Transformer-based Cross-Modal Recipe Embeddings with Large Batch Training

Jing Yang[1][0000−0002−2103−6825], Junwen Chen[1][0000−0003−1355−4128], and Keiji Yanai[1][0000−0002−0431−183X]

The University of Electro-Communications, Tokyo, Japan
`{yang-j,chen-j,yanai}@mm.inf.uec.ac.jp`

Abstract. Cross-modal recipe retrieval aims to exploit the relationships and accomplish mutual retrieval between recipe images and texts, which is clear for human but arduous to formulate. Although many previous works endeavored to solve this problem, most works did not efficiently exploit the cross-modal information among recipe data. In this paper, we present a frustratingly straightforward cross-modal recipe retrieval framework, Transformer-based Network for Large Batch Training (TNLBT) achieving high performance on both recipe retrieval and image generation tasks, which is designed to efficiently exploit the rich cross-modal information. In our proposed framework, Transformer-based encoders are applied for both image and text encoding for cross-modal embedding learning. We also adopt several loss functions like self-supervised learning loss on recipe text to encourage the model to further promote the cross-modal embedding learning. Since contrastive learning could benefit from a larger batch size according to the recent literature on self-supervised learning, we adopt a large batch size during training and have validated its effectiveness. The experimental results showed that TNLBT significantly outperformed the current state-of-the-art frameworks in both cross-modal recipe retrieval and image generation tasks on the benchmark Recipe1M by a huge margin. We also found that CLIP-ViT performs better than ViT-B as the image encoder backbone. This is the first work which confirmed the effectiveness of large batch training on cross-modal recipe embedding learning.

Keywords: cross-modal recipe retrieval, transformer, vision transformer, image generation

1 INTRODUCTION

Cross-modal recipe retrieval investigates the relation between recipe texts and recipe images to enable mutual retrieval between them. Recipe1M dataset [20, 16] is frequently used as a benchmark to evaluate the performance of cross-modal recipe retrieval frameworks. One challenge of the recipe retrieval task is that food images usually contain many non-food parts like plates and different backgrounds as noise. The recipe texts are perplexing, making them difficult to encode, since there are three components in the recipe text: title, ingredients, and instructions, the last two of which are long and structured texts. Most existing works focus just on recipe or image embedding learning,

though we believe that a framework that focuses on both recipe and image embedding learning is necessary.

As one of the early works Salvador *et al.* [20] proposed Joint Embedding (JE), projecting embeddings from both modalities into a common shared space and minimizing the cosine similarity between them to enable cross-modal recipe retrieval. GAN [7] was introduced in some recent works [27, 24, 22] for food image synthesis from embeddings to obtain more reliable cross-modal embeddings. By confirming if the same or similar embeddings can be extracted from the food images generated from cross-modal embeddings, the retrieval performance were improved. Furthermore, modality adversarial loss was introduced in ACME [24] which can narrow the modality gap between image and text embedding. Inspired by the success of GAN-based image synthesis in recipe retrieval, we also adopt GAN-based image synthesis in this work. However, regarding recipe text embedding, LSTM [8] was simply applied for encoding recipe texts in these works [27, 24, 22], which sometimes limits leveraging the information in long and structured recipe texts. In order to address this issue, some other methods [6, 26, 19] focused on improving recipe embedding learning. Authors of MCEN [6] applied attention mechanism, and Zan *et al.* [26] applied BERT [4] for recipe encoding. With the trend of Transformer [23] in natural language processing, X-MRS [9] and H-T [19] were recently proposed to adopt this technique in recipe retrieval. Authors of H-T in particular introduced a simple but effective framework with a Transformer-based structure recipe encoder and self-supervised learning, allowing the model to explore complementary information among recipe texts. Inspired by H-T, we adopt a Transformer-based recipe encoder and self-supervised learning for recipe embedding learning.

