Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 56-2 (2023) 11533–11538

Photo-based Carbohydrates Counting using Pre-trained Transformer Models

Ivan Contreras [*,**] Marti Guso [*] Aleix Beneyto [*]
Josep Vehi [*,***]

[*] Modeling, Identification and Control Engineering Laboratory, Institut d'Informatica i Applicacions, Universitat de Girona, 17003, Spain
[**] Professor Serra Húnter
[***] Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Girona, 17003, Spain

Abstract: Type 1 diabetes is a severe form of diabetes that involves inadequate insulin production by the pancreas. Therefore, patients must take adequate amounts of external insulin to balance blood glucose, where the amount of carbohydrate ingested is the major factor in calculating insulin doses correctly. However, calculating the carbohydrate content of a meal can be complicated by several factors, including patient inexperience. In this work, we propose to devise a system for automatically estimating carbohydrates from images of plated meals. The aim is to reduce patients' computational burden and to reduce the miss-estimates associated with traditional carbohydrate counting methods. The proposal introduces the generation of estimation models for carbohydrate counting using transformer-based neural networks by casting pre-trained models in a more general image context. The models are retrained using the Nutrition5k database, which contains a large diversity of meal samples with multiple examples, a wide range of carbohydrate amounts, and various ingredients. This study presents a detailed analysis of the performance of the implemented models, as well as how the meal composition and size can undermine the estimation models. The metrics show promising results, achieving a 23% error reduction over previous studies.

Keywords: Machine learning; carbohydrate estimation; transformer neural network; diabetes type 1; blood glucose;

1. INTRODUCTION

Type 1 diabetes Mellitus (T1DM) is a chronic disease that involves the destruction of beta cells in the pancreas, resulting in insulin deficiency. Insulin is a hormone necessary to regulate the metabolism of glucose uptake into the blood and maintain normal blood sugar levels American Diabetes Association (2020). Insulin deficiency results in patients with T1DM requiring external insulin doses through rapid-acting insulin injections. These doses should be administered before each meal to compensate for ingested sugars and maintain blood sugar levels within acceptable ranges. In addition, patients may need small doses of insulin throughout the day to maintain normal glucose levels. These doses can be administered as slow-acting doses or through insulin pumps.

The risk of not controlling these blood sugar levels correctly can lead to an excess (hyperglycemia) or deficiency of blood sugar (hypoglycemia). Chronic hyperglycemia is associated with long-term complications resulting in tissue damage and organ failure, decreasing life expectancy, and increased morbidity and mortality. Hypoglycemia can cause neurological deficits such as lethargy, memory loss, and disorientation, and in extreme cases, it can be fatal Nathan (1993). Therefore, management of T1DM is challenging, and therapeutic decision-making must consider several medical and lifestyle factors that must be appropriately adjusted to improve the quality of life of diabetic patients.

There are several factors that influence the amount of insulin to be administered, for example, physical activity. However, along with controlling blood glucose levels, the most important factor is a good estimation of the amount of carbohydrates (CHO) ingested in the last few hours. Thus, counting CHO in meals is crucial for proper glycemic control Evert et al. (2013). Specific training and background about the CHO content of different foods and meals are required for accurately counting CHO. Carbs estimation is difficult for T1D patients, who frequently commit counting errors due to a lack of nutritional information about the meal, inadequate portion size, or inexperience Bally et al. (2016).

The accuracy of CHO estimates made by patients and the impact of associated errors have yet to be thoroughly

[⋆] This work was partially supported by the Spanish Ministry of Universities, the European Union through Next GenerationEU (Margarita Salas), the Spanish Ministry of Science and Innovation through grant PID2019-107722RB-C22 / AEI / 10.13039/501100011033, PID2020-117171RA-I00 funded by MCIN/AEI/10.13039/501100011033 and the Government of Catalonia under 2017SGR1551 and 2020 FI_B 00965.

10.1016/j.ifacol.2023.10.445

quantified. For example, the study by Smart et al. (2009) demonstrated that a ±10 g error in estimating 60 g of CHO made no difference to postprandial glycemic control; however, other studies have reported that patients are making substantially larger miss-estimations. Literature has proposed applications with automatic support for CHO counting, but fewer patients currently use them.

