Improving Cross-Modal Recipe Embeddings with Cross Decoder

Jing Yang Junwen Chen Keiji Yanai

The University of Electro-Communications

Background: Cross-Modal Retrieval

□ Modality

- Text, image, audio, video...

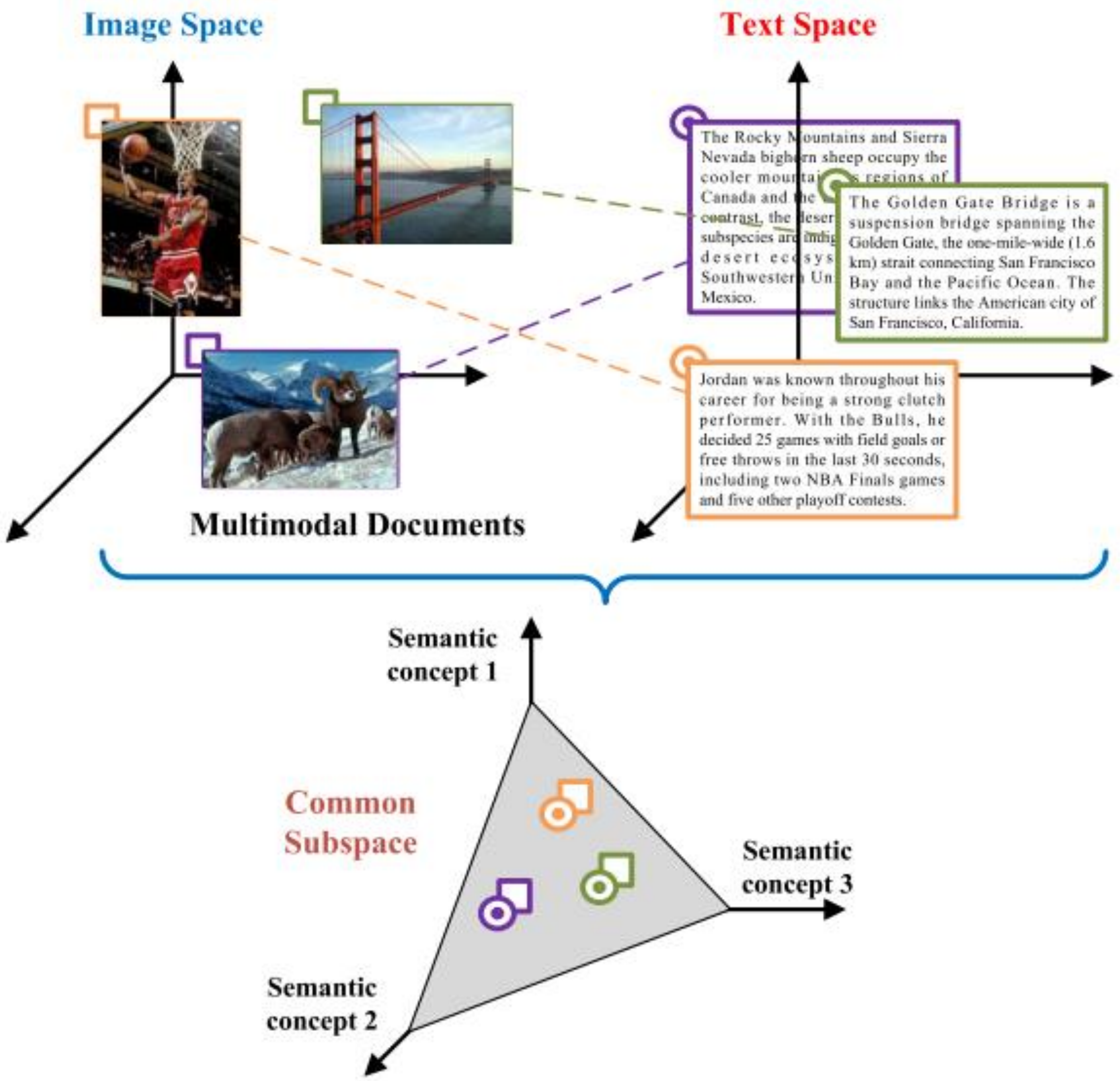
□ Cross-modal image-text retrieval

- Build the connection is difficult

➡ The gap between modalities

□ Solution

- Embeddings & Distance Learning
- A large number of data pairs



The Rocky Mountains and Sierra
Nevada bighorn sheep occupy the
cooler mountai s regions of
Canada and the
contrast, the deser
subspecies are indig
desert ecosys
Southwestern Un
Mexico.

The Golden Gate Bridge is a suspension bridge spanning the Golden Gate, the one-mile-wide (1.6 km) strait connecting San Francisco Bay and the Pacific Ocean. The structure links the American city of San Francisco, California.

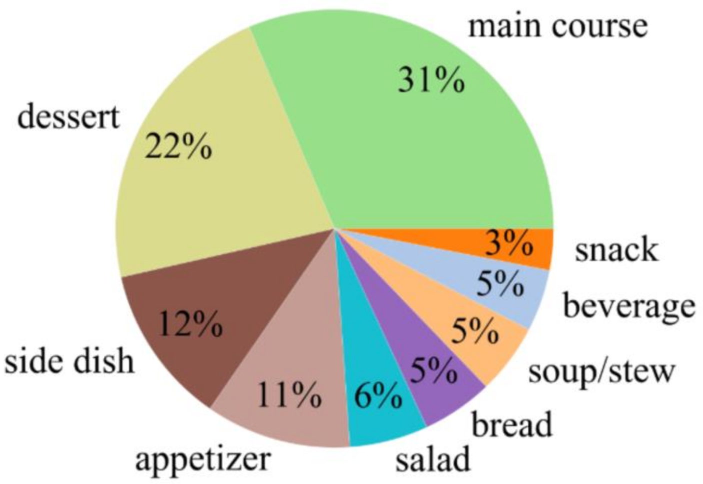
Jordan was known throughout his career for being a strong clutch performer. With the Bulls, he decided 25 games with field goals or free throws in the last 30 seconds, including two NBA Finals games and five other playoff contests.