In this paper, we propose a frustratingly straightforward Transformer-based framework that uses all the current state-of-the-art techniques. That is, we use a hierarchical transformer architecture recipe encoder for recipe embedding learning, and an adversarial network to investigate the complementary information between recipe and image embedding. In the experiments, we adopt ViT [5] and CLIP-ViT [18] as the image encoder backbone to validate the effectiveness the proposed framework. Since contrastive learning benefits from larger batch sizes according to the recent work on self-supervised learning [2] and self-supervised learning and triplet losses are used in the proposed framework, we adopt large batch training in our experiments. Furthermore, we conducted extensive experiments and ablation studies to further validate the effectiveness of our proposed framework. We discovered that large batch training was surprisingly effective for Transformer-based cross-modal recipe embedding. The results showed that our proposed framework outperformed the state-of-the-arts both in recipe retrieval (medR 1.0, R1 56.5 in 10k test set size) and image generation tasks (FID score 16.5) by a large margin. More specifically, we summarize the contributions of this work as follows:

1. We proposed a frustratingly straightforward Transformer-based cross-modal recipe retrieval framework, Transformer-based Network for Large Batch Training (TNLBT), which achieves the state-of-the-art performance on both recipe retrieval and image generation tasks.
2. We conducted experiments with large batch inspired by the effectiveness of large batch training on contrastive learning [2].

3. Through the comprehensive experiments, we confirmed that the proposed framework outperforms the current state-of-the-arts with a large margin, especially in the case of a large batch size, 768. This is the first work which confirmed the effectiveness of large batch training on cross-modal recipe retrieval as far as we know.

2 RELATED WORK

2.1 Cross-Modal Recipe Retrieval

Cross-modal retrieval aims to enable mutual retrieval between two different modalities, often image and text. The common idea of cross-modal retrieval is to embed features from two different modalities into a shared common space while keeping the distribution of corresponding embedded features close to enable mutual retrieval. A substantial issue in this task is how to narrow the gap between the various modalities [11].

Salvador *et al.*, who was the first to propose the cross-modal recipe retrieval task and the Recipe1M dataset, proposed joint embedding [20] to enable cross-modal recipe retrieval. This method was modified in AdaMine [1] by using a triplet loss [21] to improve the retrieval accuracy. In order to further exploit the information in the Recipe1M dataset, some state-of-the-art techniques like Transformer [23] and Generative Adversarial Networks (GAN) [7] are used in recipe retrieval. GAN, enables image generation conditioned on recipe embedding in recipe retrieval framework. GAN was introduced in several previous works [27, 24, 22], which improves retrieval accuracy while enabling image generation conditioned on recipe text. The issues of these previous works are, the retrieval accuracy and quality of generated images are very limited since they only adopted simple LSTM [8] for text embedding learning. However, in our proposed method we obtain generated images of better quality using Transformer.

Rather than incorporating complex networks like GAN, some research [6, 26, 19] focus on recipe embedding learning to enhance the performance of retrieval tasks. Authors of MCEN[6] introduced cross-modal attention and consistency and Zan *et al.* [26] introduced BERT [4] as a recipe encoder to enable cross-modal retrieval. Authors of X-MRS [9] introduced a Transformer [23] encoder to gain recipe embedding, further proposed the use of imperfect multilingual translations, and achieved state-of-the-art performances on retrieval tasks. Salvador *et al.* proposed a simply but effective framework H-T [19], to facilitate the power of Transformer. However, authors of these works [6, 26, 19] just simply applied bi-directional triplet loss on image and recipe features without image synthesis and reconstruction, which somewhat limited the cross-modal information learning between recipe and image features. In order to address this issue, we introduce GAN-based architecture to enhance the cross-modal embedding learning in our architecture in addition to adopting Transformer-based recipe encoders and self-supervised learning on recipe-only samples in the Recipe1M.

