Technology for Translating, Visualizing, and Generating Recipes

Multimodal Machine Translation of Chinese into English

Chan Sin-wai

First published 2026

ISBN: 9781032308425 (hbk)
ISBN: 9781032308432 (pbk)
ISBN: 9781003306948 (ebk)

9 *GenRecipe* for Generating Recipes from Videos through Deep Learning

DOI: 10.4324/9781003306948-12

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

9 *GenRecipe* for Generating Recipes from Videos through Deep Learning

Introduction

In recent years, with text, pictures, and increasingly videos, people can explore recipes in all sorts of media such as cookbooks, TV programmes, YouTube videos, etc. Popular recipe websites such as Allrecipes.com and Food.com collect large numbers of recipes containing text, pictures, drawings, photos, and videos. Visitors can browse through various categories (e.g., breakfast, desserts, chicken, Mexican cuisine, etc.) or, with the help of search engines, filter recipes by keywords.

Recent advance in artificial intelligence (AI) research and specifically the deep learning approaches to computer vision and natural language understanding have opened up new intelligent ways of sifting through recipes as well as deriving new recipes from existing ones. This is achieved by computationally training artificial neural networks (ANNs) to learn how recipes relate to one another based on their multimodal contents (text, images, videos, etc.). Typical examples that use these networks include:

- Finding recipes based on a food image
- Finding/Generating food images based on a modified recipe, and
- Finding/Generating a recipe based on a cooking video.

This chapter reviews some recent developments in deep learning for language and visual contents and outlines some major areas of use cases in the recipe domain. These areas include recipe recommendation, food image recognition, cross-modal recipe retrieval, recipe generation, and food image generation. This chapter also explores how deep learning is employed in the generation of recipes from videos with the creation of the *GenRecipe* system.

Recipe Recommendation

Sorting a large collection of recipes based on their characteristics (e.g., specific ingredients) facilitates useful functions such as finding substitute ingredients, recommending similar recipes, etc. A recipe recommender system may suggest

DOI: 10.4324/9781003306948-12

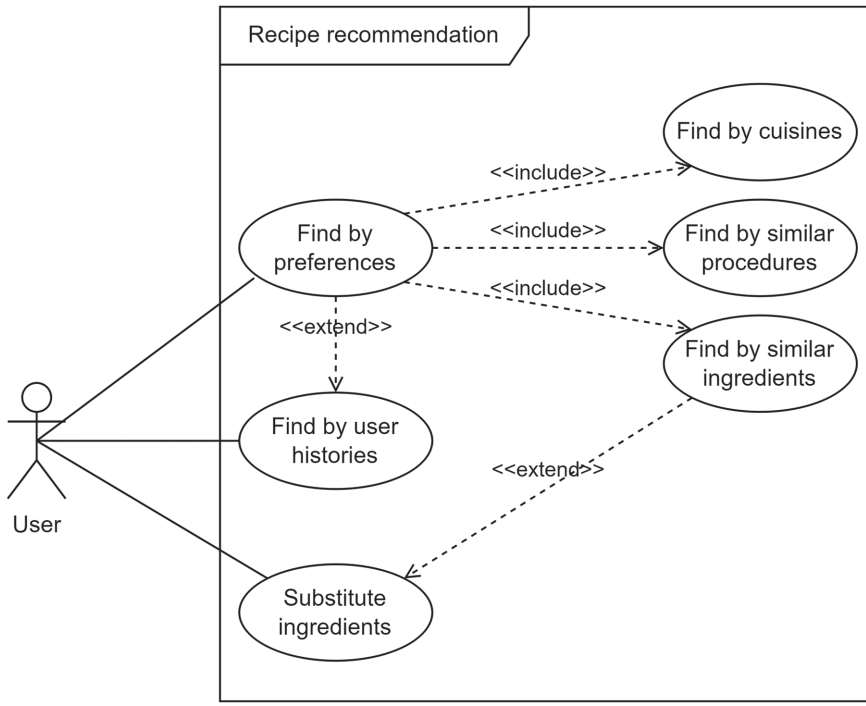

Figure 9.1 Use cases in recipe recommendation.

similar recipes to a user based on those recipes which he/she has enjoyed before. Figure 9.1 shows some use cases of such a system.

Researchers have studied ways to analyse and relate recipes for these instances using computational methods. The following subsections provide a brief review of some of these methods.

Ingredients

A major way of classifying recipes is based on the idea of a *cuisine*. Cuisines are often labelled by geographic regions or cultures. For example, on Allrecipes.com under "World Cuisine" there are Chinese, German, Indian, etc. Cuisines reflect the history, culture, geography, and food preferences of people living in different parts of the world. Similarities and differences between cuisines can often be analysed in terms of their common food ingredients. For instance, how cinnamon is used in Indian, Moroccan, and Turkish cuisines helps understand their distinct flavours. How rice is used in making fried rice, sushi, risotto dishes reflects the cooking styles in different countries. Studying different cuisines and their ingredients can inspire new combinations and techniques in one's cooking.

Su et al. (2014) studied the correlation between cuisines and ingredients based on a dataset of around 6,000 recipes downloaded from popular recipe-sharing websites. Their results revealed some limitations of the ingredient-based approach in terms of accuracy as recipes of different cuisines often share similar ingredients. For instance, they observed similarities (in terms of ingredients used) between cuisines attributable to geographic proximities such as between Chinese and Japanese as well as between Spanish and Italian.

Given a list of ingredients in a cooking recipe, what do we know about the cooked dish? Su et al. (2014) studied the cuisine classification (e.g. Chinese, Italian, etc.) of recipes and applied *associative classification* and *support vector machine* (SVM) (Hearst et al. 1998) in classifying a recipe based on the set of ingredients used. SVM is a supervised learning model for classification and regression analysis. It transforms and separates data points into different classes in higher dimensions through a hyperplane constructed optimally. Figure 9.2 illustrates the SVM method.

In their recipe recommendation method, Ueda, Takahata, and Nakajima (2011) modelled users' food preferences based on ingredients of their favourite recipes. Their method considers an ingredient that a user consumes repeatedly in his/her meals as a favourite of his/her. Recipes of these meals are broken down into ingredients for frequency calculation. Apart from favourite

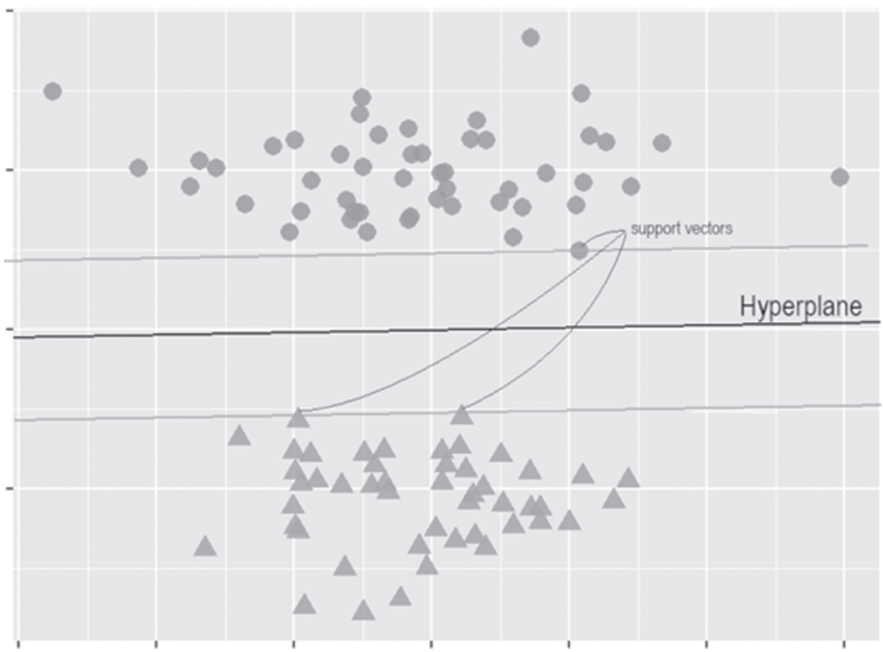

Figure 9.2 Illustration of the support vector machine method.

ingredients, the method also accounts for a user's disliked ingredients—those ingredients that a user never uses despite appearing in the user's recipe browsing history.

Teng, Lin, and Adamic (2012) considered relationships among recipes in terms of ingredient co-occurrence and substitutability in their recipe recommendation method. While ingredient co-occurrence is statistically derived from recipes, ingredient substitutability, and adjustment (in quantity) are derived from suggested modifications in users' recipe reviews.

Cooking Procedures

While characterising recipes based on ingredients is a useful approach to classifying them in many applications such as food recommendation, many recipes that share similar ingredients differ significantly in their cooking methods and procedures. L. Wang et al. (2008) modelled a recipe as a *cooking graph* in which ingredients "flow" across actions. They devised a similarity measure based on such graphs which can be applied in recipe search and recommendation. Mori, Maeta, Yamakata, et al. (2014) represented a recipe's procedural text as a directed acyclic graph. Their method involved training a word segmenter and a named entity tagger with an annotated corpus. Mori, Maeta, Sasada, et al. (2014) developed a method for generating procedural text from flow graphs and applied it to recipes.

Chang et al. (2018) used an ordered tree structure in modelling a recipe and measured procedural similarity in term of tree edit distance (Tai 1979). Figure 9.3 illustrates the idea of a tree representation of a recipe. Chang et al. (2018) devised a computational pipeline that scrapes online recipes and translates them into the tree representation. Their work was aimed at helping users sort out and choose among the different recipes for a given dish.

Yamakata et al. (2016) developed a method for representing the workflow described in a recipe as a tree diagram. Their method depends on the combined application of word segmentation, recipe term identification, and edge weight estimation. Pan et al. (2020) translated recipes into the tree representation manually. Their aim, however, is to capture sequencing information implicit in text with the help of visual information and vice versa.

Food Image Recognition

Nowadays, we can find numerous food images on the Internet. With social media, food is a popular category on picture-sharing platforms such as Instagram (Hu, Manikonda, and Kambhampati 2014) and Twitter (Yanai et al. 2019). Sorting these images which are often minimally captioned and retrieving their recipes from recipe collections is a difficult task.

Image recognition has applications in the food domain such as food logging (Kitamura, Yamasaki, and Aizawa 2008; Sahoo et al. 2019), dietary assessment (C. Liu et al. 2016; Van Asbroeck and Matthys 2020), food safety (Khan

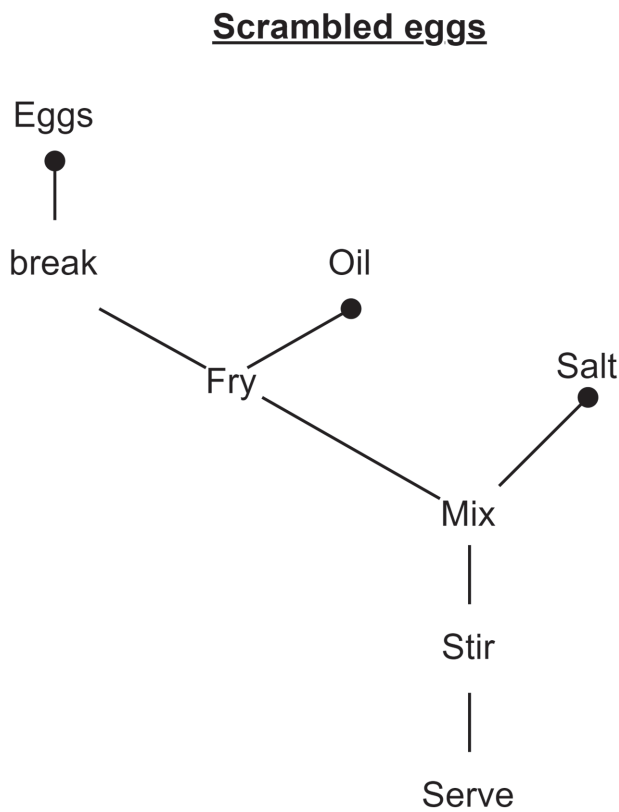

Figure 9.3 A tree representation of a recipe.

et al. 2021), etc. (See Figure 9.4). Automatic and accurate recognition of food ingredients in food images can be a convenient way of tracking our diets for health-related purposes (N. Chen et al. 2010).

