

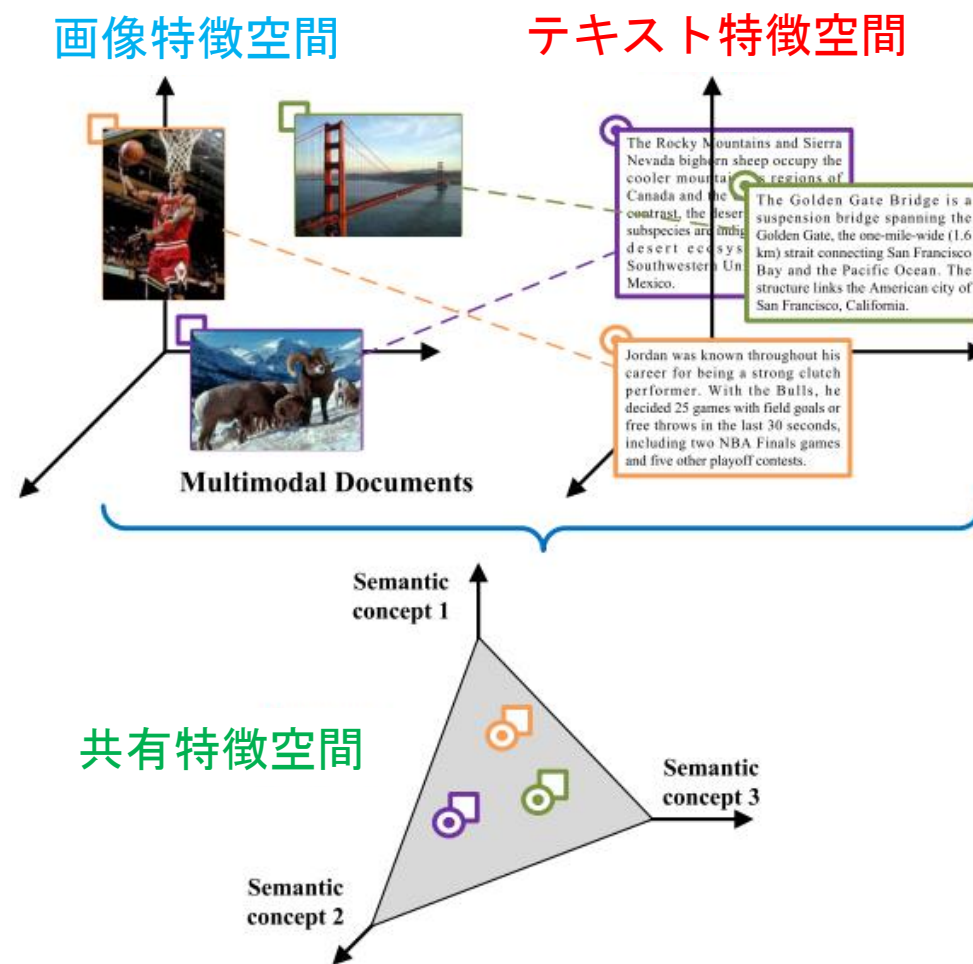
PRMU 2022

Transformerを用いた クロスモーダル レシピ検索・画像生成

電気通信大学 I類 (情報系)
楊景 柳井啓司

はじめに：クロスモーダル検索とは？


- 異なるメディアの情報源（テキスト、画像）をモダリティと呼ぶ
- 異なるモダリティの特徴を、同じ空間に埋め込み、なるべく分布を近づける
- これにより、画像・テキストの互いの検索が可能となる
→クロスモーダル検索



研究背景

- クロスモーダル検索の応用例の一つとして、レシピ検索があげられる
- レシピテキストとレシピ画像を同じ空間に埋め込むことで、クロスモーダル検索が可能となる
- Recipe1M [A] : レシピ画像とレシピテキストの100万ペアのデータセット

Query Image




→

Retrieved Recipe

Ingredients	Instructions
pasta	1. Preheat oven to 350F.
ground beef	2. Boil pasta until just cooked.
taco seasoning	3. Brown ground beef and then drain.
water	4. Add taco seasoning and water to meat and simmer for 5 minutes.
cream cheese	...
cheese	5. Put half of the shredded cheese over pasta, then cover with hamburger meat and mix gentle.
	6. Sprinkle remaining cheese over the top.
	7. Cook in the oven uncovered for 15-20 minutes.

Query Image



→

Retrieved Recipe


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

Query Recipe

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

→

Retrieved Image




Query Recipe

Ingredients	Instructions
cooked white rice	1. Roll the rice into a ball about the size of a large mini tomato.
salt	2. Wet your hands and lightly coat in salt.
shrimp	3. Divide the nori into 6 long strips, and make 6 long and narrow sushi wraps.
Broccolini	4. Remove the hard stems from the broccolini, cut to 3-4 cm lengths, parboil in salt water (not listed), then drain.
mayonnaise	5. Roll the rice in the nori seaweed, top with the shrimp, broccolini, mayonnaise, and they are done.
nori	

→

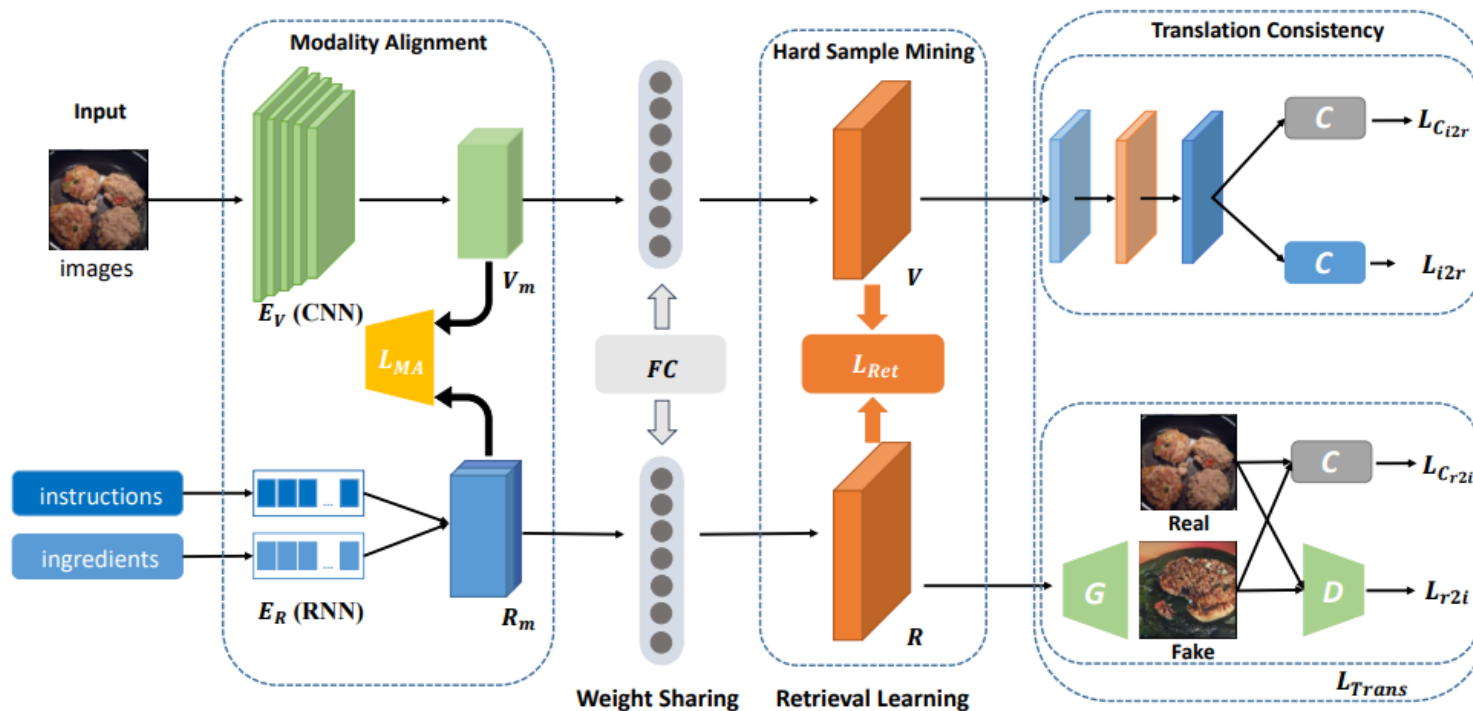
Retrieved Image



[A] A. Salvador , N. Hynes et al. Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. CVPR 2017