2.2 Food Image Synthesis

GAN [7] has been introduced and proven its effectiveness for improving recipe retrieval performance in some recent works [27, 24, 22, 9]. Authors of R2GAN [27], ACME [24]
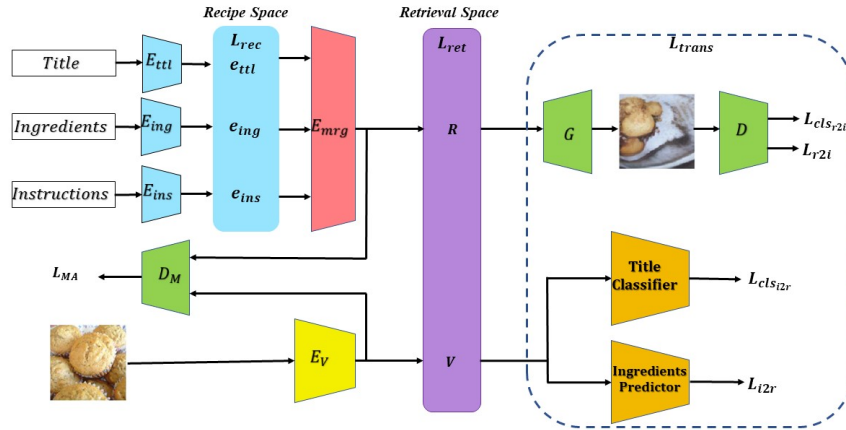
Fig. 1: **The architecture of proposed network TNLBT.** Training process is controlled by the four loss functions L_{rec}, L_{ret}, L_{MA}, L_{trans} .

and X-MRS [9] applied text-conditioned image synthesis which generates recipe images from recipe embeddings, while authors of RDE-GAN [22] proposed to disentangle image features into dish shape features, which contained only non-recipe information, and recipe embeddings and to integrate both of them to generate recipe images. Similar to ACME and RDE-GAN, we also leverage GAN in our proposed framework to improve retrieval accuracy while enabling image generation conditioned on recipe text. The biggest difference between our work and the previous works is, we applied Transformer-based encoders to further improve the quality of generated images compared to the previous works [24, 22].

3 Method

3.1 Overview

We propose a frustratingly straightforward Transformer-based framework TNLBT for cross-modal recipe retrieval which has the hierarchical Transformers for text encoding, and Vision Transformer (ViT) [5] as an image encoder. Furthermore, we propose to adopt large batch training which is confirmed to be beneficial for contrastive training and Transformer-based networks. Figure 1 shows the architecture of the proposed framework. Similar to H-T [19], our proposed framework applies hierarchical Transformer encoders and a triplet loss L_{rec} with self-supervised learning to explore complementary information in the recipe text. ViT-B [5] or CLIP-ViT [18] is used as the backbone of image encoder. The biggest difference between our work and H-T [19] is, we introduce a more sophisticated adversarial network to leverage the learned Transformer-based embeddings. Rather than just a bi-directional triplet loss is adopted

in H-T for the distance learning of image and recipe embeddings, we further introduce several loss functions to enhance the cross-modal embedding learning and enable image generation conditioned on recipe text. Inspired by the success of triplet loss [21], a triplet loss L_{ret} for distance learning in image-recipe retrieval is applied in the retrieval space. In order to mitigate the modality gap problem, we use the modality alignment loss function L_{MA} as in ACME [24]. Furthermore, we introduce image-to-recipe and recipe-to-image information recover, and a translation consistency loss function L_{trans} to keep modality-specific information. We expect that the embeddings learned in the above process correctly retains the original information and enables image generation from recipe embeddings. Finally the overall loss function can be formulated as follows:

$$L_{total} = \lambda_1 L_{rec} + \lambda_2 L_{MA} + \lambda_3 L_{trans} + L_{ret}, \tag{1}$$

where λ_1, λ_2, λ_3 are hyperparameters to control the loss balance.

3.2 Recipe Encoder and Self-Supervised Learning

The purpose of the recipe encoder, E_R (even though we adopt several encoders here, we note these as a whole recipe encoder E_R), is to encode recipe text, **t**, to recipe embeddings, **R**. In order to better explore and make use of the huge information among recipe text with components of title, ingredients, and instructions, we adopt the hierarchical Transformers as the recipe encoder similar to H-T [19]. Note that we introduced an adversarial network for image synthesis and modality alignment better to exploit the correlation in image-recipe pairs, while H-T simply applied a bi-directional triplet loss function to image-recipe pairs.