There are several studies that present different solutions to the challenge of analyzing meals from images. Vasiloglou et al. (2018) presents the application GoCarb that integrates a computer vision model using 54 typical dishes of European gastronomy. The study compares the error of professional dietitians and the model trained on the CHO estimation task achieving an average error of 14.9 grams and improving the accuracy of dietitians. A similar study with images of typical Thai cuisine meals (75k images) is presented in Chotwanvirat et al. (2021). The model estimates are compared with those of professional dietitians, providing a mean error of less than 10 grams. The algorithm uses three independent artificial neural networks for ingredient detection, segmentation, and regression. An algorithm based on a convolutional neural network (CNN) for the estimation of labeled macro nutrients in each image (750k images) is presented in Thames et al. (2021). The mean absolute relative error in the CHO estimates is 31.9%. Three different neural networks are proposed to estimate the volume of a food ingredient.

Other works focus on calorie estimation. For example, Liang and Li (2017) presents an algorithm that estimates the food calories in each image (3k images). A CNN is used for ingredient detection. Experiments showed that an average relative error of the volume and calorie estimates does not exceed 20% of the true value. Kong and Tan (2012) proposes a calorie estimation system based on three meal images. The system will classify the foods in the images and then calculate the volume. Evaluation of the system shows that it can estimate the meal volume with a maximum absolute error of 20%. Food volume is the main objective in Dehais et al. (2017), where a system is proposed to calculate the volumes of food on a plate to calculate the size of food portions. The absolute error rate achieved during the evaluation of the system is 8.2%. Other applications are more oriented to the classification of ingredients and dishes. *FoodCam* Kawano and Yanai (2015) is a mobile application aimed at classifying the ingredients of a meal from an image using an support vector machine. The system achieves an accuracy of 79.2% by classifying ingredients into one of the five most likely classes out of 100 possible. In Bossard et al. (2014), a methodology for classifying images of different meals using *Random Forests* is proposed. With the help of the proposed method, an accuracy of 50.76% is achieved in the classification task. This accuracy is much lower than that obtained with CNN models introduced later, which achieve accuracy higher than 95% Foret et al. (2021) Jia et al. (2021). There are also commercial versions, such as the LogMeal application Radeva et al. (2017), that recognize the different ingredients that make up a dish from an image of the dish and provide nutritional information on the detected elements.

To solve this challenging problem, systems based on different techniques have been proposed, most of them belonging to the field of artificial intelligence (AI). More specifically most of them are based on artificial neural networks. This paper presents a proposal for estimating the amounts of CHO from a photo of a plated meal based on pre-trained transformer models in large image databases.

2. METHODOLOGY

2.1 Data-set and preprocessing

The photos analyzed in this paper are from the Nutrition5k database by Thames et al. (2021). This database provides 20000 short videos of 5000 unique dishes made from 250 different ingredients. Each dish provides weight information on the amount of each ingredient and other nutritional information such as the amount of calories, fat, protein, and CHOs. The database is publicly accessible and is available via the internet.

The image dataset was extracted from Nutrition5k using an algorithm that takes snapshots from the videos. The first image captured in the videos was extracted for each meal. As there are 4 videos for each of the meals, four images were extracted for each unique dish. A threshold was also set for the number of grams per dish, discarding those meals containing CHO values above 100 grams. In total, a data set of approximately 17000 images was obtained. The original space color of the images was modified from *bgr* to *rgb*. The amount of CHO related to each photo has been extracted from an Excel file with the macronutrient information.

2.2 Pre-trained Transformers for Carbohydrate Estimation

CNNs have been the dominant method for computer vision applications in the last decade Voulodimos et al. (2018). However, AI methodologies are evolving fast, and new algorithms and network architectures are being devised to improve the performance and the range of applicability. Furthermore, CNNs have shown some weaknesses that may undermine their performance. For example, they have difficulties in relating distant elements of an image, or the CNN filters do not factor the relative position to the image, which results in spatially cluttered elements of an object, causing false detection of the object Bai et al. (2021). In this work, we are applying a deep learning architecture in a shape of a transformer, one of the artificial neural networks architectures that are achieving more success in the latter years. The success of transformers has been such that in many computer vision applications, CNNs are being outperformed and replaced by vision transformers (ViT) Han et al. (2022).