関連研究 (1) : 画像生成の導入

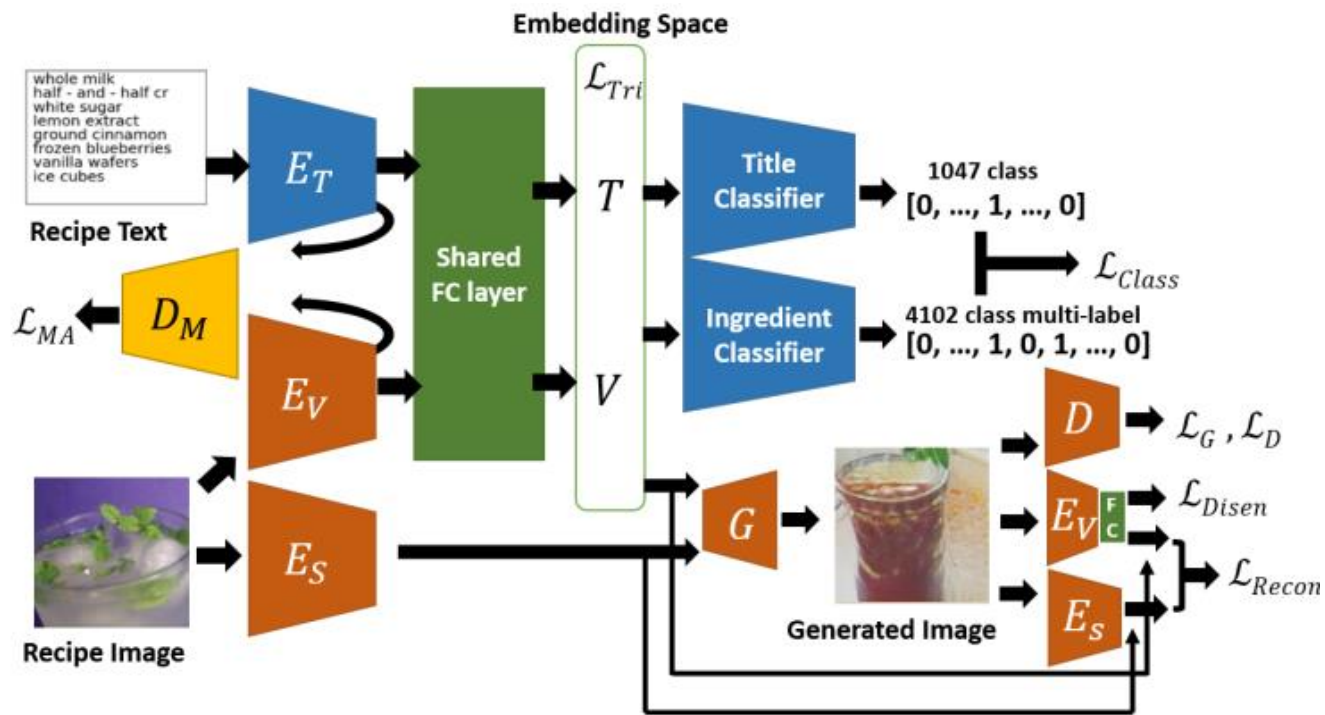
- Adversarial Cross-Modal Embedding (ACME) [CVPR2019]
- 画像生成をクロスモーダルレシピ検索に取り入れ、検索精度を向上



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

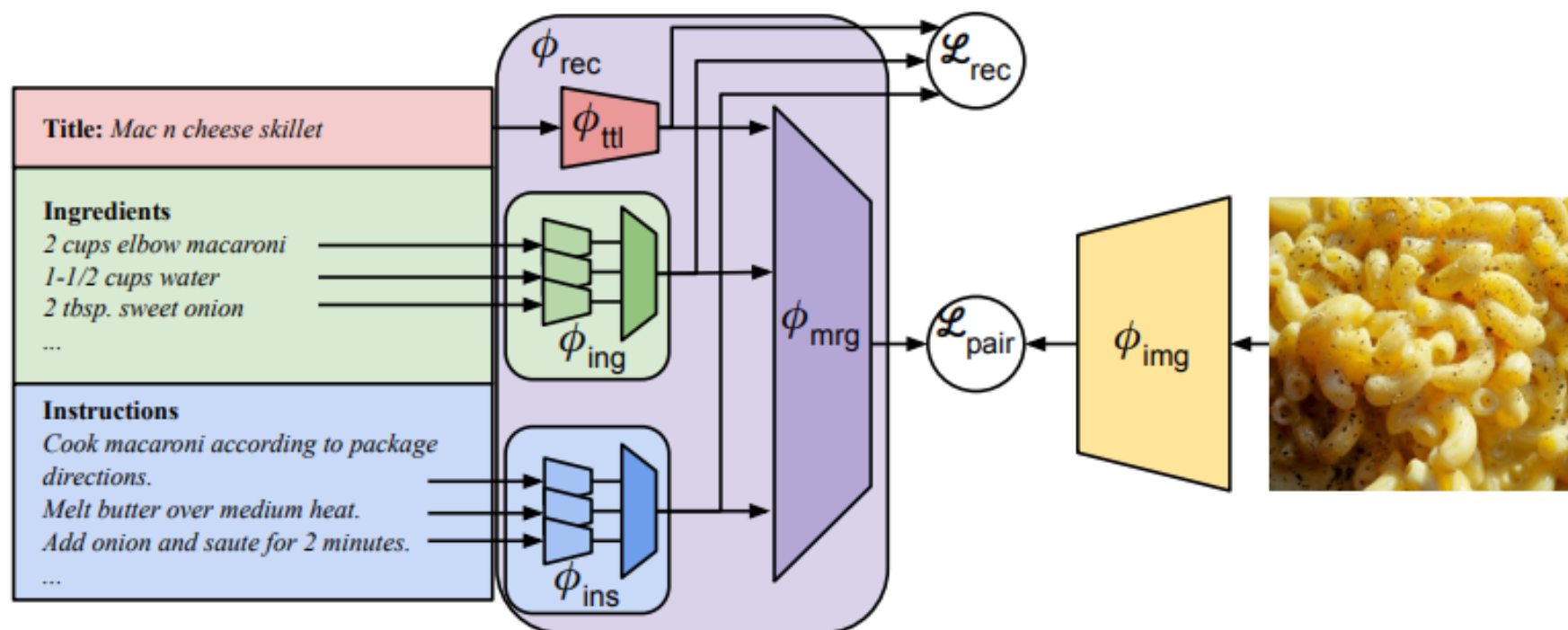
関連研究 (2): 画像エンコーダーの工夫

- Recipe Disentangling Embedding (RDE-GAN) [ACM MM 2021]
- 形状情報と画像意味情報に画像を分解し、検索精度を向上



