

Singapore Management University

## Institutional Knowledge at Singapore Management University

---

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

---

6-2022

### Decomposing generation networks with structure prediction for recipe generation

Hao WANG

Guosheng LIN

Steven C. H. HOI

*Singapore Management University*, [chhoi@smu.edu.sg](mailto:chhoi@smu.edu.sg)

Chunyan MIAO

Follow this and additional works at: [https://ink.library.smu.edu.sg/sis\\_research](https://ink.library.smu.edu.sg/sis_research)



Part of the [Databases and Information Systems Commons](#), and the [Numerical Analysis and Scientific Computing Commons](#)

---

#### Citation

WANG, Hao; LIN, Guosheng; HOI, Steven C. H.; and MIAO, Chunyan. Decomposing generation networks with structure prediction for recipe generation. (2022). *Pattern Recognition*. 126, 1-9.

Available at: [https://ink.library.smu.edu.sg/sis\\_research/6962](https://ink.library.smu.edu.sg/sis_research/6962)

This Journal Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email [cherylds@smu.edu.sg](mailto:cherylds@smu.edu.sg).

# Decomposing generation networks with structure prediction for recipe generation

Hao Wang <sup>a,c</sup>, Guosheng Lin <sup>a</sup>, Steven C.H. Hoi <sup>b</sup>, Chunyan Miao <sup>a,c,\*</sup>

<sup>a</sup> School of Computer Science and Engineering, Nanyang Technological University, Singapore

<sup>b</sup> School of Information Systems, Singapore Management University, Singapore

<sup>c</sup> Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly, NTU

Published in Pattern Recognition, 2022, 126, 108578. DOI: 10.1016/j.patcog.2022.108578

**Abstract:** Recipe generation from food images and ingredients is a challenging task, which requires the interpretation of the information from another modality. Different from the image captioning task, where the captions usually have one sentence, cooking instructions contain multiple sentences and have obvious structures. To help the model capture the recipe structure and avoid missing some cooking details, we propose a novel framework: Decomposing Generation Networks (DGN) with structure prediction, to get more structured and complete recipe generation outputs. Specifically, we split each cooking instruction into several phases, and assign different sub-generators to each phase. Our approach includes two novel ideas: (i) learning the recipe structures with the global structure prediction component and (ii) producing recipe phases in the sub-generator output component based on the predicted structure. Extensive experiments on the challenging large-scale Recipe1M dataset validate the effectiveness of our proposed model, which improves the performance over the state-of-the-art results.

**Keywords:** Text generation, Vision-and-language

## 1. Introduction

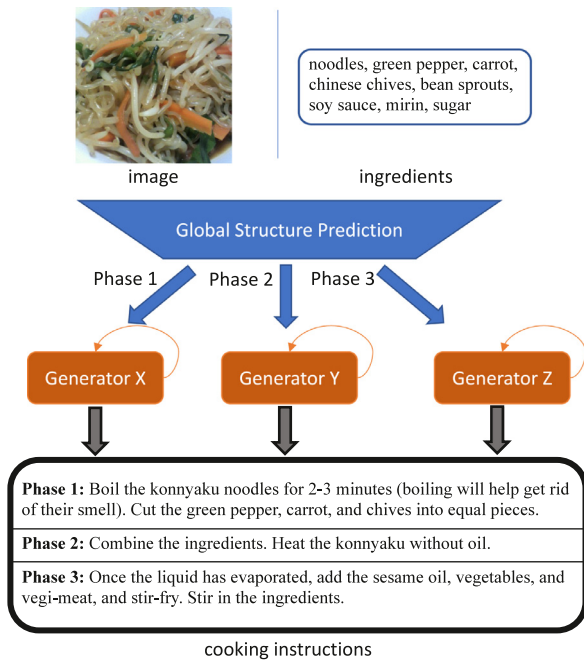
Recent food-related research works such as food image recognition [1], food retrieval [2], [3] and recipe generation [4] have raised great interests, as food is very close to people’s daily life. Recipe generation is an emerging problem, where we are interested in automatically producing recipes (cooking instructions) based on food images. In this paper, we investigate an open research task of recipe generation, and propose a novel approach to resolve this task and jointly understand the multi-modal food data including food images and recipes.

We use Recipe1M [2] dataset in this paper, which is large-scale and challenging, and only contains the static cooked food images instead of image sequence [5] for recipe generation. Since people may be curious about the exact recipe for a cooked food image and it would be hard to collect large-scale instructional video data in real world, we believe that generating cooking instructions from one single food image is of more value, compared to producing instructions from image sequence [5]. The prior recipe generation work [4] on Recipe1M dataset [2] mainly fails to explicitly learn the recipe global structure, where they give the whole cooking instruction sentences through a single decoder, and may result in some cooking steps to be missing.

Specifically, cooking instructions are one kind of procedural text, which are constructed step

by step with some format. For example, as is shown in Fig. 1, the cooking instructions are composed of several sentences, and each sentence starts with a verb in most cases. Apart from dividing the cooking instructions by sentences, we may also split them into more general phases, which represent the global structures of the cooking recipes. Imagine when people start cooking food, we may decompose the cooking procedure into some basic phases first, e.g. pre-process the ingredients, cook the main dish, etc. Then we will focus on some details, like determining which ingredients to use. While this coarse-to-fine reasoning is trivial for humans, most algorithms do not have the capacity to reason about the phase information contained in the static food image [6]. Hence it is important to guide the model to keep aware of the global structure of the recipe during generation, otherwise the generation outputs can hardly cover all the cooking details [4].

In this paper, we aim to capture the global structure of recipe and to generate the cooking instructions from one single image with a list of ingredients. The basic idea is that we first (i) assemble some of the consecutive steps to form a phase, (ii) assign suitable sub-generators to produce certain instruction phases, and (iii) concatenate the phases together to form the final recipes. We propose a novel framework of *Decomposing Generation Networks* (DGN)



**Fig. 1.** Illustration of the Decomposing Generation Networks (DGN) for recipe generation. Instead of producing instructions directly from the image and ingredient embedding [4], we first predict the instruction structure and choose different generators to match the cooking phases. And then we combine the outputs of selected sub-generators to get the final generated recipes.

with global structure prediction, to achieve the coarse-to-fine reasoning. Fig. 2 shows the pipeline of the framework. To be specific, DGN is composed of two components, i.e. the global structure prediction component and the sub-generator output component. To obtain the global structure of the cooking instruction, we input image and ingredient representations into global structure prediction component, and get the sub-generator selections as well as their orders. Then in the sub-generator output component, we adopt attention mechanism to get the phase-aware features. The phase-aware features are designed for different sub-generators and help the sub-generators produce better instruction phases.

We have conducted extensive experiments on the Recipe1M [2] dataset, and evaluated the recipe generation results by different evaluation metrics. We find our proposed model DGN outperforms the state-of-the-art methods across different metrics.