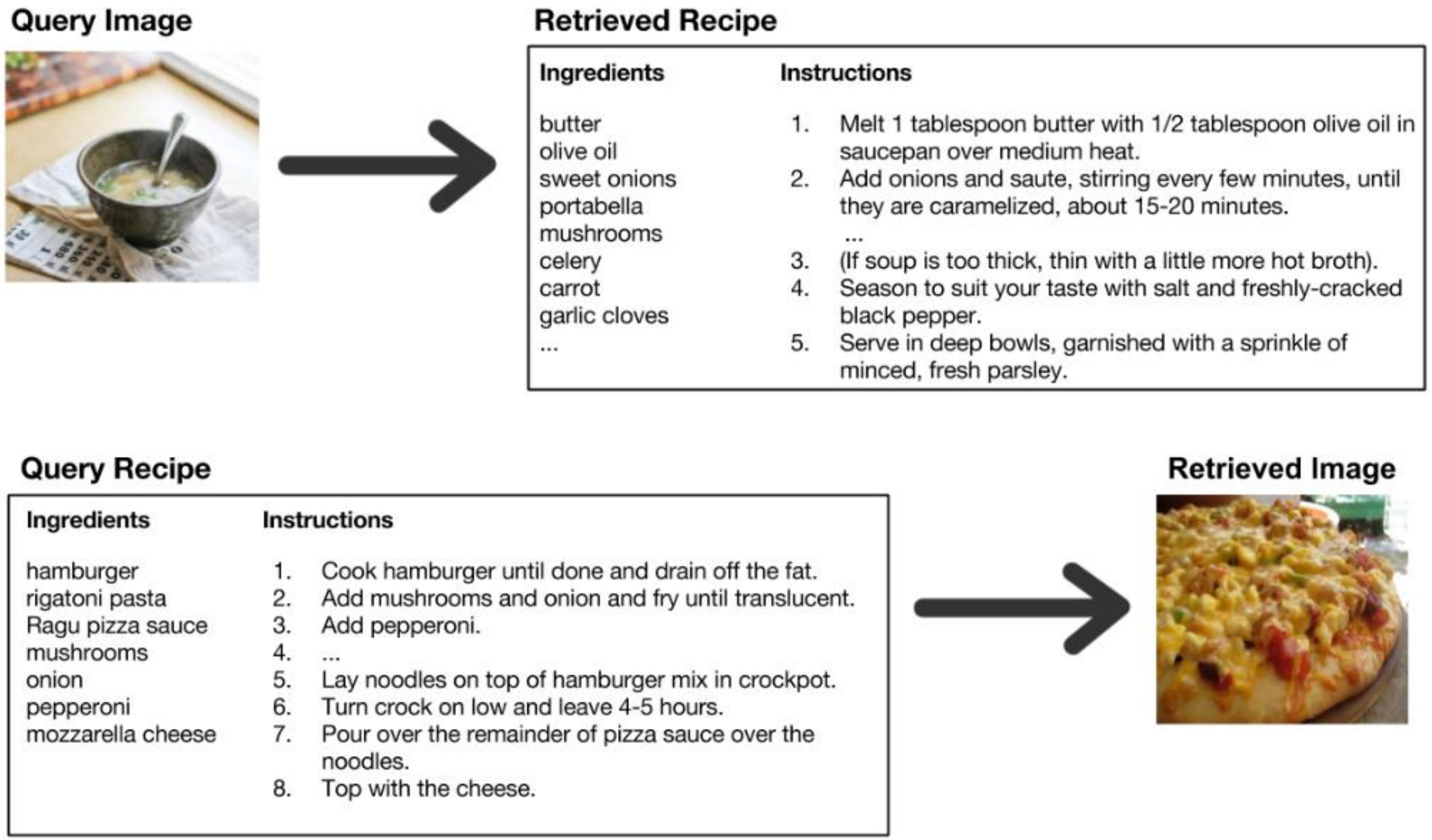
Guo, Wenzhong, Jianwen Wang, and Shiping Wang. "Deep multimodal representation learning: A survey."

Background: Recipe Retrieval and Dataset

□ Recipe1M

- One of the applications of cross-modal retrieval
- 1 million pairs of recipe images and recipe texts





Ingredients

butter
olive oil
sweet onions
portabella mushrooms
celery
carrot
garlic cloves
...

Instructions

1. Melt 1 tablespoon butter with 1/2 tablespoon olive oil in saucepan over medium heat.
2. Add onions and saute, stirring every few minutes, until they are caramelized, about 15-20 minutes.
   ...
3. (If soup is too thick, thin with a little more hot broth).
4. Season to suit your taste with salt and freshly-cracked black pepper.
5. Serve in deep bowls, garnished with a sprinkle of minced, fresh parsley.

Ingredients

hamburger
rigatoni pasta
Ragu pizza sauce
mushrooms
onion
pepperoni
mozzarella cheese

Instructions

1. Cook hamburger until done and drain off the fat.
2. Add mushrooms and onion and fry until translucent.
3. Add pepperoni.
4. ...
5. Lay noodles on top of hamburger mix in crockpot.
6. Turn crock on low and leave 4-5 hours.
7. Pour over the remainder of pizza sauce over the noodles.
8. Top with the cheese.

Salvador, Amaia, et al. "Learning cross-modal embeddings for cooking recipes and food images."

Background: Recipe Retrieval and Dataset

□ Challenge of Recipe Retrieval

- Text
  - Ingredients are diverse (and rare in dataset)
  - Instructions are detailed (or lengthy) and diverse

Query Recipe

Ingredients

butter
garlic cloves
all - purpose flour
kosher salt
milk
chicken broth
mozzarella cheese
parmesan cheese
onion
...

Instructions

1. Heat butter in 2 qt saucepan over low heat until melted
2. Add garlic.
3. Stir in flour and salt.
4. Cook, stirring constantly until bubbly.
5. Remove from heat and stir in milk and broth.

...

6. Cook uncovered at 350F 20-30 minutes until nice and bubbly.
7. Let stand 10 minutes before cutting.

Retrieved Image

SIMPLY BREAKFAST LASAGNA

Query Image

CROCK POT PIZZA

Retrieved Recipe

Ingredients

spiral shaped pasta
pepperoni
ground beef
pizza sauce
mozzarella cheese
dried parsley
onion powder
garlic

Instructions

1. Cook pasta according to package directions and drain.
2. Pour into large mixing bowl.
3. Finely chop half of the pepperoni.
4. ...
5. Pour in lightly greased casserole dish.
6. Sprinkle remaining half of cheese over top.
7. Place remaining pepperoni slices on top.
8. Sprinkle with parsley.
9. Bake in 350 degree oven until cheese bubbles.