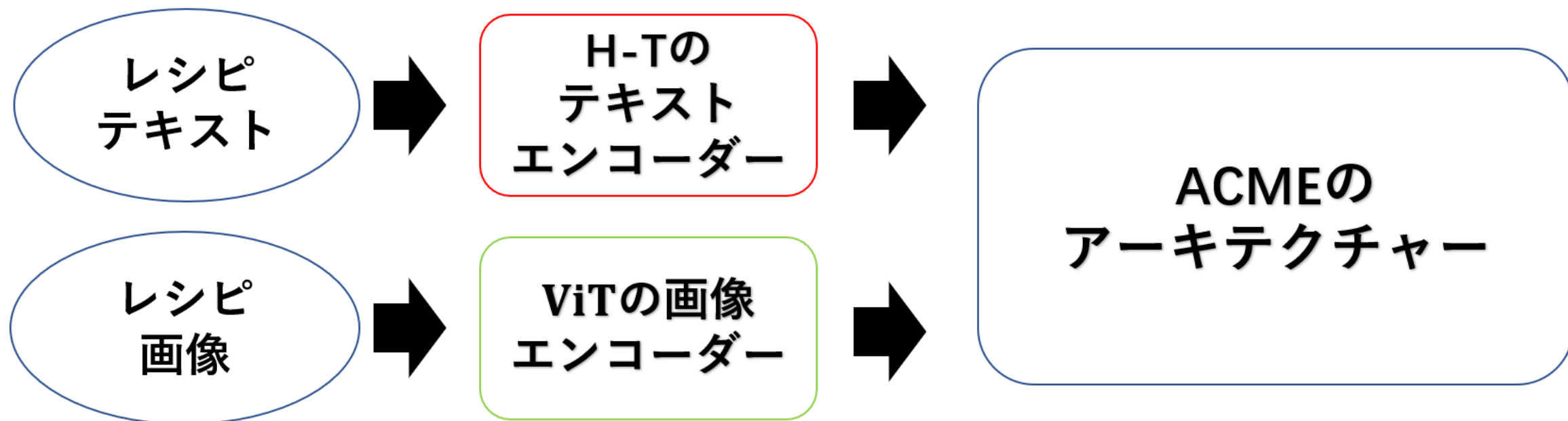
関連研究 (3): テキストエンコーダーの改良

- Hierarchical Transformers (H-T) [CVPR 2021]
- 階層的Transformerと自己教師あり学習をテキストエンコーダーに利用



研究目的

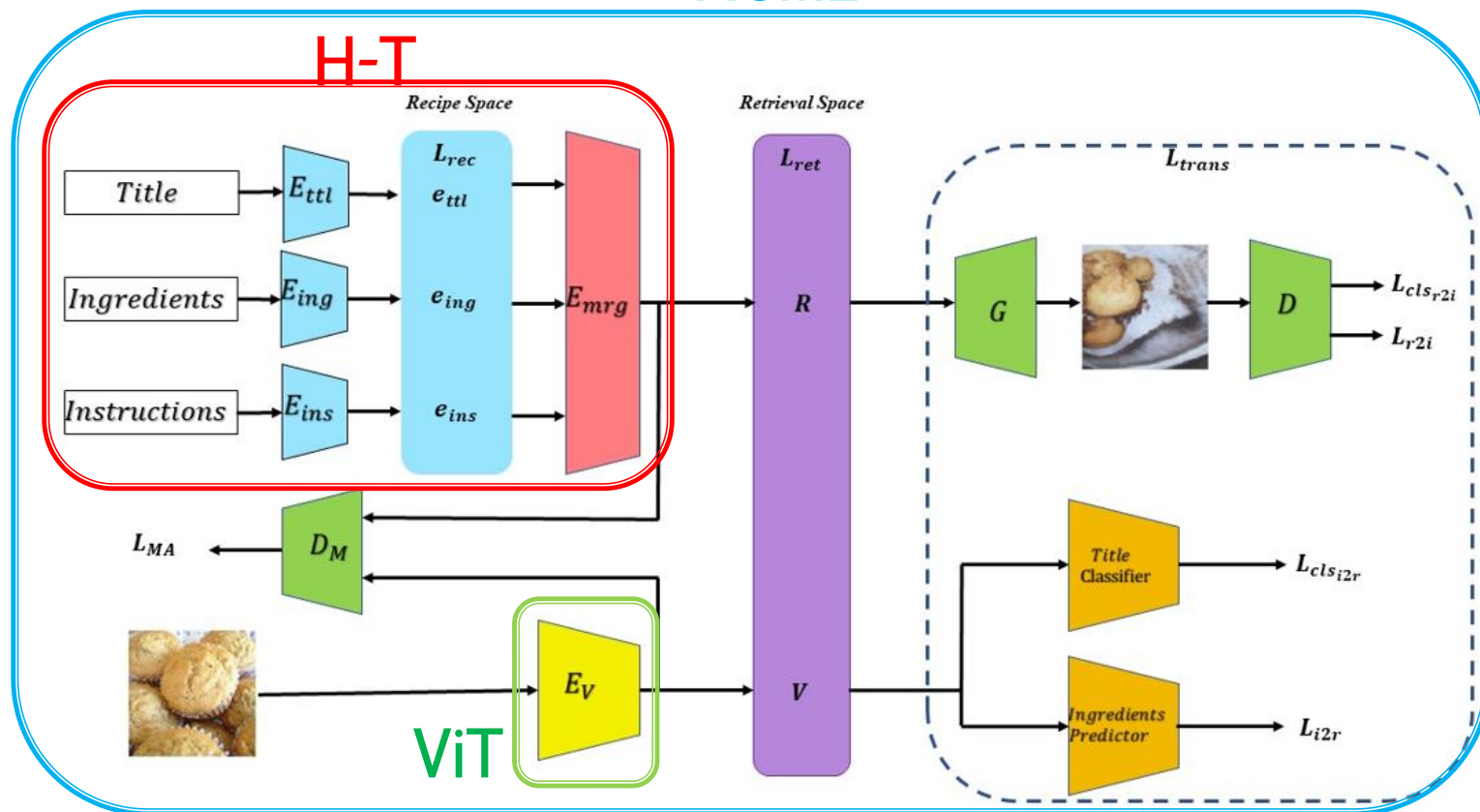
- 画像生成を行うACMEと、Transformerを利用したH-Tを融合することで、高品質なレシピ検索・画像生成のフレームワークを提案する
- 画像エンコーダーにVision Transformerを導入