## 2. Related work

### 2.1. Food computing

Our work is closely related to food computing [7], which utilizes computational methods to analyze the food data including the food images and recipes. With the development of social media and mobile devices, more and more food data become available on the Internet, the UEC Food100 dataset [8] and ETHZ Food-101 dataset [1] are proposed for the food recognition task. The previous two food datasets are restricted to the variety of data types, only have different categories of food images. YouCook2 dataset is proposed by Zhou et al. in [5], which contains cooking video data. They focused on generating cooking instruction steps from video segments in YouCook2 dataset. The latter work [9] proposed a new food dataset, Storyboarding, where the food data item has multiple images aligned with instruction steps. In their work, they proposed to utilize a scaffolding structure for the model representations. Besides, Bosselut et al. [6] generated the recipes based on the text, where they reasoned about causal effects that are not mentioned

in the surface strings, they achieved this with memory architectures by dynamic entity tracking and obtained a better understanding on procedural text.

In order to better model the relationship between recipes and food images, Recipe1M [2] has been proposed to provide richer food image, cooking instruction, ingredient, and semantic food-class information. Recipe1M contains large amounts of image-recipe pairs, which can be applied on the cross-modal food retrieval [2,3], food ingredient prediction [4] and cooking recipe generation task [4,10]. Salvador et al. [4] focused more on the ingredient prediction task. For instruction generation, they [4] generated the whole cooking instructions from given food images and ingredients through a single decoder directly, which may result in that some cooking details can be missing in some cases. Wang et al. [10] used the recipe structure information to boost the generation performance, where they first learned the sentence-level tree structures for the lengthy cooking instructions. They embedded the inferred tree structures using the graph convolution networks, and they generated the recipes based on the concatenation of image and structure representations. In [10], the recipe structure information is integrated implicitly into the generation process. In contrast, we adopt an explicit way to use the structure information, where we use different sub-generators for recipe phase generation.

It is worth noting that, [4] is used as the recipe generation task baseline on Recipe1M dataset. We take food images and ingredients as inputs in all experiments for fair comparison. Our proposed DGN approach improves the recipe generation performance by introducing the decomposing idea to the generation process. Hence, our proposed methods can be applied to many general models. We will demonstrate the details in Section 4.

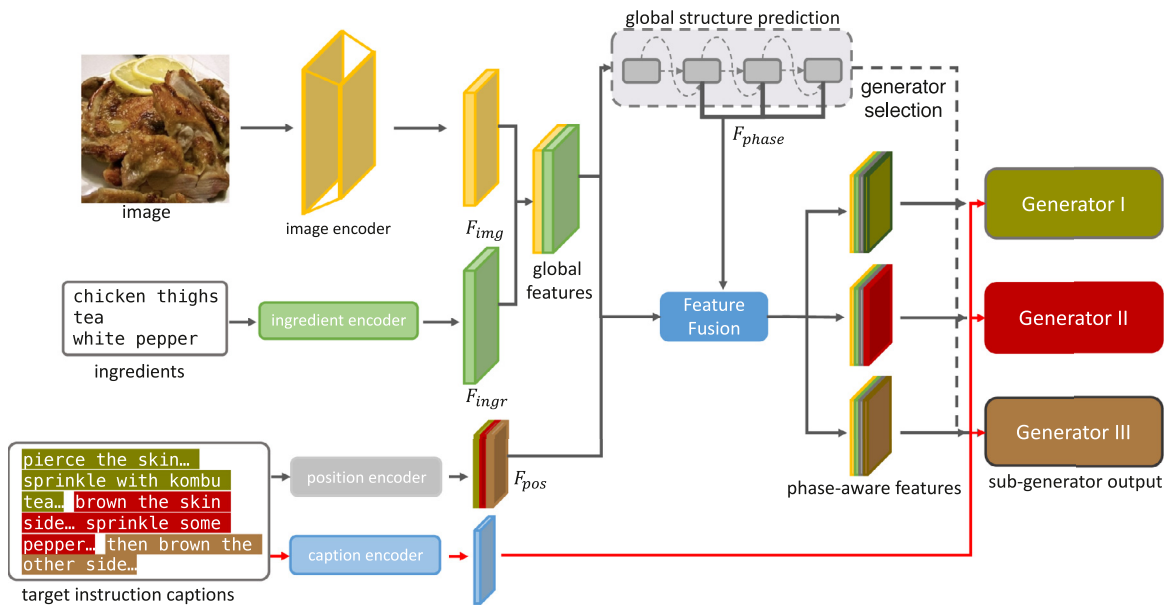
### 2.2. Text generation

Text generation is a widely researched task, which can take various input types as source information. Machine translation [11,12] is one of the representative works of text-based generation, in which the decoder takes one language text as the input and outputs another language sentences. Image-based text generation involves both vision and language, such as image captioning [13,14], visual question answering [15,16]. To be specific, the task of image captioning [13,14] is to generate textual descriptions for the given images, which is correlated with our recipe generation task. Anderson et al. [17] propose to use the attention mechanism to integrate the detected object features and improve the captioning performance. With the advent of transformer-based models [12], Li et al. [18] and Zhang et al. [19] use the detected object features as the visual tokens and the textual word tokens to train a large-scale vision-language pretrained model. It demonstrates the pretrained vision-language model help boost the captioning performance.

It is notable that the recipe generation task is clearly different from the image captioning task. Since in the food recipe dataset (Recipe1M), all ingredients are mixed and cooked in the food images, and the prevailing object detectors can hardly produce reasonable detected results. Therefore, the detector-based image captioning methods [13,14,18,19] are not applicable to the recipe generation task. Moreover, the cooking recipes contain multiple sentences, while each textual caption in image captioning datasets only has one sentence. In this paper, to address the challenging recipe generation problem, we propose to use the decomposing framework and alleviate the difficulty of generating lengthy cooking recipes.

### 2.3. Neural module networks

The idea of using neural module network to decompose neural models have been proposed for some language-vision intersec-



**Fig. 2. Decomposing Generation Networks with global structure prediction (DGN):** We take food images and the corresponding ingredients as model inputs, and obtain the image and ingredient embedding  $F_{img}$ ,  $F_{ingr}$  through a pre-trained image model CNN and the language model BERT respectively. After that, the model will be split into two branches, i.e. the global structure prediction component and the sub-generator output component. Both of them are constructed by the transformer. The global structure prediction component produces the sub-generator selections and their orders for the following branch. The sub-generator output component fuses  $F_{img}$ ,  $F_{ingr}$ , the position representations  $F_{pos}$  and the phase vector  $F_{phase}$  to obtain the input of each sub-generator, and produces different phases of the recipe.

tion tasks, such as visual question answering [20], image captioning [21], visual reasoning [22]. Neural module network has good capabilities to capture the structured knowledge representations of input images or sentences. In general, since the image layouts or questions are obviously structured, the aforementioned research works [20–22] focused on constructing better encoders with neural modules. To produce a coherent story for an image in MS COCO [23], Krause et al. [24] decomposed both images and paragraphs into their constituent parts, detecting semantic regions in images and using a hierarchical recurrent neural network to generate topic vectors with their corresponding sentences, but they generated different paragraph parts with the same decoder, which restricted the recipe generation performance. In food data of Recipe1M [2], the cooking instructions tend to be very structured as well. To generate recipes with better structures, we employ different sub-generators to produce different phases of cooking instructions.