The ViT Vaswani et al. (2017) is an artificial neural network that processes sequences of samples and, through an attention mechanism, could quantify the relationship that each sample has with the rest. Thus, the self-attention mechanism generates visual representations that do not contain the spatial constraints imposed by the convolutions. Instead, they could learn the most appropriate inductive biases depending on the task and the layer position of these mechanisms. Multiple works showed promising results in several reference applications in the computer vision area, for example, in domains such as object detection, video classification, image classification, and image generationCarion et al. (2020); Chen et al. (2021); Zhang et al. (2022); Arnab et al. (2021). Indeed, most of these implementations are well capable of rivaling or surpassing state-of-the-art performances.

On the other hand, our models are based on the concept of transfer learning, i.e., using a model trained on the task to solve a given problem by training it on a different but related task. The idea is to use the basis of a trained model to interpret the images and extract information from them. In this way, it is suggested that the model will be easier to learn the new task while requiring less training time and computational power. The base model uses a ViT-B/16 transformer architecture as an image encoder Dosovitskiy et al. (2020) that has 12 layers and manages 86M of parameters. ViT-B/16 was initially designed for image classification and although other models based on *transformers* are intended for this task, this one was chosen because their good performance in the mentioned task and its relatively simple architecture. The model ViT-B/16 was pre-trained in the classification task on the ImageNet-21k Ridnik et al. (2021) image database. Specifically, this model has been pre-trained on the task of classifying 14M images into one of the 21k possible classes. ImageNet-21k is one of the largest datasets for training classification models of public domain images currently available. The large variety of images endows the models trained with the ensemble with a generalist ability to understand images.

The original images were down scaled and cropped to the center of the meal to preserve the most significant region of the dish. The final resolution of the images used as input was 256x256 pixels. The ViT-B/16 was also modified to serve our proposal. The output of the pre-trained classification model returns a numeric value instead of one of the possible classes. Therefore, the last layer of the model corresponding to the classifier is removed and a single neuron is added in its place. This neuron is responsible for estimating the CHO values of the input image and is connected to each of the different values of the *transformer* output vector.

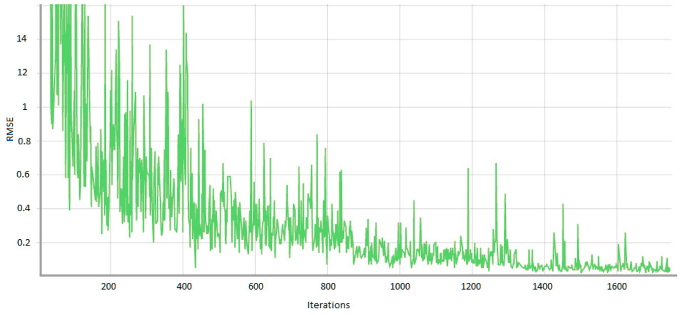

Fig. 1. Evolution of the error curve along an instance of the iterations of the training process.

3. EXPERIMENTS

In this section, we will evaluate the performance of the trained model in the task of CHO estimation from images of plated meals. To assess the robustness of the models, we used a five-fold cross-validation. First, we randomly shuffled the entire data set five times, training and testing the models with 80% and 20% of the data, respectively. An example of the evolution of the prediction error in a training execution over the data set is shown in Figure 1. Then, the average of the performance metrics reported was measured in grams and was obtained to represent the overall performance. Table 1 shows the results of the methodology. The table presents the mean absolute error (MAE), the relative error of the mean of the estimated values with respect to the mean of the CHO values (ER), and the standard deviation (SD) of the CHO predictions of the trained model. Also included are the mean squared error (MSE) and the root mean squared error (RMSE). As seen in the table, the estimation of CHOs of the model obtains promising results, fitting the estimates with acceptable error.

Table 1. Model evaluation results.

Metrics	**MAE**	**ER**	**SD**	**MSE**	**RMSE**
Values	4,65	28,07%	6,30	61,37	7,83

The mean absolute error of the model based on pre-trained transformers networks presented here is lower than the error of the model based on convolutional networks shown in the study performed on the data-set Nutrition5k Thames et al. (2021) reducing the mean absolute error by 23.8%. In order to illustrate some of the estimates, Figure 2 shows several example images from the test data set, each labeled with the actual CHO value and the CHO value predicted by the model.