S. Yang et al. (2010) tackled the problem of food image recognition by modelling spatial relationships among ingredients in a food image. Their method involves feature engineering and the use of Support Vector Machines (SVMs) (see Figure 9.2) (Hearst et al. 1998). M.-Y. Chen et al. (2012) developed a model that automatically identifies the food dish in a picture and provides nutritional information such as calories and ingredients. Their model is also based on a SVM classifier (Hearst et al. 1998).

Much progress has been made in image recognition due to advance in deep learning. While SVMs perform well with small datasets in image recognition, neural networks have outperformed SVMs in terms of accuracy when using large datasets. Neural networks are also able to manage more complex features. The development of deep learning techniques such as convolutional neural

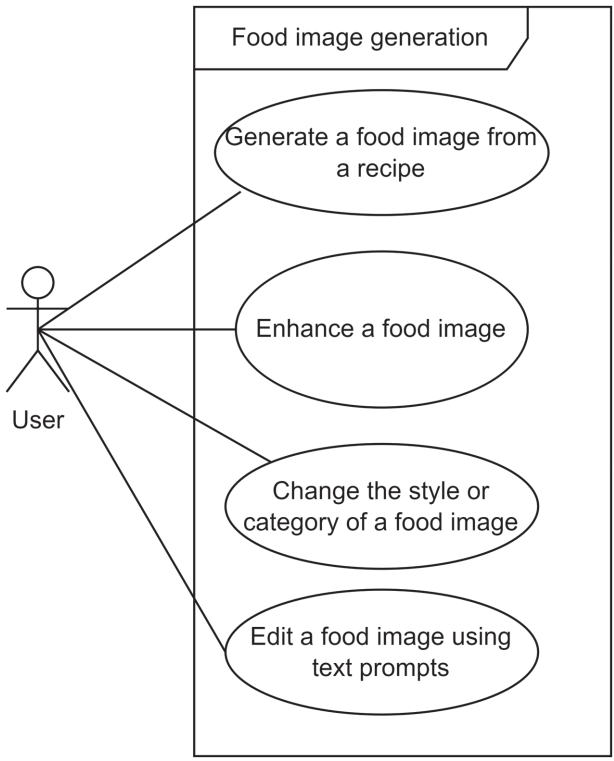

Figure 9.4 Use cases in food image recognition.

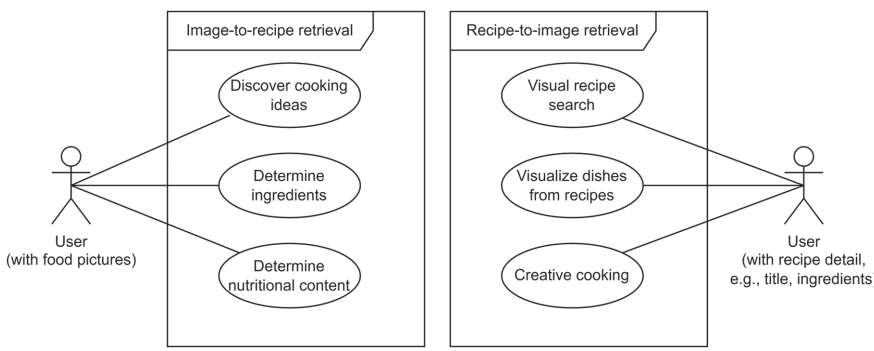

Figure 9.5 Use cases of cross-modal recipe retrieval.

networks (CNNs) has vastly increased the ability of AI in vision tasks including classification, segmentation, captioning, etc. These techniques can be readily applied in identifying food dishes and their ingredients from pictures.

CNNs can recognise complex patterns in images by employing multiple layers of artificial neurons to process image data by means of *convolution*, *pooling*, and *activation*. Convolution involves producing a feature map for a small image region such as an edge or a shape. Pooling compresses a feature map with a function such as average or maximum. Activation endows a neuron with non-linearity with a function such as a rectified linear unit (ReLU) function. Through multiple layers of convolution, pooling, and activation, CNNs are trained to recognise images based on complex features that can be learnt from large image datasets using back propagation and gradient descent methods.

Kagaya, Aizawa, and Ogawa (2014) trained a convolutional neural network (CNN) for food image recognition which attained higher accuracy than a number of SVM-based methods in a benchmark test based on a fast-food image dataset (M. Chen et al. 2009). They observed that colour features are dominant in food image recognition using their trained CNN and this echoes previous findings on handcrafted colour features for food recognition (e.g., Bosch et al. (2011)).

Jingjing Chen and Ngo (2016) applied a CNN to recognise ingredients in food images and obtain ingredient labels for text-based retrieval of recipes. While their deep learning use case involves learning the appearance of ingredients in different dishes over many example images, it does not take advantage of learning (jointly) from the corresponding recipes about the presence of individual ingredients. This issue is addressed in use cases of cross-modal deep learning, which are discussed in the next section.

Cross-modal Recipe Retrieval

Cross-modal retrieval allows users to issue a query using one type of media such as image and get results in another type such as text. Deep learning supports cross-modal retrieval that utilizes food images to retrieve recipes and vice-versa. Figure 9.6 shows some use cases of cross-modal recipe retrieval. This section reviews some deep learning approaches to cross-modal retrieval for recipes.

Visual-semantic Embedding Models for Recipes

The deep learning approach to cross-modal visual-language retrieval is underpinned by training a visual-semantic embedding model to represent visual objects based on image data and labelling text (Frome et al. 2013). Exploiting the correspondence between food images and recipes in deep learning models can lead to improved retrieval but also requires a large number of recipe-image pairs as training examples. Salvador et al. (2017) introduced a large-scale

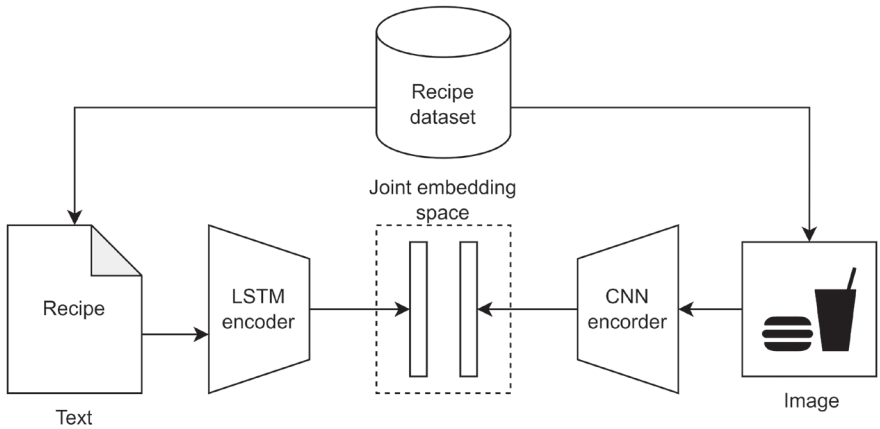

Figure 9.6 Modelling cross-modal recipe data in a joint embedding space.

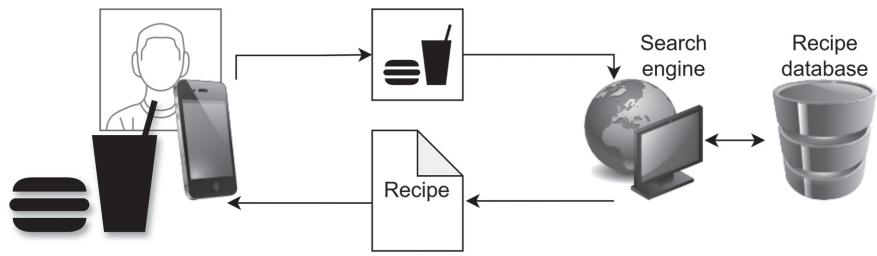

Figure 9.7 A search engine solution based on image-to-recipe retrieval.

dataset with over one million cooking recipes and 800 thousand food images. Using this dataset, they trained a deep learning model that maps pairs of recipe and image into a joint embedding space through which cross-modal (image-to-recipe or recipe-to-image) retrieval can be conducted. Figure 9.7 illustrates this approach.

Min et al. (2016) incorporated additional recipe attributes such as cuisine and course in the modelling framework for recipes. Their modelling approach can be applied in (1) cuisine classification, (2) attribute-augmented recipe image retrieval, as well as (3) ingredient and attribute inference from food images.

Other researchers have extended the cross-modal modelling approach with attention mechanism for improved performance. J.-J. Chen et al. (2018) encoded the title, ingredients, and instructions separately and leveraged

attention mechanism in addressing the cause-and-effect relationships between ingredients and actions in recipes. A joint representation was learnt together with food image data for use in cross-modal retrieval. H. Fu et al. (2020) applied attention-based learning to food images and recipe text using CNN and RNN, respectively, and then aligned the learnt representations for cross-modal consistency and retrieval.

B. Zhu, Ngo, and Chan (2021) considered the effectiveness of exploiting "noisy" recipe data from the web for cross-modal learning with image data. They found that although noisy recipe data alone do not lend themselves to more effective food recognition, their pairing with image data does improve retrieval performance. Z. Xie et al. (2021) trained a three-tier joint embedding model for cross-model retrieval between image and text involving recipe text, instruction text and food images. The model has been applied to both image-to-recipe and recipe-to-image retrieval.

A common way to find recipes is based on the ingredients to be used (Teng, Lin, and Adamic 2012). Salvador et al. (2021) introduced a hierarchical recipe transformer for encoding recipe titles, ingredients, and instructions separately and trained a joint embedding space with image data. Their experimental results with this model set new benchmarks in cross-modal recipe retrieval.

Image-to-Recipe Retrieval

In the age of social media, users may find recipe ideas by browsing through online photos of food dishes and use an AI system to retrieve recipe text for discovery and inspiration of cooking ideas. In an educational or research application, cross-modal retrieval can support exploration and research about food cultures and cooking techniques through visual examples. Many smartphone users have found it convenient to keep a picture log of their food intake for health-related purposes (Kitamura, Yamasaki, and Aizawa 2009; Kawano and Yanai 2014). Determining the nutritional content of food items in such a picture log is, however, a challenging task which requires knowing the food recipes. A potential solution is to use a search engine based on image-to-recipe retrieval for querying recipes by food pictures. Figure 9.7 illustrates this solution.

Jing-jing Chen, Ngo, and Chua (2017) applied deep learning in food recognition and recipe retrieval. Their model extracts information about ingredients, cutting, and cooking methods from a food image and uses such information in retrieving relevant recipes. This model covers over a thousand ingredients and a few dozen cooking, and several cutting, methods. In tackling the problem of predicting the relative amount of each ingredient in a dish as presented in an image, J. Li et al. (2021) proposed a learning architecture which combines food image and recipe data including titles, ingredients and instructions (cooking actions).

Recipe-to-image Retrieval

Cross-model retrieval can support text-to-image retrieval of food images—given the title and/or some ingredients of a recipe, retrieve relevant pictures of the food item or its constituent ingredients. Text-to-image retrieval is a relevant task in food applications such as recipe search (H. Xie, Yu, and Li 2010), food recommendation (W. Wang et al. 2021), recipe visualization (B. Zhu et al. 2019), creative cooking (Varshney et al. 2019), etc.