For the three components (title, ingredients, instructions) in the recipe texts, the proposed framework TNLBT encodes them separately using hierarchical Transformer encoders ($E_{ttl}, E_{ing}, E_{ins}$), with 2 layers and 4 attention heads, to obtain embeddings of title, ingredients and instructions ($e_{ttl}, e_{ing}, e_{ins}$) initially. Next, we apply self-supervised recipe loss L_{rec} as in H-T, on $e_{ttl}, e_{ing}, e_{ins}$ to explore the complementary information among them. The introduction of this loss allows us to use recipe-only image-recipe pairs to further leverage the complementary information among the three components of the recipe text. Finally, we apply a merging encoder E_{mrg} to project three components embeddings ($e_{ttl}, e_{ing}, e_{ins}$) to an unified recipe embedding, **R**.

3.3 Image Encoder

The purpose of the image encoder E_V is to encode recipe image **i** to image embedding **V**. We adopt the base size model of Vision Transformer (ViT-B) [5] pre-trained on ImageNet-21k [3] as the image encoder backbone. In order to better leverage the benefit of large batch training, we also adopt CLIP-ViT [18] as image encoder backbone where the proposed framework achieved superior performance on retrieval task.

3.4 Modality Alignment Loss

For the purpose of narrowing the gap between recipe and image embeddings, which often results in bad generalization or slow convergence, we applied modality alignment

loss L_{MA} as in ACME [24]. We adopt a discriminator D_M here to achieve this goal, which aligns the distribution of recipe embeddings $R = E_R(\mathbf{t})$ and image embeddings $V = E_V(\mathbf{i})$ during training. An adversarial loss is applied by controlling the recipe and image embeddings aligned to each other resulting in the discriminator D_M cannot distinguish the source of the given embeddings empirically. Following ACME, we also adopt WGAN-GP [10] here and the loss function is formulated as follows:

$$L_{MA} = \mathbb{E}_{\mathbf{i}\sim p(i)}[\log(1 - D_M(E_V(\mathbf{i})))] + \mathbb{E}_{\mathbf{t}\sim p(t)}[\log(1 - D_M(E_R(\mathbf{t})))] \quad (2)$$

3.5   Retrieval Loss

Following the success of triplet loss [21] in the recent works [1, 24, 22, 19], we adopt a triplet loss for distance learning. Here we obtain the recipe and image embeddings $R = E_R(\mathbf{t}), V = E_V(\mathbf{i})$. We obtain an anchor recipe embedding R_a and an anchor image embedding V_a, and obtain a negative sample with subscript n and a positive sample with subscript p to process this distance learning. The loss function is formulated as follows:

$$L_{ret} = \sum_V [d(V_a, R_p) - d(V_a, R_n) + \alpha]_+ +$$

$$\sum_R [d(R_a, V_p) - d(R_a, V_n) + \alpha]_+ , \quad (3)$$

where margin $\alpha = 0.3$, $d(.)$ is the Euclidean distance, $[x]_+ = max(x, 0)$. In addition, we adopt a hard sample mining strategy [12] to further facilitate the distance learning.

3.6   Translation Consistency Loss

The learned embeddings are useful for training but sometimes lost the modality-specific information which is meaningful and important. To alleviate this information loss, we adopt translation consistency loss following ACME [24], which ensures the learned embeddings preserve the original information across modalities. We accomplish this goal by forcing the recipe and image embeddings to recover the information in the other modalities: recipe image generation and ingredients prediction conditioned on recipe embedding and image embedding respectively. Hence, the total translation consistency loss is composed of the losses for the recipe and the image as follows:

$$L_{trans} = L_{trans_r} + L_{trans_i} \quad (4)$$

Image Generation from Recipe Embeddings   In order to preserve the modality-specific information in recipe embeddings, we aim to recover the information in image modality to ensure this property, and we set two losses as for two-fold goal: (1) we expect the generated image is as realistic as possible, (2) we expect the generated realistic image matches the target recipe. To accomplish these two goals, we adopt GAN [7] to generate images from recipe embeddings and introduce a loss L_{r2i} for goal (1) and a