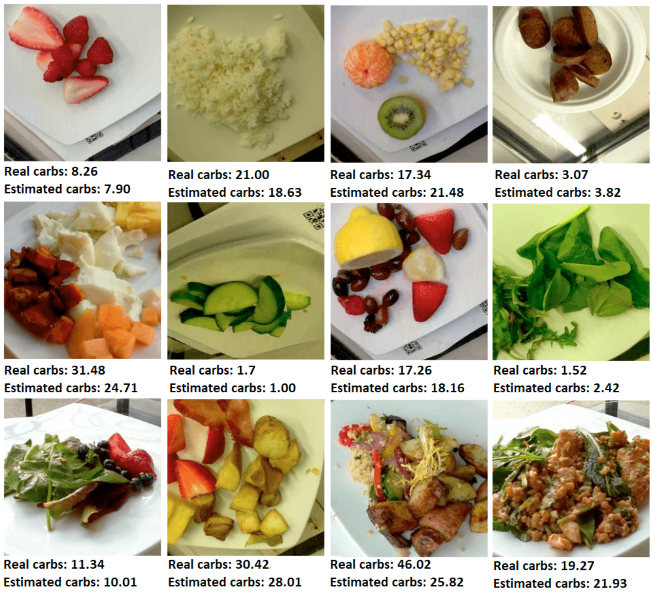

Fig. 2. Images of the data set labeled with the actual value and the predicted value of carbohydrates.

3.1 Miss-estimations and the Quantity of Carbohydrates

Figure 3 analyzes the relationship between the miss-estimation and the amount of CHOs in the meal. The objective of this graph is to check whether dishes with more CHOs make it more difficult to solve the task in the model. The figure shows the relative errors made by the model for each meal as a function of the CHOs present in the meal. In addition, the figure shows how the relative error increases as the dishes contain higher CHO values.

This indicates that meals with less CHOs make it easier for the model to estimate them.

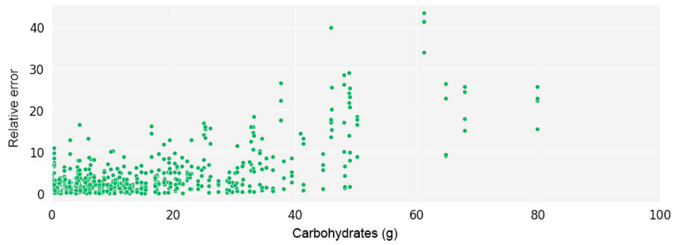

Fig. 3. Scatter plot of the relative error in the estimates depending on the amount of CHOs present in the meal.

3.2 Miss-estimations and the Number and Type of Ingredient

By their nature, some ingredients can make it more difficult for the model to estimate the amount of CHOs in an image than others. For example, their identification difficulty could vary depending on their shape, texture, or color. CHO concentration per unit volume can also be a factor in the performance of the CHO estimation model. It is, therefore, interesting to examine the errors they introduce into the model estimates for each ingredient. To accomplish this, a cumulative sum of all the absolute errors of all the estimates of the images in which the ingredient is present is calculated for each ingredient. Figure 4 displays the 50 ingredients that cause the most error, along with the mean absolute error of the model's estimates for each. The graph indicates that bulgur and chickpeas, the ingredients responsible for the largest errors, are more likely to be present in meals with large errors.

The number of ingredients present in a meal adds complexity to estimating the amount of CHOs. The greater the number of ingredients, the more elements are present in the image, and the more elements must be distinguished for an accurate estimate. So, increasing the number of ingredients could lead to an increase in complexity and thus may cause the model to perform with a different precision than simpler meals. Figure 5 shows the mean absolute error in estimates for meals classified by the number of ingredients they contain. However, the figure does not show a clear correlation between the value of the mean absolute error of the CHO estimates with the number of ingredients present in the meals analyzed.

3.3 Miss-estimations and the Density of Carbohydrates

A higher amount of CHO per unit volume in a particular ingredient may hinder the model in estimating the related CHO. To test this hypothesis, the correlation between the absolute errors of the estimations and the CHO density is examined. CHO density is obtained by dividing the total amount of CHOs in each meal, in grams, by the respective mass, also in grams. Figure 6 shows a scatter plot between the model miss-estimation for each dish and the respective CHO density.