Researchers in recipe-to-image retrieval have so far focused on the end-product of following a cooking recipe (e.g. Z. Xie et al. 2021). However, the requirements of retrieval may cover images of ingredients not only in their individual original form, but also in their work-in-progress form. Y. Zhang, Yamakata, and Tajima (2019) tackled the problem of retrieving images of ingredients during the different stages of cooking. They took a stage-wise curriculum learning approach and obtained better performance than a baseline approach which did not account for stages.

Recipe Generation

Recipe retrieval aims to find an existing recipe based on a user's query. It tries to find the most relevant recipe in collections based on attributes such as title, ingredients, cuisine, etc., or a food picture. Unlike recipe retrieval, *recipe generation* involves generating recipes based on a model of what a recipe constitutes and how different constituents are composed together into realistic coherent recipes. In other words, a recipe generation model harnesses recipe knowledge and composition to produce novel but realistic recipes.

Recipe generation can contribute to creative cooking (Varshney et al. 2019). Generating variants of the same recipe with substitute ingredients can be a useful means of satisfying people with particular dietary requirements (Fatemi et al. 2023). AI methods can also generate personalized recipes tailored to someone's historical preferences on ingredients or cooking methods (Majumder et al. 2019). Food enthusiasts may use AI to create new recipes based on examples or cooking styles as shown in food images (e.g. Salvador et al. 2019; H. Wang et al. 2019). Figure 9.8 above summarizes the applications of recipe generation.

Researchers have studied various deep learning approaches to modelling and generating cooking recipes. The following subsections review some of these approaches. Figure 9.9 shows some use cases of recipe generation.

Recipe Generation with Recurrent Neural Networks

Recurrent neural network (RNN) models have been applied to natural language generation (NLG) tasks (Gatt and Krahmer 2018) including recipe text generation (Kiddon, Zettlemoyer, and Choi 2016; Majumder et al. 2019; H. Lee et al. 2020). Given a piece of input text, an RNN can generate

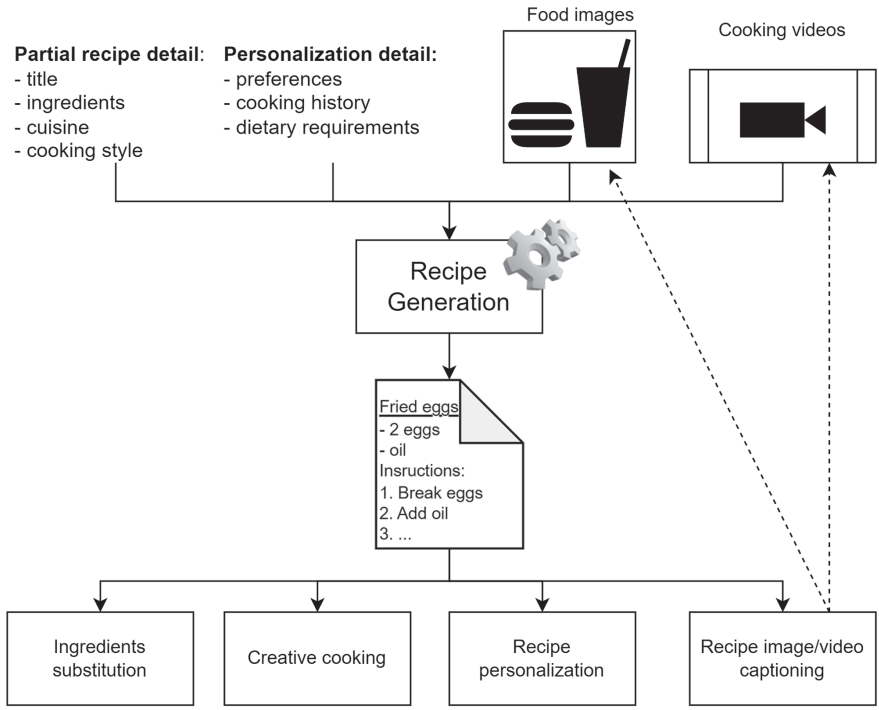

Figure 9.8 Some applications of recipe generation.

further text "word-by-word" to extend the input in a linguistically coherent manner. It does so based on a neural network model trained with a large dataset of human-generated text. Figure 9.10 illustrates the working of RNN text generation.

Training a RNN for text generation requires a large amount of text examples. Issues of using RNNs in text generation include vanishing gradients (Hochreiter and Schmidhuber 1997; Hochreiter 1998), softmax bottleneck (Z. Yang, Dai, et al. 2017), and lack of long-term memory (Hochreiter and Schmidhuber 1997). The last issue makes it difficult for RNNs to generate coherent text of extended length.

Researchers have applied RNNs in the task of recipe generation. For instance, Kiddon, Zettlemoyer, and Choi (2016) tackled the coherence problem of RNNs in natural language understanding with a *checklist mechanism* in a recipe text generation task. Given a recipe's title and the required ingredients, the mechanism tracks to ensure the inclusion of all ingredients in the generated text.

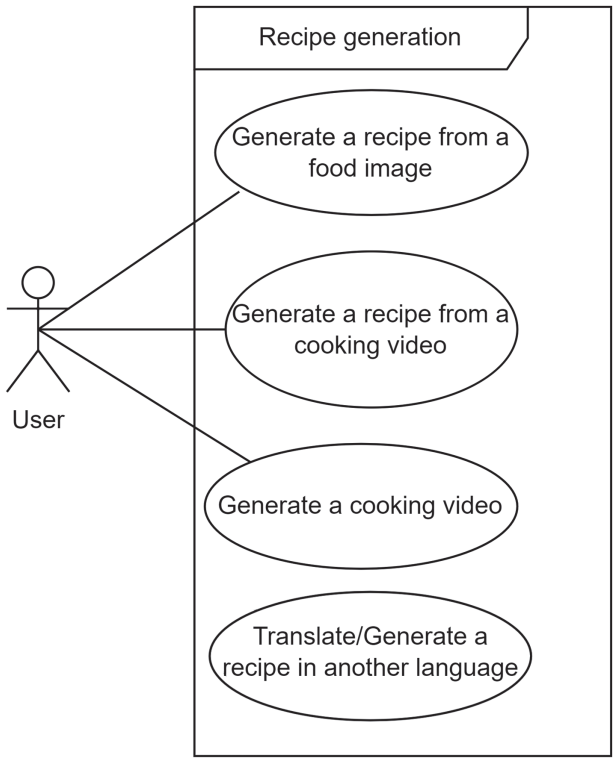

Figure 9.9 Use cases of recipe generation.

Hochreiter and Schmidhuber (1997) developed an improvement for RNNs known as Long Short-Term Memory (LSTM) that address the vanishing gradient problem. Z. Yu, Zang, and Wan (2020) applied LSTM in a recipe generation model that accounts for user preferences such as "low sugar" or "low fat".

K. Cho, Merrienboer, et al. (2014) introduced yet another improvement for RNNs known as Gated Recurrent Units (GRUs) that uses fewer parameters than LSTM. Majumder et al. (2019) applied GRUs in their recipe model which attends to a user's historical preferences in generating recipe text with incomplete ingredient details.

Recipe Generation with Large Language Models

Recurrent neural networks can learn to map sequences of input text to desired output sequences such as some translations of the input text. This has been

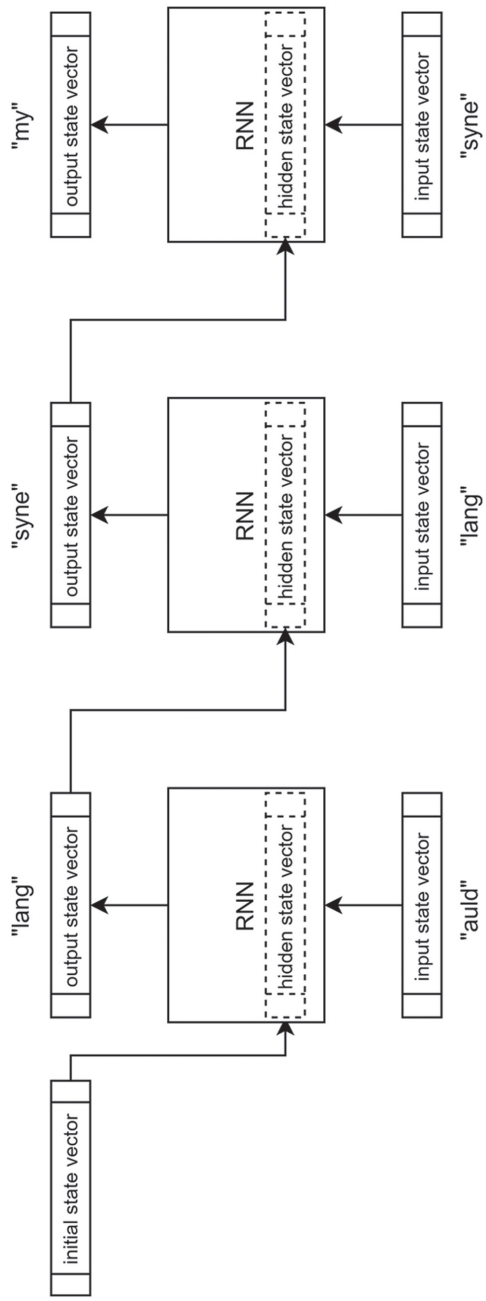

Figure 9.10 A Recurrent Neural Network (RNN) for text generation.

applied to cooking recipe text with encouraging results. However, the performance of RNNs has been superseded by a more sophisticated type of neural network known as transformer.

Transformers are feed-forward neural networks (no loops) that focus on relevant parts of an input sequence using a so-called attention mechanism. These networks support processing in parallel and are easier to train than RNNs for long sequences. When applied to language tasks such as machine translation, question answering, and natural language understanding, transformers have achieved state-of-the-art results.

Large language models (LLMs) are large-scale neural networks based on the transformers architecture with huge numbers of parameters and are typically pre-trained with large datasets based on unsupervised learning objectives. LLMs can be fine-tuned with cooking recipe datasets and used for a number of tasks. For instance, they can be used for generating plausible recipes based on narrative inputs from users.

Large language models have come of age with the arrival of models such as GPT-3 (Brown et al. 2020) and BERT (Devlin et al. 2018). These models have attained nuanced language understanding, which is necessary for conducting smooth conversations with humans, summarizing long passages, answering non-trivial questions, etc.

LLMs capitalize on unsupervised deep learning based on a vast amount of text. These models have surpassed RNNs in performance across multiple tasks. The transformer mechanism, on which LLMs are based, can scale up training much higher than RNNs can normally do. An important advantage of the transformer mechanism is that it supports parallel computations in model training and inference whereas RNNs are constrained to sequential computations. This makes LLMs vastly more efficient than RNNs to train and apply. The lack of long-term memory problem in RNNs is also mitigated by the transformer mechanism which emphasizes *attention* in the learning process and captures relevant dependencies across different parts of a text to achieve more nuanced language understanding.

LLMs are pretrained language models that are amenable to fine-tuning for target applications. Pretraining a large language model involves using large datasets of text derived from the web and various other sources of knowledge such as books. This exposes the model to general knowledge and linguistic skills such as grammar and reasoning. A pretrained model covers a great deal of things across numerous domains without focusing on particular skills domains. It can be rendered more effective at certain tasks such as answering questions or summarizing text with further training based on y small task-specific text datasets. Researchers have found this an effective approach to realizing the benefits of expensive large-scale pretraining in a wide range of specific language-based tasks. Figure 9.11 illustrates the use of a pretraining and finetuning in recipe generation.