### 3. Method

#### 3.1. Overview

In Fig. 2, we show the training flow of DGN. It is observed that the cooking instructions have obvious structures and clear formats, most cooking instruction sentences in Recipe1M dataset [2] start with a verb, e.g. *heat*, *combine*, *pierce*, etc. Since how to automatically divide the recipes into phases remains a challenging problem, we adopt a simple yet effective way to separate recipes into phases, where we use a pre-defined rule. Specifically, we split per instructions into 2 – 3 phases and try to ensure each phase shares equal sentence numbers, where one or more cooking steps (sentences) will map to one phase. This recipe segmentation rule is based on our empirical observations, since segmenting more recipe phases makes it unfeasible to build the hierarchy between cooking phases and steps. We show the example of segmented phases in Fig. 2, the recipe for the *roasted chicken* totally has five steps, which are transitioned to three phases.

Since different recipe phases have varying aspects, we adopt various types of sub-generators for phase generation. Based on the

verbs in cooking steps, we use the approach of k-means clustering to assign pseudo labels to each recipe phase. The pseudo labels indicate which sub-generator can be selected to generate certain phases. Specifically, we first extract all the verbs in recipes with spaCy [25], a Natural Language Processing (NLP) tool. Then, we can obtain the mean verb representations, which can be viewed as the representation of each phase. After that, we use k-means clustering to get pseudo labels  $G = \{g_1, \dots, g_k | Ca(g_i) \in [1, N]\}$  for phases  $\{i, \dots, k\}$  to indicate the selections of sub-generators, where  $Ca$  denotes the category of the sub-generator  $g_i$ . Note that the number of the sub-generator category  $N$  is a hyper-parameter, we do experiments with different  $N$  and show the results in Table 3.

Fig. 2 provides an overview of our proposed model, which is composed of the global structure prediction component and sub-generator output component. Our model takes food images and their corresponding ingredients as input. It uses several sub-generators for different recipe phases, allowing sub-generators to focus on different clustered recipe phases.

ResNet-50 [26] pretrained on ImageNet [27] and BERT [28] model implemented by Wolf et al. [29] are used to encode food images and ingredients respectively. We can get image and ingredient global representations  $F_{img}$  and  $F_{ingr}$ . These global representations will be fed into the global structure prediction component, to decide which sub-generators will be selected as well as their orders. To enable the interactions among sub-generators, the global structure prediction component also produces a  $P$ -dimensional phase vector  $F_{phase}$  for each of the sub-generators. Then we split the target instructions into phases and assign different position one-hot vectors  $v_p \in \mathbb{R}^3$  for each phase, which will be transformed into a  $P$ -dimensional position representations  $F_{pos}$  through a linear layer. With previous encoded features  $F_{img}$ ,  $F_{ingr}$ ,  $F_{phase}$  and  $F_{pos}$ , we can fuse them together and obtain the phase-aware features  $\mathbf{r}_i \in \mathbb{R}^P$  for sub-generator  $g_i$ .

#### 3.2. Global structure prediction component

Since the cooking instructions are divided into phases, the global structure prediction component not only needs to decide

which generators to be selected in each phase, but also is required to predict the order of the chosen sub-generators. In order to achieve the goal, we stack the transformer blocks [12] to construct our global structure prediction component. The last transformer block is followed by a linear layer and a softmax activation, to find the predictions for each step. We set hidden size  $H = 512$ , the number of heads  $n_{head} = 8$  and the number of stacked layers  $n_{layer} = 4$ , generate the sub-generator label sequence  $\{y_i, \dots, y_k\}$ .

To be specific, the transformer block contains two sub-layers with layer normalization, where the first one employs the multi-head self-attention mechanism and the second one attends to the model conditional inputs to enhance the self-attention output. The attention outputs can be computed as [12],

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where the input comes from queries  $Q$  and keys  $K$  of dimension  $d_k$ , and values  $V$  of dimension  $d_v$ . We also adopt the multi-head attention mechanism [12], which linearly maps  $Q, K, V$  with different, learned projections. These different projected results will be concatenated together and get better output values.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O, \quad (2)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \quad (3)$$

Where the projections are matrices  $W_i^Q \in \mathbb{R}^{d_k}$ ,  $W_i^K \in \mathbb{R}^{d_k}$ ,  $W_i^V \in \mathbb{R}^{d_v}$  and  $W^O \in \mathbb{R}^{d_v n_{head}}$ .

We take the global context vectors  $\{F_{img}, F_{ingr}\}$  and target recipe phase labels  $G = \{[START], g_i, \dots, g_k\}$  as inputs when training the model. We first map the discrete labels to a sequence of continuous representations  $Z$ . The model generates an output sequence  $\{y_i, \dots, y_k\}$  one element at a time. The target sequence embedding  $Z$  will be first fed into the model and processed with multi-head self-attention layers, as follows:

$$H_{self}^{attn} = \text{MultiHead}(Z, Z, Z), \quad (4)$$

We further concatenate the context vectors  $\{F_{img}, F_{ingr}\}$  together, get the conditional vector  $F_{kv}$ , which will be attended to refine previous self-attention outputs  $H_{self}^{attn}$ , which is defined as:

$$H_{cond}^{attn} = \text{MultiHead}(H_{self}^{attn}, F_{kv}, F_{kv}), \quad (5)$$

$H_{cond}^{attn}$  is the final attention outputs of each phase, which can be used as the phase vector  $F_{phase}$  for sub-generator output component. We transform  $H_{cond}^{attn}$  into  $H_{cond}'^{attn}$  for output token generation with a linear layer. The dimension of  $H_{cond}'^{attn}$  is identical with the number of sub-generator category  $N$ , the probabilities of generated tokens are  $p^{gen} = \text{softmax}(H_{cond}'^{attn})$ . Therefore, the final output tokens of global structure prediction component  $y_i = \text{argmax}(p^{gen})$ . We train the global structure prediction component with cross-entropy loss  $\mathcal{L}_{pre}$ :

$$\mathcal{L}_{pre} = \sum_{i=1}^S \ell_{cross-entropy}(p_i^{gen}, g_i), \quad (6)$$

where  $S$  is the number of instruction phases.

### 3.3. Sub-Generator output component

The sub-generator output component uses different sub-generators predicted by global structure prediction component, to produce a certain phase of the recipe, and concatenate them together to form the final cooking instruction. We stack 16 transformer blocks to construct the generator, in which 12 of them are shared blocks, and the rest 4 are independent blocks of each of

the generators. The reasons for using shared blocks lie in that the model may overfit to the limited training data and cannot generalize well, if we adopt whole independent blocks for each sub-generator.