提案手法一概要

- 画像生成を行うACMEベースに、H-Tのテキストエンコーダーを導入
- 画像エンコーダーにVision Transformer (ViT)を導入

ACME

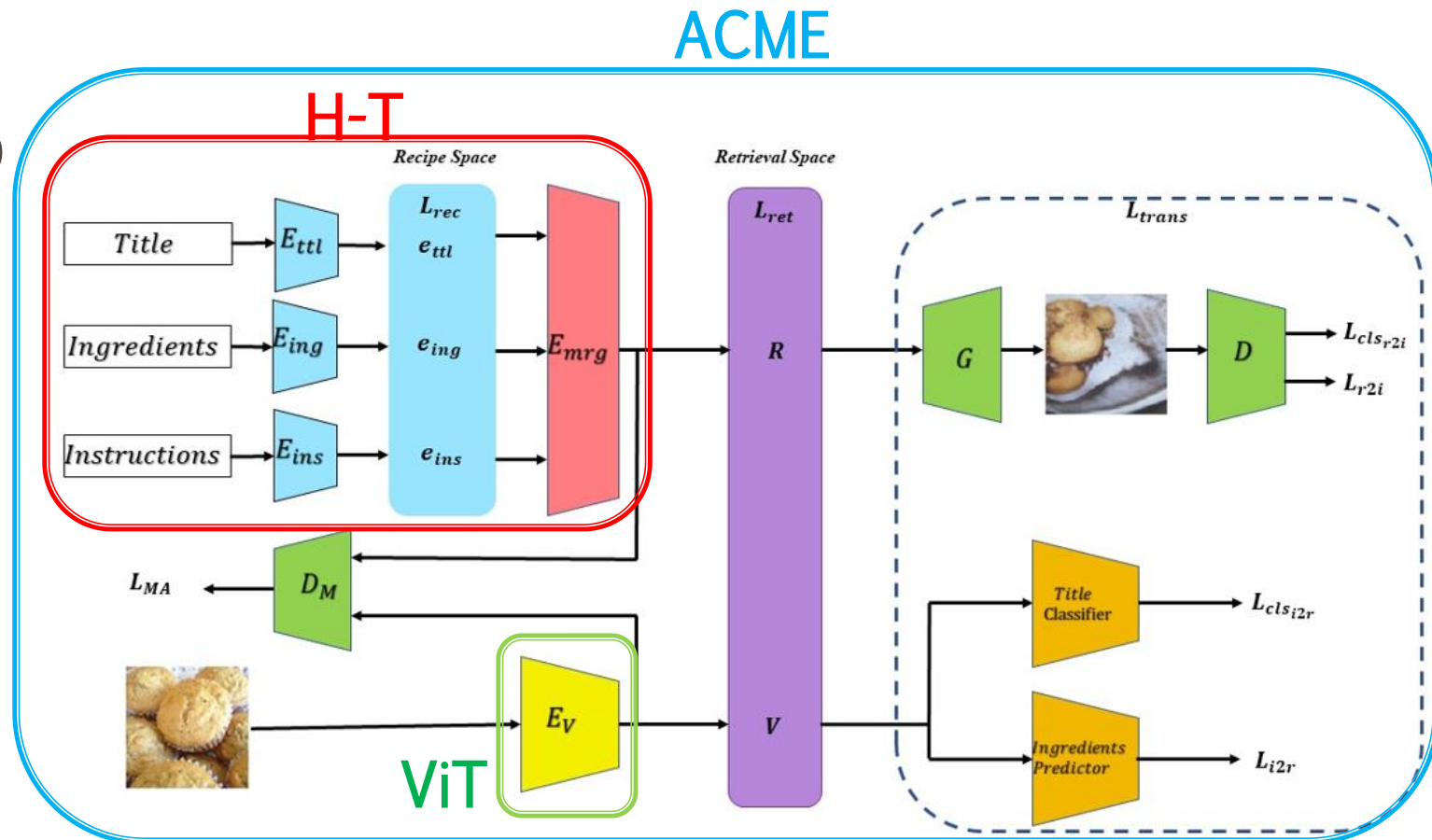


提案手法ー損失関数の概要

■ 4つの損失関数を使用

- 検索ロス: L_{ret} (クロスモーダル検索の基本 (Triplet Loss))
- レシピロス: L_{rec} (H-T)
- 敵対的ロス: L_{MA} (ACME)
- 一貫性ロス: L_{trans} (ACME)