Salvador, Amaia, et al. "Learning cross-modal embeddings for cooking recipes and food images.“ CVPR 2017

Background: Recipe Retrieval and Dataset

□ Challenge of Recipe Retrieval

- Image
  - Various plating (in bowls, on plates, on the table…)

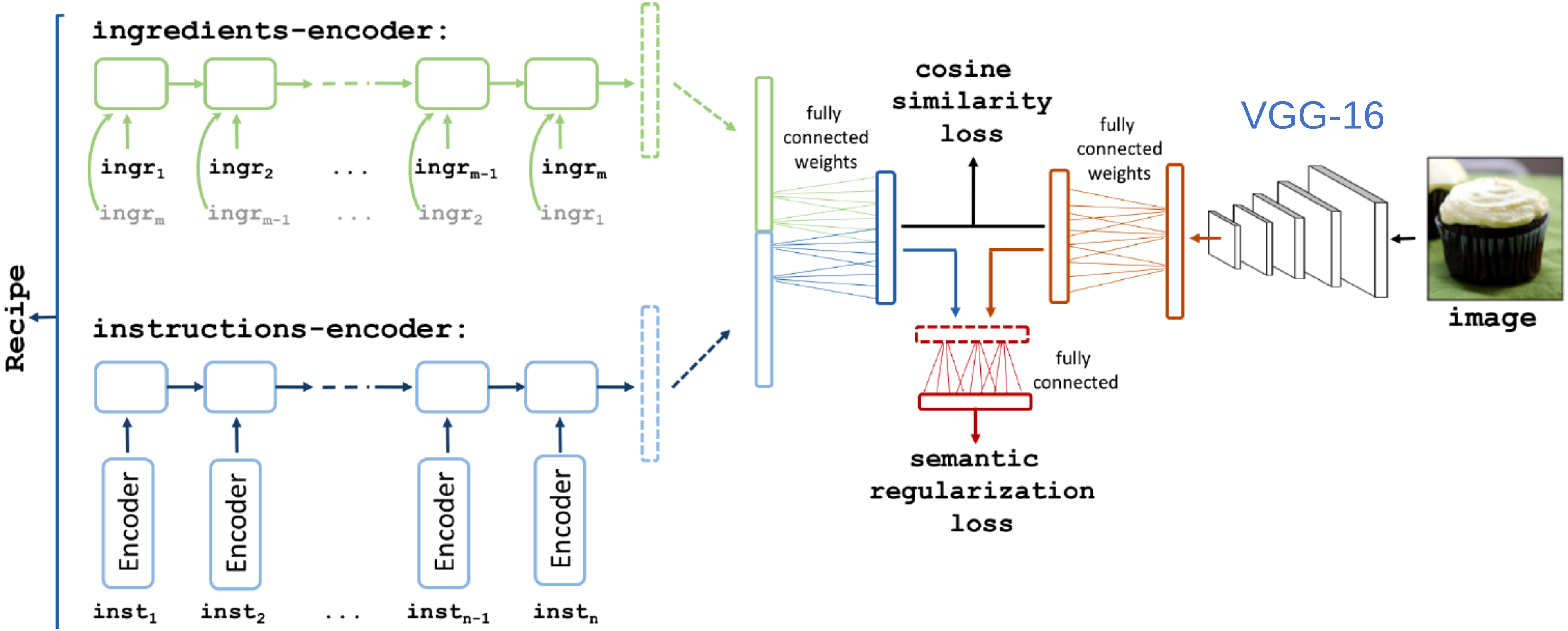
  - Different amount and background

Salvador, Amaia, et al. "Learning cross-modal embeddings for cooking recipes and food images.“ CVPR 2017

Related Works

□ Joint Embedding

- A framework with the proposal of Recipe1M
- Bidirectional LSTM for ingredients encoder
- Regular LSTM for instruction encoder



Salvador, Amaia, et al. "Learning cross-modal embeddings for cooking recipes and food images." CVPR 2017

Related Works

□ AdaMine

- Retrieval Loss (Triplet Loss)

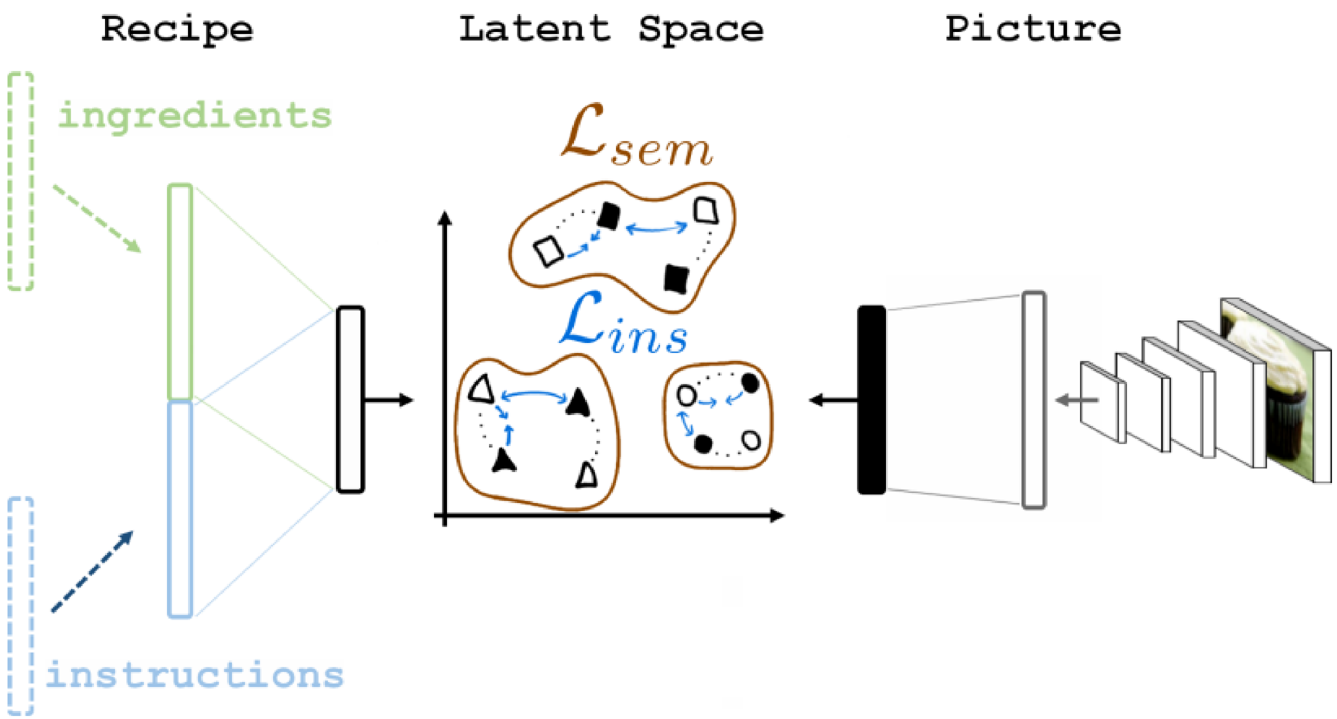
Cosine Distance

$$\ell_{ins}(\theta, x_q, x_p, x_n) = \left[d(x_q, x_p) + \alpha - d(x_q, x_n)\right]_+$$

