestore ID: 769643, Validithi and CERCA Programme/Generalitat de Catalunya. E. Aguilar acknowledges the support of CONICYT Becas Chile. P. Radeva is partially supported by ICREA Academia 2014. We acknowledge the support of NVIDIA Corporation with the donation of Titan Xp GPUs.

REFERENCES

Aguilar, E., Bolaños, M., and Radeva, P. (2019). Regularized uncertainty-based multi-task learning model for food analysis. *Journal of Visual Communication and Image Representation*, 60:360–370.

Aguilar, E. and Radeva, P. (2019a). Class-conditional data augmentation applied to image classification. In *International Conference on Computer Analysis of Images and Patterns*, pages 182–192. Springer.

Aguilar, E. and Radeva, P. (2019b). Food recognition by integrating local and flat classifiers. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 65–74. Springer.

Aguilar, E., Remeseiro, B., Bolaños, M., and Radeva, P. (2018). Grab, pay, and eat: Semantic food detection for smart restaurants. *IEEE Transactions on Multimedia*, 20(12):3266–3275.

Ali-Gombe, A. and Elyan, E. (2019). Mfc-gan: class-imbalanced dataset classification using multiple fake class generative adversarial network. *Neurocomputing*, 361:212–221.

Alliance, I. U. N. (2019). National adult nutrition survey. *Public Health*.

Anthimopoulos, M. M., Gianola, L., Scarnato, L., Diem, P., and Mougiakakou, S. G. (2014). A food recognition system for diabetic patients based on an optimized bag-of-features model. *IEEE journal of biomedical and health informatics*, 18(4):1261–1271.

Anzawa, M., Amano, S., Yamakata, Y., Motonaga, K., Kamei, A., and Aizawa, K. (2019). Recognition of multiple food items in a single photo for use in a buffet-style restaurant. *IEICE TRANSACTIONS on Information and Systems*, 102(2):410–414.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *ICML*, pages 1613–1622.

Bosch, M., Zhu, F., Khanna, N., Boushey, C. J., and Delp, E. J. (2011). Combining global and local features for food identification in dietary assessment. In *2011 18th IEEE International Conference on Image Processing*, pages 1789–1792. IEEE.

Bossard, L., Guillaumin, M., and Van Gool, L. (2014). Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*.

Bruno, V. and Silva Resende, C. J. (2017). A survey on automated food monitoring and dietary management systems. *Journal of health & medical informatics*, 8(3).

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848.

Chen, M., Dhingra, K., Wu, W., Yang, L., Sukthankar, R., and Yang, J. (2009). Pfid: Pittsburgh fast-food image dataset. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 289–292. IEEE.

Ciocca, G., Napoletano, P., and Schettini, R. (2017a). Food recognition: a new dataset, experiments and results. *IEEE Journal of Biomedical and Health Informatics*, 21(3):588–598.

Ciocca, G., Napoletano, P., and Schettini, R. (2017b). Learning cnn-based features for retrieval of food images. In *International Conference on Image Analysis and Processing*, pages 426–434. Springer.

Donadello, I. and Dragoni, M. (2019). Ontology-driven food category classification in images. In *International Conference on Image Analysis and Processing (2)*, volume 11752 of *Lecture Notes in Computer Science*, pages 607–617. Springer.

El Khoury, C. F., Karavetian, M., Halfens, R. J., Crutzen, R., Khoja, L., and Schols, J. M. (2019). The effects of dietary mobile apps on nutritional outcomes in adults with chronic diseases: A systematic review. *Journal of the Academy of Nutrition and Dietetics*.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059.

Goodfellow, I., Mirza, M., Courville, A., and Bengio, Y. (2013). Multi-prediction deep boltzmann machines. In *Advances in neural information processing systems*, pages 548–556.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Hassannejad, H., Matrella, G., Ciampolini, P., De Munari, I., Mordonini, M., and Cagnoni, S. (2016). Food image recognition using very deep convolutional networks. In *Proceedings of the 2nd International Workshop on MADiMa*, pages 41–49. ACM.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jing-jing Chen, C.-w. N. (2016). Deep-based ingredient recognition for cooking recipe retrival. *ACM Multimedia*.