We utilize each predicted sub-generator to produce one recipe phase, which requires that each of the generator inputs should be discriminative and informative enough. Therefore, we incorporate various sources of feature representations, i.e. the food image features  $F_{img}$ , the ingredient features  $F_{ingr}$ , the position representations  $F_{pos}$  and the phase vector  $F_{phase}$  ( $H_{cond}^{attn}$ ) produced by global structure prediction component.  $F_{img}$  provides the model with generation contents from the food images, which belong to a different modality, and  $F_{ingr}$  indicates the ingredients containing in the recipe, which can be reused in the generated cooking instructions. To allow the model to be aware of the generation phase, we fuse the recipe phase position representations  $F_{pos}$ , which is constructed by the position one-hot vectors  $\nu_p \in \mathbb{R}^3$  of each phase.  $F_{phase}$  is computed with the same process as  $H_{cond}^{attn}$ , which is incorporated for enhancing the interactions among different sub-generators and helps the model adapt to different generation phases.

The above four representations will be fused together to get the phase-aware features  $\mathbf{r} = \langle F_{img}, F_{ingr}, F_{pos}, F_{phase} \rangle$ , which are the inputs of sub-generators. We adopt two different ways to achieve that. The first one is that we simply concatenate these representations, and get  $\mathbf{r}_{cat}$ . In the second way, we use attention mechanism to make  $F_{img}, F_{ingr}$  attend to the concatenated embedding  $\mathbf{cat}(F_{pos}, F_{phase})$  respectively. To give attention representations, specifically we utilize projection matrix  $W_1$  and  $W_2$  on  $\mathbf{cat}(F_{pos}, F_{phase})$  and get the attention maps for  $F_{img}$  and  $F_{ingr}$ , the image and ingredient attention outputs can be formulated as:

$$F_{img}^{attn} = \text{softmax}(W_1(\mathbf{cat}(F_{pos}, F_{phase})))F_{img}, \quad (7)$$

$$F_{ingr}^{attn} = \text{softmax}(W_2(\mathbf{cat}(F_{pos}, F_{phase})))F_{ingr}, \quad (8)$$

The final attended phase-aware features  $\mathbf{r}_{attn}$  is the concatenation of  $F_{img}^{attn}$  and  $F_{ingr}^{attn}$ . It is worth noting that we involve an additional position classifier  $\mathcal{L}_{pos}$  on  $\mathbf{r}$  to ensure that it contains certain phase position information, such that the input representations  $\mathbf{r}_i$  of different sub-generators can keep aware of their positions in the recipes, and consequently the overall combined recipes from phases would be more coherent.

$$\mathcal{L}_{pos} = \sum_{i=1} \ell_{cross-entropy}(\widehat{pos}_{\mathbf{r}_i}, pos_{\mathbf{r}_i}), \quad (9)$$

We also need to input the target instruction captions  $t = \{[START], t_1, t_2, \dots, t_m\}$  for training the Transformer [12] generators, and map them to a continuous representation  $C$ . As described in Section 3.2, we utilize attention mechanism with transformer blocks:

$$\mathbf{F}_{self}^{attn} = \text{MultiHead}(C, C, C), \quad (10)$$

$$\mathbf{F}_{cond}^{attn} = \text{MultiHead}(\mathbf{F}_{self}^{attn}, \mathbf{r}, \mathbf{r}), \quad (11)$$

We use  $\mathbf{F}_{cond}^{attn}$  to generate the tokens through a linear layer and softmax activation, and we can obtain the output probabilities  $p^{token}$  among candidate tokens. For each sub-generator, we compute the training loss as follows:

$$\mathcal{L}_{gen} = \sum_{i=1}^M \ell_{cross-entropy}(p_i^{token}, t_i), \quad (12)$$

### 3.4. Training and inference

The food images, ingredients and the target instruction captions are taken as the training input of the model. We totally have three

**Table 1**

Main Results. Evaluation of DGN performance against various settings. We first show the baseline results of three different ingredient encoders, where we adopt the word embedding layer, pretrained LSTM-based model ELMo and pretrained transformer-based model BERT respectively. We use ResNet-50 as the image encoder in all the experiments. DGN is added to the baseline models as an additional branch, where we show the results of different construction ways of phase-aware features  $\mathbf{r}$ . **DGN (cat)** uses the concatenation of the provided representations for the sub-generator inputs, and **DGN (attn)** adopts the attention mechanism to enhance the representations. Moreover, we also compare our results with the state-of-the-art model SGN. We evaluate the model with perplexity (lower is better), BLEU (higher is better) and ROUGE-L (higher is better). We find the proposed DGN improves the baseline performance across all the metrics.

Methods	Ingredient Encoder	Perplexity	BLEU	ROUGE-L
<b>Baseline [4]</b>	Embedding Layer	8.06	7.23	31.8
<b>DGN (cat)</b>	Embedding Layer	7.40	9.93	34.5
<b>DGN (attn)</b>	Embedding Layer	7.34	10.51	34.9
<b>Baseline [34]</b>	ELMo	7.87	8.02	32.5
<b>DGN (cat)</b>	ELMo	7.26	9.87	35.1
<b>DGN (attn)</b>	ELMo	7.01	10.66	35.4
<b>Baseline [28]</b>	BERT	7.52	9.29	34.8
<b>DGN (cat)</b>	BERT	6.78	10.76	36.0
<b>DGN (attn)</b>	BERT	<b>6.59</b>	11.83	36.6
<b>SGN [10]</b>	BERT	6.67	<b>12.75</b>	<b>36.9</b>

loss functions, i.e. the global structure prediction loss  $\mathcal{L}_{pre}$ , sub-generator output loss  $\mathcal{L}_{gen}$  and position classification loss  $\mathcal{L}_{pos}$ , our training loss can be formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{pre} + \lambda_2 \mathcal{L}_{gen} + \lambda_3 \mathcal{L}_{pos}, \quad (13)$$

The Transformer model [12] is auto-regressive, which utilizes the previously generated tokens as additional input while generating the next [12]. Therefore, during inference time, we first feed the model with the [START] token instead of the whole target instruction captions, and then the model will output the following tokens incrementally. We run the global structure prediction component first. According to the predicted sub-generator sequence, we utilize the chosen generator for each recipe phase.

## 4. Experiments

### 4.1. Dataset and evaluation metrics

**Dataset.** We use the Recipe1M [2,4] provided official split: 252,547, 54,255 and 54,506 recipes for training, validation and test respectively. These recipes are scraped from cooking websites, and each of them contains the food image, a list of ingredients and the cooking instructions. Since Recipe1M data is uploaded by users, there have large variance and noises across the food images and recipes.

**Evaluation Metrics.** We totally adopt three different metrics for evaluation, i.e. perplexity, BLEU [30], ROUGE [31]. The prior work [4] only used perplexity for evaluation, which measures how well the probability distribution of learned words matches that of the input instructions. BLEU scores are based on an average of unigram, bigram, trigram and 4-gram precision. Here we take the averaged BLEU scores for the cooking recipes evaluation. Averaged BLEU scores not only reflect the n-gram scores, but also contain Brevity Penalty (BP), showing the impact of the generation length. However, BLEU fails to consider sentence structures [32]. In other words, BLEU cannot evaluate the performance of our global structure prediction component. ROUGE is a modification of BLEU that focuses on precision rather than recall, i.e. it looks at how many n-grams of the reference text show up in the outputs, rather than the reverse. Therefore, ROUGE can reflect the influence of the proposed global structure prediction component, which is discussed in Section 4.5.

