UNIVERSITÄT
BERN

Master of Science in Biomedical Engineering

Technology and Diabetes Management

Classifying meal images into major food categories

by

Luca Ramseyer Khalil Kilani

December 2022

Contents

Contents **iii**

1 Introduction **1**

2 Material and Methods **2**
2.1 Image data 2
2.2 CustomNet 2
2.3 ResNet-50 3

3 Results **4**

4 Discussion **7**

Bibliography **8**

Chapter 1

Introduction

Diabetes mellitus is a group of metabolic disordes, characterized by a high blood sugar level [1]. The disease is highly important, since 536 million people world wide where affected in 2021 [11]. People with this disease need to control their blood sugar level. This includes a close monitoring and assessment of a person's dietary habits. Manual control can be error susceptible and demanding for the patient. Therefore, automatic calculation of the nutrition values of a meal is desired. One possible and convenient way for a patient to determine the nutrition values of a meal, is the automatic calculation based of an image from the meal. However, to perform this calculation, a classification of the food items in the image is required beforehand. In this study we use a Convolutional Neural Network (CNN) [9] for this purpose of classification. A CNN is a type of machine learning (ML) [13] approach, which has gained a lot of attention over the last decade. The CNN is trained with the open source *Food-11 image data set* and learns to classify unseen food images into one of eleven classes. Namely: bread, dairy products, dessert, eggs, fried food, meat, noodles-pasta, rice, seafood, soup, and vegetables-fruits. The data set is used to train a custom CNN, built with PyTorch [6]. In addition, we also used a pretrained CNN to compare the performance of the custom CNN to the pretrained one.

Other research groups have already trained CNN's on open source data sets to recognize specific food in images [3]. However, to our knowledge there is no study where a CNN is used to classify the previously mentioned 11 food classes, for the purpose of nutrition calculation and management of diabetes patients. Image classification can also be done with other ML algorithms such as Random forest classification (RFC) [7] or even unsupervised methods like K-means clustering [2]. However, CNN's are currently considered to be among the most powerful algorithms for image classification. Nevertheless, to train a CNN and reach high performances, a high quantity of labeled data as well as high quality of this data is undeniable.

We hypothesize that a CNN only trained on these specific 11 food classes, can reach higher performance for food classification than a pretrained CNN such as ResNet, Inception or VGGNet. To test this hypothesis we will compare the performance of the custom built CNN, to a pretrained CNN tested on the same data set.

In this study we aim to build a CNN and train it on the Food-11 image data set. We also create so called class activation maps (CAM) [5], which visualize the parts of an image that are decisive for a given class.

Chapter 2

Material and Methods

To classify the food in an image into one of the 11 classes (bread, dairy products, dessert, eggs, fried food, meat, noodles-pasta, rice, seafood, soup, and vegetables-fruits), we used two different CNN architectures. The first one is a customized CNN that was trained with the Food-11 image data set. The second one uses the ResNet-50 architecture and was trained once with the Food-11 image data set and once a set of pretrained weights were used. For the pretrained model, only the linear classification layer has been trained again with the Food-11 data set. Further, we created CAM, which indicate the image regions which are decisive for a given classification.

2.1 Image data

The Food-11 image data set contains 16'643 food images grouped into the 11 previously mentioned food categories. The images are divided into a training, validation and a test data set. In particular, 60% of the images were used for the training set, 20% for the validation set and the last 20% as the test set. However, not every class has the same number of images. The number of images per class are listed in table 2.1.

total:	16643	meat:	2203
bread:	1721	noodles-pasta:	731
dairyproducts:	718	rice:	469
dessert:	2497	seafood:	1502
eggs:	1645	soup:	2497
fried food:	1458	vegetables-fruits:	1169

Table 2.1: Number of images per class

2.2 CustomNet

The CNN we customized with PyTorch, consist in total of four convolutional layers. In between these four layers, ReLU [10] was used as activation function and a $2x2$ Max-pool layer [4] for image size reduction. After the fourth layer, the image with all its feature maps is flattened into a one dimensional linear layer which is fully connected to the output layer.