Large language models (LLMs) such as GPT-3 have been applied to cooking recipe generation (Metz and Krishna 2022). LLMs excel in generating long coherent passages based on relevant domain knowledge in good

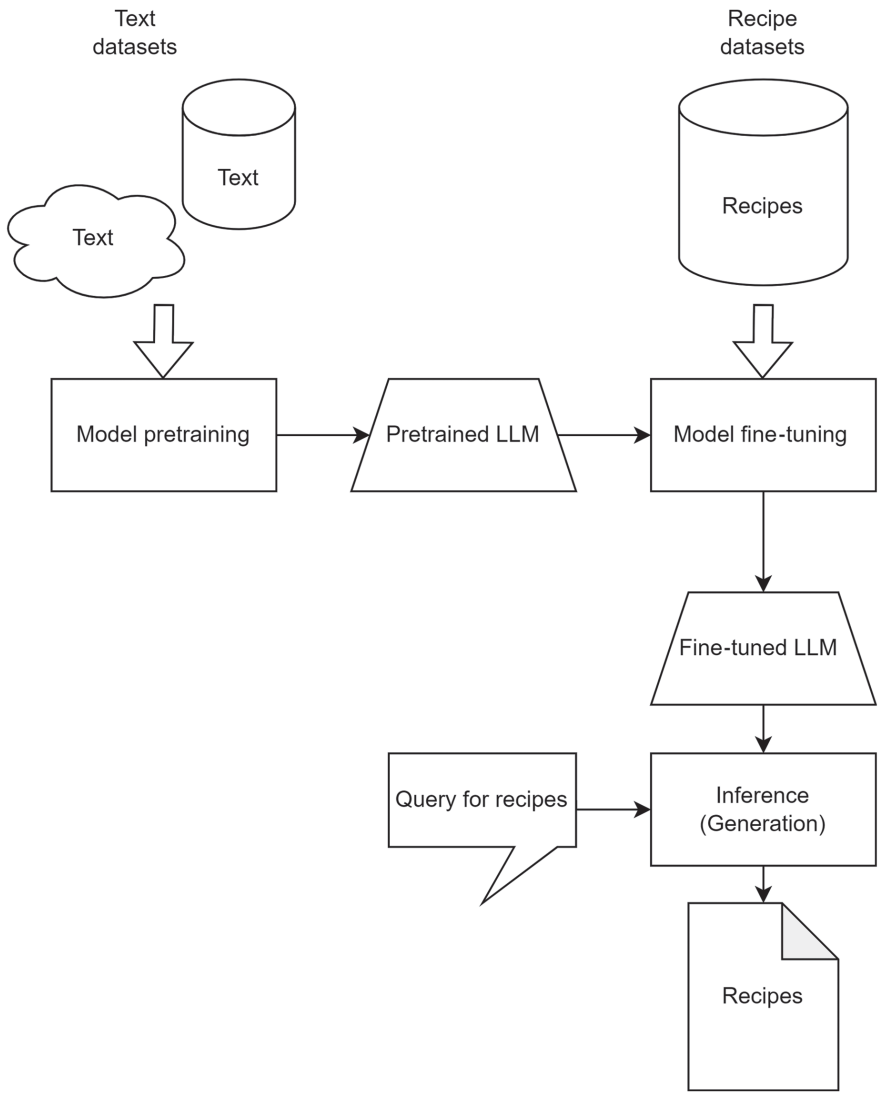

Figure 9.11 Pretraining and fine-tuning a LLM for recipe generation.

writing styles. With proper training (Christiano et al. 2017), they can also address user preferences and personalize results based on users inputs.

H. Lee et al. (2020) fine-tuned OpenAI's GPT-2 large-scale language model (Radford et al. 2019) based on the Recipe1M dataset (Marin et al. 2019) for recipe text generation, and built a web application for evaluation. Bień et al. (2020) employed a Named Entity Recogniser (NER) to extract food entities from the Recipe1M dataset and used them for fine-tuning GPT-2 for recipe generation.

Goel et al. (2022) compared a LSTM-based model trained on the RecipeDB dataset (Batra et al. 2020) and a finetuned GPT-2 model fine-tuned on the same dataset for use in a recipe generation task. The two models were evaluated for the task based on BLEU scores (Papineni et al. 2002) and the results show that the fine-tuned GPT-2 model achieves much better performance in the task.

X. Liu et al. (2022) studied the use of LLMs in generating realistic recipes. They designed a task for these models to modify the actions in a given recipe in response to an ingredient change. They also finetuned GPT-2 (Radford et al. 2019) with over a million recipes of common dishes for the task. Their experiments with the finetuned models reveal gaps between these models' and humans' understanding of recipe text in terms of things such as cooking styles and order of actions.

With large language model (LLM) based AI chatbots such as ChatGPT (OpenAI 2022), users can ask AI to select ingredients based on seasonal or geographical conditions or suggest cooking methods based on the availability of equipment.

Image-to-text Recipe Generation

Salvador et al. (2019) treated the problem of recipe generation from food images as sequence generation conditioned on a food image and its (predicted) ingredients. With a model trained on around a quarter of a million recipes with images, their system was able to outperform previous image-to-recipe retrieval approaches and even surpass human efforts.

H. Wang et al. (2019) devised ACME, or Adversarial Cross-Modal Embedding, an end-to-end framework for cross-modal retrieval and two-way translation between food images and cooking recipes. They employed an adversarial learning strategy in aligning the text and image modalities. Pan et al. (2020) built a dataset of annotated recipe workflows and trained a model with encodings of cooking step data (text and images) across workflows. Their model showed better performance in predicting relationships between cooking steps than models based on a single modality (text or image).

Generating procedural text from *a sequence of* photos is a challenge for AI research (Chandu, Nyberg, and Black 2019). This requires training AI models to understand the visual content in each photo as well as the progression of actions/events across photos. Nishimura, Hashimoto, and Mori (2019) studied the generation of recipe text from a photo sequence of cooking. This involved pretraining a joint embedding model for cooking photos and instructions using the Cookpad Image Dataset (Harashima, Someya, and Kikuta 2017).

Video-to-text Recipe Generation

Cooking videos are popular on social media and many food-related websites such as Allrecipes, help people learn cooking techniques and give useful tips

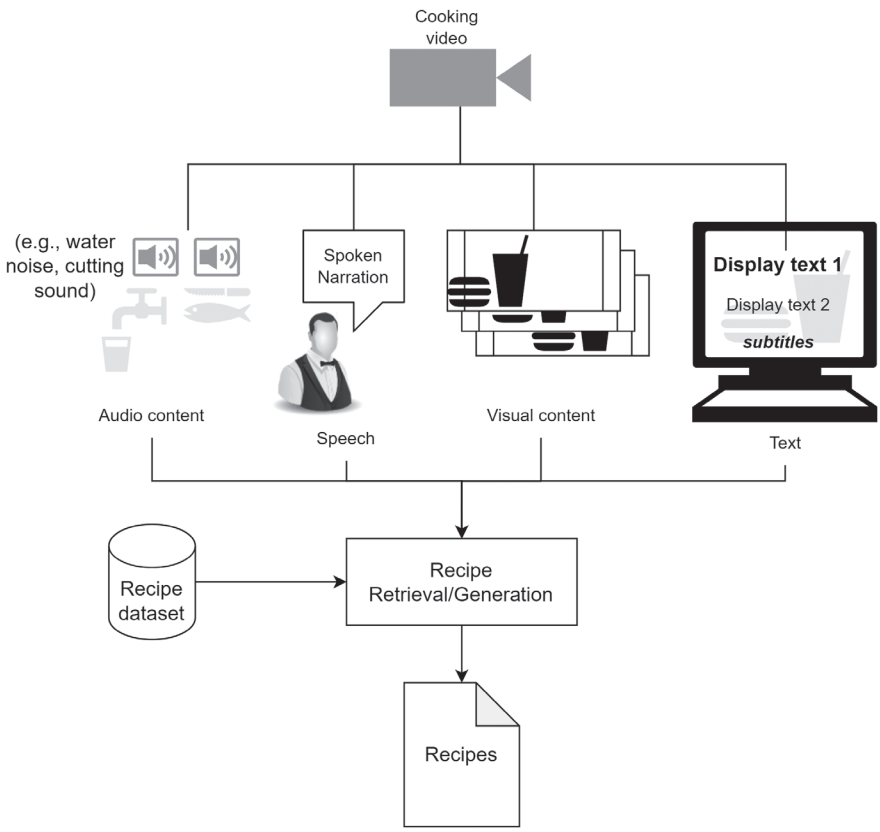

Figure 9.12 Retrieving/generating recipes based on cooking videos.

for making good and delicious foods. Browsing through these videos can provide inspiration and ideas for cooking.

Typically, a cooking video "streams" all information about ingredients, procedures, cooking methods, and other information for preparing a dish audio-visually. The visual content is often accompanied by both speech (e.g., someone narrating) and text (e.g., on-screen text display, subtitles). A straightforward transcription of spoken words (via Automatic Speech Recognition) and display text (via Optical Character Recognition) in a cooking video would suffice as a transcript to accompany the video but hardly serve as a recipe on its own. The challenge of video-to-recipe generation is to distil all the information available across modalities (audio, visual, spoken vs. written, verbal vs. non-verbal, etc.) into a procedural text and ingredient information. Figure 9.12 illustrates the multimodal task of video-to-recipe retrieval and generation.

Fujii et al. (2019) tackled the problem of generating recipe text from cooking videos on YouTube. Treating the solution as an *image-captioning* task, they

addressed the issue of maintaining consistency across recipe sentences with sequence learning. Besides, they pre-processed videos by removing irrelevant and blurry scenes with the help of object detection.

Methods to improve the quality of video to text generation have been studied by researchers. Guo et al. (2016) addressed the consistency between video content and descriptive sentences in a video captioning task with a Long Short-Term Memory (LSTM) model for attention to salient visual features in conjunction with a combined 2-D and 3-D CNN representation of video content. G. Li, Ma, and Han (2015) employed an CNN model together with a RNN model in generating descriptive captions from video clips, and applied ranking to generated candidate sentences for improved quality.

Extracting procedure knowledge from cooking videos for recipe text generation is demanding (F. Xu et al. 2020). Methods to improve the quality of video captions have previously been studied by other researchers.

Ghoddoosian, Sayed, and Athitsos (2022) tackled the problem of recognising tasks (e.g., making coffee) from instructional videos where certain actions and objects are labelled. They developed a model that learns the hierarchical relationships between tasks and actions, and in addition, the temporal segmentation of video in terms of actions. Ji et al. (2022) created a large instructional video dataset with autogenerated temporal video segmentation. The generation was enabled by a transformer-based model architecture that learnt from a fusion of the video and its transcribed text.

Seo et al. (2022) tackled the task of multimodal video captioning (MVC) by training a model to predict utterances as captions in a bi-directional (forward and backward) manner. When given an utterance in the video and the corresponding video segment, predicting the following utterance and, when given the latter, predict the former utterance. The bi-directional training ensures proper alignment of video content to generate captions. The model itself consists of a multimodal encoder coupled with a sentence decoder, both transformer-based and supported by masked language modelling (Devlin et al. 2018). It achieved state-of-the-art video captioning resulting on the YouCook2 dataset (L. Zhou, Xu, and Corso 2018) which contains 2,000 cooking videos from YouTube.

Cooking Video Moments

Beside a picture of the fully prepared dish, steps in a recipe can be accompanied by pictures/video clips which help visualize the instructions. Typically, a step-level picture/video clip would show the ingredients and actions (or effects thereof) involved in that step. However, the granularity of steps and the number of accompanying images/video clips for each step can be arbitrary. Doman et al. (2011) studied the synthesis of multimedia cooking recipes from recipe text and a database of manually-tagged video clips of cooking operations.