### 4.2. Implementation details

We utilize ResNet-50 [26] which is pretrained on ImageNet [27] as the image encoder, which takes image size of  $224 \times 224$  as input. The ingredient encoder is BERT [28], short for Bidirectional Encoder Representations from Transformers, which is a pretrained language model implemented by [29] and is one of the state-of-the-art NLP models. As the prior work setting [4], we adopt the last convolutional layer of ResNet-50, whose output dimension is 512, as the feature representations. The output embedding of BERT model will be mapped to the dimension of 512 as well. For the cooking instruction generators, different sub-generators will share 12 transformer blocks, and each of them has additional independent 4 transformer blocks with 8 multi-head attention heads. To align with [4] and achieve a fair comparison, we generation instruction of maximum 150 words. In all the experiments, we use greedy search for recipe generation.

During the training phase, since we use the teacher-forcing training strategy, we integrate the target cooking instruction into the training process. Specifically, we use the ground truth token  $t_{i-1}$  to be the model input and generate the recipe token  $y_i$ . During the inference phase, we take the previous generated token  $y_{i-1}$  as input to recursively produce the cooking recipes. It is notable that the DGN framework is not required to know ground truth phase number during testing phase. In the first step, we adopt our trained global structure prediction component, to predict the selected sub-generators as well as their order. Then in the second step, based on the predicted phase information, we further use the selected sub-generators to generate each phase of the cooking recipes.

Regarding the phase number setting of each cooking instruction, we experiment with different numbers and observe that splitting per instruction into up-to three phases has the best performance. Since the cooking step numbers range from 2 to 19, suppose that if we split too many phases for each recipe, one phase may only contain one step, which will fail to obtain the global structure information. Therefore, we assume per instruction has at most three phases.

In all the experiments, we fix the weights of the image encoder for faster training, and instead of using the predicted ingredients as conditional generator inputs [4], we take the ground truth ingredients and images as input for a fair comparison. We set  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$  in Eq. (13) to be 1, 1 and 0.1 respectively, which is based on empirical observations on validation set. The model is optimized with Adam, and the initial learning rate is set as 0.001, with 0.99 decay per epoch. The model is trained for about 25 epochs to be converged. We implement the proposed methods with PyTorch [33].

### 4.3. Baselines

We use [4] as the baseline for this paper, it is worth noting that we use the same input as our proposed model in the implementation of [4], where we also adopt food images and ground truth ingredients as inputs for fair comparisons.

Specifically, Salvador et al. [4] generate the whole cooking instructions from the cooked food images through 16 transformer blocks. In contrast, our proposed DGN extends an additional branch for the text generation process, which predicts the structures of the recipes first and then utilizes the chosen sub-generators for each phase generation. In other words, DGN can be applied to different backbone networks. We compare the difference between baseline models and the proposed DGN in Fig. 3.

To fully demonstrate the efficacy of DGN, we experiment with three different ingredient encoders to act as baseline results. The first one comes from the prior work [4], where they adopted one word embedding layer to encode the ingredients. We need to train

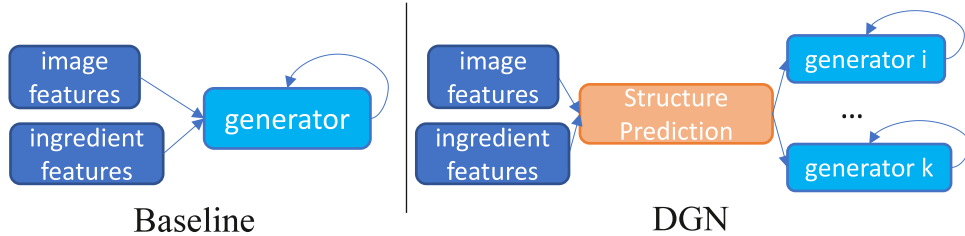


Fig. 3. The comparison of the baseline model and our proposed DGN. DGN can be applied to different backbone networks.

it from scratch. Moreover, we also adopt the pretrained models as the ingredient encoder, where we experiment with ELMo [34] and BERT [28]. ELMo and BERT are LSTM-based and transformer-based pretrained models respectively. We finetune the pretrained models during training the DGN frameworks. Note that the aforementioned three baseline models all use ResNet-50 as the image encoder, they only differ in the ingredient encoders.

#### 4.4. Main results

We show our main results of generating cooking instructions in Table 1, which are evaluated across three language metrics: perplexity, BLEU [30] and ROUGE-L [31]. Generally, models with and without DGN have an obvious performance gap. Simply using one word embedding layer for ingredient encoder performs poorly, achieving the lowest scores across all the metrics. When we replace the embedding layer with state-of-the-art pretrained language model, ELMo or BERT, the performance reasonably gets better, which highlights the significance of the pretrained model.

We then incrementally add the DGN branch to three different backbone networks. To be specific, we experiment with two ways to construct the phase-aware features  $\mathbf{r}$ , i.e. **DGN (cat)**, where  $\mathbf{r}$  is formed by the concatenation of the four representations, and **DGN (attn)**, in which we construct image and ingredient features with attention mechanism, then concatenate them together to be  $\mathbf{r}$ . First, we add **DGN (cat)** to the baseline models, surprisingly this approach can achieve more than 2 BLEU scores better than the baseline model with the word embedding layer and 1 BLEU score over state-of-the-art language model BERT, which indicates our proposed DGN is very promising and can extend to various general models. We further adopt **DGN (attn)** for recipe generation evaluation, the performance continually gets better, illustrating the usefulness of enhancing the inputs of generators. It is observed that BERT gives the best results across all the metrics among the encoder settings.

Moreover, we compare our results with the state-of-the-art recipe generation model SGN [10], where they unsupervisedly learn the sentence-level tree structure of the cooking instructions, and integrate the tree representations into the original image and ingredient representations. Although SGN [10] and our proposed DGN both use the structure information, SGN makes use of the recipe structure information in the encoder part, while our proposed DGN explicitly uses the predicted structures for the decoding process. We can see that the proposed DGN framework achieves comparable results as SGN, and even outperforms SGN at the evaluation of perplexity. To summarize, our proposed DGN framework outperforms various baseline methods across all metrics consistently, and achieves the state-of-the-art performance.

#### 4.5. Ablation studies

**The ablative influence of image and ingredient as input.** To suggest the necessity of using both image and ingredient as input, we train the model with different inputs separately. We show the

Table 2

The ablative influence of image and ingredient as input. The model is evaluated by perplexity (lower is better), BLEU (higher is better) and ROUGE-L (higher is better).

Input	Perplexity	BLEU	ROUGE-L
<b>Only Image</b>	8.16	3.72	31.0
<b>Only Ingredient</b>	7.62	5.74	32.1
<b>Image and Ingredient</b>	<b>7.52</b>	<b>9.29</b>	<b>34.8</b>

Table 3

The impact of sub-generator category number  $N$ . The model is evaluated by perplexity (lower is better), BLEU (higher is better) and ROUGE-L (higher is better).

