Cross-modal recipe analysis for fine-grained geographical mapping of food consumption from images and supermarket sales

A Thesis By

Neele Dijkstra

6926843

Utrecht University
Faculty of Science
Graduate School of Natural Sciences
MSc Artificial Intelligence

May 18, 2022

Main Supervisor Prof. Dr. Albert Salah
First External Supervisor Tinka Koster
Second External Supervisor Marjolein Verhulst

Author's declaration of originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes. I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Neele Dijkstra
6926843

Abstract

The proposed study analyses the applicability of a cross-modal deep learning model trained online recipe data (Salvador et al., 2017) by Li et al. (2020b) to recognise ingredients and their proportional amounts in food images on food images from social media to geographically map food consumption within a city. This new method can provide insights into food consumption at fine spatial granularity, contributing to global health, diminishing world hunger, and providing opportunities for local production and consumption. This study explored whether we can use cross-modal analysis of images and ingredient sets to estimate relative ingredient proportions given images from a specific geographical area. These ingredient amounts are compared to relative consumption rates given by a translated dataset of retail sales of supermarkets within the same region. Specifically, the research focuses on Baku, Azerbaijan, as case study. To study consumption in this region, a new dataset is presented, AzerFSQFood, containing food images from eating establishments around the city. A pretrained cross-modal neural network (Li et al., 2020b) is applied to the novel image dataset, and the results are compared to a new translated supermarket sales dataset, based on an existing Azerbaijani dataset from supermarkets in Baku (Zeynalov, 2020). Unfortunately, the existing cross-modal neural network performs poorly on the AzerFSQFood dataset regarding ingredient detection and relative amount prediction. This indicates that existing cross-modal models are not yet readily applicable to novel datasets, and highlights the need of available models and curated datasets. The study in addition finds a significant correlation between Foursquare images and supermarket sales around Baku in terms of relative ingredient consumption.

Author keywords
deep learning, cross-modal analysis, food consumption, geographical mapping, computer vision, natural language processing

CSS concepts

• **Information systems** → **Multimedia and multimodal retrieval**; **Web mining**; • **Computing methodologies** → **Neural networks**; • **General and reference** → *Evaluation*; • **Human-centered computing** → Visualization;

Acknowledgments

My Master's thesis would not have been here were it not for everyone who helped me. I would like to thank my main supervisor Prof. Dr. Albert Salah(Utrecht University) for his never-ending enthusiasm and ideas; I appreciate all of the input. In addition, I want to thank my external supervisors Tinka Kosterand Marjolein Verhulst(Wageningen Economic Research) for their mental support and being ever-available to read lengthy texts filled with mathematical formulas. Moreover, a big thanks to Arjen Daane (Wageningen Economic Research) for giving me the opportunity to be part of his ambitious research project. I am also forever grateful for the huge amount of work done by Berke Feliz, without whom I would still be lost in a sea of Azerbaijani phrases, and by Afke van Egmond, who nobly volunteered as second annotator. Sasha Kenjeeva probably did not know what she was getting herself into when she said I could send her Russian words for translation; I am sorry and indebted to you! I must also thank Lorenzo Bini for his help with PyTorch. Moreover, I want to thank my second reader Albert Gatt (Utrecht University) for his enthusiastic input; I hope you like the end-result. In addition, thank you to everyone at Wageningen Economic Research for my endless questions, and specifically Serdar Demirel for his Natural Language Processing crash course. Lastly, I want to thank my friends and family, but specifically my partner, who actually carried this project. Thank you, Af, for your never-ending trust in me.

Glossary

application programming interface (API) connecting service to send data from one source to another. 52–54

artificial intelligence (AI) Study of creating an artificial intelligence modelled to human intelligence. 16–18

artificial neural network (ANN) Class of machine learning models for classification and prediction that are shaped as a simplified biological neural network. 19–22, 24–27, 29, 41, 43

convolutional neural network (CNN) Artificial neural networks with convolutional layers. 25–27, 31, 38, 41–44, 49, 50

coverage of ground truth ingredient (CVG) The proportion of ground truth ingredient compounds in the ground truth that are also in the predicted label set. 59

Food and Agriculture Organisation of the United Nations (FAO) Agency of the United Nations that leads international efforts to defeat hunger. 14, 35, 46

gated recurrent unit (GRU) Recurrent neural network that can filter the important parts of input. 39

generative adversarial network (GAN) Artificial neural network that learns to generate artificial examples that closely match the real data. 43

hyperbolic tangent (tanh) Activation function for neural networks. 20

intersection over union (IOU) Intersection of sets over their disjunctive union. 61

long short-term memory (LSTM) Recurrent neural network with long- and short-term memory distinction. 28–30, 41, 50

machine learning (ML) Range of statistical techniques that can “learn" to model classification and prediction problems. 16, 18, 19, 31, 41

natural language processing (NLP) Subfield of artificial intelligence focused on modelling human language. 16, 17

network-in-network (NIN) Artificial neural network consisting of multiple networks. 38

Picture-to-amount (PITA) Convolutional neural network created by Li et al. (2020b) to detect ingredients and relative amounts in images. 50–54, 58, 60, 64, 66–71

rectified linear unit (ReLU) Activation function for neural networks. 20, 50

recurrent neural network (RNN) Artifical neural network with recurrent connections. 27, 28, 42

Rijks Instituut voor Volksgezondheid en Milieu (RIVM) Dutch governmental institute for national health and environment. 12

scale-invariant feature transform (SIFT) Method to determine most distinct parts of images that are invariant across image translation. 36

support vector machine (SVM) Mathematical technique to perform binary classification. 36, 38

Sustainable Development Goal (SDG) Goals by the United Nations to improve human life by 2030. 3, 12

United Nations (UN) Organisation of governments for peace and security. 3, 12

World Health Organization (WHO) Organisation of the United Nations for health and well-being. 12

Table of Contents

Chapter 1

Introduction

1.1 Problem statement

One of the Sustainable Development Goals (SDGs) of the United Nations (UN) is "zero hunger", meaning that by 2030, the UN wants to have ended world hunger (*Transforming our world: The 2030 agenda for sustainable development*, 2015). Needless to say, ending hunger worldwide is essential to ensure the well-being of all people on the planet. Current food consumption has had a large influence on climate change, from emitting carbon-dioxide during production to fossil fuel usage in all steps of food systems (Carlsson-kanyama & Gonza, 2009) . The general consensus among environmental researchers appears to be that food consumption patterns should shift, mainly towards a more plant-based diet, in order to combat climate change (Carlsson-kanyama & Gonza, 2009; Erp et al., 2021). However, more and better insight into food consumption patterns are crucial to achieve these goals. In other words: we need to know what people's food consumption habits are across the world. This can help not only in ending hunger, but in improving people's health and changing their eating behaviour (Erp et al., 2021; Fontanellaz et al., 2019; Mejova et al., 2015; Min et al., 2019; Mouritsen et al., 2017; Schneider et al., 2011). Moreover, it can also contribute to local production and consumption (circular economy), hereby increasing sustainable consumption and helping combat climate change, two of the other SDGs of the UN (Transforming our world: The 2030 agenda for sustainable development, 2015).

In addition, a geographical mapping of food consumption can offer many insights to academia. For anthropology and sociology, the studies of cultures and social behaviour, food is a fundamental factor in human society (Mouritsen et al., 2017). The natural sciences can benefit from food consumption mapping by gaining more insight into the bodily factors involved in the perception of food (Mouritsen et al., 2017).

There is few data on the food consumption habits across the Global South, which can be defined as ‘regions outside Europe and North America, mostly (though not all) low-income and often politically or culturally marginalised *[sic]*’ (Dados and Connell, 2012, p. 12). The data on culinary patterns that are available are often coarse and incomplete. Since the Global South encompasses precisely those countries that have the highest rates of hunger (Chishala et al., 2021), more knowledge on food consumption patterns in the Global South is vital. Moreover, it is important to gather data on food consumption in countries where not much data is yet available, because we want to not only diminish hunger but also stimulate healthy nutrition and eating behaviour, as mentioned. One of these countries where not much data is available, and which is often classified as belonging to the Global South, for example by UNESCO (OWSD & UNESCO, 2022), is Azerbaijan. Therefore, Azerbaijan was chosen as case study for the present research.

1.2 Research approach

1.2.1 Research questions

The importance of more knowledge on food consumption habits worldwide leads to the following question: *How much of each food item do people across the world consume in each region?*. Naturally, this question is too broad to unpack in one research or thesis project. A better option is to focus on a specific region to see whether a certain method can be applied there. Then, perhaps, this same methodology can be employed in future research to eventually study consumption of all food types of all humans. Therefore, I have specified my question into the following:

How can we use cross-modal analysis of Foursquare food images and recipe ingredient sets to geographically map relative food consumption across a specific city?

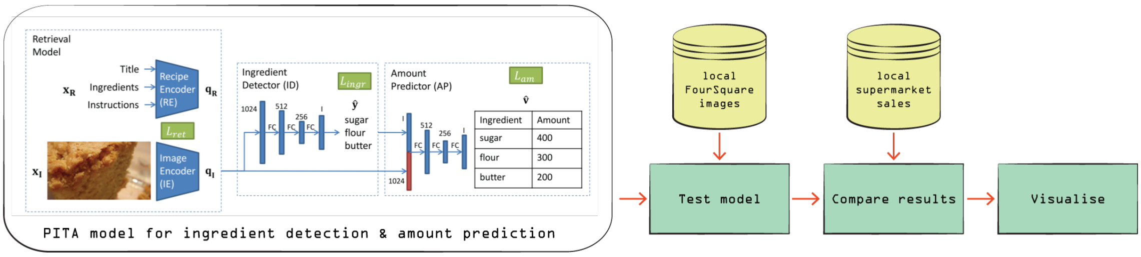

Figure 1. Data sources and methodologies of proposed study. Model architecture (left) directly taken from Li et al. (2020b, p. 10344)

This question is then subdivided into the following sub-questions:

1. How can we re-implement a pretrained cross-modal deep model to recognise the ingredients of a dish from an image?
2. How can we re-implement a pretrained cross-modal deep model to estimate relative ingredient amounts from an image?
3. How can we geographically map food consumption using Foursquare food images?
4. To what extent can the geographically mapped food consumption be validated by supermarket sales data?

1.2.2 Research methods

In the proposed study I aim to geographically map food consumption in Baku, Azerbaijan. To this end I make use of two existing datasets: the Recipe1M dataset made available by Salvador et al. (2017) and a dataset on supermarket sales in Baku (Zeynalov, 2020). In addition, I collect Foursquare images of food taken in Baku. All of these datasets are featured in Figure 1. The model made available by Li et al. (2020b) is applied to subsets of the images from Foursquare, divided according to region in Baku. Lastly, the usage of Foursquare images for food consumption estimation is validated using the dataset on supermarket sales in Baku.

In case the model indeed works as suspected, the same technique could be applied to images (from Foursquare, social media, or otherwise sourced) from other Global South cities and countries to map food consumption in these regions.

1.3 Contributions

The proposed study has the following contributions to the literature:

- A novel AzerFSQFood dataset, containing food images and associated locations from Azerbaijan, collected from Foursquare.
- Mapping of geographical food consumption that is more fine-grained than in previous work, namely on the sub-city geographical scale.
- A novel combination of ingredient detection, relative quantity estimation, and geographical mapping of food consumption.
- Demonstration of usefulness of online available image-location pairings for studying sociocultural processes.
- Insights into relative ingredient consumption in different neighbourhoods of Baku, Azerbaijan.

These contributions are of practical significance in providing organisations with a flexible and robust method of mapping food consumption for any region where food images are being posted on social media. The proposed method, analysing social media images, is less time-consuming than for example household surveys, a method now widely in use. The Foursquare images function as a proxy for social media data from other networks, which would ideally be available. The insights gained from applying the proposed model to different locations in turn helps to monitor and benefit nutrition, as mentioned, and can eventually help optimise production. This could for example stimulate countries to produce the products they consume the most or stimulate better food production in an international setting. In addition, the data can inform agriculture and food organisations on where to sell their products. Knowledge on food consumption combined with information on food production can help steer the economy towards circularity, in turn helping the environment.

The present study finds that the pretrained model used has low performance on the novel AzerFSQFood dataset. This can be due to the difficulty of the annotation task regarding food images and ingredient labels, or due to the quality of the pretrained model. Regardless, it highlights both the need for publicly available abstracted deep learning models for ingredient and amount detection and the need for publicly available, well curated datasets. In offering one novel and one heavily revised version of a dataset, respectively AzerFSQFood and the English Baku supermarket dataset, this study contributes to solving these issues. The study in addition shows that there is a moderate correlation between ingredients bought in supermarkets and food consumed in restaurants in Baku. This relationship exemplifies the usefulness of these data sources when mapping ingredient quantity consumption, but its moderate nature shows that there is a discrepancy between ingredients consumed in supermarkets and eating establishments. Alternatively, the discrepancy could be due to the translated and revised nature of the supermarket sales dataset.

1.4 Thesis structure

In the following chapters of this thesis document, I first detail my literature review, in Chapter 2. Within this related work chapter, I go over research on food consumption mapping in Section 2.1, detail the basics of machine learning methods used in computer vision in Section 2.2, and review existing food recognition models as well as explain the choice for an existing model for the present study in Section 2.3. Subsequently, I expand on the case study of the present research, Azerbaijan, in Chapter 3. Then, the methodology is explained in Chapter 4, and the results are provided in Chapter 5. Lastly, the discussion and conclusion are given in Chapter 6.

Chapter 2

Related work

2.1 Food consumption research

Food science is a very broad discipline, ranging from the social sciences studying for example healthy eating behaviour and cultural cuisines to the natural sciences studying food composition and sustainable food production (Deutsch & Miller, 2007). Due to the rise of big data, a new field in food research has developed, termed food computing (Hao et al., 2019; Jiang & Min, 2020; Min et al., 2019). Min et al. (2019) define this field as follows: '*Food computing applies computational approaches for acquiring and analysing heterogeneous* [sic] *food data from disparate sources for perception, recognition, retrieval, recommendation, and monitoring of food to address food-related issues in health, biology, gastronomy, and agronomy.*' (Min et al., 2019, p. 3). In other words, this field uses data-driven computational research to study food (Jiang & Min, 2020; Min et al., 2019), which can be done by studying a multitude of possible food interaction factors. Since the current study focuses on mapping food consumption geographically in a data-driven manner, I will only go into the main factors influencing geographical differences in food consumption. These factors can in turn inform us on the factors we need to take into account when geographically mapping food consumption. In addition, I will review the most common measures and data types used to map food consumption. From these measures, we can then take the most promising option.

2.1.1 Geographical consumption factors

In their research on Chinese cuisines, Zhu et al. (2013) discuss both geographical proximity and climate as possible correlating factors. After calculating correlations between ingredient usage in online Chinese recipes, they conclude that while climate (when controlled for geographical location) has no significant effect on ingredient usage, more geographical proximity is associated with more similar cuisines.

Wagner et al. (2014) studied spatiotemporal factors in food preferences. They analysed online recipes from Austria, computing the similarity of recipes and ingredients, as well as the correlation between geographical distance and recipe and ingredient similarity of regions. They found that indeed regions that lay closer together have similar food preferences and food preferences are different in the weekend than during the week.

These studies exemplify that there is a correlation between geographical proximity and food preference. This finding could be attributed to a multitude of factors. It is apparent that geographically close regions are more likely to share cultures, climate, economic status, and food production rates. Therefore, any of these factors could be exerting influence on eating behaviour.

2.1.1.1 Culture

There exists a wide variety of cuisines and culinary practices. This variety of cuisines is very closely linked to human society and cultures (Ahn et al., 2011; Mouritsen et al., 2017). To geographically map food consumption, therefore, one option is to study cultural practices. Similar eating behaviour in itself has in fact been deemed a "food culture" (Zhu et al., 2013)[1]. Anthropological and sociological research has studied the varied eating behaviours of different cultures and religions. (Musaiger, 1993), for instance, found that a variety of sociocultural factors such as religion, beliefs, and education have an influence on food consumption habits in Arab countries. (Nicolaou et al., 2009) studied Dutch inhabitants with Turkish and Moroccan heritages and concluded that cultural factors such as food-related hospitality exert an influence on food consumption.

In data science, knowledge on cultural eating behaviours is sometimes used as a starting point for further analysis. In their study on fish sauce consumption in

Japan, Nakano et al. (2018) for instance choose to focus on fish sauce based on the knowledge that this is an often consumed product in Japan. Similarly, Olsen et al. (2007) study fish consumption attitudes between countries, based on the knowledge that fish consumption attitudes differ between other cultures. Zhu et al. (2013) focus on similarities and differences between Chinese cuisines since they know that China houses a wide variety of cultures.

Knowledge on cultures and religions could also be used to validate data-driven research on food consumption. For instance, when studying food consumption in Saudi Arabia, a mostly Islamic culture, finding high numbers of pork consumption could be reason to doubt the validity of the analysis. In contrast, finding drops in consumption rates during Ramadan would be an expected finding.

2.1.1.2 Food production

It could be expected that food production would have an influence on food consumption, meaning that in regions where certain ingredients are produced more than others, a similar trend would be seen in terms of consumption of those ingredients. However, this relationship would only hold if consumers are indeed eating food produced in their geographical region. While globalisation is ever increasing, this is not often the case. For example, in 2018, food consumed in the Netherlands was for 75% grown abroad (Muilwijk et al., 2018). At the same time the country is the most important source of food agricultural products for seven other countries in the European Union (Jukema et al., 2021).

Research on local food consumption shows that there is a plethora of factors influencing whether consumers buy locally produced food (Bianchi & Mortimer, 2015). In addition, the factors that determine whether consumers intend to consume locally produced food are region-specific (Bianchi & Mortimer, 2015). Factors that have been found to influence local consumption behaviour are attitudes towards supporting local agribusinesses, attitudes towards supporting the local community, consumer ethnocentrism, perception of consumers on their societal debt, and the type of food bought (regular or special) (Bianchi & Mortimer, 2015; Blake et al., 2010). This large amount of factors determining whether consumers eat locally produced food underlines that the relationship between food production and food consumption in a given area is unsteady at best.

2.1.1.3 Economic status

It seems apparent that there is a link between the average economic status of a region and consumption. Some food products are more expensive than others, and more food consumption generally increases the price of eating. Indeed, multiple studies have found that income is an important (if not the most important) factor in food consumption (Chauvin et al., 2012; Choudhury et al., 2016; Musaiger, 1993; Nicolaou et al., 2009; Schneider et al., 2011).

In addition to there being a direct link between absolute income available for food and the price of edible products, relative income can also have an influence on eating behaviour (Nicolaou et al., 2009). In their study on Dutch inhabitants with Moroccan and Turkish roots, Nicolaou et al. (2009) found that since the Turkish participants living in the Netherlands *perceived* themselves as having higher income than people who stayed in Turkey, they felt able to spend more money on food.

The effect of average income in a region on eating behaviour is very apparent when it comes to food deserts. Food deserts are areas, be it a small town or a city's neighbourhood, where there is little access to fresh and healthy food that is also affordable (Choudhury et al., 2016). These areas are often also socioeconomically disadvantaged (Choudhury et al., 2016), exemplifying the fact that a lower income makes keeping a healthy diet more difficult.

Since there is an influence of socioeconomic status on food consumption, there exists a link between ethnicity and food consumption. This link has been getting the most attention in the form of "white veganism", a term that describes how veganism can be seen as a diet for the elite, requiring wealth and whiteness (Greenebaum, 2018). However, there have also been studies in which vegan people of colour argue that this is a strategy by non-vegans to make veganism seem less attainable to combat societal dietary changes (Greenebaum, 2018).

2.1.1.4 Time

Naturally, we do not eat the same meals every day. Our meals might vary throughout the week: On a workday we might eat sandwiches and pasta bolognese, whereas in the weekend we might order pizza. In addition, we might eat differently depending on the season (salad in summer and stew in winter) or whether it is a special occasion (turkey on Thanksgiving).

Wagner et al. (2014) analysed weekly temporal changes in food preferences and found that food preferences during workdays are different than during the weekend (see Section *geographical environment*). In addition, they found that food preferences change over the course of the week, with Thursday and Friday showing an increase of food preferences more similar to the weekend, and food preferences abruptly going from weekend to weekday preferences from Sunday to Monday. Kusmierczyk et al. (2016), who studied online recipe production patterns, observed a similar pattern, with ingredient preferences varying during the week.

West et al. (2013) studied web usage logs of participants searching for online recipes and found temporal changes in recipe selections. Firstly, they found that over the course of a year, recipe selection choices fluctuate around an equilibrium. Similarly, Kusmierczyk et al. (2016) found seasonal patterns in ingredient consumption. Secondly, within a month, recipe selection varies greatly (West et al., 2013). Thirdly, days that fall on or around a holiday show patterns diverging from the norm (West et al., 2013).

2.1.1.5 Food policy

A last factor with a clear intended effect on food consumption is food policy. Governments and international organisations often create guidelines for healthy consumption, such as the Dutch Voedingscentrum (Food Centre)'s food advice called the "Schijf van Vijf" (Voedingscentrum, n.d.) and the World Health Organization (WHO)'s healthy diet advice (Who, 2020). In their aforementioned SDGs the UN also advocates for more food policies (*Transforming our world: The 2030 agenda for sustainable development*, 2015). It should be noted that these food policies are not always solely based on research on nutrition, but that they can also be shaped by environmental and political factors. This is shown in the fact that the Dutch government recently removed a recommendation of eating less meat for the environment, since this recommendation was deemed too controversial (Dinther, 2021).

There exist many different forms of food policies: they can focus on recommending healthy diet choices, as mentioned, but can also come in the form of taxes on unhealthy products such as tobacco. The great variety of food policies means that their effects might also vary greatly. A study conducted by the Dutch Rijks Instituut voor Volksgezondheid en Milieu (RIVM) showed that Dutch citizens often eat less than the recommended amount of the products in the "Schijf van Vijf", and that they often eat many products that are not recommended (Rivm, 2020). This seems to indicate that the healthy diet recommendations are not effective. Watt et al. (2020), on the other hand, argue that banning price promotions of unhealthy food would stimulate a reduction in unhealthy consumption behaviour. In short, the wide variety of food policies potentially all have different effects on food consumption.

In sum, a variety of factors influence food consumption differences, among which culture, food production, economic status, time, and food policies. This means that it is very likely that food consumption differs on a small geographical scale: not only on the country-level, but also sub-country and possibly even sub-city. Therefore, for a representative mapping of food consumption, it is important to aim for an analysis with high spatial resolution.

2.1.2 Measures and data on food consumption

In addition to the multidimensional factors interacting with eating behaviour, there is a multitude of measures available for conducting data-driven research on spatiotemporal differences in food consumption. In previous studies, researchers have attempted to use household surveys (Can et al., 2015; Olsen et al., 2007; Verbeke & Vackier, 2005), production and export data (FAO, 2022), and retail data (Herranz et al., 2017; Min et al., 2019) to geographically map food consumption. However, household surveys are expensive and time-consuming. In addition, they quickly become irrelevant when societal and food patterns change. Similarly, production and export data as used by FAO (2019) are spatially not fine-grained: they only show nation-wide statistics. Therefore, a shift towards time-sensitive, inexpensively accessible online data is imminent. In various studies, researches have tried mapping food consumption by means of online recipe data (Ahn et al., 2011; Nakano et al., 2018; Said & Bellogín, 2014; Wagner et al., 2014; Zhu et al., 2013). However, for many countries in the Global South, these data are not readily available. Another, perhaps better option, is to focus on social media data. Abbar et al. (2015) and Choudhury et al. (2016) have previously aimed to study food consumption by looking at data from social media platforms Twitter and Instagram.

