

# Self-Attention and Ingredient Attention Based Model for Recipe Retrieval from Image Queries

*Matthias Fontanellaz, Stergios Christodoulidis, Stavroula Mouggiakakou*

# Contents

---

- Introduction
- Method
- Database
- Results
- Conclusion

# Introduction – Motivation

Nutrient estimation is demanding task; for professionals as well as for dedicated algorithms

## **Causes:**

- deformation, occlusion, and color change of ingredients
- high intra-class variability
- low inter-class variability

## **Why should we extract information from recipes?**

- identify ingredients in complex meal compositions
- identify the way of preparation (bake, boil, fry, deep-fry)
- more accurate nutrient content estimation based on ingredient and preparation information

# Introduction – Goal

---

**Propose a flexible and reliable method for recipe retrieval based on image queries, where our main contributions are:**

- direct encoding of the instructions, ingredients and images during training
- the utilization of multiple attention mechanisms to gain insight in the networks decision making
- lightweight architecture without recurrency or necessity of pre-processing steps for instruction encoding

# Method – Retrieval Approach

## How do we extract information from food recipes?

### □ Training:

- the model learns recipe and image representations from data
- recipe compliant representations for both modalities are aligned in a shared latent space utilizing either the cosine distance base loss proposed by [1] or the triplet loss proposed by [2]

### □ Inference

- recipe retrieval based on similarity between query image's latent representation and recipe database
- use retrieved recipe as source of ingredient and preparation prediction

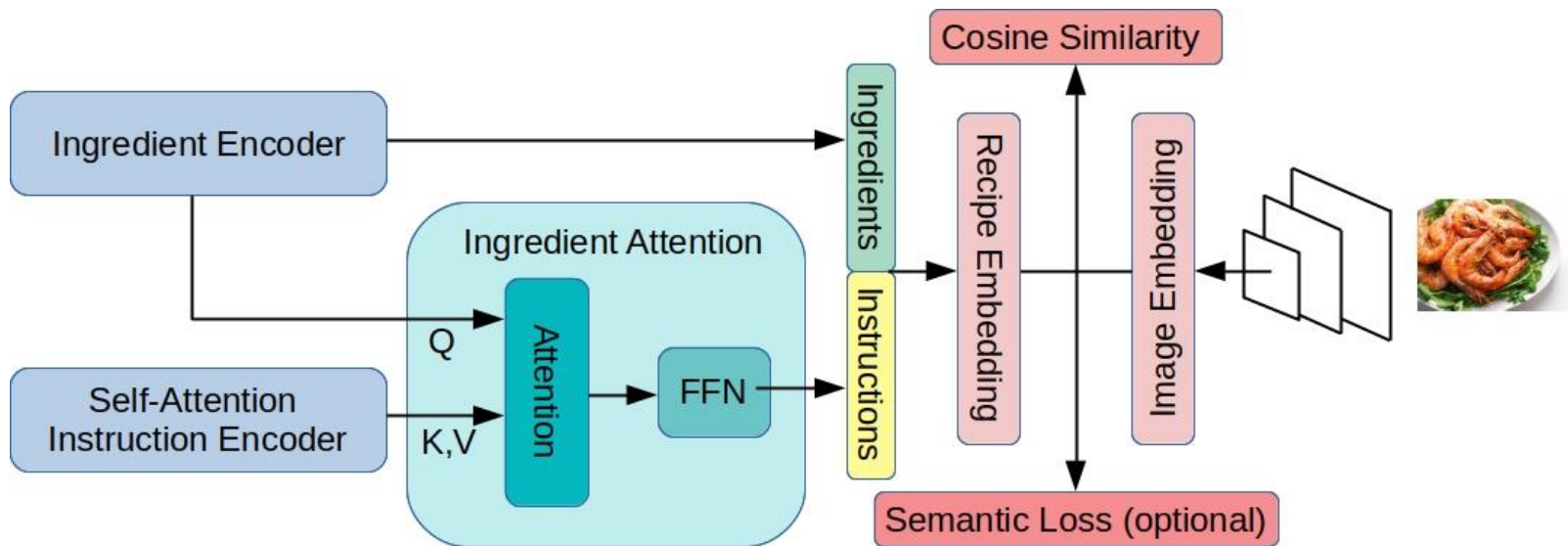
[1] J. Marin et al., "Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 2019.