The structure of this network is shown in figure 2.1.

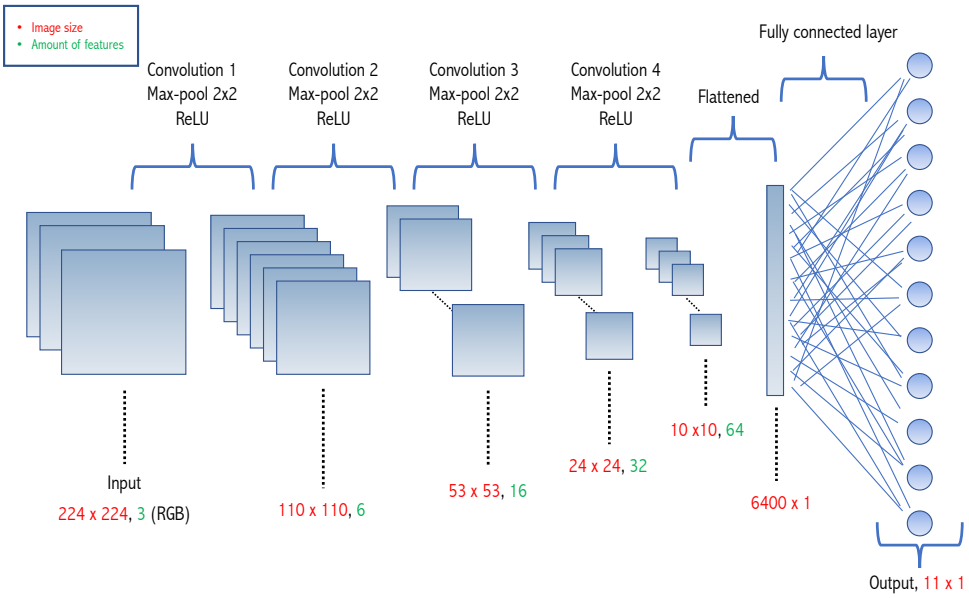

Figure 2.1: Structure of the custom CNN built with four convolutional layers, 2x2 Max-pool, ReLU activation function and one fully connected layer to the 11 output classes.

2.3 ResNet-50

The ResNet-50 architecture is a well established CNN structure and has been used successfully in different applications for image classification and segmentation [12], [8]. There are pretrained weights publicly available from training with the ImageNet database, containing over 1.2 million images. However, ResNet-50 can also be trained from scratch with your own data set.

Chapter 3

Results

Table 3.1 shows the performance of all three models used in this study. The model with the highest performance over all food classes is the pretrained ResNet with an overall accuracy of 86%. The CustomNet and the fully trained ResNet achieve only an accuracy of 31% and 30%, respectively. The class with the lowest performance over all three models is dairyproducts. The CustomNet does not make any right classifications for this class. The fully trained ResNet also performs poorly with 9% and even the pretrained ResNet only achieves a relatively low accuracy with 64%. However, also rice has an accuracy of 0% by the CustomNet as well as in the fully trained ResNet model. The pretrained ResNet still performs well for this class with 95% accuracy.

Figure 3.1 and 3.2 both show the CAM for all 11 classes and all three models, as well as the predicted outcome of each model. Red masked pixels indicate regions on which the model relied highly on its decision. Blue masked pixels indicate the opposite. One can see that the pretrained ResNet is more specific regarding the region its classifications are based on compared to the other two models.

	CustomNet	**ResNet fully trained**	**ResNet pretrained**
Accuracy of the net	31%	30%	86%
Accuracy of bread	16%	14%	82%
Accuracy of dairyproducts	0%	9%	64%
Accuracy of dessert	38%	27%	81%
Accuracy of eggs	21%	24%	76%
Accuracy of fried food	1%	9%	85%
Accuracy of meat	63%	41%	87%
Accuracy of noodles-pasta	5%	1%	98%
Accuracy of rice	0%	1%	95%
Accuracy of seafood	14%	21%	88%
Accuracy of soup	61%	53%	97%
Accuracy of vegetables-fruits	42%	65%	94%

Table 3.1: Accuracy of the models in terms of percentage of correctly classified images.