The samples are not evenly distributed with respect to carbohydrate density, however, it can be argued that those dishes with a density lower than 0.1 ratio have a reduced variability and in general a reduction in prediction errors than those that fall between 0.1 and 0.2 ratios. With respect to the higher density meals, there are relatively few instances and it is not possible to draw any conclusions with confidence, although it is likely that by analyzing a larger number of these types of meals, the upward pattern of the error variability would persist.

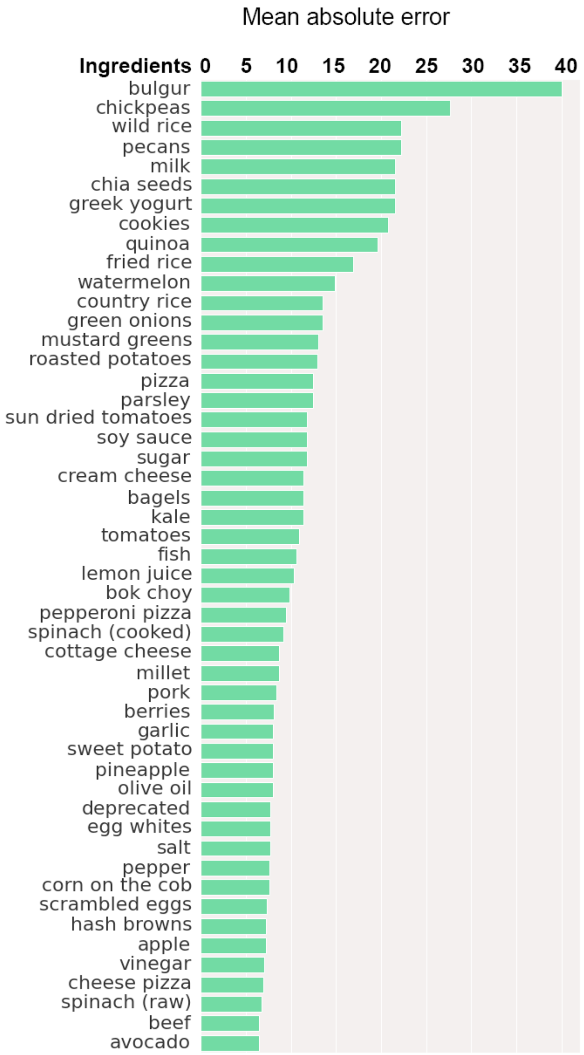

Fig. 4. Mean absolute error of the estimates of CHO in meals in which a specific ingredient is present.

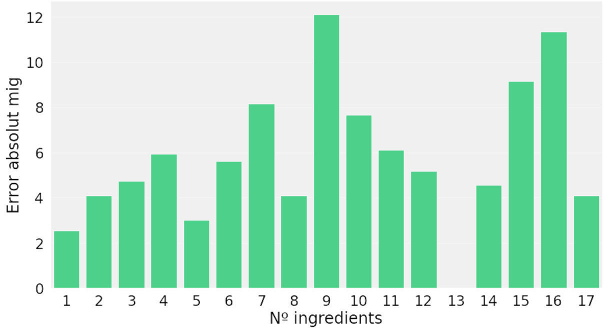

Fig. 5. Mean absolute error of the estimates as a function of number of ingredients present.

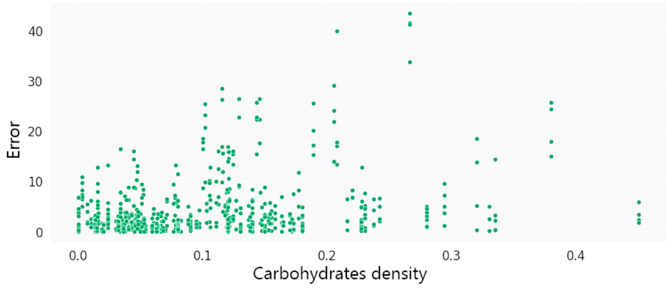

Fig. 6. Absolute error of the estimates as a function of the density of carbohydrates.