loss $L_{cls_{r2i}}$ for goal (2) following ACME [24]. During training, the discriminator D_{r2i} is used to distinguish the generated image and real image, and the generator G is used to generate food images which is conditioned on the recipe embedding $FC(E_R(r))$. The loss L_{r2i} for ensuring the generated images realistic is formulated as follows:

$$L_{r2i} = \mathbb{E}_{\mathbf{i} \sim p(i)}[\log(1 - D_{r2i}(\mathbf{i}))] + \mathbb{E}_{\mathbf{t} \sim p(t)}[\log(1 - D_{r2i}(G(\mathbf{FC}(E_R(\mathbf{t})))))] \quad (5)$$

While ensuring the generated image is as realistic as possible, we introduce loss $L_{cls_{r2i}}$ to ensure that the generated image matches the target recipe. During training, a classifier cls_{r2i} is used to encourage the generator to generate a food image with the corresponding food category to recipe embedding. $L_{cls_{r2i}}$ is simply a cross-entropy loss. Combining these two loss functions for two goals separately, the translation consistency loss for the recipe is formulated as follows:

$$L_{trans_r} = L_{r2i} + L_{cls_{r2i}} \quad (6)$$

Classification and Prediction from Image Embeddings We adopt ingredients prediction and title classification on image embeddings here for ensuring the translation consistency of images. We leverage a multi-label network here to predict the ingredients from image embeddings using a 4,102-d one-hot vector representing the existence of 4,102 different ingredients. We denote this multi-label objective as L_{i2r}.

The same as ensuring the generated image is with the correct food category in defining $L_{cls_{r2i}}$, we propose to ensure the predicted ingredients are from the food with the correct category, which means the image embeddings can be classified into the correct food category. We use $L_{cls_{i2r}}$ here to make sure the image embeddings can be classified into one correct food category of 1,047 recipe categories, where $L_{cls_{i2r}}$ is a cross-entropy loss. Combining these two loss functions, the translation consistency loss for the image is formulated as follows:

$$L_{trans_i} = L_{i2r} + L_{cls_{i2r}} \quad (7)$$

4   EXPERIMENTS

In this section, we present the experiments to validate the effectiveness of our proposed framework TNLBT both in recipe retrieval and image generation tasks, including ablation studies and comparison with the previous works. Extensive experiments are also performed to further validate the effectiveness of TNLBT.

4.1   Implementation Details

Data set.   We evaluated the performance of our proposed method on Recipe1M [20, 16] following the previous works. Following the official dataset splits, we used 238,999 image-recipe pairs for training, 51,119 pairs for validation, and 51,303 pairs for testing. **Evaluation Metrics.**   Following previous works, we evaluated retrieval performance with the median rank (medR) and R@{1,5,10} on two different test set sizes 1k and

Table 1: Comparison with the existing works on retrieval performance. The performance of recipe retrieval is evaluated on the criteria of medR($\downarrow$) and R@{1,5,10}($\uparrow$). Especially all the R@K metrics were all improved by around 10.0% in TNLBT-C compared to TNLBT-V.