Query

Retrieval Target

Dissimilar item

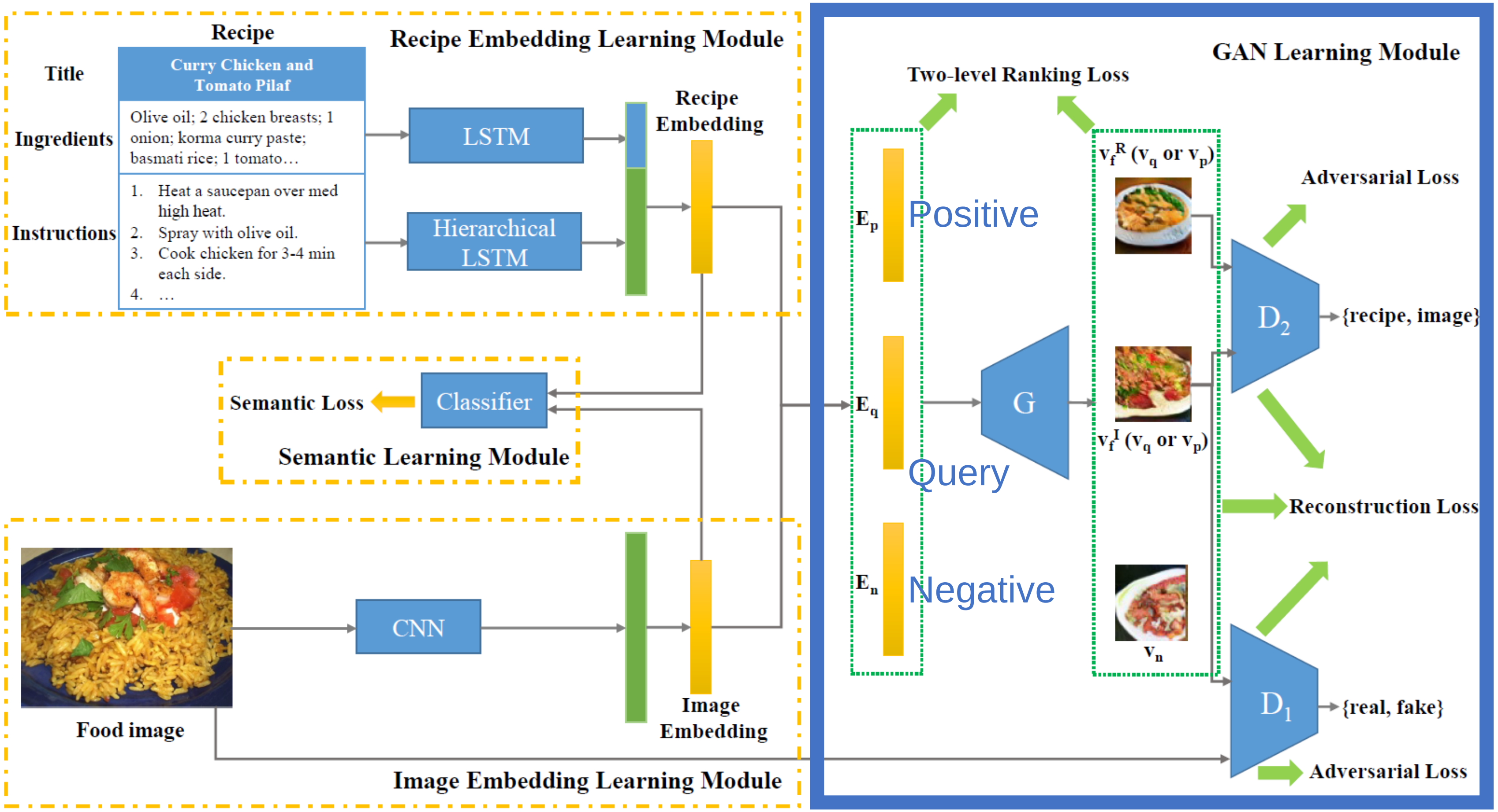


Carvalho, Micael, et al. "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings." ACM SIGIR 2018