4. DISCUSSION AND CONCLUSION

The nature of the proposed problem offers complex problems that are difficult to tackle with a standalone system such as the one proposed. The first problem is to estimate the volume of food on a plate from an image. In many cases, a large part of a portion of food on a plate is covered by the food itself. This can cause significant difficulties in estimating the quantity of portions and the amount of CHOs. Similarly, in a dish with a variety of ingredients, it may be the case that one ingredient is covered by another, which also makes it challenging to estimate portions. However, neural networks have already been shown in the past to handle a volumetric space through two-dimensional images Mildenhall et al. (2020). The results show that the transformer could solve this task quite well.

On the other hand, in the first instance, the segmentation of the photo to find the area where the food is located is another task that needs more focus. In the approach presented here, some of the images in the dataset, a portion of the photographed food portion is outside the margins. This is due to the fact that during the data preparation process, the images are cropped because they have a square aspect ratio. The images taken by the cameras are wider than the height; namely, they have an aspect ratio of 16:9, as they have a resolution of 1920x1080 pixels. On the other hand, the model works with images of size 224x224 pixels. The data set also contains images where, due to the shape of the container and the angle at which the picture is taken, a large part of the food portion is not visible. This could lead to significant errors in the model, as it is not possible to make accurate estimates of the non-visible parts of the dish. Figure 7 shows examples of images from the dataset that presents some of these problems.

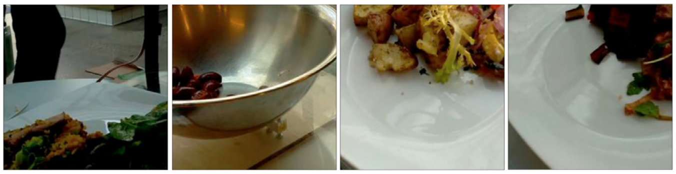

Fig. 7. Examples of images from the dataset with poor cropping

The characteristics and ingredients of meals, in general, can be very varied. For example, gastronomy varies all over the world, and there are different ingredients in the dishes of each one. Although created with the aim of being generalist, the dataset only covers some of the dishes that may exist in the world and some of the ingredients and methods of preparation. For example, the dataset does not contain any images of soup dishes. In practice, it is only possible to create a dataset that would prepare the model to deal with all possible dishes that can be presented to it. Therefore, it must be understood that the problem faced in this work is very complex, and the proposed solution is an approximation.

In this work, a CHO estimation system based on images of plated meals has been implemented to facilitate calculation related to insulin dosing in T1DM patients. A transformer has been applied, a state-of-the-art artificial neural network architecture that promises to improve results in different fields of artificial intelligence compared to previous networks. The results indicate that this architecture is valid for solving the task proposed in work and represents an improvement over solutions created with other techniques. Several analyses were applied to model performance to assess sample characteristics that may make it more difficult for the model to estimate their CHOs.

The field of AI is growing by leaps and bounds, which means that the possibilities of future work has endless branches. Our proposal is focus on transformers and architecture variations will be tested to improve the performance and provide a deeper complexity to the system. Furthermore, planned future work includes to provide additional inputs to be processed simultaneously to improve the accuracy of estimations. The Nutrition5k data set also offers a three-dimensional representation of it, that properly entered into the model, could help quantify portions. In this same line, another piece of information that could prove to be helpful is the angle at which the photograph was taken. On the other hand, Nutrition5K does not provide information about the dimensions of the plate but there are other proposals that use an everyday reference object next to the food to extract an approximation, for example a credit card, which could lead to a difference in performance of the CHO estimation algorithms. Finally, a pre-segmentation of the photo to remove the unnecessary background from the images and feed the CHO prediction model with a more polished image of the plate would benefit the performance of the model.

REFERENCES

American Diabetes Association (2020). Glycemic targets: Standards of medical care in diabetes 2020. *Diabetes Care*, 43(Supplement 1), S66–S76. doi:10.2337/dc20-S006.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6836–6846.

Bai, Y., Mei, J., Yuille, A.L., and Xie, C. (2021). Are transformers more robust than cnns? *CoRR*, abs/2111.05464.