| | 1k | | | | | | | | 10k | | | | | | | |
| | Image-to-Recipe | | | | Recipe-to-Image | | | | Image-to-Recipe | | | | Recipe-to-Image | | | |
	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10
JE [20]	5.2	24.0	51.0	65.0	5.1	25.0	52.0	65.0	41.9	-	-	-	39.2	-	-	-
R2GAN [27]	2	39.1	71.0	81.7	2	40.6	72.6	83.3	13.9	13.5	33.5	44.9	12.6	14.2	35.0	46.8
MCEN [6]	2	48.2	75.8	83.6	1.9	48.4	76.1	83.7	7.2	20.3	43.3	54.4	6.6	21.4	44.3	55.2
ACME [24]	1	51.8	80.2	87.5	1	52.8	80.2	87.6	6.7	22.9	46.8	57.9	6	24.4	47.9	59.0
SCAN [25]	1	54.0	81.7	88.8	1	54.9	81.9	89.0	5.9	23.7	49.3	60.6	5.1	25.3	50.6	61.6
IMHF [15]	1	53.2	80.7	87.6	1	54.1	82.4	88.2	6.2	23.4	48.2	58.4	5.8	24.9	48.3	59.4
RDEGAN [22]	1	59.4	81.0	87.4	1	61.2	81.0	87.2	3.5	36.0	56.1	64.4	3	38.2	57.7	65.8
H-T [19]	1	60.0	87.6	92.9	1	60.3	87.6	93.2	4	27.9	56.4	68.1	4	28.3	56.5	68.1
X-MRS [9]	1	64.0	88.3	92.6	1	63.9	87.6	92.6	3	32.9	60.6	71.2	3	33.0	60.4	70.7
TNLBT-V	1	75.1	92.3	95.3	1	75.2	92.5	95.4	2	48.0	73.7	81.5	2	48.5	73.7	81.5
TNLBT-C	**1**	**81.0**	**95.2**	**97.4**	**1**	**80.3**	**95.2**	**97.4**	**1**	**56.5**	**80.7**	**87.1**	**1**	**55.9**	**80.1**	**86.8**

Table 2: Comparison with the existing works on image generation. The performance of image generation is evaluated on the criteria of FID($\downarrow$).

Method	FID
ACME [24]	30.7
CHEF [17]	23.0
X-MRS [9]	28.6
TNLBT-V	17.9
TNLBT-C	**16.5**

Table 3: Evaluation of the importance of different components on image-to-recipe retrieval in TNLBT. * means training without recipe-only data.

Applied Components	medR	R@1	R@5	R@10
L_{ret}	2.0	43.4	70.3	79.4
$L_{ret}+L_{MA}$	2.0	41.1	68.0	77.3
$L_{ret}+L_{MA}+L_{rec}$	2.0	43.2	70.0	79.0
$L_{ret}+L_{MA}+L_{rec}+L_{trans_i}$	2.0	43.5	70.3	79.0
$L_{ret}+L_{MA}+L_{rec}+L_{trans_r}$	2.0	44.5	71.3	80.2
$L_{ret}+L_{rec}+L_{trans_r}+L_{trans_i}$	2.0	43.5	70.7	79.5
$L_{ret}+L_{MA}+L_{rec}+L_{trans_r}+L_{trans_i}$*	2.0	43.1	69.8	78.7
$L_{ret}+L_{MA}+L_{rec}+L_{trans_r}+L_{trans_i}$	2.0	44.3	70.9	79.7
TNLBT-V (large batch training)	2.0	48.0	73.7	81.5
TNLBT-C (CLIP-ViT applied)	**1.0**	**56.5**	**80.7**	**87.1**

Table 4: Comparison on image-to-recipe retrieval performance with different batch sizes adopted in TNLBT. Results are reported on rankings of size 10k.

| | TNLBT-V | | | | TNLBT-C | | | |
#batch	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10
64	3.0	36.6	64.3	74.3	2.0	48.0	75.4	83.9
128	2.0	40.9	68.3	77.6	1.4	50.1	77.1	84.9
256	2.0	44.3	70.9	79.7	1.0	53.5	79.1	86.3
512	2.0	47.1	73.4	**81.6**	1.0	55.9	80.4	86.9
768	**2.0**	**48.0**	**73.7**	81.5	**1.0**	**56.5**	**80.7**	**87.1**
1024	2.0	47.5	73.3	81.2	1.0	56.0	79.8	86.5

Table 5: Evaluation on image generation in different batch sizes on the criteria of FID($\downarrow$).