$N$	Methods	Perplexity	BLEU	ROUGE-L
<b>1</b>	<b>BERT</b>	7.52	9.29	34.8
<b>1</b>	<b>BERT+DGN</b>	6.98	10.98	35.8
<b>3</b>	<b>BERT+DGN</b>	<b>6.59</b>	<b>11.83</b>	<b>36.6</b>
<b>5</b>	<b>BERT+DGN</b>	6.95	11.15	36.0

ablation studies in Table 2, where we use a transformer for generation, instead of DGN. It can be observed that ingredient information helps more on the recipe generation, since ingredients can be directly reflected in the recipes. The model with image and ingredient as input has better performance than that of single modality input.

**The impact of sub-generator category number  $N$ .** After we get the representation of each instruction phase, we adopt k-means clustering to obtain the phase labels, which indicate the sub-generator selections. Then these labels are used for the global structure prediction component training. We show the experiment results in Table 3, where the first row shows the experiment results of BERT baseline model, the last four rows are all implemented by BERT + DGN (attn). When  $N = 1$ , we compare the results of the first and second row, the first row uses the concatenated representations of image and ingredient features, while the second row takes the enhanced phase-aware features  $\mathbf{r}$  as input, indicating the efficacy of the phase-aware features. Besides, the model with  $N = 1$  has inferior performance compared with model with  $N = 3$ , illustrating the single generator struggles to fit data from different phases. When  $N = 5$ , the model gets similar evaluation results to  $N = 1$ . That model with  $N = 5$  has poorer performance than model with  $N = 3$  may because the model does not have enough data for training, due to the more splits of the training data. Therefore, we set the hyper-parameter  $N$  to be 3.

**The impacts of DGN on the average length and vocabulary size of generated recipes.** In order to further demonstrate the effectiveness of the proposed DGN from other aspects, we perform some language analysis based on the generated outputs in the Table 4. Our DGN approach generates text of the closet average length as ground truth recipes, which are crawled from websites and written by humans. While the models without DGN generate relatively short cooking instructions, which provides the evidence for our assumptions before: using one single generator will result in some cooking details are missing. We also show some qualitative results in Fig. 4. To evaluate the diversity of the recipes, we






	Ground Truth	BERT	DGN (BERT)
 Pizza	Heat oven to 450 degrees f. Cook and stir vegetables in 1 tbsp. Dressing in skillet on medium heat 3 min. Place pizza crust on baking sheet sprayed with cooking spray; brush with remaining dressing. Top with cheese, pepperoni, vegetables and olives. Bake 10 to 12 min. Or until cheese is melted and edge of crust is golden brown. Sprinkle with thyme.	Heat oven to 450 degrees f. Toss vegetables with dressing; place in single layer on baking sheet. Bake 10 min. Or until vegetables are crisp-tender, stirring after 10 min. Top with cheese, pepperoni and olives.	<sup>1</sup> Heat oven to 450 degrees f. Cook and stir vegetables in 1/2 cup dressing in large skillet on medium heat 5 min. Or until vegetables are tender. spoon onto pizza shell; top with cheese. <sup>2</sup> Place pizza crust on baking sheet sprayed with cooking spray; spread with sauce. Top with cheese and vegetables. Bake 10 to 12 min. Or until cheese is melted and edge of crust is golden brown. <sup>3</sup> Sprinkle with pepperoni. Bake 2 to 3 min. Or until cheese is melted. Sprinkle with basil. Serve.
 Beef Stew	Combine flour, salt and pepper in a bowl. Toss beef cubes in the flour mixture to coat. In a large dutch oven, cook the butter until just starting to brown. Add the meat and onions, and cook, stirring occasionally, until the meat is browned on all sides. Add the bay leaves and allspice. Boil the water in a separate pan, then pour over the meat. Simmer, covered, for 1 1/2 hours, or until the meat is tender. Check the water level occasionally, and add more if needed. When meat is tender, remove to a serving dish.	Sprinkle meat with salt and pepper. Melt butter in dutch oven or heavy pot. Add meat and brown well on all sides. Remove meat and set aside. Add onions and allspice to dutch oven. Cover and simmer for 1 1/2 hours. Add water and bay leaves. Cover and simmer for 1 1/2 hours. Add meat and simmer for 1 1/2 hours more.	<sup>1</sup> Season the beef with salt and pepper. Melt the butter in a large skillet over medium heat. Add the beef and brown on all sides. Remove the beef from the skillet and set aside. <sup>2</sup> Add onions and allspice. Cook, stir occasionally, until onions are soft. Add bay leaves and water. Cover and simmer for 1 1/2 hours. Add more water if necessary. <sup>3</sup> Cover and cook on low for 8-10 hours. Skim off fat. Add the remaining ingredients and cook for 30 minutes more. Serve with rice or noodles. Enjoy!
 Spicy Red Beans	Place beans into a pressure cooker and fill with enough water to cover the beans by 1 inch. Bring to a boil, and boil for 2 minutes. Remove from the heat, drain and cool for 1 hour. Add water to the beans again, filling just enough to cover the beans by 1/2 inch. Season with salt, fennel seed, cayenne pepper and mixed peppercorns. Put the lid on and set the pressure to high, or 15 pounds depending on your indicator. Bring to a boil and cook for 35 minutes. Let the pressure lower naturally before releasing the lid, according to manufacturer's instructions (about 15 minutes). Serve hot.	Soak beans overnight in plenty of water. Drain and rinse beans. Place beans in a large pot with enough water to cover by 2 inches. Drain and rinse beans. Place beans in a large pot with salt, fennel seed, cayenne pepper and black pepper. Cover and simmer for 1 1/2 hours.	<sup>1</sup> Soak beans overnight in cold water. Drain and discard soaking water. Place beans in a large saucepan and cover with water. Bring to a boil, reduce heat, and simmer for 1 hour. <sup>2</sup> Add salt, fennel seeds, peppercorns. Grind to a fine powder. Add to beans and mix well. Cover and refrigerate overnight. Rewarm before using. <sup>3</sup> Cover and let stand for 1 hour. Drain and store in a cool, dry place. Makes about 6 cups.
 Apple Crumble	Peel, core and slice apples and place into medium sized mixing bowl. Add sugar, vanilla and cinnamon (or all spice) to apples and mix until apples are coated. Place apple mixture into 8x8 glass baking dish. Soften butter, stir until creamy and place in another bowl. Add flour and brown sugar until mixture looks like little peas. Cover apples with flour mixture making sure apples are coated. Place uncovered baking dish onto lower oven rack. Bake at 300 degrees for 50-55 minutes (or until corners are dark brown and crispy). Serve warm with vanilla ice cream or espresso.	Mix all ingredients together. Store in an airtight container.	<sup>1</sup> Preheat oven to 350 degrees. Peel and core apples. Cut into 1/2 inch slices. Place in a 9x13 pan. Sprinkle with sugar and cinnamon. Sprinkle the mixture over the apples. <sup>2</sup> In a small bowl, mix together the melted butter, brown sugar, and flour. Pour this over the apples. Bake for 30 minutes. Serve warm with vanilla ice cream. Enjoy! <sup>3</sup> Note: you can use any type of apples you like. I use a combination of apples, but I think it is not necessary.
 Peach and Nut Cake	Preheat oven to 350f. Grease and butter a 9x13-inch glass pan. In a food processor mix the dough ingredients. Pour into prepared pan. Mix sugar and cinnamon together. Pour half the sugar on batter. Place in 3 rows 10 slices of peaches each row. Eat 2 slices for yourself. Sprinkle the chopped nut over the peaches then top with the sugar and drizzle the butter over the sugar. Bake for 30-35 minutes.	Preheat oven to 350 degrees. Place chicken in a 9x13 baking dish. Mix soup, milk, and pepper in a bowl. Pour over chicken. Bake for 1 hour.	<sup>1</sup> Cream butter and sugar together. Sift flour, baking powder, milk, egg and salt together. Add to butter mixture alternately with the milk. Fold in pecans. <sup>2</sup> Pour batter into a greased 9x13 pan. Pour into greased 9x13 pan. Sprinkle with 1/2 cup sugar and cinnamon. Pour peaches over batter. Sprinkle with remaining 1/2 cup sugar. <sup>3</sup> Bake at 375 for 30-35 minutes. Serve warm with ice cream or whipped cream. Enjoy!