Sun et al. (2019) developed VideoBERT based on available speech recognition, video quantization, and language learning methods for learning a joint model for video and language data. An application of VideoBERT on cooking videos is to take sentences from a recipe and retrieve relevant video segments from the dataset. In order to take better advantage of the multimodal content of instructional videos, Gabeur et al. (2022) took the idea of masked modelling from BERT for language modelling to video modelling with masking applied on the basis of entire modalities and achieved gain in video retrieval performance with YouCook2 (L. Zhou, Xu, and Corso 2018) and other datasets. H. Zhang et al. (2022) gave a survey of recent works in temporal sentence grounding in videos (TSGV) which refers to the retrieval of moments of video that correspond to the semantic content in a language query. They also mentioned the potential of exploiting queries in different modalities (video, audio, speech).

Malmaud et al. (2015) developed a Hidden Markov Model (HMM) based method for aligning a sequence of recipe instructions to the auto-generated speech transcript of a corresponding cooking video. The alignment is refined by a visual recogniser trained for food items. A useful application of their method is to associate each recipe step with a video segment as a visual illustration of the step.

Recipe Translation and Multilingual Recipe Generation

The translation of cooking recipes into different languages helps make cuisines more accessible to people in different language communities around the world. It can also inspire creativity in cooking and promote sharing of cultures. A well-established application of machine learning in natural language processing is neural machine translation (NMT).

Researchers have studied the application of NMT to cooking recipes and a lot of emphasis is often put on the accuracy of translation (Hasyim et al. 2021). In contrast to phrase-based statistical machine translation (PBSMT) (Koehn, Och, and Marcu 2003), neural machine translation (NMT) based on sequence-to-sequence modelling has achieved state-of-the-art performance (Sutskever, Vinyals, and Le 2014; K. Cho, Van Merrienboer, et al. 2014; Bentivogli et al. 2016). Neural networks are often touted as more capable of handling contextual and domain-specific semantics, though significant challenges remain for NMT in some areas (Koehn and Knowles 2017; Bentivogli et al. 2016; Hasyim et al. 2021). For instance, Hasyim et al. (2021) studied the use of Google Translate in translating French-Indonesian recipe text and found issues related to the cultural context of the source and target languages.

Sato, Harashima, and Komachi (2016) compared NMT with PBSMT in translating recipes from Japanese into English. They found that PBSMT tended to make many *word order* errors whereas NMT was more prone to *substitution* errors. Both methods, however, suffered from many omission

errors. Sato, Harashima, and Komachi (2016) hinted at the need to adapt the machine translation task by considering the ingredients and order of actions in recipes.

On the other hand, NMT can be adapted for translation preferences and nuances particular to circumstantial needs. and requirements such as personal preferences, cooking skills, dietary requirements, etc. Using AI methods to generate recipes and menus in multiple languages also has applications in cooking and the restaurant business (Nobumoto et al. 2017). Neural methods for translating cooking recipes can be extended to the translation of restaurant menus, cooking blogs, subtitles in cooking videos, and so on. These methods can also contribute to the discovery and intelligent search for cooking ideas and methods.

NMT can also be extended to *multimodal machine translation* (Jiatong Liu 2021; Sulubacak et al. 2020) for cooking recipes with the help of images and videos. Mohammadshahi, Lebret, and Aberer (2019) proposed an approach to learn multimodal, multilingual embeddings for images and captions from image-caption datasets with bilingual captioning. They demonstrated improvement in the text-image retrieval tasks with English-German and English-Japanese datasets. Huang et al. (2020) exploited neural machine translation (NMT) to enrich monolingual as well as multilingual textual representations in datasets for learning multimodal multilingual embeddings for images and captions. Their experiments showed some advantage in this approach. Fei, Yu, and Li (2021) considered the use of learnt multimodal multilingual embeddings for images and captions as a means of reducing the reliance on machine translation tools in the multimodal retrieval (MMR) task. They also showed improved retrieval performance with this approach in English-German and English-Japanese datasets.

Researchers have devised neural network models of recipes that explicitly take into account the ingredients and actions and even food images and applied them in machine translation as well as other applications such as image-to-recipe retrieval. Guerrero, Pham, and Pavlovic (2021) exploited both text and image data in learning a joint representation of recipes regularised with *imperfect multilingual translations* involving multiple alphabets. Through back-translation (Sennrich, Haddow, and Birch 2016), they regularised multilingual recipe text in languages including English, German, French, Russian, and Korean, involving three alphabets (Roman, Cyrillic, and Hangul).

Food Image Generation

The recent emergence of AI methods for generating highly realistic images has inspired use cases in the food domain. Restaurants can benefit from using high quality food images in their menus and advertising materials. AI can generate dish images in diverse styles to inspire new cooking ideas. Food photographers can benefit from using AI tools to enhance their food images. As far as recipe visualization is concerned, recipe authors can supplement text with generated

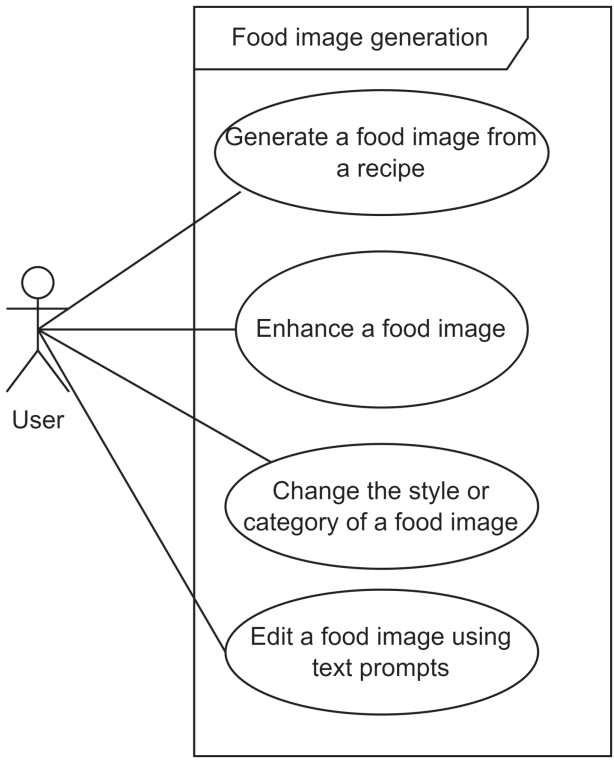

Figure 9.13 Use cases in food image generation.

images and illustrations more conveniently and/or derive variants from conventional photographs.

The following subsections review some major deep learning approaches to image generation and their applications in food image generation. Figure 9.13 shows the main use cases in food image generation.

Generative Adversarial Networks

Among the different AI approaches to food image generation, generative adversarial networks (GANs) (Goodfellow et al. 2014) feature most prominently. Since the inception of GANs (Goodfellow et al. 2014) in 2014, much progress has also been made in the generation of realistic synthesized images. A GAN consists of a pair of neural networks usually referred to as a generator and a discriminator. The former tries to generate realistic images while the latter strives to distinguish the generated (fake) images from real ones. The two networks are trained simultaneously to work against each other in such

a way toward generating highly realistic images. Figure 9.14 illustrates the training of a GAN.

The GAN approach was also extended to allow for more control over output content and support a variety of image-to-image translation tasks such as style transfer, object transfiguration, photo enhancement, etc. (Mirza and Osindero 2014; Isola et al. 2017; J.-Y. Zhu et al. 2017). Mirza and Osindero (2014) introduced additional input data (e.g., an input image) as conditions in the training of a GAN and hence the name, Conditional GAN, or CGAN. J.-Y. Zhu et al. (2017) extended the GAN approach to translating images from one domain to another without paired training examples. They called this approach CycleGAN. Both CGAN and CycleGAN have featured prominently in research on visualizing recipes (Horita et al. 2018; Papadopoulos et al. 2019; S. Wang et al. 2019; B. Zhu et al. 2019; Han et al. 2021; Z. Liu, Niu, and He 2023).

Tanno et al. (2018) demonstrated the use of a GAN for image transformation to convert a given food image from one category (e.g., ramen) to another category (e.g., curry rice) while preserving the outline shape of the food item and its plating. B. Zhu et al. (2019) used a GAN to generate synthesized

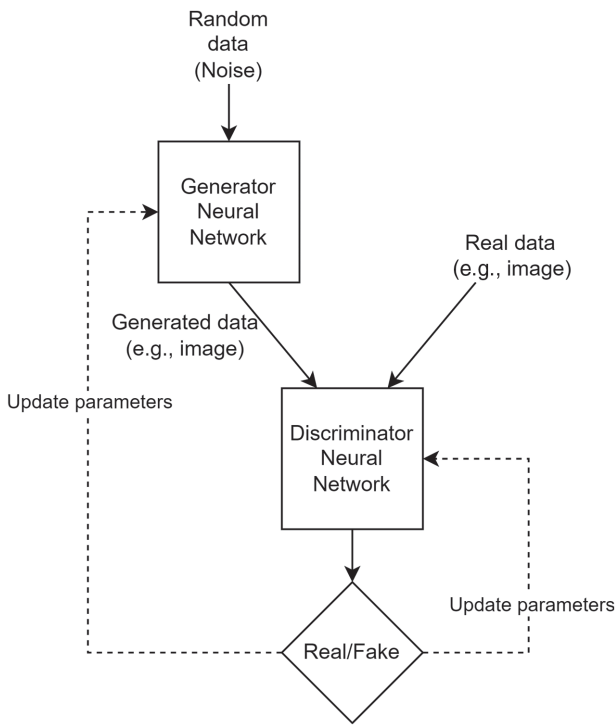

Figure 9.14 Training a Generative Adversarial Network (GAN).

thumbnail images from recipes as a means of assisting users in the browsing of recipes as well as providing cues to explain the ranking of recipes.

Papadopoulos et al. (2019) trained a GAN to generate image layers that correspond to changes to the visual appearance of a pizza through the addition and removal of layers of its toppings. The generated images can serve as a visual guide to the procedure of making a pizza. B. Zhu and Ngo (2020) trained a GAN to "simulate" cooking steps visually through food image transformations based on how cooking methods change the appearance of food ingredients. This could contribute to the ease of not only following a given recipe but also predicting the effect of modifying the recipe on the appearance of the dish.

B. Zhu and Ngo (2020) incorporated the modelling of cooking actions in CookGAN, a GAN for generating meal images from recipes. Their approach accounts for explicit interaction between ingredients and cooking actions and supports some interesting applications such as "virtual" recipe experimentation through on-the-fly modification of ingredients and instructions. Sugiyama and Yanai (2021) extended a GAN for food image generation with feature disentanglement so that serving and plate styles can be arbitrarily shaped in generated images.

Z. Liu, Niu, and He (2023) developed a GAN for generating food images based on recipe and ingredient labels. Notably, they performed experiments on generating Chinese food images based on a Chinese food dataset (Jingjing Chen and Ngo 2016) with encouraging results.

Large-Scale Text-to-Image Models

The advent of large-scale language-image pre-trained models, such as CLIP (Radford et al. 2021), facilitates the cross-referencing between language and images in downstream image tasks, and includes image generation and image editing. For food image editing, Yamamoto and Yanai (2022) applied VQGAN-CLIP (Rombach et al. 2022) to the task and experimented with dish image editing by prefixing and suffixing dish names with taste adjectives and toppings (e.g., hot-, sweet-, -with egg, -with bacon, etc.).