2.1.2.1 Production and export data

One of the most widely used resources is the Food and Agriculture Organisation of the United Nations (FAO)'s balance sheets (FAO, 2019). The FAO calculates these sheets based on the production, export, and import data of food products of a given country (FAO, 2022). A benefit of this method is that the production, import, and export data are readily available: Food consumption is inexpensive to compute in this way. However, these data are not available with higher spatial resolution than the country level, and is a very rough measure of actual consumption. Chapter 3 details the FAO balance sheet data for the case study of the current research.

2.1.2.2 Household surveys

A measure with a higher spatial resolution and a more direct connection to absolute consumption values is the use of household surveys. Various authors have used this method to map food consumption (Can et al., 2015; Olsen et al., 2007; Verbeke & Vackier, 2005). Verbeke and Vackier (2005) asked 429 respondents to answer questions on their attitudes, social norms, and perceived behavioural control regarding fish consumption. Can et al. (2015), too, studied fish consumption, but only included 127 randomly selected participants. Olsen et al. (2007), on the other hand, sampled about 4000 participants using a mixture of in-house interviews, mail and web-based surveys to study fish consumption.

Household surveys ensure that there are accurate, household-level data on the ingredients people consume. Unfortunately, this measurement is expensive as it requires surveys that are often performed in-person and on-site due to communication constraints (Min et al., 2019). This means that this method can only be performed on small participant samples (Min et al., 2019). Moreover, data collected through household surveys are time-sensitive and will not update easily. In addition, household surveys put a burden on the respondents of those surveys. They take up the time and effort of people, some of whom might be struggling already. Often, respondents are not even compensated. Moreover, household surveys suffer from non-response bias: people that refuse participation might share a characteristic (e.g. consumption habits out of the norm) that is overlooked in the study (Singer, 2006). Lastly, household surveys might provide messy data, especially in the case of interviews: unstructured interview data require a lot of time for coding and organising. All in all, household surveys are suitable for accurate and specific data of food consumption in certain regions, but less so for mapping food consumption in a time-robust and inexpensive manner across a larger geographical area.

2.1.2.3 Retail data

Food consumption can also be measured through retail data, be it from grocery stores or hospitality businesses. For instance, data from supermarket chain Walmart have been used to study consumption behaviour changes after natural disasters (Lee & Kang, 2015). Some studies use restaurant-specific data such as location and menu to perform food recognition on related food images (Min et al., 2019). For example, Herranz et al. (2017) study photos taken in restaurants and posted on social media. By taking the name, location, and menu of the restaurant into account, they create a shortlist and a probability model to recognise the dish, the restaurant, or the location. They show that restaurant information can be beneficial in all of the aforementioned tasks. Their study exemplifies that geographical information on food consumption can benefit from retail data, but that these data sources can also be combined to study food consumption. In the present study, I use data from both eating establishments and supermarket sales, as detailed in Chapter 4.

2.1.2.4 Online recipes

While in previous times most food studies still focused on household surveys and other small-scale data, since the rise of online big data, food computing is moving towards studying large scale online data sources (Min et al., 2019). Studying food consumption through online recipes has previously been attempted in various studies (Ahn et al., 2011; J. Chen & Ngo, 2016; Erp et al., 2021; Kusmierczyk et al., 2016; Min et al., 2019; Nakano et al., 2018; Said & Bellogín, 2014; Salvador et al., 2019; Wagner et al., 2014; Zhu et al., 2013). Online recipes are easy to find and process (Mouritsen et al., 2017). Secondly, they contain a lot of metadata, such as ingredients, categories, tags, user information, ratings, comments, images, nutritional value, flavour and cuisine (Min et al., 2019; Mouritsen et al., 2017). This means that online recipes can be used to study a variety of factors related to food consumption.

Firstly, they can be used to study spatial factors of food consumption. Ahn et al. (2011) study flavour networks: which flavours are likely to appear in the same dish, depending on the culture in which the dish is produced? To answer this question, they analysed online recipes from American and Korean recipe websites. They performed a network analysis of ingredients and their accompanying flavour compounds, and then checked for each of the recipe clusters whether their ingredients share compounds from the flavour network. They concluded that North American and Western European cuisine indeed does tend to include ingredients in recipes that share flavour compounds, whereas East Asian and Southern European cuisine shows the opposite pattern. Wagner et al. (2014) and Zhu et al. (2013), similarly, studied spatial patterns in food consumption and both found that geographically close regions share more similarities in food consumption patterns. Secondly, temporal patterns of food consumption can be deducted from analysing online recipes. Kusmierczyk et al. (2016) studied social and temporal factors involved in German recipe generation. Correspondingly, Wagner et al. (2014) studied temporal factors in Austrian online food preferences. Both studies found weekly temporal changes in eating preferences. Nakano et al. (2018) also studied temporal patterns in food consumption, but rather than looking at weekly patterns across all food consumption, they focused on yearly patterns in a specific product, namely Asian fish sauce consumption in Japan. Through studying online recipe search behaviour, they found an increase in fish sauce consumption in Japan. Thirdly, the images often included in online recipes can be analysed to understand food patterns. In the study of Salvador et al. (2019), they trained a machine learning (ML) model on online recipes to generate a recipe based on an image of a dish, for example. J. Chen and Ngo (2016) performed a similar analysis in creating a deep learning model to retrieve recipes from images. Both models yielded good results, with F1 scores of respectively 49.08% and 67.17%. Lastly, online recipes can be used to track health of users consuming and producing these recipes. Said and Bellogín (2014) tracked user interactions on a popular online recipe website and found significant differences between the interaction patterns of healthy and unhealthy website users.

Despite all these possible applications for online recipe analysis, Erp et al. (2021) argue in their review on natural language processing (NLP) and artificial intelligence (AI) methods for analysing food and recipes that there are some shortcomings of the method. They argue that online recipe analysis often requires collecting raw data from a variety of private sources. This means that data are noisy and require preprocessing, and connections between databases are weak. Furthermore, they argue that online recipe data are not standardised, meaning that quantity expression ranges from numerals and fractions to being spelled out, and includes different units. Lastly, they argue that recipes often do not include portion sizes or number of people, making it difficult to convert a recipe to consumption per capita. All these issues make that online recipe analysis requires a lot of preprocessing. In addition, online recipes show a bias towards complicated or special dishes: you are much less likely to find the recipe for tiramisu than for a cheese sandwich, for example. This means that not all food consumed is present in the data. Moreover, web-scraping is an ethically contested method for data collection, regarding privacy and copyright (Krotov & Silva, 2018).

Online recipe analysis also has advantages, in that the datasets are easily updated and are available for many geographical regions at once, large amounts, for some sites even more than two million (Mouritsen et al., 2017). In addition, it is suitable for many applications, as discussed. Furthermore, as the world wide web is still developing, online recipes will most likely become even more abundant and through the use of NLP and AI, linking of databases will become easier (Erp et al., 2021). In sum, online recipes have a lot of potential in food consumption analysis, with their main disadvantages being noisy data and lack of standardisation. Therefore, they are most useful for studying food consumption on a relative and population basis, and less for absolute and individual measures of eating behaviour.

2.1.2.5 Social media

Social media can also be used to study food consumption. Social media contain data on nearly every aspect of human behaviour, food not excluded: in 2019, over 300 million posts could be found using the hashtag "#food" (Salvador et al., 2019). And that covers only the posts that have this specific English hashtag. Many people use social media everywhere and at every time, meaning that it can give insights into eating patterns at every moment around the world, providing new ways to analyse geographical patterns in food consumption (Mouritsen et al., 2017). For example, as more and more people in the Global South start using social media, it offers a good alternative for studying food consumption to online recipe data, which can be sparse for regions in the Global South. In addition, organisations can use social media analyses to acquire business insights Tao et al. (2020).

Inferences from social media about food consumption can be made using different techniques. For instance, Abbar et al. (2015), Choudhury et al. (2016), Mejova et al. (2015), Ofli et al. (2017), and Sharma and Choudhury (2015) all studied health in relation to food posts on Instagram or Twitter. However, whereas Abbar et al. (2015) and Sharma and Choudhury (2015) studied textual data, Mejova et al. (2015) and Ofli et al. (2017) studied visuals from social media. These data sources come with their own opportunities and challenges, among which working with big data, NLP, computer vision, and ML techniques (Mouritsen et al., 2017). However, advances in AI and ML will probably create more opportunities for studying online big data such as social media.

Social media analysis comes with similar issues as online recipes, such as noisy data, representation bias of complex and aesthetically pleasing dishes, and ethically questionable web-scraping. Moreover, not all people might have access to social media or internet in general, causing a socioeconomic bias in social media data. Nevertheless, social media data could be a good starting point for detecting general geographical trends in food consumption. On the whole, on the whole, social media can be used to study food consumption analogous to the analysis of online recipes, but since they are used more temporally continuously and widespread, they offer more opportunities for studying data-sparse regions.

In conclusion, because of its inexpensiveness, flexibility, and widespread availability, social media data are useful in geographical food consumption mapping. In addition, a combination of social media data with online recipes can provide us with even more information. Therefore, the present study applies deep learning methods created with online recipe data on social media data to perform fine-grained spatial mapping of food consumption, as detailed in Chapter 4.

2.2 Machine learning for computer vision

ML is the study of computer science models and algorithms that can "learn" from data. The field has resulted in a wide variety of methods that can be categorised in different ways. Firstly, there is a distinction between supervised and unsupervised learning methods. Secondly, some methods are more suitable for classification or retrieval problems, whereas others are mostly applicable to regression or prediction problems. Thirdly, different algorithms are optimal for different media, such as textual or visual analysis. In addition, a popular sub-class of ML methods are deep learning models (Rina Dechter, 1986). Deep learning models are essentially a more extended version of ML models, with more depth and thus higher capabilities of abstraction and learning (LeCun et al., 2015). This means that deep learning models require less manual preprocessing and that their results are often better, but that they are also more computationally expensive (LeCun et al., 2015). In the rest of this section, I describe ML methods that are in practise often applied as deep learning models. In this section, I limit my description to the ML methods that were encountered in selecting an ingredient detection model to be applied in the present study, as not all methods from the broad ML discipline are relevant.

2.2.1 Artificial neural networks

Artificial neural networks (ANNs) are ML algorithms and the original form of neural networks. At the core of an ANN lies the perceptron (Rosenblatt, 1958). Perceptrons are roughly built in the image of a human brain, consisting of a network of nodes that can be seen as the artificial counterparts of biological neurons. Each of these nodes takes an input, x, that is altered by a function $f(x)$ resulting in output y. Depending on the shape of the ANN, this output is then passed on to different nodes. For example, the network in Figure 2 consists of an input layer, a hidden layer, and an output layer. The nodes in the input layer take the input x, apply some function to it and pass the result on to the hidden layer. The hidden layer then repeats this step, and finally, the output layer produces an end result.

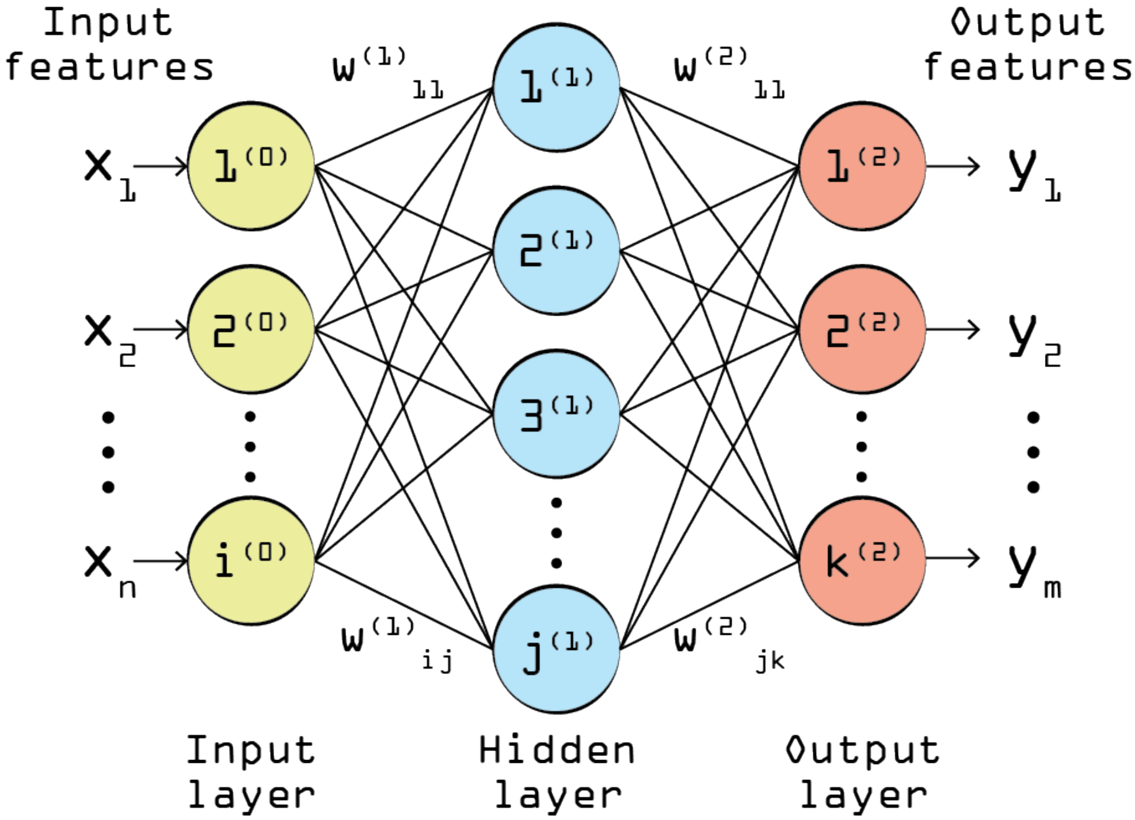

Figure 2. Schematic visualisation of an ANN with 1 input layer, 1 hidden layer, and 1 output layer, based on Bre et al. (2017, p. 4).

For a connection between a node i in a layer of n nodes, where i is any $\mathbb{N}$ between

1 and n, and a node j in a layer of m nodes, where j is any $\mathbb{N}$ between 1 and m, the activation of node i a_i is multiplied by the weight of the connection w_{ij}. The activation of node j then is the sum of all these calculated connections from the previous layer, inputted into an activation function g. This activation function can for example be the sigmoid (Equation 2.1), hyperbolic tangent (tanh) (Equation 2.2), softmax (Equation 2.3), or rectified linear unit (ReLU) (Equation 2.4) function (Nwankpa et al., 2018). The sigmoid, tanh, and ReLU functions were all partly introduced to account for non-linearity, but each of these activation functions is suitable for different types of classification and prediction problems.

$$\sigma(x) = \frac{1}{1+e^{-x}}, \tag{2.1}$$

where x is $\Sigma_{i=1}^{n} a_i w_{ij}$.

$$f(x) = \left(\frac{e^x - e^{-x}}{e^x + e^{-x}} \right) \tag{2.2}$$

$$f(x_i) = \left(\frac{\exp(x_i)}{\Sigma_j \exp(x_j)} \right) \tag{2.3}$$

$$f(x) = \max(0, x) = \begin{cases} x_i, if x_i \geq 0 \\ 0, if x_i < 0 \end{cases} \tag{2.4}$$

Figure 3 depicts an example of the calculation process within an ANN. This particular ANN gives a prediction for how likely it is that it will rain in the next hour. This prediction is made based on a score of 0 to 1 for each of the three features: wind strength, cloudiness, and temperature. The feature values for this moment of the day are inputted into the nodes of the input layer. Subsequently, to calculate the value of node $1^{(1)}$, each input layer node's activation $a_i^{(0)}$ is multiplied by connection weight w_{i1} and the resulting numbers are summed. The result is then inputted into a sigmoid activation function, which gives the activation for node $1^{(1)}$.

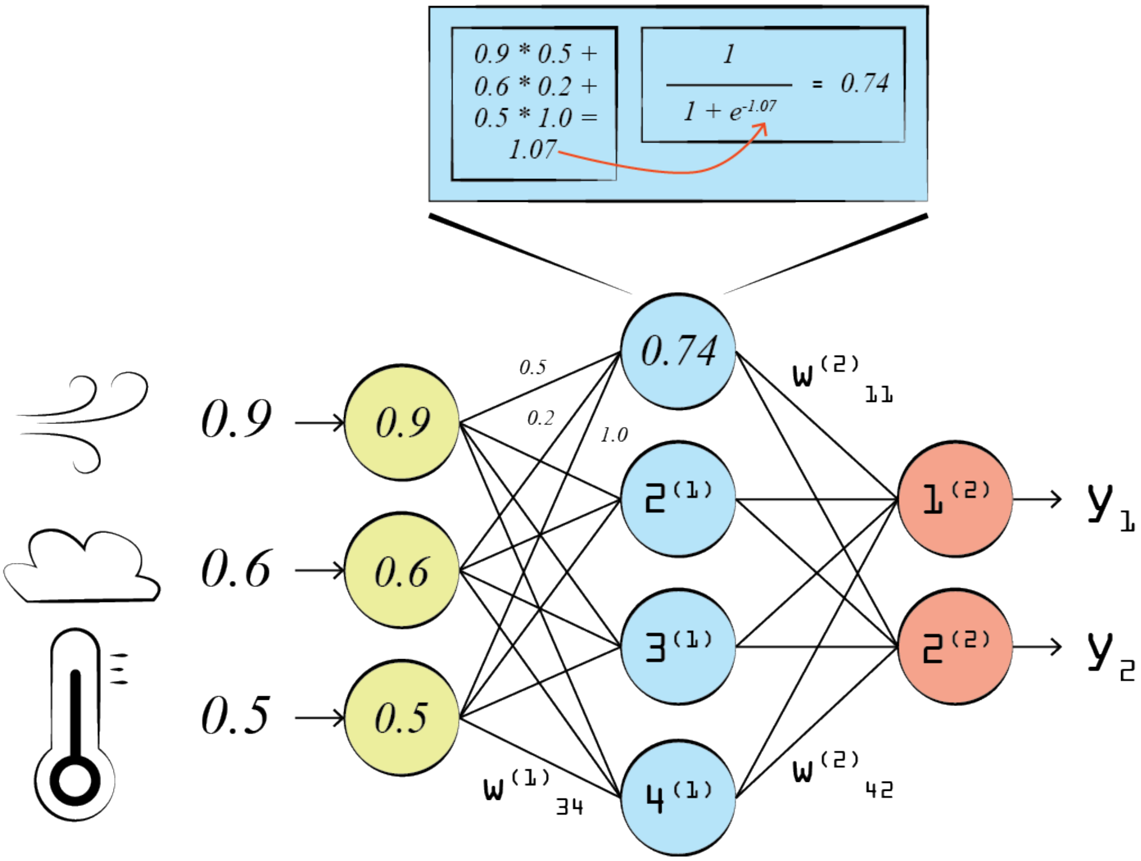

Figure 3. Visualisation of the weight summation and activation function inside a node of an ANN that is predicting rainfall, based on Bre et al. (2017, p. 4).

The example networks provided in Figures 2 and 3 are just one option of what an ANN could look like. For instance, many networks have more than one hidden layer, or have different numbers of nodes per layer. In addition, all nodes of layer 0 are connected to all nodes of layer 1, and all nodes of layer 1 are connected to all nodes of layer 2. When all nodes are connected to each other, this is called a fully-connected layer. In some networks, only some nodes in a layer L_x will be connected to some node $1^{(x+1)}$.

2.2.1.1 Training

The ANN in Figure 3 contains values for the connection weights between layer 0 and 1. The values of these weights are what determines whether the algorithm makes a correct prediction, and thus they are the focus of the training process of a neural network. An ANN is trained by means of the gradient descent and back-propagation algorithm (Rumelhart et al., 1986). Through gradient descent and back-propagation, the loss of the network is minimised and the optimal connection weights are determined. The model used in the present study is pretrained by Li et al. (2020b), as detailed in Section 4.1.

In essence, when training an ANN, we want to minimise the loss of information through the network. This problem can be operationalized as a loss function, which for linear regression is defined as the squared loss function based on the work of Gauss (1963), as shown in Equation 2.5. In the model used in the current study, other loss functions are also used, as detailed in Section 4.1, but squared loss is used here as a mathematical example.

$$Loss(h_{\mathbf{w}}) = \Sigma_{j=1}^{N}(y_j - h_{\mathbf{w}}(x_j))^2, \tag{2.5}$$

where x_j is one of N training examples, $h_{\mathbf{w}}(x_j)$ is its predicted outcome based on weight vector $\mathbf{w}$, and y_j is the real outcome of x_j. To find the global minimum of this loss function is to find the optimal weights $\mathbf{w}$ for the ANN such that the network predicts $\mathbf{y}$ from $\mathbf{x}$ with minimal error. In other words, the aim is to find $\mathbf{w}^* = \text{argmin}_{\mathbf{w}} Loss(h_{\mathbf{w}})$.

Since mathematical functions have a slope of 0 at their extrema, we can find the global minimum of the squared loss function by calculating its partial derivatives. In the case that weight vector $\mathbf{w}$ consists of two weights, w_0 and w_1, this means that we calculate two partial derivatives, shown in Equation 2.6 and 2.7 (Rumelhart et al., 1986).

$$\frac{\partial}{\partial w_0}\Sigma_{j=1}^{N}(y_j - (w_1 x_j + w_0))^2 = 0 \tag{2.6}$$

$$\frac{\partial}{\partial w_1}\Sigma_{j=1}^{N}(y_j - (w_1 x_j + w_0))^2 = 0 \tag{2.7}$$

The solutions of these equations are respectively given by Equation 2.8 and 2.9. For training linear models, solving the latter two equations is sufficient to find the minimum of the loss function, since the squared loss function is always convex, meaning that it only has one minimum (Russell & Norvig, 2010).

$$w_0 = \frac{\Sigma y_j - w_1(\Sigma x_j)}{N} \tag{2.8}$$

$$w_1 = \frac{N(\Sigma x_j y_j) - (\Sigma x_j)(\Sigma y_j)}{N(\Sigma x_j^2) - (\Sigma x_j)^2} \tag{2.9}$$

When a model is non-linear, however, as is the case when including an activation function such as the sigmoid function (equation 2.1, the loss function will not always be convex (Russell & Norvig, 2010). Thus, we need to implement a different strategy, namely gradient descent (Cauchy et al., 1847). In gradient descent, we randomly initialise a value for weight w and then alter it at each time step such that its partial derivative approaches zero. This is done by subtracting the partial derivative multiplied by a learning rate α from the current value of w, w_i, shown in Equation 2.10. This can either be done for multiple training examples at once (batch gradient descent) (Cauchy et al., 1847) or per training point (stochastic gradient descent) (Robbins, 1951), which is often faster.

$$w_i \longleftarrow w_i - \alpha \frac{\partial}{\partial w_i} Loss(\mathbf{w}) \tag{2.10}$$

The learning rate can be fixed or decrease with each step, and its size determines how fast the algorithm approaches the global minimum. A larger learning rate ensures that the gradient descent algorithm takes larger steps, making training of the model faster. However, for stochastic gradient descent, if the learning rate is too large, the gradient descent algorithm might never reach the minimum, as depicted in Figure 4. Still, a learning rate can also be too small, causing you to get stuck in a local minimum and never reaching the global minimum (see Figure 4). What a good learning rate is, depends on the exact shape of the particular loss function.

