SHOW ROOM 11
POSTER 3

Transformer-based Cross-Modal Embedding Learning with Large Batch Training

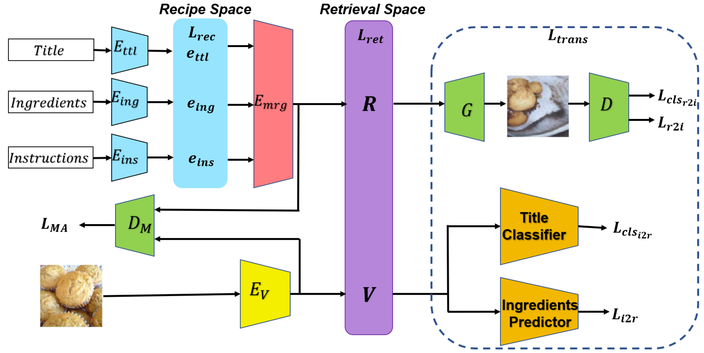
Jing Yang, Junwen Chen, Keiji Yanai
The University of Electro-Communications, Tokyo, Japan

1. Introduction

- Transformer-based framework for recipe retrieval and image generation achieving **current state-of-the-art**
- Our findings**:** Larger batch size, better retrieval performance

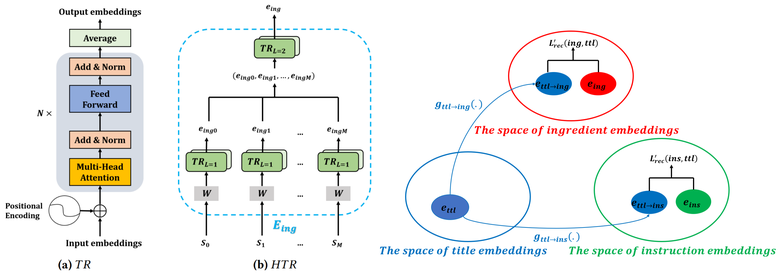
2. Framework Architecture

- **Straightforward:** Hierarchical Transformer text encoder in H-T [2], adversarial network in ACME [1] and CLIP-ViT [4]
- Four loss functions to control training process
- Introduction of CLIP **further boosted** the performance



3. Text encoder and Self-Supervised Learning

- Computing the bi-directional loss after projecting embeddings [2]



4. Comparison to Previous Works on Recipe Retrieval

- **Dataset**: Recipe1M [3], containing 1, 000, 000 of paired recipe data
- R@K: Recall rate at rank K, FID: Frechet Inception Distance
- TNLBT-V/ TNLBT-C: ViT-B/ CLIP-ViT backbone

10k test set size									
		Image-to-recipe				Recipe-to-image			
	FID	medR	R1	R5	R10	medR	R1	R5	R10
ACME (CVPR 2019)	30.7	6.7	22.9	46.8	57.9	6.0	24.4	47.9	59.0
H-T (CVPR 2021)	-	4.0	27.9	56.4	68.1	4.0	28.3	56.5	68.1
IMHF (ICMR 2021)	-	6.2	23.4	48.2	58.4	5.8	24.9	48.3	59.4
RDE-GAN (MM 2021)	-	3.5	36.0	56.1	64.4	3.0	38.2	57.7	65.8
X-MRS (MM 2021)	28.6	3.0	32.9	60.6	71.2	3.0	33.0	60.4	70.7
T-FOOD (CVPR 2022)	-	2.0	43.4	70.7	79.7	2.0	44.6	71.2	79.7
EOMA (ICMR 2022)	-	1.0	50.6	77.1	84.8	1.3	50.1	76.8	84.5
TNLBT-V	17.9	2.0	48.0	73.7	81.5	2.0	48.5	73.7	81.5
TNLBT-C	**16.5**	**1.0**	**56.5**	**80.7**	**87.1**	**1.0**	**55.9**	**80.1**	**86.8**

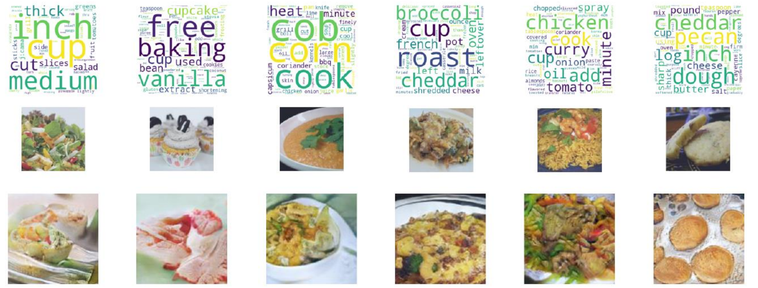
5. Visualization of Retrieval Results

- Sample highlighted in blue: query; sample highlighted in green: target
- Top retrieved are with high similarity to the query sample

Image2Recipe:

Recipe2Image:

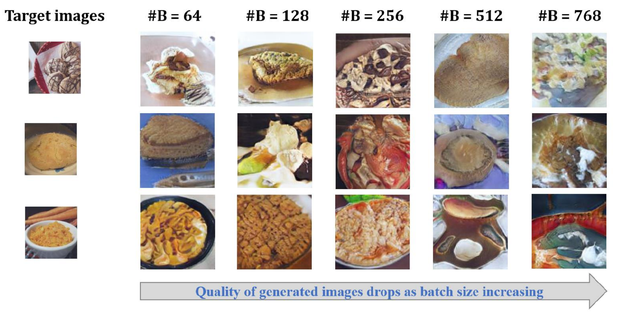
6. Image Generation Conditioned on Recipe Text



7. The Influence of Batch Size on Performance

- Larger batch size leading to higher R@1 (better retrieval performance)
- While the generated images start to collapse (worse image generation)

Image2Recipe		
#batch	MedR↓	R@1↑
64	2.0	48.0
128	1.4	50.1
256	1.0	53.5
512	1.0	55.9
768	**1.0**	**56.5**
1024	1.0	56.0



8. Reference

[1] H. Wang, D Sahoo, C. Liu, E Lim, and S. Hoi. Learning Cross-Modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images. CVPR2019
[2] A. Salvador, E. Gundogdu, L. Bazzani, and M. Donoser. Revamping Cross-Modal Recipe Retrieval with Hierarchical Transformers and Self-supervised Learning. CVPR 2021
[3] A. Salvador , N. Hynes et al. Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. CVPR 2017
[4] 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)