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

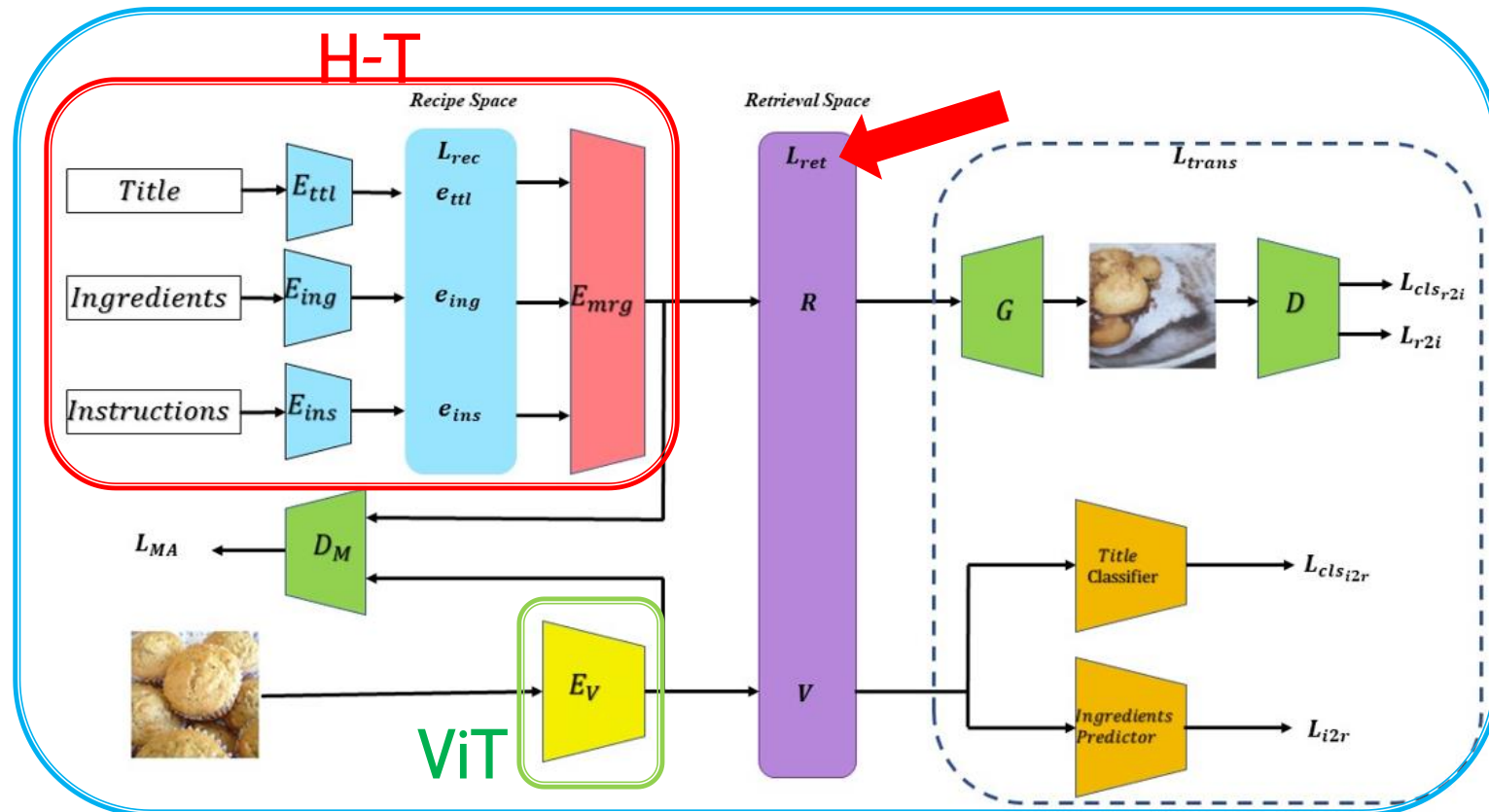


提案手法ークロスモーダル検索の精度にかかるとロス L_{ret}

- L_{ret} : ペアとなっているレシピ・画像の間の距離を縮め、そうではないペアの間の距離を伸ばす

ACME

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



提案手法ー自己教師あり学習のレシピロス L_{rec}

- L_{rec} (H-T): レシピテキストに含まれる
タイトル・成分・調理手順の間の補完的関係を探り、信頼性のある
エンベディングを作る

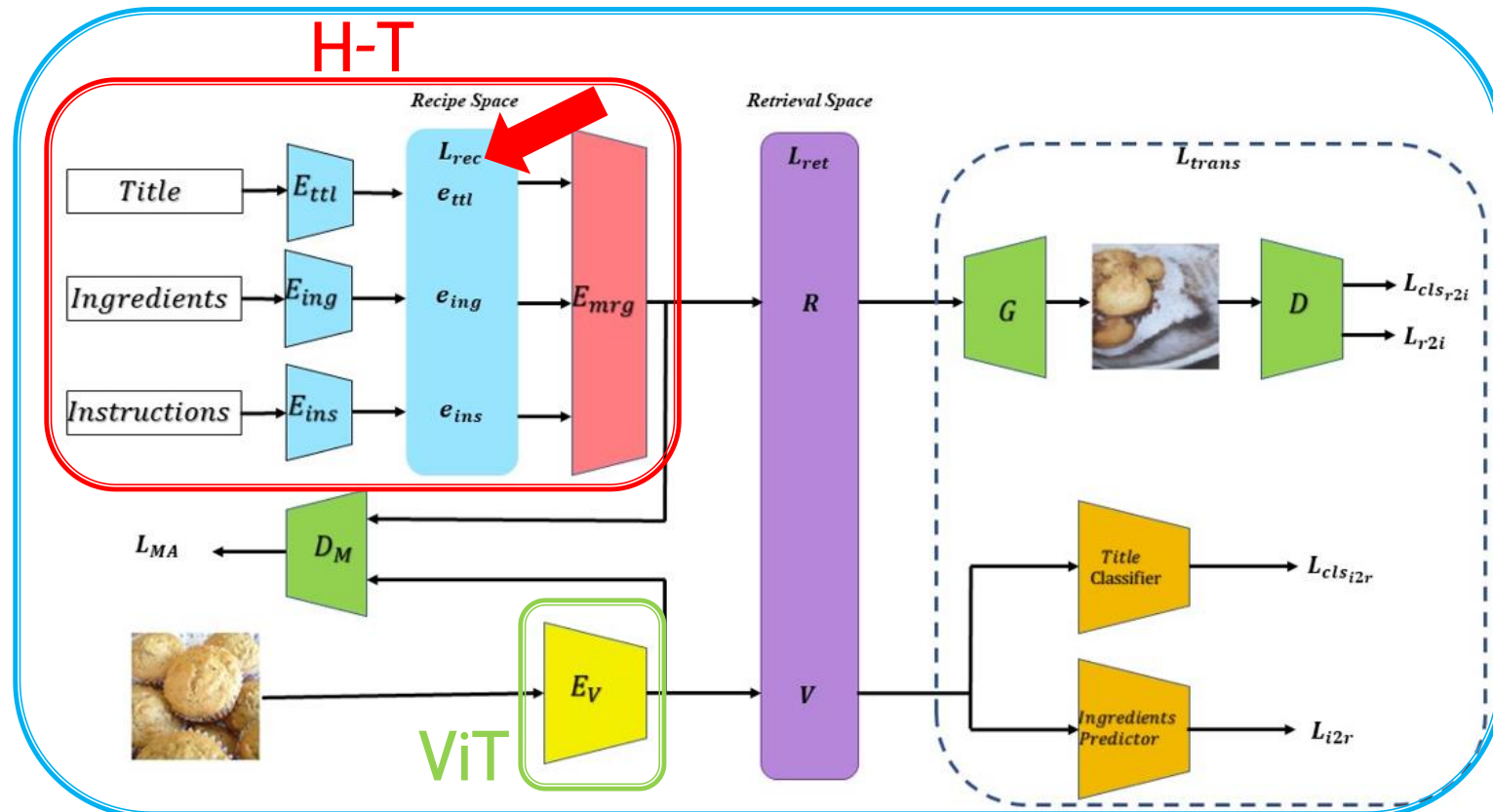
ACME

- 階層的Transformer構造

$$L_{rec} = \frac{1}{6} \sum_a \sum_b L_{bi}(a, b) \delta(a, b),$$

$a, b \in \{ttl, ing, ins\}$

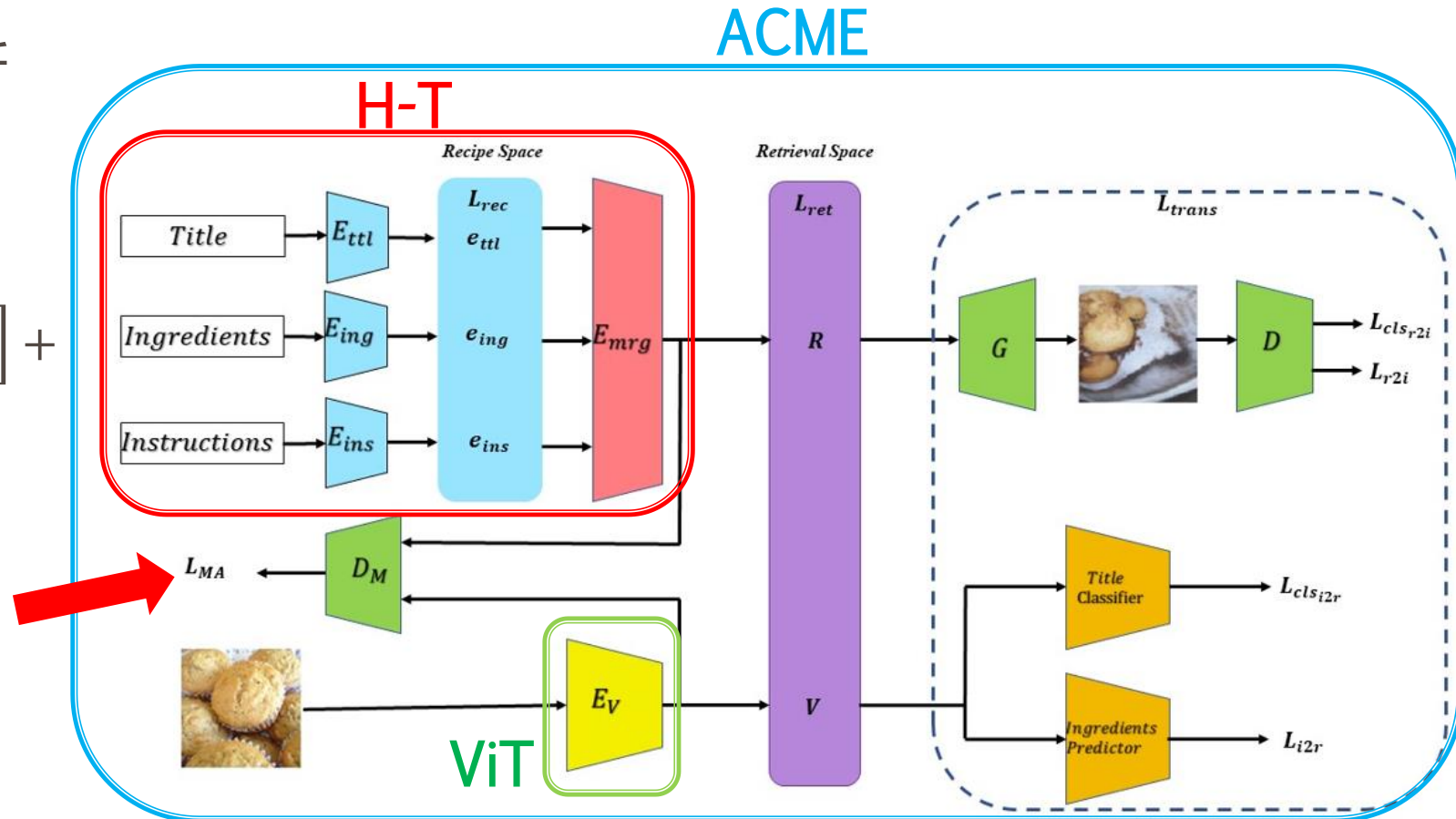
L_{bi} : 双方向のトリプレットロス



提案手法一二つのモダリティを融合するロス L_{MA}

- L_{MA} (ACME): レシピテキストとレシピ画像のモダリティ間のギャップの問題を緩和する Modality Alignment ロス
- 識別器でソースを判断

$$L_{MA} = E_{i \sim p(i)} \left[\log \left(1 - D_M(E_V(i)) \right) \right] + E_{t \sim p(t)} \left[\log \left(1 - D_M(E_T(t)) \right) \right]$$



提案手法ーエンベディングの一致性を確保するロス L_{trans}

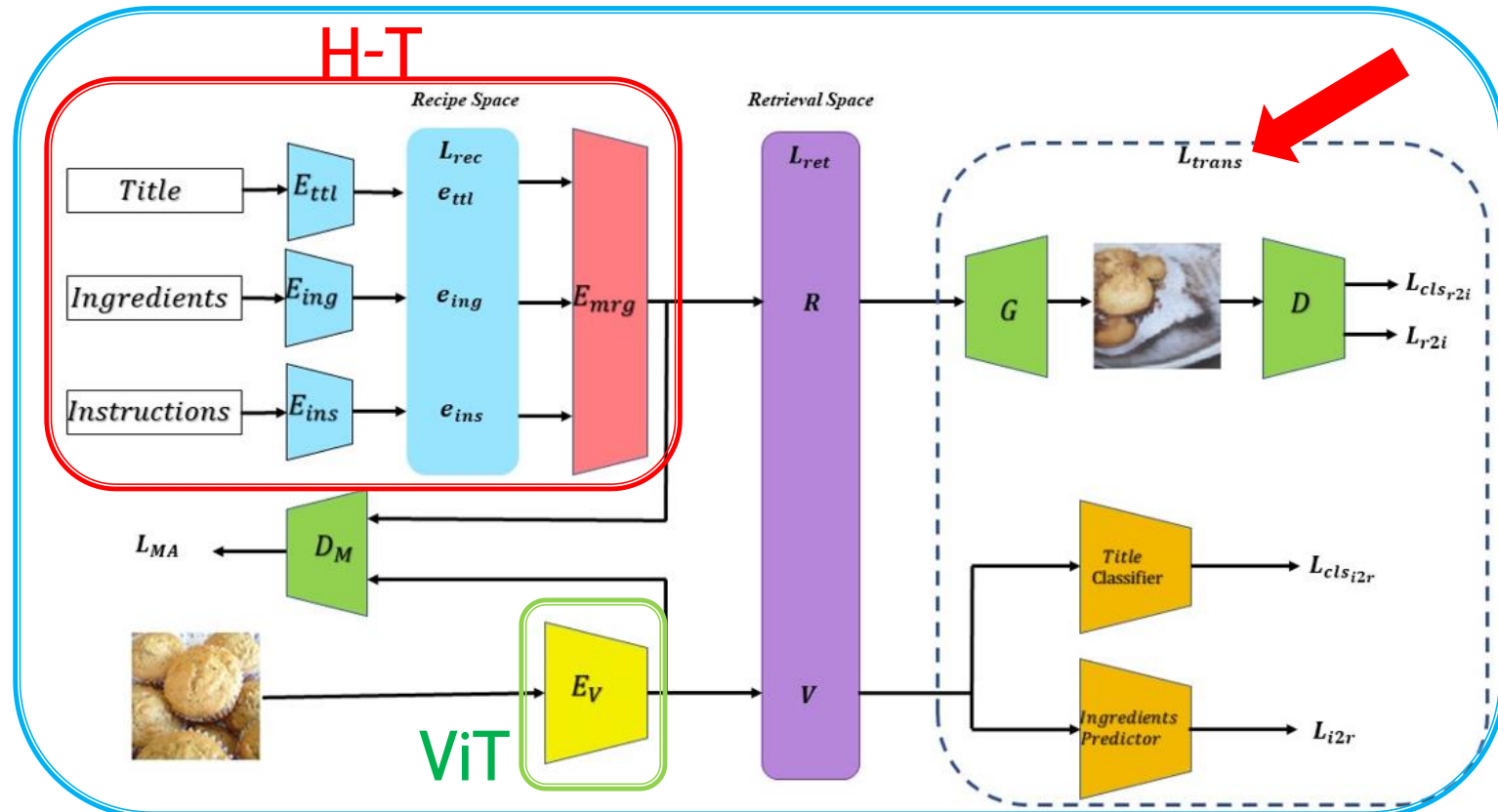
- $L_{trans}(ACME)$: 学習されたエンベディングを元の情報に復元することで、学習されたエンベディングが本来の情報を維持することを確認

ACME

$$L_{trans} = L_{trans_r} + L_{trans_i}$$