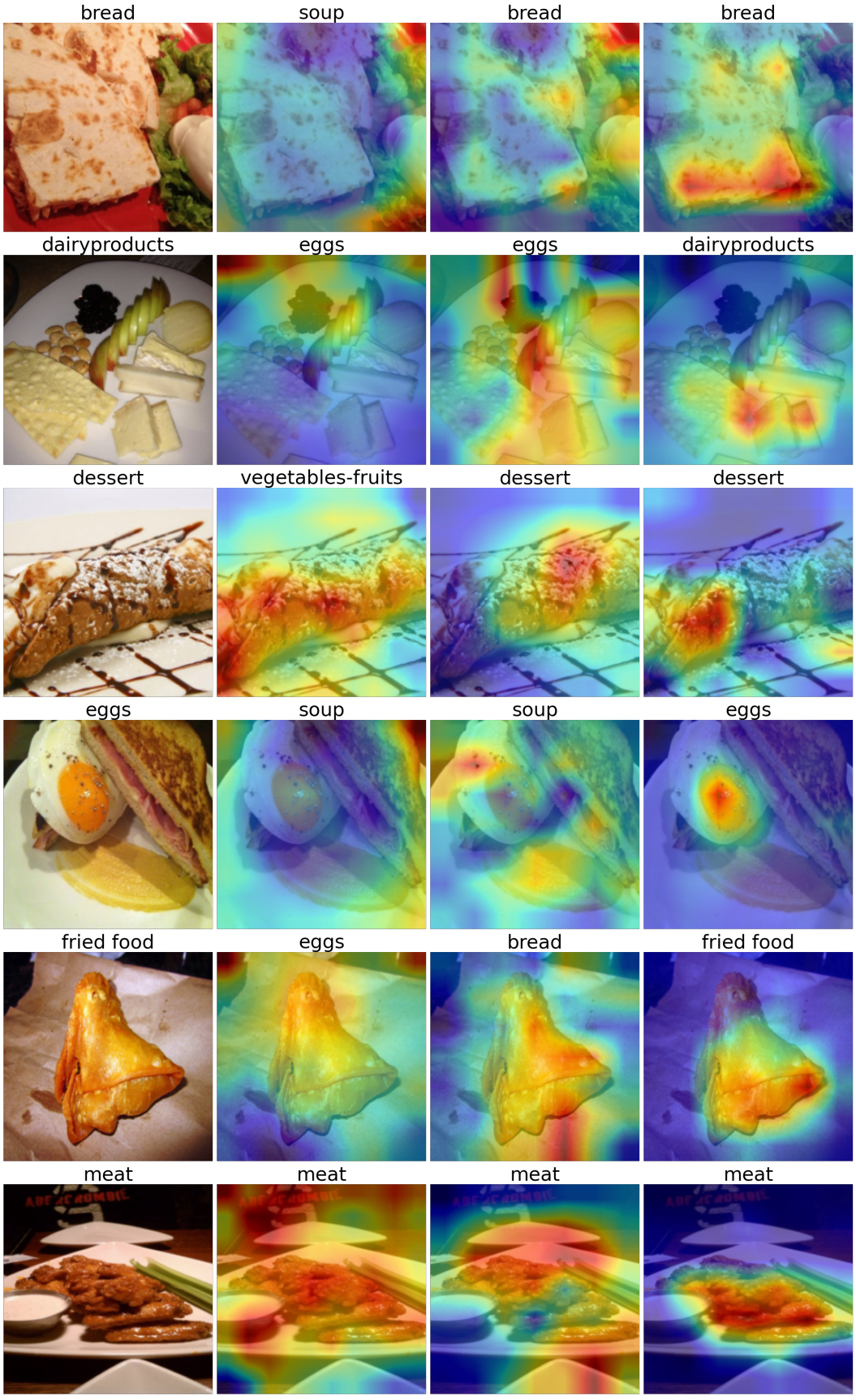

Figure 3.1: CAM and predicted outcome for all three models. From left to right: original image, CustomNet, fully trained ResNEet, pretrained ResNet.