#batch	FID (TNLBT-V)	FID (TNLBT-C)
64	**17.9**	**16.5**
128	19.3	21.9
256	27.1	30.1
512	34.4	48.6
768	46.4	69.7

10k. We reported the average metrics of 10 groups which are randomly chosen from the test set. We also evaluated the performance of image generation conditioned on recipe embeddings using Fréchet Inception Distance (FID) [13] score following previ-

Fig. 2: Generated images conditioned on different recipe embeddings. The first row are recipes used to generate images, the second row are the ground-truth images which recipe embeddings aim to recover. The third row are the generated images.

ous works [22, 9] measuring the similarity of the distribution between generated images and ground-truth images.

Training Details.   The objective is to minimize the total loss function Eq. 1. We empirically decided the hyperparameters in Eq. 1 as follows: $\lambda_1 = 0.05$, $\lambda_2 = 0.005$, $\lambda_3 = 0.002$. For ViT-B [5] image encoder backbone framework, We trained the model for 50 epochs using Adam [14] with a learning rate of 10^{-4}. For CLIP-ViT [18] image encoder backbone framework, we first trained the model for 20 epochs while freezing the image encoder with a learning rate of 10^{-4} and then we trained the model for another 100 epochs with a learning rate of 10^{-6} where the image encoder were not frozen. Since contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning [2], batch size was set to 768 during training.

4.2   Cross-Modal Recipe Retrieval

We evaluated and compared the performance of the proposed framework TNLBT with the previous works in Table 1. TNLBT with ViT-B [5] image encoder backbone (TNLBT-V) outperforms all the existing works across all metrics with large margins. We also tested the performance of our proposed framework with CLIP-ViT [18] image encoder backbone (TNLBT-C), which outperforms all the existing works across metrics with even much larger margins. When the image encoder backbone was changed from ViT-B to CLIP-ViT, all the R@K metrics were all improved by around 10.0%, and medR on 10k test set size was improved to 1.0, which has never been achieved before.

4.3   Image Generation

The proposed model generates images of 128×128 resolutions conditioned on recipe embeddings. We computed the FID [13] score to evaluate the image generation performance of our proposed method[1]. The results are reported in Table 2, where TNLBT outperformed the previous works [24, 17, 9] with a large margin on image generation[2].

---

[1] We used the open source code on `https://github.com/mseitzer/pytorch-fid`

[2] We borrow the FID scores of CHEF [17] and ACME [24] reported in X-MRS [9].

Figure 2 shows qualitative results on image generation conditioned on recipe embeddings, showing that TNLBT could generate appropriate recipe images corresponding to the given ingredients in recipes.

4.4   Ablation Studies

We also performed ablation studies to evaluate the importance of the applied components in Table 3. When L_{MA} was applied, the accuracy decreased incrementally while with more loss functions applied the accuracy increased again. However, when only L_{MA} was taken, the retrieval accuracy decreased compared to the combination of all loss functions. Hence we empirically think there is a complementary relationship among L_{MA} and the other loss terms except L_{ret}, contributing to the retrieval performance. The loss combination without L_{trans_i} applied outperforms all the other combinations by an incremental margin. Finally, when the proposed large batch training strategy and CLIP-ViT [18] applied, the retrieval accuracy was further improved.

4.5   Batch Size in Cross-Modal Embedding Learning

Since contrastive learning benefits from large batch size was reported in [2], we performed extensive experiments here to validate the influence of batch size on both recipe retrieval and image generation, where we are the first to validate this influence.

Recipe Retrieval.   Table 4 shows the image-to-recipe retrieval performance of TNLBT with different batch sizes. Retrieval performance was improved substantially when the batch size increased from 64 to 512, indicating the importance of adopting large batch during training. The improvement by increasing batch size is limited afterwards and the performance was hurt incrementally when batch size increased to 1024.

Image Generation.   We also investigated the influence of batch size on image generation performance in Table 5, where the performance of image generation hurt with batch size increasing, while the performance of recipe retrieval improved according to Table 4. Hence, we emperically believe that batch size serves as a trade-off here.