$$L_{trans_r} = L_{r2i} + L_{cls_{r2i}}$$

$$L_{trans_i} = L_{i2r} + L_{cls_{i2r}}$$



実験

- Recipe1Mの画像ーレシピペアを利用する
 - 学習データ： 238,999
 - 検証データ： 51,119
 - テストデータ： 51,303
- 残りの482,231の画像が含まれないレシピだけのサンプルも自己教師あり学習に利用
- バッチサイズ： 256 (ACME: 64、H-T: 128)
- 学習率: 0.0001
- 画像エンコーダーのバックボーン： ViT-B、 ResNet50

実験

- 検索精度・生成画像の質の確認
- バッチサイズ・各ロスが検索タスクにおける影響の確認
- 学習する損失関数:

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

- $\lambda_1 = 0.05$, $\lambda_2 = 0.005$, $\lambda_3 = 0.002$
- 1kサイズのR1を基準として, ベストモデルを更新する

実験結果（１）：レシピ検索

- Recipe1Mの画像ーレシピペアを利用する

	10k							
	Image-to-recipe				Recipe-to-image			
	medR	R1	R5	R10	medR	R1	R5	R10
ACME	6.7	22.9	46.8	57.9	6.0	24.4	47.9	59.0
H-T	4.0	27.9	56.4	68.1	4.0	28.3	56.5	68.1
RDE-GAN	3.5	36.0	56.1	64.4	3.0	38.2	57.7	65.8
Ours+ResNet	3.0	36.4	63.6	73.8	3.0	36.6	63.6	73.7
Ours + ViT	<u>2.0</u>	<u>44.3</u>	<u>70.9</u>	<u>79.7</u>	<u>2.0</u>	<u>44.6</u>	<u>70.8</u>	<u>79.5</u>

実験結果（1）：レシピ検索の例

- レシピ画像からレシピテキスト



#1



#2

Cuban Picadillo

成分:

- 1 tablespoon salt
- 1/2 teaspoon ground pepper

...