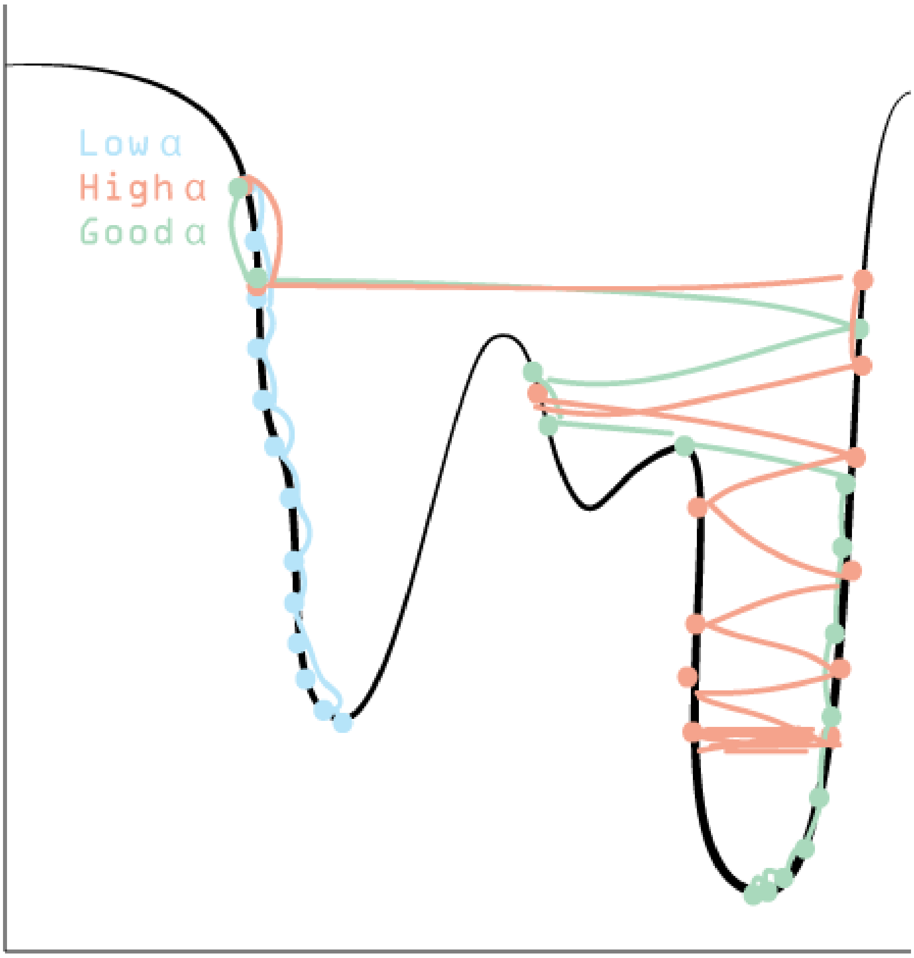

Figure 4. Visualisation of the gradient descent algorithm, with time steps shown in blue, red, and green. The blue path shows a gradient descent algorithm with a too small learning rate. The red path visualises a too large learning rate. The green path depicts a good learning rate. Based on Krittanawong et al. (2019, p. 7).

In a multi-layer neural network, the minimal loss has to be determined not only for the output layer, where we can easily compare the predicted and actual values, but also for the hidden layers (Russell & Norvig, 2010). To do this, the error is "back-propagated" from the output to the previous layers. This method is called back-propagation. In this method, each node w_i is assumed to have an influence on the error of the nodes w_j it connects to proportional to the weight of the connection w_{ij}. Therefore, the error of node w_j in layer 3 is divided across all nodes w_i in layer 2 proportional to the w_{ij} connecting them. This is done until all nodes in layer 2 have a calculated error, which is then propagated back to layer 1 for each node in the same manner, and repeated until all weights are adjusted (Russell & Norvig, 2010).

There are many different types of ANNs, and they can be applied to a variety of problems. For example, we can train a ANN to classify food ingredients according to size, shape, taste, and colour. In addition, we can use a ANN to predict the best time of the year to grow crops based on meteorological and agricultural data. However, for the purposes of this study, we are looking for a specific type of ANN that is most suitable for dealing with visual data. The model selection of the current study is detailed in Section 2.3.

2.2.2 Convolutional neural networks

For computer vision problems specifically, such as the problem in the current study, a type of ANN called a convolutional neural network (CNN) is most suitable. CNNs are ANNs as well, but include other types of layers in addition to the input, output, and hidden layers. CNNs always contain at least one convolutional layer. In this layer, a randomly initialised weight matrix called a filter or kernel is applied to the input. The input, a black and white image, can be seen as a matrix in which each element is a pixel of which the value indicates the degree of blackness at that particular location. The filter F, which has a predetermined size $\mathbb{R}^{n \times n}$, will be element-wise multiplied by each position in the input matrix X of size $\mathbb{R}^{n \times n}$, as depicted in Figure 5. For each resulting matrix, its sum will be inputted in a new matrix called a feature map of size $\mathbb{R}^{n \times n}$, such that $Y_{ij} = \Sigma(X_{ij} \odot F)$, where Y_{ij} is element i, j of feature map Y and X_{ij} is an $\mathbb{R}^{n \times n}$ sub-section of matrix X corresponding to position Y_{ij} (see Figure 5. This operation is repeated for multiple features. In this way, a convolutional layer creates a new representation of the input image, in which some features have been given more weight than others, depending on the filter used.

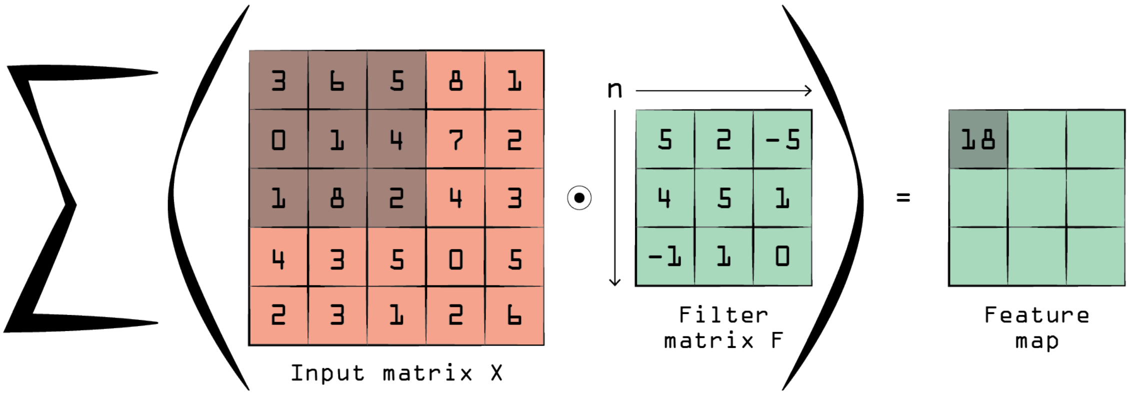

Figure 5. Abstract representation of the convolution operation in a CNN, showing an input matrix of $\mathbb{R}^{5 \times 5}$, a random filter of $\mathbb{R}^{3 \times 3}$ and a feature map of $\mathbb{R}^{3 \times 3}$. Based on Podareanu et al. (2019, p. 7).

A second layer that is often found in CNNs is a pooling layer. The goal of these layers is to down-sample the feature maps such that processing becomes more efficient. A pooling layer takes the feature maps produced by a convolutional layer and scales them down, as depicted in Figure 6. An example of a down-sampling method is the max-pooling algorithm, which was recommended by Riesenhuber and Poggio (1999) for computer vision. The max-pool operation takes the maximum number of each $\mathbb{R}^{n \times n}$ sub-matrix (for filter size n) from the feature map and inputs this into a new, smaller feature map (Serre et al., 2005).

Both the pooling and convolutional operations in a CNN can be adjusted not only in terms of filter size, but also in stride magnitude. For example, a max-pooling operation with filter size $2 * 2$ and stride 2 will first take the maximum of the leftmost $\mathbb{R}^{2 \times 2}$ sub-matrix at element A_{11} (indices starting at 1), and then the maximum of the matrix starting at element A_{13}, such as the operation in Figure 6. A max-pooling operation with filter size $\mathbb{R}^{2 \times 2}$ and stride 1, on the other hand, would first take the maximum of the $\mathbb{R}^{2 \times 2}$ sub-matrix at element A_{11}, and then the maximum of the matrix starting at element A_{12}, meaning that element A_{12} and A_{22} would feature in both sub-matrices. A higher stride leads to a smaller output feature map (both for pooling and convolution) and thus functions as down-sampling method.

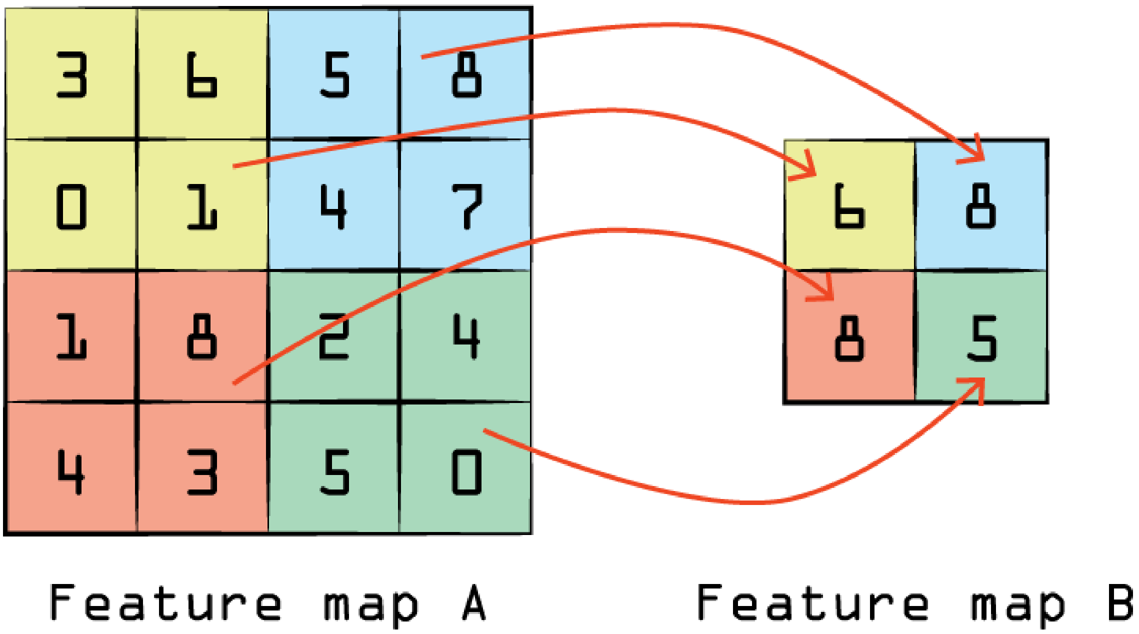

Figure 6. Abstract representation of the max-pooling operation with filter size $\mathbb{R}^{2 \times 2}$ and stride 2 in a CNN. Based on Podareanu et al. (2019, p. 7).

Figure 7 shows how in a CNN, an input image is convoluted to feature maps, which are then pooled and eventually flattened into a fully connected ANN. The particular CNN in Figure 7 tries to determine which digit is shown in the handwritten image, but CNNs can be applied to a wide variety of visual problems. For example, we can use a CNN to recognise dishes from images, which is further explored in Section 2.3. In the current study, we use a CNN to detect ingredients in images, as explained in Section 4.1.

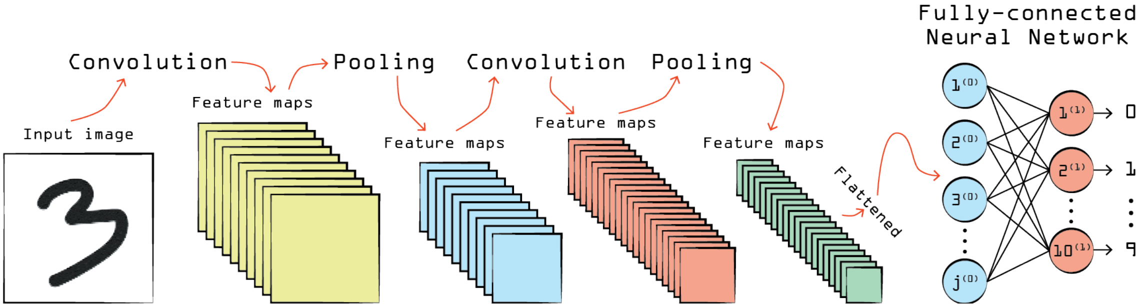

Figure 7. Abstract representation of a CNN with two convolutional and two pooling layers and a fully connected network. Based on Sekar et al. (2021, p. 3).

2.2.3 Recurrent neural networks

In addition to the visual modality of social media and online recipe analysis, which CNNs are most suitable for, the data have a textual modality. For dealing with textual data, another type of ANN is most suitable, namely recurrent neural networks (RNNs). The signal in feed-forward neural networks, such as ANNs and CNNs, only travels forwards. Contrary, in RNNs the signal travels both forwards *and* backwards (Jordan, 1986). These types of networks are especially useful for sequential data, such as language or video, since they encode the history at a previous time-step as well as the current one (Greff et al., 2017). They are trained by back-propagation through time, meaning that the gradients are computed back through all (or some, in the case of truncated back-propagation) time steps.

A special version of an RNN is a bi-directional RNN. Bi-directional RNNs were developed to account for the bi-directional nature of language. For example, not only does the start of the sentence "I like x" determine what x is (which could be anything from vacation to the colour red), but also the words that come after it: "I like x since it tastes sweet" (now it seems likely that x is a type of food) (Zhang et al., 2021). As visible in Figure 8, bi-directional RNNs include a hidden layer that processes information from end to start, rather than from beginning to end.

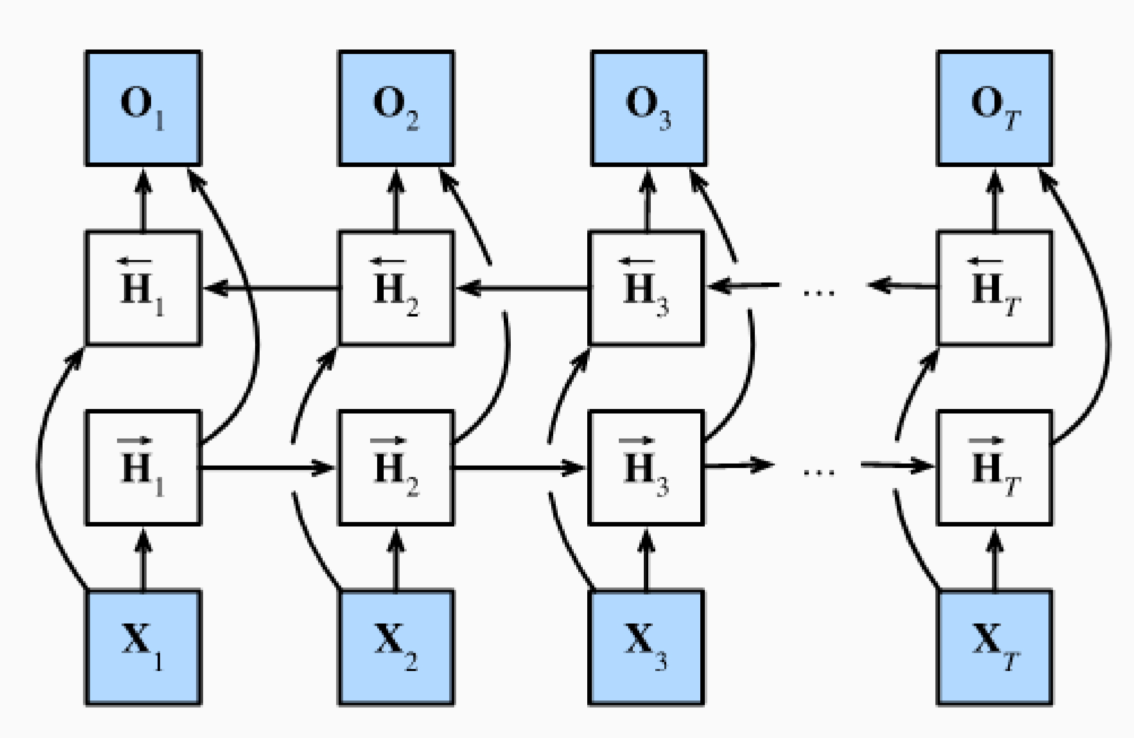
$\mathbf{O}_1$
$\mathbf{O}_2$
$\mathbf{O}_3$
$\mathbf{O}_T$
$\overleftarrow{\mathbf{H}}_1$
$\overleftarrow{\mathbf{H}}_2$
$\overleftarrow{\mathbf{H}}_3$
...
$\overleftarrow{\mathbf{H}}_T$
$\overrightarrow{\mathbf{H}}_1$
$\overrightarrow{\mathbf{H}}_2$
$\overrightarrow{\mathbf{H}}_3$
...
$\overrightarrow{\mathbf{H}}_T$
$\mathbf{X}_1$
$\mathbf{X}_2$
$\mathbf{X}_3$
$\mathbf{X}_T$

Figure 8. Abstract representation of a bi-directional RNN. Image directly taken from Zhang et al. (2021, para. 9.4.2).

2.2.3.1 Long short-term memory

Unfortunately, RNNs have the problem of vanishing gradients (Zhang et al., 2021), meaning that all information is treated as equally important. This might not always be the case, for example in processing a story where the main character is mentioned in the first line, followed by some sentences describing the surroundings of the scene. In that case, the name of the main character should be retained longer than the description of the room. A type of RNN that solves the problem of vanishing gradients is a long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997). This type of network model mimics biological long and short term memory, with short-term memory represented as a feed-forward signal and long-term memory as a recurrent feed-back signal that enables nodes to remember their previous state. Thus, the output of an LSTM feeds back into the input of the network (Greff et al., 2017), as shown in Figure 9.

In the diagram in Figure 9, the LSTM consists of four main components: 1. a forget gate, 2. an input gate, 3. candidate memory, and 4. an output gate. These components enable the LSTM to assess the importance of each item, and make a distinction between long-term memory for important items and short-term memory for less important items (Zhang et al., 2021). For example, when generating a text on a certain topic, the topic provided in the first sentence should be stored in long-term memory, since it is important for the line of the story. However, the precise wording of the currently generated sentence should be stored in short-term memory, since they are trivial later in the text. An LSTM allows to retain or forget items and in addition enables us to reset the model (for example in case of a new "section" of data).

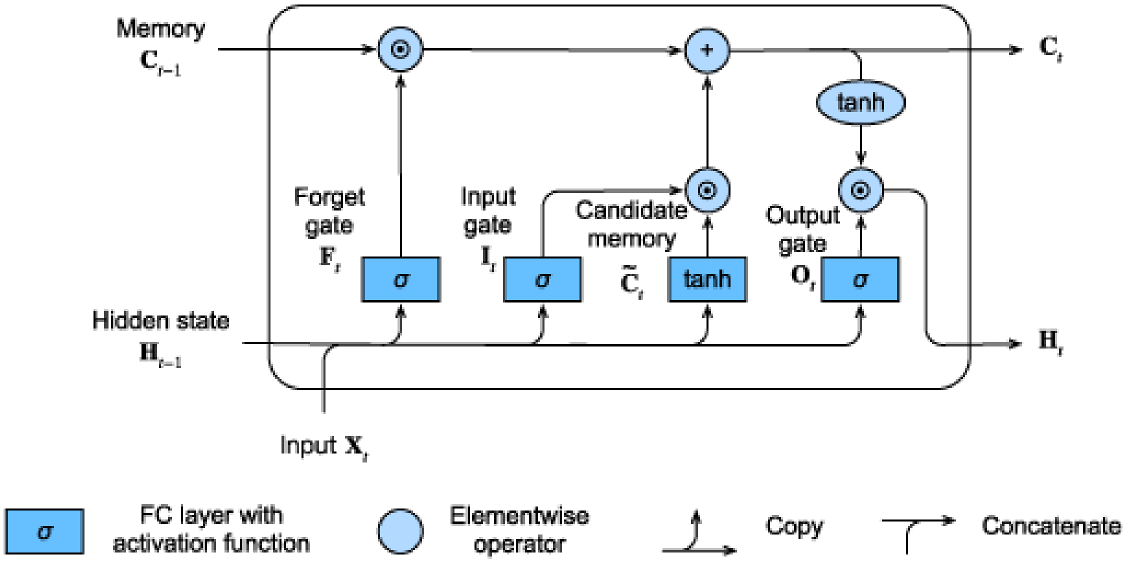

Figure 9. Abstract representation of an LSTM. Image directly taken from Zhang et al. (2021, para. 9.2.1.4).

The LSTM takes as input the new input at time step t, X_t, and the previous hidden state at time step $t-1$, namely H_{t-1}. The forget gate F_t then determines with a sigmoid function how important the previous state is and the input gate determines the importance of the current input. The activation of the input gate is then multiplied by the candidate memory and added to the memory assessed by the forget gate to result in the current cell state C_t. In addition, the output gate defines the new hidden state H_t that will be rerouted to the cell in time step $t+1$. Each of these operations function similar to the feed forward pass in an ANN, with weights determining the relative importance of each component.

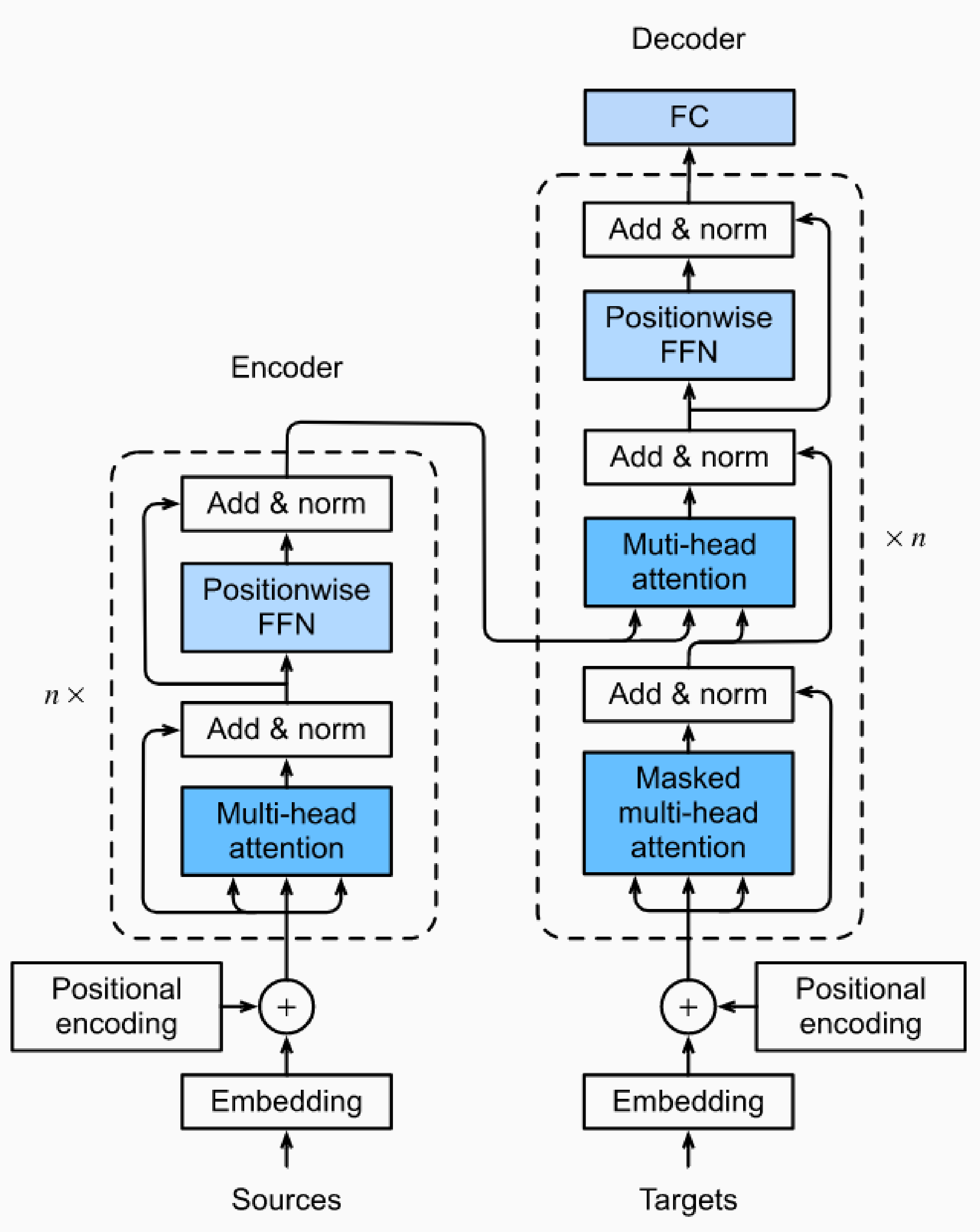

Figure 10. Abstract representation of a transformer, where FFN stands for feed forward network and FC stands for fully connected layer. Image directly taken from Zhang et al. (2021, para. 10.7.1).