5   CONCLUSIONS

In this research, we proposed a frustratingly straightforward Transformer-based Network for Large Batch Training (TNLBT) using the Hierarchical Transformer-based recipe encoder, the Vision Transformer-based image encoder, and a sophisticated adversarial network for cross-modal recipe embedding learning. We further adopted self-supervised learning to investigate the complementary information in recipe texts. Through the experiments, it was confirmed that TNLBT outperformed the state-of-the-arts on both cross-modal recipe retrieval and food image generation tasks by large margins. We also found that CLIP-ViT [18] achieved much better performance than ViT-B [5] as an image encoder backbone in TNLBT. With the extensive experiments, we found that the retrieval performance could benefit from a large batch size while the performance of image generation conditioned on recipe embeddings sometimes got hurt by a large batch size, where we were the first to validate the influence of batch size in recipe retrieval and image generation tasks.

References

1. Carvalho, M., Cadène, R., Picard, D., Soulier, L., Thome, N., Cord, M.: Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In: Proc.of ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 35–44 (2018)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proc. of International Conference on Machine Learning (2020)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proc.of IEEE Computer Vision and Pattern Recognition (2009)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics. pp. 4171–4186 (2019)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
6. Fu, H., Wu, R., Liu, C., Sun, J.: MCEN: Bridging cross-modal gap between cooking recipes and dish images with latent variable model. In: Proc.of IEEE Computer Vision and Pattern Recognition (2020)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial networks. Communication of ACM 63(11), 139–144 (2020)
8. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks 18(5-6), 602–610 (2005)
9. Guerrero, R., Xuan, H.P., Vladimir, P.: Cross-modal retrieval and synthesis (x-mrs): Closing the modality gap in shared representation learning. In: Proc.of ACM International Conference Multimedia (2021)
10. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems. pp. 5767–5777 (2017)
11. Guo, W., Wang, J., Wang, S.: Deep multimodal representation learning: A survey. IEEE Access 7, 63373–63394 (2019)
12. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6629–6640 (2017)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. of International Conference on Learning Representation (2015)
15. Li, J., Sun, J., Xu, X., Yu, W., Shen, F.: Cross-modal image-recipe retrieval via intra- and inter-modality hybrid fusion. In: Proc. of ACM International Conference on Multimedia Retrieval. p. 173–182 (2021). https://doi.org/10.1145/3460426.3463618
16. Marin, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., Torralba, A.: 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 (2019)
17. Pham, H.X., Guerrero, R., Pavlovic, V., Li, J.: CHEF: Cross-modal hierarchical embeddings for food domain retrieval. In: Proc. of the AAAI Conference on Artificial Intelligence. pp. 2423–2430 (2021)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from

natural language supervision. In: Proc. of International Conference on Machine Learning. vol. 139, pp. 8748–8763 (2021)

19. Salvador, A., Gundogdu, E., Bazzani, L., Donoser, M.: Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning. In: Proc.of IEEE Computer Vision and Pattern Recognition (2021)

20. Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., Torralba, A.: Learning cross-modal embeddings for cooking recipes and food images. In: Proc.of IEEE Computer Vision and Pattern Recognition (2017)

21. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: Proc.of IEEE Computer Vision and Pattern Recognition (2015)

22. Sugiyama, Y., Yanai, K.: Cross-modal recipe embeddings by disentangling recipe contents and dish styles. In: Proc.of ACM International Conference Multimedia (2021)

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)

24. Wang, H., Sahoo, D., Liu, C., Lim, E.p., Hoi, S.C.: Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In: Proc.of IEEE Computer Vision and Pattern Recognition. pp. 11572–11581 (2019)

25. Wang, H., Sahoo, D., Liu, C., Shu, K., Achananuparp, P., Lim, E., Hoi, S.C.H.: Cross-modal food retrieval: Learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. arXiv:2003.03955 (2020)

26. Zan, Z., Li, L., Liu, J., Zhou, D.: Sentence-based and noise-robust cross-modal retrieval on cooking recipes and food images. In: Proc. of the International Conference on Multimedia Retrieval. p. 117–125 (2020)

27. Zhu, B., Ngo, C.W., Chen, J., Hao, Y.: R2GAN: Cross-modal recipe retrieval with generative adversarial network. In: Proc.of IEEE Computer Vision and Pattern Recognition (2019)