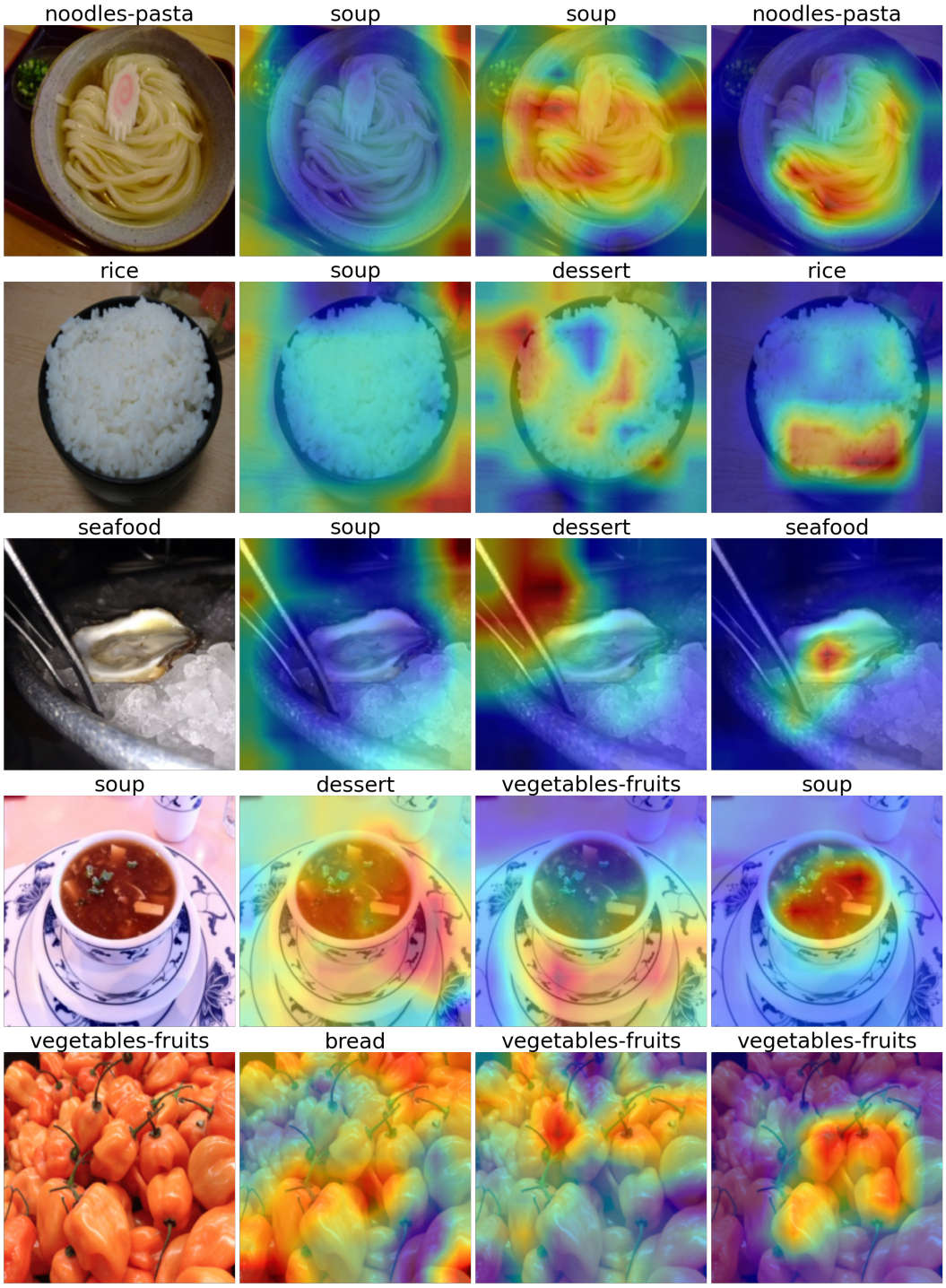

Figure 3.2: CAM and predicted outcome for all three models. From left to right: original image, CustomNet, fully trained ResNEet, pretrained ResNet.