In sum, an LSTM is very useful for modelling sequential data, such as food recipes. In addition, bi-directional LSTMs can be applied to model sequential but unordered data, such as ingredient sets. For the pretrained model used in the present study, a bi-directional LSTM was used by Li et al. (2020b) to model recipe ingredients, as explained in Section 2.3.

2.2.4 Transfer learning

In general, neural network models are trained in a supervised manner using data of a format equal to their eventual task, as started with the perceptron (Rosenblatt, 1958). For example, a CNN that is aimed at recognising bird species from images can be trained on a dataset containing images of birds and corresponding species labels. Datasets in ML are usually split up into three parts (Ripley, 1996): 1. a training set, 2. a validation set, and 3. a test set. The training set is used to train the ML model, in this case a neural network, the validation dataset is used to check generalisability, and the test set is used to test the performance of the model. Thus, in training these networks, large amounts of data are needed, which means that creating a good model is dependent on data availability for a specific task. Unfortunately, in many cases, not enough data is available. This is the case with the current study: the collected AzerFSQFood dataset contains only a moderate number of images, as detailed in Section 4.2.

One solution to lack of data on a given problem is transfer learning (Baxter et al., 1995). In transfer learning, models that are trained to solve a similar but nonidentical problem are fine-tuned using a smaller amount of data of the actual problem. For example: imagine you want to train a model to recognise fish species, but there are only small datasets available. You might then take a model that is pretrained on images and labels of bird species, and continue to train this model with a small dataset on fish species.

In the computer vision field, there exist a number of CNNs that have been proven to reach good results in image classification. A couple examples are VGG (Simonyan & Zisserman, 2015) and ResNet (He et al., 2016). Both these models are deep CNNs with many convolutional layers interspersed with max-pooling layers (He et al., 2016; Simonyan & Zisserman, 2015). The models are trained on the ImageNet database (Deng et al., 2009), an image database based on the textual WordNet database (Princeton University, 2021). For each of the 100.000+ concepts included in WordNet, ImageNet on average includes 1000 images of that concept (Deng et al., 2020). Other models that are often applied in computer vision and food computing are AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015), and Network-in-Networks (Lin et al., 2013). These pretrained architectures can be used in new computer vision problems by employing transfer learning to tune them to a new problem. This study employs the ResNet-50 (He et al., 2016) model pretrained on ImageNet (Deng et al., 2020) to encode imagery, and trained on an online recipe dataset with transfer learning by Li et al. (2020b), since ResNet-50 has proven to reach good results in food recognition (see Section 2.3).

2.2.5 Multi-modality

Despite what may seem the case, computer vision problems are rarely exclusively visual. Annotating images, for example, involves both visual data, namely the images itself, and textual data, namely the generated captions (Hao et al., 2019; Karpathy & Fei-Fei, 2015; Vinyals et al., 2015). Furthermore, as mentioned, analysing online recipes and social media can be a multi-modal problem when dealing with images, as is the case in the present study. The presence of multiple modalities calls not just for computer vision analysis, but for multi-modality analysis, in this case a combination of natural language processing and computer vision. Multi-modal analysis is also often used in food computing, for example in studying food images and corresponding recipes (Min et al., 2019).

Three most frequently used approaches for modelling multi-modal data are: 1. multi-modal fusion, where all modalities are are used in feature learning, training, and testing (for example, given both a video and an audio track of speech, try to generate written text), 2. cross-modality learning, in which data from multiple modalities are only used during feature learning (to represent single modalities based on unlabelled multi-modal data), and 3. shared representation learning, which is similar to cross-modality learning, but different modalities are used in training and testing (to learn correlations across different modalities) (Ngiam et al., 2011). In other words, in cross-modal retrieval, one can train a classifier to both retrieve modality A given modality B and to retrieve modality B given modality A (Hao et al., 2019; Li et al., 2019). Generally, multi-modal modelling aims to create a multi-modal embedding in which all involved modalities are represented. Then, the distance between these modalities is minimised to classify the data point. For example, given a recipe and a corresponding image, the goal is to place the image and the recipe in a space such that this recipe is closer to the image then any other recipe already present in the space.

These distances are minimised by ways of loss functions, such as cosine loss. These loss functions are ways by which to measure "distance" or "similarity" between features of different modalities. To optimise the loss function is to represent matching examples as close as possible in the embedding space. Often, these loss functions perform pairwise analysis, but some studies have worked with triplet loss, not only minimising the distance between two matching examples, but also maximising the distance between non-matching examples (Hao et al., 2019). This means that triplet loss often represents matching items more closely in the multi-modal embedding space.

There are two possible architectures of multi-modal analysis: single-stream and dual-stream architecture. Their structure is depicted in Figure 11. In a single-stream architecture, each modality is first separately embedded. Then, the embeddings are summed and the resulting multi-modal embedding is processed by a transformer (Bugliarello et al., 2021). The transformer establishes which parts of one modality are related to which parts of another modality, i.e. it models intra-model and inter-modal interactions(Bugliarello et al., 2021). In a dual-stream architecture, the two modalities are each first transformed and then encoded in a multi-modal embedding (Bugliarello et al., 2021). Since the present study deals with both visual and textual data, namely recipe images and text and social media images and captions, I can employ a multi-modal model with a multi-stream architecture. The approach is cross-modality learning, since first the features of both the images and text of the recipes are learned, but during training and testing, only one modality (namely, the image) is provided to return the other (the ingredient set).

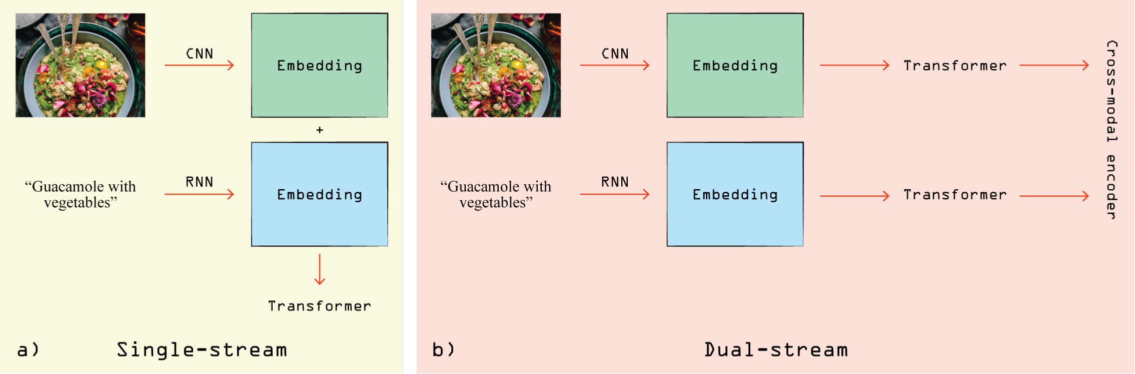

Figure 11. Abstract representation of a) single-stream and b) dual-stream architectures of multi-modal models. Based on Gatt (2021, n.p.).

2.3 Computer vision for food computing

The previously discussed machine learning methods for computer vision can be applied to mapping various aspects of food consumption. As discussed in section 2.1, food consumption can be mapped through visual and multi-modal analysis of social media images and online recipes. Min et al. (2019) mention five areas for food computing: 1. perception, the study of how people perceive food, 2. recognition, recognising food items based on visuals or other modalities, 3. retrieval, selecting food items from an offered set, 4. recommendation, optimising food recommendation systems, and 5. prediction and monitoring, for predicting a variety of factors linked to food consumption. Since the purpose of this study is to geographically map food consumption, the relevant food computing applications in this case are food recognition and retrieval. In this Section, I go into how food recognition or identification and food retrieval, for example of ingredients or dishes, can be applied to map food consumption. First, however, I review existing datasets on food consumption.

2.3.1 Existing data

There already exist a number of datasets containing both food images and recipes, which have been constructed or used in the previously described studies. Table 1 shows a comparison of these datasets. As seen in this Table, Recipe1M+ (Marin et al., 2021) and Recipe1M (Salvador et al., 2017) are the largest publicly available datasets containing both images and recipes. Therefore, I propose to use one of these two datasets for model training.

Dataset	Images	Recipes	Source (web)	Accessibility
Recipe1M+ (Marin et al., 2021)	13m	1m+	Recipe1M, Google Images	Open access
Recipe1M (Salvador et al., 2017)	800k	1m+	24+ sites	Open access
Recipes242K (Rokicki et al., 2018)	242k	242k	Allrecipes	Open access
VIREO Food-172 (J. Chen & Ngo, 2016)	110k	65k	Chinese sites	Per request
Go cooking (J. Chen et al., 2017)	61k	61k	Go Cooking	Private
Yummly-66K (Min et al., 2018)	66k	66k	Yummly	Open access
Yummly-28K (Min et al., 2017)	28k	28k	Yummly	Open access
RecipeQA (Yagcioglu et al., 2018)		20k	Instructables	Open access

Table 1. Comparison of existing datasets containing food images and recipes. The databases are sorted on amount of images (large to small).

Since I propose to analyse the usefulness of social media data in fine-grained geographical mapping of food consumption, I am also looking for a dataset containing social media images. Table 2 contains an overview of datasets of food images collected from Instagram. As my case study is Baku, Azerbaijan, I wish to apply the model to an Instagram dataset containing food images from across that city. However, as visible in Table 2, no such dataset is yet publicly available (as of my current knowledge). In addition, since Instagram down-scaled their API in 2020, there is no good way to extract location data from the platform at the time of writing. Therefore, I take to the use of a proxy dataset, collected from Foursquare. The application of the methodology on Foursquare images demonstrates the applicability of the methodology to images containing location data from other sources. To this end, I construct a dataset in this study called AzerFSQFood, by scraping food images from Azerbaijan from Foursquare.

Dataset	Data types	Images	Classes	Accessibility
Insta-1K (Ofli et al., 2017)	images, tags	3.7m	1K	Private
Instagram 1.7M (Phan & Gatica-Perez, 2017)	images, comments	1.7m		Private
Instagram800K (Rich et al., 2016)	images, tags	809k	43	Broken link
Food500 (Merler et al., 2016)	images, dish names	148k	508	Private
IFD (Kagaya & Aizawa, 2015)	images (food/n-food)	10k	2	Broken link

Table 2. Comparison of existing datasets containing food images from Instagram. The databases are sorted on number of images (large to small).

In order to make inferences about (relative) amounts of ingredient consumption, a dataset containing food volume information might also be interesting. As Table 3 shows, there are two known publicly available datasets containing volume information. However, it is also possible to train a deep model on recipe data and thereby calculate relative ingredient amounts (Li et al., 2019). This means that the Recipe1M or Recipe1M+ datasets are sufficient to make inferences about relative ingredient amounts.

Dataset	Data types	Imgs, frames	Classes	Accessibility
50 Salads (Stein & McKenna, 2013)	depth videos	518k	17	Open access
GFood3d (Myers et al., 2015)	depth videos	150k	50	Private
PFID (M. Chen et al., 2009)	images, stereo pairs, videos	5k	101	Private
NFood3d (Myers et al., 2015)	depth images	1k		Private
ECUSTFD (Liang & Li, 2017)	images, dish names, volume info	3k	19	Open access

Table 3. Comparison of existing datasets containing food images and videos with volume information. All the visuals in the datasets are originals. The databases are sorted on number of images and frames (large to small).

I use the FAO's balance sheets (FAO, 2019) regarding Azerbaijan as theoretical starting point for my analysis. In addition, I compare my results to a dataset on supermarket sales in different neighbourhoods of Baku, Azerbaijan (Zeynalov, 2020) to validate that Foursquare data can be used to measure relative ingredient consumption on a small geographical scale. This dataset contains a total of 438k purchases from 19 supermarkets in 2020 across Baku, Azerbaijan and from one supermarket outside of Baku.

2.3.2 Food identification

A first application of computer vision to consumption mapping is food identification or recognition. Food identification is a form of image classification specialised in dishes. Table 4 provides an overview of studies performing this task. In earlier research, the model was trained to recognise whether an image as depicts "food" or not, called binary food detection (Min et al., 2019). Nowadays, models are trained to be able to classify various food into their respective dishes (Min et al., 2019). This technique can for example be used to recommend healthier dishes to people who often take photos of unhealthy food (Matsuda et al., 2012).

In earlier years, food recognition from images was mostly done by using hand-crafted features such as scale-invariant feature transform (SIFT). SIFT is a method developed by Lowe (2004) to detect the most salient parts of images that are constant across scale and rotation. For example, M.-y. Chen et al. (2012) employed SIFT and a number of binary support vector machines (SVMs) to identify images of Chinese dishes. Beijbom et al. (2015), similarly, use hand-crafted features and SVMs to classify food images. However, rather than identifying a single food type for each image, they perform multi-class classification: their images feature more than one dish, each of which they want to produce labels for. Multi-class classification is more difficult than single-class classification, since it requires region detection and image segmentation. However, they only studied images of which they knew the restaurant's menu, thus the number of possible classes was very limited.

Model	Database	Labels	A@1	Image representation	Classification	Accessibility
Hoashi et al. (2010)	Food85	Single	62.52	Hand-crafted features	SVM: MKL chi-sq, one-vs-all	Private
Joutou and Yanai (2010)	Food50	Single	61.34	Hand-crafted features	SVM: MKL chi-sq, one-vs-all	Private
M.-y. Chen et al. (2012)		Single	68.3	Hand-crafted features	SVM: Adaboost, non-linear	Private
Anthimopoulos et al. (2014)	Diabetes	Single	78		SVM: Linear	Private
Bossard et al. (2014)	Food-101	Single	50.76	RF		Private
Kagaya et al. (2014)	FL	Single	73.7	CNN with LRN		Private
Kawano and Yanai (2014)	UECFOOD-256	Single	50.1	Hand-crafted features	SVM: One-vs-all, linear, AROW	Data only
Beijbom et al. (2015)	Menu-Match	Multi	77.4	Hand-crafted features	SVM: One-vs-all	Private
Bettadapura et al. (2015)		Single	63.33		SVM: SMO-MKL	Private
Ciocca et al. (2015)	UNIMIB2015	Multi	99.05	kNN		Private
Kagaya and Aizawa (2015)	FL (IFD, FCD)	Binary	99	NIN (Network-In-Network)	SVM: Adaptive	Link broken
Myers et al. (2015)	Food201-multilabel	Multi	81.4	GoogLeNet		Data only
Xu et al. (2015)	Dishes	Single	80.05	AlexNet	SVM: One-vs-all	Link broken
Lu (2016)	ImageNet-Food	Single	90	CNN		Open access
Okamoto and Yanai (2016)	UECFOOD-100	Multi	75	NIN (Network-In-Network)		Data only
Rich et al. (2016)	Instagram800K	Single	75.1	VGG-16	SVM: Linear	Link broken
J. Chen and Ngo (2016)	VIREO Food-172	Single	82.06	VGG-16		Private
Farinella et al. (2016)	UNICT-FD1200	Single	92.6	Hand-crafted features		Link broken
Merler et al. (2016)	Food500	Single	74.4	GoogLeNet		Private
Singla and Yuan (2016)	Food-11	Single	83.5	GoogLeNet		Data only
Zhou and Lin (2016)	Food-975	Single	89	GoogLeNet		Link broken
X. Chen et al. (2017)	ChineseFoodNet	Single	81.55	VGG, ResNet, DenseNet		Link broken
Ciocca et al. (2017a)	Food524DB	Single	94.96	ResNet-50		Private
Ciocca et al. (2017b)	UNIMIB2016	Multi	79	CNN and k-NN		Private
Herranz et al. (2017)	Dishes	Single	96.42	AlexNet	Probabilistic	Private
Ofli et al. (2017)	Insta-101		60.6	ResNet-50		Private
Ofli et al. (2017)	Food-101		63	ResNet-50		Private

Table 4. Comparison of existing models for food image classification. Models are chronologically ordered. A@1 = top-1 accuracy. Only studies that reported top-1 accuracy are included.

In later years, CNNs have become popular in food image classification. CNN models pretrained on general image databases such as ImageNet (Deng et al., 2009), for instance GoogLeNet (Szegedy et al., 2015), AlexNet (Krizhevsky et al., 2012), ResNet-50 (He et al., 2016) and VGG-16 (Simonyan & Zisserman, 2015) are often used in food image classification problems. For example, Myers et al. (2015) used GoogLeNet on their Food201-multi-label dataset. Kagaya and Aizawa (2015), on the other hand, used a network-in-network (NIN) model combined with multiple one-vs-all binary SVMs to perform binary food image classification, i.e. to recognise whether an image depicts food or not. A NIN is a CNN that applies a multi-layer perceptron to each pixel of the image (Lin et al., 2013). Naturally, binary food/non-food classification is easier than multi-label and even single-label food classification. However, such a binary classifier can still be useful, for example in performing preprocessing of messy data.

Relevant inferences can be made about culinary habits based on dish classification alone: M.-y. Chen et al. (2012) for instance directly infer nutritional value based on average dish composition. However, to make more accurate inferences on food consumption for geographical mapping or food recommendation and health applications, dish classification alone is not enough. In the case of the current study, the image processing model should be able to make inferences about the ingredients features in an image.

2.3.3 Ingredient recognition

An application of computer vision to food consumption that is perhaps more directly usable in geographical consumption mapping is ingredient recognition. Ingredient recognition is a type of object detection. Object detection in computer vision differs from image classification in that it yields multiple rather than one label per image, and that each label is associated with a specific location in the image rather than the image as a whole. For example, rather than classifying an image of spaghetti bolognese as "spaghetti bolognese", ingredient recognition might produce the labels "spaghetti", "tomatoes", and "mince meat".

Puri et al. (2009) aim to calculate the nutritional value of food intake of mobile phone users. Using a combination of manually segmented food images and user's audio descriptions, they train pairwise SVM classifiers to detect food items on an image, based on hand-crafted image features. They subsequently perform 3D volume estimation, as described in Section 2.3.5, to eventually calculate nutritional value. Unfortunately, their method for ingredient recognition as shown in Table 5 requires user input in the form of providing audio descriptions and including colour and size calibration checker boards in the images. In addition, their method requires manual segmentation of the images to mark the edges of multiple food times visible in the same image. Moreover, their images depict visually cleanly separated ingredients, whereas in practise, ingredients are likely to overlap partly or be mixed together. Min et al. (2017), on the other hand, use a deep model, namely the GoogLeNet (Szegedy et al., 2015) model on their Yummly-28K dataset to perform ingredient retrieval. They reach a top-50 precision of 0.871. These results show that ingredient retrieval is possible on large image datasets without requiring audio data.

Li et al. (2019) explore ingredient recognition further and compare retrieval and generation methods. For retrieval, they employ a gated recurrent unit (GRU) model. For generation, they use the ResNet-50 model, in which they replace the last layer to perform quantity estimation. They show that retrieval and generation reach similar performance, with recall rates of 0.33 and 0.35, respectively. As visible in Table 5, not many models perform only ingredient retrieval and/or recognition. Most models in fact also perform recipe retrieval, which is described in Section 2.3.4.

Model	Database	A@1	R@1	P@50	Image representation	Text representation	Architecture	Accessibility
Puri et al. (2009)		90			Hand-crafted features	SVM: Adaboost χ^2, pairwise		Private
Li et al. (2019)	Recipe1M		0.35		ResNet-50	GRU		Private
Min et al. (2017)	Yummly-28K			0.871	GoogleNet	Bag-of-Ingredients, one-vs-all logistic	Multi-stream	Private

Table 5. Comparison of existing models for cross-modal ingredient retrieval and generation from food images. Models are chronologically ordered. A@1 = top-1 accuracy, R@1 = top-1 recall, P@50 = precision at top-50.

Model	Database	A@1	R@1	Image representation	Text	Architecture	Loss function	Accessibility
Wang et al. (2015)	UPMC Food-101	85.1		VGG-19	SVM			Data only
J. Chen et al. (2017)	Go Cooking		0.048	SAN (VGG + LSTM)	LSTM	Single-stream	Rank-based	Private
Salvador et al. (2017)	Recipe1M	84.8	0.24	ResNet-50 , VGG-16	LSTM	Multi-stream	Cosine sim., semantic reg.	Open access
Carvalho et al. (2018)	Recipe1M		0.398	ResNet-50	LSTM	Multi-stream	AdaMine (double triplet)	Open access
J.-J. Chen et al. (2018)	Recipe1M		0.256	ResNet-50	LSTM	Multi-stream	Rank-based	Private
Fontanellaz et al. (2019)	Recipe1M+		0.346	ResNet-50	LSTM	Multi-stream	AdaMine (double triplet)	Private
Hao et al. (2019)	Recipe1M		0.518	ResNet-50	LSTM	Multi-stream	Adversarial	Open access
Salvador et al. (2019)	Recipe1M	55.47	**0.7547**	ResNet-50	LSTM	Single-stream	Cross-entropy, eos loss	Open access
Marin et al. (2021)	Recipe1M+	84.8	0.24	ResNet-50 , VGG-16	LSTM	Multi-stream	Cosine sim., semantic reg.	Open access

Table 6. Comparison of existing models for cross-modal recipe retrieval from food images. Models are chronologically ordered. A@1 = top-1 accuracy, R@1 = top-1 recall.

2.3.4 Recipe retrieval and generation

Recipe retrieval and generation are two applications for computer vision with regards to mapping food consumption. It is possible to train an ANN to take an image of a dish and (1) retrieve the corresponding recipe from a given set of possibilities (Jiang & Min, 2020), or (2) generate the corresponding recipe purely based on the image. This can be done in multiple steps, for instance by first generating ingredients by performing object recognition on the image and then producing the recipe based on both the image and the ingredients. Recipe retrieval requires less training data, making it a less expensive option than recipe generation (Min et al., 2019).

Table 6 shows a comparison of various ML models that either aim to retrieve or generate food recipes based on images of dishes. Note that the studies often report top-1 accuracy for recipe retrieval or generation performance, and top-1 recall for ingredient retrieval or generation. Salvador et al. (2017) and Marin et al. (2021) both train a model to pick one of a provided set of recipes based on an image, i.e. they perform recipe retrieval. Salvador et al. (2017) train a bi-directional LSTM to represent sets of ingredients in recipes in an embedding space, based on the cross-modal visual-semantic embedding method introduced by Karpathy and Fei-Fei (2015) and Vinyals et al. (2015). Similarly, Salvador et al. (2017) train a second LSTM to represent recipe instruction vectors in the same embedding space. They also train a CNN to represent images of food in the aforementioned embedding space. The concatenation of the ingredient and instructions representation is the representation of a given recipe. Lastly, they train a cross-modal embedding model to learn transformations to most closely match the embedding locations for the recipe to the embedding of the corresponding image. In other words, they make use of a multi-stream architecture to perform cross-modal analysis. This model can then be used to either retrieve the most closely-matching ingredients and instructions based on a given image or retrieve the most similar image based on a given recipe (Salvador et al., 2017). Marin et al. (2021), the same team as Salvador et al. (2017), replicated their earlier method, but now on an extended version of their database, Recipe1M+.