Related Works

□ R2GAN

- Using GAN to learn compatible cross-modal features



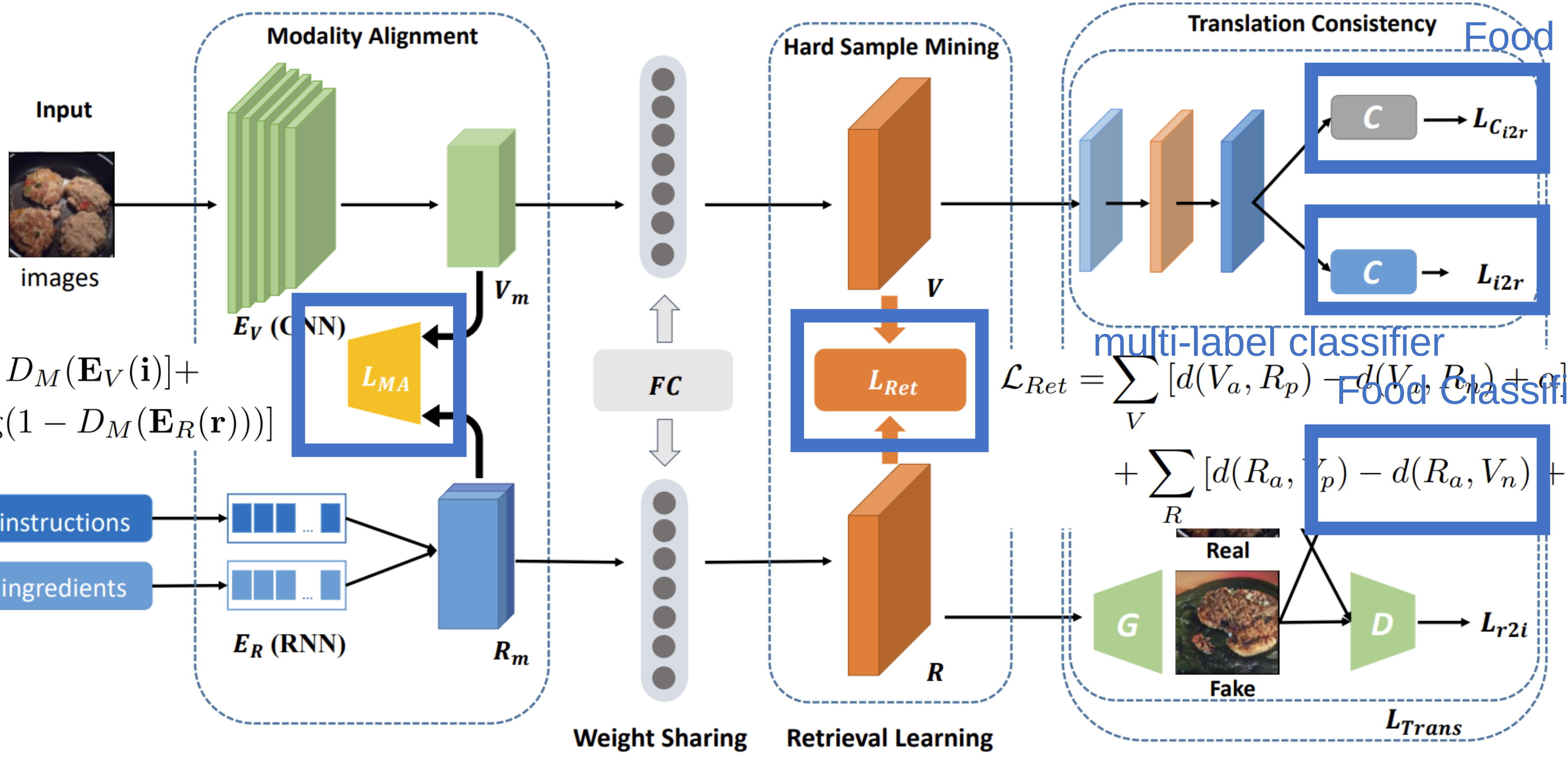
Zhu, Bin, et al. "R2gan: Cross-modal recipe retrieval with generative adversarial network." CVPR 2019

□ Adversarial Cross-Modal Embedding (ACME)

- Translation consistency losses and a new triplet loss
- Adversarial loss $\mathcal{L}_{MA}$ for modality alignment

Recover the other modality

Food Classifier



multi-label classifier

Food Classifier

$$\mathcal{L}_{MA} = \mathbb{E}_{\mathbf{i}\sim p_{image}}[\log D_M(\mathbf{E}_V(\mathbf{i})] + \mathbb{E}_{\mathbf{r}\sim p_{recipe}}[\log(1 - D_M(\mathbf{E}_R(\mathbf{r})))]$$

$$\min_{\mathbf{E}_V,\mathbf{E}_R} \max_{D_M} \mathcal{L}_{MA}$$

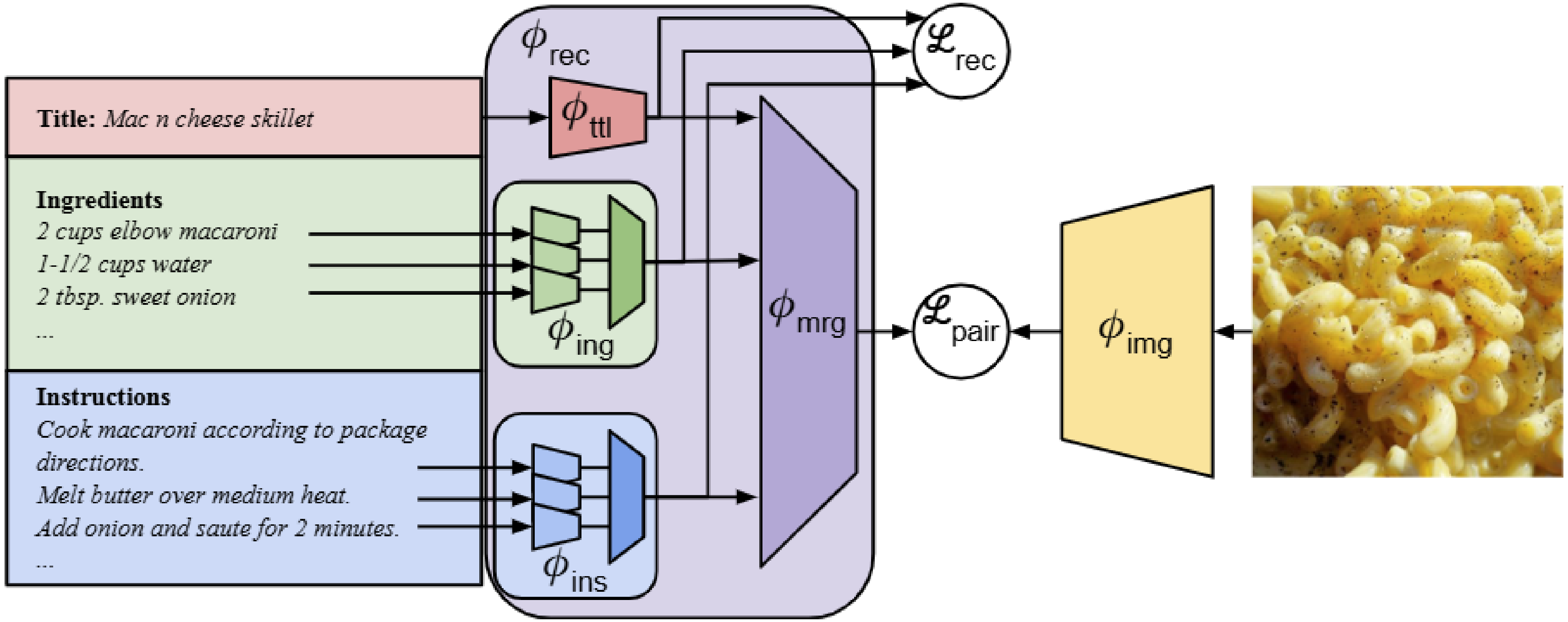
$$\mathcal{L}_{Ret} = \sum_V [d(V_a, R_p) - d(V_a, R_n) + \alpha]_+ + \sum_R [d(R_a, V_p) - d(R_a, V_n) + \alpha]_+$$

Wang, Hao, et al. "Learning cross-modal embeddings with adversarial networks for cooking recipes and food images." CVPR 2019

Related Works

□ Hierarchical Transformers (H-T)

- Hierarchical transformers to encode recipe
- Self-supervised losses on top of pairs of recipe components