Fig. 4. Analysis of generated recipes by different models. We show the generated results conditioned on three different food images, namely pizza, beef stew spicy red beans, apple crumble and peach and nut cake. The left column shows the conditioned food images, and the right three columns show the ground truth cooking instructions, baseline BERT generations and DGN generated recipes. Words with color background represent the matching parts between raw recipes and the generated recipes. In the DGN generations, we state the recipe phases with the superscript numbers.

Table 4

The impacts of DGN on the average length and vocabulary size of generated recipes. The results demonstrate that the proposed DGN increases the average length and diversity of generated cooking instructions.

Methods	Average Length	Vocab Size
Baseline [4]	69.9	3657
Baseline [28]	66.9	4521
DGN (Baseline [4])	103.1	4836
DGN (Baseline [28])	105.6	6573
Ground Truth	116.5	33,110

compute the vocabulary sizes of the generations and the ground truth, which indicates the number of unique words that appear in the text. According to the results, DGN (BERT) is actually the most diverse method apart from the ground truth. But there still remain huge gaps between the diversity of generated text and human-written text.

**The relation between the phase prediction procedure and the generated results.** Global structure prediction component is the first and basic part of our proposed DGN model, which outputs the sub-generator selections and their orders for subsequent gen-

erations. To demonstrate the relation between the phase prediction and the generated results, we experiment with DGN models trained with and without the global structure prediction loss  $\mathcal{L}_{pre}$  or the position classification loss  $\mathcal{L}_{pos}$ . It is observed that the prediction accuracy on the sub-generator selections of model trained with  $\mathcal{L}_{pre}$  is about 72%, while the accuracy of model trained without  $\mathcal{L}_{pre}$  is only 11%. Training the DGN model without  $\mathcal{L}_{pre}$  means we simply output random sub-generator selections for recipe generation. We observe clear performance drop when we do not use our phase classification results for recipe generation. It indicates the necessity and usefulness of our proposed global structure prediction component. Moreover, we also give the ablation study on our proposed  $\mathcal{L}_{pos}$ . The results demonstrate that the proposed  $\mathcal{L}_{pos}$  enables the phase-aware features to contain certain phase position information, and further improve the overall evaluation results of the generated recipes. It also validates the importance of leveraging structure for food recipe generation (Table 5).

#### 4.6. Qualitative results

We present some qualitative results from our proposed model and the ground truth cooking instructions for comparison in Fig. 4.



**Table 5**

We show the ablative influence of the global structure prediction loss  $\mathcal{L}_{pre}$  or the position classification loss  $\mathcal{L}_{pos}$  to demonstrate the relation between the phase prediction procedure and the generation results. The model is evaluated by perplexity (lower is better), BLEU (higher is better) and ROUGE-L (higher is better).

Methods	Perplexity	BLEU	ROUGE-L
<b>full DGN</b>	<b>6.59</b>	<b>11.83</b>	<b>36.6</b>
- w/o $\mathcal{L}_{pre}$	7.65	9.20	34.6
- w/o $\mathcal{L}_{pos}$	7.19	9.85	35.8

In the left column, we show the conditional food images, which come from *pizza*, *beef stew*, *spicy red beans*, *apple crumble* and *peach and nut cake* respectively. And in the right three columns, we list the true recipes, the generated recipes of BERT and that of our proposed model DGN, which uses the attended features. We indicate the recipe phases with the red number in DGN generations, and words with yellow background suggest the matching parts between raw recipes and the generated recipes.

The obvious properties of DGN generations include its average length and its ability to capture rich cooking details. First of all, we can see that DGN generates longer recipe outputs than BERT, which has a similar length as true recipes. Besides, it is observed that the phase orders predicted by the global structure prediction component make sense in the shown cases: the first instruction phase gives some instructions on pre-processing the ingredients, the middle instruction phase tends to describe the details about the main dish cooking, and the last phase often contains some concluding work of cooking.

Generally, it can be seen that DGN generates more matching cooking instruction steps with the ground truth recipes than BERT. When we go into the details, the DGN generated instructions include the ingredients used in the true recipes. Specifically, in the top row, the generated text covers the ingredients of *pepperoni*, *cheese*, *vegetables* and etc. Compared with the BERT outputs, DGN generate similar sentences at the beginning. However, DGN provides more details, e.g. in the instruction generation of *beef stew*, both BERT and DGN output the sentence of “Add onions and all-spice.”, while DGN further generate some tips: “Cook, stir occasionally, until onions are soft.”.

It is also worth noting that some of the predicted numbers are not precise enough, like in the third generated phase of *Beef Stew*, the generation output turns out to be “cook ... for 8–10 hours”, which is not aligned with common sense.

## 5. Conclusion

In this paper, we have proposed to make the generated cooking instructions more structured and complete, i.e. to decompose the recipe generation process. In particular, we present a novel framework DGN for recipe generation that leverages the compositional structures of cooking instructions. Specifically, we first predict the global structures of the instructions based on the conditional food images and ingredients, and determine the sub-generator selections and their orders. Then we construct novel phase-aware features for the input of chosen sub-generators and adopt them to produce the instruction phases, which are concatenated together to obtain the whole cooking instructions. Experimentally, we have demonstrated the advantages of our approach over traditional methods that use one single decoder to generate the long sentences, i.e. the proposed DGN can increase the diversity and average length of generated recipes. We conduct extensive experiments with ablation studies, and achieved state-of-the-art recipe generation results across different metrics in Recipe1M dataset.

Though the proposed DGN achieves promising results in Recipe1M dataset, there exists some limitations with DGN. Specifically, some of model hyper-parameters are set based on empirical experiments, such as phase number and sub-generator number selections, since it is challenging to automatically segment recipes into several phases according to context and we do not have strong supervisions.