The method of Salvador et al. (2017) was in addition replicated by Hao et al. (2019), who used the same Recipe1M database (Salvador et al., 2017) to learn cross-modal embeddings for cooking recipes and food images. They extend (Salvador et al., 2017) their technique by introducing adversarial learning. Whereas Salvador et al. (2017) employ cosine loss to learn the embedding, Hao et al. (2019) use adversarial loss. Whereas cosine loss is a pairwise comparison method, adversarial loss is a triplet loss, meaning features will be matched more closely in the embedding space. In addition, they enforced translation consistency, meaning that the embedding of modality x can be used to retrieve the matching values in modality y, making representations even more closely matched. Unfortunately, neither Salvador et al. (2017) nor Hao et al. (2019) report results on the ingredient recognition task, only on recipe retrieval.

Salvador et al. (2019) is the only model that is trained to perform recipe generation rather than recipe retrieval. They perform this task in two steps: first, the ingredients are generated from an image. Then, a second model takes both the image and ingredients, and generates a corresponding recipe. In their 2019 study, they also aim to perform cross-modal analysis, just as with their 2017 model, but this time they opt for a single-stream rather than multi-stream architecture: first, they encode the image with a CNN. Then, from the image encoding, they decode the ingredients with a transformer. Subsequently, they encode the ingredients, and lastly, they decode the instructions. They reach a top-1 accuracy of 55.47% and a recall rate of 0.75, as seen in Table 6. Salvador et al. (2017), show that they are able to perform an image to recipe retrieval task with an accuracy of 84.8%. Salvador et al. (2019), with their bottom-up single-stream approach, on the other hand, only reach an accuracy of 55.47% on the same dataset, exemplifying that recipe generation is a more difficult task than recipe retrieval. However, they do reach a higher recall rate regarding ingredient generation, namely 0.755 rather than 0.24.

All in all, the methods used in recipe generation and retrieval are very comparable, with various CNN representations for images and RNN representations for recipe instructions. Performances tend to be better for larger datasets, exemplified by the high accuracy of Marin et al. (2021) and the many studies using the (pre-Recipe1M+ (Marin et al., 2021)) largest image-recipe database, Recipe1M (Salvador et al., 2017). For the current study, recipe generation is more appropriate than retrieval, since it does not require the produced modality to be selected from a list of available options. In addition, this study is more concerned with ingredient generation than instruction generation. However, since I want to map relative ingredient consumption, I also need volume information of each detected ingredient.

2.3.5 Volume estimation

M.-y. Chen et al. (2012), who perform dish identification, use dish information combined with quantity estimation to infer nutritional value. By taking photographs of dishes using a depth camera, they are able to estimate the volume of each dish. Similarly, Ege et al. (2019) describe a previous work in which they use the inertial sensor in smartphones to estimate volume. Moreover, Myers et al. (2015) use a CNN trained on images and corresponding depth images to estimate calorie count and other nutritional quantities using a single image. However, M.-y. Chen et al. (2012) conclude that it is difficult to estimate volume using depth cameras for transparent and reflective substances such as water and cooked rice. They propose another option that can combat this problem, namely stereo images.

Puri et al. (2009), as described in Section 2.3.3, use ingredient recognition combined with stereo images to perform volume estimation. They ask users to take three photographs of their food from various angles, including checker boards for colour and size calibration. They then compare two images to form a stereo pair and use 3D reconstruction to create a 3D point cloud of the food, and, finally, the 3D scale and table plane. Similarly, Ege et al. (2019) employ the stereo cameras present on Apple iPhones to create stereo images for volume estimation. Unfortunately, neither depth cameras nor stereo images are available techniques to perform volume estimation on social media images, since depth information and stereo images will likely not be available. However, the technique used by Myers et al. (2015) only requires depth images during training and not during testing, which means it could be applied to social media images.

Other studies have aimed to perform volume estimation using only a single image. For example, Fang et al. (2018) use a generative adversarial network (GAN) to estimate food quantity based on energy and segmentation labels of the data. A GAN is a ANN that learns to generate artificial examples that closely match the truth, by trying to distinguish its own examples from real examples. A disadvantage to using their method is that it requires detailed labelling of data (Li et al., 2019).

Many studies have also aimed to estimate quantities based on reference objects in single images. Ege et al. (2019) describe their previous work of estimating food quantities using a reference object in a single image, just as (Liang & Li, 2017). Unfortunately, using predetermined reference objects is neither an option for existing social media image analysis. Another previous work by Ege et al. (2019) is estimating food quantity in multi-dish single images by selecting one dish as reference object. This method could be applicable to social media images containing multiple dishes, but in making the assumption that one dish's volume is known, the method is quite rough. A last option for single image quantity estimation with reference objects is also provided by Ege et al. (2019), namely using a boiled grain of rice as reference object.

A limit of this method is that only for dishes containing rice quantity estimation could be performed. However, perhaps a new method could include multiple options of reference objects, such as a grain of rice, a soy bean and a peanut, and pick the first available one.

Some researchers choose to infer relative rather than absolute food quantities from a single image. For example, Li et al. (2019), as mentioned in Section 2.3.3, use the images and recipes from the Recipe1M dataset (Salvador et al., 2017) to estimate relative ingredient quantity. In short, they train a CNN to retrieve recipes from images and infer the relative ingredient quantities from the information in the recipe's ingredients. They deem their method applicable to social media images and other existing image datasets.

It seems that estimating relative food amounts is a more realistic goal than absolute quantity estimation for food consumption mapping using existing online visual data. This means that Li et al. (2019)'s model would be most suitable for this study, albeit that the model is not publicly accessible. However, Li et al. (2020b) replicated their previous study and improved it by taking similarities between ingredients into account. This means that their model is penalised more for incorrectly classifying ingredients more different from the ground truth than for misclassifying the ingredient as something closer to the ground truth. For example, mixing up walnuts and almonds is less severe than mixing up tomatoes and fish. This model, too, is trained on Recipe1M, which together with Recipe1M+ is the largest publicly available multimodal recipe dataset. Since their publicly accessible model performs both ingredient detection and (relative) quantity estimation and is trained on the largest available dataset, Li et al. (2020b)'s Picture-to-amount (PITA) model is be used in this study for both these tasks.

Chapter 3

Case study

The case study for the current study is be Baku, the capital of Azerbaijan. Azerbaijan was chosen as case study because it lies in the Global South and has high rates of literacy (StatGovAz, 2021) and available food consumption, production, and retail data. This Section gives a broad overview of the demographics of Azerbaijan and Baku. In addition, it explores cuisine, culture, and social media usage.

3.1 Demographics

According to a digital report of Kemp (2021), in January 2021, Azerbaijan had a population of 10.18 million people. Of those, 56.6% lived in cities. Its capital Baku, one of the 11 cities, counts 2,3 million inhabitants as of the beginning of 2021, 22.6% of the total population, the highest number of inhabitants of all of the country's economic regions (StatGovAz, 2021). Baku consists of 12 districts. However, for the current study, I divide the world (and thus Baku) into geographical grid regions of approximately 5 by 5 kilometre. For Baku's surface area of 2000 square kilometre, that gives approximately 25 geographical regions for which to compare food consumption. Some regions can be excluded due to lack of supermarket sales data in those grids.

3.2 Consumption behaviour

Traditional Azerbaijani dishes include: 1. dolma, i.e. stuffed vegetables such as grape leaves, cabbage, or aubergine, 2. saj, i.e. an assortment of meat and vegetables, 3. pilaf or plov, i.e. a cereal such as rice combined with lamb, chicken, and/or vegetables, 4. levangi, fish or chicken with rice and stuffed with almonds, 5. kebab, i.e. barbecued meat such as lamb or veal, and 6. sweets such as baklava (AzerbaijanTravelInternational, 2021). This means that it is expected that cereals such as rice and vegetables and meat are often consumed. Figure 12 shows the volume (in 1 million kilograms) of each food type consumed in Azerbaijan in 2018, according to the FAO's balance sheets based on production, import and export data. As visible, indeed, cereals are the most consumed, followed by vegetables and milk. Meat follows after fruit, roots, alcohol and sugar. Aquatic products other than fish and seafood are the least consumed. Spices, too, make up a very small part of the total weight, but as spices have a low mass, this does not necessarily reflect frequency of spices use. Pork is hardly consumed, as is expected due to the large number of Islamic residents, which is 97.3% of the total population (CIA, 2021; 'Market and competitiveness analysis of the Azerbaijan agricultural sector: An overview', 2017). Surprisingly, alcoholic beverages are the 6th most often consumed food type, contrary to what Islamic values prescribe. In terms of production, Azerbaijan's most important crops are also cereals, followed by potatoes and vegetables, and fruits ('Market and competitiveness analysis of the Azerbaijan agricultural sector: An overview', 2017). In livestock farming, the focus is mostly on milk, beef, and sheep and goat meat, eggs and poultry. Overall, in this study I expected to find high consumption rates of cereals, vegetables, and meat.

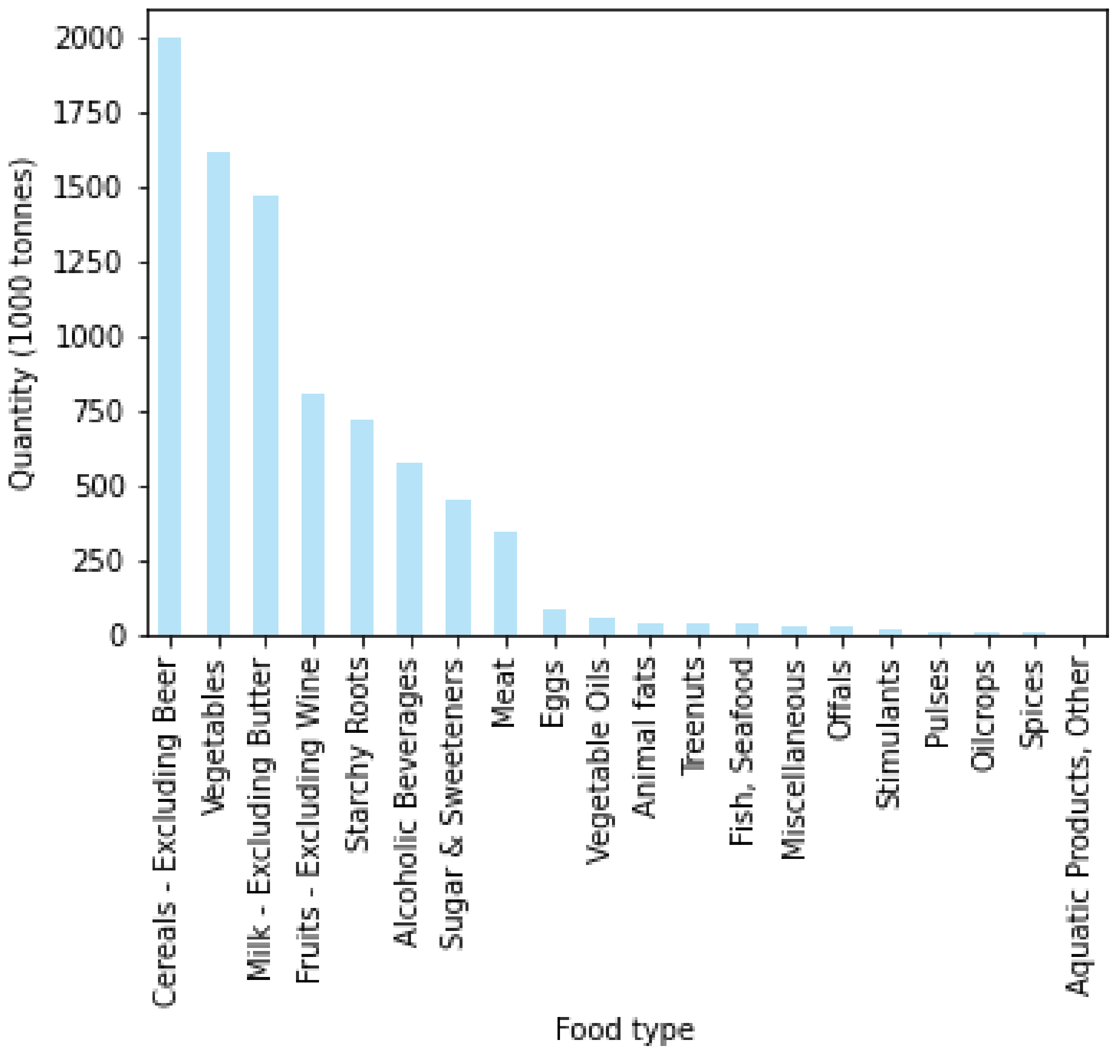

Figure 12. Food availability in Azerbaijan in 2018, in 1 million kilograms. Data: FAO (2019)

The food retail sector in Azerbaijan mostly consists of smaller shops and open markets, although in Baku, large-scale supermarket formats are gaining popularity (‘Market and competitiveness analysis of the Azerbaijan agricultural sector: An overview’, 2017). Therefore, retail data from supermarkets only include a part of the food consumed in Azerbaijan. However, in Baku, the retail data from supermarkets cover a larger part of consumption quantities. In the current study, retail data from various supermarkets in Baku represent the ground truth consumption rates for different ingredients in different sections of Baku.

3.3 Social media usage

4.3 million of Azerbaijan's 10.8 million inhabitants, equal to 42.2%, were active social media users in January 2021 (Kemp, 2021). This is close to the percentage of social media users worldwide, which is estimated at 53.6%. In addition, 8.26 million Azerbaijanis were internet users, which amounts to 81.1% of the country's population. This means that when looking at online and social media data, a large proportion but not all of Azerbaijan is represented. However, it is expected that in the capital, social media usage is higher and a closer approximation of the reference population.

The most popular social media sites in Azerbaijan, according to Kemp (2021), Instagram, with 3.5 million potential users, and Facebook, with 1.6 million potential users. According to a survey conducted by the Kaspersky Lab in 2018, Hajiyeva (2018), YouTube is even more popular, with 97% of survey respondents using it. For the current study, however, food images are collected from Foursquare, since I am interested in images rather than videos, and Instagram provides no good way of extracting location data.

Chapter 4

Methodology

The aim of this study is to perform fine-grained geographical mapping of food consumption by cross-modal analysis of Foursquare images and sets of ingredients. I did this by 1. applying a CNN pre-trained on a recipe-image dataset to novel images from Foursquare to detect ingredients, 2. inferring ingredient quantity from these images using the same model, 3. calculating ingredient proportions per local region (e.g. a city's neighbourhood), and 4. comparing these proportions to relative quantities of ingredient consumption provided by supermarket sales data.

4.1 Model design

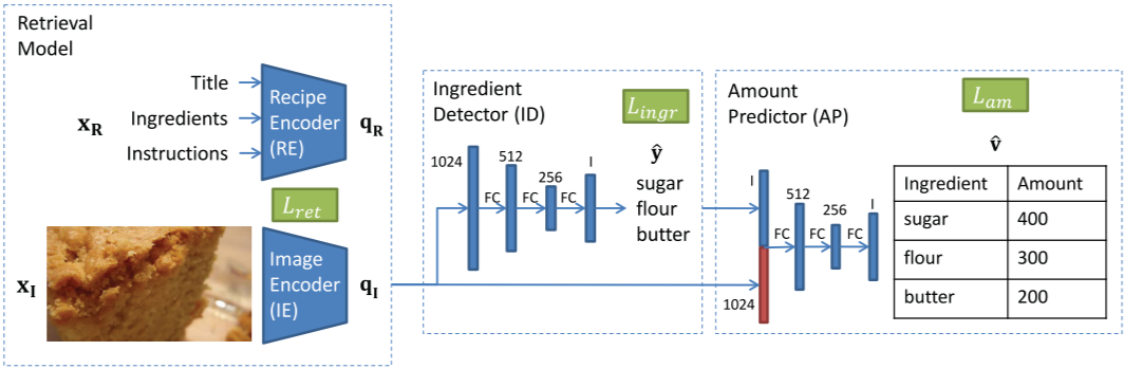

Figure 13. Schematic overview of the PITA ingredient detection and amount prediction model (Li et al., 2020b, p. 10344).

The CNN that is used to perform both ingredient detection and relative quantity estimation is an existing model from Li et al. (2020b), the Picture-to-amount (PITA) model. As visible in Figure 13, the model has a multi-stream architecture and consists of 3 parts. Firstly, it consists of a retrieval model, trained on the images and recipes from the Recipe1M (Salvador et al., 2017) dataset adjusted by Li et al. (2019). This retrieval model learns a cross-modal shared embedding space of the images and recipes. The retrieval model contains a text encoder, which is an LSTM, and an image encoder, based on a ResNet-50 model (He et al., 2016) minus the final fully connected layer. The objective of the retrieval model is to minimise cosine adversarial loss:

$$L_{ret} = \frac{1}{N}\Sigma_{i=1}^{N} max(0, m - \cos q_{R_i}, q_{I_i} + \cos q_{R_i}, u_j),\ i, j \in [1, N] \text{ and } i \neq j, \tag{4.1}$$

in which N is the batch size, m is the margin between positive and negative pair similarity, $cos(x, y)$ is the cosine similarity between x and y, q_{R_i} and q_{I_i} are respectively the text and image features of pair i, and u_j is a feature space that is negatively similar to q_{R_i}. The second part of the PITA model is an ingredient detection network, consisting of three fully connected layers, the first two with Leaky ReLU activation:

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ 0.01x, & \text{otherwise} \end{cases} \tag{4.2}$$

This ingredient detection model is trained to minimise positive sample weighted binary cross entropy loss between the true ingredients y_i and the predicted ingredients $\hat{y}_i$:

$$L_{id} = -\Sigma_{i=1}^{I}[w_i \hat{y}_i \log(p_i) + (1 - y_i) \log(1 - \hat{y}_i)], \tag{4.3}$$

in which w_i is the weight of the positive sample. The third part of the model is the amount network, consisting of three fully connected layers, the first two with Leaky ReLU activation, and the last one with softmax activation. This part of the model aims to minimise the Wasserstein distance between true ingredient amounts v and predicted ingredient amounts $\hat{v}$:

$$L_{ap} = W_p^p(\hat{v}, v, M), \tag{4.4}$$

where M is an ingredient distance matrix based on the semantic distance between ingredients. The model was trained by Li et al. (2020b) with Adam optimizer. The model is publicly available (Li et al., 2020a) and can be loaded with Python and PyTorch (PyTorch, 2019).

4.1.1 Training procedure

In this study, I did not use the text encoder, since I used the pretrained version of the model. With batch size b, image size s, and feature size f, each image batch was fed to the image encoder as a four-dimensional matrix B of size $\mathbb{R}^{b \times 3 \times s \times s}$, in which 3 denotes the number of channels per image (RGB), and within which each value stands for the pixel intensity at that point. The image encoder then returned a feature matrix of shape $\mathbb{R}^{b \times f}$. I then fed this batch of feature vectors to the ingredient detection network, resulting in ingredient matrix I of shape $\mathbb{R}^{b \times i}$, where the number of classes i (i.e., the number of ingredients) was 1362. In this matrix, each element stands for the probability that this ingredient is used to create the dish in the image. Then, I thresholded the matrix, i.e. replaced values below 0 with 0. Subsequently, I fed the thresholded matrix to the amount prediction network, resulting in amount matrix A of shape $\mathbb{R}^{b \times i}$, where each element denotes the proportional amount that the ingredient makes up of the total dish.

4.2 Data

The proposed study makes use of three data sources, of which two existing and one novel dataset. However, the existing datasets have been thoroughly preprocessed to suit the model architecture.

4.2.1 Recipe1M

The Recipe1M dataset is constructed by Salvador et al. (2017). The dataset contains more than a million cooking recipes and 800k images of food. These recipes and images are collected from more than 24 recipe websites containing recipes from around the world. For the current study, the 80k version of the Recipe1M dataset as produced by Li et al. (2019) was used. This dataset is different from Recipe1M in that it contains quantity information for ingredients in recipes, and similar ingredients are merged, resulting in 1362 ingredients.

In this study, the recipes and images of the Recipe1M dataset are not used: they were only used for training of the PITA model, conducted by Li et al. (2020b). However, the 1362 ingredients from Recipe1M are used, as they are the classes provided by the

PITA model. Li et al. (2020b) grouped these ingredients into 172 substitution groups, which I manually further fused into 62 compound ingredients, based on conceptual similarity. These compound ingredients are listed in Appendix 6.1. This was done firstly to provide a manageable number of categories for the manual annotation process, secondly because substitution groups such as "Jam, Marmelade" and "Jelly" were thought too conceptually or visually similar for a human annotator to be able to distinguish them in images, and thirdly because prediction mix-ups between these substitution groups were thought of little impact to food consumption mapping. In addition, I assigned each of the 172 substitution groups to one of the 10 merged categories in the supermarket sales dataset.

4.2.2 Supermarket sales

The original Azerbaijani supermarket sales dataset is constructed by Zeynalov (2020). This dataset contains more than 438k purchases from 21 supermarkets in 2019 across Baku, Azerbaijan. It contains 11 features: purchase id, product id, product name, product category, product price, purchase date and time, discount company, bonus card use, shop name, shop latitude, and shop longitude. The feature names, product names, and categories are in Azerbaijani.

As part of the current study, I created a new, English dataset based on the original Azerbaijani supermarket sales dataset. First, I manually translated the features and product categories. Then, I removed unnecessary features, namely discount company and bonus card use. In addition, I excluded categories that 1. denoted only drinks, 2. contained non-food items, or 3. described tobacco products, sweets, or chewing gum. Then, I used regular expressions to extract product information from the product names, namely brands, quantities, fat and alcohol percentages, and packaging, and save them as new features. The brands were extracted based on a list of brands collected from (Open Food Facts, 2022). In addition, I extracted unrecognised abbreviations based on a manually defined list.

Translation of product names was attempted in a number of ways. Firstly, I attempted to translate each product name as a whole using the unofficial Python Googletrans application programming interface (API) (Han, 2015). As this resulted in a large number of untranslated products, I then tokenised each product name into word-level uni-grams. These were then again translated from Azerbaijani to English by the unofficial Python Googletrans API (Han, 2015). Manual inspection proved that the translation quality was insufficient, and this seemed partly due to the presence of Russian (Latinised) and Turkish words in the dataset. I then proceeded to translate the untranslated tokens from Turkish, or, if no translation was found, from Russian, to English with the API. When manual inspection still showed insufficient translation, I repeated these steps with the official Google Cloud Translate API (Google, 2022). Unfortunately, the two translations were often contradicting each other and many tokens were still left untranslated. In addition to the presence of multiple languages in the dataset (some words were even in English or Italian), the absence of Turkic diacritics made that automatic translation proved difficult. In addition, support for Azerbaijani in automatic translation services is limited, perhaps due to the limited amount of speakers. Finally, a human translator with Turkish as native language and some knowledge of Azerbaijani manually translated the tokens, sometimes choosing one of the two API-generated translations, sometimes offering a new option. These translations were taken as final, although I in addition manually translated a large amount of frequently occurring words. This eventually resulted in 95.76% of product names being at least partially translated into English. After translation, some new tokens marked as brands were extracted. Subsequently, I removed English stop words, based on a list from (Bird et al., 2009). In addition, I stemmed and lemmatised the words, meaning that I reverted the English words back to their singular base form, their "lemma" (i.e., "conserved cherries" becomes "conserve cherry"). Lastly, I applied automatic spell-correction.

Based on gained knowledge of the types of product within each category, I then merged the originally 36 product categories into 10 categories, listed in Appendix 6.1. For each of these categories, I then matched the supermarket products to lemmatised and spell-corrected versions of the PITA ingredients within that category, and performed fuzzy string matching based on Levenshtein distance (Levenshtein, 1966) to select the closest matching PITA ingredient for each product. This all eventually accumulated into a dataset of 172137 purchases with 19 features.