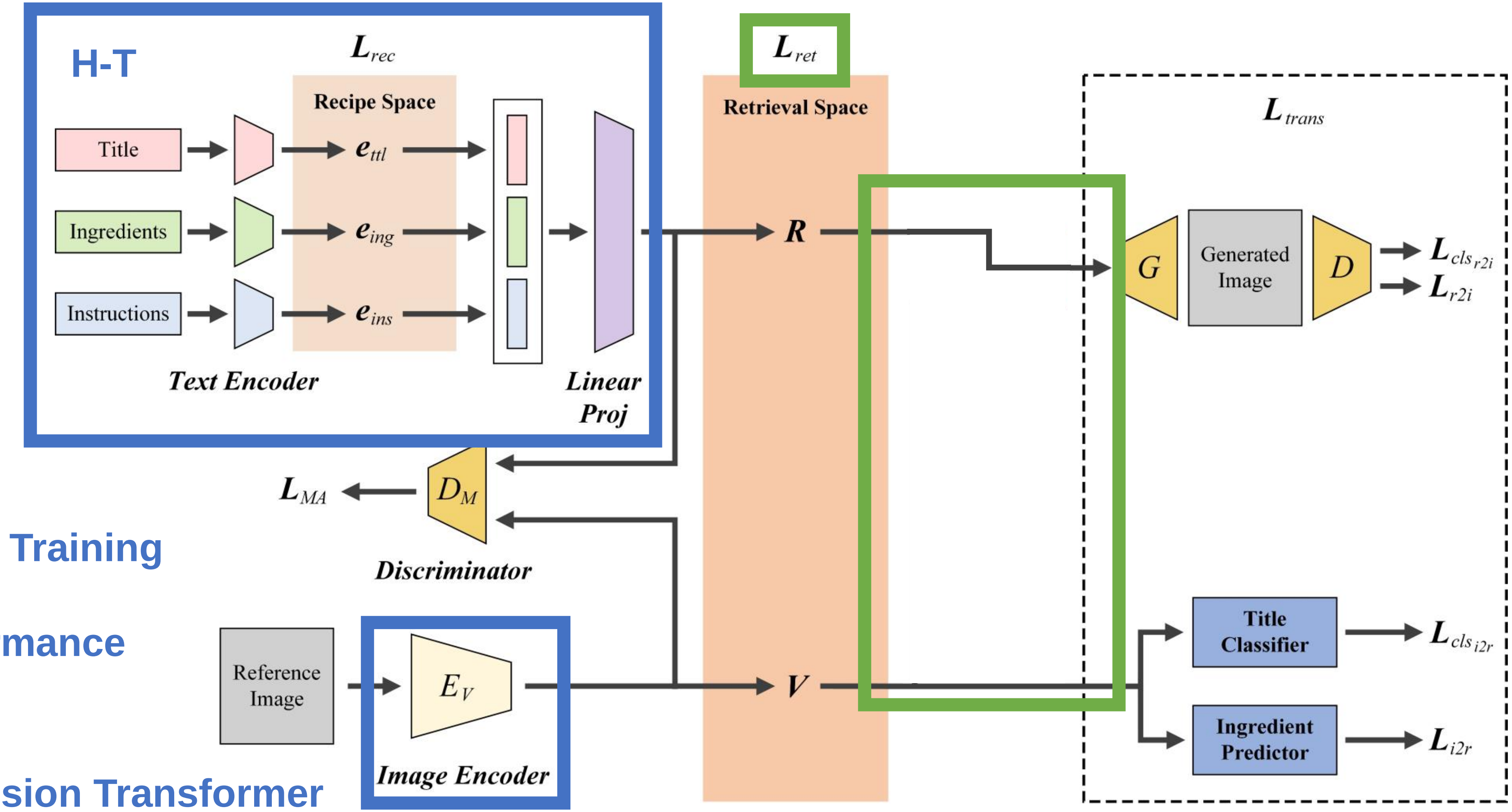


Salvador, Amaia, et al. "Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning." CVPR 2021

Motivation & Method

□ TNLBT + Dynamic Margins + Cross Decoder

- Improving the representation capability of the recipe embeddings



Large Batch Training

SOTA performance

Vision Transformer

TNLBT : Yang, Jing, Junwen Chen, and Keiji Yanai. "Transformer-Based Cross-Modal Recipe Embeddings with Large Batch Training." MMM 2023

Method: Dynamic Margins

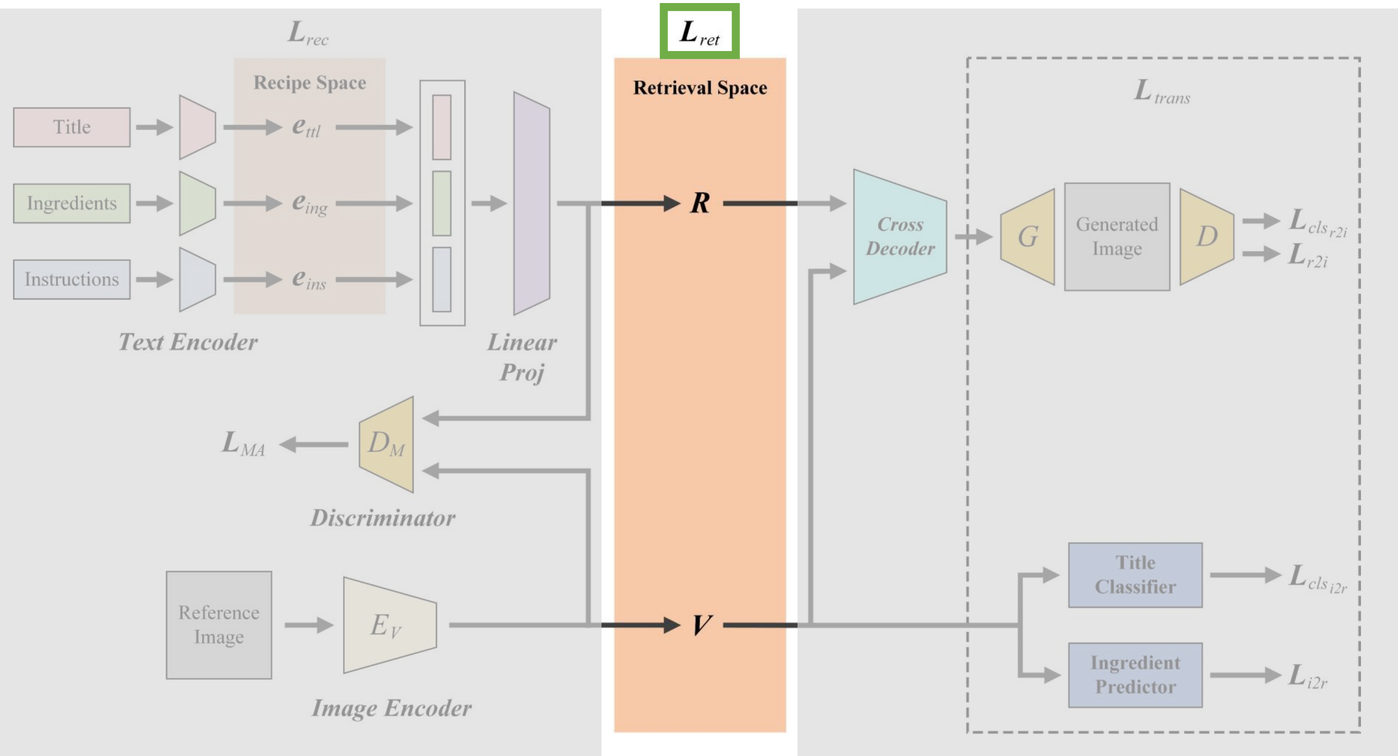
□ Distance learning with dynamic margins