[2] M. Carvalho et. al. "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings" *The 41st International ACM SIGIR Conference*, June 2018.

# Method - Network architecture (1 / 3)

Our network is a composition of

- a pretrained ResNet50
- a bi-directional LSTM for ingredient encoding
- a self-attention [1] based instruction encoder with Ingredient Attention (IA) block
- mapping layers and joint embedding space



[1] A. Vaswani et al. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008

# Method - Network architecture (2/3)

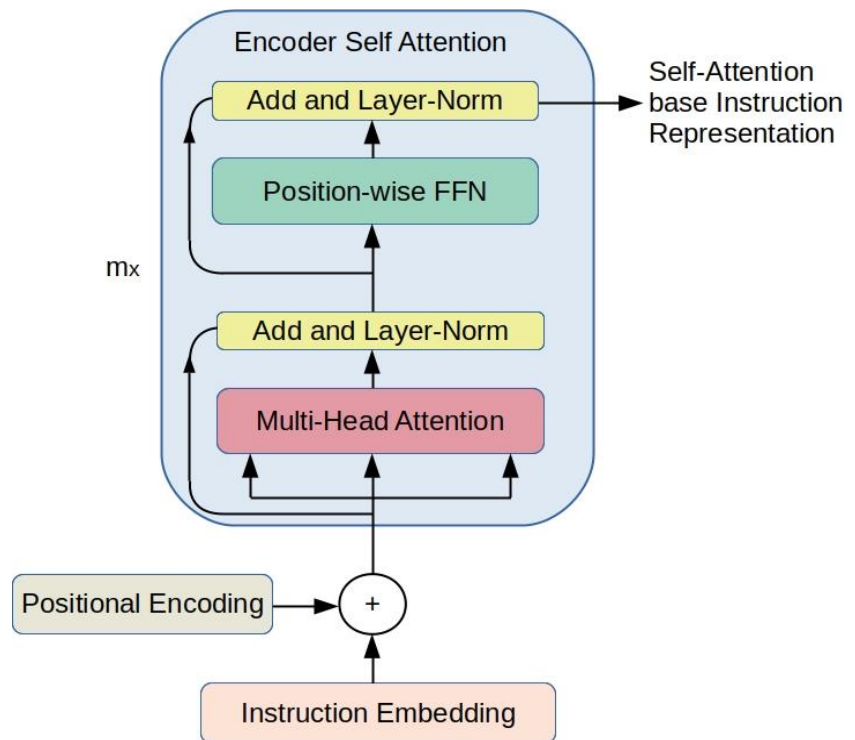
## Self-Attention based Instruction Encoder

Benefits of using Self-Attention:

- no recurrent units (lightweight)
- parallelizable on GPU
- no upstream instruction processing

Inputs and outputs

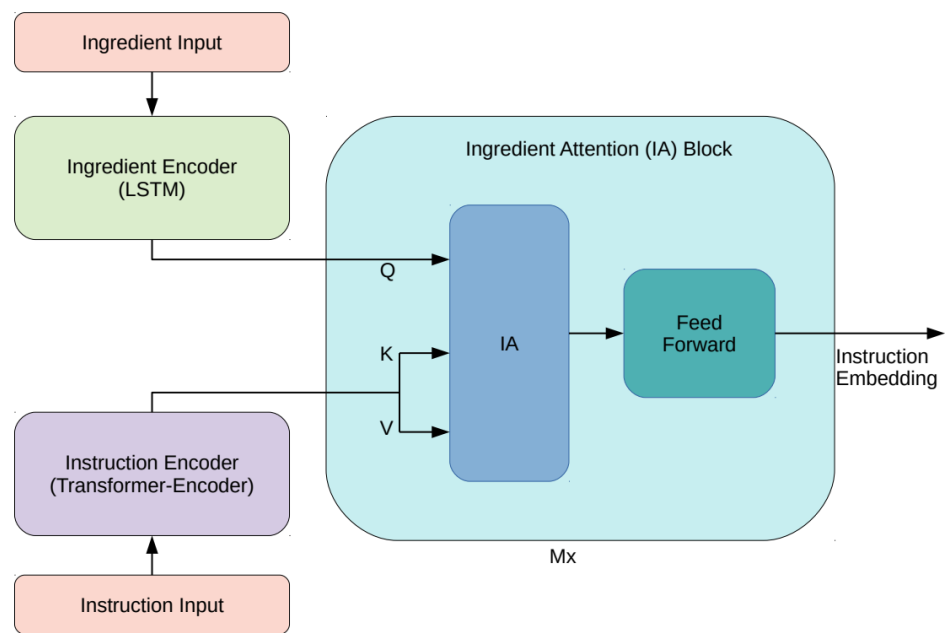
- limited the instructions word per recipe to 300
- each instruction word is represented by a 512-dimensional vector
- Size:  $B \times 300 \times 512$