The *diffusion* approach to image generation has made significant advance in terms of photorealism recently (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song et al. 2020; Rombach et al. 2022). In this approach, deep learning models are trained with examples of adding noise probabilistically to an image over many steps and learn how to reverse this process to obtain photorealistic images. Furthermore, the generative process can be conditioned based on additional inputs in text or images or both, resulting in text-to-image generation (Ramesh et al. 2022; Nichol et al. 2022), inpainting (Sohl-Dickstein et al. 2015), and text-guided image editing (Kawar et al. 2022), respectively.

Hertz et al. (2022) extended diffusion-based text-to-image generation to text-driven image editing without the need of spatial masks. They developed

an intuitive prompt-to-prompt editing framework which accounts for the spatial layout of an image to each word in its prompt. This framework supports localized as well as global editing of an image via editing its prompt; it also allows the emphasizing of a certain word for its effect on the image. Mokady et al. (2022) found a way to apply this framework to real (non-generated) images via an inversion technique for text-to-image diffusion models. This allows real images to be edited in the same way as generated images via their prompts.

The performance of text-to-image generators in producing photorealistic images based on text descriptions is subject to the choice of keywords and phrases added to a given description. Finding the appropriate keywords and phrases often depends on human intuition and experience. Researchers have proposed guidelines and techniques for prompt engineering (e.g. V. Liu and Chilton 2022; Pavlichenko, Zhdanov, and Ustalov 2022; K. Zhou et al. 2022).

Galatolo, Cimino, and Cogotti (2022) curated a diverse dataset of 300 text and image pairs as a gold standard for evaluating some recent text-to-image models. The dataset was divided into three main categories, namely painting, drawing, and realistic (including food pictures), and involved six data sources including Wikiart, Deviantart, Openverse, ImageNet Sketch, Wikimedia Commons, and COCO. Generated images were compared to the gold standard in terms of CLIP score and whether they appeared more real (not generated) than the gold standard to human evaluators. Their results show that Stable Diffusion leads the pack in both CLIP scoring and human evaluation. Borji (2022) compared Stable Diffusion, Midjourney and DALL-E 2 for their ability to generate photorealistic faces. The Fréchet Inception Distance (FID) (Heusel et al. 2017) is used as a quality metric for the generated faces.

Petsiuk et al. (2022) compared the ability of Stable Diffusion and DALL-E 2 on three text-to-image tasks (counting, shapes, faces) at three difficulty levels with the help of 20 computer science AI graduate students who provided subjective ratings (1-5) on generated images from the two models. They further devised a text-to-image benchmark for evaluating text-to-image models in a suite of 32 tasks such as "generating objects with specified colours", "handling multi-lingual prompts", etc., that cover a wide range of downstream applications. J. Cho, Zala, and Bansal (2022) developed an evaluation toolkit for four visual reasoning skills, namely object recognition, object counting, colour recognition, and spatial relation understanding in text-to-image generators. The toolkit involved an object detector trained by simulated data for each of the four skills.

Demonstrating Recipe Generation from Videos

The idea of *GenRecipe* can be better grasped with a demonstration of the production of a recipe from a video on preparing a dish "Stir-fried sliced beef with rice noodles in brown sauce" (〈乾炒牛河〉), as shown below scene by scene. Ingredients:

1. beef 100g
2. rice noodles 750g

3. onion
4. leek sprouts
5. spring onion
6. bean sprouts

Scene 1

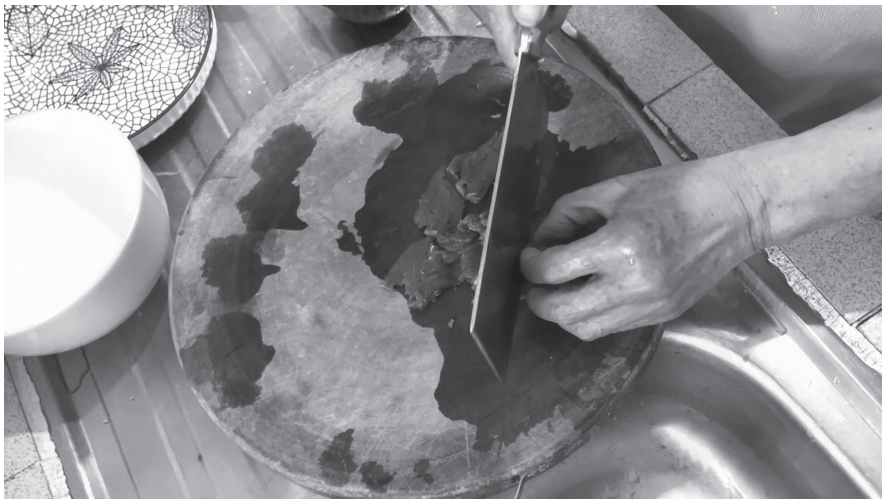

Figure 9.15 Slice the beef into pieces.

Scene 2

Figure 9.16 Marinate with salt, sugar, pepper powder and light soy sauce.

Scene 3

Figure 9.17 Add wine and soy sauce.

Scene 4

Figure 9.18 Add starch, stir, add oil.

Scene 5

Figure 9.19 Cut onion.

Scene 6

Figure 9.20 Cut leek sprout.

Scene 7

Figure 9.21 Prepare rice noodles.

Scene 8

Figure 9.22 Add oil to stir-fry the bean sprouts

Scene 9

Figure 9.23 Stir-fry the beef

Scene 10

Figure 9.24 Fry rice noodles with onion.

Scene 11

Figure 9.25 Add light soy sauce and dark soy sauce for colour.

Scene 12

Figure 9.26 Add all ingredients and keep stirring until all are cooked.

Scene 13

Figure 9.27 Sprinkle sesame seeds and serve.

Conclusion

This chapter has reviewed various deep learning approaches to the retrieval, visualization, and generation of cooking recipes and food images and how *GenRecipe* works in a video setting. Deep learning has been going from strength to strength in the food application domain. There are now many innovative use cases enabled by deep learning. Developments in AI are happening on an industrial scale at a very rapid pace at the time of writing. While technological advances are opening up new exciting use cases, there are significant issues facing deep learning and its applications including unfairness, bias, lack of explainability, large carbon footprints, etc. This calls for attention to the ethical and responsible development and deployment of this technology.

References

Batra, Devansh, Nirav Diwan, Utkarsh Upadhyay, Jushaan Singh Kalra, Tript Sharma, Aman Kumar Sharma, Dheeraj Khanna, Jaspreet Singh Marwah, Srilakshmi Kalathil, Navjot Singh *et al.* (2020) "Reciped: A Resource for Exploring Recipes", *Database*, 2020.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico (2016) "Neural versus Phrase-based Machine Translation Quality: A Case Study", *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 257–267.

Bień, Michał, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz (2020) "RecipeNLG: A Cooking Recipes Dataset for Semi-structured Text Generation", *Proceedings of the 13th International Conference on Natural Language Generation*, 22–28.

Borji, Ali (2022) "Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and Dall-e 2", *arXiv preprint arXiv:2210.00586*.

Bosch, Marc, Fengqing Zhu, Nitin Khanna, Carol J. Boushey, and Edward J. Delp (2011) "Combining Global and Local Features for Food Identification in Dietary Assessment", *2011 18th IEEE International Conference on Image Processing*, 1789–1792.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell *et al.* (2020) "Language Models are Few-shot Learners", *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Catford, J.C. (1965) *A Linguistic Theory of Translation: An Essay in Applied Linguistics*, London: Oxford University Press.

Chan, Sin-wai (2002) "The Making of *TransRecipe*: A Translational Approach to the Machine Translation of Chinese Cookbooks", in Chan Sin-wai (ed.), *Translation and Information Technology*, Hong Kong: The Chinese University Press, 3–22.

Chan, Sin-wai (ed.) (2002) *Translation and Information Technology*, Hong Kong: The Chinese University Press.

Chandu, Khyathi, Eric Nyberg, and Alan W. Black (2019) "Storyboarding of Recipes: Grounded Contextual Generation", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6040–6046.

Chang, Minsuk, Léonore V. Guillain, Hyeungshik Jung, Vivian M. Hare, Juho Kim, and Maneesh Agrawala (2018) "Recipescape: An Interactive Tool for Analyzing

Cooking Instructions at Scale", *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

Chen, Jing-jing and Chong-Wah Ngo (2016) "Deep-based Ingredient Recognition for Cooking Recipe Retrieval", *Proceedings of the 24th ACM International Conference on Multimedia*, 32–41.

Chen, Jing-Jing, Chong-Wah Ngo, Fu-Li Feng, and Tat-Seng Chua (2018) "Deep understanding of cooking procedure for cross-modal recipe retrieval" *Proceedings of the 26th ACM International Conference on Multimedia*, 1020–1028.

Chen, Jing-jing, Chong-Wah Ngo, and Tat-Seng Chua (2017) "Cross-modal Recipe Retrieval with Rich Food Attributes", *Proceedings of the 25th ACM International Conference on Multimedia*, 1771–1779.

Chen, Mei, Kapil Dhingra, Wen Wu, Lei Yang, Rahul Sukthankar, and Jie Yang (2009) "PFID: Pittsburgh Fast-food Image Dataset", *2009 16th IEEE International Conference on Image Processing (ICIP)*, 289–292.

Chen, Mei-Yun, Yung-Hsiang Yang, Chia-Ju Ho, Shih-Han Wang, Shane-Ming Liu, Eugene Chang, Che-Hua Yeh, and Ming Ouhyoung (2012) "Automatic Chinese Food Identification and Quantity Estimation", *SIGGRAPH Asia 2012 Technical Briefs*, 1–4.

Chen, Nicholas, Yun Young Lee, Maurice Rabb, and Bruce Schatz (2010) "Toward Dietary Assessment via Mobile Phone Video Cameras", *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, 2010, 106.

Cho, Jaemin, Abhay Zala, and Mohit Bansal (2022) "DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-image Generative Transformers", *arXiv preprint arXiv:2202.04053*.

Cho, Kyunghyun, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014) "Learning Phrase Representations Using RNN Encoder–decoder for Statistical Machine Translation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734.

Cho, Kyunghyun, Bart Van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio (2014) "On the Properties of Neural Machine Translation: Encoder-decoder Approaches", *arXiv preprint arXiv:1409.1259*.

Christiano, Paul F., Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei (2017) "Deep Reinforcement Learning from Human Preferences", *Advances in Neural Information Processing Systems*, 30.

Church, Kenneth W. and William A. Gale (1991) "Concordances for Parallel Texts", *Using Corpora: Proceedings of the 7th Annual Conference of the UW for the New OED and Text Research*, Oxford: Oxford University Press, 40–62.

Deeney, John J. (1995) "Transcription, Romanization, Transliteration", in Chan Sin-wai and David E. Pollard (eds.), *An Encyclopaedia of Translation: Chinese-English. English-Chinese*, Hong Kong: The Chinese University Press, 1085–1097.

DeFrancis, John (1996) *ABC Chinese-English Dictionary*, Hong Kong: The Chinese University Press.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018) "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv preprint arXiv:1810.04805*.

Fatemi, Bahare, Quentin Duval, Rohit Girdhar, Michal Drozdzal, and Adriana Romero-Soriano (2023) "Learning to Substitute Ingredients in Recipes", *arXiv preprint arXiv:2302.07960*.

Fei, Hongliang, Tan Yu, and Ping Li (2021) "Cross-lingual Cross-modal Pretraining for Multimodal Retrieval", *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3644–3650.

Frome, Andrea, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov (2013) "DeViSE: A Deep Visualsemantic Embedding Model", *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, 2121–2129.

Fu, Han, Rui Wu, Chenghao Liu, and Jianling Sun (2020) "Mcen: Bridging cross-modal gap between cooking recipes and dish images with latent variable model" *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 14570–14580.