Chapter 4

Discussion

The overall performance of the pretrained ResNet is clearly better with 86% in comparison to the CustomNet with 31% or the fully trained ResNet with 30%. The food class with the worst accuracy for the pretrained ResNEt with only 64%, is dairyproducts. For the other two models the food classes with only 10% accuracy or lower are dairyproducts, fried food, noodles-pasta and rice. To make predictions about the nutrition values of the food consumed by a diabetes patient, the classification of the food is an essential step. Therefore, the model to be used should reach a high performance for each individual food class.

Due to the similarity of the rather poor performance results between the CustomNet and the fully trained ResNet, we can make a conclusion about the data set used to train the models. In comparison to these two models, the pretrained ResNet reached rather high performance accuracy. However, this model was trained with the ImageNet data set, which includes more images. This statement is further reinforced by looking at the amount of images we had for the food classes which achieved an accuracy of under 10%. For dairyproducts, rice as well as noodles-pasta, we had less than 1'000 images. This explains to a certain degree the poor performance for this classes. Based on that we can also conclude that the architecture used for the CustomNet is probably already sufficient for the problem of this study, if one had more images to train with. One interesting finding for the CustomNet is, that most images are being classified as soup as soon as a circular structure occurs in the image. This is probably because the model has learned to recognize soup based on a bowl in the image. This corresponds also to the high amount of soup images (15%) included in the used data set. In total, we see that the model to use for the problem of this study id definitely the pretrained ResNet-50 with an overall performance of 86%.

The strongest limitation in this study is the amount of data as well as the quality of it. The comparison with the pretrained ResNet showed clearly the importance of the quantity and quality of the data used for a CNN. For example the dairyproduct images of the Food-11 image data set, do often times also included other food items. Even for a human it would be difficult to asses only one food class in such images. As next steps we suggest therefore to increase the amount of high quality data. However, based on the problem itself, we need to be able to classify more than one class per image to make proper calculations about the nutrition values of a meal.

Bibliography

[1] M. Blair. Diabetes mellitus review. Urologic nursing, 36(1), 2016.

[2] S. A. Burney and H. Tariq. K-means cluster analysis for image segmentation. International Journal of Computer Applications, 96(4), 2014.

[3] G. Ciocca, P. Napoletano, and R. Schettini. Cnn-based features for retrieval and classification of food images. Computer Vision and Image Understanding, 176:70–77, 2018.

[4] B. Graham. Fractional max-pooling. arXiv preprint arXiv:1412.6071, 2014.

[5] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei. Layercam: Exploring hierarchical class activation maps for localization. IEEE Transactions on Image Processing, 30:5875–5888, 2021.

[6] N. Ketkar and J. Moolayil. Introduction to pytorch. In Deep learning with python, pages 27–91. Springer, 2021.

[7] Y. Liu, Y. Wang, and J. Zhang. New machine learning algorithm: Random forest. In International Conference on Information Computing and Applications, pages 246–252. Springer, 2012.

[8] B. Mandal, A. Okeukwu, and Y. Theis. Masked face recognition using resnet-50. arXiv preprint arXiv:2104.08997, 2021.

[9] K. O'Shea and R. Nash. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458, 2015.

[10] A. D. Rasamoelina, F. Adjailia, and P. Sinčák. A review of activation function for artificial neural network. In 2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI), pages 281–286. IEEE, 2020.

[11] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan, C. Stein, A. Basit, J. C. Chan, J. C. Mbanya, et al. Idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes research and clinical practice, 183:109119, 2022.

[12] Z. Zahisham, C. P. Lee, and K. M. Lim. Food recognition with resnet-50. In 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), pages 1–5. IEEE, 2020.

[13] Z.-H. Zhou. Machine learning. Springer Nature, 2021.