# Method - Network architecture (3/3)

## Ingredient Attention: How to reduce Instruction Vectors and preserve diversity?

- Just one ingredient representation
- Use  $n$  different mapping matrices for generating  $n$  queries
- Calculate  $n$  ingredient attention weight sets based on similarity between query and key
- Calculate  $n$  instruction representation vectors
- Concatenate ingredient representation and instruction representation vectors





# Database

## Recipe1M+

- 1'029'720 recipes
  - Cooking instructions
  - Ingredients
  - 1048 food classes
- 13'735'679 images after augmentation

## Selection criteria

- $1 \leq \#Ingredients \leq 20$
- $\#instructions \leq 20$
- $1 \leq \#images$

## Exclusion of non-informative instructions

- empty lines or just punctuation from the counting

partition	original	refined	relative growth
train	238'399	254'238	6.6%
validation	51'119	54'565	6.7%
test	51'303	54'885	7.0%

*Number of samples for the train, validation and test set*

# Results (1 / 3)

## Evaluation metrics and results

- median rank
- recall percentage at top K













Image to Recipe					
		MedR	R@1	R@5	R@10
1k samples	Random	500.0	0.001	0.005	0.01
	JNE	5.0 ± 0.1	25.9	52.6	64.1
	AdaMine	3.0 ± 0.1	33.1	64.3	75.2
	IA	2.9 ± 0.3	34.6	66.0	76.6

*Evaluation results of JNE [1], AdaMine[2] and our Ingredient Attention based Model*

- [1] J. Marin et al., "Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 2019.
- [2] M. Carvalho et. al. "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings" The 41st International ACM SIGIR Conference, June 2018.

