How to make a Smørrebrød: Image-to-Recipe Retrieval

02501 Advanced Deep Learning in Computer Vision

Dimitrios Papadopoulos
Associate Professor, DTU Compute

10 October, 2023

Keywords/lectures: Transformers, Vision Transformers, image-to-text, CLIP

Figure 1: **Food Understanding using computer vision and NLP.** The goal of this project is to perform image-to-recipe retrieval tasks. The model will be able to retrieve a recipe (from a list of known recipes) given an image query and, in reverse, to retrieve an image (from a list of known images) given a text recipe.

1 Project description

Food is an important part of our lives. Imagine an AI agent that can look at a dish, recognize ingredients, and reliably reconstruct the exact recipe of the dish, or another agent that can read, interpret, and execute a cooking recipe to produce our favorite meal. Computer vision community has long studied image-level food classification [1, 2, 7, 8, 9, 11], and only recently focused on understanding the mapping between recipes and images using multi-modal representations [5, 10, 13, 14, 18].

Inspired by CLIP [12], the goal of this project is to retrieve a recipe (from a list of known recipes) given an image query and, in reverse, to retrieve an image (from a list of known images) given a text recipe. For this, the team will work on text and image retrieval by combining both modalities. In addition, the team will explore several additional textual information (title, instructions, ingredients) and analyze their impact.

2 Data

In this project, you will use the Food Ingredients and Recipes Dataset from kaggle [1]. The dataset consists of 13,582 images and each image comes with a corresponding title (the title of the food dish), a list of ingredients (the ingredients as they were scraped from the website), and a list of instructions (the recipe instructions to be followed to recreate the dish.)

3 Tasks

In this project, you could work on the following tasks:
Task 1: Image-to-recipe retrieval task. In this task, you are asked to build a model that is able to perform the image-to-recipe retrieval task. The model should consist of an image encoder (based on a standard CNN architecture [6, 16, 15] or even a visual transformer [4]) and a text encoder (based on a text transformer [17] or a BERT model [3]). You can get inspiration from the popular CLIP model from OpenAI [12]. The model should be trained with a triplet or a contrastive loss to learn a a joint embedding of text recipes and food images.
Task 2: Additional text modalities. In this task, you are asked to build on top of the model of Task 1 by adding extra text modalities (instructions and ingredients) when training the image-text model. You can either simply concatenate all text (title, title+ingredients, title+ingredients+instructions) or consider more advanced ways such as using one transformer for each text element (eg. BERT [3]) and then concatenate the features for all text modalities (note, you may need to project everything to a common feature space)
Task 3: Compare results with CLIP [12].
Task 4: ....

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. 2014.

[2] Xin Chen, Yu Zhu, Hua Zhou, Liang Diao, and Dongyan Wang. Chinesefoodnet: A large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743*, 2017.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Han Fu, Rui Wu, Chenghao Liu, and Jianling Sun. Mcen: Bridging cross-modal gap between cooking recipes and dish images with latent variable model. In *CVPR*, 2020.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016.

[7] Parneet Kaur, Karan Sikka, Weijun Wang, Serge Belongie, and Ajay Divakaran. Foodx-251: a dataset for fine-grained food classification. *arXiv preprint arXiv:1907.06167*, 2019.

[8] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *CVPR*, pages 5447–5456, 2018.

[9] Chang Liu, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, and Yunsheng Ma. Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. In *International Conference on Smart Homes and Health Telematics*, 2016.

[1] https://www.kaggle.com/pes12017000148/food-ingredients-and-recipe-dataset-with-images

[10] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):187–203, 2019.

[11] Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[13] Amaia Salvador, Erhan Gundogdu, Loris Bazzani, and Michael Donoser. Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning. In *CVPR*, 2021.

[14] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. 2017.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[16] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[18] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *CVPR*, 2019.