Bally, L., Dehais, J., Nakas, C.T., Anthimopoulos, M., Laimer, M., Rhyner, D., Rosenberg, G., Zueger, T., Diem, P., Mougiakakou, S., and Stettler, C. (2016). Carbohydrate estimation supported by the gocarb system in individuals with type 1 diabetes: A randomized prospective pilot study. *Diabetes Care*, 40(2), e6–e7. doi:10.2337/dc16-2173.

Bossard, L., Guillaumin, M., and Van Gool, L. (2014). Food-101 – mining discriminative components with random forests. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (eds.), *Computer Vision – ECCV 2014*, 446–461. Springer International Publishing, Cham.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, and J.M. Frahm (eds.), *Computer Vision – ECCV 2020*, 213–229. Springer International Publishing, Cham.

Chen, C.F.R., Fan, Q., and Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 357–366.

Chotwanvirat, P., Hnoohom, N., Rojroongwasinkul, N., and Kriengsinyos, W. (2021). Feasibility study of an automated carbohydrate estimation system using thai food images in comparison with estimation by dietitians. *Frontiers in Nutrition*, 8. doi:10.3389/fnut.2021.732449.

Dehais, J., Anthimopoulos, M., Shevchik, S., and Mougiakakou, S. (2017). Two-view 3d reconstruction for food volume estimation. *IEEE Transactions on Multimedia*, 19(5), 1090–1099. doi: 10.1109/TMM.2016.2642792.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Evert, A.B., Boucher, J.L., Cypress, M., Dunbar, S.A., Franz, M.J., Mayer-Davis, E.J., Neumiller, J.J., Nwankwo, R., Verdi, C.L., Urbanski, P., and Yancy, William S., J. (2013). Nutrition Therapy Recommendations for the Management of Adults With Diabetes. *Diabetes Care*, 37(Supplement 1), S120–S143. doi: 10.2337/dc14-S120.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. (2021). Sharpness-aware minimization for efficiently improving generalization.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., and Tao, D. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. doi:10.1109/TPAMI.2022.3152247.

Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision.

Kawano, Y. and Yanai, K. (2015). Foodcam: A real-time food recognition system on a smartphone. *Multimedia Tools and Applications*, 74(14), 5263–5287. doi: 10.1007/s11042-014-2000-8.

Kong, F. and Tan, J. (2012). Dietcam: Automatic dietary assessment with mobile camera phones. *Pervasive and Mobile Computing*, 8(1), 147–163. doi: 10.1016/j.pmcj.2011.07.003.

Liang, Y. and Li, J. (2017). Computer vision-based food calorie estimation: dataset, method, and experiment. *CoRR*, abs/1705.07632. URL `http://arxiv.org/abs/1705.07632`.

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., and Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*.

Nathan, D.M. (1993). Long-term complications of diabetes mellitus. *New England journal of medicine*, 328(23), 1676–1685.

Radeva, P., Bolaños Solà, M., Soriano Oliú, J.L., and Aguilar, E. (2017). Logmeal: Un entorn basat en el reconeixement del menjar per millorar els hàbits saludables. URL `https://logmeal.es/`.

Ridnik, T., Baruch, E.B., Noy, A., and Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. *CoRR*, abs/2104.10972.

Smart, C.E., Ross, K., Edge, J.A., Collins, C.E., Colyvas, K., and King, B.R. (2009). Children and adolescents on intensive insulin therapy maintain postprandial glycaemic control without precise carbohydrate counting. *Diabetic Medicine*, 26(3), 279–285. doi:10.1111/j.1464-5491.2009.02669.x.

Thames, Q., Karpur, A., Norris, W., Xia, F., Panait, L., Weyand, T., and Sim, J. (2021). Nutrition5k: Towards automatic nutritional understanding of generic food. *CoRR*, abs/2103.03375.

Vasiloglou, M., Mougiakakou, S., Reber Aubry, E., Bokelmann, A., Fricker, R., Gomes, F., Guntermann, C., Meyer, A., Studerus, D., and Stanga, Z. (2018). A comparative study on carbohydrate estimation: Gocarb vs. dietitians. *Nutrients*, 10. doi:10.3390/nu10060741.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). "deep learning for computer vision: A brief review". *Computational Intelligence and Neuroscience*, 7068349, 13.

Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., and Guo, B. (2022). Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11304–11314.