- Adjust the learning difficulty of retrieval loss

$$\alpha \rightarrow \alpha_{dm}$$

- Increase α_{dm} during training

$$L_{ret} = \sum_{V} [d(V_a, R_p) - d(V_a, R_n) + \boxed{\alpha_{dm}}]_+ + \sum_{R} [d(R_a, V_p) - d(R_a, V_n) + \boxed{\alpha_{dm}}]_+$$

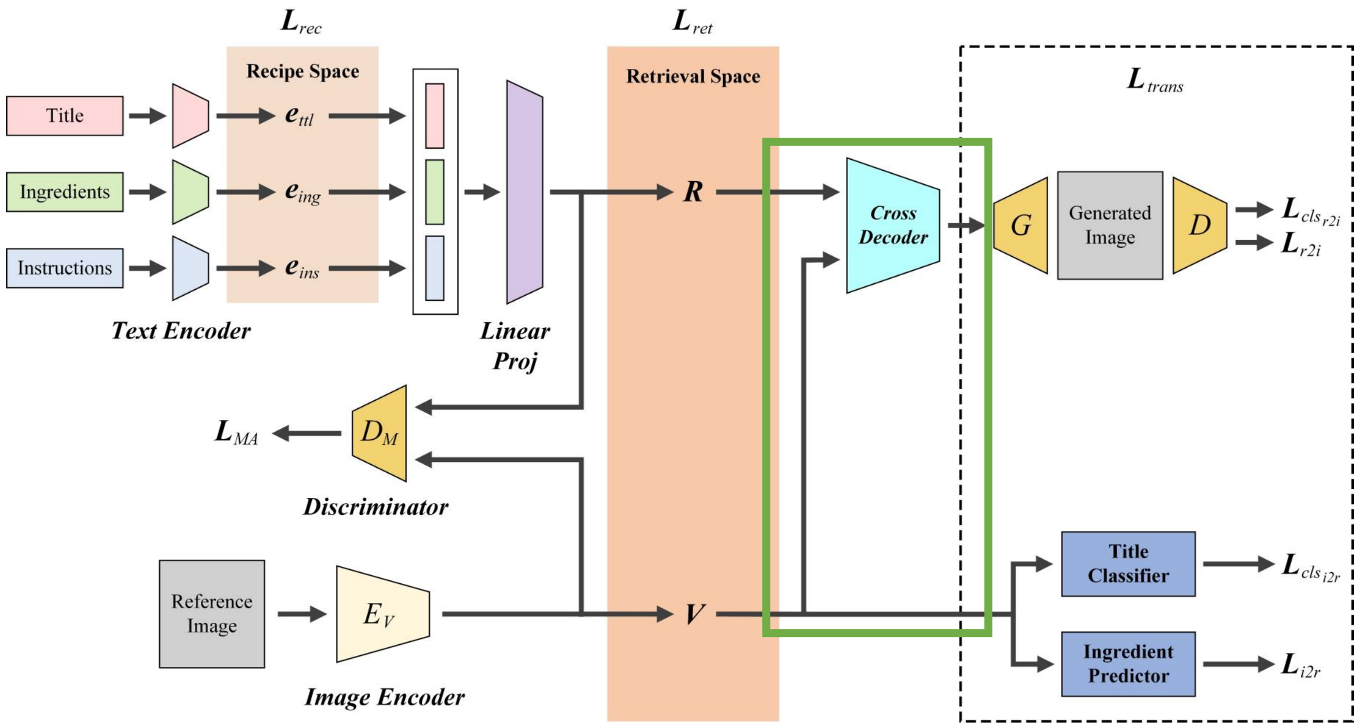


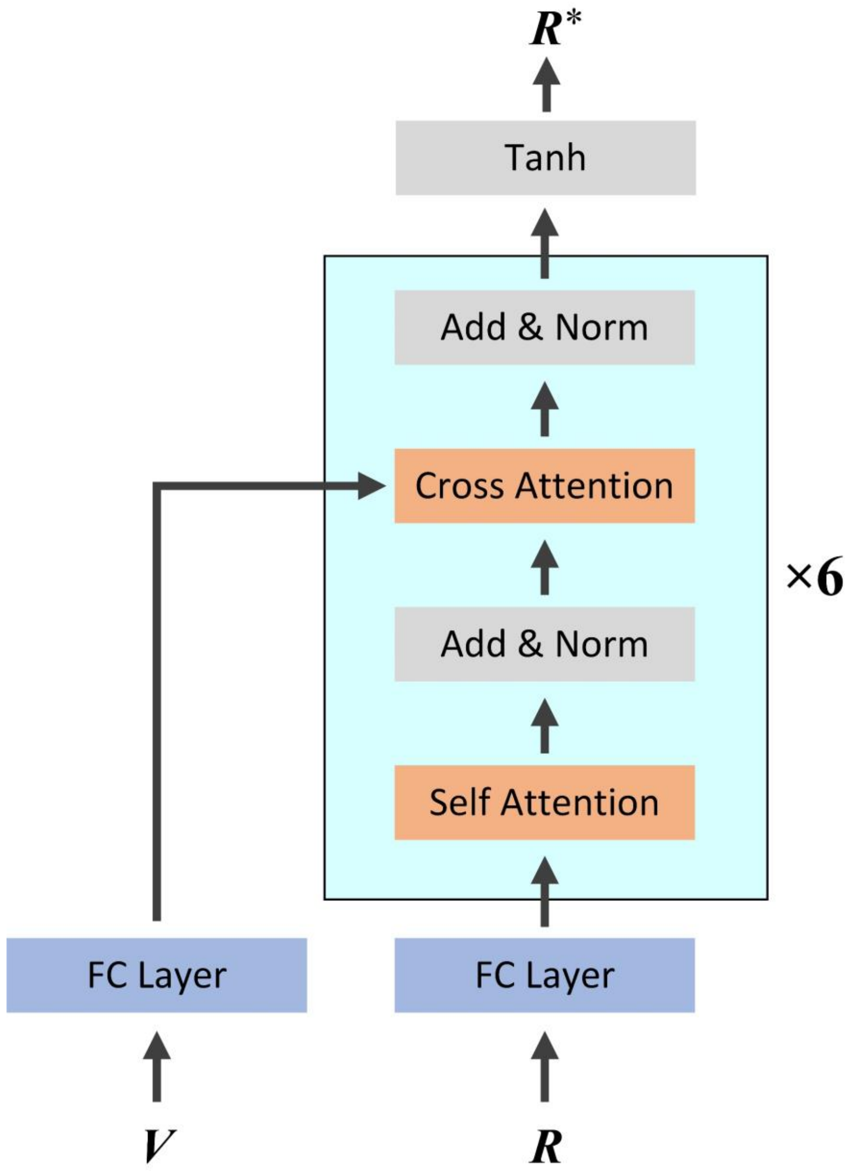
Method: Cross Decoder

□ Improving cross-modal recipe embeddings

- Fusing before Generating

$$R^{*} = CrossDec(FC(R), FC(V))$$





Experiments: Evaluation Metric

□ Test Process

- Randomly select recipe-image pairs from test set
  - 1k setting and 10k setting

□ medR

- Median rank of the closest ground truth result in the list

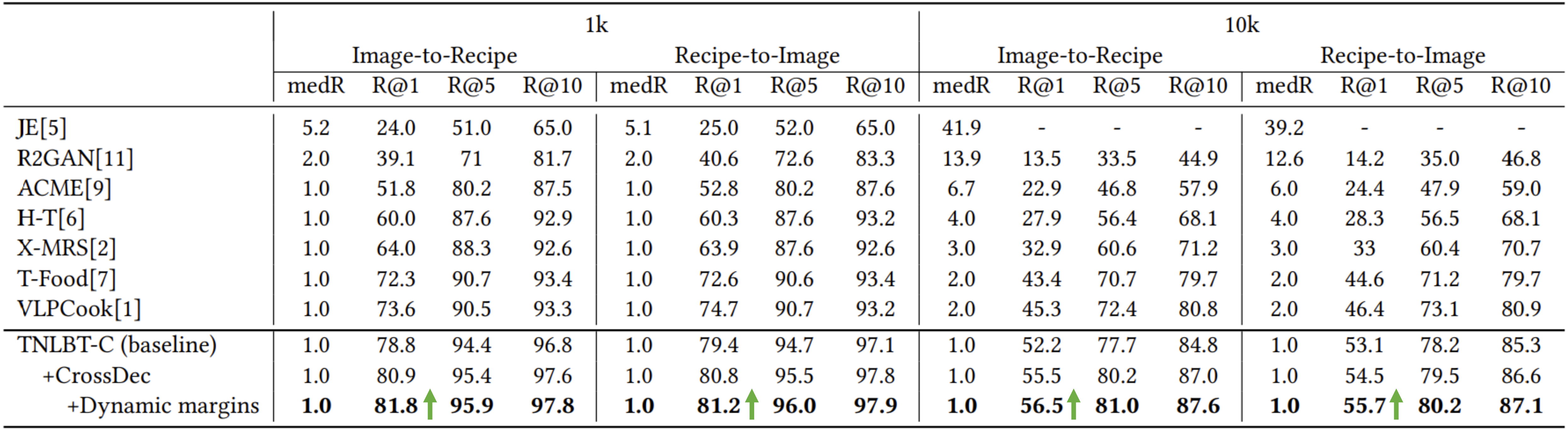
□ R@K

- Recall percentage at top K (R@1, R@5, R@10)

Experiments: Results

□ Comparison with state-of-the-art methods

- Randomly select recipe-image pairs from test set



	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[5]	5.2	24.0	51.0	65.0	5.1	25.0	52.0	65.0	41.9	-	-	-	39.2	-	-	-
R2GAN[11]	2.0	39.1	71	81.7	2.0	40.6	72.6	83.3	13.9	13.5	33.5	44.9	12.6	14.2	35.0	46.8