4.2.3 AzerFSQFood

As image dataset, I collected photos from Foursquare.com using the Foursquare Places API (Foursquare, 2021b). This API returns places in a specified location with user-posted photos geographically tagged with those places. I repeatedly searched for "Dining and Drinking" places in each of the supermarket locations within Baku, Azerbaijan, with a radius of 9 km. The repetition was needed because the API returns a maximum of 50 places at once and I aimed for collecting as many eating establishments as possible to get the highest number of images. After removing duplicates, this resulted in 217 eating establishments. To each place, I then applied the Foursquare Get Place Photos API (Foursquare, 2021a) to collect a maximum of 50 images containing food and drinks taken at each restaurant. This resulted in 3225 images. I then manually filtered out photographs that 1. did not contain largely visible unpacked food items, 2. were of bad quality in terms of lighting or resolution (i.e. when I as human annotator was not able to visually identify the food items in the image), 3. were duplicates, and 4. were non-real photographs (cartoons or collages). After removing eating establishments without images, the resulting dataset contains 177 eating establishments with 2644 images.

The photographs in the dataset contain one or more dishes, sometimes accompanied by drinks or other non-food items. No cropping or repositioning of the images was applied, meaning that the images vary in object-to-background ratio and object positioning. In addition, lighting condition and resolution differ within the dataset.

I manually annotated a random stratified sample of approximately 20% of all images with the ingredient compounds created from the PITA ingredients. For each image, I listed the ingredient compounds that I estimated to be involved in the making of the dish, in order of quantity. This resulted in 381 annotated images. Subsequently, a second annotator annotated a sample of approximately 20% of the first annotated sample, resulting in 78 doubly annotated images. The second annotation was done to check the quality of the annotation, through inter-annotator reliability, measured by Krippendorf's alpha (Krippendorf, 1989), as reported in Section 5.2.1.

4.3 Ethical considerations

The conducted study concerns various ethical dilemmas. Firstly, web-scraping comes with privacy concerns. Luckily, no human faces were included in the photographs, since the Foursquare API only retrieved images depicting food. Secondly, the conducted data-driven analysis only offers the ability to detect patterns in data rather than enabling explanations as to why these patterns have developed (Kar & Dwivedi, 2020). The model employed in this study might suffer from overfitting, and external validity may be lowered, limiting its generalisability (Kar & Dwivedi, 2020). Therefore, in this study I compare the quantitative findings to the cultural theory explained in Section 3. In addition, no conclusions on Azerbaijani food consumption habits are drawn. Rather, inferences are focused on the applicability of cross-modal Foursquare food image analysis to mapping food consumption and the applicability of the model to other food image sources. Thirdly, because of my Dutch background, data from other countries are handled carefully and with attention to potential bias. It is for example important to realise that food consumption culture, such as what and where is consumed and how food is produced, differs between countries. Moreover, since translation is subjective, translating datasets might introduce bias. It is therefore important to note that the study, although set up to be unbiased and objective, always to some degree reflects the (subconscious) views of the researcher.

Chapter 5

Experiments and results

5.1 Experiments

I applied Li et al. (2020b)'s PITA model to food images of the novel AzerFSQFood dataset, to determine whether this pretrained model, selected on the basis of a literature study into food detection models, is able to handle real-life raw datasets of food images. This model allows for inferences of the type $p(\hat{g}_i|I)$, where I is a food image and $\hat{g}_i$ the predicted amount of the i^{th} ingredient, given by $\hat{g}_i = \frac{C}{M}m_i$, with M as the total mass of all ingredients, C as a normalising constant, and m_i as the mass of ingredient i. The model is tested by manual annotation of a 20% test dataset obtained from the AzerFSQFood dataset.

To analyse the usability of online food photographs taken at eating establishments to geographically map food consumption, I compared the ingredients depicted in these photographs to supermarket sales data on these ingredients. Based on the geographical extremes of the eating establishments in AzerFSQFood, Baku is divided into 25 grid spaces of approximately 4 by 5 km each, as shown in Figure 14.

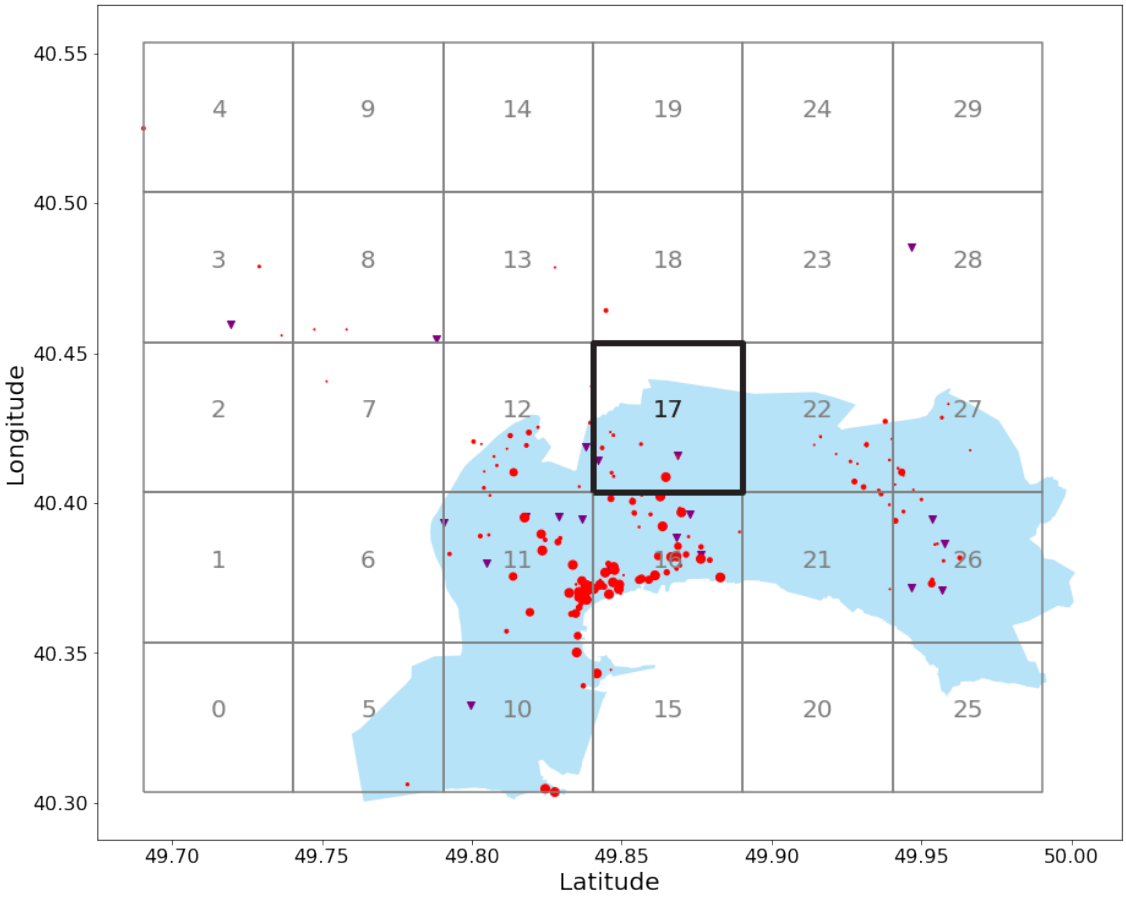

Figure 14. Map of the city centre of Baku, Azerbaijan (in blue) with the supermarket locations (purple triangles), the eating establishments (red circles, scaled according to amount of ingredients), and the grids (grey), with grid 17 highlighted (black).

For grid number 17, highlighted in Figure 14, I annotated all photographs associated with eating establishments situated in this area. Subsequently, I determined per ingredient compound its amount proportion, equal to the sum of all occurrences of this ingredient in across the annotated subset. In addition, I determined the proportion of each ingredient compound as featured in the purchases of the supermarkets located in grid 17. The relative amounts of these ingredient compounds as indicated by the supermarket sales and the online photographs were then compared.

5.2 Quantitative analysis

5.2.1 Inter-coder reliability

For the AzerFSQFood dataset, I scored inter-coder reliability with Krippendorf's alpha (Krippendorf, 1989) combined with Jaccard distance (Jaccard, 1901) as distance metric, respectively depicted in Equation 5.1 and 5.2:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{5.1}$$

and

$$J(A, B) = \frac{|A \cap B|}{|A \cup B}, \tag{5.2}$$

in which D_o is the observed disagreement and D_e is the expected chance disagreement between coders. I chose this method over Cohen's and Fleiss kappa and other distance metrics, because Krippendorf's alpha is well suited for multi-class and multi-label data, and Jaccard distance can be used to compare two unordered sets (Artstein & Poesio, 2008). This resulted in a Krippendorf's Alpha of 0.458, which signifies an inter-coder reliability of moderate size, since Krippendorf's Alpha ranges from zero (chance agreement) to one (complete agreement).

5.2.2 Model performance

I assess the performance of the PITA model (Li et al., 2020b) on the AzerFSQFood dataset through various metrics. I calculate these metrics for the annotated subset of the AzerFSQFood dataset, where the annotated (#62) ingredient compound labels Y are the true labels, and the model-generated labels $\hat{Y}$ are the predicted labels. First, I calculate the subset accuracy or exact match ratio using SciKit-learns function "accuracy_score" (Pedregosa et al., 2011), which calculates accuracy according to this function:

$$Acc = \frac{1}{n} \Sigma_{i=1}^{n} I(Y_i = \hat{Y}_i), \tag{5.3}$$

where n is the number of samples, Y_i is the true value for sample i and $\hat{Y}_i$ is the prediction. In this formula, only completely identical label sets are accepted: i.e., only if the predicted labels $\hat{y}$ for sample x are all equal to the true labels in y, sample x is marked as correctly predicted. This strict method resulted in an accuracy of 0.0%, meaning that none of the predicted ingredient label sets were identical to the true ingredient label sets in the annotation.

For a more complete picture of the model performance, I use precision, recall, and F1-score. Firstly, I calculate recall, which corresponds to a metric proposed in Li et al. (2020b), the coverage of ground truth ingredient (CVG):

$$CVG = \frac{Y \cap \hat{Y}}{Y}, \tag{5.4}$$

in which the number of common ingredient compound labels between ground truth set y and predicted set $\hat{y}$ is divided by the total number of labels in the ground truth. The micro-average of the recall is 0.0671, meaning that a (class-based) average of 6% of ground truth labels was also found in the prediction, indicating low model performance on the AzerFSQFood data regarding ingredient detection.

In Figure 15, the recall for each of the ingredient compounds for which recall > 0 is shown. As depicted, recall is best for sugar. These results should be regarded in relation to the confusion matrices of these ingredients, as they indicate the balance of labels in the annotated and predicted data. The confusion matrices are shown in Figure 16.

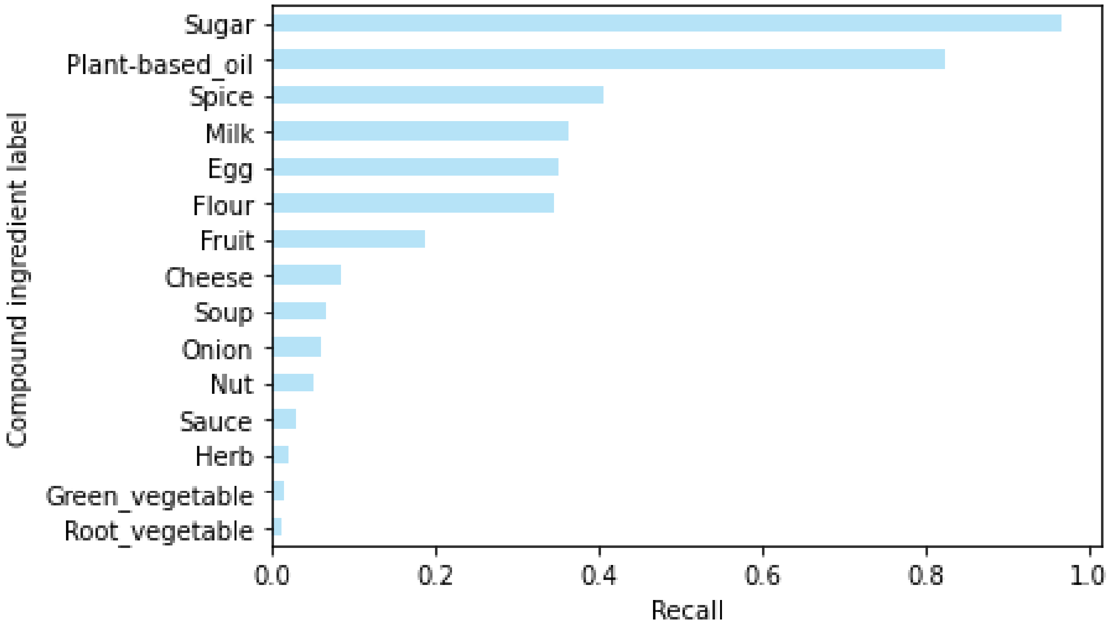

Figure 15. Recall for compound ingredients with recall over 0

Secondly, I calculate precision:

$$Precision = \frac{Y \cap \hat{Y}}{\hat{Y}} \tag{5.5}$$

This gives a micro-average of 0.071, meaning that a (class-based) average of 7% of predicted labels was also found in the ground truth, indicating that the model often reports ingredients not indicated by the human annotators.

Thirdly, I calculate F1 score:

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{5.6}$$

The micro-F1 score is 0.068. With a range of zero to one (one being an optimal F1 score), this means that the model is unable to correctly label ingredient classes on the images: it both reports ingredients that are not present and fails to recognise ingredients that are in the dish.

Figure 16 shows the confusion matrices for the PITA model on AzerFSQFood for each of the ingredient compounds. As we can see, some ingredients such as green vegetables have more false negatives, whereas for other ingredients, such as flour, false positives are more abundant. In addition, some ingredients, such as mint, are according to the annotators never present in the ground truth dataset. Rare ingredients such as these are harder to detect since the model is penalised for false positives and thus will not quickly report them. These rare ingredient compounds are kept in the analysis to be able to make quantitative comparisons between the AzerFSQFood ingredient proportions and the supermarket sales ingredient proportions, to explore food consumption patterns in Baku.

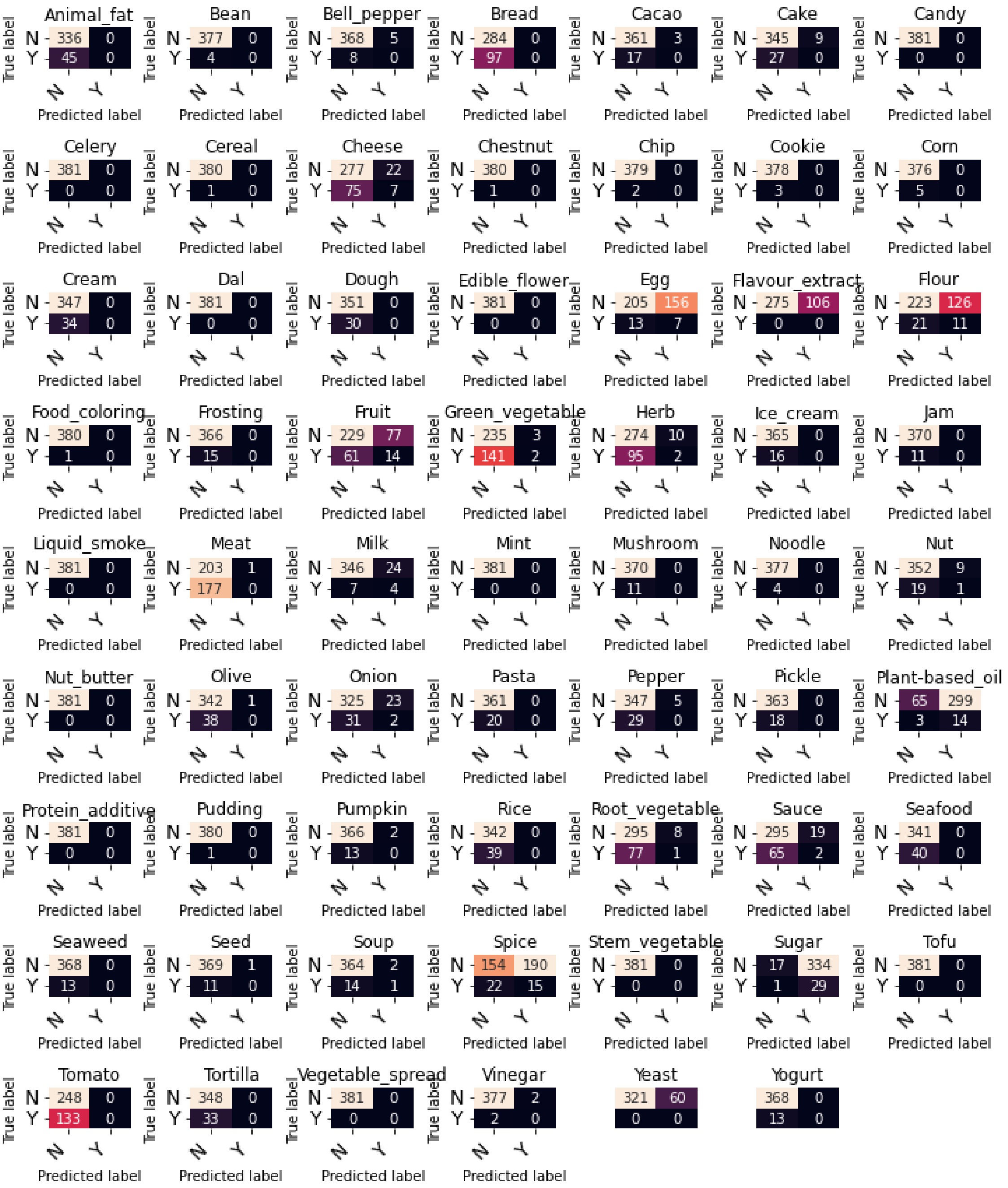

Figure 16. Confusion matrices for all 62 compound ingredients

As done by Li et al. (2020b), I also calculate the intersection over union (IOU), also named the Jaccard index (Jaccard, 1901). This metric is defined as the common number of ingredient compound labels divided by the union of ground truth and predicted labels:

$$J(Y, \hat{Y}) = \frac{Y \cap \hat{Y}}{Y + \hat{Y} - Y \cap \hat{Y}} \tag{5.7}$$

The micro-average of the IOU or Jaccard index is 0.035. This means that a class-based average of 3.5% of labels are found both in the ground truth and predicted datasets, meaning that there is an average 96.5% of labels that are either not detected or falsely predicted. This indicates that the model performs poorly at recognising the correct ingredients in the AzerFSQFood images.

To measure the amount prediction performance, I calculate the Kendall tau statistic (Kendall, 1945) for each sample for which the common labels are more than one. The Kendall tau statistic is defined as:

$$\tau = \frac{P - Q}{\sqrt{(P + Q + T) * (P + Q + U)}}, \tag{5.8}$$

in which P and Q are the number of concordant and discordant pairs, respectively, and T and U the number of pairs only in Y and $\hat{Y}$, respectively. I calculate this statistic with Scipy's "kendalltau" function (Virtanen et al., 2020). The unweighted macro average of the tau statistic for these samples is 0.037, for a range between -1 (completely different sets) and 1 (identical sets). This means that on average, the true and predicted ingredient label sets are moderately equally ordered.

5.2.3 Supermarket-restaurant correlation

I compare the supermarket sales data to annotations of the AzerFSQFood dataset. I do this for grid 17, as specified in the experiments. I determine the Pearson correlation statistic r as defined in Equation 5.9, in which $cov(X, Y)$ is the covariance between X and Y, and σ is the standard deviation.

$$r = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \tag{5.9}$$

The Pearson correlation, calculated with SciPy (Virtanen et al., 2020), is 0.381 with p-value 0.006. This is a moderate correlation (ranging from 0 to 1) that is significant at the significance level α of 0.01.

Figure 17 shows the ingredient proportional amounts for grid 17 from the annotated AzerFSQFood set and the supermarket sales data. As visible, supermarket-annotation similarities for ingredient compounds such as cream and pepper are quite high, whereas for compounds such as Yogurt and Fruit, they are very low. In the AzerFSQFood dataset, meat and green vegetables are the most abundant, whereas in the supermarket sales dataset, fruit and meat are the most present.

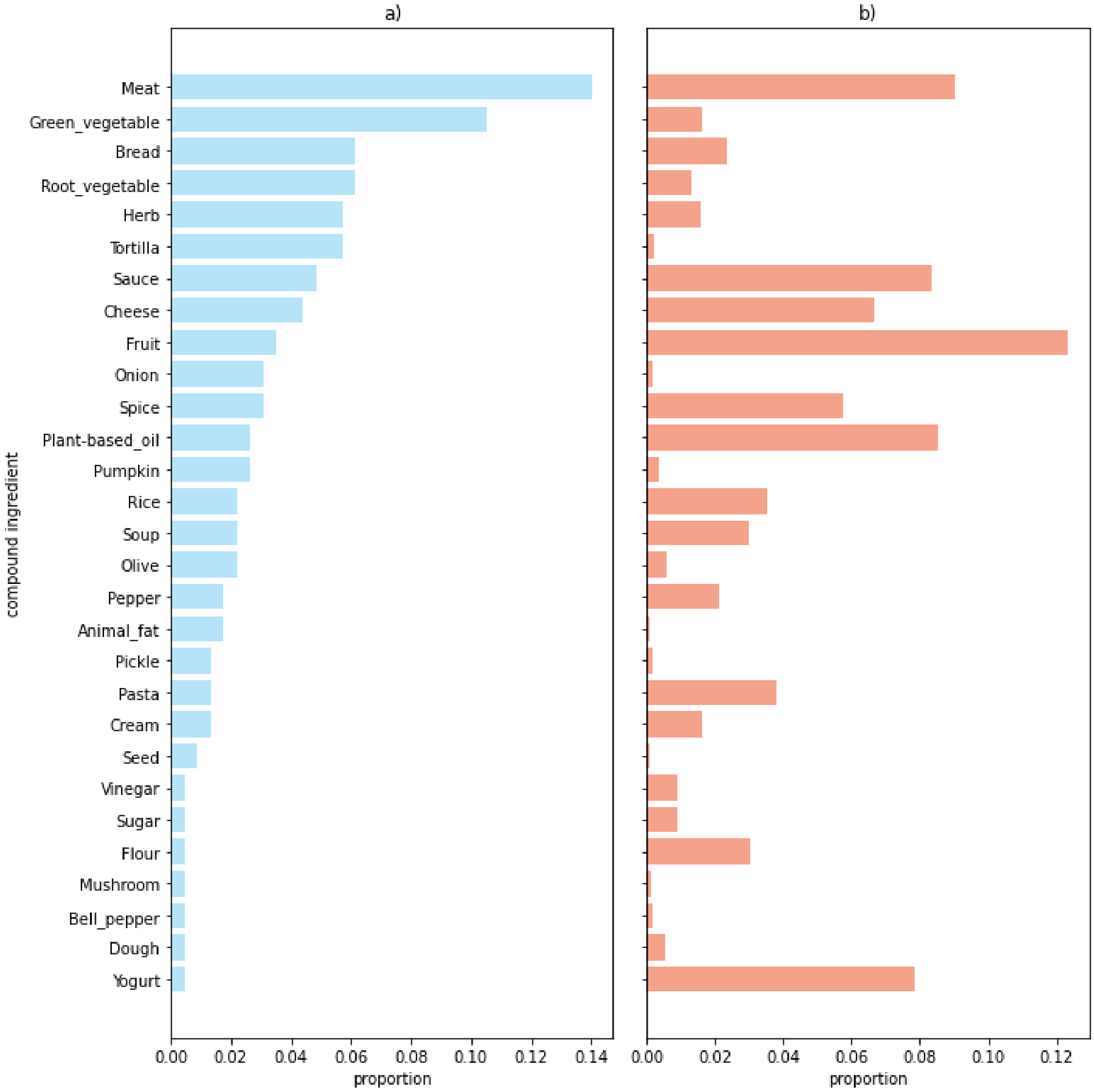

Figure 17. Proportional ingredient compound amounts for grid 17 from a) annotated FSQ images and b) supermarket sales.