Our proposed DGN can give structured representations for lengthy paragraphs. In this paper, we show its effectiveness in food domain, and it can also be adapted to other fields that require to generate long sentences. In the future work, we will try to extend it to some other fields.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This research is supported, in part, by the National Research Foundation (NRF), Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-003) and under its NRF Investigatorship Programme (NRF1 Award No. NRF-NRF105-2019-0002). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore. This research is also supported, in part, by the Singapore Ministry of Health under its National Innovation Challenge on Active and Confident Ageing (NIC Project No. MOH/NIC/COG04/2017 and MOH/NIC/HAIG03/2017), and the MOE Tier-1 research grants: RG28/18 (S) and RG22/19 (S).

## References

- [1] L. Bossard, M. Guillaumin, L. Van Gool, Food-101—mining discriminative components with random forests, in: European Conference on Computer Vision, Springer, 2014, pp. 446–461.
- [2] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, A. Torralba, Learning cross-modal embeddings for cooking recipes and food images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3020–3028.
- [3] H. Wang, D. Sahoo, C. Liu, E.-p. Lim, S.C.H. Hoi, Learning cross-modal embeddings with adversarial networks for cooking recipes and food images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11572–11581.
- [4] A. Salvador, M. Drozdal, X. Giro-i Nieto, A. Romero, Inverse cooking: recipe generation from food images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10453–10462.
- [5] L. Zhou, C. Xu, J.J. Corso, Towards automatic learning of procedures from web instructional videos, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
- [6] A. Bosselut, O. Levy, A. Holtzman, C. Ennis, D. Fox, Y. Choi, Simulating action dynamics with neural process networks, in: Proceedings of the 6th International Conference for Learning Representations (ICLR), 2018.
- [7] W. Min, S. Jiang, L. Liu, Y. Rui, R. Jain, A survey on food computing, ACM Comput. Surv. (CSUR) 52 (5) (2019) 1–36.
- [8] Y. Matsuda, H. Hoashi, K. Yanai, Recognition of multiple-food images by detecting candidate regions, in: Multimedia and Expo (ICME), 2012 IEEE International Conference on, IEEE, 2012, pp. 25–30.
- [9] K. Chandu, E. Nyberg, A.W. Black, Storyboarding of recipes: grounded contextual generation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6040–6046.
- [10] H. Wang, G. Lin, S.C.H. Hoi, C. Miao, Structure-aware generation network for recipe generation from images, in: European Conference on Computer Vision, Springer, 2020, pp. 359–374.
- [11] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.
- [12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
- [13] X. Xiao, L. Wang, K. Ding, S. Xiang, C. Pan, Dense semantic embedding network for image captioning, Pattern Recognit. 90 (2019) 285–296.

- [14] J. Wang, W. Wang, L. Wang, Z. Wang, D.D. Feng, T. Tan, Learning visual relationship and context-aware attention for image captioning, *Pattern Recognit.* 98 (2020) 107075.
- [15] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, J. Tan, Cross-modal knowledge reasoning for knowledge-based visual question answering, *Pattern Recognit.* 108 (2020) 107563.
- [16] Z. Bai, Y. Li, M. Woźniak, M. Zhou, D. Li, DecomVQANet: decomposing visual question answering deep network via tensor decomposition and regression, *Pattern Recognit.* 110 (2021) 107538.
- [17] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.
- [18] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., Oscar: object-semantics aligned pre-training for vision-language tasks, in: *European Conference on Computer Vision*, Springer, 2020, pp. 121–137.
- [19] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, VinVL: revisiting visual representations in vision-language models, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5579–5588.
- [20] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Neural module networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 39–48.
- [21] X. Yang, H. Zhang, J. Cai, Learning to collocate neural modules for image captioning, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4250–4260.
- [22] D.A. Hudson, C.D. Manning, Compositional attention networks for machine reasoning, in: *International Conference on Learning Representations*, 2018.
- [23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: *European Conference on Computer Vision*, Springer, 2014, pp. 740–755.
- [24] J. Krause, J. Johnson, R. Krishna, L. Fei-Fei, A hierarchical approach for generating descriptive image paragraphs, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 317–325.
- [25] M. Honnibal, I. Montani, spaCy2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.
- [26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Ieee, 2009, pp. 248–255.
- [28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, *NAACL-HLT (1)*, 2019.
- [29] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al., Transformers: state-of-the-art natural language processing, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.
- [30] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2002, pp. 311–318.
- [31] C.-Y. Lin, ROUGE: a package for automatic evaluation of summaries, in: *Text Summarization Branches Out*, 2004, pp. 74–81.
- [32] C. Callison-Burch, M. Osborne, P. Koehn, Re-evaluation the role of Bleu in machine translation research, in: *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.
- [33] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch(2017).
- [34] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, *arXiv preprint arXiv:1802.05365* (2018).

**Hao Wang** is a PhD student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include multimodal analysis and computer vision.

**Guosheng Lin** is currently an Assistant Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include computer vision and machine learning.

**Steven C. H. Hoi** is currently the Managing Director of Salesforce Research Asia, and an Associate Professor (on leave) of the School of Information Systems, Singapore Management University, Singapore. Prior to joining SMU, he was an Associate Professor with Nanyang Technological University, Singapore. He received his Bachelor degree from Tsinghua University, PR China, in 2002, and his PhD degree in computer science and engineering from The Chinese University of Hong Kong, in 2006. His research interests are machine learning and data mining and their applications to multimedia information retrieval (image and video retrieval), social media and web mining, and computational finance, etc., and he has published over 150 refereed papers in top conferences and journals in these related areas. He has served as the Editor-in-Chief for *Neurocomputing Journal*, general co-chair for ACM SIGMM-Workshops on Social Media (WSM'09, WSM'10, WSM'11), program co-chair for the fourth Asian Conference on Machine Learning (ACML'12), book editor for "Social Media Modeling and Computing", guest editor for ACM Transactions on Intelligent Systems and Technology (ACM TIST), technical PC member for many international conferences, and external reviewer for many top journals and worldwide funding agencies, including NSF in US and RGC in Hong Kong. He is an IEEE Fellow and ACM Distinguished Member.

**Chunyan Miao** is the chair of School of Computer Science and Engineering in Nanyang Technological University (NTU), Singapore. Dr. Miao is a Full Professor in the Division of Information Systems and Director of the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), School of Computer Engineering, Nanyang Technological University (NTU), Singapore. She received her PhD degree from NTU and was a Postdoctoral Fellow/Instructor in the School of Computing, Simon Fraser University (SFU), Canada. She visited Harvard and MIT, USA, as a Tan Chin Tuan Fellow, collaborating on a large NSF funded research program in social networks and virtual worlds. She has been an Adjunct Associate Professor/Associate Professor/Founding Faculty member with the Center for Digital Media which is jointly managed by The University of British Columbia (UBC) and SFU. Her current research is focused on human-centered computational/ artificial intelligence and interactive media for the young and the elderly. Since 2003, she has successfully led several national research projects with a total funding of about 10 Million dollars from both government agencies and industry, including NRF, MOE, ASTAR, Microsoft Research and HP USA. She is the Editor-in-Chief of the *International Journal of Information Technology* published by the Singapore Computer Society.