# Results (2/3)

## Recipe retrieval performance

	Sample 1	Sample 2	Sample 3
Query	 <ol style="list-style-type: none"> <li>1. marinade</li> <li>2. whole chicken</li> <li>3. salt</li> <li>4. garlic powder</li> <li>5. onion powder</li> </ol>	 <ol style="list-style-type: none"> <li>1. cooked chicken breasts</li> <li>2. grapes</li> <li>3. feta cheese</li> <li>4. celery</li> <li>5. dried onion flakes</li> <li>6. cashews</li> <li>7. etc.</li> </ol>	 <ol style="list-style-type: none"> <li>1. water</li> <li>2. crabmeat</li> <li>3. parmesan cheese</li> <li>4. lemon juice</li> <li>5. mayonnaise</li> <li>6. black pepper</li> </ol>
Top-1 Retrieval	 <ol style="list-style-type: none"> <li>1. whole chicken</li> <li>2. ground black pepper</li> <li>3. onion powder</li> <li>4. cayenne pepper</li> <li>5. garlic powder</li> <li>6. garlic cloves</li> <li>7. dried thyme</li> <li>8. butter</li> </ol>	 <ol style="list-style-type: none"> <li>1. cooked chicken breasts</li> <li>2. celery</li> <li>3. mayonnaise</li> <li>4. sour cream</li> <li>5. fresh rosemary</li> </ol>	 <ol style="list-style-type: none"> <li>1. 2 can Cream of Chicken</li> <li>2. 2 can Cream of mushroom</li> <li>3. Cooked chicken</li> <li>4. egg noodles</li> <li>5. onion</li> <li>6. ream cheese</li> <li>7. etc.</li> </ol>
Top-2 Retrieval	 <ol style="list-style-type: none"> <li>1. chicken drumsticks</li> <li>2. sea salt</li> <li>3. ground black pepper</li> <li>4. garlic powder</li> <li>5. onion powder</li> <li>6. dried oregano</li> <li>7. dried chipotle powder</li> <li>8. ground cumin</li> <li>9. coconut sugar crystals</li> </ol>	 <ol style="list-style-type: none"> <li>1. tuna</li> <li>2. celery</li> <li>3. sweet pickle</li> <li>4. mayonnaise</li> </ol>	 <ol style="list-style-type: none"> <li>1. fettuccine</li> <li>2. cooked chicken breasts</li> <li>3. cream of chicken soup</li> <li>4. sour cream</li> <li>5. onions</li> <li>6. dried parsley</li> <li>7. white wine</li> <li>8. cheddar cheese</li> <li>9. salt</li> </ol>
Top-3 Retrieval	 <ol style="list-style-type: none"> <li>1. whole chicken</li> <li>2. lime</li> <li>3. paprika</li> <li>4. cumin</li> <li>5. salt and pepper</li> </ol>	 <ol style="list-style-type: none"> <li>1. 1 can Canned tuna</li> <li>2. celery</li> <li>3. dill</li> <li>4. lemon juice</li> <li>5. mayonnaise</li> <li>8. salt and pepper</li> </ol>	 <ol style="list-style-type: none"> <li>1. chicken</li> <li>2. noodles</li> <li>3. mayonnaise</li> <li>4. onions</li> <li>5. salt</li> <li>6. cream of mushroom soup</li> <li>7. milk</li> <li>8. cheddar cheese</li> </ol>

# Results (3/3)

## Attention heatmaps

- we used  $n=2$  mapping matrices for generating 2 ingredient queries
- Ingredient Attention might help the network to handle high intra-class variability or low inter-class variability by
  - concentrating on shared ingredients (intra-class)
  - focusing on meal specific ingredients (inter-class)
- generic instructions such as preparing common ingredients are mostly ignored



**Ingredients**  
couscous  
skinless chicken breasts  
tomatoes  
onion  
extra virgin olive oil  
chicken broth  
carrots  
zucchini  
bay leaves  
whole cloves  
cinnamon  
turmeric  
chili powder  
salt and black pepper

### query 0

clean the chicken breasts with cold water lemon/lime and a pinch of salt  
rinse the chicken breasts and pat dry cut as desired saute the onion with  
oil until tender add to the the onion the chicken breasts cloves cinnamon stir-fry  
until the chicken turns to golden brown combine the chicken the tomatoes red chili  
pepper curry or turmeric carrots zucchini cook for five minutes add the chicken broth bay leaves  
salt and black pepper cook for 15 minutes meanwhile prepare the couscous following direction from the package  
in a serving dish place the chicken and sauce in the middle and the  
couscous around it serve it warm EOR

### query 1

clean the chicken breasts with cold water lemon/lime and a pinch of salt  
rinse the chicken breasts and pat dry cut as desired saute the onion with  
oil until tender add to the the onion the chicken breasts cloves cinnamon stir-fry  
until the chicken turns to golden brown combine the chicken the tomatoes red chili  
pepper curry or turmeric carrots zucchini cook for five minutes add the chicken broth bay leaves  
salt and black pepper cook for 15 minutes meanwhile prepare the couscous following direction from the package  
in a serving dish place the chicken and sauce in the middle and the  
couscous around it serve it warm EOR

least important

most important

# Conclusion

---

- In this work we introduced self-attention in the context of the recipe retrieval task
- Ingredient Attention model
  - outperforms baselines
  - converges faster
- Using the sets of ingredient attention weights, we gain insight into the networks thinking
- Our attention-based method is suitable for a standalone applications
- Elimination of noisy instructions increases the effective dataset sizes

Thank you for the attention!  
Questions?



matthias.fontanellaz@artorg.unibe.ch



DTR\_ARTORG



[https://www.artorg.unibe.ch/research/aihn/index\\_eng.html](https://www.artorg.unibe.ch/research/aihn/index_eng.html);  
<https://go-food.tech/>