The FAO balance sheet dataset, as mentioned, is currently the state-of-art in food consumption mapping. The data in those sheets available on Azerbaijani food consumption indicated that the most consumed products (in terms of kilograms) are cereals, vegetables, and milk products, whereas the least consumed ingredients are oil crops, spices, and aquatic products, as depicted in Figure 12. This is partially reflected in the AzerFSQFood ingredient proportions, with meat and green vegetables being the first second most consumed ingredients. In the supermarket sales data, although green vegetables took up a smaller proportion, meat and yogurt (a milk product) do show up as among the most consumed.

5.3 Qualitative analysis

To explore the PITA model performance more in depth, I perform a qualitative analysis of the ingredient detection and amount prediction results. Figure 18 depicts two images: one with the best possible Kendall τ value, one with the worst. The labels are ranked in order of quantity (1=highest relative amount). We can see that for a less complicated dish, the prediction is better, although far from perfect.

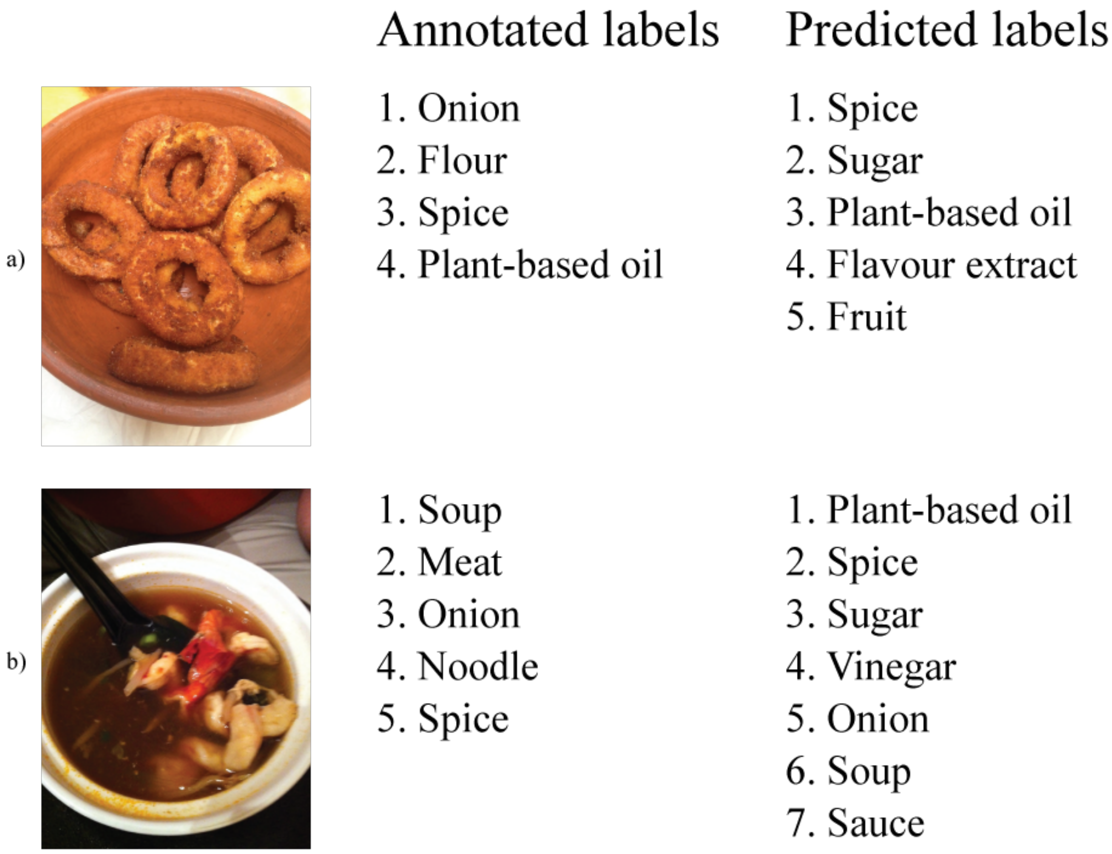

Figure 18. Annotated and predicted label sets for images with a) best and b) worst Kendall τ for ordered label sets, of respectively 1 and -1.

To demonstrate the application of the proposed model to map food consumption in a spatially high-resolution manner, I present Figure 19. This shows for one ingredient, in this case meat, the proportional amount of supermarket sales per grid space, projected on a map of Baku (grey) and its city centre (blue). Grid spaces without supermarkets or meat consumption are transparent. The Figure indicates that there might be differences in meat consumption on a small geographical scale.

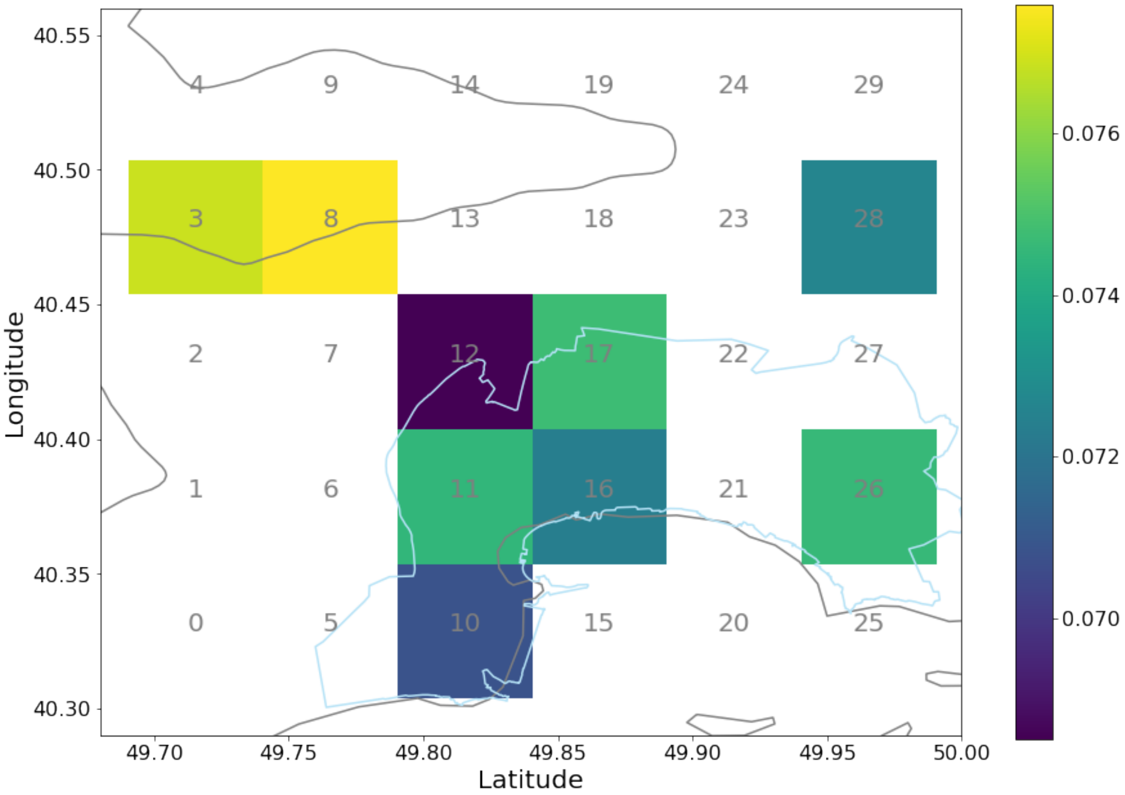

Figure 19. Choropleth of proportional amount of meat supermarket sales for each grid space with supermarket sales data.

Chapter 6

Discussion and conclusion

In the present study, I researched the applicability of a computer vision model on online images to geographically map food consumption. To that end, I applied an existing model, namely the PITA model developed by Li et al. (2020b), on a novel dataset, the AzerFSQFood dataset. This dataset contains food images collected from Foursquare from eating establishments around Baku, Azerbaijan. To assess the usability of online food images for mapping food consumption, I compared the human-annotated relative ingredient amounts to proportional ingredient amounts of a heavily preprocessed dataset on supermarket sales in Baku. I found that the PITA model, in contrast to its performance on the Recipe1M dataset (Salvador et al., 2017), reaches low subset accuracy, precision, and recall on the AzerFSQFood dataset. In addition, I found that the correlation between proportional ingredient amounts from supermarket sales data and from annotated AzerFSQFood images is significant.

The performance of the PITA model on the AzerFSQFood dataset is measured with subset accuracy, recall, precision, F1 score, Jaccard index, and the Kendall tau statistic. The extremely low score on the subset accuracy means that no ingredient sets were perfectly recognised. In addition, only a small proportion of ground truth ingredients were recognised well, as exemplified by the micro-average of recall. Moreover, the low precision score indicates a large percentage of false positives. These latter two scores are combined into the micro-F1 score, which is of a similar order of magnitude. The Jaccard index gives an idea of the proportion of ingredients that are both in the ground truth and predicted sets over the total amount of ingredients in both sets. This score, too, is quite low, indicating again that a low proportion of ingredients is detected. Lastly, the Kendall tau statistic indicates the quality of amount prediction, by comparing the ordered ingredient sets. The low score of this statistic exemplifies the low quality of amount predictions.

The low effectiveness of the PITA model on the AzerFSQFood dataset can be explained through a variety of factors related to the dataset: firstly, through the quality of the AzerFSQFood data annotation; secondly, by the multitude of images featuring multiple dishes; thirdly, by the presence of new food items and dishes in the AzerFSQFood dataset that were less featured in the Recipe1M dataset. Alternatively, the low accuracy of the model can be explained through features of the PITA model itself: firstly, through the lack of training of the PITA model on the AzerFSQFood dataset; secondly, through the low coding quality of the PITA model. In the paragraphs below, I will elaborate on each of these factors.

The AzerFSQFood dataset annotation showed low intercoder reliability. This indicates that the ground truth labels used in assessing the performance of the PITA model are not very reliable. The high level of disagreement between the coders might be caused by the difficulty of distinguishing all ingredients used in a dish, not just the visible ones. This offers another explanation, next to some ingredients being less frequently used than others, as to why some ingredient compounds are much more often labelled by the coders (as visible in the confusion matrices in Figure 16). In the future, when studies deal with a coding task of such high difficulty, firstly, more than two coders should be used; secondly, there should be strict rules on the definition of each ingredient category; and thirdly, a pilot coding session should be conducted to resolve any disputes.

In the AzerFSQFood dataset, as opposed to in the Recipe1M dataset, many images feature more than one dish. As described in Section 2.3, labelling multiple dishes in an image is a more difficult task than labelling one dish. Since the PITA model was trained on the Recipe1M dataset, it was not trained to recognise multiple dishes in one image. Although the model performs multi-labelling, this is in terms of ingredients, which are dependent labels, since they are used together in the same dish. The labelling of multiple dishes, however, involves detecting independent labels in one sample/image, which is a different task. In the present study, when annotating the AzerFSQFood images containing multiple dishes, no distinction was made between ingredients in one dish and ingredients in another dish. In future research dealing with raw datasets such as images scraped from online platforms, we should either perform image segmentation or train the model to recognise multiple dishes per image and then recognise for each dish the involved ingredients.

The low performance of the PITA model on the AzerFSQFood dataset could also be specific to the type of food depicted in the AzerFSQFood images. Since the photographs in this dataset were taken at eating establishments around Baku in Azerbaijan, they predominantly feature Azerbaijani dishes, although they do include international dishes such as American fast food and Japanese sushi. These Azerbaijani dishes might not have been included in the training data of the model, the Recipe1M dataset (Salvador et al., 2017). This means that the model is not as accustomed to recognising the ingredients and dishes featured in the AzerFSQFood dataset. To combat this, large and therefore valuable publicly available food image datasets such as Recipe1M+ (Marin et al., 2021) should be expanded with newly collected image and recipe data from other (virtual and geographical) locations. The dataset presented in this study, AzerFSQFood, is a good addition to this, but requires improved manual annotations.

The low model performance could alternatively be due to the lack of training and fine-tuning of the PITA model on the AzerFSQFood dataset. Because the aim of the research was to check the applicability of the PITA model on raw, novel datasets in order to assess its usefulness in fast-paced geographical mapping of food consumption, the model was not trained on the new dataset. In practise, decision makers wanting to get an updated image of what food consumption looks like in new geographical areas will not train a new model, but rather will want to apply an existing model that is highly abstracted. In addition, due to time constraints of the present research project, further improvement of the PITA model through training was not possible. Computer vision studies aiming to create an optimal food recognition and amount detection model should in the future train and fine-tune their models on large and varied datasets, in order to create the most flexible type of model which can then be applied to novel data.

In the present study, I chose to use the PITA model developed by Li et al. (2020b) based on it being the only publicly available model performing both multi-label ingredient detection and relative amount estimation. However, when working with the model, it became clear that the model lacked something in terms of coding quality. Despite the data and the code being openly accessible (Li et al., 2020a), the code is unstructured and undocumented. This means that it is hard to pinpoint whether the most up-to-date version of the model was used, and whether some alterations to the code were still necessary. Still, their model was of great value to this research, because it was publicly available and performed relative amount estimation. The fact that it was so difficult to find a publicly available model for the purposes of this study underlines the importance of open research and code and data sharing. This allows researchers in the field to build on each other and to progress faster.

I intended to measure correlation between pita prediction and supermarket sales in terms of relative ingredient amounts, but this was not deemed informative because of the bad performance of the PITA model. Therefore, I looked at the correlation between an annotated subset of the AzerFSQFood data and a subset of the supermarket sales data, namely that of grid 17 14. Unfortunately, using the correlation meant that the amount comparison had to be done in a count-based manner, rather than volume-based estimations as given by both the PITA predictions and the supermarket sales. The moderate but significant correlation between supermarket sales and food images within this geographical space regarding relative ingredient amounts indicate that there is indeed a relationship between what people buy in supermarkets and what people eat in restaurants. This relationship might be partially causational, since eating establishments might buy their supply in local supermarkets. Interestingly, there is a discrepancy between the most bought items in supermarkets and in the images of AzerFSQFood, as shown by Figure 17. This is reflected by the correlation being relatively small, and might be explained by the fact that the supermarkets featured in the supermarket sales dataset do not sell all food categories. For instance, seafood (fish and shellfish) were not included in the supermarket sales data, but were found in the AzerFSQFood images. This might be due to one or multiple factors: Azerbaijani people could eat different foods in eating establishments than at home; they might buy products such as seafood in separate stores or food markets; or there could be a distinction between courses people tend to eat at home (breakfast and lunch) and courses people tend to eat at restaurants (dinner).

Another important factor about the supermarket sales dataset is its extensive preprocessing, including natural language translation. Because the dataset contained words in Azerbaijani, Turkish, and Russian, amongst others, translation was difficult and not automated. In addition, there is no good Azerbaijani translation service available on the internet. Therefore, translation was eventually done by a native Turkish speaker, a native Russian speaker, and myself, a native Dutch speaker, none of whom speak Azerbaijani. Then, the translated products were compared in a category-based manner to the PITA ingredient labels, and for each product, the closest PITA label in terms of edit distance was chosen. These two processes (manual translation and fuzzy string matching) mean that the English supermarket sales dataset as used in this paper is not directly derived from its Azerbaijani original. This means that it is not possible to draw conclusive inferences about relative food consumption in Baku based on the supermarket data. However, this study assumes that the English supermarket sales dataset is a good enough reflection of the Azerbaijani supermarket sales to give an idea of relative ingredient consumption and to make comparisons with annotated Foursquare images.

The comparison between the supermarket sales and eating establishment ingredient proportion data is further complicated by the unbalanced nature of the classes. Some ingredient compounds are never labelled in the ground truth data of AzerFSQFood, whereas others are found in the majority of images. This means that rare ingredients are not only harder to detect with the computer vision model, but that comparisons with the supermarket dataset are also skewed, since just a couple of occurrences of these rare labels in the supermarket data will cause a discrepancy with the AzerFSQFood data. It would be better for future studies to either focus on a balanced set of ingredients or penalise the computer vision model more for low recall than for precision, in order to detect rare ingredients. In addition, quantitative comparisons between data sources in terms of relative ingredient amount should be done on balanced datasets so as not to skew the correlation's results.

It was hypothesised that the most eaten ingredients in Baku according to the supermarket sales and AzerFSQFood data would correlate with the most eaten ingredients according to FAOstat's balance sheets. Qualitative comparisons showed that although some of the most consumed ingredient indeed aligned between data sources, there was also a discrepancy. This might as well be caused by Azerbaijani people buying products such as seafood in separate stores or food markets. However, since the FAOstat food dataset contains food categories that represent raw ingredients such as cereals, whereas the ingredient compounds consist of processed foods such as pasta, it is not possible to draw a direct (quantitative) comparison between the three data sources.

A last limitation of the present study is the information loss caused by re-coding the 1362 PITA ingredients and their 172 substitution groups into 62 ingredient compounds. This was done to facilitate annotation speed, in accordance with the time limit of the current project. However, because of this, within-group prediction errors were not included in analysis. In a future, more extensive project, it would be better to look at the 172 substitution groups again. This will also make comparison with Li et al. (2020b)'s results easier.

In short, the study showed that the predictive value of the PITA model on the AzerFSQFood dataset was low, both in terms of ingredient detection and relative amount prediction. This might be caused by shortcomings of the AzerFSQFood dataset and its annotation, indicated by low inter-coder reliability, or by limitations of the PITA model and its (lack of) training. In addition, the study showed a correlation regarding relative ingredient amount consumption between supermarket sales and Foursquare food images around Baku, Azerbaijan, meaning that these data sources could indeed be used alongside each other to geographically estimate food consumption. In the future, researchers should improve the quality of the AzerFSQFood dataset regarding both images and annotation, and fine-tune a cross-modal ingredient detection and amount prediction model.

6.1 Conclusion

The current study aims to explore the usability of online images to automate fine-grained geographic mapping of food consumption. To that end, an existing cross-modal machine learning model trained on a recipe-image dataset to detect food ingredients and estimate their relative amounts was selected from available sources. This model was applied to a novel dataset created in this study, AzerFSQFood. This dataset consists of food images taken at eating establishments at Baku, Azerbaijan, and collected from Foursquare. In addition, manually annotated relative ingredient amounts in the AzerFSQFood dataset were compared to relative ingredient amounts in supermarket sales data from Baku. The model proved to have low performance, both in terms of ingredient detection and relative amount estimation, on the AzerFSQFood dataset, indicating that its usability for the proposed problem is low. However, a moderate correlation between supermarket sales data and annotated data from eating establishments indicates that these data sources are interesting to further explore in geographical food consumption mapping. In addition, this study highlights the importance of open source research data and model source code.

Bibliography

Abbar, S., Mejova, Y. & Weber, I. (2015). You tweet what you eat: Studying food consumption through Twitter. *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3197–3206. https://doi.org/10.1145/2702123.2702153

Ahn, Y. Y., Ahnert, S. E., Bagrow, J. P. & Barabási, A. L. (2011). Flavor network and the principles of food pairing. *Scientific Reports*, *1*, 1–7. https://doi.org/10.1038/srep00196

Anthimopoulos, M. M., Gianola, L., Scarnato, L., Diem, P. & 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. https://doi.org/10.1109/jbhi.2014.2308928

Artstein, R. & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

AzerbaijanTravelInternational. (2021). Azerbaijani cuisine (6 famous Azerbaijani dishes) [Accessed: 03-12-2021]. https://ati.az/food/azerbaijani-cuisine/

Baxter, J., Caruana, R., Mitchell, T., Pratt, L. Y., Silver, D. L. & Thrun, S. (1995). Learning to learn: Knowledge consolidation and transfer in inductive systems [Accessed: 02-12-2021]. https://plato.acadiau.ca/courses/comp/dsilver/NIPS95%5C_LTL/transfer.workshop.1995.html

Beijbom, O., Joshi, N., Morris, D., Saponas, S. & Khullar, S. (2015). Menu-Match: Restaurant-specific food logging from images. *2015 IEEE Winter Conference on Applications of Computer Vision*, 844–851.

Bettadapura, V., Thomaz, E., Parnami, A., Abowd, G. D. & Essa, I. (2015). Leveraging context to support automated food recognition in restaurants. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 580–587.

Bianchi, C. & Mortimer, G. (2015). Drivers of local food consumption: A comparative study. *British Food Journal*, *117*(9), 2282–2299. https://doi.org/10.1108/bfj-03-2015-0111

Bird, S., Loper, E. & Klein, E. (2009). NLTK: Natural language toolkit.

Blake, M. K., Mellor, J. & Crane, L. (2010). Buying Local Food: Shopping Practices, Place, and Consumption Networks in Defining Food as "Local". *Annals of the Association of American Geographers*, *100*(2), 409–426. https://doi.org/10.1080/00045601003595545

Bossard, L., Guillaumin, M. & Gool, L. V. (2014). Food-101 – Mining discriminative components with random forests. *European Conference on Computer Vision*, 446–461.

Bre, F., Gimenez, J. & Fachinotti, V. (2017). Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings*, *158*, 1–23. https://doi.org/10.1016/j.enbuild.2017.11.045

Bugliarello, E., Cotterell, R., Okazaki, N. & Elliott, D. (2021). Multimodal pretraining unmasked: A Meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, *9*, 978–994. https://doi.org/10.1162/tacla00408

Can, M. F., Günlü, A. & Can, H. Y. (2015). Fish consumption preferences and factors influencing it. *Food Science and Technology*, *35*(2), 339–346.

Carlsson-kanyama, A. & Gonza, A. D. (2009). Potential contributions of food consumption patterns to climate change. *The American Journal of Clinical Nutrition*, *89*(5), 1704s–1709s. https://doi.org/10.3945/ajcn.2009.26736AA.1704S

Carvalho, M., Cadène, R., Picard, D., Soulier, L., Thome, N. & Cord, M. (2018). Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, 35–44. https://doi.org/10.1145/3209978.3210036

Cauchy, A. et al. (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, *25*, 536–538.

Chauvin, N. D., Mulangu, F. & Porto, G. (2012). *Food production and consumption trends in Sub-Saharan Africa: Prospects for the transformation of the agricultural sector* (No. *4*), United Nations Development Programme. https://doi.org/10.1080/10455752.2016.1245915

Chen, J. & Ngo, C.-W. (2016). Deep-based ingredient recognition for cooking recipe retrieval. *MM '16: Proceedings of the 24th ACM international conference on Multimedia*, 32–41.

Chen, J.-J., Ngo, C.-W., Feng, F.-L. & Chua, T.-S. (2018). Deep understanding of cooking procedure for cross-modal recipe retrieval. *MM '18: Proceedings of the 26th ACM international conference on Multimedia*, 1020–1028. https://doi.org/10.1145/3240508.3240627

Chen, J., Pang, L. & Ngo, C.-w. (2017). Cross-modal recipe retrieval: How to cook this dish? *Proceedings of the International Conference on Multimedia Modeling*, 588–600.

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

Chen, M.-y., Yang, Y.-h., Ho, C.-j., Wang, S.-h., Liu, S.-m., Chang, E. & Yeh, C.-h. (2012). Automatic Chinese food identification and quantity estimation. *SA '12: SIGGRAPH Asia 2012 Technical Briefs*, 1–4. https://doi.org/10.1145/2407746.2407775

Chen, X., Zhu, Y., Zhou, H., Diao, L. & Wang, D. (2017). ChineseFoodNet: A large-scale image dataset for Chinese food recognition. *ArXiv Preprint*, 1–8.