Joutou, T. and Yanai, K. (2009). A food image recognition system with multiple kernel learning. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 285–288. IEEE.

Kaur, P., Sikka, K., Wang, W., Belongie, S., and Divakaran, A. (2019). Foodx-251: A dataset for fine-grained food classification. *arXiv preprint arXiv:1907.06167*.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584.

Khan, S., Hayat, M., Zamir, S. W., Shen, J., and Shao, L. (2019). Striking the right balance with uncertainty. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 103–112.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Lee, K.-H., He, X., Zhang, L., and Yang, L. (2018). Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5447–5456.

Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., and Ma, Y. (2016a). Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. In *International Conference on Smart Homes and Health Telematics*, pages 37–48. Springer.

Liu, W., Wen, Y., Yu, Z., and Yang, M. (2016b). Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7.

Louizos, C. and Welling, M. (2017). Multiplicative normalizing flows for variational bayesian neural networks. In *ICML-Volume 70*, pages 2218–2227. JMLR. org.

Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., and Malossi, C. (2018). Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*.

Martinel, N., Foresti, G. L., and Micheloni, C. (2018). Wide-slice residual networks for food recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 567–576. IEEE.

Matsuda, Y., Hoashi, H., and Yanai, K. (2012). Recognition of multiple-food images by detecting candidate regions. In *2012 IEEE International Conference on Multimedia and Expo*, pages 25–30. IEEE.

Meyers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., and Murphy, K. P. (2015). Im2calories: towards an automated mobile vision food diary. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1233–1241.

Ming, Z.-Y., Chen, J., Cao, Y., Forde, C., Ngo, C.-W., and Chua, T. S. (2018). Food photo recognition for dietary tracking: System and experiment. In *International Conference on Multimedia Modeling*, pages 129–141. Springer.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Molchanov, D., Ashukha, A., and Vetrov, D. (2017). Variational dropout sparsifies deep neural networks. In *ICML-Volume 70*, pages 2498–2507. JMLR. org.

Nag, N., Pandey, V., and Jain, R. (2017). Health multimedia: Lifestyle recommendations based on diverse observations. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 99–106. ACM.

Nielsen, C. and Okoniewski, M. (2019). Gan data augmentation through active learning inspired sample acquisition. In *Proceedings of the IEEE conference on computer vision and pattern recognition Workshops*, pages 109–112.

Odena, A., Olah, C., and Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. In *ICML-Volume 70*, pages 2642–2651. JMLR. org.

Sahoo, D., Hao, W., Ke, S., Xiongwei, W., Le, H., Achananuparp, P., Lim, E.-P., and Hoi, S. C. (2019). Foodai: Food image recognition via deep learning for smart food logging.

Sensoy, M., Kaplan, L., and Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In *Advances in neural information processing systems*, pages 3179–3189.

Shaham, T. R., Dekel, T., and Michaeli, T. (2019). Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4570–4580.

Subhi, M. A., Ali, S. H., and Mohammed, M. A. (2019). Vision-based approaches for automatic food recognition and dietary assessment: A survey. *IEEE Access*, 7:35370–35381.

Tanno, R., Okamoto, K., and Yanai, K. (2016). Deepfoodcam: A dcnn-based real-time mobile food recognition system. In *Proceedings of the 2nd International Workshop on MADiMa*, pages 89–89. ACM.

Wang, Y., Chen, J.-j., Ngo, C.-W., Chua, T.-S., Zuo, W., and Ming, Z. (2019). Mixed dish recognition through multi-label learning. In *Proceedings of the 11th Workshop on Multimedia for Cooking and Eating Activities*, CEA '19, page 1–8, New York, NY, USA. Association for Computing Machinery.

Wu, H., Merler, M., Uceda-Sosa, R., and Smith, J. R. (2016). Learning to make better mistakes: Semantics-aware visual food recognition. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 172–176. ACM.

Yanai, K. and Kawano, Y. (2015). Food image recognition using deep convolutional network with pre-training and fine-tuning. In *2015 IEEE International Conference on Multimedia And Expo Workshops (ICMEW)*, pages 1–6. IEEE.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.