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

goFOOD™
ARTIFICIAL INTELLIGENCE MEETS NUTRITION

Matthias Fontanellaz, Stergios Christodoulidis, Stavroula Mougiakakou

ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland

Background and Aims

Direct computer vision based-nutrient content estimation is a demanding task, due to deformation and occlusions of ingredients, as well as high intra-class and low inter-class variability between meal classes. In order to tackle these issues, we propose a system for recipe retrieval from images, where the recipe representation is generated with the aid of Ingredient Attention (IA). Utilizing self-attention [1] and IA, we are able to

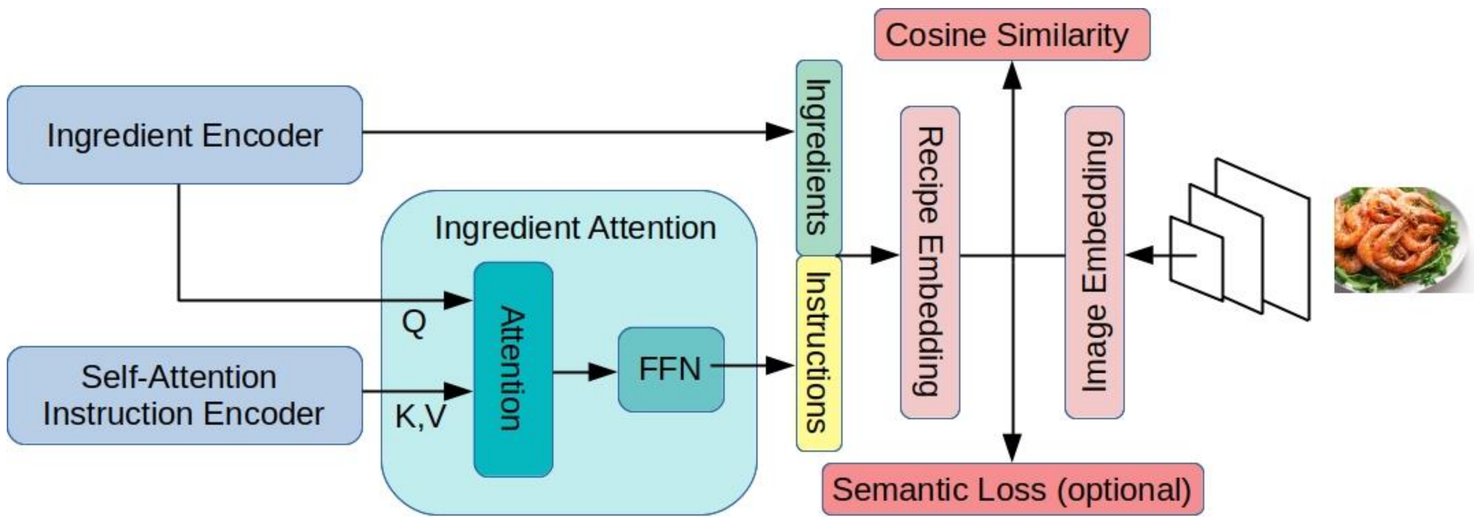
- directly process raw recipe text
- tackle the problem of high intra-class and low inter-class variability
- gain insight into which parts of instructions are of importance with respect to a certain ingredient list

Methodology

The main objective of our IA model is to align recipe representation and corresponding image representation (in terms of cosine similarity) in a joint embedding space.

Our network is a composition of a

- pretrained ResNet50
- bi-directional LSTM for ingredient encoding
- self-attention based instruction encoder with Ingredient Attention block, leading to a light-weight network structure
- mapping layers and joint embedding space



Overview of our proposed Ingredient Attention based alignment network

Recipe1M+ [2]

Key Data

- 1’029’720 recipes
- 13’735’679 images after image augmentation
- 1048 different food classes

Selection Criteria (for Recipes to be valid member of one of the train, validation or test set)

- 1 ≤ #Ingredients ≤ 20
- #instructions ≤ 20
- 1 ≤ #images

By excluding instructions which are composed solely of punctuation or spaces, we were able to increase the effective train, validation and test set sizes using the same selection criteria as described.

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

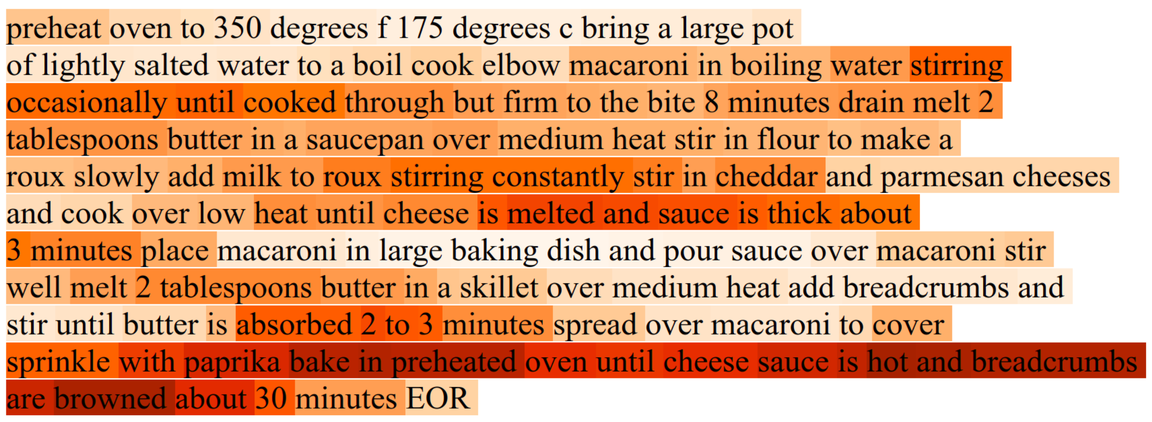
Results

We have (re-)implemented and trained all models utilizing Keras with Tensorflow back-end and our data-preprocessing

	MedR	R@1	R@5	R@10
JNE [2]	5.0±0.1	25.9	52.6	64.1
AdaMine [3]	3.0±0.1	33.1	64.3	75.2
IA	2.9±0.1	34.9	66.0	76.6

Comparison between our two baseline implementations (Joint Neural Embedding JNE and AdaMine) and our novel IA

Qualitative results, such as attention heatmaps are a useful tool for visualizing the networks thinking about how ingredients should be processed.



Foci on instructions for a mac and cheese recipe based on IA. Dark red means strong focus

Conclusions

- Utilizing self-attention empowers our model to directly process raw instruction input without any upstream instruction sentence embedding
- With IA, we are able to unveil internal focus in the text processing path by observing attention weights
- In future experiments we will demonstrate the power of IA for better handling high intra-class and low inter-class variability by not only evaluation scores (MedR, R@K) but also with qualitative results such as attention heatmaps

UNIVERSITÄT BERN
ARTORG CENTER
BIOMEDICAL ENGINEERING RESEARCH

INSELSPITAL
UNIVERSITÄTSSPITAL BERN
HOPITAL UNIVERSITAIRE DE BERNE
BERN UNIVERSITY HOSPITAL

References

1. A. Vaswani, et al. 2017. Attention Is All You Need. In Advances in Neural Information Processing Systems, pages 5998–6008
2. J. Marín, et al. 2018. Recipe1M: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. In IEEE Trans. Pattern Anal. Mach. Intell. (2019
3. M. Carvalho, et al. 2018. Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-Image Embeddings. ACM, New York, NY, USA, 35-44. DOI:https://doi.org/10.1145/3209978.3210036