Fujii, Tatsuki, Yuichi Sei, Yasuyuki Tahara, Ryohei Orihara, and Akihiko Ohsuga (2019) "'Never Fry Carrots without Chopping' Generating Cooking Recipes from Cooking Videos Using Deep Learning Considering Previous Process", *International Journal of Networked and Distributed Computing*, 7(3): 107–112.

Fung, Kam Ling 馮金陵, Lee Ngan Woon 李銀煥, and Hui Choi Yip 許彩葉 (1994) 《清爽涼拌》(*Cold Dishes*), Hong Kong: Food Paradise Publishing Co. 飲食天地出版社.

Gabeur, Valentin, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid (2022) "Masking Modalities for Cross-modal Video Retrieval", *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1766–1775.

Galatolo, Federico A., Mario G.C.A. Cimino, and Edoardo Cogotti (2022) "TeTImEval: A Novel Curated Evaluation Data Set for Comparing Text-to-image Models", *arXiv preprint arXiv:2212.07839*.

Gatt, Albert and Emiel Krahmer (2018) "Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation", *Journal of Artificial Intelligence Research*, 61: 65–170.

Ghoddoosian, Reza, Saif Sayed, and Vassilis Athitsos (2022) "Hierarchical Modeling for Task Recognition and Action Segmentation in Weakly-labeled Instructional Videos", *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1922–1932.

Goel, Mansi, Pallab Chakraborty, Vijay Ponnaganti, Minnet Khan, Sritanaya Tatipamala, Aakanksha Saini, and Ganesh Bagler (2022) "Ratatouille: A Tool for Novel Recipe Generation", *2022 IEEE 38th International Conference on Data Engineering Workshops (ICDEW)*, 107–110.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David WardeFarley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014) "Generative Adversarial Networks", *arXiv preprint arXiv:1406.2661*.

Gove, Philip Babcock (comp.) (1986) *Webster's Third New International Dictionary*, Springfield, Mass.: Merriam-Webster Inc.

Guerrero, Ricardo, Hai X. Pham, and Vladimir Pavlovic (2021) "Cross-modal Retrieval and Synthesis (X-MRS): Closing the Modality Gap in Shared Subspace Learning", *Proceedings of the 29th ACM International Conference on Multimedia*, 3192–3201.

Guo, Zhao, Lianli Gao, Jingkuan Song, Xing Xu, Jie Shao, and Heng Tao Shen (2016) "Attention-based LSTM with Semantic Consistency for Videos Captioning", *Proceedings of the 24th ACM International Conference on Multimedia*, 357–361.

Han, Fangda, Guoyao Hao, Ricardo Guerrero, and Vladimir Pavlovic (2021) "Multi-attribute Pizza Generator: Cross-domain Attribute Control with Conditional StyleGAN", *arXiv preprint arXiv:2110.11830*.

Harashima, Jun, Yuichiro Someya, and Yohei Kikuta (2017) "Cookpad Image Dataset: An Image Collection as Infrastructure for Food Research", *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1229–1232.

Hasyim, Muhammad, Firman Saleh, Rudy Yusuf, and Asriani Abbas (2021) "Artificial Intelligence: Machine Translation Accuracy in Translating FrenchIndonesian Culinary Texts", *International Journal of Advanced Computer Science and Applications*, 12 (3).

Haynes, Colin (1998) *Breaking down the Language Barriers*, London: Aslib.

Hearst, Marti A., Susan T. Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf (1998) "Support Vector Machines", *IEEE Intelligent Systems and Their Applications*, 13(4): 18–28.

Hertz, Amir, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or (2022) "Prompt-to-prompt Image Editing with Cross Attention Control", *arXiv preprint arXiv:2208.01626*.

Heusel, Martin, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter (2017) "GANs Trained by a Two Time-scale Update Rule Converge to a Local Nash Equilibrium", *Advances in Neural Information Processing Systems*, 30.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020) "Denoising Diffusion Probabilistic Models", *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Hochreiter, Sepp (1998) "The Vanishing Gradient Problem during Learning Recurrent Neural Nets and Problem Solutions", *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 6(2): 107–116.

Hochreiter, Sepp and Jürgen Schmidhuber (1997) "Long Short-term Memory", *Neural Computation*, 9(8): 1735–1780.

Horita, Daichi, Ryosuke Tanno, Wataru Shimoda, and Keiji Yanai (2018) "Food Category Transfer with Conditional Cyclegan and a Large-scale Food Image Dataset", *Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management*, 67–70.

Hu, Fu 胡附(1984)《數詞和量詞》(*Numerals and Measure-words*), Shanghai: Shanghai Education Press 上海教育出版社.

Hu, Yuheng, Lydia Manikonda, and Subbarao Kambhampati (2014) "What We Instagram: A First Analysis of Instagram Photo Content and User Types", *Proceedings of the International AAAI Conference on Web and Social Media*, 8: 595–598.

Huang, Po-Yao, Xiaojun Chang, Alexander Hauptmann, and Eduard Hovy (2020) "Forward and Backward Multimodal NMT for Improved Monolingual and Multilingual Cross-modal Retrieval", *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 53–62.

Ifrah, Georges (2000) *The Universal History of Numbers: From Prehistory to the Invention of the Computer*, tr. David Bellos, E.F. Harding, Sophie Wood, and Ian Monk, New York: Wiley.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros (2017) "Image-to-image Translation with Conditional Adversarial Networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134.

Ji, Lei, Chenfei Wu, Daisy Zhou, Kun Yan, Edward Cui, Xilin Chen, and Nan Duan (2022) "Learning Temporal Video Procedure Segmentation from an Automatically

Collected Large Dataset", *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1506–1515.

Kagaya, Hokuto, Kiyoharu Aizawa, and Makoto Ogawa (2014) "Food Detection and Recognition Using Convolutional Neural Network", *Proceedings of the 22nd ACM International Conference on Multimedia*, 1085–1088.

Kawano, Yoshiyuki and Keiji Yanai (2014) "Foodcam: A Real-time Mobile Food Recognition System Employing Fisher Vector", *MultiMedia Modeling: 20th Anniversary International Conference, MMM 2014, Dublin, Ireland, January 6-10, 2014, Proceedings,* Part II 20, 369–373.

Kawar, Bahjat, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani (2022) "Imagic: Text-based Real Image Editing with Diffusion Models", *arXiv preprint arXiv:2210.09276*.

Khan, Rijwan, Santosh Kumar, Niharika Dhingra, and Neha Bhati (2021) "The Use of Different Image Recognition Techniques in Food Safety: A Study", *Journal of Food Quality*, 1–10.

Kiddon, Chloé, Luke Zettlemoyer, and Yejin Choi (2016) "Globally Coherent Text Generation with Neural Checklist Models", *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 329–339.

Kitamura, Keigo, Toshihiko Yamasaki, and Kiyoharu Aizawa (2008) "Food Log by Analyzing Food Images", *Proceedings of the 16th ACM International Conference on Multimedia*, 999–1000.

Kitamura, Keigo, Toshihiko Yamasaki, and Kiyoharu Aizawa (2009) "Foodlog: Capture, Analysis and Retrieval of Personal Food Images via Web", *Proceedings of the ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities*, 23–30.

Koehn, Philipp and Rebecca Knowles (2017) "Six Challenges for Neural Machine Translation", *Proceedings of the First Workshop on Neural Machine Translation*, 28–39.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu (2003) "Statistical Phrase-based Translation", *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 127–133.

Lee, Helena, Ke Shu, Palakorn Achananuparp, Philips Kokoh Prasetyo, Yue Liu, Ee-Peng Lim, and Lav R. Varshney (2020) "RecipeGPT: Generative Pretraining Based Cooking Recipe Generation and Evaluation System", *Companion Proceedings of the Web Conference 2020*, 181–184.

Lee, Ngan Woon 李銀煥 (1996) 《家常老幼靚湯》(*Soup for the Whole Family*), Hong Kong: Sea Shore Book Company 海濱圖書公司.

Li, Guang, Shubo Ma, and Yahong Han (2015) "Summarization-based Video Caption via Deep Neural Networks", *Proceedings of the 23rd ACM International Conference on Multimedia*, 1191–1194.

Li, Jiatong, Fangda Han, Ricardo Guerrero, and Vladimir Pavlovic (2021) "Picture-to-amount (pita): Predicting Relative Ingredient Amounts from Food Images", *2020 25th International Conference on Pattern Recognition (ICPR)*, 10343–10350.

Liu, Chang, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, and Yunsheng Ma (2016) "Deepfood: Deep Learning-based Food Image Recognition for Computer-aided Dietary Assessment", *Inclusive Smart Cities and Digital Health: 14th International Conference on Smart Homes and Health Telematics, ICOST 2016, Wuhan, China, May 25-27, 2016. Proceedings 14*, 37–48.

Liu, Jiatong (2021) "Multimodal Machine Translation", *IEEE Access*.

Liu, Vivian and Lydia B. Chilton (2022) "Design Guidelines for Prompt Engineering Text-to-image Generative Models", *CHI Conference on Human Factors in Computing Systems*, 1–23.

Liu, Xiao, Yansong Feng, Jizhi Tang, Chengang Hu, and Dongyan Zhao (2022) "Counterfactual Recipe Generation: Exploring Compositional Generalization in a Realistic Scenario", *arXiv preprint arXiv:2210.11431*.

Liu, Zhiming, Kai Niu, and Zhiqiang He (2023) "ML-CookGAN: Multi-label Generative Adversarial Network for Food Image Generation", *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s): 1–21.

Majumder, Bodhisattwa Prasad, Shuyang Li, Jianmo Ni, and Julian McAuley (2019) "Generating Personalized Recipes from Historical User Preferences", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5976–5982.

Malmaud, Jonathan, Jonathan Huang, Vivek Rathod, Nicholas Johnston, Andrew Rabinovich, and Kevin Murphy (2015) "'What's Cooking'? Interpreting Cooking Videos using Text, Speech, and Vision" *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 143–152.

Marin, Javier, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba (2019) "Recipe1m+: A Dataset for Learning Cross-modal Embeddings for Cooking Recipes and Food Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1): 187–203.

Metz, Cade and Priya Krishna (2022) "Can AI Write Recipes Better Than Humans?" *The Seattle Times* www.seattletimes.com/business/can-ai-write-recipes-better-than-humans/, retrieved on 17 April 2023.

Min, Weiqing, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz (2016) "Being a Supercook: Joint Food Attributes and Multimodal Content Modeling for Recipe Retrieval and Exploration", *IEEE Transactions on Multimedia*, 19(5): 1100–1113.

Mirza, Mehdi and Simon Osindero (2014) "Conditional Generative Adversarial Nets", *arXiv preprint arXiv:1411.1784*.

Mohammadshahi, Alireza, Rémi Lebret, and Karl Aberer (2019) "Aligning Multilingual Word Embeddings for Cross-modal Retrieval Task", *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, 11–17.

Mokady, Ron, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or (2022) "Null-text Inversion for Editing Real Images Using Guided Diffusion Models", *arXiv preprint arXiv:2211.09794*.

Mori, Shinsuke, Hirokuni Maeta, Tetsuro Sasada, Koichiro Yoshino, Atsushi Hashimoto, Takuya Funatomi, and Yoko Yamakata (2014) "Flowgraph2text: Automatic Sentence Skeleton Compilation for Procedural Text Generation", *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, 118–122.

Mori, Shinsuke, Hirokuni Maeta, Yoko Yamakata, and Tetsuro Sasada (2014) "Flow Graph Corpus from Recipe Texts", *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2370–2377.

Nagao, Makoto (1984) "A Framework of a Mechanical Translation between Japanese and English by Analogy", in Alick Elithorn and Ranan Barneiji (eds.), *Artificial and Human Intelligence*, Amsterdam: North-Holland Publishing Company, 73–80.