ACME[9]	1.0	51.8	80.2	87.5	1.0	52.8	80.2	87.6	6.7	22.9	46.8	57.9	6.0	24.4	47.9	59.0
H-T[6]	1.0	60.0	87.6	92.9	1.0	60.3	87.6	93.2	4.0	27.9	56.4	68.1	4.0	28.3	56.5	68.1
X-MRS[2]	1.0	64.0	88.3	92.6	1.0	63.9	87.6	92.6	3.0	32.9	60.6	71.2	3.0	33	60.4	70.7
T-Food[7]	1.0	72.3	90.7	93.4	1.0	72.6	90.6	93.4	2.0	43.4	70.7	79.7	2.0	44.6	71.2	79.7
VLPCook[1]	1.0	73.6	90.5	93.3	1.0	74.7	90.7	93.2	2.0	45.3	72.4	80.8	2.0	46.4	73.1	80.9
TNLBT-C (baseline)	1.0	78.8	94.4	96.8	1.0	79.4	94.7	97.1	1.0	52.2	77.7	84.8	1.0	53.1	78.2	85.3
+CrossDec	1.0	80.9	95.4	97.6	1.0	80.8	95.5	97.8	1.0	55.5	80.2	87.0	1.0	54.5	79.5	86.6
+Dynamic margins	**1.0**	**81.8** ↑	**95.9**	**97.8**	**1.0**	**81.2** ↑	**96.0**	**97.9**	**1.0**	**56.5** ↑	**81.0**	**87.6**	**1.0**	**55.7** ↑	**80.2**	**87.1**

3.8% 2.3% 8.2% 4.9%

Conclusion

- We introduce a **Cross Decoder** to improve the representation capability of the cross-modal recipe embeddings

- We introduce **dynamic margins** into the retrieval distance learning to adjust the learning difficulty

- The results on the Recipe1M dataset show that our method outperforms the **state-of-the-art** methods

UEC