調理手順:

mash the salt, pepper and garlic together in a mortar...

Mexican Couscous

成分:

- 1 cup couscous
- 1 1/2 cups chicken broth

...

調理手順:

add everything together and cook on low heat until ...

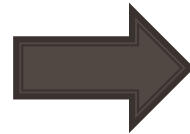
実験結果（1）：レシピ検索の例

- レシピテキストからレシピ画像

タイトル:
cuban picadillo

成分:
1 tablespoon salt
1/2 teaspoon ground pepper
...

調理手順:
1) mash the salt, pepper and garlic together in a mortar until well blended.
2) add the tomato sauce and let it simmer for a minute.
...



#1



#2



#3



実験結果（１）：レシピ検索の例ー上位になる検索結果

- WordCloud: ワードの頻度でそのワードの大きさを決める
ー複雑な文書データの図による可視化

本来のレシピテキスト

タイトル:

honey-broiled figs with ricotta

成分:

1 cup fresh ricotta

2 tablespoons honey

...

調理手順:

preheat the broiler and position a rack 6 inches from the heat.

in a food processor, combine the ricotta with the 2 tablespoons of honey and puree until very smooth.

...

可視化されたレシピテキスト



実験結果（1）：レシピ検索の例ー上位に

- WordCloud: ワードの頻度でそのワー
ー複雑な文書データの図による可視化



honeyが
よく使わ
れている

お菓子の
レシピっ
ぽい

Ricottaが
よく使わ
れている



本来のレシピテキスト

タイトル:

honey-broiled figs with ricotta

成分:

1 cup fresh ricotta

2 tablespoons honey

...

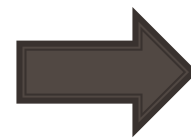
調理手順:

preheat the broiler and position a rack 6 inches from the heat.

in a food processor, combine the ricotta with the 2 tablespoons of honey and puree until very smooth.

...

可視化されたレシピテキスト



実験結果（1）：レシピ検索の例ー上位になる検索結果

- ViT-Bを画像エンコーダーに利用した検索例（検索対象の上位三位まで表示）
- 青枠：クエリー，緑枠：検索目標

目標レシピテキスト順位：1



目標レシピ画像順位：1



目標レシピテキスト順位：31



目標レシピ画像順位：27



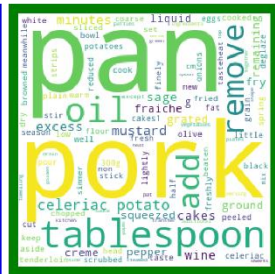
レシピ画像からレシピテキスト

レシピテキストからレシピ画像

実験結果（1）：レシピ検索の例ー上位になる検索結果

- ViT-Bを画像エンコーダーに利用した検索例（検索対象の上位三位まで表示）
- 青枠：クエリー，緑枠：検索目標

目標レシピテキスト順位：1



上位の検索対象
はお菓子の
レシピ

目標レシピ画像順位：1



上位の検索対象
はお菓子の画像

目標レシピテキスト順位：31



目標レシピ画像順位：27



レシピ画像からレシピテキスト

レシピテキストからレシピ画像

実験結果（1）：各ロスの影響

- $L_{ret}: RET$, $L_{rec}: REC$, $L_{MA}: MA$, $L_{trans_r}: R2I$, $L_{trans_i}: I2R$
- RET+MA+REC+TR*は, recipe onlyデータを学習に利用しない

	10k							
	Image-to-recipe				Recipe-to-image			
	medR	R1	R5	R10	medR	R1	R5	R10
RET+MA+R2I+I2R	2.0	41.9	68.5	77.5	2.0	42.7	68.9	77.7
RET+MA+REC	2.0	43.2	70.0	79.0	2.0	43.6	70.0	78.9
RET+MA+REC+I2R	2.0	43.5	70.3	79.0	2.0	44.0	70.4	79.1
RET+MA+REC+R2I	2.0	44.5	71.3	80.2	2.0	44.9	71.3	80.1
RET+MA+REC+R2I+I2R*	2.0	43.1	69.8	78.7	2.0	43.6	70.0	78.8
RET+MA+REC+R2I+I2R	2.0	44.3	70.9	79.7	2.0	44.6	70.8	79.5

実験結果（1）：バッチサイズの影響

- 画像エンコーダーがResnet50の場合
- バッチサイズが386以上に増大しても，精度の改良が限られている

	10k							
Batch size	Image-to-recipe				Recipe-to-image			
	medR	R1	R5	R10	medR	R1	R5	R10
64	3.7	31.2	58.3	69.2	3.4	31.9	58.7	69.5
128	3.0	35.0	62.6	73.6	3.0	35.4	62.6	73.1
256	3.0	36.4	63.6	73.8	3.0	36.6	63.6	73.7
386	3.0	34.9	61.2	71.0	3.0	34.8	61.0	70.8
512	3.0	37.0	63.9	73.7	3.0	37.3	63.8	73.3

実験結果（1）：バッチサイズの影響

- 画像エンコーダーがViT-Bの場合
- バッチサイズが512に増大しても，精度が上がる

	10k							
Batch size	Image-to-recipe				Recipe-to-image			
	medR	R1	R5	R10	medR	R1	R5	R10
64	3.0	34.0	60.8	71.1	3.0	34.0	60.5	70.8
128	2.0	41.3	68.3	77.5	2.0	41.5	68.5	77.7
256	2.0	44.3	70.9	79.7	2.0	44.6	70.8	79.5
386	2.0	46.3	73.0	81.2	2.0	46.6	73.0	81.2
512	2.0	47.1	73.4	81.6	2.0	47.4	73.5	81.5

実験結果（2）：画像生成—生成画像の質の比較

■ レシピテキストからの画像生成

クエリーレシピ

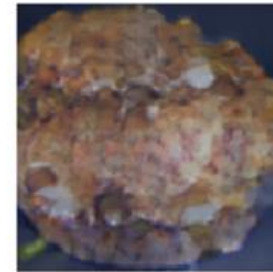
ターゲット画像

ACME

Ours + ResNet

Ours + ViT

spray washed homestyle french
half milk **cup** butter
pan potato remaining whole
cheese potatoes sprinkle
corn fries water
onion stir
ounce box minute cheddar



sifted water heaping muffin
spray cereal teaspoon together buds salt
mix cereal together buds salt
flour sugar extra
bake soda baking
bran minutes batter
crisco buttermilk light egg



butter teaspoon
unsalted pounds half
spoon **squash** minutes
packed sheet black
large ground squash salt
cut cooled water
sugar melted
light fresh baking total
medium brown kosher
caramelized tablespoons