Chishala, B. H., Mofya-mukuka, R., Chabala, L. M. & Kuntashula, E. (2021). Zero hunger. In J. Gill & M. Smith (Eds.), *Geosciences and the sustainable development goals* (pp. 31–51). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-38815-7_2

Choudhury, M. D., Sharma, S. & Kiciman, E. (2016). Characterizing dietary choices, nutrition, and language in food deserts via social media. *CSCW '16: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1157–1170. https://doi.org/10.1145/2818048.2819956

CIA. (2021). Azerbaijan [Accessed: 03-12-2021]. https://www.cia.gov/the-world-factbook/countries/azerbaijan/

Ciocca, G., Napoletano, P. & Schettini, R. (2015). Food recognition and leftover estimation for daily diet monitoring. *Proceedings of the International Conference on Image Analysis and Processing*, 334–341. https://doi.org/10.1007/978-3-319-23222-5

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

Dados, N. & Connell, R. (2012). The global south. *American Sociological Association: Contexts*, *11*(1), 12–13. https://doi.org/10.1177/1536504212436479

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255. https://doi.org/10.1167/9.8.1037

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2020). About ImageNet [Accessed: 03-11-2021]. https://www.image-net.org/about.php

Deutsch, J. & Miller, J. (2007). Food studies: A multidisciplinary guide to the literature. *Choice Reviews Online*, *45*(03), 393–401. https://doi.org/10.5860/choice.45.03.393

Dinther, M. v. (2021). Ministerie vindt minder vlees eten te omstreden voor campagne over klimaatbewustzijn [Accessed: 19-10-2021]. *De Volkskrant*. https://www.volkskrant.nl/nieuws-achtergrond/ministerie-vindt-minder-vlees-eten-te-omstreden-voor-campagne-over-klimaatbewustzijn~bee34bfb/

Ege, T., Ando, Y., Tanno, R., Shimoda, W. & Yanai, K. (2019). Image-based estimation of real food size for accurate food calorie estimation. *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 274–279. https://doi.org/10.1109/mipr.2019.00056

Erp, M. V., Reynolds, C., Maynard, D., Starke, A., Martín, R. I., Andres, F., Leite, M. C. A., Toledo, D. A. D., Rivera, X. S., Trattner, C., Brewer, S., Martins, C. A., Kluczkovski, A., Frankowska, A., Bridle, S., Levy, R. B., Rauber, F., Tereza Da Silva, J. & Bosma, U. (2021). Using natural language processing and artificial intelligence to explore the nutrition and sustainability of recipes and food. *Frontiers in Artificial Intelligence*, *3*(621577), 1–8. https://doi.org/10.3389/frai.2020.621577

Fang, S., Shao, Z., Mao, R., Fu, C., Kerr, D. A., Boushey, C. J., Delp, E. J. & Zhu, F. (2018). Single-view food portion estimation: Learning image-to-energy mappings using generative adversarial networks. *25th IEEE International Conference on Image Processing (ICIP)*, 251–255.

FAO. (2019). Food balances (2014-) [Accessed: 19-10-2021]. https://www.fao.org/faostat/en/%5C#data/FBS

FAO. (2022). Crop, livestock and food statistics - methodology [Accessed: 13-10-2021]. https://www.fao.org/food-agriculture-statistics/statistical-domains/crop-livestock-and-food/methodology/en/

Farinella, G. M., Allegra, D., Moltisanti, M., Stanco, F. & Battiato, S. (2016). Retrieval and classification of food images. *Computers in Biology and Medicine*, *77*, 23–39. https://doi.org/10.1016/j.compbiomed.2016.07.006

Fontanellaz, M., Christodoulidis, S. & Mougiakakou, S. (2019). Self-attention and ingredient-attention based model for recipe retrieval from image queries. *MADiMa '19: Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management*, 25–31. https://doi.org/10.1145/3347448.3357163

Foursquare. (2021a). Get place photos API [Accessed: 26-04-2022]. https://developer.foursquare.com/reference/place-photos

Foursquare. (2021b). Places API [Accessed: 08-12-2021]. https://developer.foursquare.com/docs/places-api-overview

Gatt, A. (2021). What part of "understand" don't you "see"? Exploring the visual grounding capabilities of deep multimodal models.

Gauss, K. F. (1963). *Theory of the motion of the heavenly bodies moving about the sun in conic sections*.

Google. (2022). Google cloud translation [Accessed: 26-04-2022]. https://cloud.google.com/translate

Greenebaum, J. (2018). Vegans of color: Managing visible and invisible stigmas. *Food, Culture and Society*, *21*(5), 1–18. https://doi.org/10.1080/15528014.2018.1512285

Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R. & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 1–12. https://doi.org/10.1109/tnnls.2016.2582924

Hajiyeva, G. (2018). Which social media networks are most popular in Azerbaijan? [Accessed: 02-12-2021]. https://caspiannews.com/news-detail/which-social-media-networks-are-most-popular-in-azerbaijan-2018-12-22-30/

Han, S. (2015). Googletrans [Accessed: 26-04-2022]. https://github.com/ssut/py-googletrans

Hao, W., Sahoo, D., Liu, C., Lim, E.-P. & Hoi, S. C. (2019). Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 11564–11573. https://doi.org/10.1109/cvpr.2019.01184

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1002/chin.200650130

Herranz, L., Jiang, S. & Xu, R. (2017). Modeling restaurant context for food recognition. *IEEE Transactions on Multimedia*, *19*(2), 430–440.

Hoashi, H., Joutou, T. & Yanai, K. (2010). Image recognition of 85 food categories by feature fusion. *2010 IEEE International Symposium on Multimedia*, 296–301.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société vaudoise des sciences naturelles*, *37*, 547–579.

Jiang, S. & Min, W. (2020). Food computing for multimedia. *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, 4782–4784. https://doi.org/10.1145/3394171.3418544

Jordan, M. I. (1986). *Serial order: A parallel distributed processing approach* (tech. rep.). University of California, Institute for Cognitive Science. San Diego, CA, USA.

Joutou, T. & Yanai, K. (2010). A food image recognition system with multiple kernel learning. *Proceedings of the IEEE International Conference on Image Processing*, 285–288.

Jukema, G., Ramaekers, P. & Berkhout, P. (2021). *De Nederlandse agrarische sector in internationaal verband* (tech. rep.). Wageningen University and Research; CBS. https://www.rijksoverheid.nl/documenten/rapporten/2021/01/22/de-nederlandse-agrarische-sector-in-internationaal-verband

Kagaya, H. & Aizawa, K. (2015). Highly accurate food/non-food image classification based on a deep convolutional neural network. *International Conference on Image Analysis and Processing*, 350–357. https://doi.org/10.1007/978-3-319-23222-5_43

Kagaya, H., Aizawa, K. & Ogawa, M. (2014). Food detection and recognition using convolutional neural network. *Proceedings of the 22nd ACM international conference on Multimedia*, 1085–1088. https://doi.org/10.1145/2647868.2654970

Kar, A. K. & Dwivedi, Y. K. (2020). Theory building with big data-driven research – Moving away from the “what ” towards the “why ”. *International Journal of Information Management*, *54*(June), 102205. https://doi.org/10.1016/j.ijinfomgt.2020.102205

Karpathy, A. & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3128–3137. https://www.cv-foundation.org/openaccess/content%5C_cvpr%5C_2015/papers/Karpathy%5C_Deep%5C_Visual-Semantic%5C_Alignments%5C_2015%5C_CVPR%5C_paper.pdf%5C%0Ahttp://www.cv-foundation.org/openaccess/content%5C_cvpr%5C_2015/papers/Karpathy%5C_Deep%5C_Visual-Semantic%5C_Alignments%5C_2015%5C_CVPR%5C_paper.pdf

Kawano, Y. & Yanai, K. (2014). FoodCam-256: A large-scale real-time mobile food recognition system employing high-dimensional features and compression of classifier weights. *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, (1), 761–762. https://doi.org/10.1145/2647868.2654869

Kemp, S. (2021). Digital 2021: Azerbaijan [Accessed: 02-12-2021]. https://datareportal.com/reports/digital-2021-azerbaijan

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, *33*(3), 239–251.

Krippendorf, K. (1989). Content analysis. In E. Barnouw, G. Gerbner, W. Schramm, T. L. Worth & L. Gross (Eds.), *International encyclopedia of communication* (pp. 403–407). Oxford University Press.

Krittanawong, C., Johnson, K., Rosenson, R., Wang, Z., Aydar, M., Baber, U., Min, J., Tang, W., Halperin, J. & Narayan, S. (2019). Deep learning for cardiovascular medicine: A practical primer. *European heart journal*, *40*, 1–15. https://doi.org/10.1093/eurheartj/ehz056

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems 25*, 1097–1105. https://doi.org/10.1201/9781420010749

Krotov, V. & Silva, L. (2018). Legality and ethics of web scraping. *Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018*, 1–5.

Kusmierczyk, T., Trattner, C. & Nørvåg, K. (2016). Understanding and predicting online food recipe production patterns. *HT '16: Proceedings of the 27th ACM Conference on Hypertext and Social Media*, 243–248. https://doi.org/10.1145/2914586.2914632

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Lee, J.-g. & Kang, M. (2015). Geospatial big data: Challenges and opportunities. *Big Data Research*, *2*, 74–81. https://doi.org/10.1016/j.bdr.2015.01.003

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*, 707–710.

Li, J., Guerrero, R. & Pavlovic, V. (2019). Deep cooking: Predicting relative food ingredient amounts from images. *MADiMa '19: Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management*, 2–6. https://doi.org/10.1145/3347448.3357164

Li, J., Han, F., Guerrero, R. & Pavlovic, V. (2020a). Dataset [Accessed: 01-12-2021]. http://foodai.cs.rutgers.edu:2020/static/dataset.html

Li, J., Han, F., Guerrero, R. & Pavlovic, V. (2020b). Picture-to-amount (PITA): Predicting relative ingredient amounts from food images. *Proceedings - International Conference on Pattern Recognition*, 10343–10350. https://doi.org/10.1109/icpr48806.2021.9412828

Liang, Y. & Li, J. (2017). Computer vision-based food calorie estimation: Dataset, method, and experiment. *ArXiv Preprint*, 1–7.

Lin, M., Chen, Q. & Yan, S. (2013). Network in network. *ArXiv Preprint*, 1–10.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Lu, Y. (2016). Food image recognition by using convolutional neural networks (CNNs). *ArXiv Preprint*, 1–6. http://arxiv.org/abs/1612.00983

Marin, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I. & Torralba, A. (2021). Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images [Downloaded: 04-10-2021]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(1), 187–203. https://doi.org/10.1109/tpami.2019.2927476

Market and competitiveness analysis of the Azerbaijan agricultural sector: An overview. (2017). *Master Plan for Promoting Investments in the Azerbaijan Agricultural Sector*, 1–33.

Matsuda, Y., Hoashi, H. & Yanai, K. (2012). Recognition of multiple-food images by detecting candidate regions. *2012 IEEE International Conference on Multimedia and Expo*, 25–30. https://doi.org/10.1109/icme.2012.157

Mejova, Y., Haddadi, H., Noulas, A. & Weber, I. (2015). #FoodPorn: Obesity patterns in culinary interactions. *DH '15: Proceedings of the 5th International Conference on Digital Health 2015*, 51–58. https://doi.org/10.1145/2750511.2750524

Merler, M., Wu, H., Uceda-Sosa, R., Nguyen, Q.-B. & Smith, J. R. (2016). Snap, Eat, repEat: A food recognition engine for dietary logging. *Proceedings of the International Workshop on Multimedia Assisted Dietary Management*, 31–40.

Min, W., Bao, B.-k., Mei, S., Zhu, Y., Rui, Y. & Jiang, S. (2018). You are what you eat: Exploring rich recipe information for cross-region food analysis. *IEEE Transactions on Multimedia*, *20*(4), 950–964. https://doi.org/10.1109/tmm.2017.2759499

Min, W., Jiang, S., Liu, L., Rui, Y. & Jain, R. (2019). A survey on food computing. *ACM Computing Surveys*, *52*(5), 92:1–92:36. https://doi.org/10.1145/3329168

Min, W., Jiang, S., Sang, J., Wang, H., Liu, X. & Herranz, L. (2017). Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration. *IEEE Transactions on Multimedia*, *19*(5), 1100–1113.

Mouritsen, O. G., Edwards-Stuart, R., Ahn, Y. Y. & Ahnert, S. E. (2017). Data-driven methods for the study of food perception, preparation, consumption, and culture. *Frontiers in ICT*, *4*(15), 1–5. https://doi.org/10.3389/fict.2017.00015

Muilwijk, H., Westhoek, H. & De Krom, M. (2018). *Voedsel in Nederland: Verduurzaming bewerkstelligen in een veelvormig systeem* (tech. rep. april). PBL Planbureau voor de Leefomgeving. The Hague, Netherlands. https://www.

pbl.nl/sites/default/files/cms/publicaties/pbl-2018-notitie-voedsel-in-nederland-3239.pdf

Musaiger, A. O. (1993). Socio-cultural and economic factors affecting food consumption patterns in the Arab countries. *The Journal of the Royal Society for the Promotion of Health*, *113*(2), 68–74. https://doi.org/10.1177/146642409311300205

Myers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J. & Murphy, K. (2015). Im2Calories: Towards an automated mobile vision food diary. *Proceedings of the IEEE International Conference on Computer Vision*, 1233–1241.

Nakano, M., Sato, H., Watanabe, T., Takano, K. & Sagane, Y. (2018). Mining online activity data to understand food consumption behavior: A case of Asian fish sauce among Japanese consumers. *Food Science and Nutrition*, *6*(4), 791–799. https://doi.org/10.1002/fsn3.622

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. & Ng, A. Y. (2011). Multimodal deep learning. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, 689–696.

Nicolaou, M., Doak, C. M., van Dam, R. M., Brug, J., Stronks, K. & Seidell, J. C. (2009). Cultural and social influences on food consumption in Dutch residents of Turkish and Moroccan origin: A qualitative study. *Journal of Nutrition Education and Behavior*, *41*(4), 232–241. https://doi.org/10.1016/j.jneb.2008.05.011

Nwankpa, C. E., Ijomah, W., Gachagan, A. & Marshall, S. (2018). *Activation functions: Comparison of trends in practice and research for deep learning*. http://arxiv.org/abs/1811.03378

Ofli, F., Aytar, Y., Weber, I., Hammouri, R. A. & Torralba, A. (2017). Is Saki #delicious? The food perception gap on Instagram and its relation to health. *WWW '17: Proceedings of the 26th International Conference on World Wide Web*, 509–518. https://doi.org/10.1145/3038912.3052663

Okamoto, K. & Yanai, K. (2016). An automatic calorie estimation system of food images on a smartphone. *MADiMa 2016 - Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, co-located with ACM Multimedia 2016*, 63–70. https://doi.org/10.1145/2986035.2986040

Olsen, S. O., Scholderer, J., Brunsø, K. & Verbeke, W. (2007). Exploring the relationship between convenience and fish consumption: A cross-cultural study. *Appetite*, *49*, 84–91. https://doi.org/10.1016/j.appet.2006.12.002

Open Food Facts. (2022). Open food facts [Accessed: 26-04-2022]. https://world.openfoodfacts.org/

OWSD & UNESCO. (2022). Countries in the global south [Accessed: 18-05-2022]. https://owsd.net/sites/default/files/OWSD%5C%20138%5C%20Countries%5C%20-%5C%20Global%5C%20South.pdf

Pedregosa, K., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Phan, T.-T. & Gatica-Perez, D. (2017). #Healthy #Fondue #Dinner: Analysis and inference of food and drink consumption patterns on Instagram. *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*, 327–338.

Podareanu, D., Codreanu, V., Aigner, S., Leeuwen, C. & Weinberg, V. (2019). *Best practice guide - deep learning* (tech. rep.). PRACE. https://doi.org/10.13140/RG.2.2.31564.05769

Princeton University. (2021). WordNet: A lexical database for English [Accessed: 03-11-2021]. https://wordnet.princeton.edu/

Puri, M., Zhu, Z., Yu, Q., Divakaran, A. & Sawhney, H. (2009). Recognition and volume estimation of food intake using a mobile device. *WACV: 2009 Workshop on Applications of Computer Vision*, 1–8. https://doi.org/10.1109/wacv.2009.5403087

PyTorch. (2019). PyTorch [Accessed: 26-04-2022]. https://pytorch.org/

Rich, J., Haddadi, H. & Hospedales, T. M. (2016). Towards bottom-up analysis of social food. *Proceedings of the International Conference on Digital Health Conference*, 110–120.

Riesenhuber, M. & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Rina Dechter. (1986). Learning while searching in constraint-satisfaction-problems. *Proceedings of the 5th National Conference on Artificial Intelligence*, 178–183. www.aaai.org

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge university press.

Rivm. (2020). *Eet en drinkt Nederland volgens de richtlijnen Schijf van Vijf?: Resultaten van de voedselconsumptiepeiling 2012-2016* (tech. rep.). Rijksinstituut voor Volksgezondheid en Milieu. Bilthoven.

Robbins, H. E. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *22*, 400–407.

Rokicki, M., Trattner, C. & Herder, E. (2018). The impact of recipe features, social cues and demographics on estimating the healthiness of online recipes. *Pro-*

ceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018), 310–319.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hintont, G. E. & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.7551/mitpress/1888.003.0013

Russell, S. & Norvig, P. (Eds.). (2010). *Artificial intelligence: A modern approach* (3rd ed.). Pearson Education, Inc.

Said, A. & Bellogín, A. (2014). You are what you eat! Tracking health through recipe interactions categories and subject descriptors. *RSWeb 2014: Proceedings of the 6th Workshop on Recommender Systems and the Social Web*.

Salvador, A., Drozdzal, M., Giro-i-nieto, X. & Romero, A. (2019). Inverse cooking: Recipe generation from food images. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10453–10462.

Salvador, A., Hynes, N., Marin, J., Weber, I. & Torralba, A. (2017). Learning cross-modal embeddings for cooking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3068–3076. https://doi.org/10.1109/cvpr.2017.327

Schneider, U. A., Havlík, P., Schmid, E., Valin, H., Mosnier, A., Obersteiner, M., Böttcher, H., Skalský, R., Balkovič, J., Sauer, T. & Fritz, S. (2011). Impacts of population growth, economic development, and technical change on global food production and consumption. *Agricultural Systems*, *104*(2), 204–215. https://doi.org/10.1016/j.agsy.2010.11.003

Sekar, K., Gopinath, S., Sakthivel, K. & Lalitha, S. (2021). *Design and Implementation of a Deep Convolutional Neural Networks Hardware Accelerator*, *1964*, 1–7. https://doi.org/10.1088/1742-6596/1964/5/052008

Serre, T., Wolf, L. & Poggio, T. (2005). Object recognition with features inspired by visual cortex. *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, *2*, 994–1000. https://doi.org/10.1109/cvpr.2005.254

Sharma, S. S. & Choudhury, M. D. (2015). Measuring and characterizing nutritional information of food and ingestion content in Instagram. *WWW 2015 Companion: Proceedings of the 24th International Conference on World Wide Web*, 115–116. https://doi.org/10.1145/2740908.2742754

Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *Iclr*, 1–14. http://arxiv.org/abs/1409.1556

Singer, E. (2006). Introduction: Nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 637–645. https://doi.org/10.1093/poq/nfl034

Singla, A. & Yuan, L. (2016). Food/non-food image classification and food categorization using pre-trained GoogLeNet model. *MADiMa '16: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, 3–11.

StatGovAz. (2021). Population of Azerbaijan [Accessed: 03-12-2021]. https://www.stat.gov.az/source/demoqraphy/ap/

Stein, S. & McKenna, S. J. (2013). Combining embedded accelerometers with computer vision for recognizing food preparation activities. *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2013)*, 729–738.

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *91*(8), 1–9. https://doi.org/10.1109/cvpr.2015.7298594

Tao, D., Yang, P. & Feng, H. (2020). Utilization of text mining as a big data analysis tool for food science and nutrition. *Comprehensive Reviews in Food Science and Food Safety*, *19*(2), 875–894. https://doi.org/10.1111/1541-4337.12540

Transforming our world: The 2030 agenda for sustainable development (tech. rep.). (2015). United Nations. New York, NY, USA, Springer.

Verbeke, W. & Vackier, I. (2005). Individual determinants of fish consumption: Application of the theory of planned behaviour. *Appetite*, *44*, 67–82. https://doi.org/10.1016/j.appet.2004.08.006

Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. (2015). Show and tell: A neural image caption generator. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3156–3164. https://doi.org/10.1109/cvpr.2015.7298935

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Voedingscentrum. (n.d.). Schijf van vijf. https://www.voedingscentrum.nl/Assets/Uploads/voedingscentrum/Documents/Professionals/Schijf%20van%20Vijf/Schijf%20van%20Vijf%20poster%20met%20bollen%202019%20.pdf

Wagner, C., Singer, P. & Strohmaier, M. (2014). Spatial and temporal patterns of online food preferences. *WWW '14 Companion: Proceedings of the 23rd*

International Conference on World Wide Web, 553–554. https://doi.org/10.1145/2567948.2576951

Wang, X., Kumar, D., Thome, N., Cord, M. & Precioso, F. (2015). Recipe recognition with large multimodal food dataset. *IEEE International Conference on Multimedia & Expo (ICME)*, 1–7. https://doi.org/10.1109/icmew.2015.7169757

Watt, T. L., Beckert, W., Smith, R. D. & Cornelsen, L. (2020). Reducing consumption of unhealthy foods and beverages through banning price promotions: What is the evidence and will it work? *Public Health Nutrition*, *23*(12), 2228–2233. https://doi.org/10.1017/s1368980019004956

West, R., White, R. W. & Horvitz, E. (2013). From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. *World Wide Web*, 1–11.

Who. (2020). Healthy diet [Accessed: 19-10-2021]. https://www.who.int/news-room/fact-sheets/detail/healthy-diet

Xu, R., Herranz, L., Jiang, S., Wang, S., Song, X. & Jain, R. (2015). Geolocalized modeling for dish recognition. *IEEE Transactions on Multimedia*, *17*(8), 1187–1199.

Yagcioglu, S., Erdem, A., Erdem, E. & Ikizler-Cinbis, N. (2018). RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1–14.

Zeynalov, N. (2020). Supermarket-dataset [Downloaded: 14-10-2021]. https://www.kaggle.com/nicatheynal/supermarketdataset

Zhang, A., Lipton, Z. C., Li, M. & Smola, A. J. (2021). Dive into deep learning [Accessed: 03-11-2021]. *arXiv preprint arXiv:2106.11342*. https://d2l.ai/index.html

Zhou, F. & Lin, Y. (2016). Fine-grained image classification by exploring bipartite-graph labels. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1124–1133.

Zhu, Y.-x., Huang, J., Zhang, Z.-k., Zhang, Q.-m., Zhou, T. & Ahn, Y.-y. (2013). Geography and similarity of regional cuisines in China. *PLoS ONE*, *8*(11), 1–8. https://doi.org/10.1371/journal.pone.0079161

Appendix

Appendix A: Ingredient compound labels

1. Animal fat
2. Bean
3. Bell pepper
4. Bread
5. Cacao
6. Cake
7. Candy
8. Celery
9. Cereal
10. Cheese
11. Chestnut
12. Chip
13. Cookie
14. Corn
15. Cream
16. Dal
17. Dough
18. Edible flower
19. Egg
20. Flavour extract
21. Flour
22. Food coloring
23. Frosting
24. Fruit
25. Green vegetable
26. Herb

27. Ice cream
28. Jam
29. Liquid smoke
30. Meat
31. Milk
32. Mint
33. Mushroom
34. Noodle
35. Nut
36. Nut butter
37. Olive
38. Onion
39. Pasta
40. Pepper
41. Pickle
42. Plant-based oil
43. Protein additive
44. Pudding
45. Pumpkin
46. Rice
47. Root vegetable
48. Sauce
49. Seafood
50. Seaweed
51. Seed
52. Soup
53. Spice
54. Stem vegetable
55. Sugar
56. Tofu
57. Tomato
58. Tortilla
59. Vegetable spread
60. Vinegar
61. Yeast
62. Yogurt

Appendix B: Supermarket ingredient categories

1. Dairy
2. Fruits
3. Grains & products
4. Meals
5. Meat
6. Oils & sauces
7. Pastry & sweeteners
8. Spices
9. Stimulants
10. Vegetables