Newmark, Peter (1981) *Approaches to Translation*, Oxford: Pergamon Press.

Newmark, Peter (1988) *A Textbook of Translation*, Hertfordshire: Prentice-Hall.

Nichol, Alexander Quinn, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen (2022) "GLIDE: Towards

Photorealistic Image Generation and Editing with TextGuided Diffusion Models", *International Conference on Machine Learning*, 16784–16804.

Nishimura, Taichi, Atsushi Hashimoto, and Shinsuke Mori (2019) "Procedural Text Generation from a Photo Sequence", *Proceedings of the 12th International Conference on Natural Language Generation*, 409–414.

OpenAI (2022) "Introducing ChatGPT", accessed on 26 April 2023, available at https://openai.com/blog/chatgpt.

Pan, Liang-Ming, Jingjing Chen, Jianlong Wu, Shaoteng Liu, Chong-Wah Ngo, Min-Yen Kan, Yugang Jiang, and Tat-Seng Chua (2020) "Multi-modal Cooking Workflow Construction for Food Recipes", *Proceedings of the 28th ACM International Conference on Multimedia*, 1132–1141.

Papadopoulos, Dim P., Youssef Tamaazousti, Ferda Ofli, Ingmar Weber, and Antonio Torralba (2019) "How to Make a Pizza: Learning a Compositional Layer-based Gan Model", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8002–8011.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002) "Bleu: A Method for Automatic Evaluation of Machine Translation", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Pavlichenko, Nikita, Fedor Zhdanov, and Dmitry Ustalov (2022) "Best Prompts for Text-to-image Models and How to Find Them", *arXiv preprint arXiv:2209.11711.*

Petsiuk, Vitali, Alexander E. Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A. Plummer, Ori Kerret *et al.* (2022) "Human Evaluation of Text-to-image Models on a Multi-task Benchmark", *arXiv preprint arXiv:2211.12112.*

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1985) *A Comprehensive Grammar of the English Language*, New York: Longman Group Ltd.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever *et al.* (2019) "Language Models Are Unsupervised Multitask Learners", *OpenAI Blog*, 1(8): 9.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark *et al.* (2021) "Learning Transferable Visual Models from Natural Language Supervision", *International Conference on Machine Learning*, 8748–8763.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen (2022) "Hierarchical Text-conditional Image Generation with Clip Latents", *arXiv preprint arXiv:2204.06125.*

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer (2022) "High-resolution Image Synthesis with Latent Diffusion Models", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Sagart, Laurent (1999) *The Roots of Old Chinese*, Amsterdam and Philadelphia: John Benjamins Publishing Company.

Sahoo, Doyen, Wang Hao, Shu Ke, Wu Xiongwei, Hung Le, Palakorn Achananuparp, Ee-Peng Lim, and Steven C.H. Hoi (2019) "FoodAI: Food Image Recognition via Deep Learning for Smart Food Logging", *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2260–2268.

Salvador, Amaia, Michal Drozdzal, Xavier Giro-i Nieto, and Adriana Romero (2019) "Inverse Cooking: Recipe Generation from Food Images", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10453–10462.

Salvador, Amaia, Erhan Gundogdu, Loris Bazzani, and Michael Donoser (2021) "Revamping Cross-modal Recipe Retrieval with Hierarchical Transformers and Self-supervised Learning", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15475–15484.

Salvador, Amaia, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba (2017) "Learning Cross-modal Embeddings for Cooking Recipes and Food Images", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3020–3028.

Sato, Takayuki, Jun Harashima, and Mamoru Komachi (2016) "Japanese-English Machine Translation of Recipe Texts", *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, 58–67.

Schäffner, Christina and Helen Kelly-Holmes (1995) *Cultural Functions of Translation*, Clevedon: Multilingual Matters.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016) "Improving Neural Machine Translation Models with Monolingual Data", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 86–96.

Seo, Paul Hongsuck, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid (2022) "End-to-end Generative Pretraining for Multimodal Video Captioning", *arXiv preprint arXiv:2201.08264*.

Simoons, Frederick J. (1991) *Food in China: A Cultural and Historical Inquiry*, Boca Raton: CRC Press.

Sohl-Dickstein, Jascha, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli (2015) "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics", *International Conference on Machine Learning*, 2256–2265.

Song, Yang, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole (2020) "Score-based Generative Modeling through Stochastic Differential Equations", *arXiv preprint arXiv:2011.13456*.

Su, Han, Ting-Wei Lin, Cheng-Te Li, Man-Kwan Shan, and Janet Chang (2014) "Automatic Recipe Cuisine Classification by Ingredients", *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 565–570.

Sugiyama, Yu and Keiji Yanai (2021) "Cross-modal Recipe Embeddings by Disentangling Recipe Contents and Dish Styles", *Proceedings of the 29th ACM International Conference on Multimedia*, 2501–2509.

Sulubacak, Umut, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann (2020) "Multimodal Machine Translation through Visuals and Speech", *Machine Translation*, 34: 97–147.

Sun, Chen, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid (2019) "VideoBERT: A Joint Model for Video and Language Representation Learning", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7464–7473.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014) "Sequence to Sequence Learning with Neural Networks", *Advances in Neural Information Processing Systems*, 27: 3104–3112.

Tai, Kuo-Chung (1979) "The Tree-to-tree Correction Problem", *Journal of the ACM (JACM)*, 26(3): 422–433.

Tam, Ng Wai Fung 譚吳威鳳 (1997) 《超值營養食譜》(*Healthy Money-saving Meals*), Hong Kong: Wan Li Book Co. Ltd 萬里機構and Food Paradise Publishing Co. 飲食天地出版社.

Tanno, Ryosuke, Daichi Horita, Wataru Shimoda, and Keiji Yanai (2018) "Magical Rice Bowl: A Real-time Food Category Changer", *Proceedings of the 26th ACM International Conference on Multimedia*, 1244–1246.

Teng, Chun-Yuen, Yu-Ru Lin, and Lada A. Adamic (2012) "Recipe Recommendation Using Ingredient Networks", *Proceedings of the 4th Annual ACM Web Science Conference*, 298–307.

Treffry, Diana (ed.) (1998) Collins English Dictionary, 4th edition. Glasgow: HarperCollins Publishers.

Ueda, Mayumi, Mari Takahata, and Shinsuke Nakajima (2011) "Recipe Recommendation Method Based on User's Food Preferences", *Proceedings of the IADIS International Conference on E-Society*, 591–594.

Van Asbroeck, Stephanie and Christophe Matthys (2020) "Use of Different Food Image Recognition Platforms in Dietary Assessment: Comparison Study", *JMIR Formative Research*, 4(12): e15602.

Varshney, Lav R., Florian Pinel, Kush R. Varshney, Debarun Bhattacharjya, Angela Schörgendorfer, and Y-M Chee (2019) "A Big Data Approach to Computational Creativity: The Curious Case of Chef Watson", *IBM Journal of Research and Development*, 63(1): 7–11.

Vasconcellos, Muriel (1996) *Recent Trends in Machine Translation*, London: Aslib.

Venuti, Lawrence (1995) *The Translator's Invisibility: A History of Translation*, London and New York: Routledge.

Vinay, Jean-Paul and Jean Darbelnet (1954, 1995) *Comparative Stylistics of French and English: A Methodology for Translation*, Juan C. Sager and M.-J. Hamel (tr. and ed.), Amsterdam and Philadelphia: John Benjamins Publishing Company.

Wang, Hao, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven C.H. Hoi (2019) "Learning Cross-modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11572–11581.

Wang, Liping, Qing Li, Na Li, Guozhu Dong, and Yu Yang (2008) "Substructure Similarity Measurement in Chinese Recipes", *Proceedings of the 17th International Conference on World Wide Web*, 979–988.

Wang, Su, Honghao Gao, Yonghua Zhu, Weilin Zhang, and Yihai Chen (2019) "A Food Dish Image Generation Framework Based on Progressive Growing GANs", *Collaborative Computing: Networking, Applications and Worksharing: 15th EAI International Conference, CollaborateCom 2019, London, UK, August 19-22, 2019, Proceedings 15*, 323–333.

Wang, Wenjie, Ling-Yu Duan, Hao Jiang, Peiguang Jing, Xuemeng Song, and Liqiang Nie (2021) "Market2Dish: Health-aware Food Recommendation", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1): 1–19.

Weellisch, Hans (1975) *Transcription and Transliteration*, Silver Spring, MD: Institute of Modern Language.

Wikitionary (CC BY-SA 3.0) https://en.wiktionary.org/wiki/recipe accessed: 2023-04-27

Xie, Haoran, Lijuan Yu, and Qing Li (2010) "A Hybrid Semantic Item Model for Recipe Search by Example", *2010 IEEE International Symposium on Multimedia*, 254–259.

Xie, Zhongwei, Ling Liu, Lin Li, and Luo Zhong (2021) "Learning Joint Embedding with Modality Alignments for Cross-modal Retrieval of Recipes and Food Images", *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, 2221–2230.

Xu, Frank F., Lei Ji, Botian Shi, Junyi Du, Graham Neubig, Yonatan Bisk, and Nan Duan (2020) "A Benchmark for Structured Procedural Knowledge Extraction from Cooking Videos", *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, 30–40.

Yam, Lisa方任利莎 (1997)《方太食譜之魚蝦蟹》(*Lisa Yam's Cook Book: Seafood*), Hong Kong: Ming Pao Press Ltd. 明窗出版社.

Yamakata, Yoko, Shinji Imahori, Hirokuni Maeta, and Shinsuke Mori (2016) "A Method for Extracting Major Workflow Composed of Ingredients, Tools, and Actions from Cooking Procedural Text", *2016 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 1–6.

Yamamoto, Kohei and Keiji Yanai (2022) "Text-based Image Editing for Food Images with CLIP", *Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management*, 29–37.

Yanai, Keiji, Kaimu Okamoto, Tetsuya Nagano, and Daichi Horita (2019) "Large-scale Twitter Food Photo Mining and its Applications", *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 77–85.

Yang, Shulin, Mei Chen, Dean Pomerleau, and Rahul Sukthankar (2010) "Food Recognition Using Statistics of Pairwise Local Features", *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2249–2256.

Yu, Zhiwei, Hongyu Zang, and Xiaojun Wan (2020) "Routing Enforced Generative Model for Recipe Generation", *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3797–3806.

Zhang, Deqin 張德鑫 (1999) 《數裏乾坤》(*The Mysterious World of Numerals*), Beijing: Peking University Press 北京大學出版社.

Zhang, Yixin, Yoko Yamakata, and Keishi Tajima (2019) "Categorization of Cooking Actions Based on Textual/Visual Similarity", *Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management*, 42–49.

Zhou, Kaiyang, Jingkang Yang, Chen Change Loy, and Ziwei Liu (2022) "Learning to Prompt for Vision-language Models", *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, Luowei, Chenliang Xu, and Jason J. Corso (2018) "Towards Automatic Learning of Procedures from Web Instructional Videos", *Thirty-second AAAI Conference on Artificial Intelligence*.

Zhu, Bin, Chong-Wah Ngo, and Wing-Kwong Chan (2021) "Learning from Web Recipe-image Pairs for Food Recognition: Problem, Baselines and Performance", *IEEE Transactions on Multimedia*, 24: 1175–1185.

Zhu, Bin and Chong-Wah Ngo (2020) "CookGAN: Causality based text-to-image synthesis", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5519–5527.

Zhu, Bin, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao (2019) "R2gan: Cross-modal Recipe Retrieval with Generative Adversarial Network", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11477–11486.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A Efros (2017) "Unpaired Image-to-image Translation Using Cycle-consistent Adversarial Networks", *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232.