実験結果（2）：画像生成

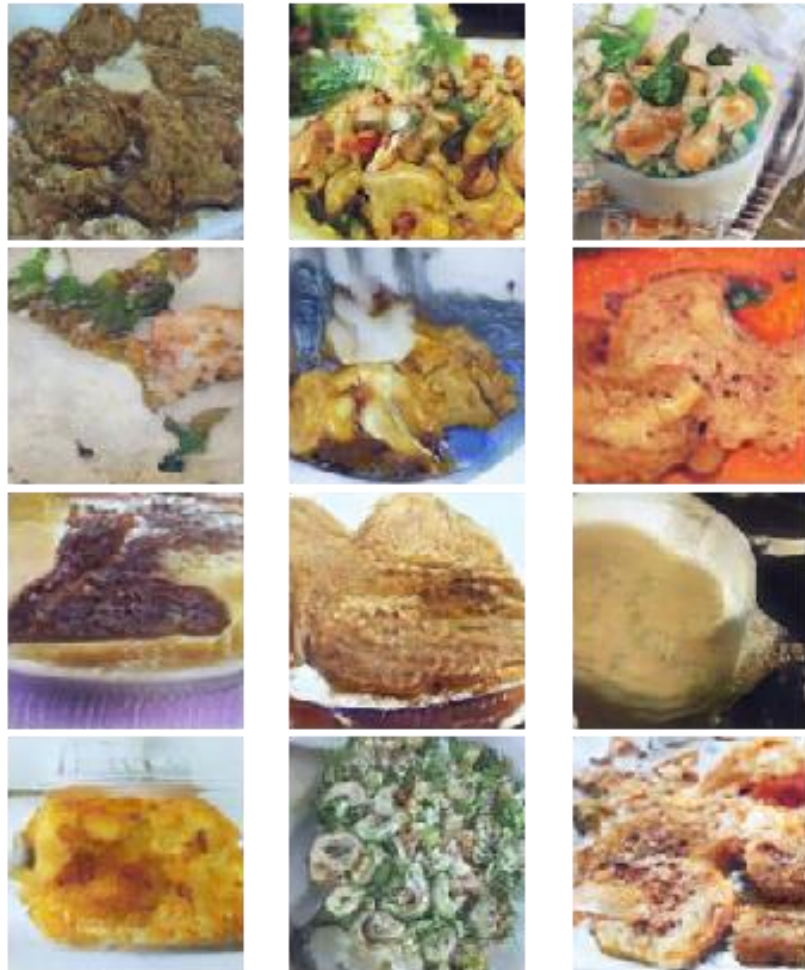
- 定量評価：FIDスコア（低ければ低いほど生成画像の質が良い）
- ViTよりResNet50の方が画像生成における良い結果が得られた
- 生成画像の質が大幅に改良された

手法	FID
ACME	30.7
Ours + ResNet	23.3
Ours + ViT	<u>27.1</u>

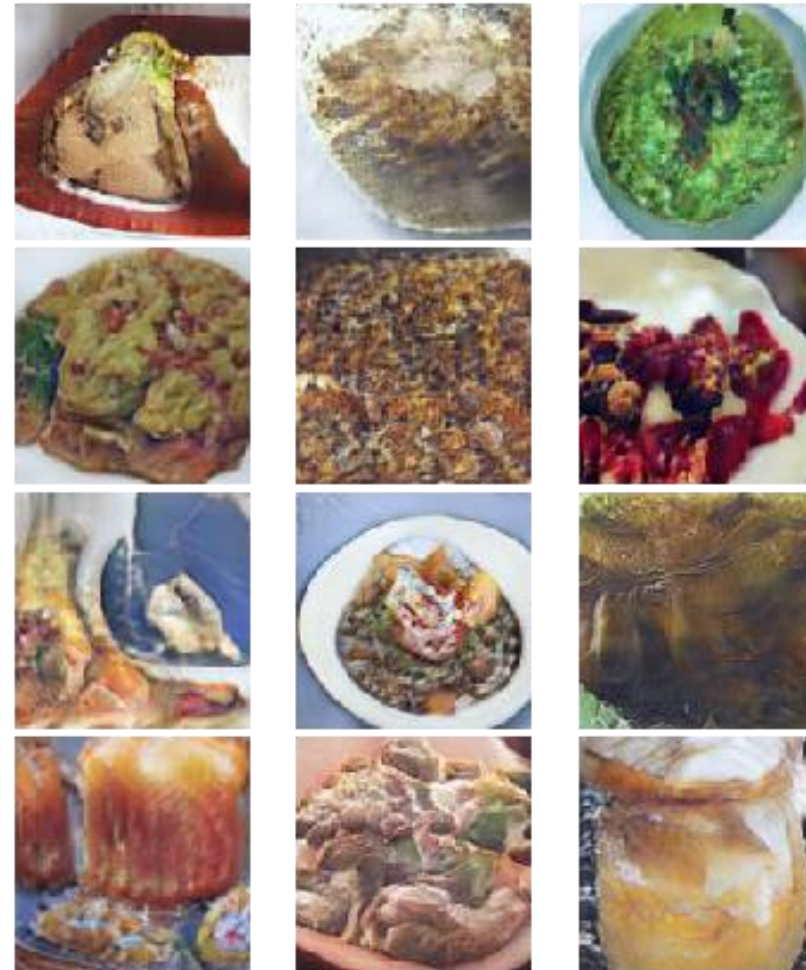
FID scoreが
25以下まで
改善

実験結果（２）：画像生成—生成画像の一部の例

ResNet50バックボーン



ViTバックボーン



最後に

まとめ

- ACMEとH-Tを利用したレシピ検索フレームワークを提案
- Vision Transformerの導入により、検索精度が大幅に向上
- 画像生成・レシピ検索の二つのタスクにおいて**最高精度**を達成

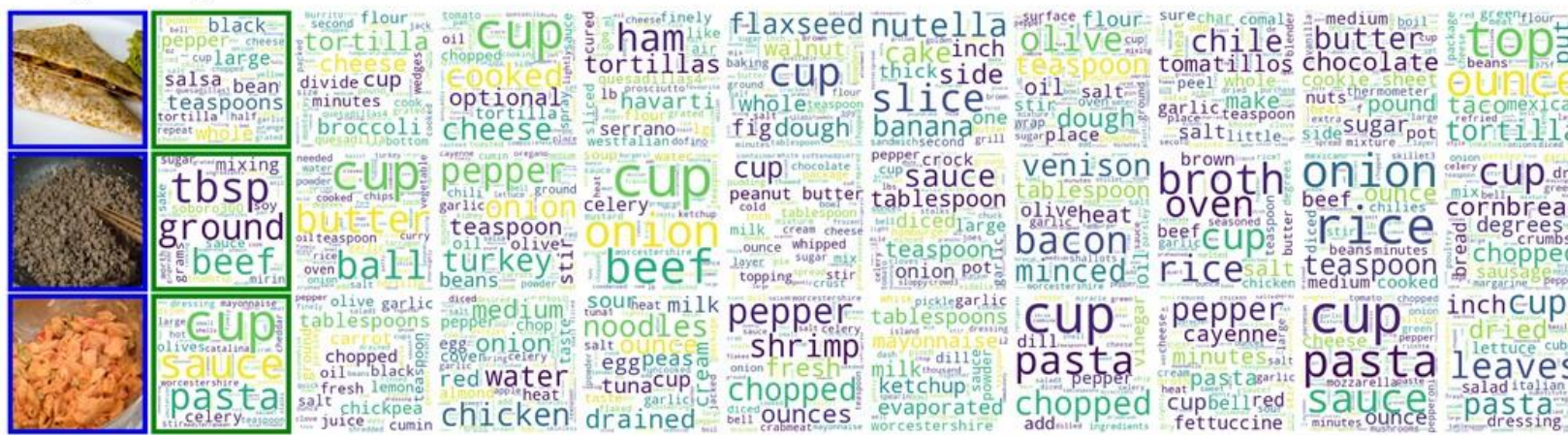
今後の課題

- RDE-GAN手法の適用を検討（画像からのノイズ除去など）



補足資料-ViTバックボーンへの検索例

Image2Recipe:



Recipe2Image:



補足資料-Resnetバックボーンでの検索例

Image2Recipe:



Recipe2Image:

